Genetic Epidemiology 32 : 647–657 (2008)

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

Azra Kurbasic[1]* and Ola Hössjer[2]

[1]*Centre for Mathematical Sciences, Department of Mathematical Statistics, Lund University, Box 118 SE-211 Lund, Sweden*
[2]*Department of Mathematics, Stockholm University, Stockholm, Sweden*

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 single nucleotide polymorphisms (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 (LD) 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 LD 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] Am. J. Hum. Genet. 61: 1179–1188) for pure linkage analysis. *Genet. Epidemiol*. 32:647–657, 2008.

Key words: hidden Markov model; linkage; association; linkage disequilibrium; SNPs

Contract grant sponsor: National Research School in Genomics and Bioinformatics; Contract grant sponsor: Swedish Research Council; Contract grant number: 626-2002-6286.
*Correspondence to: Azra Kurbasic, Centre for Mathematical Sciences, Department of Mathematical Statistics, Lund University, Box 118 SE-211 Lund, Sweden. E-mail: azra@maths.lth.se
Received 14 December 2006; Accepted 10 March 2008
Published online 15 May 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20339

INTRODUCTION

Mapping genes due to increased identical by descent (IBD) sharing has become widespread in 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, 2004]. 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 Abecasis and Wigginton [2005], Boyles et al. [2005], Broman and Feingold [2004], Browning et al. [2004], Huang et al. [2004] and Schaid et al. [2002, 2004].

There are different suggestions of what to do for dense marker maps. One possibility is to cluster SNPs, setting intracluster genetic distances to zero, thus treating each cluster as a single marker with no internal recombination [Abecasis and Wiggington, 2005]. Another possibility is to cluster SNPs, then retaining only afew SNPs from each cluster with low pairwise LD, and finally proceed with linkage analysis, assuming all SNPs are in LE and have nonzero genetic distances, [Matise et al., 2003; 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 moderates 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 the presence of the linkage component. Several methods for joint linkage and association analysis have recently been proposed by a number of authors, for instance [Cantor et al., 2005; Fulker et al., 1999; Göring and Terwilliger, 2000; Hössjer, 2005a; Jung et al., 2005; Li et al., 2005; Sham et al., 2000; Pérez-Enciso, 2003; Xiong and Jin, 2000]. The methods in these articles 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 article 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 Lth-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 $x_1 < \cdots < x_K$ 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 $1, \ldots, f$ and nonfounders labels $f+1, \ldots, n$. We let $b_{i,2k-1}$ be the maternal allele of individual k at x_i and $b_{i,2k}$ the paternal allele ($1 \le k \le n, 1 \le i \le K$). 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 $\mathcal{T}_i \subset \{1, \ldots, n\}$ 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 the data we might know the phase though if k is a nonfounder.) Notice that $T_i = \emptyset$ is possible, in which case we interpret x_i as a marker locus with completely missing data. Let

$$\boldsymbol{M} = (M_1, \ldots, M_k)$$

be marker data at all loci for all individuals and $m = 2(n-f)$ the number of meioses in the pedigree. Define $v_i = (v_{i1}, \ldots, v_{im})$ 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 lth meioses [Donnelly, 1983], and $b_i = (b_{i1}, \ldots, b_{i,2f})$ 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 $w_1, \ldots, w_K$. 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 $\boldsymbol{M}$ using the multipoint distribution

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

MARKOV MODEL FOR MULTIPOINT PROBABILITY

Usually, (1) is evaluated by means of the Lander-Green algorithm, assuming LE 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; Rabiner, 1989]. For this we need to assume no chiasma interference, so that $\{v_i\}$ becomes a Markov chain with transition probabilities

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

where θ_i is the recombination fraction between x_i and x_{i+1} and $|v_{i+1} - v_i| = \sum_{l=1}^m |v_{i+1,l} - v_{il}|$ 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 Lth order Markov chain, i.e.

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

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

Then

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

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) and (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 $B_i=(b_{i-L_i+1},\ldots,b_i)$ and

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

for $i=1,\ldots,K$. 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 ($d_1=\cdots=d_K=2$) is

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

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

For ease of presentation, we have assumed L to be constant. It is straightforward though to generalize the HMM algorithm to allow for haplotypes of varying length, with L smaller in regions of low LD.

