Genetic Epidemiology 30: 220–230 (2006)

Haplotype Association Analysis for Late Onset Diseases Using Nuclear Family Data

Chun Li[1]* and Michael Boehnke[2]

[1]*Department of Biostatistics, Center for Human Genetics Research, Vanderbilt University, Nashville, Tennessee*
[2]*Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan*

In haplotype-based association studies for late onset diseases, one attractive design is to use available unaffected spouses as controls (Valle et al. [1998] Diab. Care 21:949–958). Given cases and spouses only, the standard expectation-maximization (EM) algorithm (Dempster et al. [1977] J. R. Stat. Soc. B 39:1–38) for case-control data can be used to estimate haplotype frequencies. But often we will have offspring for at least some of the spouse pairs, and offspring genotypes provide additional information about the haplotypes of the parents. Existing methods may either ignore the offspring information, or reconstruct haplotypes for the subjects using offspring information and discard data from those whose haplotypes cannot be reconstructed with high confidence. Neither of these approaches is efficient, and the latter approach may also be biased. For case-control data with some subjects forming spouse pairs and offspring genotypes available for some spouse pairs or individuals, we propose a unified, likelihood-based method of haplotype inference. The method makes use of available offspring genotype information to apportion ambiguous haplotypes for the subjects. For subjects without offspring genotype information, haplotypes are apportioned as in the standard EM algorithm for case-control data. Our method enables efficient haplotype frequency estimation using an EM algorithm and supports probabilistic haplotype reconstruction with the probability calculated based on the whole sample. We describe likelihood ratio and permutation tests to test for disease-haplotype association, and describe three test statistics that are potentially useful for detecting such an association. *Genet. Epidemiol.* 30:220–230, 2006. © 2006 Wiley-Liss, Inc.

Key words: EM algorithm; haplotype reconstruction; disease-haplotype association

Contract grant sponsor: Pilot and Feasibilty Grant; Contract grant sponsor: NIH; Contract grant number: HG00376.

*Correspondence to: Chun Li, Center for Human Genetics Research, Vanderbilt University Medical Center, 519 Light Hall, Nashville, TN 37232-0700. E-mail: chun.li@vanderbilt.edu

Received 4 May 2005; Accepted 27 July 2005
Published online 8 February 2006 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20139

INTRODUCTION

Association analysis is a powerful method to map genes for complex diseases [Risch and Merikangas, 1996]. Because multimarker haplotypes may yield more information than single markers do, haplotype-based association analyses have the potential to be more powerful than those based on single markers.

A variety of sampling designs may be used in association studies, including the case-control design, case-parents trios, discordant sib pairs, or more general pedigrees. One attractive design, particularly for late onset diseases, is to use available unaffected spouses as controls [Valle et al., 1998]. Given cases and spouses only, the standard expectation-maximization (EM) algorithm can be used to estimate haplotype frequencies [Excoffier and Slatkin, 1995; Long et al., 1995]. When offspring for some of the spouse pairs are available, their genotypes provide additional information about the haplotypes of the parents [Valle et al., 1998]. It is then inefficient to ignore offspring genotype information. A possible solution is to reconstruct haplotypes for the subjects based on their own genotypes and those of their spouse and offspring, and conduct association analysis using the haplotypes reconstructed with high confidence. However, this approach often will result in loss of information from subjects whose haplotypes cannot be reconstructed with high confidence, further leading to inefficiency. It also can introduce bias in haplotype frequency estimation since some genotype patterns are intrinsically harder to resolve.

In addition to case-spouse control pairs, we often will have case-offspring pairs, cases without family information, and unrelated controls ascertained independently of the cases. These individuals provide additional information on haplotype frequency estimation and help increase the power to detect disease-haplotype association. For such data, we propose a unified, likelihood-based method of haplotype inference. For haplotype frequency estimation, we introduce an EM algorithm that makes use of offspring genotype information when available to apportion the contribution of ambiguous haplotypes for the subjects, resulting in more efficient haplotype frequency estimation. For subjects without offspring genotype information, haplotypes are apportioned as in the standard EM algorithm for case-control data. The method also supports probabilistic haplotype reconstruction with the probability calculated on the basis of the whole sample. To test for disease-haplotype association, we introduce likelihood ratio and permutation tests, and describe three other potentially useful test statistics.

METHODS

First, we outline the data structure, notation, and the hypotheses of interest. Second, for a sample composed of cases, spouse or unrelated controls, and offspring of some spouse pairs, cases, or controls, we introduce an EM algorithm to estimate haplotype frequencies. Third, we introduce likelihood ratio and permutation tests to test for disease-haplotype association, and define three permutation test statistics. Fourth, we extend our method to X-linked markers. Fifth, we apply Bayes' rule to haplotype reconstruction, using frequencies estimated based on the whole sample. Sixth, we describe a simulated population of disease-associated and non-disease-associated haplotypes that will be used in our computer simulations.

DATA STRUCTURE, NOTATION, AND HYPOTHESES OF INTEREST

In haplotype association analysis in a case-control study, we compare the distributions of haplotypes in the case and control groups. Some individuals from the two groups may be related by having offspring in common, which is the situation when we sample unaffected spouses of the cases and their offspring for late onset diseases [Valle et al., 1998]. In what follows, we assume the data consist of three classes of pedigree structures: nuclear families with known affection status for the parents (for example, pedigree P_1 in Fig. 1); nuclear families with only one parent's genotype available (for example, pedigree P_2 in Fig. 1); and individuals (singletons) with known affection

