On Haplotype Reconstruction for Diploid Populations *

Jian Zhang
EURANDOM, Den Dolech 2
5612 AZ, Eindhoven, and
Computational Molecular Biology
Max-Planck-Institute for Molecular
Genetics, Ihnestrasse 73
D-14195 Berlin (Dahlem)

Martin Vingron
Computational Molecular Biology
Max-Planck-Institute for Molecular
Genetics, Ihnestrasse 73
D-14195 Berlin (Dahlem)

Margret R. Hoehe
Genome Research
Max-Delbrück-Center for Molecular Medicine
Robert-Rössle 10, D-13092 Berlin

June 18, 2001

Address for correspondence: Dr. Martin Vingron, Computational Molecular Biology, Max-Planck-Institute for Molecular Genetics, Ihnestrasse 73, D-14195 Berlin (Dahlem). E-mail: vingron@molgen.mpg.de; Phone: +49-30-8413-1150; Fax: +49-30-8413-1152.

Abstract

The problem of inferring haplotype pairs directly from the unphased genotype data is crucial in a number of haplotype analyses such as association study and linkage disequilibrium mapping for diploid populations. In literature there are mainly two popular approaches to this problem. One is what called maximum resolution approach and another is the EM algorithm based on the genotype likelihood. In this paper we intend to improve these two different approaches by introducing a general complete-data-likelihood framework. Based on this framework, we develop two kinds of estimators, namely, the maximum profile-likelihood (MPL) and Bayesian estimators. We demonstrate that under certain conditions, the MPL estimator will attain a maximum resolution. This implies under certain condition, when the maximum resolution is unique, the maximum resolution estimator is also a MPL estimator. On the other hand, the genotype likelihood based EM estimator is simply the maximum marginal-likelihood estimator of haplotype frequencies. As an alternative to the coalescent model based estimator (Stephens, Smith and Donnelly, 2001), we introduce the minimum evolution estimator. We solve the optimization problems arising from our procedures by using the evolutionary Monte Carlo (EMC) algorithm, a recent developed Markov chain Monte Carlo algorithm. Our approaches are tested on some real and simulated data sets. Interestingly, for the African-American substance-dependent data set, using our procedures we obtain almost the same substance-dependent haplotype group as that in Hoehe et al.(2000) although these procedures may give different haplotype assignments. Overall our procedures have the following advantages over the existing estimators: (1) As an improvement on the maximum resolution algorithm, our procedures can take into account the count and structual information of the genotypes; (2) unlike the EM algorithm, our procedures allow one to handle the genotypes with a large number of amibiguous loci by directly estimating the values of ambiguous loci rather than the frequencies of all possible haplotypes; (3) our minimum evolution estimator allows one to use the phylogenetic information from the data without assuming a coalescent model.

Key words: Haplotype reconstruction, profile likelihood, genotypes, multiple SNPs, and Markov chain Monte Carlo.

Running head: Haplotype reconstruction

1 Introduction

An entire human genomic reference sequence has been released recently. It is now known that individual humans differ from one another by about one base pair per thousand. These differences, called single nucleotide polymorphisms (SNPs), are believed to closely link to the human genetic bases. Recently, more than 1.4 million SNPs have been mapped in the human reference sequence. To turn this new genomic information into an engine of pharmaceutical discovery, one of the next genomics efforts will focus on exploring and using DNA sequence (or genetic) variations relative to this genomic sequence, among individuals and between human populations (see Baltimore, 2001; The Genome International Sequencing Consortium, 2001). The particular combination of such variants presented in some defined regions or sites is often described by a haplotype. Haplotype based genome-wide association and linkage studies are becoming increasingly important in understanding the history and organization of the human genome, as well as in uncovering and analyzing candidate genomic regions related to complex disease. See Clark et al. (1998); Nickerson et al.(1998); Kruglyak (1999); McPeek and Strahs (1999); Service et al. (1999); Bonnen et al. (2000); Davidson, 2000; Hoehe et al. (2000); Lam et al. (2000); Liu et al. (2000); Templeton et al. (2000); Fallin et al. (2001), Pritchard (2001), Rannala and Reeve (2001), and among others. In particular, a recent study showed that some complex diseases are correlated to grouping and interaction of several SNPs, rather than any individual SNP (Drysdale et al., 2000). This is because even if each SNP may only contribute a small amount to the phenotypes in these diseases, the joint contribution of multiple SNPs can be still significant. A strategy for searching for genetic variants of small effect is therefore essential, and association studies are seen to address this need.

In diploid organism there are two haplotypes for each individual at the region of interest, because there are two (not completely identical) copies of each chromosome. For a large scale screen of dipliod populations, it is not feasible to examine two copies of this region separately to obtain the haplotype pair. So often only certain mixtures of these haplotype pairs, called unphased genotypes, are available. This phase information can be established by genotyping the family members of each individual. But for many cases, these family members are simply not available or not enough to solve the ambiguity completely (see, e.g., Hodge, Boehnke and Spence, 1999). There are also other experimental methods such as long-range allele-specific PCR (Clark et al., 1998). Unfortunately these methods are often cost-prohibitive for a large scale screen of populations. To overcome these difficulties, in the last decade, several methods have been proposed for the inference of haplotypes from unphased genotype data. Two of them are now popular. One is now called the maximum resolution approach (Clark, 1990; Gusfield, 2000). The other is the EM approach based on the likelihood of unphased genotype data (Excoffier and Slatkin, 1995; Hawley and Kidd, 1995; Long et al., 1995; Hoehe et al., 2000). The first approach is motivated by the fact that, for any haplotype that is common enough that homozygotes can be found in the sample, the sample is expected to have several heterozygotes bearing one copy of that haplotype (Clark et al., 1998). This approach, restricted to haplotypes of biallelic sites, focus on the estimation of haplotype pairs instead of haplotype frequencies. It has not taken full advantage of the count information (i.e. the information from the repeated observations) on some genotypes. In contrast, the second approach has a clear statistical background and is not restricted to biallelic allels. It first uses the EM algorithm to estimate haplotype frequencies, then assigns the individual a haplotype pair with the highest frequency among all the possible pairs consistent with the genotype. Although there are several Monte Carlo studies to indicate that both methods are promising (Clark, 1990; Fallin and Schork, 2000), Stephens, Smith and Donnelly (2001) showed that the EM can outperform the maximum resolution method under a coalescent model. It is not surprising because the latter has ignored the count information. For the African-American substance-dependent data (Hoehe et al., 2000), we found the EM and maximum resolution methods may give some different haplotype assignments. Naturally we ask whether this affects the conclusion of the followed haplotype based statistical analysis.

In an effort to address these issues, Stephens, Smith and Donnelly (2001) presented a new statistical method by assuming a coalescent model for the underlying haplotypes. Here, in the same spirit and under the minimal model assumption of Hardy-Weinberg equilibrium, we try to improve the maximum resolution method using the count information of some genotypes. For this purpose, we introduce a joint likelihood for the phases of genotypes as well as for the haplotype frequencies. We point out that both the maximum resolution and EM approaches have close connections to this new framework. Relying on this framework, two alternative haplotype estimation procedures are presented in this paper. The first one is on the basis of a profile likelihood derived from the above joint likelihood. This profile likelihood usually assigns the relatively higher likelihood to the haplotype which appear in both homozygotes and heterozygotes. Unlike the EM approach, in our procedure we give our attention directly to the haplotype assignment rather than to the haplotype frequencies. We show that under certain condition, the maximum resolution estimator can be simply derived from our procedure. However, the ideas of the maximum resolution and the profile likelihood are essentially different. The profile likelihood intends to identify as many older haplotypes hidden in genotypes as possible, whereas the maximum resolution trys to find the maximum number of descendants from the initial known haplotyes. To introduce our second procedure, we develop a Bayesian model. In this framework, the EM haplotype frequency estimator is simply the mode of the marginal posterior distribution of all possible haplotype frequencies. In the same spirit, in our second procedure we estimate the genotype phases by the mode of their marginal joint posterior distribution. We also present a method to incorporate prior information into the MPL estimators. Finally, as an alternative to the coalescent method of Stephens, Smith and Donnelly (2001), we introduce a minimum evolution procedure which is based on an implicit assumption about evolution, namely that evolutionary change is rare. In this procedure, we intend to find a haplotype assignment which has a minimum length of its phylogenetic tree.

