Statistica Neerlandica (2011) Vol. 65, nr. 3, pp. 239–258
doi:10.1111/j.1467-9574.2011.00480.x

Estimating and testing linkage disequilibrium from sib data

Marianne Jonker* and Aad van der Vaart†

Department of Mathematics, VU University Amsterdam, Netherlands

Zoltan Bochdanovits

Department of Clinical Genetics, VUMC, Amsterdam, Netherlands

Dorret Boomsma

Department of Biological Psychology, VU University Amsterdam, Netherlands

We consider estimation and testing of linkage equilibrium from genotypic data on a random sample of sibs, such as monozygotic and dizygotic twins. We compute the maximum likelihood estimator with an EM-algorithm and a likelihood ratio statistic that takes the family structure into account. As we are interested in applying this to twin data we also allow observations on single children, so that monozygotic twins can be included. We allow non-zero recombination fraction between the loci of interest, so that linkage disequilibrium between both linked and unlinked loci can be tested. The EM-algorithm for computing the maximum likelihood estimator of the haplotype frequencies and the likelihood ratio test-statistic, are described in detail. It is shown that the usual estimators of haplotype frequencies based on ignoring that the sibs are related are inefficient, and the likelihood ratio test for testing that the loci are in linkage disequilibrium.

Keywords and Phrases: twin studies, LD, EM-algorithm, likelihood ratio test, haplotype, phase ambiguity.

1 Introduction

Interest in gene-gene interactions that underlie human phenotypic variation is growing (Phillips, 2008; Cordell, 2009) and novel approaches to detect them have been proposed (Ritchie *et al.*, 2001; Zhang and Liu, 2007). Although these methods have been aimed at circumventing the need for an exhaustive search based on all possible gene pairs, nonetheless genome-wide screens of epistasis are still lacking due to the need for extensive multiple testing correction and the subsequent loss of power.

*m.a.jonker@vu.nl
†a.w.vander.vaart@vu.nl

Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

A possible alternative approach would be a two-stage design based on an easily detectable signature of epistatic interactions. Only pairs of loci that exhibit the desired property would be formally tested for joint association with a phenotype in the second stage. This approach would greatly reduce the number of tests performed and hence would increase the statistical power to detect gene-gene interactions.

When specific combinations of alleles at two different loci jointly affect a certain (disease) phenotype (by epistatic interactions) than these specific combinations of alleles will be overrepresented in a sample enriched for the phenotype (cases) and underrepresented in a sample depleted from the phenotype (controls). Consequently, the two loci involved will be in linkage disequilibrium (LD). This holds for both physically linked and unlinked loci. The mathematical details of this genomic 'signature' of gene-gene interaction are firmly established in population genetic literature (Pollak, 1979). LD between two loci can be detected in genotype data and used as a predictor of gene-gene interactions. Pre-screening for a deviation from LD may therefore reveal gene-gene interactions.

Measures of linkage disequilibrium between two loci are defined in terms of haplotype frequencies in the population (e.g. Balding, 2006). The standard method for inference on linkage disequilibrium is to measure the genotypes in a random sample of individuals from the population, and estimate the haplotype frequencies by maximum likelihood or compute a likelihood ratio for testing LD. The computation of the maximum likelihood estimator based on the genotypes at the two loci without phase information can be carried out by the EM-algorithm, as explained in Excoffier and Slatkin (1995) and Slatkin and Excoffier (1996). In this note we consider the alternate situation that we observe the genotypes of a random sample of n sib pairs. In view of the dependence between the sibs, viewing them as $2n$ individuals from the population and applying the standard approach would give incorrect significance levels. Also it turns out that even for estimating the marginal (single-locus) frequencies the empirical estimators (the fractions of alleles among the $2n$ sibs) are inefficient relative to the maximum likelihood estimator, even though unbiased.

Standard, multi-purpose packages for genetic inference, such as GENEHUNTER, can also produce estimates of haplotype frequencies, but make the assumption that the loci are in linkage equilibrium, at least in the founders of the pedigrees. For our present purpose and small pedigrees, this would lead to large biases, as explained in Schaid *et al.* (2002), or even be useless to test for disequilibrium. Becker and Knapp (2002, 2004) consider the estimation of haplotype frequencies in a parents' population from genotypic information on parents and their children, and their software package FAMHAP can deal with missing data. Putter, Meulenbelt and van Houwelingen (2007) focus on the relative efficiency of haplotype frequency estimation in sibships compared to unrelated individuals, and describe an algorithm to estimate the haplotype frequencies in the parents' (and sib) population based on sib data only. However, all three papers assume that the recombination fraction between the loci is equal to zero. In the present note we consider the problem with a general recombination fraction, which obviously complicates the algorithm. In the application to finding epistatic interactions it is desired to test linkage equilibrium for unlinked loci.

In this article the recombination fraction between the two loci is assumed known. In practice it is set equal to 0.5 for loci on different chromosomes, or is determined from a genetic map or approximated from a physical map. The robustness of the estimation and testing methods against deviations of the estimated recombination fraction from the true value was evaluated by means of a simulation study. The results show that the effect of small to medium misspecifications of the fraction hardly effects the estimates of the haplotype frequencies and the level of the likelihood ratio test for testing linkage equilibrium (LE) between the loci, but assuming a zero-recombination fraction (as in the algorithms presented by Becker and Knapp (2002, 2004) and Putter *et al.* (2007)) while the loci are actually unlinked may yield biased estimates of haplotype frequencies and an invalid test.

Besides on computation of estimates of LD, we focus in this article on the likelihood ratio statistic for testing for linkage disequilibrium, again based on unphased genotype data of siblings only. This requires the computation of the maximum likelihood estimators for the haplotype frequencies under the general model and under the null hypothesis of LE, for which we derive the EM-algorithm. Furthermore, it requires a conditioning argument for the likelihood of the observed data (on the sibs) versus the likelihood for the (imaginary) data consisting of parents and sibs.

Our research was motivated by the work of Bochdanovits *et al.* (2008). In their research it is shown that pre-screening for a deviation from LD may reveal gene-gene interactions. Specific mouse data (mouse recombinant inbred lines (RILs)) is used. LD has been quantified between all pairs of physically unlinked loci for which genotype is publicly available. Given the specific nature of the mouse data (RILs) it was possible to estimate LD with a simple Pearson correlation coefficient and highly significant deviations from equilibrium were found. The gene pairs that showed excess LD were tested for association against a set of publicly available phenotype data and a significant interaction effect was found between genes involved in detoxification and voluntary ethanol consumption in mice. This biologically plausible result demonstrated the validity of the concept that ascertainment of unlinked gene pairs that are in significant LD can be expected to non-additively affect quantitative phenotypic variation, as predicted by population genetic theory (Pollak, 1979).

The description of the EM-algorithm and the computation of the likelihood ratio statistic are the subject of sections 2 and 5. In addition we show in section 3 how to include also a random sample of single children in the analysis, which is important for instance in applications to twin studies. In section 4 we briefly compare (for estimating the marginal frequencies) the efficiency of the estimates based on sibs with the estimate based on a random sample of an equal number of unrelated individuals. In section 6 the results of a simulation study are presented. This simulation study was performed to evaluate the reliability of the proposed estimation and testing methods and to show the effect of using a (slightly) misspecified recombination fraction. In this section, also the results of the application of the EM-algorithm and LD-testing on data from a study on childhood depression are presented.

Code for an implementation in the R package is available from the authors.

