METHODS

Haplotyping and Estimation of Haplotype Frequencies for Closely Linked Biallelic Multilocus Genetic Phenotypes Including Nuclear Family Information

K. Rohde1* and R. Fuerst2
1Max-Delbrueck-Centrum for Molecular Medicine, Berlin, Germany
2German Resource Center for Genome Research, Berlin, Germany

For the SNP 2000 Special Issue

With the discovery of single nucleotide polymorphisms (SNP) along the genome, genotyping of large samples of biallelic multilocus genetic phenotypes for (fine) mapping of disease genes or for population studies has become standard practice. A genetic trait, however, is mainly caused by an underlying defective haplotype, and populations are best characterized by their haplotype frequencies. Therefore, it is essential to infer from the phase-unknown genetic phenotypes in a sample drawn from a population the haplotype frequencies in the population and the underlying haplotype pairs in the sample in order to find disease predisposing genes by some association or haplotype sharing algorithm. Haplotype frequencies and haplotype pairs are estimated via a maximum likelihood approach by a well-known expectation maximization (EM) algorithm, adapting it to a large number (up to 30) of biallelic loci (SNP), and including nuclear family information, if available, into the analysis. Parents are treated as an independent sample from the population. Their genotyped offspring reduces the number of potential haplotype pairs for both parents, resulting in a higher accuracy of the estimation, and may also reduce computation time. In a series of simulations our approach of including nuclear family information has been tested against both the EM algorithm without nuclear family information and an alternative approach using GENEHUNTER for the haplotyping of the families, using the locus-by-locus allele counts of the sample. Our new approach is more precise in haplotyping in cases of a high number of heterozygous loci, whereas for a moderate number of heterozygous positions in the sample all three different approaches gave the same perfect results. Hum Mutat 17:289–295, 2001. © 2001 Wiley-Liss, Inc.

KEY WORDS: SNP; haplotype frequency estimation; expectation maximization algorithm; nuclear family information; haplotyping

INTRODUCTION

With the advent of the SNP technology, large samples of biallelic multilocus genetic phenotypes (i.e., genotypes with the phase at heterozygous loci not known) are common practice. These samples can be used either to estimate the haplotype frequencies in the population the samples were drawn from, or to find the most likely haplotype pair for each individual in the sample, or both.

In order to differentiate between populations, the occurring haplotypes and the accuracy of their estimated frequencies are essential.

For finding associations between a given genetic trait and some haplotype [Zhao et al., 2000; Fallin and Schork, 2000] by case control or transmission disequilibrium test (TDT) studies, the

Received 16 October 2000; accepted revised manuscript 11 January 2001.

*Correspondence to: Klaus Rohde, Bioinformatics, Max-Delbrueck-Centrum for Molecular Medicine, Robert-Roessle-Str. 10, D-13092 Berlin, Germany. E-mail: rohde@mdc-berlin.de
right haplotype pair for each individual counts, even if the estimated haplotype frequency in the population may not be perfectly true.

The maximum likelihood estimation of the haplotype frequencies in the population is carried out by a standard EM algorithm [Weir, 1990; Xie and Ott, 1993; Excoffier and Slatkin, 1995; Hawley and Kidd, 1995] under the assumption of Hardy Weinberg equilibrium (HWE) and random mating in the population. The assumption of HWE makes it possible to describe the probability of each genetic phenotype in the sample by a sum of underlying haplotype pairs with their probability, which is given as the product of the haplotype frequencies for each pair. The EM algorithm finds then, starting at given haplotype frequencies, in an iterative manner, those haplotype frequencies which make the drawn sample (the genetic phenotypes) the most likely. That those estimated haplotype frequencies reflect the population is only true if the population is in HWE and random mating and the sampling error is not too high [Fallin and Schork, 2000].