The optimization problems arising from our procedures are very hard. We solve these computational problems using the evolutionary Monte Carlo (EMC) algorithm, a recent developed Markov chain Monte Carlo algorithm. Our procedures are tested on some real and simulated data sets. We conclude from these studies that our procedures have the following advantages over the existing ones: (1) As an improvement on the maximum resolution algorithm, our procedures can take into account the count and structural information of the genotypes; (2) unlike the EM, our procedures allow one to handle the genotypes with a large number of amibiguous loci by directly estimating the values of ambiguous loci rather than the frequencies of all possible haplotypes; (3) in addition, the minimum evolution approach allows one to use the phylogenetic information without assuming a coalescent model.

2 Haplotype reconstruction

2.1 Notations

Suppose that we are interested in a region of u_0 sites specified by a reference allelic vector $(r_1, r_2, \cdots, r_{u_0})^T$. Each allele r_u is a marker (with a single or several DNA base pairs) in the reference genomic sequence. For example, for the SNP case, r_u is a single DNA base pair. For diploid populations, each individual has two copies of this region. These copies have their own allelic vectors, say, $H^{(v)} = (h_1^{(v)}, \cdots, h_{u_0}^{(v)})^T$, $v = 1, 2$. We call these vectors the genetic haplotype vectors. For simplicity, we assume these sites are all biallelic. Then, given the reference haplotype vector, the genetic haplotype vectors of each individual can be simply expressed as two zero-one vectors. In these vectors, a site has 0 and 1 according to whether its genetic haplotype is identical with or different from the reference. Hereafter we call these zero-one vectors the haplotype vectors or simply haplotypes. For the ease of notation, we use the same symbols $H^{(v)}$ to represent these new vectors. As we pointed out before, in practice, for each individual, we may observe only the mixture of its genetic haplotype pair, namely, the genetic genotype vector $G = (g_1, g_2, \cdots, g_m)^T$, in the sense that we don't know which copy of the region each g_i comes from. Similarly, given the reference haplotype vector, each genotype vector can be transformed into a vector with components $0, 1$ and 2. Here, a site has 0, 1 and 2, according to whether on this site the haplotype pair is homozygous and identical with the reference, or homozygous but different from the reference, or heterozygous. Here, we call this new vector the genotype vector (or simply genotype) and denote it by the same symbol G. A genotype vector with h heterozygous sites has 2^{h-1} possible decompositions into pairs of haplotype vectors, say $H_{G,j} = (H_{G,j}^{(1)}, H_{G,j}^{(2)})^T$, $j = 1, \cdots, 2^{h-1}$. In particular, a genotype vector G with at most one heterozygous site can be decoded directly. Let a denote a 2^{h-1}-dimentional vector taking values in $\{e_j : j = 1, \cdots, 2^{h-1}\}$, where e_j is the 2^{h-1}-dimensional unit vector with all components being 0 except the j-th. Let $a = e_j$ if we assign the haplotype pair $H_{G,j}$ to G. Sometimes we denote e_j simply by j. We call a the phase (parameter) of G. Note that the underlying phase is often unknown.

Suppose that we have an independent genotype sample $\mathbf{G} = (G_1, \cdots, G_n)$ of size n. Assume that G_i has c_i possible haplotype decompositions, namely, the set $\mathcal{H}_i = \{(H_{ij}^{(1)}, H_{ij}^{(2)}) : j = 1, \cdots, c_i\}$, and the phase a_i. Suppose that $\{H_{ij}^{(v)} : v = 1, 2; j = 1, \cdots, c_i; i = 1, \cdots, n\}$ has only k_0 different haplotype vectors, say, $\{H_{0k} : k = 1, \cdots, k_0\}$. Then, given $\mathbf{G}$, under the assumption of Hardy-Weinberg equilibrium, we have the "complete-data likelihood"

$$L(\mathbf{G}|\mathbf{p}, \mathbf{a}) = \prod_{i=1}^{n} \prod_{j=1}^{c_i} p_{ij}^{a_{ij}}$$

where $\mathbf{a} = (a_1, \cdots, a_n)^T$, $\mathbf{p} = (p_1, p_2, \cdots, p_{k_0})^T$ is the population frequency vector of $\{H_{0k} : k = 1, \cdots, k_0\}$; and $p_{ij} = p_k^2$ if $H_{ij}^{(1)} = H_{ij}^{(2)} = H_{0k}$, and $p_{ij} = 2p_k p_l$ if $H_{ij}^{(1)} = H_{0k}$, $H_{ij}^{(2)} = H_{0l}$ and $k \neq l$. A simple calculation shows that it is proportional to the product

$$\prod_{k=1}^{k_0} p_k^{s_k}$$

with

$$s_k = \sum_{i=1}^{n} \sum_{j=1}^{c_i} \sum_{v=1}^{2} a_{ij} \delta(H_{ij}^{(v)} = H_{0k})$$

where $\delta(\cdot)$ is the indicator of a set. Obviously, $\sum_{k=1}^{k_0} s_k = 2n$. Furthermore, we have the marginal likelihood of $\mathbf{p}$

$$L(\mathbf{G}|\mathbf{p}) = \prod_{i=1}^{n} \{ \sum_{(H_1,H_2)\in\mathcal{H}_i} p_{H_1} p_{H_2} \} \tag{2.1}$$

where p_{H_1} and p_{H_2} are the population frequencies of H_1 and H_2, respectively.

2.2 EM approach

The EM approach is an algorithm of attempting to find $\mathbf{p}$ that maximizes the marginal likelihood (2.1). It reconstructs haplotype pairs by choosing the most probable ones, given the genotype data and the estimated population haplotype frequencies $\hat{\mathbf{p}}$. Here, we often assume that individuals with the same genotype also have the same underlying haplotype pairs (Excoffier and Slatkin, 1995).

As pointed out in Stephen, Smith and Donnelly (2001), to implement the EM algorithm, we need to store the haplotype frequency variables for every possible haplotype in the sample. A genotype with k ambiguous loci has 2^{k-1} possible decompositions. This implies that the storage requirement increases exponentially with the number of ambiguous loci. We must impose the limit on the number of ambiguous loci in practice. For example, K. Rohde and R. Fürst in their webservice (http://made.bioinf.mdc-berlin.de) imposed limits of 30 and 10 on the numbers of loci and ambiguous loci in each genotype, respectively. The other shortcoming in this algorithm is that the resulting estimator is a local minimum which may strongly depend on the starting point.

2.3 Maximum resolution approach

Clark (1990) proposed an algorithm for the haplotype assignment, which includes two steps: First form the initial set of the haplotypes from the "self-resolved" genotypes (i.e., those genotypes with at most one heterozygous position). Then, choose a known haplotype at time to see whether any of the unresolved genotypes is a mixture of a known haplotype with a complementary haplotype, and if it is the case, update the known haplotype set by adding in the complementary haplotype. Repeat this procedure until all the unresolved genotypes are resloved or no further genotypes can be resolved. Obviously the solution depends on the order in which the known haplotypes are chosen in the second step. The larger the number of resolved ambiguous genotypes (i.e. the resolution), the better the solution is. Gusfield (2000) presented a maximum resolution algorithm based on linear programming.

To address the possible way for improving this method, we consider the following examples.

