DOI: 10.1534/genetics.107.079525

A New Method for Haplotype Inference Including Full-Sib Information

Xiang Dong Ding,*,† Henner Simianer* and Qin Zhang†,[1]

**University of Goettingen, Institute of Animal Breeding and Genetics, 37075 Goettingen, Germany and* †*State Key Laboratories of Agrobiotechnology, Key Laboratory for Animal Breeding and Genetics of Ministry of Agriculture of China, College of Animal Science and Technology, China Agricultural University, Beijing, 100094, China*

Manuscript received July 26, 2007
Accepted for publication September 18, 2007

ABSTRACT

Recent literature has suggested that haplotype inference through close relatives, especially from nuclear families, can be an alternative strategy in determining linkage phase and estimating haplotype frequencies. In the case of no possibility to obtain genotypes for parents, and only full-sib information being used, a new approach is suggested to infer phase and to reconstruct haplotypes. We present a maximum-likelihood method via an expectation-maximization algorithm, called FSHAP, using only full-sib information when parent information is not available. FSHAP can deal with families with an arbitrary number of children, and missing parents or missing genotypes can be handled as well. In a simulation study we compare FSHAP with another existing expectation-maximization (EM)-based approach (FAMHAP), the conditioning approach implemented in FBAT and GENEHUNTER, which is only pedigree based and assumes linkage equilibrium. In most situations, FSHAP has the smallest discrepancy of haplotype frequency estimation and the lowest error rate in haplotype reconstruction, only in some cases FAMHAP yields comparable results. GENEHUNTER produces the largest discrepancy, and FBAT produces the highest error rate in offspring in most situations. Among the methods compared, FSHAP has the highest accuracy in reconstructing the diplotypes of the unavailable parents. Potential limitations of the method, *e.g.*, in analyzing very large haplotypes, are indicated and possible solutions are discussed.

WITH the discovery of single-nucleotide polymorphisms (SNPs) 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 often is more efficient than separate analyses of individual markers (Dawson *et al.* 2002). Considerable research effort has been devoted to algorithms that infer haplotype phase from genotype data.

There are a growing number of articles on haplotype inference for unrelated individuals (Clark 1990; Excoffier and Slatkin 1995; Stephens *et al.* 2001), but more and more studies show that haplotype inference through close relatives, especially from nuclear families, can be an alternative strategy, as family information can reduce phase ambiguity and improve the efficiency of haplotype frequency estimates (Hodge *et al.* 1999; Rohde and Fuerst 2001; Becker and Knapp 2002; Schaid 2002). However, these methods consider mainly those nuclear families with both parents and one child (trios). When diseases with onset in adulthood or in old age are studied, it may be impossible to obtain genotypes for markers in the parents of the affected offspring, so that only full-sib information is available, which also may be true for other reasons. Obviously, it is essential to develop efficient approaches to handle such families.

The existing computational methods for haplotyping fit into two categories: statistical methods and rule-based methods. The rule-based approaches (Qian and Beckman 2002; Li and Jiang 2003; Gao *et al.* 2004; Baruch *et al.* 2006) are deterministic and fast and thus can handle large pedigrees with dense markers. However, they normally do not provide numerical assessments of the reliability of their results, and the utility of rule-based approaches for nuclear families remains unknown (Niu 2004). On the other hand, statistical approaches are flexible in tackling nuclear families (Rohde and Fuerst 2001; Becker and Knapp 2002; Ding *et al.* 2006), although they are time-consuming and thus may not be suitable for large pedigrees.

Maximum likelihood via the expectation-maximization (EM) algorithm (Dempster *et al.* 1977) is a widely used statistical approach for haplotype inference. Excoffier and Slatkin (1995) were the first to propose a maximum-likelihood-based approach for haplotype frequency estimation for unrelated individuals. EM-based approaches without assuming linkage equilibrium among the loci were suggested for various types of complete (Rohde and Fuerst 2001; Becker and Knapp 2002) or incomplete (Ding *et al.* 2006) nuclear family

[1]*Corresponding author:* College of Animal Science and Technology, China Agricultural University, Beijing, 100094, China.
E-mail: qzhang@cau.edu.cn

data. Their performance was shown to be superior to that of the Lander–Green algorithm (Lander and Green 1987) implemented in GENEHUNTER (Kruglyak *et al.* 1996), which as well as other linkage analysis programs assumes complete linkage equilibrium between the loci (Becker and Knapp 2002; Ding *et al.* 2006).

Several methods have been suggested for haplotype inference using sibship data (Becker and Knapp 2004; Horvath *et al.* 2004; Liu *et al.* 2006), which have their own strengths and weaknesses. In this article, we propose a new maximum-likelihood-based method for haplotype reconstruction and estimation of haplotype frequencies using full-sib families, which allows genetic markers to be in linkage disequilibrium and assumes that no recombination occurs between the markers.

In our study, we first introduce the general idea of the new algorithm. Most of the technical details are presented in the appendix. We then report the outcome 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. In the discussion we provide arguments to explain the better statistical properties of our procedure compared with the established methods, and we discuss options to overcome practical problems and limitations, *e.g.*, missing genotypes and the restriction in the number of loci processed simultaneously.

METHODS

Definitions: We consider a series of N closely linked polymorphic loci. For $N = 3$, a possible phase-unknown 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, denoted as G_i, is then 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 phase-unknown genotype, several diplotypes are possible. So for one family f with n_f full sibs, and their phase-unknown genotype combination defined as YP_f, there are several possible diplotype combinations, one of which can be represented as $(G_1, G_2, \ldots, G_{n_f})_i$, termed full-sib haplotype set (FSHS), where G_1 denotes the diplotype of sib 1 in the ith FSHS of family f.

The likelihood function: Following similar arguments presented by Excoffier and Slatkin (1995), for a sample of m families with only full sibs, the likelihood function of the population haplotype frequencies is defined as

$$L(p_1, p_2, \ldots, p_n) = \prod_{f=1}^{m} \sum_{i=1}^{S_f} P_f(G_1, G_2, \ldots, G_{n_f})_i, \quad (1)$$

where $p_1, p_2, \ldots, p_n$ are the population frequencies of all haplotypes, and $\sum_{i=1}^{n} p_i = 1$. $(G_1, G_2, \ldots, G_{n_f})_i$ is the ith FSHS for family f with n_f full sibs, and S_f is the number of possible FSHS in family f.

