Original Paper

Human Heredity

Hum Hered 2006;62:12–19
DOI: 10.1159/000095598

Received: March 3, 2006
Accepted after revision: July 19, 2006
Published online: September 5, 2006

Haplotype Reconstruction and Estimation of Haplotype Frequencies from Nuclear Families with Only One Parent Available

Xiangdong Ding[a, b] Qin Zhang[b] Christine Flury[a] Henner Simianer[a]

[a]Institute of Animal Breeding and Genetics, University of Goettingen, Goettingen, Germany;
[b]State Key Laboratories of Agrobiotechnology, Key Laboratory of Animal Genetics and Breeding, Ministry of Agriculture, College of Animal Science and Technology, China Agricultural University, Beijing, China

Key Words
SNP · Haplotype frequency estimation · Haplotype inference · EM algorithm · Incomplete nuclear family

Abstract
Recent literature has suggested that haplotype inference through close relatives, especially from nuclear families can be an alternative strategy in determining the linkage phase. In this paper, haplotype reconstruction and estimation of haplotype frequencies via expectation maximization (EM) algorithm including nuclear families with only one parent available is proposed. Parent and his (her) child are treated as parent-child pair with one shared haplotype. This reduces the number of potential haplotype pairs for both parent and child separately, resulting in a higher accuracy of the estimation. In a series of simulations, the comparisons of PHASE, GENEHUNTER, EM-based approach for complete nuclear families and our approach are carried out. In all situations, EM-based approach for trio data is comparable but slightly worse error rate than PHASE, our approach is slightly better and much faster than PHASE for incomplete trios, the performance of GENEHUNTER is very bad in simple nuclear family settings and dramatically decreased with the number of markers being increased. On the other hand, the comparison result of different sampling designs demonstrates that sampling trios is the most efficient design to estimate haplotype frequencies in populations under same genotyping cost.

Introduction

With the discovery of single nucleotide polymorphisms (SNP) along the genome, genotyping of large samples of biallelic multilocus genetic phenotypes for fine mapping of complex traits has become standard practice. Both simulation and empirical studies have demonstrated that statistical analysis based on haplotypes may be more efficient than separate analysis of individual markers [1]. Considerable research effort has been devoted to algorithms that infer haplotype phase from genotype data.

There is a growing number of articles on haplotype inference for unrelated individuals [2–4], however, these methods can not make use of family information effectively. Inferring haplotypes based on close relatives can be an alternative strategy, which can reduce haplotype ambiguity and improve the efficiency for haplotype frequency estimates [5–7]. Rohde and Fuerst [6] proposed an expectation-maximization (EM) algorithm [8] for the maximum likelihood estimation of haplotype frequencies using nuclear family information. They found that reconstruction based on maximum likelihood estimates including child information performed better than reconstruction based on maximum likelihood estimates using only individual information. However, in cases where we can obtain only the genotypes of one parent because the other parent is not available for study or not

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0001–5652/06/0621–0012$23.50/0

Accessible online at:
www.karger.com/hhe

Xiangdong Ding
Institute of Animal Breeding and Genetics, University of Goettingen
Albrecht-Thaer-Weg 3
DE–37075 Goettingen (Germany)
Tel. +49 551 39 5623, Fax +49 551 39 5587, E-Mail xding@gwdg.de

cooperative, the approaches mentioned above cannot handle such data, new approach should be developed.

Maximum likelihood implemented via EM algorithm is a very popular method for haplotype inference. In this paper, we propose a new maximum likelihood based method for haplotype reconstruction and estimation of haplotype frequencies for closely linked multilocus systems for nuclear families with only one parent available and an EM algorithm to obtain the corresponding maximum likelihood estimates. The simultaneous consideration of parent-child-pairs reduces the number of potential haplotype pairs substantially. We will provide results of a simulation study showing that our approach results in a higher accuracy of the estimation of population haplotype frequencies and of reconstructed individual haplotypes, as well as a reduced computation time. Possible applications of the suggested method will be discussed.