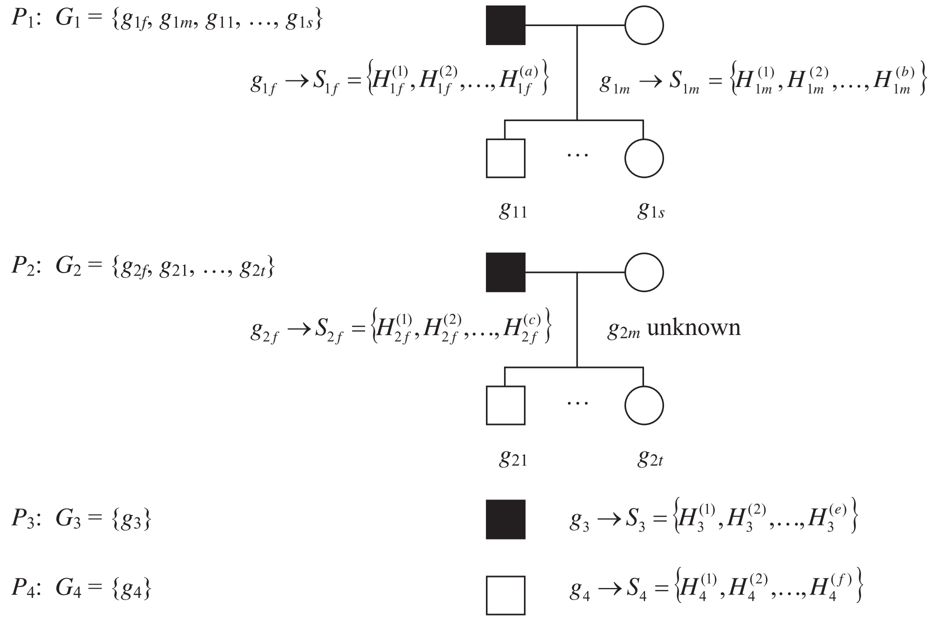
P_1: $G_1 = \{g_{1f}, g_{1m}, g_{11}, \ldots, g_{1s}\}$
$g_{1f} \rightarrow S_{1f} = \{H_{1f}^{(1)}, H_{1f}^{(2)}, \ldots, H_{1f}^{(a)}\}$
$g_{1m} \rightarrow S_{1m} = \{H_{1m}^{(1)}, H_{1m}^{(2)}, \ldots, H_{1m}^{(b)}\}$
…
g_{11}
g_{1s}
P_2: $G_2 = \{g_{2f}, g_{21}, \ldots, g_{2t}\}$
$g_{2f} \rightarrow S_{2f} = \{H_{2f}^{(1)}, H_{2f}^{(2)}, \ldots, H_{2f}^{(c)}\}$
g_{2m} unknown
…
g_{21}
g_{2t}
P_3: $G_3 = \{g_3\}$
$g_3 \rightarrow S_3 = \{H_3^{(1)}, H_3^{(2)}, \ldots, H_3^{(e)}\}$
P_4: $G_4 = \{g_4\}$
$g_4 \rightarrow S_4 = \{H_4^{(1)}, H_4^{(2)}, \ldots, H_4^{(f)}\}$

Fig. 1. Examples of pedigree structures used in haplotype analysis. P_1: a nuclear family with s offspring; P_2: a nuclear family with t offspring and one parent's genotype unavailable; P_3: an affected singleton; P_4: an unaffected singleton. Individual multilocus genotypes are denoted as g. The haplogenotypes consistent with the subjects' genotypes are listed after arrows.

status (for example, pedigrees P_3 and P_4 in Fig. 1). Each parent in a nuclear family may be affected or unaffected. Examples of singletons include unrelated cases and unrelated controls. For a nuclear family, if we do not have offspring genotype information, we assume the parents are genetically independent and effectively form two singleton pedigrees. Singletons and parents in nuclear families contribute their haplotypes to the case and control groups according to their affection status, and we refer to them as subjects in this paper. Let n_1 and n_2 be the numbers of subjects in the case and control groups, respectively. In this paper, we ignore the affection status of the offspring, a reasonable choice for late onset diseases, but perhaps not for earlier onset diseases.

We first consider haplotypes defined on a dense set of m markers on an autosome and assume Hardy-Weinberg equilibrium. When a haplotype region is small, recombination is rare; to simplify calculation, we assume recombination does not occur within our sampled individuals. We denote multilocus genotypes (phase unknown) as g, haplogenotypes (phase known) as H, and haplotypes as h. A haplogenotype H consists of two unordered haplotypes, which we denote $H=\{h_1, h_2\}$. For simplicity, we often call multilocus genotypes as genotypes.

Let $p(h)$ and $q(h)$ be the frequencies of haplotype h in the case and control groups, respectively, and let $\mathbf{p}$ and $\mathbf{q}$ be the corresponding vectors of haplotype frequencies. For a haplogenotype $H=\{h_1, h_2\}$, Hardy-Weinberg equilibrium implies that the probabilities of H in the case and control groups are $p(H)=2p(h_1)p(h_2)$ and $q(H)=2q(h_1)q(h_2)$ if $h_1 \neq h_2$, and $p(H)=[p(h_1)]^2$ and $q(H)=[q(h_1)]^2$ if $h_1=h_2$. We wish to test the null hypothesis H_0: $p(h)=q(h)$ for all haplotypes h against the alternative hypothesis H_a: $p(h) \neq q(h)$ for some h.

For X-linked markers, a male has only one haplotype, and the probability of a haplogenotype for a female can be derived similarly as above under Hardy-Weinberg equilibrium. The null and alternative hypotheses are the same as those for autosomes.

AUTOSOMAL HAPLOTYPE FREQUENCY ESTIMATION

Consider a nuclear family like P_1 in Figure 1. In P_1, the father is affected and has multilocus genotype g_{1f} and the mother is unaffected and has multilocus genotype g_{1m}. Let $S_{1f}=\{H_{1f}^{(1)}, H_{1f}^{(2)}, \ldots, H_{1f}^{(a)}\}$ be the set of haplogenotypes that are consistent with the father's genotypes and similarly $S_{1m}=\{H_{1m}^{(1)}, H_{1m}^{(2)}, \ldots, H_{1m}^{(b)}\}$ be the set for the mother. If there are no missing genotypes, $1 \leq a,\ b \leq 2^{m-1}$. If one or more marker genotypes are missing, the set of consistent haplogenotypes can be large unless the two alleles at the marker can be inferred from the spouse and offspring. For a haplogenotype $H_{1f}^{(i)}$, we denote its two haplotypes as $h_{1f,1}^{(i)}$ and $h_{1f,2}^{(i)}$, that is $H_{1f}^{(i)}=\{h_{1f,1}^{(i)}, h_{1f,2}^{(i)}\}$.