The EM algorithm: The EM algorithm iterates between the expectation step and the maximization step until the haplotype frequency estimations converge (*i.e.*, when the changes in haplotype frequency in consecutive iterations are less than some small value).

To implement the EM algorithm, a set of initial values is required. It is assumed that given the phase-unknown genotypes of family f, all the possible FSHSs for family f have the same probability; *i.e.*,

$$P_f^{(0)}(G_1, G_2, \ldots, G_{n_f})_i = 1/S_f. \quad (2)$$

These initial probabilities are used in the likelihood function to calculate the initial likelihood value. According to Ceppellini *et al.*(1955) and Smith (1957), the population haplotype frequencies can be calculated in the first and in all subsequent iterations as

$$p^{(g+1)}(h_t) = \left(\sum_{f=1}^{m} 2n_f\right)^{-1} \sum_{f=1}^{m} \sum_{i=1}^{S_f} \delta_{it} P_f^{(g)}(G_1, G_2, \ldots, G_{n_f})_i, \quad (3)$$

where $2n_f$ is the total number of haplotypes being considered in the ith FSHS $(G_1, G_2, \ldots, G_{n_f})_i$ in family f and δ_{it} is an indicator variable equal to the number of times that haplotype t is present in the ith FSHS; its possible values are $0, 1, \ldots, 2n_f$.

In the expectation step in the gth iteration, the haplotype frequencies obtained in the previous iteration are used to calculate the probability of each possible FSHS for family f as

$$P_f^{(g+1)}(G_1, G_2, \ldots, G_{n_f})_i = \frac{P_f^{(g)}(G_1, G_2, \ldots, G_{n_f})_i}{\sum_{k=1}^{S_f} P_f^{(g)}(G_1, G_2, \ldots, G_{n_f})_k}, \quad (4)$$

where

$$P_f^{(g)}(G_1, G_2, \ldots, G_{n_f})_i = \sum_{j=1}^{7} \left\{ P_f^{(g)}((G_1, G_2, \ldots, G_{n_f})_i \mid f_j, m_j) P^{(g)}(f_j, m_j) \right\}. \quad (5)$$

Here, we give only a brief explanation of Equation 5; for details see the appendix. We first inferred the possible parental combinations based on FSHS, and according to the posterior parental information the probability of FSHS is calculated. For one FSHS, there often will be several possible parental combinations, so $P^{(g)}(f_j, m_j)$ is the probability of the jth parental combination given the estimates of population haplotype frequencies in the gth iteration, and $P_f^{(g)}((G_1, G_2, \ldots, G_{n_f})_i | f_j, m_j)$ is the probability of FSHS conditional on the jth possible parental combination.

Iterating between the E-step, using Equation 4 to update probabilities of all FSHSs, and the M-step, using Equation 3 to calculate all haplotype frequencies, the

EM algorithm yields the maximum-likelihood estimates of the population haplotype frequencies when an adequate convergence criterion is reached.

In addition to the estimation of haplotype frequencies, haplotype reconstruction is another objective of haplotype inference. Using the probability of each possible FSHS obtained in the expectation step Equation 4 after convergence, the conditional probabilities of these FSHSs for a full-sib family with phase-unknown genotype combination YP_f can be calculated after the conversion of all probabilities as

$$P\{(G_1, G_2, \ldots, G_{n_f})_i \mid (p_1, p_2, \ldots, p_n), YP_f\} = \frac{P(G_1, G_2, \ldots, G_{n_f})_i}{\sum_{k=1}^{S_f} P(G_1, G_2, \ldots, G_{n_f})_k}. \quad (6)$$

The one with the highest probability is the most likely FSHS in full-sib family *f*, and subsequently the most likely diplotype for each member in this family can be obtained. On the other hand, there should be several possible parental combinations for this most likely FSHS, where a probability can be assigned to each possible parental combination (see the APPENDIX). The one with the highest probability will be regarded as the most likely parental combination, and diplotypes of parents will be easily obtained.

SIMULATION STUDY

Simulated data: To evaluate our approach, we carried out a series of simulation studies. We simulated haplotypes using Schaffner's simulation program (SCHAFFNER *et al.* 2005) based on a coalescent model. The parameters used for the simulation were: chromosome segment length, 1 Mb; mutation rate, 1.5×10^{-8}; recombination rate, 1×10^{-8}; effective population size, 10,000; and number of sampled chromosomes, 1000. 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. With the assumption of no recombination, the diplotype of their offspring was generated by randomly picking one of the two haplotypes of the father and the mother, respectively. For full-sib families, the information on both parents was omitted after generating the children. Markers are thinned to obtain the required 1 SNP per 8-kb density that was used throughout this study. In the different scenarios, haplotypes of 5, 10, or 20 SNPs were considered, the number of full-sib families was varied between 15 and 60, and the number of offspring in each family was varied between 2 and 20, respectively. For each scenario, 100 replicates were generated and analyzed, and every data set was expected to be in Hardy–Weinberg equilibrium.

Approaches to be compared: In our study, we compared our approach FSHAP with the following three approaches:

a. FAMHAP estimates haplotype frequencies from unrelated individuals or simple nuclear families with an arbitrary number of children with the EM algorithm (BECKER and KNAPP 2004). The frequencies are the frequencies in the founders, *i.e.*, those of the parents of the nuclear families and/or the individuals (single-person families). FAMHAP provides only the most likely diplotypes of both parents. The diplotypes of offspring must be inferred again on the basis of their own genotype and parental diplotypes.

Sibships with two missing parents can be treated as well, and these are regarded as nuclear families in which parental genotype information is missing at all loci, but frequencies are still estimated with respect to the parental generation (BECKER and KNAPP 2004). However, the frequencies in the parental generation are identical to those in the offspring generation due to Hardy–Weinberg equilibrium.

b. FBAT was initially designed by HORVATH *et al.* (2004) for implementing a broad class of family-based association tests. For multiple tightly linked markers, the haplotypes are first reconstructed via a conditioning approach and association testing is then applied (HORVATH *et al.* 2004). Haplotype FBAT can deal with nuclear families, sibships, etc., when using the additional program HaploInfo (http://www.biostat.harvard.edu/~fbat/haploinfo.htm). This analysis provides haplotype population frequencies and diplotypes of both parents and offspring.
c. GENEHUNTER is a widely used software for linkage analysis (KRUGLYAK *et al.* 1996), which makes full use of the pedigree information. After convergence the program provides information only on the most likely diplotype and does not give its posterior probability. Although very popular and technically suited to handle sibship data, GENEHUNTER uses pedigree information only assuming genotypes to be in full linkage equilibrium.