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 $h_{ik}=h_k$, $k=1,\ldots,2f$. 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, $k=1,\ldots,f$. Let π_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

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

A phase switch of founder k corresponds to changing W to $(\pi_k(B),v+c_k)$, and this configuration cannot 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

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

of size 2^f. Here, $\pi_k^0(B)$ is the identity transformation, $\pi_k^1(B)=\pi_k(B)$ and $\circ$ denotes function composition. Hence, each element of (7) corresponds to switching founder phase of those founders k for which $\xi_k=1$.

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 linkage 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 than (6).

SCORE FUNCTIONS

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

$$\begin{aligned} H_{0i}: &\quad \tau \text{ unlinked to } x_i, \\ H_{1i}: &\quad \tau = x_i, \varepsilon \neq 0, \end{aligned} \quad (9)$$

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

$$\begin{aligned} H_{0i}: &\quad \tau = x_i, \varepsilon = 0, \\ H_{1i}: &\quad \tau = x_i, \varepsilon \neq 0. \end{aligned} \quad (10)$$

The latter formulation (10) was originally introduced by Whittemore [1996] for linkage tests statistics. The advantage of (10) over (9) is that the parameter ε 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 $b=(b_1,\ldots,b_{2f})$ and $v=(v_1,\ldots,v_m)$ be arbitrary founder allele and inheritance vectors and $w=(b,v)$. Write

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

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

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

[1] In Hössjer [2005a], such an association component was included, and the additional requirement "alleles at τ 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.

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

$$l(x_i, \varepsilon) = \log P_{i,\varepsilon}(\boldsymbol{M}|Y) = \log\left(\sum_w P(\boldsymbol{M}|w)P_{i,\varepsilon}(w)\right) \quad (12)$$

and assume that $\varepsilon = (\varepsilon_1, \varepsilon_2)$ involves two parameters. The first one, ε_1, corresponds to association tests and the second one, ε_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. $l'(x_i, \varepsilon)$ is the partial derivative of $l(x_i, \varepsilon)$ with respect to ε. For pure association tests, the derivative is taken with respect to one component, ε_1, so that $S = S_1$ is a scalar. For pure linkage tests, the derivative is taken with respect to ε_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 ε, so that $S = (S_1, S_2)$ is a vector. For more details on likelihoods leading to (13), see Hössjer [2003, 2005b], McPeek [1999] and Whittemore [1996] for linkage tests Clayton [1999], Shih and Whittemore [2002] and Whittemore and Tu [2000] for association tests and Hössjer [2005a] for combined association and linkage tests.

We consider association and linkage score functions of the form

$$S_1(w) = \sum_{k=1}^{n} \omega_k(b_{2k-1} + b_{2k}),$$

$$S_2(w) = \sum_{1 \le k < l \le n} \omega_{kl}\mathrm{IBD}_{kl}, \quad (14)$$

where IBD_{kl} is the number of alleles shared IBD by k and l and ω_k and ω_{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,

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

In general, test procedures based on S_1^{NF} are less powerful than those using S_1, since information is lost by conditioning on founder alleles. On the other hand, S_1^{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^{NF}), pure linkage ($S = S_2$) or combined linkage and association ($S = (S_1, S_2)$ or (S_1^{NF}, S_2)). The standardized version[2] of S at locus x_i is

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

where $\Sigma_i = \mathrm{Var}_0(S(w_i)|Y)$. Both μ_i and Σ_i are scalars for pure association and linkage score functions, but a 1×2 vector and a 2×2 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 at locus x_i.

For incomplete marker data, we do not observe $Z_i(w_i)$. Instead, we use the multipoint distribution (1) and define[3] the family score

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

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

$$P_{i,\varepsilon}(w) = P_{i,0}(w)(1 + Z_i(w)\gamma_i\varepsilon^{\mathrm{T}}), \quad \varepsilon = (\varepsilon_1, \varepsilon_2) \in \Theta, \quad (19)$$

where $\gamma_i = \mathrm{diag}(\gamma_{i1}, \gamma_{i2})$ is a given diagonal 2×2 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), ε_1 and ε_2 quantify strength of association and linkage, respectively, without having direct interpretation in terms of genetic model parameters. The parameter space Θ is defined by the requirements $\varepsilon_2 \ge 0$ and $P_{i,\varepsilon}(w) \ge 0$ for all w. Notice that we only allow for positive values of ε_2, corresponding to increased allele sharing among affected for binary phenotypes. On the other hand, both positive and negative values of ε_1 are allowed for, since either of the two alleles at x_i may be associated with the disease. It is possible to restrict Θ further as in pure association models ($\varepsilon_2 = 0$) and pure linkage models ($\varepsilon_1 = 0$). 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