2 EM-algorithm

Consider a nuclear family as in Figure 1, where we denote by (Y_1, Y_2) and (Y_3, Y_4) the ordered haplotypes of the father and the mother, and by (X_1, X_2) and (X_3, X_4) the ordered haplotypes of the two children, ordered by paternal and maternal origins. All haplotypes refer to two loci, which we assume to be biallelic, with alleles denoted by 0 and 1. Thus all variables Y_i and X_j take their values in the set

$$\mathcal{Y} := \mathcal{X} := \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}.$$

We assume that the parents' haplotypes Y_1, Y_2, Y_3, Y_4 are i.i.d. random vectors with relative frequencies $h_{00}, h_{01}, h_{10}, h_{11}$. (For notational convenience we write h_{kl} for h_y with $y = \begin{pmatrix} k \\ l \end{pmatrix}$.) The independence can be justified by assuming random mating in the populations of the parents and the parents' parents.

The haplotype frequencies in the children's population can be computed from the haplotype frequencies in the parents' population and the recombination fraction, by conditioning on the event a recombination had or had not taken place. The (k, l)-haplotype frequency in the children's population is given by

$$(1-\theta)h_{kl} + \theta h_{k.} h_{.l}, \quad k, l \in \{0, 1\},$$

for θ the recombination fraction between the loci. Here $h_{k.} = h_{k0} + h_{k1}$ and $h_{.l} = h_{0l} + h_{1l}$ are the marginal frequencies for the two loci, which are the same in the parents' and children's populations. The haplotype frequencies in the children's population can be seen to factorize over the loci if and only if the h_{kl} factorize ($h_{kl} = h_{k.} h_{.l}$), i.e. the parents' population is in linkage equilibrium (LE) if and only if the children's population is. Write, for $j = 1, 2, 3, 4$,

$$X_j = \begin{pmatrix} X_{j1} \\ X_{j2} \end{pmatrix}, \quad Y_j = \begin{pmatrix} Y_{j1} \\ Y_{j2} \end{pmatrix}.$$

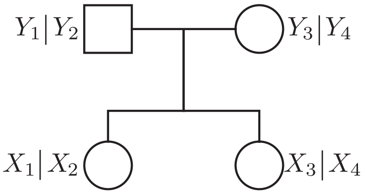

Fig. 1. Notation. The variables Y_1, Y_3 and X_1, X_3 are the paternal haplotypes of the parents and the sibs, respectively. The remaining variables are the maternal haplotypes.

Then the unordered genotypes of the two children at locus 1 can be written as $\{X_{11}, X_{21}\}$ and $\{X_{31}, X_{41}\}$; and the unordered genotypes at locus 2 as $\{X_{12}, X_{22}\}$ and $\{X_{32}, X_{42}\}$.

As mentioned the assumption that Y_1, Y_2, Y_3, Y_4 are i.i.d. is satisfied under random mating in the parents' and the parents' parents generations, assuming that generations can indeed be separated. In addition we assume that given Y_1, Y_2, Y_3, Y_4 the distribution of X_1, X_2, X_3, X_4 is determined by the usual segregation model in genetics. Specifically X_1, X_2, X_3, X_4 are conditionally independent (expressing that the four meioses involved are independent); the first coordinates of X_1 and X_3 are randomly chosen from the first coordinates of the parents Y_1 and Y_2, and the second coordinate is chosen from the same parent with probability $1-\theta$ and otherwise from the other parent; finally X_2, X_4 derive from Y_3, Y_4 in the same manner. The recombination fraction θ is assumed given. The total set of observations are the *unordered* genotypes derived from a random sample of size n from the distribution of (X_1, X_2, X_3, X_4). We want to estimate the haplotype frequencies $h_{00}, h_{01}, h_{10}, h_{11}$ in the parents' population or test that they factorize over the two loci. We do not observe the parents' genotypes.

The maximum likelihood estimator of the haplotype frequencies can be computed using the EM-algorithm, with (the sample of) variables Y_1, Y_2, Y_3, Y_4, X_1, X_2, X_3, X_4 as the full data and the unordered genotypes

$$\{X_{11}, X_{21}\}, \{X_{31}, X_{41}\}, \{X_{12}, X_{22}\}, \{X_{32}, X_{42}\}$$

as the observed data. This requires two conditioning or reconstruction steps.

1. From the observed data to the ordered pairs of haplotypes (X_1, X_2) and (X_3, X_4) of the children.
2. From $(X_1, X_2), (X_3, X_4)$ to the full data Y_1, Y_2, Y_3, Y_4, X_1, X_2, X_3, X_4.

Because of the dependency between the children both steps have to be performed on the set of all four individuals, although of course intermediate formulas factorize due to the random mating assumption and the assumption of independent meioses and segregation.

This allows to compute for each of the 16 possible values of (Y_1, Y_2) the four conditional probabilities

$$q(x_1 \,|\, y_1, y_2) := \mathrm{P}(X_1 = x_1 \,|\, Y_1 = y_1,\, Y_2 = y_2), \quad x_1, y_1, y_2 \in \mathcal{Y}.$$

These 64 numbers (many of which are zero) are simple functions of the recombination fraction and, given the numerical value of the latter, can be stored at the beginning of the algorithm. The conditional distribution of X_3 given (Y_1, Y_2) is identical to the conditional distribution in the display, and X_1 and X_3 are independent given (Y_1, Y_2). Furthermore, the vectors (Y_1, Y_2, X_1, X_3) and (Y_3, Y_4, X_2, X_4) are i.i.d., under the assumed symmetry between the sexes. (If we want to make a difference between paternal and maternal origins, then the distributions of the two vectors still have the same form, but the 64 probabilities can be different.)

The joint distribution of the two paternal haplotypes of the children can now be computed as

$$\begin{aligned} p_h(x_1,x_3) &:= \mathrm{P}_h(X_1=x_1, X_3=x_3) \\ &= \sum_{y_1\in\mathcal{Y}}\sum_{y_2\in\mathcal{Y}} \mathrm{P}(X_1=x_1 \mid Y_1=y_1,\, Y_2=y_2)\mathrm{P}(X_3=x_3 \mid Y_1=y_1,\, Y_2=y_2)h_{y_1}h_{y_2} \\ &= \sum_{y_1\in\mathcal{Y}}\sum_{y_2\in\mathcal{Y}} q(x_1 \mid y_1,y_2)q(x_3 \mid y_1,y_2)h_{y_1}h_{y_2}. \end{aligned}$$

This sum has 16 terms, in principle.

The data on a single child (two unordered genotypes) can be summarized as a count in a (3×3)-table (see Table 1), with the three possible genotypes $\{0,0\},\{0,1\}$, $\{1,1\}$ for the two loci on the two dimensions. For 8 of the 9 cells in the table, the genotype of at least one locus is homozygous, and it is possible to recover the *unordered* pair of haplotypes with certainty from the pair of genotypes. E.g. if $\{0,1\}$ and $\{0,0\}$ are the genotypes of an individual at locus 1 and locus 2, then

$$\{0,1\}\times\{0,0\} \to \left\{\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}\right\}.$$

The remaining (middle) cell in the (3×3)-table is the combination $\{0,1\}\times\{0,1\}$, which corresponds to two possible pairs of unordered haplotypes:

$$\{0,1\}\times\{0,1\} \to \left\{\begin{pmatrix}0\\1\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}\right\} \quad \text{or} \quad \left\{\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1\\1\end{pmatrix}\right\}. \tag{1}$$

To take account of the family structure we need to map back further into the *ordered* pairs of haplotypes (X_1,X_2) and (X_3,X_4). This should be done jointly for the two children, since the other child may be informative about the parents, who are informative on the haplotypes of the first child.

This step entails ordering of the unordered pairs of haplotypes for children in the 8 boundary cells and resolution and ordering of haplotypes for the doubly heterozygous children. Say that a child is double heterozygous (DHZ) if it is heterozygous at both loci and not DHZ (NHZ) otherwise. There are four cases:

Table 1. Lay-out of the data on one sib

	{0,0}	{0,1}	{1,1}
{0,0}			
{0,1}		DHZ	
{1,1}			

Notes: The two borders refer to the three possible (unordered) genotypes $\{0,0\},\{0,1\},\{1,1\}$ at the two loci. A sib belongs to one of the nine entries in the table. Doubly heterozygous (DHZ) sibs are counted in the middle cell, marked DHZ.