Methods

Definitions

We consider a series of N closely linked polymorphic loci. The individual genotype is the set of N single locus genotypes at those loci without any phase information, for N = 3, a possible genotype of individual i is Y_i = (12; 34; 56). A haplotype is defined as the ordered series of alleles on one of the homologous chromosomes of one individual, e.g. for Y_i, a possible first haplotype is H_{i1} = (1 4 5). The diplotype is defined as a particular combination of two haplotypes, e.g. $G_i = (H_{i1}, H_{i2})$ = (1 4 5, 2 3 6). Note that for a given genotype several diplotypes are possible.

For a parent-child pair composed of one parent and his (her) child, the child must inherit a haplotype from the parent and they will have at least one possible common haplotype, assuming that there are no recombination events (the case of recombination will be addressed in the discussion). Consider for example parent i and his (her) child j with genotype Y_i = (12; 34; 56) and Y_j = (11; 33; 56), their combined genotype can be symbolized as

$$YP_f = \begin{pmatrix} 12;34;56 \\ 11;33;56 \end{pmatrix}$$

At the first locus, the parent and the child share the common allele 1, at the second locus they share allele 3, and at the third locus they share either allele 5 or 6, so the possible common haplotypes between this parent-child pair are H = (1 3 5) or H = (1 3 6). The possible diplotypes for this parent-child pair then are

G_i = (1 3 5, 2 4 6) and G_j = (1 3 5, 1 3 6)

or

G_i = (1 3 6, 2 4 5) and G_j = (1 3 6, 1 3 5),

respectively. Thus, the haplotypes of a parent-child pair can be represented as (H_C,H_P,H_O), where H_C denotes the common haplotype and H_P and H_O represent the other two haplotypes of the parent and child, respectively, and is termed as parent-child haplotype pair (PCHP). For the given example, the two possible PCHPs are (135, 246, 136) and (136, 245, 135).

The Likelihood Function

Following similar arguments as presented by Excoffier and Slatkin [3], for a sample of m parent-child pairs, the likelihood function of the population haplotype frequencies is defined as

$$L(p_1, p_2, \ldots, p_h) = \prod_{f=1}^{m} \left(\sum_{i=1}^{S_f} P(H_C, H_P, H_O)_i \right) \quad (1)$$

where $p_1, p_2, \ldots, p_h$ are the population frequencies of all haplotypes with $\sum_{i=1}^{h} p_i = 1$, S_f is the number of possible common haplotypes for parent-child pair f, or equivalently the number of possible PCHPs, and $P(H_C,H_P,H_O)_i$ is the probability of the i-th possible PCHP $(H_C,H_P,H_O)_i$ for parent-child pair f.

The EM Algorithm

The EM algorithm is an iterative method to find the maximum likelihood estimates of haplotype frequencies in the observed data. In the expectation step, a set of initial or actual haplotype frequency values is used to calculate the probabilities of all possible PCHPs for all parent-child pairs. Then based on these expected PCHP probabilities, haplotype frequency estimations can be updated in the maximization step. The EM algorithm iterates between these two steps until haplotype frequency estimations converge (i.e., when the changes in haplotype frequency in consecutive iterations are less than some small value).

To obtain initial values of $p_1, p_2, \ldots, p_h$, it is assumed that for parent-child pair f all the possible PCHPs have the same probability, i.e.

$$P_f^{(0)}(H_C, H_P, H_O)_i = 1/S_f \quad (2)$$

These PCHP probabilities are used in eq. (1) to calculate the initial likelihood value. According to Ceppellini et al. [9], Smith [10] and Rohde and Fuerst [6] the population haplotype frequencies can be calculated in the first and in all subsequent iterations as

$$p^{(g+1)}(H_t) = \frac{1}{3m} \sum_{f=1}^{m} \sum_{i=1}^{S_f} \delta_{it} P_f^{(g)}(H_C, H_P, H_O)_i \quad (3)$$

where δ_{it} is an indicator variable equal to the number of times that haplotype H_t is present in the i-th possible genotype pair of parent-child pair f, its possible value is 0, 1, 2 or 3.