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

with a constant depending on marker data $\boldsymbol{M}$, but not on ε. Hence, the score vector l_i' obtained from (20) is (18). The

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

[3]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 $P_i(\bar{w}) = P(\bar{w}_i = \bar{w}|\boldsymbol{M}) = \sum_{\bar{W}_i:\bar{w}_i=\bar{w}} P(\bar{W}_i|\boldsymbol{M})$.

diagonal entries of $\gamma_i = \mathrm{diag}(\gamma_{i1}, \gamma_{i2})$ 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 $(\boldsymbol{M}_1, Y_1), \ldots, (\boldsymbol{M}_N, Y_N)$. The total retrospective log likelihood for all families is

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

where $\bar{Z}_{ji}$ and $\gamma_{ji} = \mathrm{diag}(\gamma_{ji1}, \gamma_{ji2})$ are the likelihood score vectors and weight matrices of the jth family at locus x_i. The parameter space $\Theta = \cap_{j=1}^{N}\Theta_j$ is the intersection of the parameter spaces of all individual families. This definition of Θ guarantees that (19) is nonnegative for all N families, which makes all terms of (21) well defined.

Let

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

be the maximum likelihood estimator of ε at x_i and define the log LR statistic

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

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] and Table I.

A more easily computable approximation of T_i, which also motivates above-mentioned asymptotic distributions, is obtained by replacing $l(x_i, \cdot)$ in (22) by its second-order Taylor expansion around the origin. This quadratic approximation of $l(x_i, \cdot)$ is a function of $(\varepsilon_1, 0)$ for pure association tests $(T_i = T_{1i})$, $(0, \varepsilon_2)$ for pure linkage tests $(T_i = T_{2i})$ and $(\varepsilon_1, \varepsilon_2)$ for combined association and linkage tests $(T_i = T_{\mathrm{combined},i})$. Keeping in mind the restriction $\varepsilon_2 \geq 0$, this yields

$$T_{1i} \approx -(l_i')^2 / l_i'',$$

$$T_{2i} \approx -1_{\{l_i' \geq 0\}}(l_i')^2 / l_i'',$$

$$T_{\mathrm{combined},i} \approx wX^{\mathrm{T}}Xw^{\mathrm{T}} + vX^{\mathrm{T}}Xv^{\mathrm{T}}1_{\{Xv^{\mathrm{T}} \geq 0\}}, \tag{23}$$

where

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

$$l_i'' = -\sum_{j=1}^{N} \bar{Z}_{ji}\gamma_{ji}^2\bar{Z}_{ji}^{\mathrm{T}},$$

$X = l_i'(-l_i'')^{-1/2}$, X^{T} is the transpose of X, and $v = (v_1, v_2)$ and $w = (-v_2, v_1)$ are two orthogonal unit vectors such that v is proportional to the vector $(0,1)(-l_i'')^{-1/2}$.

TABLE I. Quantiles $\lambda_\alpha = F^{-1}(1-\alpha)$ of different distributions F

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

A simplified combined association and linkage test is obtained by summing (or taking some other weighted average) of separate association and linkage test statistics. One possibility is to use

$$T_{1i} + T_{2i} \tag{24}$$

as test statistics. In general, (24) differs from a combined linkage and association test $T_{\mathrm{combined},i}$, unless (i) the parameter space can be written as $\Theta = [c_1, c_2] \times [c_3, c_4]$ for some constants $c_1, \ldots, c_4$ and (ii) the likelihood in (21) factorizes as $P_{i,\varepsilon}(\boldsymbol{M}_j|Y_j) = P_{i,(\varepsilon_i,0)}(\boldsymbol{M}_j|Y_j) * P_{i,(0,\varepsilon_2)}(\boldsymbol{M}_j|Y_j)$ for each family $j = 1, \ldots, N$. However, in general both (i) and (ii) fail. Despite of this, (24) is often a good approximation of $T_{\mathrm{combined},i}$. Indeed, if l_i'' is diagonal (as for complete marker data), it follows from (23) that the quadratic approximation of $T_{\mathrm{combined},i}$ equals the sum of the quadratic approximations of T_{1i} and T_{2i}.