1. Both children NHZ. We have two unordered pairs of haplotypes $\{X_1, X_2\}$ and $\{X_3, X_4\}$, which can be ordered in four different ways, with probabilities given in Table 2.
2. First child NHZ, second child DHZ. We have an unordered pair of haplotypes $\{X_1, X_2\}$ and an ambiguous set of unordered haplotypes

   $$\left\{\begin{pmatrix} X_{31} \\ X_{32} \end{pmatrix}, \begin{pmatrix} X_{41} \\ X_{42} \end{pmatrix}\right\}, \quad \text{or} \quad \left\{\begin{pmatrix} X_{31} \\ X_{42} \end{pmatrix}, \begin{pmatrix} X_{41} \\ X_{32} \end{pmatrix}\right\}.$$

   There are eight ways to form these in two ordered pairs of haplotypes, given in Table 3 with their relative probabilities.
3. First child DHZ, second child NHZ. There are eight possibilities.
4. Both children DHZ. There are 16 possibilities.

Thus we have described the conditional distribution of $(X_1, X_2), (X_3, X_4)$ given the data. (The mentioned numbers of possibilities are maxima; for concrete observations many of them may coincide.)

Finally we describe the steps of the EM-algorithm. The conditional distribution of $(Y_1, Y_2), (Y_3, Y_4)$ given $(X_1, X_2), (X_3, X_4)$ can be factorized as the product of the

Table 2. Ordering of $\{X_1, X_2\}, \{X_3, X_4\}$ in the situation that both pairs are NHZ

$\{z_1, z_2\}, \{z_3, z_4\}$	Probability proportional to
$(z_1, z_2), (z_3, z_4)$	$p_h(z_1, z_3)p_h(z_2, z_4)$
$(z_1, z_2), (z_4, z_3)$	$p_h(z_1, z_4)p_h(z_2, z_3)$
$(z_2, z_1), (z_3, z_4)$	$p_h(z_2, z_3)p_h(z_1, z_4)$
$(z_2, z_1), (z_4, z_3)$	$p_h(z_2, z_4)p_h(z_1, z_3)$

Notes: If the observed values are $\{z_1, z_2\}, \{z_3, z_4\}$, then there are four possible orderings, given in the left column. These have conditional probabilities proportional to the expression in the right column.

Table 3. Ordering and resolution of $\{X_1, X_2\}, \{X_3, X_4\}$ in case the first pair is NHZ, the second DHZ

$\{z_1, z_2\}, \{z_{31}, z_{41}\}, \{z_{32}, z_{42}\}$	Probability proportional to
$(z_1, z_2), (z_3, z_4)$	$p_h(z_1, z_3)p_h(z_2, z_4)$
$(z_1, z_2), \left(\begin{pmatrix} z_{31} \\ z_{42} \end{pmatrix}, \begin{pmatrix} z_{41} \\ z_{32} \end{pmatrix}\right)$	$p_h\left(z_1, \begin{pmatrix} z_{31} \\ z_{42} \end{pmatrix}\right) p_h\left(z_2, \begin{pmatrix} z_{41} \\ z_{32} \end{pmatrix}\right)$
$(z_1, z_2), (z_4, z_3)$	$p_h(z_1, z_4)p_h(z_2, z_3)$
$(z_1, z_2), \left(\begin{pmatrix} z_{41} \\ z_{32} \end{pmatrix}, \begin{pmatrix} z_{31} \\ z_{42} \end{pmatrix}\right)$	$p_h\left(z_1, \begin{pmatrix} z_{41} \\ z_{32} \end{pmatrix}\right) p_h\left(z_2, \begin{pmatrix} z_{31} \\ z_{42} \end{pmatrix}\right)$
$(z_2, z_1), (z_3, z_4)$	$p_h(z_2, z_3)p_h(z_1, z_4)$
$(z_2, z_1), \left(\begin{pmatrix} z_{31} \\ z_{42} \end{pmatrix}, \begin{pmatrix} z_{41} \\ z_{32} \end{pmatrix}\right)$	$p_h\left(z_2, \begin{pmatrix} z_{31} \\ z_{42} \end{pmatrix}\right) p_h\left(z_1, \begin{pmatrix} z_{41} \\ z_{32} \end{pmatrix}\right)$
$(z_2, z_1), (z_4, z_3)$	$p_h(z_2, z_4)p_h(z_1, z_3)$
$(z_2, z_1), \left(\begin{pmatrix} z_{41} \\ z_{32} \end{pmatrix}, \begin{pmatrix} z_{31} \\ z_{42} \end{pmatrix}\right)$	$p_h\left(z_2, \begin{pmatrix} z_{41} \\ z_{32} \end{pmatrix}\right) p_h\left(z_1, \begin{pmatrix} z_{31} \\ z_{42} \end{pmatrix}\right)$

Notes: If the observed values are $\{z_1, z_2\}$, $\{z_{31}, z_{41}\}, \{z_{32}, z_{42}\}$, then there are eight possible orderings, given in the left column. These have conditional probabilities proportional to the expression in the right column.

conditional distribution of (Y_1, Y_2) given (X_1, X_3) and the conditional distribution of (Y_3, Y_4) given (X_2, X_4), which have the same form. These conditional distributions can be determined using Bayes rule:

$$\begin{aligned}&\mathrm{P}_h(Y_1 = y_1, Y_2 = y_2 \mid X_1 = x_1, X_3 = x_3)\\ &\quad = \frac{\mathrm{P}(X_1 = x_1 \mid Y_1 = y_1, Y_2 = y_2)\mathrm{P}(X_3 = x_3 \mid Y_1 = y_1, Y_2 = y_2)h_{y_1}h_{y_2}}{p_h(x_1, x_3)}\\ &\quad = \frac{q(x_1 \mid y_1, y_2)q(x_3 \mid y_1, y_2)h_{y_1}h_{y_2}}{p_h(x_1, x_3)}.\end{aligned}$$

The conditional probabilities $q(x_i \mid y_1, y_2)$ and the probabilities $p_h(x_1, x_3)$ were obtained before.

The likelihood for observing a sample of observations of the full data Y_1^i, Y_2^i, Y_3^i, $Y_4^i, X_1^i, X_2^i, X_3^i, X_4^i$ (for $i = 1, \ldots, n$) takes the form

$$\prod_{i=1}^{n} h_{Y_1^i} h_{Y_2^i} h_{Y_3^i} h_{Y_4^i} p(X_1^i, X_2^i, X_3^i, X_4^i \mid Y_1^i, Y_2^i, Y_3^i, Y_4^i). \tag{2}$$

Here the second term, the conditional density of $X_1^i, X_2^i, X_3^i, X_4^i$ given Y_1^i, Y_2^i, Y_3^i, Y_4^i is free of the parameter (the haplotype frequencies), and hence can be dropped. The logarithm of the remaining part can be written

$$L(h) = \sum_{k,l} N_{kl} \log h_{kl},$$

where N_{kl} is the total number of haplotypes $\binom{k}{l}$ carried by the parents.

In the E-step of the EM-algorithm we compute the conditional expectation

$$E_0(L(h) \mid DATA) = \sum_{k,l} \log h_{kl} E_0(N_{kl} \mid DATA).$$

The subscript 0 indicates that we compute the expectation using the current iterate of the haplotype frequencies. In the M-step we find the haplotype frequencies h that maximize this expression. Since the preceding display precisely gives a multinomial likelihood with 'observed values' $E_0(N_{kl} \mid DATA)$, the M-step is seen to be

$$\hat{h}_{kl} = \frac{1}{4n} E_0(N_{kl} \mid DATA). \tag{3}$$

The difficulty is to compute the conditional expectation in this display.