In the expectation step at the g-th iteration, the haplotype frequencies obtained in the previous iteration is used to calculate the probability of each possible PCHP as

$$P_f^{(g+1)}(H_C, H_P, H_O)_i = \frac{P^{(g)}(H_C, H_P, H_O)_i}{\sum_{j=1}^{S_f} P^{(g)}(H_C, H_P, H_O)_j} \quad (4)$$

where

$$P^{(g)}(H_C,H_P,H_O)_i = p^{(g)}(H_C)_i \, p^{(g)}(H_P)_i \, p^{(g)}(H_O)_i \quad (5)$$

according to eq. (5), only the population frequency of every PCHP can be obtained, and after the transformation based on eq. (4), the

population probability of each possible PCHP for one particular parent-child pair *f* is finally calculated.

Iterating between the E-step, using eq. (4) and (5) to calculate all PCHP probabilities, and the M-step, using eq. (3) to calculate all haplotype frequencies, until convergence yields the maximum likelihood estimate of the population haplotype frequencies.

Haplotype Reconstruction

After the EM algorithm reaches convergence, the population frequencies of all haplotypes in the population can be calculated. For association and TDT studies, the correct haplotype reconstruction is critical, because the occurrences of haplotypes in cases and controls, or haplotypes transmitted or non-transmitted in nuclear families need to be counted. For a parent-child pair with genotype combination YP_f, there will be several possible diplotypes, and subsequently a corresponding PCHP (H_C,H_P,H_O) to each possible diplotype GP_i. So the probability of each possible diplotype can be computed given the population haplotype frequencies and the genotype pair:

$$P\left(GP_i \middle| (p_1, p_2, \ldots, p_h), YP_f\right) = \frac{P\left(H_C, H_P, H_O\right)}{\sum_{j=1}^{S_f} P\left(H_C, H_P, H_O\right)_j} \quad (6)$$

The diplotype with the maximum conditional probability is the most likely diplotype for genotype pair *f* and can be split in the most likely diplotypes for parent and child.

Simulation Study

Simulated Data

In order to evaluate our approach, we carried out a series of simulation studies. We simulated haplotypes using Schaffner's simulation program [11] based on a coalescent model that incorporates variation in recombination rates. The parameters used for the simulation were: chromosome segment length: 1 Mb, mutation rate: $1.5 \cdot 10^{-8}$, variable recombination rate: $1 \cdot 10^{-8}$, effective population size: 10,000, number of sampled chromosomes: 120 or 180. From the simulated haplotypes, the diplotypes of related individuals were produced as follows: we first combined two randomly chosen haplotypes to be the diplotype of the first parent and two other randomly chosen haplotypes to form the diplotype of the second parent. The diplotype of their offspring was generated by randomly picking one of the two haplotypes of the father and mother, respectively. For incomplete families, the information on the missing parent was omitted after generating the child. Markers are thinned to obtain the required 1 SNP per 8 kb density that was used throughout the present study. In the different scenarios, haplotypes of 5, 10, or 20 SNPs were considered, respectively, and the number of families was varied between 30 and 45. For each scenario, 100 replicates were generated and analyzed.

Approaches to Be Compared

We compared four different approaches for analysis:

a) Complete-family-EM: the maximum likelihood estimation via the EM algorithm using complete nuclear family information with both parental genotypes available proposed by Rhode and Fuerst [6]. This algorithm makes use both of linkage disequilibrium (LD) and pedigree information.

b) PHASE: Stephens et al. [4] introduced two Bayesian approaches, which were implemented in the program PHASE [4, 12], PHASE was initially designed for unrelated individuals, but now the PHASE program can handle trio data as well [13]. Similarly, the incomplete trios can be analyzed by PHASE by setting one whole parent in a trio as missing.

c) GENEHUNTER: It is a very popular software for linkage analysis [14] which makes full use of the pedigree information but, as was noted by Schaid et al. [15], assumes that genetic markers are in linkage equilibrium. GENEHUNTER after convergence only provides information on the most likely haplotype and does not give its posterior probability.

d) Our approach: it only differs from the 'complete-family-EM' approach in that one parent is missing. The parameters were estimated with the approaches described in the method section and thus account both for LD and pedigree information.

Criteria

To evaluate the quality of haplotype frequency estimation, the indices I_F and I_H were used.

