A rapid conditional enumeration haplotyping method in pedigrees

Guimin Gao, Ina Hoeschele

Genet. Sel. Evol. 40 (2008) 25–36
© INRA, EDP Sciences, 2008
DOI: 10.1051/gse:2007033

Available online at:
www.gse-journal.org

Original article

A rapid conditional enumeration haplotyping method in pedigrees

Guimin GAO[1], Ina HOESCHELE[2*]

[1] Department of Biostatistics, Section on Statistical Genetics, University of Alabama at Birmingham, Birmingham, Alabama 35294, USA
[2] Virginia Bioinformatics Institute and Department of Statistics, Virginia Tech, Blacksburg, Virginia 24061, USA

(Received 19 February 2007; accepted 27 July 2007)

Abstract – Haplotyping in pedigrees provides valuable information for genetic studies (*e.g.*, linkage analysis and association study). In order to identify a set of haplotype configurations with the highest likelihoods for a large pedigree with a large number of linked loci, in our previous work, we proposed a conditional enumeration haplotyping method which sets a threshold for the conditional probabilities of the possible ordered genotypes at every unordered individual-marker to delete some ordered genotypes with low conditional probabilities and then eliminate some haplotype configurations with low likelihoods. In this article we present a rapid haplotyping algorithm based on a modification of our previous method by setting an additional threshold for the ratio of the conditional probability of a haplotype configuration to the largest conditional probability of all haplotype configurations in order to eliminate those configurations with relatively low conditional probabilities. The new algorithm is much more efficient than our previous method and the widely used software SimWalk2.

haplotyping / pedigree / conditional probability / likelihood

1. INTRODUCTION

Haplotyping in a pedigree involves the consideration of the Space of All Consistent Haplotype Configurations (SACHC) for the pedigree based on all observed data (genotype data and pedigree structure). For a larger pedigree with a larger number of linked loci, the size of SACHC is too large for an exact method to be feasible. Most configurations in SACHC typically have very

* Corresponding author: inah@vt.edu

small conditional probabilities, so that only a relatively small subset of configurations with high conditional probabilities (or likelihood) is relevant [4]. Identifying a subset of configurations with the highest likelihoods and estimating their conditional probabilities in SACHC is an important computational step for genetic studies such as the calculation of haplotype frequencies and the estimation of identity-by-descent matrices. Likelihood-based sampling methods are often employed to infer the most likely haplotype configuration or a set of configurations with the highest likelihoods for a large pedigree with a large number of loci (*e.g.*, [7, 10]). These methods are flexible but can have high CPU time requirements and may converge very slowly. Some rule-based algorithms (*e.g.*, [1, 6, 8]) can be applied to large pedigrees, but these algorithms often assume zero recombinants or are more appropriate for pedigree data with a small expected number of recombinations [3], such as high density marker data in a short chromosomal region.

In our previous work [4], we proposed a conditional enumeration method based on computations of conditional probabilities and likelihood, and on setting a threshold λ ($\lambda < 1$) for the conditional probabilities of the possible ordered genotypes at every unordered individual-marker. It is often efficient to identify a set of configurations with the approximately highest likelihoods in SACHC. However, the computing time of this method can increase substantially, when (1) threshold λ is set very close to 1, (2) the pedigree contains a high proportion of homozygous genotypes and is less informative, or (3) inter-marker distances is large (say $\geqslant 5$ cM) and the pedigree contains a large number of recombinations which can increase the haplotype uncertainty of the individuals. In this study, we describe a rapid haplotyping algorithm based on a modification of the conditional enumeration method. The modified enumeration method is more efficient than the original method for large pedigrees with large numbers of loci. We compare the modified method by simulation in large pedigrees with the original method and with a sampling method implemented in the software SimWalk2 [10, 11], which is widely used for haplotyping in large pedigrees. SimWalk2 identifies a single haplotype configuration that is often nearly optimal.

2. METHODS

In this study, we assume linkage equilibrium between markers in the founders of the pedigree and we also assume that all individuals in a pedigree have been genotyped for all markers without genotype errors. We use the same

notation as in our previous work [4]. The combination of a specific individual and a specific marker locus is termed an individual-marker. The genotype of some individual-markers in non-founders can be ordered by their parents' genotypes. The observed data after this partial reconstruction are denoted by **D**. Let **U** denote all the remaining heterozygous individual-markers in a pedigree, each with an unordered genotype. Assume that the size of **U** is t. To reconstruct a haplotype configuration for the entire pedigree, one needs to assign an ordered genotype for each individual-marker in **U**.