These three approaches were compared with FSHAP, which is specially designed for haplotype inference using families with only full sibs and can handle arbitrary numbers of full sibs. The parameters were estimated with the approaches described in METHODS and thus account both for linkage disequilibrium (LD) and for pedigree information.

FAMHAP, FBAT, and FSHAP allow genetic markers to be in linkage disequilibrium and assume that no recombination occurs between the markers in the generation leading to the full-sib groups. Although the inappropriateness of using GENEHUNTER to reconstruct haplotypes from markers in LD has been identified (SCHAID

et al. 2002), it was used here as a lower-bound reference for the performance of FAMHAP, FBAT, and FSHAP.

Criteria: The efficiencies of the different approaches were evaluated with two sets of performance indexes. The first set, including indexes I_F and I_H, is related to the evaluation of the population haplotype frequency estimation. I_F measures the discrepancy between the estimated and true simulated sample haplotype frequencies and was defined by Stephens *et al.* (2001) as

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

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

Identification rate I_H examines whether all haplotypes present in the sample are identified in the estimated haplotypes. In a sample with N individuals, the minimum frequency for every true haplotype must be $\geq(2N)^{-1}$, which can be used as a lower threshold value for determining the existence of a haplotype; *i.e.*, a haplotype is accepted to be detected only if its estimated frequency is $>(2N)^{-1}$. On the basis of this, Excoffier and Slatkin (1995) suggested the statistic

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

where k_{true} is the number of true haplotypes in the sample, k_{found} is the number of identified haplotypes with frequency above the threshold value in the sample, and k_{missed} is the number of true haplotypes not identified in the sample. 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.

There are two options for the definition of true haplotype frequency. The first one is the relative frequency of haplotype i in the entire ("true") population, and the second one is the relative frequency of haplotype i in the sample (*i.e.*, in the sibships). The methods compared in our study all make use of the same data. Accuracy of parameter estimation is a combination of (i) sampling and (ii) estimation conditional on the sample. Since we are interested only in the differences between methods, only step ii is relevant; therefore a comparison conditional on the drawn samples seems appropriate.

The second set of indexes, including error rate and I_R, is related to the evaluation of the haplotype reconstruction.

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

Although the phase-unknown genotypes of parents are not available, they can be inferred according to the information of offspring. However, the father and the mother cannot be definitely assigned due to their unknown genotypes; only the reconstructed parental diplotypes are taken into account to be compared with true parental diplotypes in the calculation of error rate in our approach. For FAMHAP, FBAT, and GENEHUNTER, the reconstructed diplotypes for father and mother were assigned to the most similar true genotypes of the parents, respectively. The following combinations were compared: (i) reconstructed father–true father and reconstructed mother–true mother and (ii) reconstructed father–true mother and reconstructed mother–true father. The more similar combination was accepted and used as basis for calculation of error rate in parents.

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 diplotype, and FAMHAP only provides that for parents, the statistic I_R can be given only by FSHAP and FBAT.

Where appropriate, contrasts of the means of simulation results between different estimation methods were tested with a conventional t-test using SAS 9.1 (SAS Institute 2004).

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

RESULTS

We simulated four scenarios with identical genotyping costs: 60 families with two sibs, 30 families with four sibs, 20 families with six sibs, and 15 families with eight sibs. All the approaches deal with the same data sets with haplotypes of 10 SNPs. The results of our comparisons with respect to the performance of FSHAP, FAMHAP, FBAT, and GENEHUNTER from these scenarios are shown in Figure 1. In most cases, our new method for haplotype inference using sibship data (FSHAP) has the smallest discrepancy and lowest error rate and the highest identification rate. Only in some situations, the performance of FSHAP is close to FAMHAP, *e.g.*, the discrepancy in the first scenario of 60 families with two sibs and the identification rate in the third scenario of 20 families with six sibs. In the estimation of haplotype frequencies, GENEHUNTER produces the largest discrepancy and the lowest identification rate, and in the haplotype reconstruction, FBAT produces the highest error rate in offspring in most situations.

In the parental haplotype reconstruction using offspring's information, as shown in Figure 1, FSHAP,

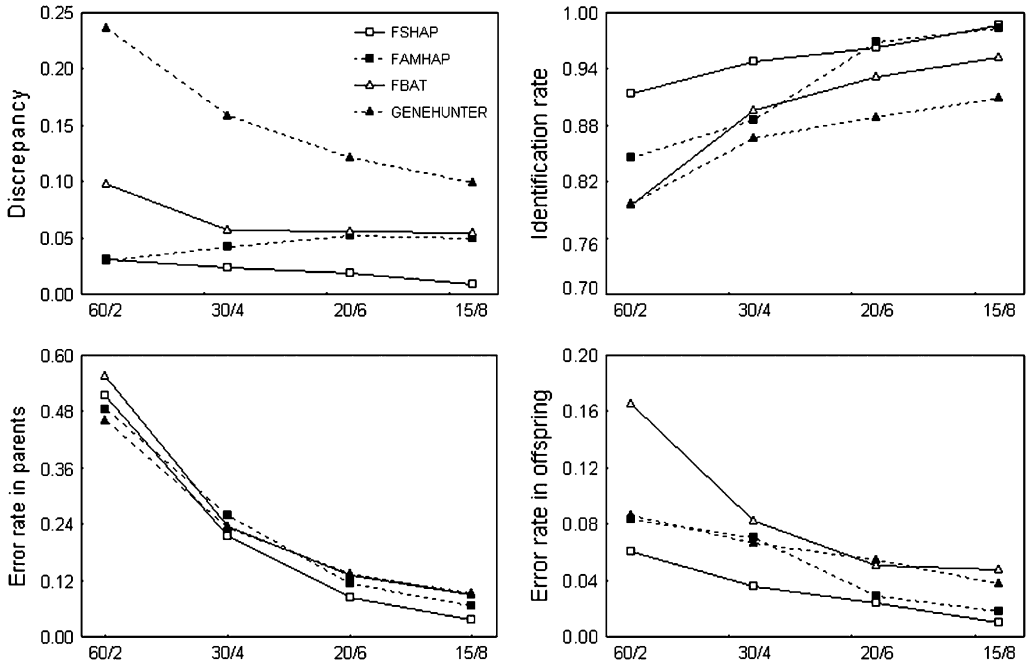

FIGURE 1.—Comparison of haplotype frequency estimation and haplotype reconstruction of FSHAP, FAMHAP, FBAT, and GENEHUNTER from four scenarios with identical genotyping cost: 60 families with two sibs (denoted as 60/2), 30 families with four sibs, 20 families with six sibs, and 15 families with eight sibs in 100 data sets; haplotypes of 10 SNPs are simulated.