I_F is used to examine the discrepancy between the estimated frequencies and the actual frequencies. It was defined by Excoffier and Slatkin [3] as one minus half of the sum of absolute difference between estimated and true haplotype frequencies, i.e.

$$I_F = 1 - \frac{1}{2}\sum_{i=1}^{h}\left|\hat{p}_i - p_i\right| \quad (7)$$

where the $\hat{p}_i$ and p_i are the estimated and the true simulated frequency in the sample for the *i*-th haplotype, respectively. I_F varies between 0 and 1. The more accurate the estimation is, the closer I_F will be close to 1.

I_H is used to examine whether all the haplotypes present in the population are identified in the estimated haplotypes. In a population with *N* individuals, the minimum frequency for every true haplotype must be greater than or equal to $(2N)^{-1}$ which can be used as a lower threshold value for determining the existence of a haplotype, i.e. a haplotype is only accepted to be detected if its estimated frequency is above $(2N)^{-1}$. Based on this, Excoffier and Slatkin [3] suggested the statistic

$$I_H = \frac{2\left(k_{true} - k_{missed}\right)}{k_{true} + k_{found}} \quad (8)$$

where k_{true} is the number of true haplotypes, k_{found} is the number of identified haplotypes with frequency above the threshold value, and k_{missed} is the number of true haplotypes not identified. I_H also varies between 0 and 1. When all true haplotypes are identified, it will be 1, and when none of the true haplotypes are identified, it will be 0.

The best-guess haplotype frequencies are provided in the output of complete-family-EM, PHASE and our approach, which can be directly used to calculate I_F and I_H. Although they are not explicitly provided by GENEHUNTER, their frequencies can be calculated based on counting haplotypes in the reconstructed diplotypes of parents and children.

For accuracy of haplotype reconstruction of individual genotypes, error rate and haplotype reconstruction reliability I_R were used.

If the most likely diplotype of an individual is the same as the simulated true genotype, this individual will be considered as correctly haplotyped. The error rate is the proportion of not completely correctly haplotyped individuals in the population.

Even if the most likely diplotype of an individual is the correct one, the posterior probability of this diplotype may be substantially smaller than one. The overall quality of the haplotype reconstruction procedure can be evaluated with the average posterior probability of correctly reconstructed haplotypes, which is denoted as I_R. Since GENEHUNTER does not provide the posterior probability of the most likely haplotype, the statistic I_R could not be given for the GENEHUNTER analysis.

The reconstructed diplotypes of children are not available from PHASE, while they can be reconstructed using their own genotypes and the inferred diplotypes of their parents. For incomplete trios, the available parents are used, and for complete trios, the parent with higher posterior probability will be chosen if both parents are correctly haplotyped.

In the calculation of the error rate and I_R for the incomplete trio design, although the reconstructed diplotypes of missing parents can be provided by PHASE and GENEHUNTER, taking them into account will decrease the performance of these algorithms. Therefore the information of missing parent was discarded. Similarly, they were discarded in the calculation of I_F and I_H for GENEHUNTER.

Computing time of the algorithms was measured in seconds on an IBM server (SUSE Linux 9.2 and 3 GHz Intel Xeon processor).

Results

Performance

The comparisons of the EM-based approach with the Bayesian approach for complete and incomplete trios are shown in table 1 and 2. For complete trios, EM-based algorithms are generally equivalent to PHASE except that the values of I_H from complete-family-EM are higher than those from PHASE.

For incomplete trios, our approach performs comparable to PHASE in estimation of haplotype frequencies and haplotype reconstruction. However, the computing time is 2000 to 4000 times higher with PHASE compared to our approach. When reconstructing haplotypes of 10 and 20 SNPs, respectively, computing time of PHASE is

Table 1. Comparison of efficiency of complete-family-EM, PHASE and GENEHUNTER based on frequency discrepancy I_F, I_H, error rate and inference reliability I_R in case of 30 trios from 100 data sets

	Complete-family-EM	PHASE	GENEHUNTER
Number of SNP = 10			
I_F	0.9676	0.9683	0.9257
I_H	0.9938	0.9308	0.7920
Error rate	0.0040	0.0042	0.1557
I_R	0.9936	0.9907	
Computing time, s	0.2100	9.8700	0.4020
Number of SNP =20			
I_F	0.9676	0.9497	0.8552
I_H	0.9938	0.8943	0.6242
Error rate	0.0048	0.0046	0.3659
I_R	0.9974	0.9913	
Computing time, s	42.1440	36.5400	0.4500

Parameter I_R was not available for GENEHUNTER.