Let $\{M_1, M_2, \ldots, M_t\}$ be a specific ordering of the individual-markers in **U**. Let m_i denote an ordered genotype assigned to individual-marker M_i, then a set of assignments $\{m_1, m_2, \ldots, m_t\}$ is a haplotype configuration for **U**. The joint probability of this configuration conditional on the observed data (**D**) is [4]

$$\Pr(m_1, m_2, \ldots, m_t \,|\, \mathbf{D}) = \prod_{i=1}^{t} p_i, \tag{1}$$

where $p_i = \Pr(m_i \,|\, m_1, \ldots, m_{i-1}, \mathbf{D})$ denotes the probability of an assigned ordered genotype m_i at individual-marker M_i, conditional on a set of assignments, $m_1, m_2, \ldots, m_{i-1}$, at the first $i-1$ individual-markers $M_1, M_2, \ldots, M_{i-1}$, and observed data **D**. Also, m_i is one of the two possible ordered genotypes m_i^l and m_i^s, where m_i^l (m_i^s) has the larger (smaller) conditional probability p_i^l (p_i^s) at individual-marker M_i, and $p_i^j = \Pr(m_i^j \,|\, m_1, \ldots, m_{i-1}, \mathbf{D})$ for $j = s, l$, with $p_i^s \leqslant p_i^l$, $p_i^s + p_i^l = 1$, and $p_i^l \geqslant 0.5$. Probability p_i is equal to one of the conditional probabilities p_i^s and p_i^l, so that $p_i \leqslant p_i^l$. Under the assumption of linkage equilibrium between markers in the founders, probabilities p_i, p_i^s and p_i^l can be calculated by an approximation method using only the informative flanking markers of the individual under consideration and its parents and offspring [4].

In our previous conditional enumeration haplotyping method (see [4] for details), we set a threshold λ for the conditional probabilities of ordered genotypes at every individual-marker, and assigned (one or two) ordered genotypes to each individual-marker in **U** sequentially by using an optimal (marker) search process. After the first $i-1$ individual-markers $\{M_1, M_2, \ldots, M_{i-1}\}$ have been assigned ordered genotypes, for each set of assignments $\{m_1, m_2, \ldots, m_{i-1}\}$ to these $i-1$ individual-markers, we temporarily treat each of the remaining individual-markers (not including the first $i-1$ individual-markers) in **U** as M_i, and calculate the corresponding conditional probability p_i^l for each of these M_i. We find the individual-marker with the highest conditional probability p_i^l among all the remaining individual-markers in **U**, and assign this

individual-marker to M_i. This procedure is called an optimal (marker) search process. At the individual-marker M_i, if $p_i^l \geqslant \lambda$, we delete the ordered genotype m_i^s, otherwise, both ordered genotypes, m_i^l and m_i^s are retained. After all individual-markers in **U** have been processed by this algorithm, we can obtain a subset of haplotype configurations with approximately the highest likelihoods. When setting $\lambda = 0.5$, the conditional enumeration haplotyping method becomes a conditional probability haplotyping method [4] which is very fast and identifies a single haplotype configuration by assigning a single ordered genotype m_i^l to each individual-marker M_i, and the optimal (marker) search process generates an optimal reconstruction order [4], $\{M_1, M_2, \ldots, M_i\}$.

Here, we propose a more efficient modified conditional enumeration haplotyping method by setting an additional threshold α for the conditional probabilities of haplotype configurations for **U** to eliminate some configurations with low conditional probabilities.

For the haplotype configuration $\{m_1, m_2, \ldots, m_t\}$, let q_i denote the ratio of conditional probability p_i to the larger conditional probability p_i^l at individual-marker M_i, *i.e.*, $q_i = p_i/p_i^l$ and $q_i \leqslant 1$. We define the important quantity Q_i as the product of $q_1,\ q_2, \ldots,\ q_i$ ($Q_i = \prod_{k=1}^{i} q_k$). For any integer $i \leqslant t$, we have $Q_i \geqslant Q_t$.