The variable N_{kl} can be written as a sum $N_{kl} = \sum_{i=1}^{n} N_{kl}^i$ over the n families. Because the families are independent, the conditional expectation of this sum given the DATA can be written as a sum over the families as well, conditioning the ith variable N_{kl}^i on the observed values $\{X_{11}^i, X_{21}^i\}, \{X_{31}^i, X_{41}^i\}, \{X_{12}^i, X_{22}^i\}$, $\{X_{32}^i, X_{42}^i\}$ for that family. Here we may condition first on the ordered genotypes $(X_1^i, X_2^i), (X_3^i, X_4^i)$, where we can use that the vectors $(Y_1^i, Y_2^i, X_1^i, X_3^i)$ and $(Y_3^i, Y_4^i, X_2^i, X_4^i)$ are independent. For a typical family i, we have, with superscript i omitted from the alleles from the second line onwards,

$$\begin{aligned}
&E_0[N^i_{kl}\,|\,\{X^i_{11},X^i_{21}\},\{X^i_{31},X^i_{41}\},\{X^i_{12},X^i_{22}\},\{X^i_{32},X^i_{42}\}]\\
&\quad = E_0\left[E_0(1_{Y_1=\binom{k}{l}}+1_{Y_2=\binom{k}{l}}\,|\,X_1,X_3)+E_0(1_{Y_3=\binom{k}{l}}+1_{Y_4=\binom{k}{l}}\,|\,X_2,X_4)\right.\\
&\qquad\left.|\,\{X_{11},X_{21}\},\{X_{31},X_{41}\},\{X_{12},X_{22}\},\{X_{32},X_{42}\}\right].
\end{aligned}$$

Here the inner conditional expectations can be written

$$\begin{aligned}
&E_0(1_{Y_1=\binom{k}{l}}+1_{Y_2=\binom{k}{l}}\,|\,X_1,X_3)=2\mathrm{P}_0\left(Y_1=\binom{k}{l}\,|\,X_1,X_3\right)\\
&\quad =2h_{0,kl}\frac{\sum_{y_2\in\mathcal{Y}}q(X_1\,|\,\binom{k}{l},y_2)q(X_3\,|\,\binom{k}{l},y_2)h_{0,y_2}}{p_0(X_1,X_3)}=:2h_{0,kl}\frac{\phi_{0,kl}}{p_0}(X_1,X_3).
\end{aligned}$$

Here $h_{0,kl}$ are the current iterates of the haplotype frequencies h_{kl}. The second inner conditional expectation has the same form, but with (X_1,X_3) replaced by (X_2,X_4), whence

$$\begin{aligned}
\hat{h}_{kl}=h_{0,kl}\frac{1}{2n}\sum_{i=1}^{n}E_0[&\frac{\phi_{0,kl}}{p_0}(X^i_1,X^i_3)+\frac{\phi_{0,kl}}{p_0}(X^i_2,X^i_4)\\
&|\,\{X^i_{11},X^i_{21}\},\{X^i_{31},X^i_{41}\},\{X^i_{12},X^i_{22}\},\{X^i_{32},X^i_{42}\}].
\end{aligned}$$

The remaining conditional expectation consists of averaging over the resolutions of the observed data into the ordered haplotypes of the sibs, which involves the four different combinations $\{NHZ,DHZ\}^2$ with their $4,8,8,16$ subcases, considered previously.

As written above, every iteration of the algorithm requires recomputation of a sum over all families, which will make the algorithm slow with large samples. Because there are only $9=3^2$ possible realizations of the data on one sib (such as $\{x_{11},x_{21}\},\{x_{12},x_{22}\}$ for the first sib), there are only 81 possible realizations of the data on one family. Furthermore, identification of observations that are equal after permuting the sibs leaves only $45=(1/2)9(9+1)$ different realizations. Thus the formulas can actually be condensed in no more than 45 different cases, for any sample size. The symmetrization in the last conditioning step can also be simplified with this reduction in mind. Still the algorithm will be slower than the EM-algorithm for the case of observing a random sample from a population, as there iterations involve only the resolution of the DHZ sibs, as in Equation 1.

3 Single children

Suppose that next to the sib data as in Figure 1 we also have observations on single children. This situation arises for instance in twin studies, where dizygotic twins contribute sibs, but monozygotic twins should be considered single children.

The easiest method to include single children is to consider them as sibs with the second (imaginary) sib missing. The EM-algorithm then needs an additional conditioning step for the single children. The EM-algorithm in the preceding section ended with conditioning the variable $(\phi_0/p_0)(X_1,X_3)+(\phi_0/p_0)(X_2,X_4)$ on $\{X_{11},X_{21}\}$, $\{X_{31},X_{41}\},\{X_{12},X_{22}\},\{X_{32},X_{42}\}$. We now condition this variable on $\{X_{11},X_{21}\}$, $\{X_{12},X_{22}\}$ instead. We may first condition on the ordered pair (X_1,X_2). Given this pair the variables X_3 and X_4 possess densities proportional to $u\mapsto p_h(X_1,u)$ and $u\mapsto p_h(X_2,u)$, respectively. Hence,

$$E_0\left(\frac{\phi_{0,kl}}{p_0}(X_1,X_3)+\frac{\phi_{0,kl}}{p_0}(X_2,X_4)\,|\,(X_1,X_2)\right)$$
$$=\sum_{u\in\mathcal{X}}\left(\frac{\phi_{0,kl}}{p_0}(X_1,u)\frac{p_0(X_1,u)}{\sum_{u\in\mathcal{X}}p_0(X_1,u)}+\frac{\phi_{0,kl}}{p_0}(X_2,u)\frac{p_0(X_2,u)}{\sum_{u\in\mathcal{X}}p_0(X_2,u)}\right)$$
$$=\left[\frac{\sum_{y_2\in\mathcal{Y}}q\left(X_1\,|\begin{pmatrix}k\\l\end{pmatrix},y_2\right)h_{0,y_2}}{p_0(X_1)}+\frac{\sum_{y_2\in\mathcal{Y}}q\left(X_2\,|\begin{pmatrix}k\\l\end{pmatrix},y_2\right)h_{0,y_2}}{p_0(X_2)}\right].$$

Here p_0 is the current estimate of the marginal density $p_h(x_1)=\sum_{u\in\mathcal{X}}p_h(x_1,u)$ of X_1 (with some abuse of notation denoted by the same symbol as the joint density of (X_1,X_3), but with one argument). In the case that the child is NHZ, the unordered pair $\{X_1,X_2\}$ is observed, and no further conditioning is necessary, as the expression in the display is already a function of this unordered pair. In the case that the child is DHZ, there are two possible unordered pairs given the data (which is $\{0,1\},\{0,1\}$), given in Equation 1, with conditional probabilities proportional to

$$p_0\left(\begin{pmatrix}0\\1\end{pmatrix}\right)p_0\left(\begin{pmatrix}1\\0\end{pmatrix}\right)\quad\text{and}\quad p_0\left(\begin{pmatrix}0\\0\end{pmatrix}\right)p_0\left(\begin{pmatrix}1\\1\end{pmatrix}\right).$$

The expression in the preceding display, with X_1,X_2 the pair $\begin{pmatrix}0\\1\end{pmatrix},\begin{pmatrix}1\\0\end{pmatrix}$ or the pair $\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1\\1\end{pmatrix}$, must be averaged over these two possibilities.