Table 2. Comparison of efficiency of our approach, PHASE and GENEHUNTER based on frequency discrepancy I_F, I_H, error rate and inference reliability I_R in case of 30 incomplete trios from 100 data sets

	Our approach	PHASE	GENEHUNTER
Number of SNP = 10			
I_F	0.9529	0.9495	0.8405
I_H	0.9636	0.9576	0.6629
Error rate	0.0357	0.0367	0.2877
I_R	0.9533	0.9530	
Computing time, s	0.1140	458.3880	0.3000
Number of SNP = 20			
I_F	0.9262	0.9259	0.8402
I_H	0.9134	0.9134	0.4302
Error rate	0.0356	0.0363	0.6152
I_R	0.9669	0.9515	
Computing time, s	1.9680	4689.0660	0.3300

Parameter I_R was not available for GENEHUNTER.

Table 3. Performance of GENEHUNTER under different number of markers in the case of 30 complete and incomplete trios from 100 data sets

	Complete trios			Incomplete trios		
Number of SNP:	5	10	20	5	10	20
I_F	0.9658	0.9257	0.8552	0.8430	0.8405	0.8402
I_H	0.9068	0.7920	0.6242	0.7598	0.6629	0.4302
Error rate	0.0603	0.1557	0.3659	0.1032	0.2877	0.6152

increased from 458 to 4689 s, indicating that computing time of PHASE may become prohibitive for larger haplotypes or data sets.

Although the reconstructed diplotypes of the missing parent can be given by PHASE, the accuracy for the missing parent is very low, the values of error rate are 0.67 and 0.84 in case of 10 SNPs and 20 SNPs, respectively.

With the assumption of markers in linkage equilibrium, the performance of GENEHUNTER is poor under the coalescent model, especially in the case of GENEHUNTER dealing with incomplete trios. On the other hand, as shown in table 1 and 2, GENEHUNTER is significantly affected by the number of SNP. The error rate is dramatically increased from 0.1657 to 0.3517 with the number of SNP being increased from 10 to 20, and the values of I_F and I_H are decreased as well. This is further proved by the results in table 3.

Efficiency of Different Sampling Designs

We compared three different sampling designs for haplotype inference: complete nuclear families, incomplete nuclear families and unrelated individuals. To compare their efficiency, we sampled 30 trios for complete-family-EM, 45 incomplete trios for our approach and 90 unrelated individuals for PHASE. The total number of individuals to be genotyped is uniformly 90 for each sampling design.

As shown in table 1 and 2, the performance of complete-family-EM and our approach is equivalent to PHASE respectively for complete and incomplete trios. Therefore, the differences observed between designs can be attributed to the different data structure and degree of pedigree information and is not caused by differences due to the estimation procedure. As shown in table 4, according to all criteria, the sampling design of complete nuclear families performs uniformly best, and sampling unrelated individuals performs worst.

Efficiency of Number of Children of Each Family

So far, PHASE can not handle those families with more than one child. In Rohde and Fuerst [6], increasing the number of children in nuclear families will result in the improvement of the efficiency of haplotype inference, because more children can provide more family information. The same conclusion is obtained in our simulation studies. As shown in table 5, the performance of GENEHUNTER and our approach in case of 30 families with two children each is higher than in case of 45 families with one child each. Note that the sampling and genotyping cost are same in these two cases. In general, adding a second child improves the quality of estimates. This is especially so for the strategies that performed poorly in the one child scenario of GENEHUNTER.