Let T denote the largest conditional probability of all haplotype configurations for **U** (T is unknown), and let R denote the ratio of the conditional probability of the haplotype configuration $\{m_1, m_2, \ldots, m_t\}$ to T, *i.e.*, $R = \Pr(m_1, m_2, \ldots, m_t \,|\, \mathbf{D})/T$ and $R > 0$. If R is very small (*e.g.*, $R < 0.001$, then the conditional probability $\Pr(m_1, m_2, \ldots, m_t \,|\, \mathbf{D})$ is very small relative to the largest conditional probability T, and the configuration $\{m_1, m_2, \ldots, m_t\}$ can be ignored when our purpose is to identify a set of configurations with the highest likelihoods. We describe an approximation method to estimate the upper bound of R.

Corresponding to the configuration $\{m_1, m_2, \ldots, m_t\}$, we reconstruct another haplotype configuration $\{m_1^l,\ m_2^l,\ \ldots,\ m_t^l\}$ for **U** in the same order $\{M_1, M_2, \ldots, M_t\}$, but each ordered genotype m_i^l is chosen with the larger conditional probability $\Pr(m_i^l \,|\, m_1^l, \ldots, m_{i-1}^l, \mathbf{D}) \geqslant 0.5$ at each individual-marker M_i ($i = 1, 2, \ldots, t$). The conditional probability of configuration $\{m_1^l, m_2^l, \ldots, m_t^l\}$ is $\Pr(m_1^l, m_2^l, \ldots, m_t^l \,|\, \mathbf{D}) = \prod_{i=1}^{t} \Pr(m_i^l \,|\, m_1^l, \ldots, m_{i-1}^l, \mathbf{D})$.

Note that probability $\Pr(m_i^l \mid m_1^l, \ldots, m_{i-1}^l, \mathbf{D})$ is different from probability p_i^l $(= \Pr(m_i^l \mid m_1, \ldots, m_{i-1}, \mathbf{D}))$. Since $\Pr(m_1^l, m_2^l, \ldots, m_t^l \mid \mathbf{D}) \leqslant T$, we have

$$\begin{aligned} R = \frac{\Pr(m_1, m_2, \ldots, m_t \mid \mathbf{D})}{T} &\leqslant \frac{\Pr(m_1, m_2, \ldots, m_t \mid \mathbf{D})}{\Pr(m_1^l, m_2^l, \ldots, m_t^l \mid \mathbf{D})} \\ &= \frac{\prod_{i=1}^{t} p_i}{\prod_{i=1}^{t} p_i^l} \cdot \frac{\prod_{i=1}^{t} p_i^l}{\Pr(m_1^l, m_2^l, \ldots, m_t^l \mid \mathbf{D})} = Q_t \cdot \frac{\prod_{i=1}^{t} \Pr(m_i^l \mid m_1, \ldots, m_{i-1}, \mathbf{D})}{\prod_{i=1}^{t} \Pr(m_i^l \mid m_1^l, \ldots, m_{i-1}^l, \mathbf{D})} \\ &= Q_t \prod_{i=1}^{t} r_i = Q_t r, \end{aligned}$$

where $r_i = \Pr(m_i^l \mid m_1, \ldots, m_{i-1}, \mathbf{D}) / \Pr(m_i^l \mid m_1^l, \ldots, m_{i-1}^l, \mathbf{D})$ and $r = \prod_{i=1}^{t} r_i$. Hence we obtain $R \leqslant Q_t r$. For any $i \leqslant t$, since $Q_i \geqslant Q_t$, we have

$$R \leqslant Q_i r. \tag{2}$$

From $\Pr(m_i^l \mid m_1^l, \ldots, m_{i-1}^l, \mathbf{D}) \geqslant 0.5$, we have $r_i \leqslant 2$ and $r \leqslant 2^t$. But we can find a smaller and more useful approximate upper bound on r. Consider the two haplotype configurations $\{m_1, m_2, \ldots, m_t\}$ and $\{m_1^l, m_2^l, \ldots, m_t^l\}$ described above. For a specific i $(\leqslant t)$, at each individual-marker M_j $(j = 1, \ldots, i-1)$ among the first $i-1$ individual-markers $\{M_1, M_2, \ldots, M_{i-1}\}$, the assignment m_j^l to M_j in the latter configuration is the ordered genotype with the larger probability $\Pr(m_j^l \mid m_1^l, \ldots, m_{j-1}^l, \mathbf{D})$ at the individual-marker M_j conditional on the assignments $\{m_1^l, m_2^l, \ldots, m_{j-1}^l\}$ to the individual-markers $\{M_1, \ldots, M_{j-1}\}$. But the assignment m_j for M_j in the former configuration may be the ordered genotype with the smaller probability at the individual-marker M_j conditional on the assignments $\{m_1, m_2, \ldots, m_{j-1}\}$ at the individual-markers $\{M_1, \ldots, M_{j-1}\}$. Based on pedigree knowledge, at the i-th individual-marker M_i, with very high probability,