FAMHAP, FBAT, and GENEHUNTER perform poorly in the case of families with only two full sibs, where the error rate in parents remains above 0.45. This performance is not so helpful to further analysis. However, their performance improves rapidly with the number of offspring being increased; especially, the error rate in parents from our approach (FSHAP) and FAMHAP decreases faster than that from FBAT and GENEHUNTER. For the calculation of error rate in parents generally our approach performs better than the other three methods, whereas the performance of FBAT is very close to that of GENEHUNTER.

As expected, the efficiency of all the approaches can be improved by increasing the number of offspring in each family (Figure 1), which provides more family information to exclude more redundant FSHSs and parental combinations. The only exception is that the discrepancy of haplotype frequencies from FAMHAP does not decrease as in other approaches but increases slightly.

This point is further illustrated by Table 1. For the second scenario of 30 families with only four sibs each, even when the genotyping cost is double after the number of families is increased to 60, the performance of FSHAP and FAMHAP is still lower than that in the fourth scenario of 15 families with only eight sibs each. On the other hand, it also can be seen from Table 1 that the improvement of efficiency of FSHAP and FAMHAP is very small by increasing only the number of families, and the identification rate is not increased but decreased a little bit.

Our approach and FBAT can provide a posterior probability for the most likely diplotype. As shown in Table 2, observing more offspring in families is also helpful to improve the reliability of inference for parents and offspring. For only two full sibs, there are a lot of possible parental combinations, which make the reliability of inference for parents very low, but for multiple sibs, more redundant parental combinations are excluded and the posterior probability of the most likely diplotype will be increased. Table 2 shows that the reliability of FBAT is apparently higher than that of FSHAP in most situations. This mainly is a con-

TABLE 1

Efficiency of an increasing number of families *vs.* efficiency of an increasing number of offspring from FSHAP and FAMHAP (10 SNPs)

	FSHAP			FAMHAP		
	30 families with 4 sibs	60 families with 4 sibs	15 families with 8 sibs	30 families with 4 sibs	60 families with 4 sibs	15 families with 8 sibs
Discrepancy	0.0234	0.0181	0.0092	0.0420	0.0305	0.0495
Identification rate	0.9475	0.9462	0.9874	0.8864	0.8454	0.9832
Error rate in parents	0.2128	0.2111	0.0367	0.2575	0.2614	0.0667
Error rate in offspring	0.0352	0.0341	0.0100	0.0706	0.0688	0.0177

TABLE 2

The reliability of haplotype reconstruction for parents and offspring from FSHAP and FBAT (10 SNPs)

No. of families	No. of offspring	FSHAP		FBAT	
		Parent	Offspring	Parent	Offspring
60	2	0.6371	0.9385	0.7563	0.9862
30	4	0.8965	0.9632	0.9129	0.9924
20	6	0.9595	0.9726	0.9508	0.9926
15	8	0.9793	0.9818	0.9688	0.9956

sequence of the higher error rate of FBAT compared with FSHAP. FBAT identifies less haplotypes correctly, but those correctly identified on average have a higher posterior probability than the ones identified with FSHAP.

Generally, the discrepancy and error rate for the EM-based methods increase when more SNPs are included, since in this case the number of possible haplotypes increases exponentially while the average amount of information for estimation from the same data set decreases. As shown in Table 3, the efficiencies of FSHAP, FAMHAP, and FBAT all decrease when the number of SNPs is increased from 10 to 20. However, the decrease with FSHAP is very moderate compared to that with FAMHAP and FBAT.

It is also indicated from Table 3 that the impact of the number of SNPs on the running time for the three EM-based approaches of FSHAP, FAMHAP, and FBAT is very large compared to that for GENEHUNTER. In the case of 10 SNP loci, FSHAP is the fastest one among these four approaches, but the average running time increases dramatically from 0.17 sec to 9.66 sec when the number of SNPs is increased from 10 to 20. Compared to FSHAP, FBAT becomes much slower, where the average running time exponentially increases to 161 sec from 3.45 sec. However, a progressive-extension technique was implemented in FAMHAP (Becker and Knapp 2004), which makes FAMHAP very fast for the large number of SNP loci. Similarly, the Lander–Green algorithm (Lander and Green 1987) keeps the speed of GENEHUNTER almost stable with the doubling of the number of loci.

TABLE 4

The impact of the number of offspring on the running time (*s*) of FSHAP, FAMHAP, FBAT, and GENEHUNTER (10 SNPs)

No. of families	No. of offspring	FSHAP	FAMHAP	FBAT	GENEHUNTER
60	2	0.6176	0.6901	14.7200	2.5799
30	4	0.1749	0.4701	3.4500	1.3301
20	6	0.1199	0.2300	2.2101	3.1700
15	8	0.0963	0.8001	1.8900	54.6101

The running time is also affected by the number of children in families since more redundant parental combinations can be excluded to improve speed by using multiple sibs for FSHAP, FAMHAP, and FBAT. Therefore, the running time of these three approaches is decreased when the number of children is increased from two to six (Table 4). However, this advantage will be counteracted by the enumeration of all haplotype configurations of more children; *e.g.*, the running time of FAMHAP is suddenly increased as sib size is increased to eight. It is also indicated from Table 4 that FSHAP performs faster than FAMHAP, FBAT, and GENEHUNTER. FAMHAP is the second fastest approach.

In the case of 10 and 20 SNPs, some families cannot be handled by FAMHAP due to too many possible haplotypes. Therefore FAMHAP with a progressive-extension technique being implemented is used throughout our study, and FAMHAP without a progressive-extension technique is denoted as FAMHAP_nit. A small number of SNPs (5) and varying numbers of children in families (4–8 or 2–10, respectively) are assumed to compare the performance of FSHAP, FAMHAP, FAMHAP_nit, and FBAT. We simulated two scenarios: 60 families with 4–8 sibs and 60 families with 2–10 sibs, all data sets with haplotypes of 5 SNPs.