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 $i = L+1, \ldots, K$.

Various methods for estimating haplotype frequencies exists, see e.g. Douglas et al. [2001], Hodge et al. [1999] and Niu et al. [2002]. We use the software Haplotyper [Niu et al., 2002].

We will simulate marker data conditionally on phenotypes, see, for instance Boehnke [1986], Ott [1989], Ploughman and Boehnke [1989] and Terwilliger et al. [1993]. For one single family, simulation from $P_{i,\varepsilon}(\boldsymbol{M}|Y)$ can be achieved in two steps as follows:

1. Generate $\boldsymbol{b} = (b_1, \ldots b_K)$ and $(\boldsymbol{v} = v_1, \ldots v_K)$ from $P_{i,\varepsilon}(\boldsymbol{b}, \boldsymbol{v}|\boldsymbol{Y})$.
2. Compute $\boldsymbol{M} = (M_1, \ldots M_K)$ by spreading founder alleles from $\boldsymbol{b}$ according to $\boldsymbol{v}$ 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, $S_2 = S_{\mathrm{pairs}}$, the linkage score function is 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) for combined association and linkage test. In case of pure association and linkage $\gamma_{ji} = 1/\sqrt{N}$. The parents of each family are numbered 1,2, the children 3,4, the phenotype vector is denoted $Y = (Y_1, \ldots, Y_4)$, 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 II–IV. In the region studied we did not see much LD for markers more than 0.1 cM apart.

TABLE II. Information about the SNPs used in the study, located along Chromosome 9 and spanning approximatively 0.45 cM

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 III. Estimated population frequencies of haplotypes of length two [Niu et al., 2002].

	Markers								
Haplotype	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

Under the null hypothesis, we simulate marker data for 500 families with $Y = (1, 0, 1, 1)$. We generate founder haplotypes using estimated population frequencies from Table III. Simulation results from one data set are depicted in Figure 4. As test statistics we use $T_{\max} = \max_{1 \le i \le K} T_i$ for the given choice of association and/or linkage method. In order to compute significance levels, we generated marker data for $J = 1{,}000$ replicate data sets of the same type. Based on the replicate test statistics T_i^j and $T_{\max}^j$, $i = 1, \ldots, K$, $j = 1, \ldots, J$, *pointwise* and *regionwide* significance levels for threshold t are estimated as

$$\alpha_{pt}(t) = \frac{1}{JK} \sum_{j=1}^{J} \sum_{i=1}^{K} 1_{\{T_i^j \ge t\}}$$

and

$$\alpha(t) = \frac{1}{J} \sum_{j=1}^{J} 1_{\{T_{\max}^j \ge t\}},$$

respectively. Figures 1 and 2 depict pointwise and Figure 3 regionwide significance curves when $L = 0$ and $L = 1$.

Among the association LR tests (T_1 and T_1^{NF}) and linkage LR test (T_2), T_2 is by far most sensitive to incorrect LD

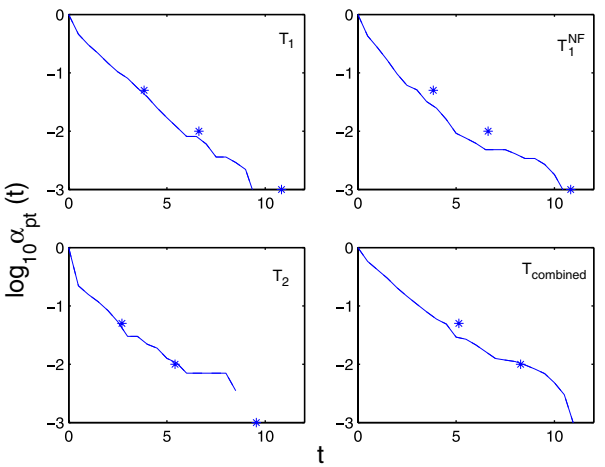

Fig. 1. Pointwise significance curve using four statistics when $Y = (1, 0, 1, 1)$, $\mathcal{T}_i = \{1, 2, 3, 4\}$ and $J = 1{,}000$. T_{combined} is the LR statistic when $S = (S_1^{\mathrm{NF}}, S_2)$. Nominal significance levels from Table I are included as $*$ in all subplots. Curves when $L = 0$ and $L = 1$ cannot be discerned.