Table 4. Comparisons of efficiency of different sampling design: nuclear families (complete-family-EM for 30 trios), incomplete nuclear families (our approach for 45 incomplete trios) and unrelated individuals (PHASE for 90 individuals) with 20 SNPs

Design:	Complete-family-EM 30 complete trios	Our approach 45 incomplete trios	PHASE 90 individuals
I_F	0.9676	0.9337	0.9267
I_H	0.9938	0.9523	0.8090
Error rate	0.0048	0.0351	0.1127
I_R	0.9974	0.9718	0.8168

Discussion

Likelihood-based approaches via the EM algorithm and Bayesian approaches are prevailing in haplotype inference, and it is difficult to give a general conclusion about which one is more efficient. Although Stephens et

Table 5. Comparisons of haplotype frequency estimation and haplotype reconstruction with different number of children in incomplete nuclear families from our approach and GENEHUNTER

	Our approach		GENEHUNTER	
	45 families with 1 child each	30 families with 2 children each	45 families with 1 child each	30 families with 2 children each
I_F	0.9337	0.9383	0.8326	0.8704
I_H	0.9523	0.9539	0.4471	0.6172
Error rate	0.0351	0.0242	0.5808	0.3471
I_R	0.9718	0.9772		

Parameter I_R was not available for GENEHUNTER.

al. [4] demonstrated that PHASE outperformed EM by a significant margin under a coalescent model, and Marchini et al. [13] further proved that the performance of an EM-based approach for trio data is comparable but with a slightly worse error rate than PHASE. Zhang et al. [16] and Xu et al. [17] revealed that PHASE and EM-based methods exhibited similar performances in their simulated datasets. The results of our study confirm the conclusion of Marchini et al. [13] in the case of trio data, that complete-family-EM proposed by Rohde and Furst [6] is comparable to PHASE, albeit with a slightly higher error rate. However, in the case of incomplete trios, our approach is slightly more accurate and substantially faster than PHASE.

As shown in the study of Marchini et al. [13] PHASE is the slowest one in dealing with 30 trios with 2% missing data and 187 SNPs in the comparison with wphase, HAP, HAP2 and PL-EM. It takes PHASE 3 h and 32 min for the whole 100 datasets, and as shown in our study, it takes PHASE 42.144 seconds per dataset for 30 trios with 20 SNPs, while the computing time dramatically increased to 4689.066 s when one parent is randomly assumed missing. So it can be imagined that the computing time for PHASE will become prohibitive if among the 30 trios in the study of Marchini et al. [13] some are also assumed to have one parent missing at random.

For large chromosome segments, PHASE uses the partition-ligation strategy [18] to divide it into a lot of units to improve the speed and accuracy, this may be the reason that PHASE runs faster than complete-family-EM in case of 20 SNPs (table 1). However, for incomplete trios, PHASE tries to make inference of the diplotypes of the missing parents, which will take a long time to deduce the computing time dramatically increased compared with complete trios, while our approach uses the common haplotypes between parent and child to kick out much more unnecessary information, it makes our approach very fast, not exponentially increased as PHASE with SNPs being increased. So far, the number of loci handled by complete-family-EM is up to 30 [6]. Similarly, partition-ligation strategy can be introduced into complete-family-EM and our approach, too. It can make complete-family-EM to deal with more than 30 loci, and improve the efficiency and accuracy of complete-family-EM and our approach as well.

Although GENEHUNTER is extended to be capable of haplotype phase reconstruction using pedigree data [14], Becker and Knapp [5] showed that GENEHUNTER is not suitable for low-information content SNP marker in simplex nuclear family settings. The same conclusion can be made in our simulation study, where GENEHUNTER performs much worse than complete-family-EM, PHASE and our approach in the same case, respectively. Even when the number of offspring in each family is increased, the performance of GENEHUNTER is still poor. On the other hand, GENEHUNTER is easily affected by the number of SNP, with the number of SNP being increased, the performance of GENEHUNTER is dramatically decreased. Because GENEHUNTER assumes that linked markers are in linkage equilibrium, it means that too much information will be lost for the high linkage disequilibrium SNPs based on coalescent model, and with only the pedigree information being used, more and more haplotypes will be wrongly inferred with the number of markers being increased.