We, however, are not so much interested in population studies, but focus on the association of (complex) genetic traits to some haplotypes. For association and TDT studies the correct haplotyping is important because we count here occurrences of haplotypes pairs with their probability, which is given as the product of the haplotype frequencies for each pair. The EM algorithm finds then, starting at given haplotype frequencies, in an iterative manner, those haplotype frequencies which make the drawn sample (the genetic phenotypes) the most likely. That those estimated haplotype frequencies reflect the population is only true if the population is in HWE and random mating and the sampling error is not too high [Fallin and Schork, 2000].

In order to do this, we intend to include nuclear family information into the analysis, where the haplotype frequency estimation is done by the EM algorithm over the independent parents, and the children's genetic phenotype is used to exclude those haplotype pairs from the analysis, which are however possible for the parents but contradictory for the children. Exclusion of those “misleading” haplotype pairs will improve the finding of the correct underlying haplotype pairs in the sample, but on the other hand, may also serve to improve the haplotype frequency estimation, if the population is in HWE and random mating.

Our approach using nuclear family information is especially designed to be complemented by a TDT test [Spielman and Ewens, 1996], since at the end of the EM algorithm for the sample of nuclear families, we remain with the most likely haplotype pair constellation for every family in the sample with the transmitted and non-transmitted haplotypes for the children.

**METHOD**

Without sampling nuclear families, the sample data consist of a set of $k$ distinct phenotypes $g_i$ with their multiplicity $m_i$, and

$$
\sum_{i=1}^{k} m_i = m
$$

where $m$ is the number of individuals in the sample. The likelihood function of the data, given the genotypes is

$$
L(\text{data} | g_1, \ldots, g_k) = \prod_{i=1}^{k} P(g_i)^{m_i}
$$

where we have already skipped a multinomial factor which drops out later on, and $P(g_i)$ is the frequency of $g_i$ in the population. In case of biallelic loci (e.g., wild type and mutated allele), the maximal number of different haplotypes $l$ with $n$ loci is $l = 2^n$, whereas the number of different genotypes (haplotype pairs) $e_j$, leading to the $j$-th genetic phenotype with $z$ heterozygous loci is given by $e_j = 2^{j-1}$.

Assumption of random mating and Hardy Weinberg equilibrium (HWE) leads to the equation

$$
P(g_e) = L_k(y_1, \ldots, y_l) = \sum_{i=1}^{l} \sum_{j=1}^{z} c_{ij} y_i y_j,
$$

with the constraint

$$
\sum_{i=1}^{l} y_i = 1
$$

where $y_i$ is the frequency of haplotype $i$ (with which it is drawn from the population), and
If haplotype pair \((i,j)\) is compatible to phenotype \(g_i\),
\[ c^k_i = \begin{cases} 1 & \text{if haplotype pair } (i,j) \text{ is compatible} \\ 0 & \text{else} \end{cases} \]

serves to skip all those haplotype pairs from the sum, which cannot be explained (which are not compatible) with the genetic phenotype \(g_k\).

\(L_k(y_1,...,y_l)\) has two meanings. On one hand it is the probability of phenotype \(g_k\), given the haplotype frequencies, on the other hand it is a sum over all possible haplotype pair configurations which phenotype \(g_k\) may take on, weighted with its probability, in this way playing the role of an “ensemble sum of states.”

Those haplotype frequencies, which make the sample the most likely one under the constraint, are given by the maximum of the likelihood function under the constraint, as function of the \(y_i\). Solving the system of equations with the Lagrange multiplicator \(\lambda\)

\[
\frac{\partial}{\partial \log(y_{y_j})} \sum_{s=1}^{k} m_s \log(\sum_{i,j} c^k_{ij} y_{y_j}) + \lambda (1 - \sum_{i=1}^{l} y_{y_i}) = 0,
\]

and the additional partial derivative with respect to \(\lambda\), one gets the expectation maximization recursion

\[
y^{(t+1)}_{y_i} = \frac{1}{2m} \sum_{s} m_s \sum_{i,j} c^k_{ij} y^{(t)}_{y_j} y^{(t)}_{y_j} (1 - \sum_{i,j} c^k_{ij} y^{(t)}_{y_j} y^{(t)}_{y_j}), \quad r = 1,...,l
\]

where the haplotype frequency \(y_{y_i}\) for haplotype \(r\) in the step \(t+1\) is given by all haplotype frequencies a step before and the factor \(c^k_{ij}\) counts, how often (0, 1 or 2 times) haplotype \(r\) occurs in the haplotype pair \((i,j)\).