As shown in Table 5, FSHAP performs significantly better compared to the other approaches in most situations. The values of discrepancy from FAMHAP and FAMHAP_nit are not different, whereas the performance of FAMHAP is significantly better than that of FAMHAP_nit with respect to identification rate and haplotype reconstruction.

TABLE 3

The impact of the number of SNPs on the efficiency of FSHAP, FAMHAP, FBAT, and GENEHUNTER (30 families with four sibs each)

	FSHAP		FAMHAP		FBAT		GENEHUNTER	
No. of SNPs	10	20	10	20	10	20	10	20
Discrepancy	0.0234	0.0552	0.0420	0.0649	0.0568	0.1207	0.1586	0.1598
Identification rate	0.9475	0.9133	0.8864	0.7832	0.8964	0.8126	0.8659	0.8234
Error rate in parents	0.2128	0.2403	0.2575	0.3346	0.2348	0.3112	0.2312	0.2447
Error rate in offspring	0.0352	0.0828	0.0706	0.2073	0.0818	0.2126	0.0660	0.1512
Running time (*s*)	0.1749	9.6677	0.4701	1.3001	3.4500	160.9800	1.3301	2.1198

TABLE 5

Comparison of efficiency of FSHAP, FAMHAP, FAMHAP_nit, and FBAT in the case of families with different numbers of children (five SNPs) from 100 data sets

	FSHAP	FAMHAP	FAMHAP_nit	FBAT
	60 families with 4–8 children each (on average 5.90 children/family)			
Discrepancy	0.0043 (0.0006)[a]	0.0209 (0.0009)[b]	0.0190 (0.0009)[b]	0.0223 (0.0011)[b]
Identification rate	0.9794 (0.0042)[a]	0.9136 (0.0084)[b]	0.8401 (0.0132)[c]	0.9707 (0.0046)[a]
Error rate in parents	0.1125 (0.0030)[a]	0.1311 (0.0094)[b]	0.2470 (0.0250)[b]	0.1326 (0.0043)[c]
Error rate in offspring	0.0080 (0.0011)[a]	0.0192 (0.0016)[b]	0.1355 (0.0190)[c]	0.0204 (0.0015)[b]
Running time (*s*)	0.4209	0.8001	0.9301	9.9700
	60 families with 2–10 children each (on average 5.93 children/family)			
Discrepancy	0.0035 (0.0005)[a]	0.0250 (0.0012)[b,c]	0.0233 (0.0012)[b]	0.0282 (0.0012)[c]
Identification rate	0.9829 (0.0039)[a]	0.9271 (0.0080)[b]	0.8539 (0.0138)[c]	0.9658 (0.0051)[d]
Error rate in parents	0.1528 (0.0030)[a]	0.1716 (0.0043)[b]	0.2608 (0.0245)[c]	0.2084 (0.0070)[d]
Error rate in offspring	0.0054 (0.0008)[a]	0.0186 (0.0020)[b]	0.1333 (0.0210)[c]	0.0354 (0.0016)[d]
Running time (*s*)	0.2142	6.4800	6.4900	2.9600

FAMHAP_nit: FAMHAP without the progressive-extension technique being implemented. Standard error is in parentheses.
[a,b,c,d] Means with different superscripts within one row differ significantly at the $\alpha < 0.01$ significance level.

DISCUSSION

One limitation of the EM algorithm is that it cannot handle a large number of loci due to the memory constraint. So far, FSHAP can handle up to 30 loci. Recently, some strategies were proposed to overcome this problem. Clayton (1999) first implemented progressive-extension (PE) techniques in his program SNPHAP. Becker and Knapp (2004) further used PE in FAMHAP, which gives reliable approximations of the maximum-likelihood estimates for up to 63 SNP loci. However, the number of bits required for the type of data prevents this technique from handling more loci (Becker and Knapp 2004). The PE technique does not guarantee that the true EM estimates for individual haplotypes are obtained (Qin *et al.* 2002), which in the case of FAMHAP is not critical since the main objective of the program is to estimate haplotype frequencies rather than to reconstruct individual haplotypes. However, the results of our study (Table 5) indicate that the progressive-extension technique will not impair the performance of FAMHAP in haplotype frequency estimation, while it can make FAMHAP perform better in haplotype reconstruction. A possible explanation is that despite the good haplotype frequency estimates FAMHAP without progressive-extension techniques might be unable to pinpoint which family has which haplotype (T. Becker, personal communication).

Another widely used strategy for a large number of loci is the partition-ligation (PL) algorithm proposed by Niu *et al.* (2002). PL was first implemented together with Gibbs sampling to estimate haplotype phases for a large number of SNPs, and Qin *et al.* (2002) further combined it with the EM algorithm to handle large sets of loci. The PL–EM of Qin *et al.* (2002) is currently implemented for unrelated individuals only, but can also be integrated in our approach.

Although both FAMHAP and FSHAP are EM-based approaches, there are two crucial steps in FSHAP that make it perform better than FAMHAP, both with respect to computing speed and accuracy of haplotype inference:

i. In our approach, a collapse technique is used to infer possible parental haplotype combinations. It starts from the possible parental haplotype combinations based on a single pair of full sibs and then goes through all additional full sibs, excluding those haplotype combinations not being compatible with the extra children (see the appendix). In FAMHAP a complete list of possible parental haplotype combinations is set up first, which will be very large when parental multilocus genotypes are missing. Afterward, those diplotypes not compatible with the children's genotypes are excluded. This strategy is much slower and more memory demanding than the one implemented in FSHAP, especially when the number of loci is large.
ii. In our approach, the probability of each parental configuration is calculated according to different mating designs, which will give different weight to each parental haplotype combination. Further, a multinominal distribution is used to calculate the joint probability of a sibship's diplotype given each posterior parental combination, which makes effective use of family information (see the appendix). By this the information on parents and sibships is updated in each iteration simultaneously, which makes the estimation of haplotype frequencies more accurate and the inference of haplotype reconstruction in parents and offspring more reliable compared to FAMHAP, which originally was developed primarily for haplotype frequency estimation rather than for individual haplotype reconstruction.