$$\Pr(m_i^l \mid m_1, \ldots, m_{i-1}, \mathbf{D}) \leqslant \Pr(m_i^l \mid m_1^l, \ldots, m_{i-1}^l, \mathbf{D}), \tag{3}$$

or $r_i \leqslant 1$ (this inequality was confirmed in our data simulation). Even though for some individual-marker M_i inequality (3) may not hold, since both probabilities in inequality (3) are greater than 0.5, the two probabilities should be very close to each other. Thus from the definition $r = \prod_{i=1}^{t} r_i$, we obtain $r \leqslant 1$ approximately, and from inequality (2), for any $i \leqslant t$, we have

$$R \leqslant Q_i. \tag{4}$$

Given a small threshold 10^{α} ($10^{\alpha} < 1$; *e.g.*, $\alpha = -3$), for haplotype configuration $\{m_1, m_2, \ldots, m_t\}$, if we can find an integer i ($\leqslant t$), such that $Q_i \leqslant 10^{\alpha}$, then R will be very small and the configuration is ignorable and can be deleted when haplotyping in the pedigree. Since Q_i is calculated from the conditional probabilities of the first i assigned individual-markers in **U**, M_1, M_2, ..., M_i, by utilizing only these conditional probabilities (with no need for calculating the conditional probabilities at the remaining individual-markers, $M_{i+1}, \ldots$, M_t) we can infer whether the corresponding configuration can be deleted from SACHC. This elimination of configurations produces considerable saving in the computing time required for haplotyping.

Based on this principle for haplotype configuration elimination, we now modify our previous conditional enumeration haplotyping method. The new algorithm employs two user-determined threshold parameters: threshold λ for the conditional probabilities of ordered genotypes at every individual-marker ($\lambda \geqslant 0.5$) [4] and threshold 10^{α} for the ratio of the conditional probability of a haplotype configuration to T ($\alpha < 0$ and $10^{\alpha} \leqslant (1 - \lambda)/\lambda$, see the Appendix).

Suppose that ordered genotypes have been assigned to the first $i - 1$ individual-markers, for each set of assignments $\{m_1, m_2, \ldots, m_{i-1}\}$ to these $i - 1$ individual-markers, we find the individual-marker M_i with the highest conditional probability p_i^l among all the remaining individual-markers in **U**. And then we assign ordered genotypes to individual-marker M_i as follows ($i = 1, 2, \ldots, t$):

1. When $p_i^l \geqslant \lambda$, assign m_i^l to individual-marker M_i.
2. When $p_i^l < \lambda$, if assigning m_i^s to individual-marker M_i produces $Q_i \leqslant 10^{\alpha}$, then we only assign m_i^l to individual-marker M_i, otherwise we retain both ordered genotypes, m_i^l and m_i^s, for individual-marker M_i.

After all individual-markers in **U** have been processed with this algorithm, we will have obtained a set of haplotype configurations SACHC* ($\subseteq$ SACHC) for the pedigree. The elements (configurations) of SACHC* can be ranked by their likelihoods, and SACHC* will always contain a subset of configurations which have approximately the highest likelihoods among all configurations in SACHC of the pedigree. This subset of configurations with approximately the highest likelihoods can be obtained by eliminating configurations with lower likelihoods in SACHC*, as desired. The likelihood of a configuration can be calculated with the method described in [11] by adopting Haldane's model of recombination.

The number of haplotype configurations retained in SACHC*, the accuracy and the computing time of the modified conditional enumeration method can all be controlled with the chosen values for thresholds λ and α, and increase

with increasing absolute values of λ and α. When λ approaches 1 and α approaches $-\infty$ (10^{α} approaches 0), the modified conditional enumeration haplotyping method approaches an exhaustive enumeration method (exact method). The exhaustive enumeration method is computationally expensive or infeasible for large pedigrees or large numbers of loci.