Generally one starts the iteration under the assumption of equal prior probability for all haplotype frequencies, in which case the haplotype frequencies on the right hand side can be canceled. As a special feature, we take in the first iteration step only the most likely haplotype pair of the sum (and if there are more equally likely, then the sum of them) into account, and find the next estimation of the haplotype frequencies by counting over all those most likely haplotype pairs in the sample. We have found that the haplotype frequencies found in this way are close to the global solution of the EM algorithm and provide an excellent starting point for further iterations, if one excludes the possibility that the iterations will get stuck for haplotypes which are not under the most likely haplotype pairs in the sample and have therefore a zero frequency after this step. This choice of a starting point avoids the general practice of starting the routine several times in order to exclude local optima, and works quite well as the performance of our approach in our simulation studies (see Tables 1 and 2) will prove.

The iteration stops as soon as the improvement of the likelihood function or of the haplotype frequencies are smaller than a pre-set limit.

Introduction of nuclear family information into the expectation maximization algorithm is done in three steps:

1. For the expectation maximization step, use only the genetic phenotypes of the parents, since children are not drawn independently from the population.
2. From the lists of compatible haplotype pairs for each parent in a family, construct a (longer) list of all pairs of compatible haplotype pairs for each couple. That means that the sum over different genetic phenotypes in the former expectation maximization step will be replaced by a sum over all families in the sample, with their couples of compatible haplotype pairs with their probability.
3. At this stage, the genetic phenotype information of the children is used to remove all couples of haplotype pairs from the family’s list, which are contradictory to the genetic phenotype of the children. If, for example, both parents are completely heterozygous, only one completely homozy-
gous offspring would reduce the multitude of compatible couples of haplotype pairs to one exactly determined haplotype pair couple. This step serves additionally as a kind of pedigree check. If it happens that after comparing the children’s genetic phenotype, the list of non contradictory couples of haplotype pairs for a family is empty, then there is an error in the pedigree, or we have found a recombination. In both cases, these families should be omitted from the analysis. However, the possibility of a recombination may be valuable information. For closely linked SNP, a recombination in nuclear families should be a rare event, even at higher recombination rates.

Having done this, the expectation maximization step reads

$$y_r^{(t+1)} = \frac{1}{4f} \sum_{j=1}^{f} \sum_{j=1}^{f} \sum_{i=1}^{f} c_{ijw} c_{ijw}^{(t)} y_{ijw}^{(t)} y_{ijw}^{(t)} y_{ijw}^{(t)} y_{ijw}^{(t)}$$

where \( f \) is the number of families in the sample, and

$$c_{ijw}^{(t)} = \begin{cases} 1 & \text{if in family } s \text{ the compatible couple of } \text{haplotype pairs } (ij,uv) \text{ is non-contradictory to the children} \\ 0 & \text{else} \end{cases}$$

and \( z_{ijuw}^{(t)} \) counts, how often (0, 1, 2, 3, or 4 times) haplotype \( r \) occurs in the couple of haplotype pairs \( (ij,uv) \).

RESULTS OF SIMULATION STUDIES

In order to evaluate our approach, we carried out a series of simulation studies. Under assumption of equally frequent haplotypes in the population, we simulated samples of 25 families with one or two children. Three sets of haplotypes were chosen, from low to high numbers of heterozygous positions in the sample, for a choice of 10 different, equally frequent haplotypes with 10 loci, each, (see Table 1) and a choice of four different, equally frequent haplotypes with four loci, each, (Table 2). In this way, we can test the influence of both the amount of heterozygosity (phase-unknown positions) in the sample and the number of loci per haplotype on the performance. The three choices of the 10 locus haplotypes with a frequency of 0.1, each, are: (1111111112; 1111111121; 1111112111; ...; 1211111111; 2111111111) with an allele frequency for the mutated allele 2, \( P(2) = 0.1 \), at each locus and a frequency of heterozygous positions \( P(\text{het}) = 0.18 \) (first row in Table 1); (1111111111; 1111111111; 1111111122; ...; 1222222222) with increasing allele frequencies \( P(2) = 0.0 \) (0.25) 0.75 at successive loci, leading to \( P(\text{het}) = 0.33 \) (second row in Table 1); and (1111111111; 2111112222; 2211112222; ...; 1111222222) with allele frequency \( P(2) = 0.50 \) at each locus and \( P(\text{het}) = 0.50 \) (third row in Table 1).