There are three sampling designs for haplotype inference: complete nuclear families, parent-child pairs and unrelated individuals. If the primary objective of a study is to estimate population haplotype frequencies, it can be concluded from our results that sampling parent-child pairs clearly is more efficient than sampling unrelated individuals, despite the fact that two unrelated individuals contribute four independent haplotypes (i.e. 2 independent haplotypes per genotyped individual) while with a parent-child pair only three independent haplotypes are sampled (1.5 independent haplotypes per genotyped individual). This reduction of effective information by 25% is more than balanced by the possibility to identify the common haplotype and to infer the complementary unique parental haplotypes in the parent-child case. Adding the second parent (complete-family-EM) leads to a further improvement, where four independent haplotypes can be observed in a complete family (1.33 independent haplotypes per genotyped individual), more common haplotyes are used.

On the other hand, haplotype analysis is more used to assess the association between multiple markers and traits of interest. In this case, the correct haplotyping is important because we count the occurrences of haplotypes in cases and controls. Although sampling trios or incomplete trios is still more accurate than sampling unrelated individuals, the number of informative cases and controls is decreased. Here three scenarios are considered for the classical unrelated case-control study:

(1) N/6 trios where either the parent or child is a case and N/6 trios where either the parent or child is a control

(2) N/4 one parent-child pairs where either the parent or child is a case and N/4 one parent-child pairs where either the parent or child is a control

(3) N/2 unrelated cases and N/2 unrelated controls

As shown in table 4, although the haplotype reconstruction under scenario 1 and 2 are more accurate and powerful than under scenario 3, the informative sample size from the first two scenarios is decreased by 1/3 and 1/2 to scenario 3 reducing the power of case-control test. Therefore, the efficiency of these sampling designs should balance these two situations. Further studies on the efficiency of these three sampling designs for case-control design will be carried out in the future.

As in other family-based haplotype reconstruction methods it also is assumed here that within a nuclear family recombination between any two loci does not occur in the considered chromosome segments [19]. Therefore, we make an error if a recombination does take place. The magnitude of this error can be evaluated based on the following example:

Consider a nuclear family with the genotype pair

$$YP_f = \begin{pmatrix} 12;34;56 \\ 11;33;55 \end{pmatrix}$$

Our algorithm would clearly reconstruct the diplotypes for this parent-child pair to be G_i = (1 3 5, 2 4 6) and G_j = (1 3 5, 1 3 5) under the assumption that no recombination has occurred. If, however, a recombination had occurred in the second marker interval, the correct diplotype is G_i = (1 3 6, 2 4 5) and G_j = (1 3 5, 1 3 5).

In this case, we see that only two out of the four possible haplotypes are correctly reconstructed, which affects both the haplotype frequency estimation in the population and the individual haplotype reconstruction. It should be noted that in this sampling structure recombinations are not detectable, but only can be taken into account based on their respective probabilities. In the example above, the probability of the first or second diplotype conditional on the given genotype pair would be $(1 - r)$ and r, respectively, where r is the recombination rate between the second and third locus. With many loci, this leads to a 'combinatorial explosion' of possible recombined haplotypes each having a minimal conditional probability, and thus makes the estimation problem numerically intractable.

Consider a haplotyping study for a set of loci spanning a chromosome segment of length x Morgan. If, as it is typical for such studies, x is assumed to be small, we may assume that each crossing over is equivalent to a recombination, and that it is sufficiently accurate to assume a Poisson distribution of crossing over events. Then the probability of having a recombination in such an interval is $1 - e^{-x} \approx x$ for small values of x. As shown with the example above, recombination will cause at most that half of the correct haplotypes in the affected parent-child pair are misidentified in the fully informative case. Therefore, an upper limit for the rate of wrongly reconstructed haplotypes due to recombination in a sample of parent-child-pairs approximately is $x/2$. If, as in many applications of SNP haplotyping, the chromosome segment considered is of length 1 centiMorgan or less, the error rate due to recombination is <0.005, which certainly is acceptable considering other sources of error, e.g. the technical error rate in genotyping processes [20].

In practical situations, incomplete data on some individuals due to failure of typing one (or more) of the considered loci is very common in every lab. To handle such an incompletely genotyped individual in our mixed haplotyping approach, we first list all the possible genotypes at this missing locus. If his (her) parent or child does not miss the same locus, we can determine one allele for this individual to be equal to one of the observed alleles in the parent or child. This will reduce the number of possible genotypes. When inferring this individual's diplotype, each of its possible genotypes has a corresponding most likely diplotype with a conditional probability, so the one with the maximum conditional probability among these most likely diplotypes is considered as the final diplotype, and its corresponding genotype is the final genotype.