In the modified method, we calculate the conditional probabilities for individual-markers in **U** by an approximation method [4], and we use inequality (4) which is only approximately true. Therefore, to guarantee the accuracy of the method, one should choose high absolute values for threshold parameters λ and α subject to maintaining an acceptable computing time. We recommend that the value of λ be set larger than 0.65, and that α ($\alpha < 0$) be set according to the average distance (d) between adjacent markers, with a decrease in the absolute value of α for an increase in d. For example, if $d \leqslant 2$ cM, we can set $\alpha \leqslant -1.0$; if $d \geqslant 5$ cM we can set α as large as -0.3 ($10^{-0.3} \approx 0.5$).

3. SIMULATION STUDIES AND RESULTS

To evaluate the performance of the modified method (abbreviated below as the "modified method"), we compared this method with our original conditional enumeration haplotyping method ("original method") and the widely used software SimWalk2 by analyzing three simulated pedigrees with different inter-marker distances (results from additional simulation studies evaluating our original method and comparing it to SimWalk2 can be found in [4]). The three simulated pedigrees had 163, 450 and 198 members with 18, 30 and 18 founders over 5, 8 and 6 generations, and a single linkage group consisting of 10, 10 and 20 bi-allelic markers with allele frequency of 0.5 and inter-marker distance of 10 cM, 5cM and 1.5 cM, respectively. Each father had two spouses, and each full sib family had three children.

Table I presents the haplotyping results from the analyses of the three pedigrees with the modified and the original conditional enumeration haplotyping methods. For the same λ value, when setting a sufficiently small value for α, the modified method identified a set of top haplotype configurations with the sum of likelihood ratios nearly identical to that of the set of corresponding top configurations identified by the original method (top configurations are those configurations with the estimated highest likelihoods, and a likelihood ratio is the ratio of the likelihood of a top configuration to that of the true configuration). However, the modified method uses much less computing time. The computing time of the original method can become unacceptably long. For example, in the analysis of the 198-member pedigree, when setting $\lambda > 0.973$,

Table I. Comparison of the modified ("Modified") and the original conditional enumeration haplotyping method ("Original") based on analyses of three simulated pedigrees.

N[a]	cM[b] (Loci[c])	Method	λ	α	Sum of likelihood ratios of top configurations[d]		
					100	2000	Time[e]
163	10 (10)	Original	0.835	-	1.339 e8	5.807 e8	4:15:20
		Modified	0.835	−2.0	1.338 e8	5.807 e8	0:06:47
			0.96	−2.2	1.435 e9	5.153 e9	0:58:57
			0.99	−2.2	1.435 e9	5.155 e9	1:01:34
450	5 (10)	Original	0.78	-	5.826 e13	4.781 e14	50:05:55
		Modified	0.78	−1.5	5.826 e13	4. 781 e14	0:31:13
			0.95	−1.32	5.826 e13	4.841 e14	0:22:30
			0.98	−1.75	6.870 e13	5.225 e14	2:26:50
198	1.5 (20)	Original	0.973	-	618.452	1298.1	53:04:28
		Modified	0.973	−3.0	618.452	1298.1	0:08:11
			0.99	−2.8	818.384	2202.01	0:07:24
			0.995	−3.0	818.384	2302.67	0:10:35

[a] N denotes the number of individuals in the pedigree.
[b] Distance between adjacent markers.
[c] The number of loci in the (single) linkage group.
[d] The sums of the likelihood ratios of the top 100 and 2000 configurations, where top configurations are those with the estimated highest likelihoods; likelihood ratio is the ratio of the likelihood of a top configuration to that of the true configuration. 1.339 e8 denotes 1.339×10^8.
[e] Time h:min:s on 2.00 GHz Intel (R) Xeon(TM) CPU (1 047 546 KB RAM, MS Window 2000).

the computing time (not listed in Tab. I) is much more than 53 h; in this case, the modified method (with $\lambda = 0.99$ or 0.995) identified a set of haplotype configurations quickly (in less than 11 min) whose sum of likelihood ratios was much higher than that from the original method (with $\lambda = 0.973$).

We note that in the analysis of the 198-member pedigree using the original method, when setting $\lambda \leqslant 0.970$, the computing time is very short ($\leqslant$ 0:07:41, see also Tab. II), but when setting $\lambda \geqslant 0.973$, the computing time increases substantially. The reason is that at many individual-markers in **U**, the larger conditional probabilities of the ordered genotypes are less than 0.973 but greater than 0.970. When setting $\lambda = 0.973$, two ordered genotypes are retained for each of these individual-markers, and the computing time increases exponentially with the number of these individual-markers. However when setting $\lambda \leqslant 0.970$, we only keep one ordered genotype for each of these individual-markers.