Example 2.1 *Suppose that we have four different genotypes* $000, 001, 022$ *and* 220, *two of which are ambiguous.*

Obviously, there are two solutions which attain the maximum resolution: one is $\{(000,000),(001,001),(011,000),(110,000)\}$; *the other is* $\{(000,000),(001,001),(010,001),(100,010)\}$. *When the optimal solution is not unique, which one is better?*

Example 2.2 *Suppose that we have four different genotypes* $0001, 1001, 2201,$ *and* 1122 *with counts (i.e., multiplicities)* 1, $n_2(\geq 1)$, 1 *and* 1, *respectively.*

Using the maximum resolution approach, we obtain a unique optimal solution $\{(0001, 0001), (1001, 1001), (1101, 0001), (1101, 1110)\}$ *in which* 1001 *has no heterozygous descendant. Here we first take* $\{0001, 1001\}$ *as the initial known haplotype set, from which we choose* 0001 *for resolving* 2201*. Then update the known haplotype set by adding in the complementary haplotype* 1101*. Finally* 1122 *is resolved by* 1101*. Note that the count information is completely ignored in this algorithm. This is in contradiction with the rationale of the approach mentioned in the Introduction. For example, set* $n_2 = 10$*, then the haplotype* 1001 *already has a much higher frequency than the other known haplotype in the initial haplotype set,* 0001*. According to the coalescent theory in population genetics, the expected rank of a haplotype by age is the same as the rank by its frequency, and older haplotypes will tend to have more mutational connections than younger ones (see, e.g., Posada and Crandall, 2001). This implies that* 2201 *is more probably resolved by* 1001 *than by* 0001 *according to the rationale of the maximum resolution. However, choosing* 1001*, we fail to resolve all genotypes according to the Clark concept. Now the question here is of whether the maximum resolution concept is reasonable when the repeated observations are available.*

2.4 Profile likelihood approach

To address the issues arising from the previous subsections, in this subsection we first propose the maximum likelihood (ML) estimator $(\mathbf{p}, \mathbf{a})$ on the basis of the complete-data-likelihood. Then we show how the haplotype can be reconstructed from this ML estimator.

By definition, the maximum likelihood estimator of $(\mathbf{p}, \mathbf{a})$, namely, $(\hat{\mathbf{p}}, \hat{\mathbf{a}})$ is that value of $(\mathbf{p}, \mathbf{a})$ that maximizes $L(\mathbf{G}|\mathbf{p}, \mathbf{a})$. To simplify the computation, we further assume that $\mathbf{G}$ comprises m different genotypes with the counts $n_1, \cdots, n_m$, respectively. All possible haplotype pair decompositions of these different genotypes are denoted by $\tilde{H}_{uj}^{(v)}$, $v = 1, 2$, $j = 1, \cdots, \tilde{c}_u$, respectively for $u = 1, \cdots, m$. In Appendix (1) , we prove the following proposition.

Proposition 2.1 *The ML estimator assigns the same haplotype pairs to identical genotype observations.*

Intuitively, this is reasonable because we have only the genotype information and we are unable to distinguish the underlying haplotype pairs for two individuals of the same genotypes. This simple property substantially reduces the search space of $\mathbf{a}$ and leads to the following simple scheme to calculate the ML estimator.

First, let $\tilde{\mathbf{a}}$ be the vector of the phase parameters for m different genotype vectors. Define the profile log-likelihood of $\tilde{\mathbf{a}}$ by

$$l(\mathbf{G}|\tilde{\mathbf{a}}) \propto \sum_{k=1}^{k_0} \frac{\tilde{s}_k}{2n} \log \frac{\tilde{s}_k}{2n}$$

where

$$\tilde{s}_k = \tilde{s}_k(\tilde{\mathbf{a}}) = \sum_{u=1}^{m} \sum_{j=1}^{\tilde{c}_u} \sum_{v=1}^{2} n_u \tilde{a}_{uj} \delta(\tilde{H}_{uj}^{(v)} = H_{0k}) \tag{2.2}$$

is a linear function of $\tilde{\mathbf{a}}$. Obviously, $l(\mathbf{G}|\tilde{\mathbf{a}})$ is covex in $\tilde{\mathbf{a}}$.

Let $\hat{\tilde{\mathbf{a}}}$ be that value of $\tilde{\mathbf{a}}$ that maximizes the foregoing profile likelihood. Then the ML estimator $\hat{\mathbf{a}}$ can be expressed as $\hat{a}_i = \hat{\tilde{a}}_u$, for $\sum_{t=0}^{u-1} n_t < i \leq \sum_{t=0}^{u} n_t$, $u = 1, \cdots, m$. Moreover, for $k = 1, \cdots, k_0$, substituting $\hat{\tilde{a}}_u$, $u = 1, \cdots, m$, into $\tilde{s}_k$, we have $\hat{\tilde{s}}_k$ and $\hat{p}_k = \hat{\tilde{s}}_k / 2n$.

Now the haplotype pairs can be reconstructed from the above phase estimator as follows. Given m haplotype sets $\{(\tilde{H}_{uj}^{(1)}, \tilde{H}_{uj}^{(2)}) : j = 1, \cdots, \tilde{c}_u\}$, $u = 1, \cdots, m$, for $(j_1, \cdots, j_m)$ (called an assignment), we calculate $\tilde{s}_k$ through counting the times that H_{0k} appears in the set of $\tilde{H}_{uj_u}^{(1)}, \tilde{H}_{uj_u}^{(2)}$, $u = 1, \cdots, m$, with each pair $(\tilde{H}_{uj_u}^{(1)}, \tilde{H}_{uj_u}^{(2)})$ being repeated n_u times. This is similar to the what called gene-counting method (Excoffier and Slatkin, 1995). Then we calculate the profile log-likelihood for assignment $(j_1, \cdots, j_m)$

$$l(\mathbf{G}|(j_1, \cdots, j_m)) = \sum_{k=1}^{k_0} \frac{\tilde{s}_k}{2n} \log \frac{\tilde{s}_k}{2n} \tag{2.3}$$

Note that $-l(\mathbf{G}|(j_1, \cdots, j_m))$ is just an entropy. We look for an optimal assignment in the sense that it maximizes the above profile log-likelihood or minimizes the above entropy. The corresponding haplotype estimator is called the maximum profile likelihood (MPL) estimator.

To justify the advantage of our procedure over the maximum resolution method, we apply our procedure to Examples 2.1 and 2.2.

Example 2.1 (continued). Suppose that these four genotypes have the counts $n_1(\geq 1)$, 1, 1, and 1, respectively. Let $p_1, \cdots, p_6$ be the population frequencies of $000, 001, 011, 010, 110$ and 100, respectively. Then for $n_1 = 1$ and 3, the MPL assignment $(1,1,1,1)$ (which stands for the set of haplotype pairs $\{(000,000), (001,001), (011,000), (110,000)\}$) is unique. For $n_1 = 1$, we have $\hat{p}_1 = 1/2$, $\hat{p}_2 = 1/4$, $\hat{p}_3 = 1/8$, $\hat{p}_4 = 0$, $\hat{p}_5 = 1/8$, and $\hat{p}_6 = 0$. For $n_1 = 3$, we have $\hat{p}_1 = 2/3$, $\hat{p}_2 = 1/6$, $\hat{p}_3 = 1/12$, $\hat{p}_4 = 0$, $\hat{p}_5 = 1/12$, and $\hat{p}_6 = 0$. The PL method can solve the problem mentioned in the last subsection. Furthermore the solutions are consistent with those derived from the EM algorithm of Hoehe et al. (2000).

Example 2.2 (continued). Similar to Example 2.1, if $n_2 = 1$, the MPL assignment is $(1,1,1,1)$ (i.e., $\{(0001,0001), (1001,1001), (1101,0001), (1101,1110)\}$), which also attains the maximum resolution. However, if $n_2 = 10$, then there are two MPL assignments, $(1,1,2,1)$ and $(1,1,2,2)$ (i.e., the sets $\{(0001,0001), (1001,1001), (1001,0101), (1111,1100)\}$ and $\{(0001,0001), (1001,1001), (1001,0101), (1110,1101)\}$). Both have not attained the maximum resolution. This is not surprising because the haplotype 1001 has a greater frequency than 0001. The solutions are also consistent with those from the EM algorithm.

From the above two examples, we see that the PL method intends to choose the haplotypes with higher frequecies to resolve the unsolved genotypes. This is consistent with the principle of population genetics. Therefore, the criterion of maximum resolution may be not reasonable when there exist multiple counts. Example 2.2 also indicates that if the MPL estimator is not a solution of the maximum resolution, the solution could be not unique.