Liu *et al.*(2006) proposed another EM-based approach for haplotype inference from sibship data, which was not included in the comparison in our study. Liu *et al.*(2006) report that their approach performs slightly better than FAMHAP and that the variability of discrepancy of their performance is small with the sample size. However, only sibships with two children were taken into account in their study. The approach proposed by Liu *et al.*(2006) is similar to our approach by considering different parental mating designs; however, the calculation of posterior parental combinations is different. On the other hand, Liu *et al.*(2006) do not make effective use of the joint information of full sibs given the parental configuration; therefore we expect our approach to be more efficient with increasing family sizes.

Our study proves that including nuclear family information will improve not only the correctness of haplotype reconstruction but also the accuracy of haplotype frequency estimates as discussed in other studies (Rohde and Fuerst 2001; Becker and Knapp 2002; Schaid 2002). Especially for our approach FSHAP the parental information can also be inferred accurately when the number of offspring is increased. It will be especially helpful for research in multiparous species like pigs, dogs, fish, and many lab animals, where it is easy to collect families with multiple siblings.

Theoretically, our approach can deal with sibships of arbitrary size. However, families with an excessively large number of children cannot be handled due to the limitation of computing memory. On the other hand, increasing the number of children is not always helpful to improve the efficiency of our approach. As shown in Table 6, the improvement is very small when the number of children is increased from 8 to 12 and 15, and the performance of our approach is decreased when the number of children is increased to 20.

As Figure 2 shows, it is difficult to have a functional relationship between the running time and sib size, because the composition of phase-unknown genotypes of sibs plays a more important role than sib size. To illustrate this we give three scenarios: (i) families with all four children having genotype (12; 12; 12; 12; 12), (ii) families with four children having different genotypes each, and (iii) families with eight children having different genotypes each. The running time of our approach in the first case is higher than that in the second and third cases because the number of possible parental combinations is much higher in the first case than in the second and third cases, even though the sib size in the first two cases is identical and doubled in the third case.

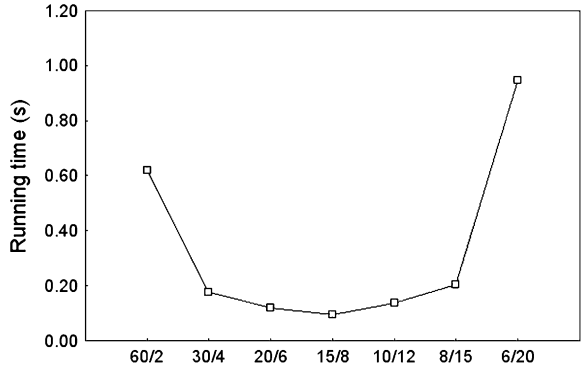

Figure 2.—The impact of the number of children in families on the running time (*s*) of FSHAP from seven scenarios with identical genotyping cost: 60 families with 2 sibs (denoted as 60/2), 30 families with 4 sibs, 20 families with 6 sibs, 15 families with 8 sibs, 10 families with 12 sibs, 8 families with 15 sibs, and 6 families with 20 sibs in 100 data sets; haplotypes of 10 SNPs are simulated.

Comparing the results obtained here to the ones reported by Ding *et al.*(2006), it can be concluded that with the same burden of genotyping complete nuclear families are more informative than sibships. In their simulation based on the same program (Schaffner *et al.* 2005), the error rate of complete-family EM proposed by Rohde and Fuerst (2001) is 0.4% in the case of 30 trios and 10 SNPs (Ding *et al.* 2006), which is less genotyping cost and a much lower error rate than the 3.5% of FSHAP in the case of 30 sibships with four sibs each and 10 SNPs.

In practical situations, incomplete data on some individuals due to failure of typing for one (or more) of the component loci is very common in every lab. Our approach can easily handle such a situation. For an individual with a missing locus, we first list all the

TABLE 6

The impact of the number of children on the efficiency of FSHAP (10 SNPs)

No. of families	No. of offspring	Discrepancy	Identification rate	Error rate in parents	Error rate in offspring	Running time (*s*)
60	2	0.0316	0.9138	0.5131	0.0603	0.6176
30	4	0.0234	0.9475	0.2128	0.0352	0.1749
20	6	0.0182	0.9628	0.0825	0.0237	0.1199
15	8	0.0092	0.9874	0.0367	0.0100	0.0963
10	12	0.0077	0.9923	0.0175	0.0100	0.1368
8	15	0.0098	0.9880	0.0188	0.0152	0.2029
6	20	0.0113	0.9882	0.0692	0.0581	0.9470

possible genotypes at this missing locus, where the information of other sibs of this individual can be used to exclude some impossible genotypes. Thus this individual will have several possible phase-unknown genotypes. When inferring this individual's diplotype, each of her (his) phase-unknown genotypes has a corresponding most likely diplotype with a conditional probability, so the one with the highest probability among these most likely diplotypes is considered as the final diplotype, and its corresponding phase-unknown genotype is the final multilocus genotype.

As in other family-based haplotype reconstruction methods, it also is assumed that within a nuclear family recombination does not occur in the considered chromosome segments (Hodge *et al.* 1999). When recombination events do occur among loci, it will make it complex to infer the parental combinations on the basis of the information of sibs. However, for tightly linked loci, recombination is an unlikely event. Moreover, recent studies (Patil *et al.* 2001; Gabriel *et al.* 2002) have shown that the human genome can be partitioned into large blocks with high LD and relatively low recombination, separated by short regions of low LD. Therefore, if the markers within the same haplotype block are analyzed together, it is reasonable to assume that there is no recombination among these markers (Wang *et al.* 2002).

Although FSHAP was initially designed for families with only full sibs, it can also deal with sibships with parents. According to the principle of FSHAP, the available parent will help to exclude redundant parental combinations and to improve the efficiency of FSHAP.

Furthermore, our approach can also be used in mixed-data structures, consisting, *e.g.*, of complete nuclear families (two parents and at least one child) (Rohde and Fuerst 2001), incomplete nuclear families (one parent and at least one child) (Ding *et al.* 2006), sibships with an arbitrary number of children (this study), and single individuals (Excoffier and Slatkin 1995). All of these four methods are implemented via an EM algorithm and are similar in the likelihood function. Hence, they can be unified in one framework for mixed-data structures, which will be done in a future study.

At the moment, FSHAP runs only under Linux, and it is available on request from the authors.