Let s be the number of offspring in family P_1 and $G_1=\{g_{1f}, g_{1m}, g_{11}, \ldots, g_{1s}\}$ be the multilocus genotypes of all individuals in the family. The likelihood for the family is

$$\Pr(G_1|\mathbf{p},\mathbf{q}) = \sum_{i=1}^{a}\sum_{j=1}^{b} p(H_{1f}^{(i)})q(H_{1m}^{(j)}) \times \prod_{v=1}^{s} \Pr(g_{1v}|H_{1f}^{(i)}, H_{1m}^{(j)}).$$

Notice that we use the haplotype frequencies of the case group for the father because he is affected and those of the control group for the mother because she is unaffected. Under the assumption of no recombination, the calculation of $\Pr(g_{1v}|H_{1f}^{(i)}, H_{1m}^{(j)})$ is straightforward. Given the parental haplotypes $H_{1f}^{(i)}=\{h_{1f,1}^{(i)}, h_{1f,2}^{(i)}\}$ and $H_{1m}^{(j)}=\{h_{1m,1}^{(j)}, h_{1m,2}^{(j)}\}$, an offspring can equally likely have one of the four possible haplogenotypes: $\{h_{1f,1}^{(i)}, h_{1m,1}^{(j)}\}$, $\{h_{1f,1}^{(i)}, h_{1m,2}^{(j)}\}$, $\{h_{1f,2}^{(i)}, h_{1m,1}^{(j)}\}$, and $\{h_{1f,2}^{(i)}, h_{1m,2}^{(j)}\}$. Then the conditional probability $\Pr(g_{1v}|H_{1f}^{(i)}, H_{1m}^{(j)})$ is the proportion of these haplogenotypes that are consistent with g_{1v}.

The above approach can be modified for nuclear families with only one parent's genotype available, like P_2 in Figure 1. In P_2, the mother is unaffected with genotype unavailable. Her set of possible haplogenotypes consists of all possible haplogenotypes. When there is only one offspring ($t=1$), we may equivalently use haplotype frequencies for the mother to calculate the likelihood

$$\Pr(G_2|\mathbf{p},\mathbf{q}) = \sum_{i=1}^{c}\sum_{h} p(H_{2f}^{(i)})q(h)\Pr(g_{21}|H_{2f}^{(i)}, h)$$

where the second summation is over all haplotypes and g_{21} is the genotype of the single offspring. The calculation of $\Pr(g_{21}|H_{2f}^{(i)}, h)$ is similar as above. Given the father's haplogenotype $H_{2f}^{(i)}$ and the mother's haplotype h, the offspring can equally likely have one of two possible haplogenotypes, and $\Pr(g_{21}|H_{2f}^{(i)},h)$ is the proportion of these haplogenotypes that are consistent with g_{21}.

For a parent-offspring family such as P_2 (Fig. 1), one might want to use offspring genotypes to narrow down the list of possible haplogenotypes for the parent, and then treat the parent as a singleton with the narrowed list of haplogenotypes. In general, this is not a good approach because it is possible that a parental haplogenotype is consistent with an offspring's genotype but the offspring haplotype that is supposed to come from the other parent is very rare in the population. This makes the parental haplogenotype have a smaller probability to start with than if the parent were a singleton. Thus, an appropriate approach is to take into account the frequency of the haplotype from the other parent as described above.

For an individual for whom we do not have offspring genotype information, like the affected singleton in P_3 or the unaffected singleton in P_4 (Fig. 1), the probability calculation simplifies to that appropriate for a case-control study. Let $S_3 = \{H_3^{(1)}, H_3^{(2)}, \ldots H_3^{(e)}\}$ and $S_4 = \{H_4^{(1)}, H_4^{(2)}, \ldots H_4^{(f)}\}$ be the sets of all haplogenotypes that are consistent with their multilocus genotypes G_3 and G_4 respectively. The likelihood for the singletons are

$$\Pr(G_3|\mathbf{p}) = \sum_{i=1}^{e} p(H_3^{(i)})$$

and

$$\Pr(G_4|\mathbf{q}) = \sum_{i=1}^{f} q(H_4^{(i)}).$$

We now describe an EM algorithm for calculating the maximum likelihood estimates (MLEs) of the haplotype frequencies. After the kth iteration, let $p^{(k)}(h)$ and $q^{(k)}(h)$ be the estimated frequencies of haplotype h in the case and control groups, respectively, and let $p^{(k)}(H)$ and $q^{(k)}(H)$ be the corresponding probabilities for haplogenotype H.

Consider the father in P_1 (Fig. 1). If $a = 1$ and $h_{1f,1}^{(1)} = h_{1f,2}^{(1)}$ (that is, he is homozygous at all markers), he contributes two copies of this haplotype to the case group; if $a = 1$ and $h_{1f,1}^{(1)} \neq h_{1f,2}^{(1)}$ (that is, he is homozygous at all but one marker), he contributes one copy of each haplotype. If $a > 1$ and there are no missing genotypes, all haplotypes $h_{1f,j}^{(i)}$ ($i = 1, \ldots, a$; $j = 1, 2$) are different. At the $(k+1)$th iteration, using Bayes' rule, the posterior probability of haplogenotype $H_{1f}^{(i)} = \{h_{1f,1}^{(i)}, h_{1f,2}^{(i)}\}$ ($i = 1, \ldots, a$) is

$$\begin{aligned}\Pr&\left(H_{1f}^{(i)}|G_1; \mathbf{p}^{(k)}, \mathbf{q}^{(k)}\right)\\ &= \frac{\Pr(H_{1f}^{(i)}|\mathbf{p}^{(k)}, \mathbf{q}^{(k)})\Pr(G_1|H_{1f}^{(i)}; \mathbf{p}^{(k)}, \mathbf{q}^{(k)})}{\Pr(G_1|\mathbf{p}^{(k)}, \mathbf{q}^{(k)})}\\ &= \frac{p^{(k)}(H_{1f}^{(i)})\sum_{j=1}^{b} q^{(k)}(H_{1m}^{(j)})\prod_{v=1}^{s}\Pr(g_{1v}|H_{1f}^{(i)}, H_{1m}^{(j)})}{\sum_{l=1}^{a}\sum_{j=1}^{b} p^{(k)}(H_{1f}^{(l)})q^{(k)}(H_{1m}^{(j)})\prod_{v=1}^{s}\Pr(g_{1v}|H_{1f}^{(l)}, H_{1m}^{(j)})}.\end{aligned} \tag{1}$$