To conclude this section, we show that there is a close connection between the maximum resolution method and our method if all genotypes have a single count. Note that Gusfield (2000) showed the computation of the maximum resolution estimator is NP hard. This together with the following proposition implies the computation of the MPL estimator is also NP hard. We begin with some notations. Following Gusfield (2000), we say that a haplotype H_1 can resolve a genotype G if there is another haplotype H_2 that can combine with H_1 to form G. H_1 and H_2 are called resolved haplotypes. For a genotype A, let $\mathbf{H}(A)$ denote the set of the haplotypes derived from all possible haplotype pairs of A. Suppose that we have a genotype sample, in which t_1 different genotypes are completely homozygous (i.e., no heterozygous site), t_2 different haplotypes have a single heterozygous site, and t_3 different haplotypes have 2 or more heterozygous sites. These three groups are denoted by sets $\mathbf{G}_1$, $\mathbf{G}_2$ and $\mathbf{G}_3$, respectively. As pointed out before, the haplotype pairs for the genotypes in the first two sets can be directly resolved. The sets of these haplotypes are denoted by $\mathbf{H}(\mathbf{G}_1)$ and $\mathbf{H}(\mathbf{G}_2)$, respectively. Let $\mathbf{H}_0$ denote the set of the current resolved haplotypes. Let $\mathbf{G}_0$ denote the current unresolved genotype vectors. Fill a prelimary list of haplotypes, for example, by setting $\mathbf{H}_0 = \mathbf{H}(\mathbf{G}_1 \cup \mathbf{G}_2)$ and $\mathbf{G}_0 = \mathbf{G}_3$. Then screen $\mathbf{G}_0$ for a $A \in \mathbf{G}_0$ which can be resolved by a $H_1 \in \mathbf{H}_0$. Let (H_1, H_2) be the corresponding haplotype pair assigned to A. Update $\mathbf{H}_0$ and $\mathbf{G}_0$ by adding H_2 to $\mathbf{H}_0$ and removing A from $\mathbf{G}_0$. Repeat this procedure until there is no resolvable genotype in $\mathbf{G}_0$ or $\mathbf{G}_0$ is empty. Apply the above procedure repeatedly until there is no resolvable genotype in $\mathbf{G}_0$. The final $\mathbf{G}_0$ is called the 'orphan' set. Clearly the final $\mathbf{G}_0$ depends on the order in which genotypes in $\mathbf{G}_2$ and $\mathbf{G}_3$ are called. The maximum resolution algorithm attempts to minimize the number of the remain 'orphans'.

We show the following proposition in Appendix 2.

Proposition 2.2 *Under the following conditions, attaining a maximum resolution is a necessary condition for an assignment being a MPL estimator.*

(C1) $\mathbf{G}_3$ can be resolved completely by applying the above inference procedure;

(C2) for any $A \in \mathbf{G}_1 \cup \mathbf{G}_2 \cup \mathbf{G}_3$, any haplotype from $\mathbf{H}(A)$ can only resolve one of the remaining genotypes in $\mathbf{G}_3$.

We use Example 2.2 to help readers to grasp the main idea of the above proposition. In Example 2.2, we have two different assignments, $(1,1,1,1)$ and $(1,1,2,1)$. The first one divides the assigned haplotypes into two groups, $\{(100,100)\}$ and $\{(000,000),(110,000),(110,111)\}$ with the profile log-likelihood $\{2/8\log(2/8)+3/8\log(3/8)+2/8\log(2/8)+1/8\log(1/8)\} = -1.3208$. The second one splits the assigned haplotypes into three groups, $\{(000,000)\}$, $\{(100,100),(100,010)\}$ and $\{(110,111)\}$ with the profile log-likelihood $\{2/8\log(2/8)+3/8\log(3/8)+3/8\log(1/8)\} = -1.494$. The second one has an 'orphan' group which decreases the profile-likelihood. In this example, the maximum resolution assignment is unique and thus is a MPL estimator. However, it is possible that a maximum resolution is not a MPL estimator. For instance, in Example 2.1, $(1,1,2,2)$ is a maximum resolution assignment, but not a MPL estimator. The condition **(C1)** can not be further relaxed in general. The examples in Subsection 3.2 will show that if the condition **(C1)** doesn't hold, the MPL and maximum resolution estimators can be completely different.

2.5 Bayesian approach

In light of (5.2), we start with the joint likelihood $\prod_{k=1}^{k_0} p_k^{\tilde{s}_k(\tilde{\mathbf{a}})}$ where $\tilde{s}_k(\tilde{\mathbf{a}})$ is defined in (2.2). We adopt $\prod_{k=1}^{k_0} p_k^{\alpha_k - 1}$ and $\prod_{u=1}^{m} 1/\tilde{c}_u$ as the priors for the haplotype frequencies and phase parameters respectively, where $\alpha_k > 0, 1 \leq k \leq k_0$ are some prespecified constants. Then the posterior distribution of $(\mathbf{p}, \mathbf{a})$ is

$$p(\mathbf{p}, \mathbf{a}|\mathbf{G}) = D_0^{-1} \prod_{k=1}^{k_0} p_k^{\tilde{s}_k + \alpha_k - 1} \tag{2.4}$$

where

$$D_0 = D_0(\mathbf{G}) = \sum_{u=1}^{m} \sum_{j_u=1}^{\tilde{c}_u} \frac{\prod_{k=1}^{k_0} \Gamma(\tilde{s}_k(\mathbf{a}(j_1, \cdots, j_m)) + \alpha_k)}{\Gamma(\sum_{k=1}^{k_0} (\tilde{s}_k(\mathbf{a}(j_1, \cdots, j_m)) + \alpha_k))}.$$

Here $\Gamma(\cdot)$ is a Gamma function, $\mathbf{a}(j_1, \cdots, j_m) = (a_1, \cdots, a_m)^T$ and a_u is a $\tilde{c}_u$-dimensional unit vector with all components being zero except the j_uth. The marginal posterior distribution of $\mathbf{p}$ is

$$D_0^{-1} \sum_{u=1}^{m} \sum_{j_u=1}^{\tilde{c}_u} \prod_{k=1}^{k_0} p_k^{\tilde{s}_k(\mathbf{a}(j_1, \cdots, j_m)) + \alpha_k - 1}$$

which, when all $\alpha_k = 1$, is proportinal to the genotype likelihood (see, e.g., Excoffier and Slatkin, 1995), on which the EM haplotype frequency estimator is based. This implies that this EM estimator is just the mode of the marginal posterior distribution of $\mathbf{p}$. Similarly, we have the following marginal posterior distribution for $\mathbf{a}$, namely,

$$\begin{aligned} p(\mathbf{a}|\mathbf{G}) &= \frac{1}{D_0}\frac{\prod_{k=1}^{k_0}\Gamma(\tilde{s}_k(\mathbf{a}(j_1,\cdots,j_m)) + \alpha_k)}{\Gamma(2n + \sum_{k=1}^{k_0}\alpha_k))} \\ &\approx \exp\left\{\sum_{k=1}^{k_0}(\tilde{s}_k(\mathbf{a}) + \alpha_k)\log\frac{\tilde{s}_k(\mathbf{a}) + \alpha_k}{2n + \sum_k \alpha_k}\right\}. \end{aligned} \tag{2.5}$$

Its mode, say $\mathbf{a}_p$, can be used as a phase estimator. In particular, we can write

$$\mathbf{a}_p = \text{argmax}_{\mathbf{a}(j_1,\cdots,j_m)}\frac{\prod_{k=1}^{k_0}\Gamma(\tilde{s}_k(\mathbf{a}(j_1,\cdots,j_m)) + \alpha_k)}{\Gamma(2n + \sum_{k=1}^{k_0}\alpha_k)}$$

The last term in (2.5) can be viewed as a way to incorporate the prior information into the PL estimator. For this purpose, we note that for each fixed $\mathbf{a}$, (2.4) gives the following Bayesian haplotype frequency estimator (i.e., the mean of the posterior distribution)

$$\left(\frac{\tilde{s}_1(\mathbf{a}) + \alpha_1}{2n + 2md}, \cdots, \frac{\tilde{s}_k(\mathbf{a}) + \alpha_{k_0}}{2n + 2md}\right)^T.$$

Substituting this estimator into (2.4) leads to the last term in (2.5). So we have the following modification of the profile likelihood (2.3),

$$\sum_{k=1}^{k_0}\frac{\tilde{s}_k(\mathbf{a}) + \alpha_k}{2n + 2md}\log\frac{\tilde{s}_k(\mathbf{a}) + \alpha_k}{2n + 2md}. \tag{2.6}$$

Now a new phase estimator can be simply produced by maximizing this new likelihood.