TABLE IV. Table displaying the correlation coefficient r^2 of LD between markers

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

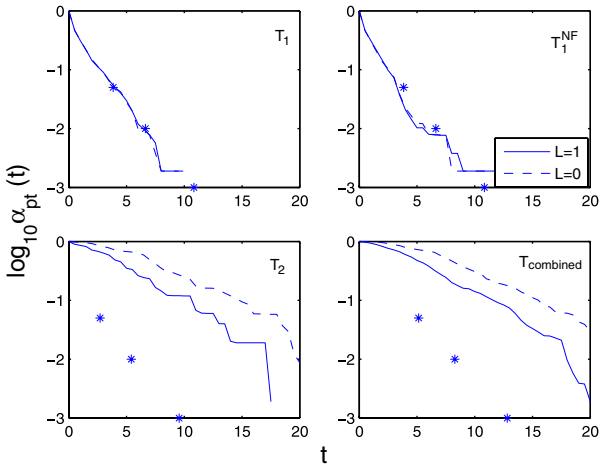

Fig. 2. Pointwise significance curve assuming either $L = 0$ or $L = 1$, using four statistics with $\mathcal{T}_{\text{combined}}$, Y and J as in Figure 1 and $\mathcal{T}_i = \{3, 4\}$. Nominal significance levels from Table I are included as $*$ in all subplots.

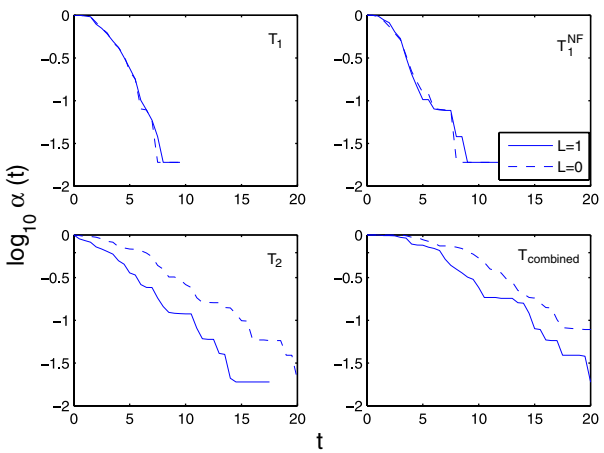

Fig. 4. LR statistic T for one data set simulated under H_0 and with Y, $\mathcal{T}_i$ and T_{combined} as in Figure 2.

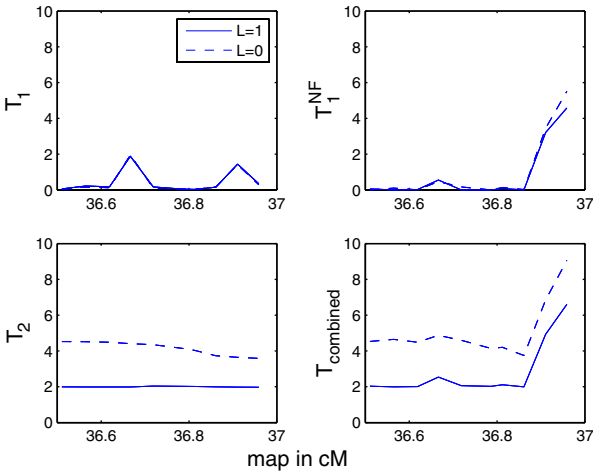

Fig. 3. Regionwide significance curve with T_{combined}, Y, $\mathcal{T}_i$ and J as in Figure 2.

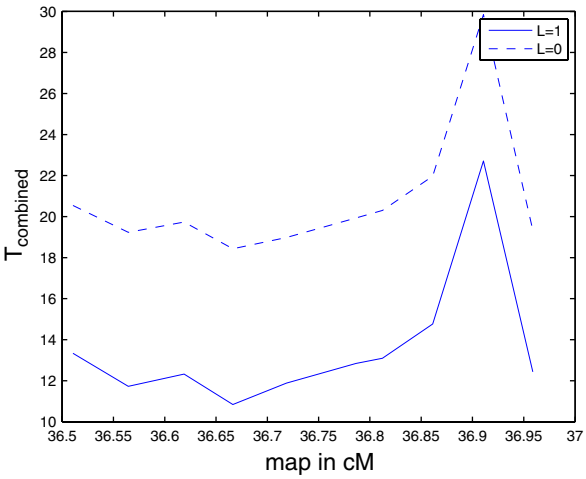

Fig. 5. LR statistic T for combined association and linkage analysis for data simulated under H_1 with marker 9 as disease locus and penetrance values $(f_0, f_1, f_2) = (0.10, 0.11, 0.12)$. $S = (S_1, S_2)$, $Y = (1, 0, 1, 1)$ and $\mathcal{T}_i = \{3, 4\}$.