The suggested approach can also be used in mixed data structures, consisting e.g. of complete nuclear families (2 parents, one child) [6], incomplete nuclear families (1 parent, one child), and single individuals [3]. Applications to other family structures (e.g. fullsibs without parents) are in preparation.

Acknowledgements

We are grateful for the constructive criticism of anonymous reviewers on an earlier version of the manuscript. One of the authors (X.D.D.) is funded by the FUGATO program of the German Federal Ministry of Education and Research, one of the authors (C.F.) was funded by the German Research Foundation (DFG).

References

1 Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, Dibling T, Tinsley E, Kirby S, Carter D, Papaspyridonos M, Livingstone S, Ganske R, Lõhmussaar E, Zernant J, Tõnisson N, Remm M, Mägi R, Puurand T, Vilo J, Kurg A, Rice K, Deloukas P, Mott R, Metspalu A, Bentley DR, Cardon LR, Dunham I: A first generation linkage disequilibrium map of human chromosome 22. Nature 2002;418:544–548.

2 Clark AG: Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 1990;7:111–112.

3 Excoffier L, Slatkin M: Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Boil Evol 1995;12:921–927.

4 Stephens M, Smith NJ, Donnelly P: A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 2001;68:978–989.

5 Becker T, Knapp M: Efficiency of haplotype frequency estimation when nuclear family information is included. Hum Hered 2002; 54:45–53.

6 Rohde K, Fuerst R: Haplotyping and estimation of haplotype frequencies for closely linked biallelic multilocus genetic phenotypes including nuclear family information. Human mutation 2001;17:289–295.

7 Schaid DJ: Relative efficiency of ambiguous vs. directly measured haplotype frequencies. Genet Epidemiol 2002;23:426–443.

8 Dempster AP, Laird NM, Rubin DB: Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc (Ser B) 1977; 391:1–38.

9 Ceppellini R, Siniscalco M, Smith CAB: The estimation of gene frequencies in a random mating population. Ann Hum Genet 1955; 20:97–115.

10 Smith CAB: Counting methods in genetical statistics. Ann Hum Genet 1957;21:254–276.

11 Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D: Calibrating a coalescent simulation of human genome sequence variation. Genome Res 2005;15:1576–1583.

12 Stephens M, Donnelly P: A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 2003;73:1162–1169.

13 Marchini J, Cutler D, Patterson N, Stephens M, Eskin E, Halperin E, Lin S, Qin ZS, Munro MH, Abecasis G, Donnelly P, for the International HapMap Consortium: A comparison of phasing algorithms for trios and unrelated individuals. Am J Hum Genet 2006;78:437–450.

14 Kruglyak L, Daly MJ, Reeve-Daly MP, Lander EL: Parametric and nonparametric linkage analysis: A unified multipoint approach. Am J Hum Genet 1996;58:1347–1363.

15 Schaid DJ, McDonnell SK, Wang L, Cunningham JM, Thibodeau SN: Caution on pedigree haplotype inference with software that assumes linkage equilibrium. Am J Hum Genet 2002;71:992– 995.

16 Zhang S, Pakstis AJ, Kidd KK, Zhao H: Comparisons of two methods for haplotype reconstruction and haplotype frequency estimation from population data. Am J Hum Genet 2001;69:906–914.

17 Xu CF, Lewis K, Cantone KL, Khan P, Donnelly C, White N, Crocker N, Boyd PR, Zaykin DV, Purvis IJ: Effectiveness of computational methods in haplotype prediction. Hum Genet 2002;110:148–156.

18 Niu T, Qin ZS, Xu X, Liu JS: Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 2002;70:157–169.

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

20 Ewen KR, Bahlo M, Treloar SA, Levinson DF, Mowry B, Barlow JW, Foote SJ: Identification and analysis of error types in high-throughput genotyping. Am J Hum Genet 2000;67:727–736.