This is the father's contribution of haplotype $h_{1f,j}^{(i)}$ ($j = 1, 2$) to the case group through haplogenotype $H_{1f}^{(i)}$. If one or more marker genotypes are missing, his contribution of a haplotype may come through several haplogenotypes. Since the sum of the posterior probabilities (1) for all possible haplogenotypes is one, the father's total contribution of haplotypes is two. Similarly, we can calculate the mother's haplotype contributions to the control group.

For family P_2 (Fig. 1), when there is only one offspring ($t = 1$), we may directly use haplotype frequencies for the mother. The father's contribution can be simplified to

$$\begin{aligned}\Pr&\left(H_{2f}^{(i)}|G_2; \mathbf{p}^{(k)}, \mathbf{q}^{(k)}\right)\\ &= \frac{p^{(k)}(H_{2f}^{(i)})\sum_h q^{(k)}(h)\Pr(g_{21}|H_{2f}^{(i)}, h)}{\sum_{l=1}^{c}\sum_h p^{(k)}(H_{2f}^{(l)})q^{(k)}(h)\Pr(g_{21}|H_{2f}^{(l)}, h)}\end{aligned} \tag{2 a}$$

and the mother's contribution of haplotype h to the control group can be calculated as

$$\begin{aligned}\Pr&\left(h|G_2; \mathbf{p}^{(k)}, \mathbf{q}^{(k)}\right)\\ &= \frac{q^{(k)}(h)\sum_{i=1}^{c} p^{(k)}(H_{2f}^{(i)})\Pr(g_{21}|H_{2f}^{(i)}, h)}{\sum_{i=1}^{c}\sum_h p^{(k)}(H_{2f}^{(i)})q^{(k)}(h)\Pr(g_{21}|H_{2f}^{(i)}, h)}.\end{aligned} \tag{2 b}$$

For the singleton in P_3 (Fig. 1), if $e = 1$, the haplotypes are known and contributed directly to the case group. If $e > 1$, at the $(k+1)$th step, the posterior probability of $H_3^{(i)} = \{h_{31}^{(i)}, h_{32}^{(i)}\}$ ($i = 1, \ldots, e$) is

$$\Pr(H_3^{(i)}|G_3, \mathbf{p}^{(k)}) = \frac{p^{(k)}(H_3^{(i)})}{\Pr(G_3|\mathbf{p}^{(k)})} = \frac{p^{(k)}(H_3^{(i)})}{\sum_{l=1}^{e} p^{(k)}(H_3^{(l)})} \tag{3}$$

and this is the contribution of haplotype $h_{3j}^{(i)}$ ($j = 1,2$) to the case group through $H_3^{(i)}$. The total contribution of haplotypes again is two. Note that the apportionment (2) is the same as in the standard EM algorithm for a case-control study [Excoffier and Slatkin, 1995; Long et al., 1995]. The posterior probabilities of haplogenotypes for the singleton in P_4 (Fig. 1) can be similarly calculated, with haplotype frequencies $\mathbf{q}^{(k)}$ replacing $\mathbf{p}^{(k)}$.

In both families P_1 and P_2, if we had ignored the offspring genotype information, we would have calculated the parents' contributions assuming they were singletons. We achieve better apportionment of their contributions by using additional information provided by the offspring genotypes.

Once we have calculated contributions of haplotypes for all the subjects (singletons and parents in nuclear families), we update haplotype frequency estimates in the case and control groups. For a haplotype h, the updated estimate $p^{(k+1)}(h)$ is the sum of all contributions of h to the case group divided by $2n_1$, where n_1 is the number of cases. Similarly, $q^{(k+1)}(h)$ is the sum of all contributions of h to the control group divided by $2n_2$, where n_2 is the number of controls. We repeat this process until the frequency estimates converge. This is an example of an allele-counting algorithm [Ceppellini et al., 1955], and also can be shown to be an EM algorithm. To help ensure that the frequency estimates are MLEs rather than local maxima, we restart the algorithm with a variety of non-zero starting values.

Under the null hypothesis, we do a similar calculation. Here, all subjects contribute to a single group and only one set of haplotype frequencies will be estimated. At each iteration, the updated frequency for haplotype h is the sum of contributions to h from all subjects divided by $2(n_1+n_2)$.

Note that in (1), (2a), and (2b), the conditional probabilities $\Pr(g_{1v}|H_{1f}^{(i)}, H_{1m}^{(j)})$ and $\Pr(g_{2v}|H_{2f}^{(i)}, h)$ do not depend on haplotype frequencies and can be calculated only once and stored. As a result, although missing marker genotypes for subjects will increase the number of consistent haplogenotypes and result in longer computation time, missing marker genotypes for offspring have relatively small effect on computation since offspring genotypes are used only in these conditional probabilities.