We specify the constants α_k, $k = 1, \cdots, k_0$, by the following way.

We first assign the same weight $2d$ to the m observed different genotypes (see Subsection 2.1), where d is constant with a default value 1. Assume that the m genotypes are equally important and that the phases of these genotypes are unknown. Then, the structures of these genotypes imply that the importance of each candidate haplotype in different genotypes should be different. So, for each of these genotype, we distribute the weight $2d$ equally to all its candidate haplotype vectors. Then for $k = 1, \cdots, k_0$, let α_k be the summation of all the weights we put on the haplotype H_{0k} which appears in n candidate haplotype sets, $\mathcal{H}_i$, $i = 1, \cdots, n$ mentioned in Subsection 2.1. These constants are called the pseudo-counts. For instance, for Example 2.1, letting $d = 1$, we have the pseudo-counts $(3, 5/2, 1/2, 1, 1/2, 1/2)$ for $000, 001, 011, 010, 110$ and 100, respectively.

Note that these pseudo-counts yield a prior haplotype frequency estimator, namely, $(\alpha_1/2md, \cdots, \alpha_{k_0}/2md)^T$. A prior phase estimator is obtained simply by assigning each individual a haplotype pair with the highest prior frequency among all the possible candidate haplotype pairs.

2.6 Minimum evolution approach

Note that very recently Stephens, Smith and Donnelly (2001) developed a parametric method for reconstructing haplotypes in that the underlying haplotypes are assumed to follow a coalescent model. Here we instead model the haplotype structure by a phylogenetic tree. We try to find a haplotype assignment with a minimum length of its phylogenetic tree.

For each feasible haplotype assignment, we build a phylogenetic tree for its associated haplotypes, termed a feasible tree. Then, we select an optimal tree according to its length (i.e., the sum of all branch lengths). To reduce the burden of computation, we adopt a modified UPGMA procedure to build the tree. Similarly, we can also use the neighbour-joining procedure. See Page and Holmes (1998) for the introduction of the UPGMA procedure. In theory, the other tree building methods can be applied here. But these methods are often extremely time demanding. In our procedure, we use a new distance for haplotypes, termed the information distance (Zhang and Vingron, 2001). More specifically, we first calculate the pairwise percentage identities for m different haplotypes, say q_{ij}, $1 \leq i, j \leq m$. Then we define the distance between the ith and jth haplotypes by the information distance between the probability vectors $(q_{i1}, \cdots, q_{im})^T$ and $(q_{j1}, \cdots, q_{jm})^T$, namely,

$$\begin{aligned} d(i,j) &= \sum_{k=1}^{m} q_i \log(q_i/q_{\{i,j\},k}) + (1-q_i)\log((1-q_i)/(1-q_{\{i,j\},k})) \\ &\quad + \sum_{k=1}^{m} q_j \log(q_j/q_{\{i,j\},k}) + (1-q_j)\log((1-q_j)/(1-q_{\{i,j\},k})) \end{aligned}$$

where $q_{\{i,j\},k}$ is the average of q_{ik} and q_{jk}. The advantage of this new distance over the traditional Hemming distance is when comparing two haplotypes, we use not only the similarity between them but also their similarities to the other haplotypes. This point has been justified by Zhang and Vingron (2001). In our procedure we also adopt a weighted arithmetic average distance for the cluster pair, C_i and C_j, namely

$$d(C_i, C_j) = \frac{1}{\sum_{k_1 \in C_i} s_{k_1} \sum_{k_2 \in C_j} s_{k_2}} \sum_{k_1 \in C_i, k_2 \in C_j} s_{k_1} s_{k_2} d(k_1, k_2)$$

by taking account of the multiple count information, where s_k is the count of the kth haplotype. It is easily seen in $d(C_i, C_j)$ that the higher the count of a haplotype, the larger weight this haplotype will have.

2.7 Computation

Note that according to (2.5), the Bayesian estimator can be approximated by the minimum of (2.6), which is very similar to (2.3). This means that we only need to solve the problem of how to calculate the MPL and minimum evolution estimators. The computation includes two steps. First, for a haplotype assignment we calculate the objective function (the profile likelihood in the MPL case and the length of the associated tree in the minimum evolution case). Then we optimize this function with respect to the assignment.

We begin with m different genotypes $G_1, \cdots, G_m$ with $v_1, \cdots, v_m$ ambiguous loci, repectively. As stated before, although the EM approach does use the count information, it is limited by the requirement of storing 2^{k-1} variables for each genotype with k ambiguous loci. In contrast, for the PL and minimum evolution approaches, we only need to optimize an objective function with $\sum_{i=1}^{m}(v_i - 1)$ variables. Of course, this advantage can not be fully taken if we for example use $l(\mathbf{G}|\mathbf{a})$ as the objective function directly. We need to reexpress $l(\mathbf{G}|\mathbf{a})$ as a function with the $\sum_{i=1}^{m}(v_i - 1)$ ambiguous loci as variables. For this purpose, let $\mathbf{z}$ denote a $\sum_{i=1}^{m}(v_i - 1)$ dimensional variable in which each component takes value of 0 or 1. From $\mathbf{z}$, we can construct the haplotype pair (H_{1i}, H_{2i}) for each G_i, $i = 1, \cdots, m$ as follows: All resolved positions of G_i are set the same in both H_{1i} and H_{2i}. The first ambiguous position of G_i is set 1 and 0 in H_{1i} and

H_{2i}, respectively. The remains of ambiguous positions of G_i are set $z_{k_i+1}, \cdots, z_{k_i+v_i-1}$, respectively in H_{1i}, where $k_i = \sum_{j=1}^{i-1}(v_j - 1)$. The ambiguous positions of G_i are set in H_{2i} to the opposite of the entry in H_{1i}. This implies that there is one-by-one correspondence between z and the assignment $(j_1, \cdots, j_m)$ in (2.3). Moreover, for each $\mathbf{z}$, we identify the different haplotypes and calculate their counts from these H_{1i}, H_{2i}, $i = 1, \cdots, m$. Then, the profile log-likelihood in (2.3) can be written in the form $l(\mathbf{G}|\mathbf{z})$, a function of $\mathbf{z}$. Now the optimal assignment can be obtained by maximizing $l(\mathbf{G}|\mathbf{z})$. Although, compared with the original optimization problem, the number of the operational variables are significantly reduced, the new one is still a very hard optimization problem in a high dimensional space. In particular, the new objective function $l(\mathbf{G}|\mathbf{z})$ with many subtle local maxima is no longer a convex function. Here, we apply a recently developed MCMC algorithm, called the evolutionary Monte Carlo (Liang and Wong, 2000), to solve the problem.