We thank two anonymous reviewers for their constructive comments, Lin Wang for help in using the program HaploInfo to read the output of FBAT, and Tim Becker for helpful advice. This research was supported by the Functional Genome Analysis in Animal Organisms program of the German Federal Ministry of Education and Research, the Förderverein Biotechnologieforschung e.V. Bonn, Lohmann TierzuchtGmbH Cuxhaven, and by The National Key Basic Research Program of China (grant no. 2006CB102104).

LITERATURE CITED

Baruch, E., J. I. Weller, M. Cohen-Zinder, M. Ron and E. Seroussi, 2006 Efficient inference of haplotypes from genotypes on a large animal pedigree. Genetics **172:** 1757–1765.

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

Becker, T., and M. Knapp, 2004 Maximum-likelihood estimation of haplotype frequencies in nuclear families. Genet. Epidemiol. **27:** 21–32.

Ceppellini, R., M. Siniscalco and C. A. B. Smith, 1955 The estimation of gene frequencies in a random mating population. Ann. Hum. Genet. **20:** 97–115.

Clark, A. G., 1990 Inference of haplotypes from PCR-amplified samples of diploid populations. Mol. Biol. Evol. **7:** 111–122.

Clayton, D., 1999 A generalization of the transmission/disequilibrium test for uncertain haplotypes. Am. J. Hum. Genet. **65:** 1170–1177.

Dawson, E., G. R. Abecasis, S. Bumpstead, Y. Chen, S. Hunt *et al.*, 2002 A first generation linkage disequilibrium map of human chromosome 22. Nature **418:** 544–548.

Dempster, A. P., N. M. Laird and D. B. Rubin, 1977 Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B **391:** 1–38.

Ding, X. D., Q. Zhang, C. Flury and H. Simianer, 2006 Haplotype reconstruction and estimation of haplotype frequencies from nuclear families with only one parent available. Hum. Hered. **62:** 12–19.

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

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy *et al.*, 2002 The structure of haplotype blocks in the human genome. Science **296:** 2225–2229.

Gao, G., I. Hoeschele, P. Sorensen and F. Du, 2004 Conditional probability methods for haplotyping in pedigrees. Genetics **167:** 2055–2065.

Hodge, S. E., M. Boehnke and M. A. Spence, 1999 Loss of information due to ambiguous haplotyping of SNPs. Nat. Genet. **21:** 360–361.

Horvath, S., X. Xu, S. Lake, E. Silverman, S. Weiss *et al.*, 2004 Tests for associating haplotypes with general phenotype data: application to asthma genetics. Genet. Epidemiol. **26:** 61–69.

Kruglyak, L., M. J. Daly, M. P. Reeve-Daly and E. L. Lander, 1996 Parametric and nonparametric linkage analysis: a unified multipoint approach. Am. J. Hum. Genet. **58:** 1347–1363.

Lander, E. S., and P. Green, 1987 Construction of multilocus genetic linkage maps in humans. Proc. Natl. Acad. Sci. USA **84:** 2363–2367.

Li, J., and T. Jiang, 2003 Efficient inference of haplotypes from genotypes on a pedigree. J. Bioinform. Comput. Biol. **1:** 41–69.

Liu, P. Y., Y. Lu and H. W. Deng, 2006 Accurate haplotype inference for multiple linked single-nucleotide polymorphisms using sibship data. Genetics **174:** 499–509.

Niu, T., 2004 Algorithms for inferring haplotypes. Genet. Epidemiol. **27:** 334–347.

Niu, T., Z. S. Qin, X. Xu and J. S. Liu, 2002 Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am. J. Hum. Genet. **70:** 157–169.

Patil, N., A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi *et al.*, 2001 Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science **294:** 1719–1723.

Qian, D., and L. Beckman, 2002 Minimum-recombinant haplotyping in pedigrees. Am. J. Hum. Genet. **70:** 1434–1445.

Qin, Z. S., T. Niu and J. S. Liu, 2002 Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. Am. J. Hum. Genet. **71:** 1242–1247.

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

SAS Institute, 2004 *SAS 9.1.3 Help and Documentation*. SAS Institute, Cary, NC.

Schaffner, S. F., C. Foo, S. Gabriel, D. Reich, M. J. Daly *et al.*, 2005 Calibrating a coalescent simulation of human genome sequence variation. Genome Res. **15:** 1576–1583.

Schaid, D. J., 2002 Relative efficiency of ambiguous vs. directly measured haplotype frequencies. Genet. Epidemiol. **23:** 426–443.

Schaid, D. J., S. K. McDonnell, L. Wang, J. M. Cunningham and S. N. Thibodeau, 2002 Caution on pedigree haplotype inference with software that assumes linkage equilibrium. Am. J. Hum. Genet. **71:** 992–995.

Smith, C. A. B., 1957 Counting methods in genetical statistics. Ann. Hum. Genet. **21:** 254–276.

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

Wang, N., J. M. Akey, K. Zhang, K. Chakraborty and L. Jin, 2002 Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. Am. J. Hum. Genet. **71:** 1227–1234.

Communicating editor: C. Haley

APPENDIX

Inferring parental combination using sibs: For one population, there are seven types of diplotype combinations as shown in Table A1 for any two individuals: (1) both are identical homozygotes; (2) one is a homozygote and the other is a heterozygote, with one common haplotype; (3) both are identical heterozygotes; (4) both are heterozygotes with one common haplotype; (5) both are different homozygotes; (6) one is a homozygote and the other is a heterozygote, without a common haplotype; and (7) both are heterozygotes without a common haplotype.

Calculation of joint probability of the children's diplotypes using parental information: For one family with known parental combination the probability of each possible diplotype of the children have can be obtained as shown in Table A3, and the joint probability of n children in this family is multinominal,

$$P(G_1, \ldots, G_n \mid f, m) = \frac{n!}{y_1! \ldots y_k!} p_1^{y_1} \ldots p_k^{y_k}, \tag{A1}$$

where $(G_1, \ldots, G_n)$ is the set of diplotypes of all children in this family, k is the number of types of diplotypes among all children in this family (its maximum value is 4 as shown in Table A3), y_i $(i = 1, \ldots, k)$ is the number of children with diplotype i, and p_i is the corresponding probabilities of diplotype i shown in Table A3.