TESTING FOR DISEASE-HAPLOTYPE ASSOCIATION

Let $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$ be the MLEs of the haplotype frequencies under the alternative hypothesis. Given $(\hat{\mathbf{p}}, \hat{\mathbf{q}})$, the likelihood for a nuclear family like P_1 in Figure 1 is

$$L_a(P_1) = \Pr(G_1|\hat{\mathbf{p}}, \hat{\mathbf{q}}) = \sum_{i=1}^{a}\sum_{j=1}^{b} \hat{p}(H_{1f}^{(i)})\hat{q}(H_{1m}^{(j)}) \prod_{v=1}^{s} \Pr(g_{1v}|H_{1f}^{(i)}, H_{1m}^{(j)})$$

and the likelihood for singletons like P_3 and P_4 in Figure 1 are $L_a(P_3) = \sum_{i=1}^{e} \hat{p}(H_3^{(i)})$ and $L_a(P_4) = \sum_{i=1}^{f} \hat{q}(H_4^{(i)})$. Under the null hypothesis of equal haplotype frequencies, their likelihood $\mathrm{L}_0(\cdot)$ can be similarly calculated using the MLEs estimated based on the combined group.

To test for disease-haplotype association, we may calculate the likelihood-ratio test statistic

$$T = 2\left(\sum_{P} \log L_a(P) - \sum_{P} \log L_0(P)\right) \quad (4)$$

where the sums are over all independent pedigrees P, and compare T to the chi-squared distribution with $N-1$ degrees of freedom, where N is the number of haplotypes truly present in the population. In reality, we often do not know N. One might instead use the number $N_{\max}$ of all theoretically possible haplotypes. However, as expected, simulations suggest that when N is much smaller than $N_{\max}$, as may be the case for a dense set of markers owing to linkage disequilibrium (LD), the likelihood-ratio test using $N_{\max}-1$ degrees of freedom can be very conservative (data not shown). One might also use the number N_{obs} of "observed" haplotypes. This also is inappropriate because N_{obs} tends to be smaller than real N and thus leads to anti-conservative tests; in addition, ignoring the inherent variation in estimating N_{obs} may also lead to anti-conservative tests. Even if we know N, the asymptotics may not work well for a moderate number of markers because the number of parameters can be large compared to the number of subjects; our simulations on seven biallelic markers using 400 spouse pairs showed that the results can be very conservative (data not shown).

Alternatively, we may carry out a permutation test by permuting the affection status of the subjects—singletons and parents in nuclear families. Specifically, to generate a permuted data set, we randomly assign n_1 subjects to the case group and the other n_2 subjects to the control group. The observed test statistic then can be compared with the statistics calculated for the permuted data sets to assess significance. The log-likelihood ratio (4) is one possible choice of test statistic. In this situation, since a permuted data set does not change the maximum likelihood of the data under the null, we can equivalently base our test on the log-likelihood $\sum_P \log L_a(P)$ under the alternative hypothesis.

In an effort to increase the power to detect disease-haplotype association, we define three additional test statistics, in which we combine haplotypes into a few categories, and calculate statistics based on the newly defined categories. By combining haplotypes, we seek to consolidate signals and reduce variation, so as to increase power to detect disease-haplotype association. First, since a disease-predisposing variant may have originated on a single founder haplotype, we compare each haplotype with the combined category consisting of all the other haplotypes, calculate chi-squared statistics for the resulting 2×2 tables, and choose the largest statistic. We call this the "best-haplotype" statistic. Second, because there may exist >1 founder disease haplotypes, we also compare every two haplotypes with the combined category of the rest, and choose the largest chi-squared statistic for the resulting 2×3 tables. We call this the "best-two-haplotype" statistic. Third, for complex diseases, it is likely that a disease predisposing variant emerged long ago and recombinations between the variant and a tightly linked marker have occurred. Alternatively, multiple independent founder variants may be present. In either of these situations, a disease variant may be associated with multiple haplotypes. Hence, we also combine the haplotypes with higher frequency estimates in the case group than in the control group into one category, and the remaining haplotypes into a second category, and then calculate chi-squared statistic for the resulting 2×2 table. We call this the "high-vs-low" statistic. For these statistics, the best haplotype or grouping of the haplotypes may vary among permuted data sets.

The permutation test is computationally intensive since the EM algorithm under the alternative hypothesis must be carried out for each permuted data set. However, because the additional information of offspring genotypes often makes the number of consistent haplogenotypes for a parent much smaller (see Results), it is often much faster to carry out the test using offspring information than not.

X-LINKED HAPLOTYPE ANALYSIS

Our method can be extended to haplotype analysis on the X chromosome. For X chromosome markers, with no missing genotype and assuming no recombination and no genotype error, paternal haplogenotype is observed and maternal haplogenotype is resolved if offspring genotypes are available. With missing genotypes, our method can be easily modified to analyze the data. In this situation, since a male offspring does not inherit X chromosome from his father, a spouse pair with only male offspring is effectively unrelated for the purpose of this analysis, and the family can be broken into two pedigrees—the father as a singleton and the mother-son as a pedigree. For a mother-son pedigree, the offspring genotypes are used to narrow down the set of possible haplogenotypes for the mother.

When haplotype frequencies are tallied for calculating test statistics, a male will contribute one haplotype, while a female will contribute two haplotypes. If we suspect gender could play a role in the disease risk, the permutations need to be carried out separately for men and women. For this situation, it is ideal to have a balanced subject ascertainment with equal numbers of men and women in the case and control groups.

PROBABILISTIC HAPLOTYPE RECONSTRUCTION

Although we do not need to infer individual haplotypes for detecting association of a haplotype with disease, knowledge of haplotypes may be useful for further analyses. For example, once we have detected association of a haplotype with disease, we may wish to carry out further phenotypic evaluation on individuals carrying the associated haplotype, or to sequence a subset of these individuals to find disease-predisposing variants. If a subject is heterozygous at ≤ 1 marker, his/her haplotypes are already known. For those who have >1 possible haplogenotypes, we reconstruct haplotypes probabilistically.

For autosomal markers, the equations (1)–(3) provide the basis for our haplotype reconstruction. For example, for the father in P_1 in Figure 1, the (posterior) probability that his haplogenotype is $H_{1f}^{(i)}$ is given in (1), with $\mathbf{p}^{(k)}$ and $\mathbf{q}^{(k)}$ replaced with the MLEs $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$. This calculation of posterior probability is efficient in the sense that it is based on the MLEs of the parameters, which in turn are estimated on the basis of the whole data set. Posterior probability for X-linked haplotypes can be similarly calculated.