Table II. Comparison among the original and modified conditional enumeration haplotyping methods (denoted by "Original" and "Modified", respectively) and SimWalk2 (2.83) based on analyses of the 163-member and 198-member pedigrees.

N[a]	cM[b] (Loci[c])	Methods	λ	α	Highest log-likelihood (Number[d])	Time[e]
163	10 (10)	Original	0.835	-	−266.223 (17)	4:15:20
		Modified	0.98	−2.2	−265.221 (18)	0:58:57
		SimWalk2	-	-	−271.001 (1)	1:09:11
198	2.0 (15)	Original	0.97	-	−281.575 (16)	0:07:41
		Modified	0.995	−3.0	−281.575 (33)	0:10:35
		SimWalk2	-	-	−369.891 (1)	160:42:34

[a] N denotes the number of individuals in the pedigree.
[b] Distance between adjacent markers.
[c] The number of loci in the (single) linkage group.
[d] The number of haplotype configurations with the estimated highest log-likelihood (*e.g.*, for the 163-member pedigree the original method identified 17 configurations with the same log-likelihood of −266.233).
[e] Time on 2.00 GHz Intel (R) Xeon(TM) CPU (1 047 546 KB RAM, MS Window 2000).

We also note that the original and modified methods were run with many different values for thresholds λ and α. In Tables I and II below we only present the results for some representative values of the thresholds.

Table II presents results on the comparison of the modified method with the original method and SimWalk2 (2.83), based on analyses of the 163- and 198-member pedigrees. Table II shows that the modified method can identify a set of haplotype configurations with much higher log-likelihood and in much shorter time when compared to SimWalk2 which identifies a single configuration. For the 198-member pedigree with denser markers, the modified method identified 33 configurations with the same log-likelihood of −281.575 in about 10 min, while SimWalk2 identified a single configuration with the log-likelihood of −369.891 in about 160 h.

4. DISCUSSION

The modified conditional enumeration haplotyping method is an efficient algorithm for large pedigrees and large numbers of loci, in particular for the case of tightly linked markers, where the existing sampling methods are always computationally intensive.

For a large pedigree with high proportion of uninformative markers, we can control the computing time more effectively by setting a (user-determined)

control parameter (n_c) for the maximum number of retained haplotype configurations (the maximum size of SACHC*, *e.g.*, $n_c = 10\,000$). After the first $i-1$ unordered individual-markers $M_1, M_2, \ldots, M_{i-1}$ in $\mathbf{U}$ have been assigned ordered genotypes, if the total number of retained haplotype configurations exceeds n_c, the algorithm will adjust the values for thresholds λ and α so that only a single ordered genotype (the one with larger conditional probability p_i^l at M_i) is retained for each of the remaining unordered individual-markers in $\mathbf{U}$. This step can reduce the computing time dramatically. We note that the enumeration haplotyping methods use an optimal (marker) search process and assign ordered genotypes at each step to the individual-marker which has the most information in the corresponding individual and its parents and offspring among all remaining individual-markers in $\mathbf{U}$.

In this contribution, we have assumed linkage equilibrium between markers and that all individuals in a pedigree have been genotyped for all markers. We have work in progress extending our methods to pedigrees with missing marker data while accounting for founder allele frequencies and marker-marker linkage disequilibrium among high-density single nucleotide polymorphism (SNP) markers in the founders of a pedigree. The extension of the haplotyping method to deal with missing data also involves developing an efficient genotype elimination algorithm for large pedigrees with large numbers of loops for which the existing methods may not work well or be computationally infeasible (*e.g.*, [2, 5, 9]; O'Connell 2006, personal communications). We will report on this extension in a later communication.

The modified haplotyping method described above was implemented in a C/C++ program, which is available upon request from the first author for academic research.

ACKNOWLEDGEMENTS

This research was supported by grant R01 GM66103-01 (to I. Hoeschele) and grant R01 GM073766 from the National Institute of General Medical Sciences, USA, and partly supported by grants R01 ES09912 and U54 CA100949 from the National Institutes of Health, USA.

REFERENCES

[1] Baruch E., Weller J.I., Cohen-Zinder M., Ron M., Seroussi E., Efficient inference of haplotypes from genotypes on a large animal pedigree, Genetics 172 (2006) 1757–1765.