The evolutionary Monte Carlo algorithm works by simulating a population of Markov chains in parallel, where a different temperature is attached to each chain. The population is updated by mutation, crossover and exchange operators, and the updates are accepted or rejected according to the Metropolis rule. More specifically, given the current population $\mathbf{Z} = \{\mathbf{z}_1, \cdots, \mathbf{z}_N\}$ and a temperature ladder $\mathbf{t} = \{t_1, \cdots, t_N\}$, we construct a Boltzmann distribution for the population $\mathbf{Z}$ by

$$f(\mathbf{Z}) \propto \exp\{-\sum_{i=1}^{N} l(\mathbf{G}|\mathbf{z}_i)/t_i\}.$$

We sample the next population by the following two steps: (1). Apply the mutation or the crossover operator to $\mathbf{Z}$ with probability p_m and $1 - p_m$, respectively. (2). Exchange $\mathbf{z}_i$ with $\mathbf{z}_j$ for N pairs (i, j) with i being sampled uniformly on $\{1, \cdots, N\}$ and $j = i \pm 1$ with probability $w(\mathbf{z}_j|\mathbf{z}_i)$, where $w(\mathbf{z}_{i+1}|\mathbf{z}_i) = w(\mathbf{z}_{i-1}|\mathbf{z}_i)$ $= 0.5$ and $w(\mathbf{z}_2|\mathbf{z}_1) = w(\mathbf{z}_{N-1}|\mathbf{z}_N) = 1$. The details of these operators can be found in Appendix 3. In this paper, we set $t_i = t_h - (t_h - t_l)i/N$, $i = 1, \cdots, N$, where t_h and t_l are the highest and lowest temperatures, respectively. As in Liang and Wong (2000), we choose t_h and t_l such that $\mathrm{V}ar(l(\mathbf{G}|\mathbf{z}_i))(t_h - t_l)^2 = O(1)$ or simply by checking whether the overall acceptance rates of mutation, crossover and exchange operations are around 0.50.

3 Applications

3.1 Substance-dependent individuals and controls

To test a potential role of the human μ opioid receptor gene (OPRM1) in substance dependence, Hoehe et al.(2000) analyzed all known functionally relevant regions of this prime candidate gene by multiplex sequence comparison in 172 African-American cases and controls. They obtained 172 genotypes on 25 variable positions in this gene, in which using the EM method they predicted 52 different haplotypes. These haplotypes were classified by similarity clustering into two functionally related groups, one of which is associated with substance dependence. The haplotype reconstruction is crucial in this analysis. As pointed out in Example 2.2, the haplotype reconstructions based on the maximum resolution, EM or PL may be not unique. Naturally we are concerned with how large these different haplotype assignments affect on the conclusion made in Hoehe et al.(2000). In our opinion, this practical issue can not be fully answered by comparing the error rates of these methods for a genotype data set with known haplotypes which are simulated from a coalescent model. This is because we do not know the model which the real data really come from.

To this end, we first apply the PL approach to this data set. Note that the z for this data set is 196−variate. After several preliminary trials, we decide to choose the population size $N = 20$, the highest and lowest temperatures, $t_h = 0.02$ and $t_l = 0.001$, and the mutation and crossover parameters, $p_m = 0.2$, $p_0 = 0.004, p_1 = 0.008$ and $p_2 = 0.01$ in our algorithm. According to Liang and Wong (2000), the convergence of the Markov chains can be diagnosed using the Gelman-Rubin statistic (Gelman and Rubin, 1992). The first 1000 iterations are discarded as initial "burn-in" iterations. Then we make $M = 10^6$ iterations, a long run to get enough samples for inferring the maximum value of the profile log-likelihood $l(\mathbf{G}|\mathbf{z})$. Note that the overall acceptance rates of mutation, crossover and exchange operations are 0.584974, 0.365175, and 0.598087. In each iteration, we get a population in which the last sample is kept for the optimization task. This leads to $M = 10^6$ samples of $l(\mathbf{G}|\mathbf{z})$, say $l(\mathbf{G}|\mathbf{z}_1), \cdots, l(\mathbf{G}|\mathbf{z}_M)$. The EMC sampled maximum is defined as $\text{argmax}_{\mathbf{Z}_k} l(\mathbf{G}|\mathbf{z}_k)$.

Table 3.1 presents the haplotypes with counts reconstructed from the 172 genotypes. Figure 3.1 (a) and (b) show two hierarchical clusterings for these haplotypes using the information distance with and without weighting. Note that these results are independent of the initial haplotype population in the algorithm. Interestingly, comparing Table 3.1 and Figure 3.1 with Table 2 and Figure 4 in Hoehe et al. (2000), we can see that only 46 haplotypes are predicted here, 6 less than that of Hoehe et al. (2000). But surprisingly these haplotypes can still be classified into two groups: group one $\{27, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46\}$, group two made up by the remaining haplotypes. Group one is almost same as the substance-dependent group in Hoehe et al. (2000) in the following sense: both have 13 members, of which 11 haplotypes are common; both are significantly more frequent in substance-dependent individuals; and both have the same characteristic pattern of sequence variants. Note that the weighting in the information distance has some effects only on the tree structure of the second group (see the two trees in Figure 3.1).

It is easy to see that both our solution and that of Hoehe et al. (2000) attain the maximum resolution. In fact, we have found several other solutions which also attain the maximum resolution. But we do not know how many different solutions of the maximum resolution may exist. This makes the maximum resolution approach difficult to be used in this data set.

We then apply the Bayesian maximum posterior estimator defined by minimizing (2.6) to this data. The prior $\alpha_k, k = 1, \cdots, k_0$ are specified following the scheme in Subsection 2.5. Fortunately, the solution is the same as what is derived from the PL approach.

Finally, to employ the minimum evolution approach, we choose the population size $N = 20$, the highest and lowest temperatures, $t_h = 0.02$ and $t_l = 0.001$, and the mutation and crossover parameters, $p_m = 0.2$, $p_0 = 0.004$, $p_1 = 0.008$ and $p_2 = 0.01$ in our minimum evolution algorithm. We make $M = 8 \times 10^5$ iterations. The overall acceptance rates of mutation, crossover and exchange operations are 0.192804, 0.129297, and 0.591254. This results in 52 different haplotypes listed in Table 3.2. The dendrogram based our modified UPGMA clustering is presented in Figure 3.2. These haplotypes can be divided into two groups. Group one is dominated by the haplotypes of substance-dependent individuals in the sense that it is made up of three subgroups: (A) $\{32, 34, 37, 38, 46, 52\}$; (B) $\{22, 26, 33, 36, 42, 43, 48, 50\}$; and (C) $\{9, 18, 20, 21, 27, 28, 29, 31, 35, 44, 45, 47, 49, 51\}$. Subgroup (A) contains only haplotypes of substance-dependent individuals only. These six haplotypes feature the same constellation of five polymorphic sites as in Hoehe et al. (2000). With one exception Subgroup (B) contains only haplotypes of substance-dependent individuals. With four exception Subgroup (C) also contains only haplotypes of substance-dependent individuals. However, both Subgroups (B) and (C) present some different patterns from Subgroup (A). Group two corresponds to a mixture of cases and controls.

In summary, the above analyses convince us that a same pattern features with 5 conserved heterozygous polymorphic sites can be derived from the EM, PL, Bayesian, and minimum evolution approaches. The minimum evolution approach also suggests the other patterns.

3.2 Simulated data sets

In this subsection, we test our method on a randomly generated data of up to 50 genotypes, 10 initially resolved and 40 ambiguous, each containing 15 loci. Gusfield (2000) used this data set to test his maximum resolution algorithm under more extreme conditions than one would expect in realistic data. Here, different from him, we use this data set to show how large the difference between the maximum resolution and our proposal is under some extreme conditions. This implies that the more genotypes should be sampled when the genotypes can be not fully resolved completely by the maximum resolution.

We first assume that all the genotypes are single. Gusfield (2000) showed that only 26 genotypes can be resolved by the maximum resolution algorithm. In contrast, our procedure resolves all genotypes. The dimension of $\mathbf{z}$ in our algorithm is 268. We choose the population size $N = 20$, the highest and lowest temperatures, $t_h = 0.008$ and $t_l = 0.0004$, and the mutation and crossover parameters, $p_m = 0.2$, $p_0 = 0.001$, $p_1 = 0.002$, $p_2 = 0.004$ in our algorithm. As before, the first 1000 iterations are discarded as initial "burn-in" iterations. Then similar to the last subsection, we make $M = 3 \times 10^6$ iterations to get a simulated maximum. The overall acceptance rates of mutation, crossover and exchange operations are 0.606994, 0.53044, and 0.555151. 63 different haplotypes are predicted. Table 3.3 shows these haplotypes with their counts. The result is quite different from Gusfield (2000). This convinces us that our procedure and the maximum resolution algorithm could give totally different results when the genotypes can not be resolved completely according to the Clark's rule.