For example, in the case of $n = 4$, $(G_1, \ldots, G_n) = (h_a h_a, h_a h_a, h_a h_b, h_b h_b)$, and the parental combination known as $(h_a h_b \times h_a h_b)$, we can obtain the values of k, y_1, y_2, and y_3 as

$$\begin{pmatrix} k = 3 \\ y_1 = 2 \\ y_2 = 1 \\ y_3 = 1 \end{pmatrix}.$$

Then according to Table A3 and Equation A1, the joint probability of these children is

$$\frac{4!}{(2!) \times (1!) \times (1!)} \left(\frac{1}{4}\right)^2 \frac{1}{2} \frac{1}{4} = \frac{3}{32}.$$

.

TABLE A1

Type and diplotype combinations for two individuals

Type of combination	1	2	3	4	5	6	7
Diplotype of individual 1	$h_a h_a$	$h_a h_a$	$h_a h_b$	$h_a h_b$	$h_a h_a$	$h_a h_a$	$h_a h_b$
Diplotype of individual 2	$h_a h_a$	$h_a h_b$	$h_a h_b$	$h_a h_c$	$h_b h_b$	$h_b h_c$	$h_c h_d$

This system of types of diplotype combinations can also be used when analyzing families with only full sibs. In our method, a collapse technique is used to calculate the probabilities of parental combinations. For two full sibs, as shown in Table A2, their possible diplotype combinations can be listed first, and for each possible combination, several possible parental combinations can be inferred, and the probability of each parental combination can be calculated given the population haplotype frequencies. For multiple sibs, those diplotype combinations that are not compatible with the extra offspring are excluded.

TABLE A2

Type and probability of parental combinations according to diplotypes of two sibs

Two sibs: Diplotype combination	Parent		
	Diplotype combinations	Type	Probability P(f, m)
(h_ah_a,h_ah_a)	$(h_ah_a \times h_ah_a)$	1	p_a^4
	$(h_ah_a \times h_ah_k)$	2	$2p_a^3 \sum_{k=1,k\neq a}^{n} p_k$
	$(h_ah_k \times h_ah_k)$	3	$4p_a^2 \sum_{k=1,k\neq a}^{n} p_k^2$
	$(h_ah_k \times h_ah_l)$	4	$4p_a^2 \sum_{\substack{k=1\\k\neq a}}^{n} \sum_{\substack{l=1\\l\neq a,k}}^{n} p_kp_l$
$(h_ah_a,\ h_ah_b)$	$(h_ah_a \times h_ah_b)$	2	$2p_a^3p_b$
	$(h_ah_b \times h_ah_b)$	3	$4(p_ap_b)^2$
	$(h_ah_b \times h_ah_k)$	4	$4p_a^2p_b \sum_{k=1,k\neq a,b}^{n} p_k$
(h_ah_b,h_ah_b)	$(h_ah_a \times h_ah_b)$, $(h_bh_b \times h_ah_b)$	2	$2p_a^3p_b + 2p_ap_b^3$
	$(h_ah_b \times h_ah_b)$	3	$4(p_ap_b)^2$
	$(h_ah_b \times h_ah_k)$ $(h_ah_b \times h_bh_k)$ $(h_ah_k \times h_bh_k)$	4	$4p_a^2p_b \sum_{k=1,k\neq a,b}^{n} p_k + 4p_ap_b^2 \sum_{k=1,k\neq a,b}^{n} p_k + 4p_ap_b \sum_{k=1,k\neq a,b}^{n} p_k^2$
	$(h_ah_a \times h_bh_b)$	5	$(p_ap_b)^2$
	$(h_ah_a \times h_bh_k)$, $(h_bh_b \times h_ah_k)$	6	$2p_a^2p_b \sum_{\substack{k=1\\k\neq a,b}}^{n} p_k + 2p_ap_b^2 \sum_{\substack{k=1\\k\neq a,b}}^{n} p_k$
	$(h_ah_k \times h_bh_l)$, $(h_ah_l \times h_bh_k)$	7	$8p_ap_b \sum_{\substack{k=1\\k\neq a,b}}^{n} \sum_{\substack{l=1\\l\neq a,b,k}}^{n} p_kp_l$
(h_ah_b,h_ah_c)	$(h_ah_b \times h_ah_c)$	4	$4p_a^2p_bp_c$
	$(h_ah_a \times h_bh_c)$	6	$2p_a^2p_bp_c$
	$(h_ah_k \times h_bh_c)$	7	$4p_ap_bp_c \sum_{\substack{k=1\\k\neq a,b,c}}^{n} p_k$
(h_ah_a,h_bh_b)	$(h_ah_b \times h_ah_b)$	3	$4(p_ap_b)^2$
(h_ah_a,h_bh_c)	$(h_ah_b \times h_ah_c)$	4	$4p_a^2p_bp_c$
(h_ah_b,h_ch_d)	$(h_ah_d \times h_bh_c)$, $(h_ah_c \times h_bh_d)$	7	$8p_ap_bp_cp_d$

TABLE A3

Diplotype and its probability of offspring according to parental combination

Parent		Offspring	
Type	Combination	Diplotype	Probability
1	$h_ah_a \times h_ah_a$	h_ah_a	1
2	$h_ah_a \times h_ah_b$	h_ah_a	1/2
		h_ah_b	1/2
3	$h_ah_b \times h_ah_b$	h_ah_a	1/4
		h_ah_b	1/2
		h_bh_b	1/4
4	$h_ah_b \times h_ah_c$	h_ah_a	1/4
		h_ah_c	1/4
		h_ah_b	1/4
		h_bh_c	1/4
5	$h_ah_a \times h_bh_b$	h_ah_b	1
6	$h_ah_a \times h_bh_c$	h_ah_b	1/2
		h_ah_c	1/2
7	$h_ah_d \times h_bh_c$	h_ah_b	1/4
		h_ah_c	1/4
		h_bh_c	1/4
		h_bh_d	1/4

Calculation of the probability of the FSHS: For family f with n_f sibs, there are several possible FSHSs, and for each FSHS, there are also several possible parental combinations for each FSHS; then the probability of the ith FSHS in family f can be calculated as

$$P(G_1, G_2, \ldots, G_{n_f})_i = \sum_{j=1}^{7} \{P((G_1, G_2, \ldots, G_{n_f})_i \mid f_j, m_j)P(f_j, m_j)\}, \quad \text{(A2)}$$

where j is the type of parental combination shown in Table A2, and $P(f_j, m_j)$ can be obtained by using the expression listed in Table A2. $P((G_1, G_2, \ldots, G_{n_f})_i \mid f_j, m_j)$ is the conditional probability of the ith FSHS given the parental combination.