For each subject, the haplogenotype with the highest posterior probability may be assigned as his/her reconstructed haplogenotype. We call this method "probability-vote" reconstruction. To be confident in our reconstruction, we may want to assign a reconstructed haplogenotype only if its posterior probability exceeds a pre-specified level β, say 99%. We call this method "level-β" reconstruction. For both methods, if a reconstructed haplogenotype is correct, we call it a success. If offspring genotypes are available, we expect to improve the reconstruction success rate.

SIMULATION OF POPULATION OF HAPLOTYPES

To assess our methods, we carried out computer simulations. For our simulations, we constructed a population of disease-associated and non-disease-associated autosomal haplotypes defined on five biallelic markers by using an evolutionary algorithm first described by Devlin and Risch [1995] and modified by Lange and Boehnke [2004].

We assumed a founder population of 1000 individuals, admixed from two subpopulations each of size 500. Founder marker data were randomly generated under linkage equilibrium within subpopulations. However, the allele frequencies were very different in the two subpopulations, resulting in LD among the markers in the mixed founder population. The first subpopulation was simulated using allele frequencies .90 and .10 for each marker, while the second subpopulation was simulated using allele frequencies .10 and .90 for each marker. The disease-predisposing variant was completely linked to the third marker, and came only from subpopulation 1 with frequency .40. In the mixed founder population, all five markers had equally frequent alleles, and the disease allele frequency was .20. When growing the population, to shorten the simulation time, we intentionally used a large value ($\theta = .10$) for recombination fraction between adjacent markers. We grew the founder population exponentially over 50 generations to reach a final population of about 500,000 individuals. In the final population, LD between the disease gene and the five markers as measured by D' were .77, .74, .83, .62, and .76, and as measured by r^2 were .13, .15, .17, .09, and .13, respectively. All 32 possible haplotypes were present in the final population. In addition, the disease variant could be found on all 32 haplotypes, but at different frequencies.

RESULTS

TYPE I ERROR RATE

We carried out computer simulations to check if the type I error rate is under control for our test. We simulated 10,000 replicate data sets of 400 affected-unaffected spouse pairs with one offspring, and with two random genotype missing rates (0% and 10%). We compared the likelihood-ratio test statistic with the chi-squared distribution with 31 degrees of freedom because there were 32 haplotypes in our simulated population. For the permutation tests, we generated 999 random permutations for each replicate data set and used $(x+1)/1{,}000$ as the P-value estimate, where x is the number of permuted data sets that were no less extreme than the replicate. Table 1 lists the estimated type I error rates at significance level $\alpha = .01$. The likelihood-ratio test appeared to be anti-conservative when offspring genotype information is used, probably because the test statistic converges to the limiting chi-squared distribution slowly due to reduced number of sampling units (that is, two subjects join to form a family) and more complicated likelihood structure for each unit; our simulations with 2,000 case-spouse-offspring trios showed the type I error rate was well under control (data not shown). Given that the tests are correlated to some extent, especially for the three statistics introduced in this paper, the results for the permutation tests are within the expected range of random fluctuation due to sampling, and suggest that type I error rates are under control.

POWER

We also carried out simulations to compare the power to detect disease-haplotype association with and without offspring genotype information, and among the different test procedures we introduced. We considered three disease models (Table 2) and two random genotype missing rates (0%, 10%). The models are additive, dominant,

TABLE 1. Estimated type I error rates (%) at significance level $\alpha = 1\%$ with random missing genotype rates 10% and 0%

Missing genotype rate:	10%		0%	
Use offspring genotype:	No	Yes	No	Yes
Likelihood-ratio test:	0.81	1.70	0.92	1.83
Permutation test:				
Log-likelihood	1.10	1.03	0.89	1.00
High vs. low	0.97	1.07	1.07	1.20
Best haplotype	1.03	1.03	0.96	1.00
Best two haplotypes	1.01	1.05	1.01	0.95

Note: Based on our simulated population and 10,000 replicate data sets of 400 affected-unaffected spouse pairs each with one offspring.

and recessive, with disease allele frequency .20, disease prevalence .10, and relative risk $\lambda_S = 1.20$ for sibs of an affected individual. For each disease model and each genotype missing rate, we simulated 2,500 replicate data sets of 400 affected-unaffected spouse pairs with one offspring. Table 3 lists the estimated power to detect disease-haplotype association at significance level $\alpha = .01$.

As expected, for all the tests, the power to detect disease-haplotype association was higher when using offspring genotype information than when not using it; the two exceptions in Table 3 are likely due to random variation in simulations, and repeated simulations have shown the correct order. The power also was higher when the genotype missing rate was lower. Missing genotype data appeared to have a slightly stronger adverse effect on the power of the three test statistics based on combining haplotypes, presumably because the performance of these statistics relies on correct grouping of haplotypes, which depends on the accuracy of haplotype frequency estimates. Nonetheless, for our simulated population, these statistics showed great potential in detecting disease-haplotype association. With the same disease allele frequency, prevalence, and sib relative risk, it was easier to detect disease-haplotype association when the disease model was additive or dominant than when it was recessive. This is because when the models have same modest disease allele frequency, same sibling relative risk, and same disease prevalence, a recessive model tends to yield smaller difference in disease allele frequency in the cases and controls than a dominant or additive model.

TABLE 2. Disease models used in simulations

	p	f_0	f_1	f_2
Model 1: Additive	.20	.055	.166	.277
Model 2: Dominant	.20	.051	.186	.186
Model 3: Recessive	.20	.084	.084	.480

Note: p is the population frequency of the disease-predisposing allele and f_i is the penetrance for the genotype with *i* copies of that allele. All models have disease prevalence 10% and single-locus sibling relative risk $\lambda_S = 1.2$.

Since 32 haplotypes were present in our simulated population, the correct degrees of freedom for the likelihood-ratio test were 31. For the situations we considered, the permutation test based on the log-likelihood statistic had power similar to the likelihood-ratio test using the correct degrees of freedom. Limited simulations suggested that this also was true for populations with fewer existent haplotypes (data not shown). Thus, in real applications in which we may not know the correct degrees of freedom because of the uncertainty of haplotype data, the permutation test based on the log-likelihood statistic may be a good alternative to the likelihood-ratio test.