To see the influence of the mutiple counts on the haplotype reconstruction, we then consider the same genotypes as above but with multiple counts as shown in Table 3.3. To apply our algorithm, we choose the population size $N = 20$, the higest and lowest temperatures, $t_h = 0.003$, $t_l = 0.00015$, and the mutation and crossover parameters $p_m = 0.2$, $p_0 = 0.001$, $p_1 = 0.002$, $p_2 = 0.004$. As before, after 1000 initial "burn-in" iterations, we make $M = 3 \times 10^6$ iterations to get a simulated maximum. The overall acceptance rates of mutation, crossover and exchange operations are 0.430877, 0.501228, and 0.517761.

Table 3.5 shows the haplotypes with their counts reconstructed from the genotypes. 63 different haplotypes are predicted. The result is quite different from the case where all these genotypes are single. This implies the multiple count information does have a big effect on the haplotype reconstruction. Note the result is again different from the maximum resolution estimator.

In Tables 3.6 and 3.7 we show the difference between the PL and minimum evolution approaches by a simulated genotype data set. Table 3.6 shows this data set and the haplotypes derived by the minimum evolution approach. For comparison, in Table 3.7 we give the haplotypes derived from the same genotypes as in Table 3.6 relying on the PL approach. Note that the difference is obivious because Table 3.6 gives 29 different haplotypes while Table 3.7 presents only 20 different haplotypes. Note that for this data set we set the highest and lowest temperatures, $t_h = 0.08$ and $t_l = 0.004$; and $p_m = 0.2$, $p_0 = 0.002$, $p_1 = 0.002$, $p_2 = 0.008$ in our algorithm.

4 Discussions

Reconstructing haplotypes from genotypes is a basic step in any large scale screens of human populations for significant DNA polymorphisms. We have introduced three new statistical approaches for haplotype reconstruction, termed the profile likelihood, Bayesian and minimum evolution approaches. Like the maximum resolution and EM methods, these new approaches are nonparametric in that we have used the minimum model assumption of Hardy-Weinberg equilibrium of the underlying haplotype populations. Our procedures have two advantages over the maximum resolution or EM methods: (1) Unlike the maximum resolution approach, we take into account both the multiple count information and certain structual information of genotypes via the complete-data likelihood, the prior specification and the mimimum evolution principle; (2) in contrast to the EM method, our procedures can reconstruct haplotypes from heavily ambiguous genotypes with the help of the evolutionary Monte Carlo algorithm. Unlike the EM algorithm, our evolutionary Monte Carlo algorithm intend to find a global maximum. Of course, we should expect some similarity between the profile likelihood and EM methods because both methods stem from the same complete-data likelihood. We also find the following unexpected connection between the profile likelihood and the maximum resolution methods. Under a certain condition, the profile likelihood estimator is consistent with that derived from the maximum resolution method. According to our limited experiences, this could be true even under the more general condition where all the genotypes can be completely resolved and the solution of the maximum resolution is unique. As shown in Figure 2 of Stephen, Smith and Donnelly (2001), in reality, the above condition may be not true and then we have a difficulty getting the the maximum resolution algorithm to consistently provide a unique solution.

The time for calculating the PL and Bayesian estimators is varied from several minutes to several hours depending on the complexity of the data. However, the calculation of the minimum evolution estimator is very time-demanding. It may need several days if you want to get a good solution. So if the data under investigation come from a coalescent model approximately, it is better to use the coalescent approach of Stephens, Smith and Donnelly (2000).

In literature, there simply do not exist enough real data sets, with known haplotypes, to allow sensible statistical comparisons of different methods (Stephens, Smith and Donnelly, 2000). When we tackle the real data, it seems better to apply all current available methods to avoid the possible bias from any single method.

We have applied our procedures to both some real and simulated data sets. For the genotypes of African-American substance-dependent individuals and controls, via the PL and Bayesian approaches, we obtained 46 haplotypes, 6 fewer than those obtained by the EM method (Hoehe, et al., 2000). Fortunately, these haplotypes can be classified into two groups, one of which is mainly from substance dependent individuals and is almost the same as that obtained in Hoehe et al. (2000). A similar result is derived by the minimum evolution approach. Thus, our analysis further support the main result of Hoehe et al. (2000).

To conclude this section, we note that although in this paper we discuss only the haplotypes of biallelic loci, both our methods and algorithm can be easily extended to cope with other types of loci like microsatellite loci. It is also straightforward to modify our methods to allow them to deal with missing genotype data in some individuals at some loci.

5 Appendices

(1). **Proof of Proposition 2.1.** Note that, given $\mathbf{p}$, $L(\mathbf{G}|\mathbf{p},\mathbf{a})$ attains the maximum $L(\mathbf{G}|\mathbf{p},\mathbf{a}(\mathbf{p}))$ at $\mathbf{a}=\mathbf{a}(\mathbf{p})$ with

$$a_i = a_u^* = e_{j_u}^{(u)} \tag{5.1}$$

for $\sum_{t=0}^{u-1} n_t < i \le \sum_{t=0}^{u} n_t$, $1 \le u \le m$, where $n_0 = 0$, $\quad j_i = \text{argmax}_j p_{ij}$. (5.1) means that given $\mathbf{p}$, we should assign the same haplotype pairs to identical genotype observations. This implies that $L(\mathbf{B}|\mathbf{p},\mathbf{a}(\mathbf{p}))$ is proportional to $\prod_{k=1}^{k_0} p_k^{s_k^*}$, which is less than

$$\prod_{k=1}^{k_0} \left(\frac{s_k^*}{2n}\right)^{s_k^*}$$

where

$$s_k^* = \sum_{u=1}^{m}\sum_{j=1}^{\tilde{c}_u}\sum_{v=1}^{2} n_u a_{uj}^* \delta(\tilde{H}_{uj}^{(v)} = H_{0k}).$$

Thus, for a positive constant d_0,

$$\begin{aligned}
\max_{\mathbf{p},\mathbf{a}} L(\mathbf{G}|\mathbf{p},\mathbf{a}) &\le \max_{\mathbf{p}}\max_{\mathbf{a}} L(\mathbf{G}|\mathbf{p},\mathbf{a}) \\
&\le d_0 \max_{\tilde{a}_1,\cdots,\tilde{a}_m} \prod_{k=1}^{k_0} \left(\frac{\tilde{s}_k}{2n}\right)^{\tilde{s}_k}
\end{aligned} \tag{5.2}$$

where $\tilde{s}_k$ is defined by replacing a_{uj}^* by $\tilde{a}_{uj}$, and $\tilde{a}_1,\cdots,\tilde{a}_m$ are the phase parameters for the m different genotype vectors. This completes the proof.

(2). **Proof of Proposition 2.2.** It is obvious that any scheme of haplotype pair assignment can only divides the assigned haplotypes of $\mathbf{G}_1 \cup \mathbf{G}_2 \cup \mathbf{G}_3$ into the four kinds of possible groups, say, $\mathbf{F}_1$, $\mathbf{F}_2$, $\mathbf{F}_3$ and $\mathbf{F}_4$. There is no common haplotype between any two of these groups. See the example presented at the end of Subsection 2.4. $\mathbf{F}_1$ has a single element from $\mathbf{G}_1$; $\mathbf{F}_2$ has three elements, among which two are from $\mathbf{G}_1$ and the remaining is from $\mathbf{G}_2$; $\mathbf{F}_3$ of size f_3 includes one element from $\mathbf{G}_1$ and the others from $\mathbf{G}_3$; and $\mathbf{F}_4$ of size f_4 is a subset of $\mathbf{G}_3$. Suppose for each $k = 1,2,3,4$, there are d_k $\mathbf{F}_k$-type groups. $\mathbf{F}_4$ is an unresolved (orphan) group. Then the corresponding profile log-likelihood is equal to

$$\begin{aligned}
& \{\frac{2d_1}{2n}\log\frac{2}{2n} + \frac{2d_2}{2n}\log\frac{3}{2n} \\
+ & \frac{d_3}{2n}\log\frac{3}{2n} + \frac{d_3(f_3-1)}{2n}\log\frac{2}{2n} \\
+ & \frac{d_3}{2n}\log\frac{1}{2n} + \frac{d_4(f_4-1)}{2n}\log\frac{2}{2n} + \frac{2d_4}{2n}\log\frac{1}{2n}\} \\
= \quad & \{\frac{2d_1 + d_3(f_3-1) + d_4(f_4-1)}{2n}\log\frac{2}{2n} \\
+ & \frac{2d_2+d_3}{2n}\log\frac{3}{2n} + \frac{d_3+2d_4}{2n}\log\frac{1}{2n}\}
\end{aligned}$$

which attains the minimum when $d_4 = 0$ and the above assumptions **(C1)** and **(C2)** hold. This implies that a necessary condition for an assignment being a MPL estimator is that there is no unresolved group.