4 Single locus frequencies

To test the hypothesis of linkage equilibrium we must also compute the maximum likelihood estimator under the null hypothesis that the haplotype frequencies factorize over the loci. In this case the full likelihood Equation 2 can be factorized in two parts referring to parameters and observations of the two loci, and the observations for the two loci are stochastically independent. (Indeed $h_{Y_j}=h_{Y_{j1}.}h_{.Y_{j2}}$ for every $j=1,\ldots,4$, where $h_{k.}$ and $h_{.l}$ are the marginal frequencies and Y_{j1} and Y_{j2} the alleles of parent j at loci 1 and 2.) It follows that the null maximum likelihood estimators of the marginal frequencies $h_{k.}$ and $h_{.l}$ are the maximum likelihood estimators based on the data concerning the two loci separately.

The empirical estimators

$$\frac{1}{4n}\sum_{i=1}^{n}(X_{11}^i+X_{21}^i+X_{31}^i+X_{41}^i) \quad \text{and} \quad \frac{1}{4n}\sum_{i=1}^{n}(X_{12}^i+X_{22}^i+X_{32}^i+X_{42}^i)$$

are unbiased estimators for the marginal frequencies $h_{1.}$ and $h_{.1}$. However, these are not the maximum likelihood estimators under our assumptions. In fact, the variance of the empirical estimators suffers from the dependence between the sib-information $X_{11}^i+X_{21}^i$ and $X_{31}^i+X_{41}^i$, which renders them much less efficient than the maximum likelihood estimators, as shown in Figure 2. This picture also shows that a random sample of $2n$ unrelated children allows better estimates than a sample of n sib pairs, but the loss in efficiency is not big.

The maximum likelihood estimators of the marginal frequencies can be computed through the EM-algorithm by similar, but simpler, arguments as before. We keep the notation as given in Figure 1, but now let the variables $(Y_1, Y_2), (Y_3, Y_4)$, (X_1, X_2), (X_3, X_4) refer to the ordered genotypes of parents and children at a single locus.

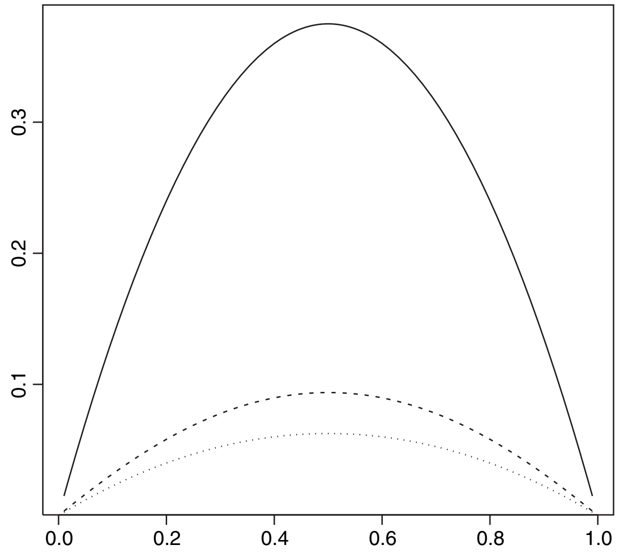

Fig. 2. Variance of the empirical estimator, the maximum likelihood estimator, and the empirical estimator based on a random sample of $2n$ unrelated children of a marginal frequency (top to bottom) as a function of the marginal frequency on the horizontal axis, per observation. The middle curve gives the inverse Fisher information, whereas the other two curves are the exact variances.

Thus these variables take their values in the set $\{0,1\}$, with Y_1, Y_2, Y_2, Y_4 i.i.d. with unknown probability $h = \mathrm{P}(Y_j = 1)$, and the conditional distribution of X_1, X_2, X_3, X_4 given Y_1, Y_2, Y_3, Y_4 completely determined by the segregation process. We observe the unordered genotypes $\{X_1, X_2\}$ and $\{X_3, X_4\}$ of the two sibs, in a random sample of n families.

The variables (X_1, X_3) and (X_2, X_4) are i.i.d. with density given in Table 4. The distribution of $\{X_1, X_2\}, \{X_3, X_4\}$ can be obtained from this by listing the map from the pairs $(X_1, X_2), (X_3, X_4)$ into the pairs $\{X_1, X_2\}, \{X_3, X_4\}$. This correspondence is given in Table 5, with $\{0,0\}, \{0,1\}, \{1,1\}$ denoting the three possible values of $\{X_1, X_2\}$.

The full likelihood takes the form

$$\prod_{i=1}^{n}\left(\prod_{j=1}^{4}(1-h)^{1-Y_j^i}h^{Y_j^i}\right)p(X_1^i, X_3^i\,|\,Y_1^i, Y_2^i)p(X_2^i, X_4^i\,|\,Y_3^i, Y_4^i) \propto (1-h)^{4n-N}h^N,$$

for $N = \sum_{i=1}^{n}\sum_{j=1}^{4} Y_j^i$ the total number of Y_j^i that are equal to 1. An iteration of the EM-algorithm becomes

$$\begin{aligned}
\hat{h} &= \frac{1}{4n}E_0(N\,|\,DATA)\\
&= \frac{2}{4n}\sum_{i=1}^{n} E_0[E_0(Y_1^i\,|\,X_1^i, X_3^i) + E_0(Y_3^i\,|\,X_2^i, X_4^i)\,|\,\{X_1^i, X_2^i\}, \{X_3^i, X_4^i\}]\\
&= \frac{1}{2n}\sum_{i=1}^{n} E_0\left[\frac{\psi_0}{p_0}(X_1^i, X_3^i) + \frac{\psi_0}{p_0}(X_2^i, X_4^i)\,|\,\{X_1^i, X_2^i\}, \{X_3^i, X_4^i\}\right],
\end{aligned}$$

Table 4. Density of (X_1, X_3)

$x_1 x_3$	$p_h(x_1, x_3)$
0,0	$(1-h)^2 + h(1-h)/2$
0,1	$h(1-h)/2$
1,0	$h(1-h)/2$
1,1	$h(1-h)/2 + h^2$

Table 5. Origin of observed data on a single locus from ordered genotypes of the two children

$\{x_1, x_2\}$	$\{x_3, x_4\}$	(x_1, x_2), (x_3, x_4)
{0,0}	{0,0}	(0,0),(0,0)
{0,0}	{0,1}	(0,0),(0,1) or (0,0),(1,0)
{0,0}	{1,1}	(0,0),(1,1)
{0,1}	{0,0}	(0,1),(0,0) or (1,0),(0,0)
{0,1}	{0,1}	(0,1),(0,1) or (1,0),(0,1) or (0,1),(1,0) or (1,0),(1,0)
{0,1}	{1,1}	(0,1),(1,1) or (1,0),(1,1)
{1,1}	{0,0}	(1,1),(0,0)
{1,1}	{0,1}	(1,1),(0,1) or (1,1),(1,0)
{1,1}	{1,1}	(1,1),(1,1)

Note: The 16 ordered genotypes on the right give rise to the nine unordered genotypes on the left.

Table 6. Values of $E_h(Y_1 \,|\, X_1, X_3)p_h(X_1, X_3)$

x_1, x_3	$\psi_h(x_1, x_3)$
0,0	$h(1-h)/4$
0,1	$h(1-h)/4$
1,0	$h(1-h)/4$
1,1	$h(1-h)/4+h^2$

where ψ_0 is the value at the current iterate of the function ψ_h defined by the equation $E_h(Y_1 \,|\, X_1, X_3) = \psi_h(X_1, X_3)/p_h(X_1, X_3)$, and given in Table 6. By arguments similar as before the right side can be reexpressed as

$$\hat{h} = \frac{1}{n}\sum_{i=1}^{n} \frac{\psi_0(X_1^i, X_3^i)p_0(X_2^i, X_4^i) + \psi_0(X_1^i, X_4^i)p_0(X_2^i, X_3^i) + \psi_0(X_2^i, X_4^i)p_0(X_1^i, X_3^i) + \psi_0(X_2^i, X_3^i)p_0(X_1^i, X_4^i)}{p_0(X_1^i, X_3^i)p_0(X_2^i, X_4^i) + p_0(X_1^i, X_4^i)p_0(X_2^i, X_3^i) + p_0(X_2^i, X_4^i)p_0(X_1^i, X_3^i) + p_0(X_2^i, X_3^i)p_0(X_1^i, X_4^i)}.$$

The four terms of the sums in the numerator and denominator are different only in the case that both sibs are heterozygous. As shown in Table 5 in the other cases these sums can be reduced to two terms or one term.