assumptions in the analysis, followed by T_1^{NF} and T_1. The reason why T_1^{NF} is somewhat 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 4. For T_1 there is no difference between analyses under LD and LE assumption when parents are unaffected or have unknown phenotypes. This holds regardless of which parents 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 have been genotyped for markers. There is inflation of T_2 when one parent is untyped for markers (data not shown), but even more so when both parents are untyped. 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.

As seen from Figures 1 and 2, the nominal P-value approximation from Table I is very good when $\mathcal{T}_i = \{1, 2, 3, 4\}$ and for pure association tests when $\mathcal{T}_i = \{3, 4\}$ but poor for pure linkage and combined linkage and association tests. These discrepancies between the nominal and empirical P-values deserve further study.

When simulating under the alternative hypothesis we have 1,000 sib pair families and choose one marker as disease locus, see Figure 5. The genetic model corresponds to a low penetrant disease with a rather small (7.8%) disease allele frequency. 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 HMM, 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.

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

$$S = S(W) \tag{25}$$

is a score function depending on marker data from all L_i loci included in $W = (B, v)$. We conjecture that (25) is preferable over (14) in particular for coarser marker maps such that the disease locus is located between two neighboring markers. It this case, the founder marker haplotypes 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 causal variant, the disease alleles are more associated to phenotypes than any other (combination of) marker alleles, indicating that a single locus association component of S will work well. When S_1 is single locus, information from other markers 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.)

Our approach, including the computational savings, can be extended to handle different recombination rates between men and women. The crucial point is that the transition probabilities (2) of the inheritance vector Markov chain is generalized so that the product structure (A8) in the Appendix is retained.

The present implementation of the algorithm is computationally intensive, in spite of the speedups. Indeed, in order to analyze N pedigrees of the same kind along a chromosome segment with K SNPs, the number of operations to compute T_i at all markers is proportional to $NK \cdot \mathrm{CC}$, where CC, the proportionality constant in (8) is given in Table V for a number of different pedigrees. It is evident that the computational complexity is affected primarly by the number of founders. For a small pedigree with two founders, a short chromosomal region (say $K = 10$) and a data set with $N = 1{,}000$ individuals, values of L up to 3–4 seem to be within reach. However, if J Monte Carlo replicate data sets (with the same K and N) are generated in order to compute significant levels, the number of operations is proportional to $JNK \cdot \mathrm{CC}$. Then if $J = 1{,}000$ values of L up to 2 seem more reasonable.

TABLE V. Proportionality constant (8) of computational complexity (CC), for one pedigree when $K = 1$

Pedigree	n	f	m	CC
Trio	3	2	2	$6 \cdot 16^L$
Sib pair	4	2	4	$32 \cdot 16^L$
Sib trio	5	2	6	$160 \cdot 16^L$
Sib quartet	6	2	8	$768 \cdot 16^L$
Uncle-nephew	6	3	6	$96 \cdot 64^L$
First cousins	8	4	8	$256 \cdot 256^L$
Second cousins	12	6	12	$1536 \cdot 4096^L$

For a data set consisting of N identical pedigrees, analyzed at K markers, the computational complexity is thus proportional to NK CC.

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 Abecasis et al. [2002] and Gudbjartsson et al. [2000]. (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].

The perhaps most natural extension for real data is to combine (4) and (5). At regions where neighboring SNPs are tightly linked they may be grouped into clusters and treated as a single marker, with L small. In other regions where neighboring SNPs have looser linkage but still show some degree of LD, it seems more reasonable to retain individuals SNPs as markers and use a larger L. In this way, it is possible to augment the grouping approach of Abecasis and Wigginton (2005).

A great challenge is to develop faster versions or approximations of our methodology such that genome-wide scans based on all polymorphic SNPs become feasible. For instance, if $K = 6 \times 10^6$, SNPs are used along the human genome, at an average distance of 0.5 kb, and if LD 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. The extent to which such a reduction of L affects significance level is an interesting topic for future research.