The respective choices for the four locus haplotypes of equal frequency 0.25 are (1112; 1121; 1211; 2111) with allele frequency \( P(2) = 0.25 \) at each locus and a frequency of heterozygous positions \( P(\text{het}) = 0.38 \) (second row in Table 2); (1111; 1112; 1122; 1222) with increasing \( P(2) = 0.0 \) (0.25) 0.75 for successive loci and \( P(\text{het}) = 0.31 \).
One must not mistake the frequency of heterozygous positions in the sample \( P(het) \) with the frequency of heterozygotes in the sample \( p_{het} \). The frequency of heterozygotes in the sample for \( n \) different haplotypes with probability \( p_i \) is \( p_{het} = 1 - \sum_{i=1}^{n} p_i^2 \), that means 0.9 for the set of 10, and 0.75 for the set of four equally frequent haplotypes, irrespective of the special choice of the haplotypes. The heterozygotes, however, may differ in the fraction of heterozygous (phase-unknown) positions in the genotype; a heterozygote may carry at least one heterozygous position and, at most, the total number of loci in the genotype. This fraction of heterozygous positions in the genotype depends clearly on the special choice of the haplotypes. The frequency of the heterozygous positions in the sample \( P(het) \) is the sum over all heterozygotes with their frequency \( p_ip_j \) multiplied by the fraction of heterozygous positions in their genotype \((i,j)\).

The complexity of the haplotype estimation depends not so much on the frequency of heterozygotes in the sample (they follow the HWE), but on the frequency of heterozygous position in the sample \( P(het) \).

Our choice of the haplotype sets may seem artificial, however the choice of equally frequent haplotypes ensures a maximum number of heterozygotes in the sample, and the special choice of the haplotype patterns span a range from a relative low fraction of heterozygous positions (0.18, only two of 10 positions heterozygous in the genotype) up to a high fraction of heterozygous positions (0.50; the highest possible with two alleles under HWE at one locus). Real populations, with frequencies of haplotypes high enough to have a reasonable chance to be drawn in samples of 100 individuals, may lie somewhere in this range.

For each different set of conditions, we simulated 1000 replicates and tested the performance of our approach with inclusion of nuclear families together with two alternative approaches, i) the genuine haplotype estimation, using the genetic phenotype information of only the parents (part of our routine), and ii) using GENEHUNTER [Kruglyak et al., 1996] for haplotyping the families in the samples, inferring the allele frequencies at each locus by counting alleles locus by locus in the parents of each sample.

If one compares the genuine EM algorithm, using only parent information, with our new approach of including nuclear family information, on the basis of the same number of families in the sample, one could argue that our new approach would need additional genotyping of the children, in this way providing additional information. In order to analyze this effect, we have compared a simulated set of 50 parents without children, with our 25 family sample set with two children in the family, with both samples having the same burden of genotyping. The results of the simulation for the 50 family samples without offspring information are given in parenthesis in the first columns of Tables 1 and 2.