Observations on single children can again be included by considering these as sibs with one child missing. Instead of $E_0[Y_1^i + Y_3^i \,|\, \{X_1^i, X_2^i\}, \{X_3^i, X_4^i\}]$ a single child contributes a term to the sum of the form

$$E_0\left[\frac{\psi_0}{p_0}(X_1^i, X_3^i) + \frac{\psi_0}{p_0}(X_2^i, X_4^i) \,|\, \{X_1^i, X_2^i\}\right] = \frac{\sum_u \psi_0(X_1^i, u)}{p_0(X_1^i)} + \frac{\sum_u \psi_0(X_2^i, u)}{p_0(X_2^i)},$$

where p_0 is the marginal density of X_1 under the current iterate.

5 Likelihood ratio test

The existence of linkage equilibrium can be tested by the likelihood ratio test, which compares the likelihood of the observed data at the maximum likelihood estimators of the haplotype frequencies under the general model and under the null hypothesis of LE between the loci under the assumption that $h_{kl} = h_{k.}h_{.l}$ for $k, l \in \{0, 1\}$). In our situation, where the parents are missing, the ratio of the likelihoods of the observed data under two parameters can be computed from the corresponding ratio of likelihoods of the 'full' data by the general formula

$$\frac{p_\eta(W)}{p_{\eta_0}(W)} = E_{\eta_0}\left(\frac{r_\eta(Z)}{r_{\eta_0}(Z)} \,|\, W\right).$$

Here Z is a 'full' observation with density r_η and W a transformation of Z, with density p_η. We apply this formula with η and η_0 equal to the maximum likelihood estimators $\hat{h}$ and $\tilde{h}$ of the haplotype frequencies in the parents' population under full and null hypotheses, with Z the sample of variables $Y_1^i, Y_2^i, Y_3^i, Y_4^i$, $X_1^i, X_2^i, X_3^i, X_4^i$, and with W the sample of observed unordered genotypes $\{X_{11}^i, X_{21}^i\}$, $\{X_{31}^i, X_{41}^i\}$, $\{X_{12}^i, X_{22}^i\}, \{X_{32}^i, X_{42}^i\}$ (for $i = 1, \ldots, n$). The ratio of the likelihoods can then be written in the form (cf. Equation 2)

$$E_{\tilde{h}}\left[\prod_{i=1}^{n}\frac{\hat{h}_{Y_1^i}\hat{h}_{Y_2^i}\hat{h}_{Y_3^i}\hat{h}_{Y_4^i}}{\tilde{h}_{Y_1^i}\tilde{h}_{Y_2^i}\tilde{h}_{Y_3^i}\tilde{h}_{Y_4^i}}\,|\,DATA\right].$$

As before the term $p(X_1^i, X_2^i, X_3^i, X_4^i \mid Y_1^i, Y_2^i, Y_3^i, Y_4^i)$, which according to Equation 2 appears in both numerator and denominator of the ratio has cancelled out, because it is independent of the haplotype frequencies. By independence of the different nuclear families the above expression can be rewritten as

$$\prod_{i=1}^{n} E_{\tilde{h}}\left[\frac{\hat{h}_{Y_1^i}\hat{h}_{Y_2^i}\hat{h}_{Y_3^i}\hat{h}_{Y_4^i}}{\tilde{h}_{Y_1^i}\tilde{h}_{Y_2^i}\tilde{h}_{Y_3^i}\tilde{h}_{Y_4^i}}\,|\,\{X_{11}^i, X_{21}^i\}, \{X_{31}^i, X_{41}^i\}, \{X_{12}^i, X_{22}^i\}, \{X_{32}^i, X_{42}^i\}\right].$$

We first condition on the ordered genotypes $(X_1^i, X_2^i), (X_3^i, X_4^i)$, which yields

$$\prod_{i=1}^{n} E_{\tilde{h}}\left[E_{\tilde{h}}\left(\frac{\hat{h}_{Y_1^i}\hat{h}_{Y_2^i}\hat{h}_{Y_3^i}\hat{h}_{Y_4^i}}{\tilde{h}_{Y_1^i}\tilde{h}_{Y_2^i}\tilde{h}_{Y_3^i}\tilde{h}_{Y_4^i}}\,|\,(X_1^i, X_2^i), (X_3^i, X_4^i)\right)\right.$$
$$\left.|\,\{X_{11}^i, X_{21}^i\}, \{X_{31}^i, X_{41}^i\}, \{X_{12}^i, X_{22}^i\}, \{X_{32}^i, X_{42}^i\}\right].$$

After some algebra and using the fact that the probability $p(X_1, X_2, X_3, X_4 \mid Y_1, Y_2, Y_3, Y_4)$ does not depend on the haplotype frequencies, we can reduce the inner expectation in the previous display to

$$\frac{p_{\hat{h}}(X_1^i, X_3^i)p_{\hat{h}}(X_2^i, X_4^i)}{p_{\tilde{h}}(X_1^i, X_3^i)p_{\tilde{h}}(X_2^i, X_4^i)}.$$

This shows that the ratio of the likelihoods is equal to

$$\prod_{i=1}^{n} E_{\tilde{h}}\left[\frac{p_{\hat{h}}(X_1^i, X_3^i)p_{\hat{h}}(X_2^i, X_4^i)}{p_{\tilde{h}}(X_1^i, X_3^i)p_{\tilde{h}}(X_2^i, X_4^i)}\,|\,\{X_{11}^i, X_{21}^i\}, \{X_{31}^i, X_{41}^i\}, \{X_{12}^i, X_{22}^i\}, \{X_{32}^i, X_{42}^i\}\right].$$

The latter expression also follows without algebra from the general formula with the full data Z taken equal to the pairs $(X_1^i, X_3^i), (X_2^i, X_4^i)$. The computation of the expectation in the last expression consists of averaging over the resolutions of the observed data into the ordered haplotypes of the sibs, with the appropriate weights. The details for this step are already given in section 2.

Under the null hypothesis of LE two times the log likelihood ratio statistic is asymptotically chi-squared distributed with 1 degree of freedom. The null hypothesis is rejected for values of this statistic larger than the α upper quantile of the χ_1^2-distribution.

6 Simulation studies and application to depression data

We evaluated the reliability of the estimation and testing methods in two simulation studies. We were particularly interested in the effect of misspecifiying the recombination fraction θ.

6.1. First simulation study, estimation

In the first simulation study we focussed on estimating the haplotype frequencies in the parents' and children's populations. Estimation of the haplotype frequency in the parents' population is of interest, because the likelihood ratio test statistic for testing LE (in the parents' population) is based on these estimates.

The genotype data of 1000 sibling pairs in the children's population were simulated by gene-dropping: first the haplotypes in the parents' population were simulated and thereafter the haplotypes of the children were found by 'dropping down' these haplotypes according to Mendel's laws and using a prespecified value of the recombination fraction. For each child the unordered genotypes at the two loci were inferred from their haplotypes, and estimates were computed solely based on these genotypes and family structure.