For the situations we considered, the permutation test based on the high-vs-low statistic, which results from combining haplotypes according to over- and under-representation in the case group, had the highest power to detect disease-haplotype association among the tests we carried out. Limited simulations on populations with different LD patterns suggested that when offspring genotypes were available and the genotype missing rate was low, the power of the permutation test based on the high-vs-low statistic was often the highest, and when offspring genotypes were not available or the genotype missing rate was very high, the power of the permutation test based

TABLE 3. Estimated power (%) at significance level $\alpha = 1\%$ with random missing genotype rates 10% and 0%

	Additive model				Dominant model				Recessive model			
Missing genotype rate:	10%		0%		10%		0%		10%		0%	
Use offspring genotype:	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Likelihood-ratio test:	51	62	56	66	46	57	52	61	18	24	20	27
Permutation test:												
Log-likelihood	54	55	56	57	50	50	51	52	19	18	20	19
High vs. low	53	80	70	87	49	76	64	83	19	35	27	43
Best haplotype	20	43	33	56	17	38	29	53	7	15	11	23
Best two haplotypes	23	48	36	63	18	44	32	59	7	16	12	25

Note: Based on our simulated population and 2,500 replicate data sets of 400 affected-unaffected spouse pairs each with one offspring. The three models are described in Table 2.

on the log-likelihood statistic may be substantially higher (data not shown). Among the three statistics, the power of the best-haplotype and best-two-haplotype statistics was lower than that of the high-vs-low statistic, presumably due to being unable to combine signals effectively for the population used in our simulations, which had multiple disease-associated haplotypes. However, these two tests are potentially powerful for alternative population histories in which the disease variant is associated with a small number of haplotypes.

Offspring genotypes provide information about parental haplotypes. As a reference, we also estimated the power of the likelihood-ratio test under the ideal situation of having observed the haplotypes. For the situations we considered, the power under the ideal situation was very similar to that for the likelihood-ratio test with one offspring (data not shown), suggesting that even one offspring can provide useful information with respect to disease-haplotype association.

HAPLOTYPE RECONSTRUCTION

For each replicate data set, we also estimated haplotype reconstruction success rates for the two reconstruction methods we described, and calculated the average success rates over the 2,500 replicates we used in power estimation (Table 4). As expected, the reconstruction success rates were higher when using offspring genotypes than when not using them. For example, for the models we considered, with 10% missing genotypes, the success rate for reconstructing haplogenotypes with ≥99% posterior probability increased from 23.0% with no offspring to 42.9% with one offspring, 61.7% with two offspring, and 81.7% with four offspring, while the success rate for reconstructing haplogenotypes based on highest posterior probability increased from 78.1% with no offspring to 88.1% with one offspring, 93.2% with two offspring, and 97.4% with four offspring. Higher genotype missing rates had an adverse effect on reconstruction success rates, but could be compensated for by using more offspring, if available. Schaid [2002] and Becker and Knapp [2002] also observed the benefits of offspring genotypes on parental haplotype reconstruction.

TABLE 4. Average reconstruction success rates and number of consistent haplogenotypes at random genotype missing rates 10% and 0%

	Probability vote (%)		Level-99 (%)				Average number of consistent haplogenotypes	
Missing genotype rate:	Success		Success		Non-success			
	10%	0%	10%	0%	10%	0%	10%	0%
♯ offspring:								
0	78.1	88.4	23.0	37.9	0.4	0.4	10.2	5.2
1	88.1	97.8	42.9	84.0	0.3	0.2	4.2	2.0
2	93.2	98.8	61.7	91.6	0.2	0.1	2.5	1.5
3	95.9	99.2	73.7	95.4	0.1	<0.1	1.8	1.3
4	97.4	99.5	81.7	97.4	<0.1	<0.1	1.5	1.1

Note: Based on our simulated population and under Model 1 in Table 2. The numbers are averages over 2,500 replicates. For the probability-vote method, the non-success rate can be calculated as (1 − success rate). Results under other models are similar. Haplotypes are defined on five biallelic markers.

To illustrate the advantage of having offspring genotypes, we also calculated the average number of consistent haplogenotypes for a parent (Table 4). As the number of offspring increased, the set of consistent haplogenotypes for a parent became smaller and provided more certain inference. As a by-product, many fewer iterations of the EM algorithm were required for convergence, and computation time was greatly reduced.

DISCUSSION

In association studies for some late onset diseases, there are significant advantages in using unaffected spouses as controls. Because we often are unable to obtain genotypes for parents of the subjects, we cannot carry out the transmission/disequilibrium test [Spielman et al., 1993]. The alternative of using unaffected sibs as controls generally results in low power due to overmatching [Boehnke and Langefeld, 1998]. Thus, a well-designed case-control study is an attractive alternative. Spouse controls tend to be convenient to obtain and to be reasonably well matched environmentally and genetically to the cases, as long as gender does not play an important role in disease risk [Valle et al., 1998]. Spouses also may already have been obtained as part of an ongoing linkage study. For haplotype analysis, if offspring genotypes are available, we also are able to take advantage of spouse genotypes mutually for both members of a spouse pair.

In this paper, we described methods for efficient haplotype analysis for case-control data with some subjects forming spouse pairs and offspring genotypes available for some subjects. We proposed a likelihood-based method of haplotype inference, which works in a unified way with three types of subjects: subjects as couples with offspring genotype information available; subject-offspring pairs; and subjects as singletons without offspring information. Our method enables efficient haplotype frequency estimation using an EM algorithm and supports haplotype reconstruction with posterior probability calculated based on the whole sample. Similar methods have been used in the context of population haplotype frequency estimation [Boehnke, 1991; Rohde and Fuerst, 2001].