[2] Du F.X., Hoeschele I., A note on genotype and allele elimination in complex pedigrees with incomplete genotype data, Genetics 156 (2000) 2051–2062.
[3] Fishelson M., Dovgolevsky N., Geiger D., Maximum likelihood haplotyping for general pedigrees, Hum. Hered. 59 (2005) 41–60.
[4] Gao G., Hoeschele I., Sorensen P., Du F.X., Conditional probability methods for haplotyping in pedigrees, Genetics 167 (2004) 2055–2065.
[5] Henshall J.M., Tier B., Ker R.J., Estimating genotypes with independently sampled descent graphs, Genet. Res. 78 (2001) 281–288.
[6] Li J., Jiang T., Computing the minimum recombinant haplotype configuration from incomplete genotype data on a pedigree by integer linear programming, J. Comp. Biol. 12 (2005) 719–739 [http://www.cs.ucr.edu/~jili/haplotyping.html].
[7] Lin S., Skrivanek Z., Irwin M., Haplotyping using SIMPLE: caution on ignoring interference, Genet. Epidemiol. 25 (2003) 384–387.
[8] O'Connell J.R., Zero-recombinant haplotyping: applications to fine mapping using SNPs, Genet. Epidemiol. 19 (Suppl. 1) (2000) S64–S70.
[9] O'Connell J.R, Weeks D.E., An optimal algorithm for automatic genotype elimination, Am. J. Hum. Genet. 65 (1999) 1733–1740.
[10] Sobel E., Lange K., Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker sharing statistics, Am. J. Hum. Genet. 58 (1996) 1323–1337.
[11] Sobel E., Lange K., O'Connell J.R., Weeks D.E., Haplotyping algorithm, in: Speed T.P., Waterman M.S. (Eds.), IMA volumes in mathematics and its applications, Vol. 81, Genetic mapping and DNA sequencing, Springer-Verlag, New York, 1996, pp. 89–110.

APPENDIX: RELATIONSHIP OF TWO THRESHOLDS λ AND 10^{α}

In the modified method, after a set of ordered genotypes $\{m_1, m_2, \ldots, m_{i-1}\}$ have been assigned to the first $i-1$ individual-markers in **U**, we decide whether or not we should delete m_i^s at individual-marker M_i based on two steps (see also main text): (A) Conditional probability p_i^l at the *single* individual-marker M_i is compared to threshold λ; if $p_i^l \geqslant \lambda$, then we delete m_i^s. (B) When $p_i^l < \lambda$, the product $Q_i = \prod_{k=1}^{i} q_k$ is compared to threshold 10^{α} where Q_i is calculated from *a group of* conditional probabilities (p_k and p_k^l, $k = 1, \ldots, i$) at *a set of* i ($\geqslant 2$) individual-markers, under the assumption that m_i^s was assigned to individual-marker M_i; if $Q_i \leqslant 10^{\alpha}$, then delete m_i^s.

However, in step (B) a special case can occur, where $p_i^l < \lambda$, $Q_{i-1} = 1$, and $Q_i = q_i = p_i / p_i^l$ (*e.g.*, when the set of ordered genotypes $\{m_1^l, m_2^l, \ldots, m_{i-1}^l\}$ are assigned to the first $i-1$ individual-markers). In this case, if assigning m_i^s to individual-marker M_i produces $Q_i = q_i \leqslant 10^{\alpha}$, according to step (B), we should delete m_i^s, but here we do not hope to delete m_i^s because Q_i ($= q_i$) only

contains information from the conditional probabilities (p_i and p_i^l) at the single individual-marker M_i and deleting m_i^s by use of Q_i in step (B) would be equivalent to decreasing the value of threshold λ without using additional information from the conditional probabilities at the first $i-1$ individual-markers. In this situation, step (A) suffices because p_i^l has already contained the information from the conditional probabilities at the *single* individual-marker M_i. To avoid deleting m_i^s by step (B) in the special case ($p_i^l < \lambda$, $Q_{i-1} = 1$, and $Q_i = q_i$), in the modified method we set a limit for 10^α, $10^\alpha \leqslant (1-\lambda)/\lambda$. Then in step (B), when assigning m_i^s to individual-marker M_i, we have $Q_i = q_i = (1-p_i^l)/p_i^l > (1-\lambda)/\lambda$, so $Q_i > 10^\alpha$, and m_i^s will not be deleted in the special case.