Data simulation and haplotype estimation was repeated 1000 times. The accuracy of the estimates was summarized by the weighted (sample) mean square error (MSE):

$$\text{MSE} = \frac{1}{1000} \sum_{m=1}^{1000} \sum_{i=0}^{1} \sum_{j=0}^{1} \left(\frac{\hat{h}_{ij}^m - h_{ij}}{h_{ij}} \right)^2$$

for $\hat{h}_{ij}^m$ the maximum likelihood estimator of the true haplotype frequency h_{ij} in the mth simulation. The MSE can be written as a sum of squared weighted sample biases of the four haplotype estimators,

$$\sum_{i=0}^{1} \sum_{j=0}^{1} \left(\frac{\hat{h}_{ij} - h_{ij}}{h_{ij}} \right)^2,$$

with $\hat{h}_{ij}$ the mean of all 1000 estimates $\hat{h}_{ij}^m, m = 1, \ldots, 1000$ and a sum of weighted (sample) variances of the four estimates:

$$\sum_{i=0}^{1} \sum_{j=0}^{1} \frac{1}{1000} \sum_{m=1}^{1000} \left(\frac{\hat{h}_{ij}^m - \hat{h}_{ij}}{h_{ij}} \right)^2.$$

The MSE and the squared bias for estimators of haplotype frequencies in children's and parents' populations computed in 30 different settings are reported in Table 7. The sum of variances equals the difference between the MSE and the sum of squared biases and can therefore be computed from the values given in the table. The five columns of the table correspond to five different simulation scenarios, each determined by a vector of haplotype frequencies and a recombination fraction in the parents' population. In the five scenarios the vector $(h_{00}, h_{01}, h_{10}, h_{11})$ and the recombination fraction were equal to $(0.17, 0.13, 0.25, 0.45)$ and $\theta = 0.5$; $(0.01, 0.09, 0.09, 0.81)$ with $\theta = 0.5$ and $\theta = 0.3$; and $(0.4, 0.1, 0.1, 0.4)$ with $\theta = 0.5$ and $\theta = 0.3$, respectively. The rows of the table correspond to different values of the recombination fraction used to compute the estimates; the odd rows give the MSE and the

Table 7. Weighted means square errors and the summed squared weighted biases for the different simulation studies

	Simulation 1		Simulation 2		Simulation 3		Simulation 4		Simulation 5	
θ	MSE	bias^2	MSE	bias^2	MSE	bias^2	MSE	bias^2	MSE	bias^2
0.0	0.875	0.868	0.082	0.028	0.072	0.021	0.033	0.028	0.023	0.018
	1.477	1.470	0.082	0.028	0.072	0.021	1.860	1.847	0.716	0.706
0.1	0.860	0.854	0.068	0.009	0.061	0.006	0.019	0.014	0.010	0.004
	1.417	1.409	0.081	0.011	0.072	0.008	1.407	1.393	0.353	0.342
0.2	0.846	0.839	0.063	0.002	0.058	0.001	0.009	0.004	0.006	0.000
	1.347	1.337	0.091	0.003	0.083	0.001	0.921	0.888	0.114	0.091
0.3	0.831	0.824	0.062	0.000	0.058	0.000	0.005	0.000	0.005	0.000
	1.260	1.248	0.113	0.000	0.105	0.000	0.401	0.374	0.014	0.000
0.4	0.819	0.812	0.060	0.000	0.057	0.000	0.006	0.002	0.007	0.002
	1.164	1.150	0.144	0.000	0.137	0.000	0.090	0.068	0.081	0.066
0.5	0.822	0.815	0.053	0.000	0.051	0.000	0.004	0.000	0.027	0.023
	1.094	1.077	0.172	0.000	0.168	0.000	0.023	0.000	0.228	0.210

Note: The values for the children's population are at the odd rows, and those for the parents' population at the even rows.

squared bias for the haplotype estimates in the children's population whereas the even rows show these values for the parents' population. For the first, second and fourth column the last two rows give estimates based on the true value and the other rows estimates based on a misspecified model, whereas for the third and fifth column the seventh and eighth rows refer to the correctly specified model. It is clear that the bias of the estimates grows with the misspecification of the recombination fraction; so incorrectly assuming that the recombination fraction equals zero, like the existing methods do (Becker and Knapp (2002, 2004) and Putter *et al.* (2007)), is not sensible. Furthermore, the estimates are worse for the parents' population than for the children's population. Although the mean squared error and the bias vary with the haplotype frequencies there is no indication that misspecification of the recombination fraction is worse for some haplotype frequencies. However, only a few settings have been considered and strong conclusions concerning this can not been made.

We performed a second study to evaluate the effect of slight misspecification of the recombination fraction. This is of interest, since nowadays the position estimates of genetic markers are fairly precise. We simulated the data on 1000 sibling pairs using haplotype frequencies $(0.17, 0.13, 0.25, 0.45)$ in the first study and $(0.4, 0.1, 0.1, 0.4)$ in the second one and recombination fraction $\theta = 0.25$ for both, and computed the maximum likelihood estimates of the haplotype frequencies in the children's population using five different recombination fractions: $\theta =$ 0.20, 0.23, 0.25, 0.27, 0.30. The weighted mean square errors based on 1000 replications and the corresponding sum of squared weighted sample biases were equal to 0.675, 0.670, 0.668, 0.667, 0.666 and 0.663, 0.658, 0.656, 0.655, 0.654, respectively in the first study and in the second study the MSE equalled 0.00642, 0.00626, 0.00618, 0.00612, 0.00612 and 1000 times the squared weighted sample biases were 0.0826, 0.0193, 0.0145, 0.0249, 0.1267. We conclude that the estimation method is robust against small deviations of the assumed recombination fraction from the true value.

6.2. Second simulation study, testing

In the second simulation study we evaluated the reliability of the likelihood ratio test for testing linkage equilibrium by determining its level for a range of values of the haplotype frequencies and recombination fraction in the parents' population. The data were simulated under the null hypothesis of linkage equilibrium, so the haplotype frequencies were chosen so that $h_{kl} = h_{k.}h_{.l}$ for $k, l \in \{0, 1\}$.

In the first simulation the genotypes of 10,000 sibling pairs were simulated under the assumptions that $h_{00} = h_{01} = h_{10} = h_{11} = 0.25$ and $\theta = 0.5$. The likelihood ratio test statistic based on the genotypes of the sibs was compared to the 0.05 upper-quantile of the chi-squared distribution with 1 degree of freedom to decide whether the null hypothesis was rejected or not. This whole procedure, simulation, estimation and testing, was repeated 10,000 times. The fractions of rejected hypotheses, an estimate of the level of the test, were equal to 0.0296, 0.0493 and 0.0508 when using the three recombination fractions 0.0, 0.4 and 0.5 in the algorithm of section 5 respectively. Thus misspecification of the recombination fraction (the first two of the three cases) lead to a conservative test. In a second simulation study the true recombination fraction was lowered from 0.5 to 0.3 (with the haplotype frequencies unchanged at 0.25). The estimated levels of the likelihood ratio tests were 0.0332, 0.0487, 0.0504, 0.0530 and 0.0557 when using the recombination fractions 0.0, 0.25, 0.30, 0.35 and 0.5, respectively, thus showing conservativeness for underspecification of the recombination fraction and a slight increase in the level under overspecification. In a third simulation the vector of haplotype frequencies was taken equal to $(0.2 \times 0.4 = 0.08, 0.2 \times 0.6 = 0.12, 0.8 \times 0.4 = 0.32, 0.8 \times 0.6 = 0.48)$ and the recombination fraction $\theta = 0.5$. This lead to estimated levels 0.1224, 0.0484 and 0.0494 when using recombination fractions 0.0, 0.4 and 0.5, respectively, thus showing a much larger true level than the nominal level under underspecification of the recombination fraction, in contrast to the first two studies.

Based on these results we conclude that the level of the test is robust against small deviations of the assumed recombination fraction from the true value, but not against large deviations. Assuming that the recombination fraction equals zero when the true value is close to 0.5 can lead to invalid tests, where the direction of the deviation from the nominal level depends on the haplotype frequencies.