Since we are more interested in the correct haplotyping of the individuals in the sample than

<table>
<thead>
<tr>
<th></th>
<th>( P(het) )</th>
<th>EM only par.</th>
<th>EM nucl. fam.</th>
<th>GENEHUNTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 child</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.31</td>
<td>100*</td>
<td>100*</td>
<td>95.7</td>
<td></td>
</tr>
<tr>
<td>0.38</td>
<td>100*</td>
<td>100*</td>
<td>85.9</td>
<td></td>
</tr>
<tr>
<td>0.50</td>
<td>87.6</td>
<td>94.7</td>
<td>76.7</td>
<td></td>
</tr>
<tr>
<td>2 children</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.31</td>
<td>100* (100)*</td>
<td>100*</td>
<td>97.7</td>
<td></td>
</tr>
<tr>
<td>0.38</td>
<td>100* (100)*</td>
<td>100*</td>
<td>93.2</td>
<td></td>
</tr>
<tr>
<td>0.50</td>
<td>87.1 (87.5)</td>
<td>96.5</td>
<td>87.2</td>
<td></td>
</tr>
</tbody>
</table>

\( P(het) \) is the frequency of heterozygous positions in the sample, and * means not a single false haplotype in all simulations; a) EM only par.—Expectation maximization only over parents (values for 50 couples in parenthesis); b) EM nucl. fam.—Expectation maximization using nuclear family information; c) GENEHUNTER—Haplotyping of the sample via GENEHUNTER with allele frequencies calculated locus by locus from the sample.
in the correct haplotype frequencies in the population, we compare the number of correctly estimated haplotypes in the parents for each of the simulated samples, keeping in mind that perfectly estimated haplotypes in the sample reflect, of course, the frequency with which the haplotypes were drawn via the individual genetic phenotypes from the sample.

**DISCUSSION OF SIMULATIONS**

The results of our simulations in Tables 1 and 2 show, not surprisingly, the trend that all three different approaches, EM (only parents), EM (nuclear families), and GENEHUNTER perform better if the number of heterozygous positions in the sample is lower. A general trend is, at least in our confined set of simulations, that both EM algorithms, with and without family information, perform better than GENEHUNTER. Both family based approaches, EM (nuclear families) and GENEHUNTER, work better with two offspring rather than one. However, this effect is definitely more striking for GENEHUNTER.

For low to moderate frequencies of heterozygous positions in the sample, \( P(\text{het}) \), both EM algorithms are comparable, with a high rate of up to 100%, and GENEHUNTER follows closely. That means, that in these cases the choice of the approach is not so important, and the inclusion of family information (and with that, additional genotyping) is not worthwhile, if one is interested only in haplotyping and simple association to a genetic trait. However, if one plans a subsequent TDT test, one needs children and their correctly estimated haplotype pairs anyway, and should use our approach, which is more stable under all simulated conditions.

For a higher number of heterozygous positions in the sample, our new approach, EM with inclusion of nuclear family information, definitely performs better than both the EM (only parents) and GENEHUNTER, which is obviously not designed for this kind of problem, and even estimates haplotypes, which were not in the simulated sample.

One may observe the trend that, in contrast to GENEHUNTER, both EM algorithms perform better with the set of the 10 locus haplotypes than with that of the four locus haplotypes. A reason for this might be found in the different basic approaches; the EM algorithms work right from the start with haplotypes, whereas GENEHUNTER relies on locus-by-locus allele frequencies.

By comparing the EM algorithm only over parents (50 families) with the EM algorithm over nuclear families (25 families with two children, each), we find that doubling the sample size in EM (only parents) gives no definite improvement of the haplotyping. This is not really surprising, since a higher sample size will reduce the sampling error and lead to more exactly estimated haplotype frequencies in the population, however in most cases, with only minor changes in the haplotype frequencies. Therefore the more imprecise haplotype frequencies found with a smaller sample size are already sufficient to find the exact most likely haplotype pairs for each individual in the sample.

In general, we may conclude that, especially if one has in mind a subsequent TDT test of linkage and association to a genetic trait, our approach of including nuclear family information into the EM algorithm of haplotype frequency estimation should be advisable. The programs are written in C, run under Solaris™ and Linux, and are available on request from the authors (rohde@mdc-berlin.de). The limit of the calculations is at present not so much the number of loci per haplotype (up to 30), but more the number of heterozygous positions in the genotypes, which may give rise to an enormous number of possible (to be processed) haplotype pairs, placing high demand on memory and computing power.

**REFERENCES**