For parent-offspring pairs with the other parent's affection status unavailable, two approaches are available. One approach is to assume the affection status of the other parent was unaffected; this is acceptable for diseases with low prevalence because the probability of misclassification will be low. An alternative approach is to calculate population haplotype frequencies as weighted combinations of case and control frequencies using disease prevalence as the weight; this may be good for diseases with moderate prevalence but it relies on disease prevalence estimation of the population.

We described likelihood ratio and permutation tests to test for disease-haplotype association, and defined four statistics for the permutation test. Not surprisingly, simulations showed that all tests were more powerful when using offspring genotype information than when not using it. For the likelihood-ratio test, the correct degrees of freedom is the number of haplotypes truly present in the population less one. However, we generally do not know the number of existing haplotypes; using the maximum possible number of haplotypes as the degrees of freedom generally results in a conservative test, while using the "observed" number of haplotypes often leads to an anti-conservative test. In this situation, simulations suggested that the permutation test based on the log-likelihood statistic may be a good alternative.

For the situations we considered, the permutation test based on the high-vs-low statistic, which results from combining haplotypes according to over- and under-representation in the case group, had the highest power to detect disease-haplotype association. This is presumably because this statistic is best suited for our simulated population, which has multiple disease-associated haplotypes. In populations in which there are only one or two disease-associated haplotypes, we expect the best-haplotype or best-two-haplotype statistics also will be powerful. In some other situations, especially in the absence of offspring genotypes or if the genotype missing rate is high, these three statistics may have lower power than the log-likelihood because the performance of these statistics probably relies more strongly on the accuracy of haplotype frequency estimates. The performance of our tests may also depend on the LD between a disease-predisposing variant and the markers of haplotypes, and on the LD among the markers for the population under study. Populations with different LD patterns might yield different results.

In the permutation tests, we permuted affection status across all subjects. If all subjects are case-spouse control pairs, then permuting affection status within spouse pair may be an alternative to our permutation procedure. These two permutation procedures will lead to different but highly correlated results. However, often we will have some cases without family information and an additional set of controls ascertained independently of the cases. Permuting within families will make these subjects non-informative for our analysis. Thus, we chose to permute across all subjects as we would do for a case-control study.

Haplotypes often cannot be inferred with certainty. Missing genotypes in the data introduce additional uncertainty. For late onset diseases, genotypes from spouses and offspring can provide useful information about the haplotypes of the subjects. Besides helping increase the power to detect disease-haplotype association, offspring genotypes can significantly improve haplotype reconstruction success rates. If per-genotype efficiency is the goal, however, genotyping offspring may not be justified [Becker and Knapp, 2002]. But genotypes of offspring often are available due to the need for other analyses such as studying association between candidate genes and disease-related phenotypes in the offspring of cases as a risk population. In this situation, incorporating available offspring genotypes into the disease-haplotype association analysis becomes desirable, and we proposed an efficient method towards this goal.

The population we used in our simulation study was artificial, although it was generated to mimic approximately the history of an admixed and then isolated population. We used this simulated population to demonstrate the advantage of using offspring genotype information in detecting disease-haplotype association and in reconstructing haplotypes. To assess the generality of our results, we also carried out limited simulations using populations with other LD patterns. The advantage of using offspring genotypes was also demonstrated in those populations, although the relative merits of the test statistics were different.

Finally, we note that our method can be extended to more general situations. Although we presented the method for nuclear families and singletons, in principle, it can be extended to any pedigree structures as long as we know the pedigree founders' affection status and choose not to use the affection status information on the non-founders. We focused on haplotypes defined on a dense set of markers among which recombination is rare, and therefore ignored recombination. If one wishes to analyze a larger chromosomal region and allow for recombination, the method can in principle be modified to achieve this goal. In this situation, recombination fractions come into play only through the conditional probabilities, which can be calculated and stored before the EM algorithm is carried out.

Haplotype analysis of late onset diseases is often more difficult than that of early onset diseases due to lack of parental genotype information. When available, offspring genotypes can offer information on the haplotypes of the subjects. However, mixed pedigree structures often coexist in a data set, with offspring information available for only a subset of subjects and the number of offspring varying across families. In this paper, we described a unified method to analyze such mixed types of pedigree structures appropriately, leading to more efficient use of available data and more power to detect disease-haplotype association. We hope this method will be useful for disease gene discovery for late onset diseases. Software for simulation and data analysis is available upon request from Chun Li, the first author.

ACKNOWLEDGMENTS

We gratefully acknowledge support from a startup fund and a Pilot and Feasibility Grant from the Vanderbilt Diabetes Center (to C.L.) and NIH Grant HG00376 (to M.B.). We thank Ethan Lange for providing source code for generating population haplotypes and William Duren for providing help in programming. We thank Laura Scott and two anonymous reviewers for their valuable comments.

REFERENCES

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

Boehnke M. 1991. Allele frequency estimation from data on relatives. Am J Hum Genet 48:22–25.

Boehnke M, Langefeld CD. 1998. Genetic association mapping based on discordant sib pairs: the discordant-alleles test. Am J Hum Genet 62:950–961.

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

Devlin B, Risch N. 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311–322.

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

Lange EM, Boehnke M. 2004. The haplotype runs test: the parent-parent-affected offspring triodesign. Genet Epidemiol 27: 118–130.

Long JC, Williams RC, Urbanek M. 1995. An E-M algorithm and testing strategy for multiplelocus haplotypes. Am J Hum Genet 56:799–810.

Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. Science 273:1516–1517.

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

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

Spielman RS, McGinnis RE, Ewens WJ. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52: 506–516.

Valle T, Tuomilehto J, Bergman RN, Ghosh S, Hauser ER, Eriksson J, Nylund SJ, Kohtamäki K, Tuomilehto-Wolf E, Toivanen L, Vidgren G, Ehnholm C, Blaschak J, Langefeld CD, Watanabe RM, Magnuson V, Ally DS, Hagopian WA, Ross E, Buchanan TA, Collins F, Boehnke M. 1998. Mapping genes for non-insulin dependent diabetes mellitus: design of the Finland-United States Investigation of NIDDM Genetics (FUSION) study. Diab Care 21:949–958.