6.3. Application to depression data

This work was partly motivated by the investigation of epistatic interaction between genes associated with childhood depression/anxiety. Two physically unlinked single nucleotide polymorphisms (SNPs) that have been previously shown to jointly affect phenotypic variation were tested to be in linkage disequilibrium. The SNP rs902790 in GPR156, a GABA(B) related G-protein coupled receptor, and the SNP rs1979370 in DNAI2 have been predicted to have gene-gene interactions from a genome-wide screen of two-locus population differentiation. Subsequently, a significant interaction effect between the two genes and childhood depression/anxiety has been found.

Table 8. Estimates of the haploptype frequencies of SNPs in GPR156 and DNA12 genes of a Dutch twin cohort, under full and null hypothesis

Full		Null	
0.0858553	0.01626504	0.09364481	0.00853542
0.8305960	0.06728365	0.8228222	0.07499757

Therefore, we test here the hypothesis that these two non-synonymous SNPs are in linkage disequilibrium in the same cohort where the phenotypic association has been found. Based on a cohort of 483 monozygotic and 476 dizygotic twin pairs (see, Boomsma, van Beẅsterveldt and Hudziak (2005) for a description of this cohort) we estimated the four haplotype frequencies under both full model and null hypothesis and $\theta = 0.5$ as given in Table 8. Furthermore, we calculated the p-value of the likelihood ratio test of the null hypothesis of linkage equilibrium to be $p = 0.21$, indicating a lack of evidence of LD between the two SNPs (if the recombination fraction θ is taken equal to 0.0 the p-value would be 0.26). We also calculated the estimates based on only the monozygotic twin pairs, and found differences in only the second decimal of the estimates, indicating only a minor difference between monozygotic and dizygotic twins.

The prediction that the two epistatically interacting SNPs should be in LD could not be confirmed in this dataset. However, it is interesting to note that in the original paper it was the double heterozygous genotype that exhibited the highest levels of depression/anxiety, driving the epistatic effect in the phenotypic association. However, the double heterozygous genotype class is a mixture of all four haplotypes, possibly explaining the lack of epistasis induced LD in this particular situation.

7 Discussion

In this article we derived the likelihood ratio test for testing linkage disequilibrium from unphased genotypic data from a sample of siblings lacking genotypic information on their parents. It was essential to take the family structure into account, as otherwise the significance level of the test would be incorrect. To this aim the genotypic information on the parents was viewed as missing data, after which the maximum likelihood estimators of the haplotype frequencies under the null hypothesis of LE and under the full model could be computed with the EM-algorithm, and the likelihood ratio statistic by a similar algorithm.

Our algorithm differs from algorithms in standard packages by not assuming that the parents' population is in LE, and differs from algorithms in Becker and Knapp (2004) and Putter *et al.* (2007) by not assuming that the recombination fraction between loci of interest is equal to zero. The algorithm is more involved and more time consuming than in the last references, because fewer of the 64 numbers of $q(x_1 \mid y_1, y_2)$ for $x_1, y_1, y_2 \in \mathcal{X}$ are equal to zero. One application of the algorithm is in searching for epistatic effects of linked or unlinked loci that have lead to LD in the population by selection. Wrongly assuming that the recombination fraction is equal to zero may yield biased estimates of the haplotype frequencies and wrong test results. This was illustrated in a simulation study. The simulation study also showed that a slight misspecification of the recombinaton fraction hardly affected the haplotype estimates and the level of the chi-squared test for LE. So, even if the recombination fraction is not exactly known and a (rough) estimate is used, the results are reliable.

The maximum likelihood estimators of the haplotype-frequencies in the parents' population are computed with an EM-algorithm. Confidence regions for these frequencies can be constructed from the likelihood ratio test-statistic for testing the null hypothesis that $h_{kl}=h^0_{kl}$ for $k,l=0,1$ for given haplotype frequencies $h^0_{kl},k,l=0,1$. The $(1-\alpha)100\%$-confidence region for the haplotype frequencies equals all values for $h^0_{kl},k,l=0,1$ for which the null hypothesis would not be rejected (with the level of the test equal to α).

In section 4 we have compared the efficiency of estimating single locus frequencies based on sibs and based on a random sample of an equal number of unrelated individuals. If available, a random sample is preferable, but the loss in efficiency is small *provided* the maximum likelihood estimator as discussed in this article is used on the sibs and not the more obvious empirical estimator. These results are in line with the results obtained by Putter *et al.* (2007).

The algorithm presented in this article is restricted to two loci and applicable to sib-data only. However, the ideas behind the algorithm apply to multiple loci and more general pedigrees. Extension to multiple loci is particularly straightforward, although computational efficiency in the case of a large number of loci will require efficient pre-analysis of the possible genotypes given the data, as is done in algorithms for unrelated individuals. The proposed method is computationally heavy and might therefore not be feasible for genome-wide analysis. (In our current R implementation testing one pair of loci takes about 1/2 a second on a laptop with Intel T7300 2 GHz processor.) Interesting pairs or trios of loci could be selected at forehand in order to speed up the computations. These pairs or trios may be found based on previous findings (in literature), biological relationships (pathways) or by using a fast algorithm in a first step. Ignoring the sibling correlation in the estimation of haplotype frequencies yields unbiased estimates, and the computations are much faster. After selecting the most interesting pairs and trios the haplotype frequencies can be re-estimated, now incorporating the correlations between the siblings.

References

Balding, D. (2006), A tutorial on statistical methods for population association studies, *Nature Reviews Genetics* **7**, 781–791.

Becker, T. and M. Knapp (2002), Efficiency of haplotype frequency estimation when nuclear family information is included, *Human Heredity* **54**, 45–53.

Becker, T. and M. Knapp (2004), Maximum-likelihood estimation of haplotype frequencies in nuclear families, *Genetic Epidemiology* **27**, 21–32.

Bochdanovits, Z., D. Sondervan, S. Perillous, T. van Beijsterveldt, D. Boomsma and P. Heutink (2008), Genome-wide prediction of functional gene-gene interactions inferred from patterns of genetic differentiation in mice and men. *PLoS ONE* **3**, e1593.

Boomsma, D. I., C. E. M. van Beijsterveldt and J. J. Hudziak (2005), Genetic and environmental influences on anxious/depression during childhood: a study from the Netherlands twin register, *Genes Brain and Behavior* **4**, 466–481.

Cordell, H. J. (2009), Detecting gene-gene interactions that underlie human diseases, *Nature Reviews Genetics* **10**, 392–404.

Excoffier, L. and M. Slatkin (1995), Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population, *Molecular biology and evolution* **12**, 921–927.

Phillips, P. C. (2008), Epistasis – the essential role of gene interactions in the structure and evolution of genetic systems, *Nature Reviews Genetics* **9**, 855–867.

Pollak, E. (1979), Some models of genetic selection, *Biometrics* **35**, 119–137.

Putter, H., I. Meulenbelt and J. C. van Houwelingen (2007), Relative efficiency of haplotype frequency estimation in sibshipsand nuclear families compared to unrelated individuals, *Human Heredity* **64**, 52–62.

Ritchie, M. D., L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupontt, F. F. Parl and J. H. Moore (2001), Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer, *American Journal of Human Genetics* **69**, 138–147.

Schaid, D., S. Mc Donnell, L. Wang, J. Cunningham and S. Thibodeau (2002), Caution on pedigree haplotype inference with software that assumes linkage equilibrium, *American Journal of Human Genetics* **71**, 992–995.

Slatkin, M. and L. Excoffier (1996), Testing for linkage disequilibrium in genotypic data using the expectation-maximization algorithm, *Heredity* **76**, 377–383.

Zhang, Y. and J. S. Liu (2007), Bayesian inference of epistatic interactions in case-control studies, *Nature Genetics* **39**, 1167–1173.

Received: October 2009. Revised: December 2010.