However, for pure linkage tests, a much coarser SNP density is sufficient to yield high information content at all loci and thus a faster algorithm.

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. We are grateful for the reviewers' helpful comments.

REFERENCES

Abecasis GR, Wigginton JE. 2005. Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. Am J Hum Genet 77:754–767.

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. 2002. Merlin—apid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97–101.

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, 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, Feingold E. 2004. SNPs made routine. Nat Methods 1:104–105.

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. Nat Genet 29:229–232.

Donnelly KP. 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. Nat Genet 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 genome-wide 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. Nat Genet 25:12–13.

Hodge SE, Boehenke M, Spence MA. 1999. Loss of information due to ambiguous haplotyping of SNPs. Nat Genet 21:360–361.

Hössjer O. 2003. Determining inheritance distributions via stochastic penetrances. J Am Stat Assoc 98:1035–1051.

Hössjer O. 2005a. Combined association and linkage analysis for general pedigrees and genetic models. Stat Appl Genet Mol Biol 4: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, 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.

Matise TC, Sachidanandam R, Clark AG, Kruglyak L, Wijssman E, Kakol J, Buyske S, Chui B, Cohen P, de Toma C, Ehm M, Glanowski S, He Chunsheng, Heil J, Markianos K, McMullen I, Pericak-Vance MA, Silbergleit A, Stein L, Wagner M, Wilson AF, Winick JD, Winn-Deen ES, Yamashiro CT, Cann HM, Lai E, Holden AL. 2003. A 39-centimorgan-resolution human single-nucleotide polymorhism linkage map and screening set. Am J Hum Genet 73:271–284.

McPeek 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 86:4175–4178.

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, Hebbring 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 Epidmiol 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 Res 15:1592–1593.
Whittemore A. 1996. Genome scanning for linkage: an overview. Biometrics 59:704–716.
Whittemore A, Halpern J. 1994. A class of tests for linkage using affected pedigree members. Biometrics 50:118–127.
Whittemore AS, Tu I-P. 2000. Detection of disease genes by use of family data. I. Likelihood-based theory. Am J Hum Genet 66:1328–1340.
Xiong M, Jin L. 2000. Combined linkage and linkage disequilibrium mapping for genome screens. Genet Epidemiol 19:211–234.

APPENDIX

FORWARD-BACKWARD ALGORITHM

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

$$P(W_{i+1}|W_i) = P(v_{i+1}|v_i)P(B_{i+1}|B_i) = P(v_{i+1}|v_i)\prod_{k=1}^{2f} P(b_{i+1,k}|h_{ik}), \tag{A1}$$

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

$$P(W_i) = 2^{-m}\prod_{k=1}^{2f} P(h_{ik}), \tag{A2}$$

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

Introduce

$$\boldsymbol{M}_i^- = (M_1, \ldots, M_i),$$

$$\boldsymbol{M}_i^+ = (M_{i+1}, \ldots, M_K),$$

$$\alpha_i(W) = P(\boldsymbol{M}_i^-, W_i = W),$$

$$\beta_i(W) = P(\boldsymbol{M}_i^+|W_i = W),$$

for $i = 1, \ldots, K$. 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

$$P(W_i = W|\boldsymbol{M}) \propto \alpha_i(W)\beta_i(W), \tag{A3}$$

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

$$\alpha_i = (\alpha_i(W)),\ \ 1 \times |\mathcal{W}_i| \text{ forw. prob. vector},$$

$$\beta_i = (\beta_i(W)),\ \ 1 \times |\mathcal{W}_i| \text{ backw. prob. vector},$$

$$P_i = (P_i(W)),\ \ 1 \times |\mathcal{W}_i| \text{ abs. prob. vector},$$

$$Q_i = (Q_i(W, W')),\ \ |\mathcal{W}_i| \times |\mathcal{W}_{i+1}| \text{ transition matrix},$$

$$D_i = \text{diag}(D_i(W)),\ \ |\mathcal{W}_i| \times |\mathcal{W}_i| \text{ diagonal matrix}, \tag{A4}$$

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

$$\alpha_{i+1} = \alpha_i Q_i D_{i+1},\ \ i = 1, \ldots, K-1,$$

$$\beta_i^{\mathrm{T}} = Q_i D_{i+1}\beta_{i+1}^{\mathrm{T}},\ \ i = K-1, \ldots, 1, \tag{A5}$$

where β_i^{T} is the transpose of β_i, with initial conditions

$$\beta_K = (1, \ldots, 1),$$

$$\alpha_1 = P_1 D_1.$$

Q_i and P_i are deduced from (A1) and (A2) and

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

Here, (b, v) is the part of $W = (B, v)$ containing founder alleles at the rightmost locus of B only and $(b, v) \to M_i$ 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 (30) 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 neighboring pairs of markers, giving a total complexity (6).

REDUCED COMPUTATIONAL COMPLEXITY

Let $\overline{\mathcal{W}}_i$ 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 $1 \times |\overline{\mathcal{W}}_i|$, $|\overline{\mathcal{W}}_i| \times |\overline{\mathcal{W}}_i|$ and $|\overline{\mathcal{W}}_i| \times |\overline{\mathcal{W}}_{i+1}|$.

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

$$\overline{Q}_i(\overline{W},\overline{W'}) = P(\overline{W}_{i+1}=\overline{W'}|\overline{W}_i=\overline{W}) = \sum_{W'\in\overline{W}} Q_i(W,W'), \tag{A6}$$

provided the RHS of (A6) 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

$$\overline{P}_i(\overline{W}) = P(\overline{W}_i=\overline{W}) = 2^f P_i(W),$$

$$\overline{D}_i(\overline{W},\overline{W}) = P(M_i|\overline{W}_i=\overline{W}) = 1_{\{(b,v)\to M_i\}}, \tag{A7}$$

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 nonreduced state space.

It turns out that the recursive computation (30) 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

$$Q_i = Q_{i1}Q_{i2}\dots Q_{i,2f+m}. \tag{A8}$$

For $k=1,\dots,2f$, Q_{ik} updates the kth 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 jth 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

$$\overline{Q}_{ik}(\overline{W},\overline{W}') = \sum_{W'\in\overline{W}'} Q_{ik}(W,W'). \tag{A9}$$

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 (A9) along a column or row of $\overline{Q}_{ik}$ is at most 2, it follows that the recursive updating

$$\bar{\alpha}_{i+1} = \bar{\alpha}_i\overline{Q}_i\overline{D}_{i+1},$$

$$\bar{\beta}_i^{\mathrm{T}} = \overline{Q}_i\overline{D}_{i+1}\bar{\beta}_{i+1}^{\mathrm{T}},$$

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

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

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

DERIVATION OF (20)

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

$$P_i(w) = C^{-1}L_i(w)P_{i,0}(w), \tag{A10}$$

where $C = P_0(\boldsymbol{M}) = \sum_w P_{i,0}(w)L_i(w)$. Let $L(x_i,\varepsilon) = P_{i,\varepsilon}(\boldsymbol{M}|Y)$ denote the retrospective likelihood. Combining (A10), (19) and (17) we get

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

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

SIMULATING FROM $P_{i,\varepsilon}(\boldsymbol{b},\boldsymbol{v}\,|\,\mathbf{Y})$

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

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

where $\boldsymbol{h}_k = (b_{1k},\dots,b_{Kk})$ is the kth founder haplotype of length K, containing alleles at all marker loci. Hence, $\boldsymbol{v}$ and each $\boldsymbol{h}_k$ 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 $P_{i,\varepsilon}(w|Y)$.

(b) Given v_i, generate $\boldsymbol{v}^- = (v_1,\dots,v_{i-1})$ and $\boldsymbol{v}^+ = (v_{i+1},\dots,v_K)$ from $P(\boldsymbol{v}^-|v_i)P(\boldsymbol{v}^+|v_i)$. 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 $\boldsymbol{B}^- = (B_1,\dots,B_{i-1})$ and $\boldsymbol{B}^+ = (B_{i+1},\dots,B_K)$ from $P(\boldsymbol{B}^-|B_i)P(\boldsymbol{B}^+|B_i)$. 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

$$P(b_{i-L_i,k}|b_{i-L_i+1,k},\dots,b_{i,k}) = \frac{g_{ik}}{\sum \tilde{g}_{ik}} \tag{A11}$$

in direction from right to left. In the denominator of (A11), 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 1(a), we may use a biological model, with ε containing penetrances and disease allele frequencies. An alternative is to start with a score function S and disease allele frequencies and then use (19).