(3). **The mutation, crossover and exchange operators.** Note that in the mutation operator in our algorithm, a new vector $\mathbf{y}$ is generated by randomly selecting a member, say $\mathbf{z}_k$, from the population $\mathbf{Z}$ and by randomly mutating some components of $\mathbf{z}_k$ from 0 to 1 or from 1 to 0. $\mathbf{Z}$ is replaced by the proposed population $\mathbf{Y} = \{\mathbf{z}_1, \cdots, \mathbf{y}, \cdots, \mathbf{z}_N\}$ with probability $\min\{1, r_m\}$, where r_m is the Metropolis-Hastings ratio, $r_m = f(\mathbf{Y})/f(\mathbf{Z})$. In the crossover operator, a new pair of vectors, say $\mathbf{y}_i, \mathbf{y}_j$, are two "offspring" of a pair $\mathbf{z}_i, \mathbf{z}_j$ $(i \neq j)$ selected from $\mathbf{Z}$ according to a roulette wheel procedure. See Liang and Wong (2000) for the details. $\mathbf{Z}$ is updated by the proposal population $\mathbf{Y} = \{\mathbf{z}_1, \cdots, \mathbf{y}_i, \cdots, \mathbf{y}_j, \cdots, \mathbf{z}_N\}$ with probability $\min\{1, r_m\}$ where r_m is the Metropolis-Hastings ratio,

$$r_m = \frac{f(\mathbf{Y})P((\mathbf{y}_i, \mathbf{y}_j)|\mathbf{Y})}{f(\mathbf{Z})P((\mathbf{z}_i, \mathbf{z}_j)|\mathbf{Z})}$$

where

$$\begin{aligned} P((\mathbf{z}_i, \mathbf{z}_j)|\mathbf{Z}) &\propto \exp\{l(\mathbf{G}|\mathbf{z}_i)/t_i\} + \exp\{l(\mathbf{G}|\mathbf{z}_j)/t_j\} \\ P((\mathbf{y}_i, \mathbf{y}_j)|\mathbf{Y}) &\propto \exp\{l(\mathbf{G}|\mathbf{y}_i)/t_i\} + \exp\{l(\mathbf{G}|\mathbf{y}_j)/t_j\}. \end{aligned}$$

In the exchange operator, we change the order of two randomly selected $\mathbf{z}_i$ and $\mathbf{z}_j$ (without changing the order of t_i and t_j) with probability $\min\{1, r_e\}$, where

$$r_e = \exp\{(-l(\mathbf{G}|\mathbf{z}_i) + l(\mathbf{G}|\mathbf{z}_j))(1/t_i - 1/t_j)\}.$$

Acknowledgements

We are very grateful to Prof. F. Liang, The National University of Singapore, for allowing us using the part of his codes on the evolutionary Monte Carlo algorithm, and to Dr. Klaus Rohde, Max Delbrueck Center for Molecular Medicine, for showing us how to specify the prior haplotype frequency estimator. We would like also thank Dr. T. Müller and Dr. J. Stoye, Max-Planck-Institute for Molecular Genetics, for their helpful comments.

References

Baltimore, D. (2001) Our genome unveiled. *Nature* **409**, 814-816.

Bonnen, P.E., Story, D., and et al. (2000) Haplotypes at ATM identify coding-sequence variation and indicate a region of extensive linkage disequilibrium. *Am. J. Hum. Genet.* **67,** 1437-1451.

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

Clark, A., Weiss, K. and et al. (1998) Haplotype structure and population genetic inference from nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.* **63**, 595-612.

Davidson, S. (2000) Research suggests importance of haplotypes over SNPs. *Nat. Biotechnology* **18,** 1134-1135.

Drysdale, C.M., McGraw, D.W., and et al. (2000) Complex promoter and coding region β_2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness.*PNAS* **97,** 10483-10488.

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

Hawley, M.E. and Kidd, M.E. (1995) HAPLO: A program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *The Jornal of Heredity* **86**, 409-411.

Hodge, S.E., Boehnke, M. and Spence, M. (1999) Loss of information due to ambiguous haplotyping of SNPs. *Nat. Gent.* **21,** 360-361.

Fallin, D., Cohen, A., and et al. (2001). Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. *Genome Research* **11,** 143-151.

Fallin, D. and Schork, N. (2000) Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *Am. J. Hum. Genet.* **67**, 947-959.

THE GENOME INTERNATIONAL SEQUENCING CONSORTIUM (2001) Initial sequencing and analysis of the human genome. *Nature* **409,** 860-921.

Gusfield, D. (2000) A practical algorithm for optimal inference of haplotypes from diploid populations. *ISMB 2000 Proceedings, Eighth International Conference on Intelligent Systems for Molecular Biology*, edited by Altman, R., Bailey, T.L. et al., AAAI Press, pp. 183-189.

Hoehe, M.R., Köpke, K., and et al. (2000) Sequence variability and candidate gene analysis in complex disease: association of μ opioid receptor gene variation with substance dependence. *Human Molecular Genetics* **9,** 2895-2908.

Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* **22,** 139-144.

Lam, J.C., Roeder, K., and Devlin, B. (2000) Haplotype fine mapping by evolutionary tree. *Am. J. Hum. Genet.* **66**, 659-673.

Liang, F. and Wong, W. (2000) Evolutionary Monte Carlo: Applications to model sampling and change point problem. *Statist. Sinica* **10** 317-342.

Liu, J., Sabatti, C., and et al. (2000) Bayesian analysis of haplotypes for linkage disequilibrium mapping.*Technical Report,* Department of Statistics, Harvard University.

Long et al.(1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am. J. Hum. Genet.* **56** 799-810.

Nickerson D.A., Taylor, S.L., Weiss, K.M., Clark, A.G., and et al. (1998) Genome resequenencing and variation analysis in a 9.7 kb region of the human lipoprotein lipase gene. *Nat. Genet.* **19,** 233-240.

McPeek, M.S. and Strahs, A. (1999) Assessing linkage disequilibrium using the decay of haplotype sharing with application to fine-scale genetic mapping. *Am. J. Hum. Genet.* **65,** 858-875.

Page R.D.M. and Holmes E.C. (1998) *Molecular Evolution. Blackwell Science,* Oxford, London.

Pritchard, J.K.(2001) Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69,** 124-137.

Posada, D. and Crandall, K.A. (2001) Intraspecific gene genealogies: trees grafting into networks. *Trends in Ecology & Evolution* **16**, 37-45.

Rannala, B. and Reeve, J.P. (2001) High-resolution multiple linkage- disequilibrium mapping in the context of a human genome sequence. *Am. J. Hum. Genet.* **69,** 159-178.

Service, S.K., Temple Lang, D.W., Freimer, N.B., and Sandkuijl, L.A. (1999) Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. *Am. J. Hum. Genet.* **64**, 1728-1738.

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

Templeton, A.R., Weiss, K.M., and et al. (2000) Cladistic structure within the human lipoprotein lipase gene and its implications for phenotypic association studies. *Genetics* **156**, 1259-1275.

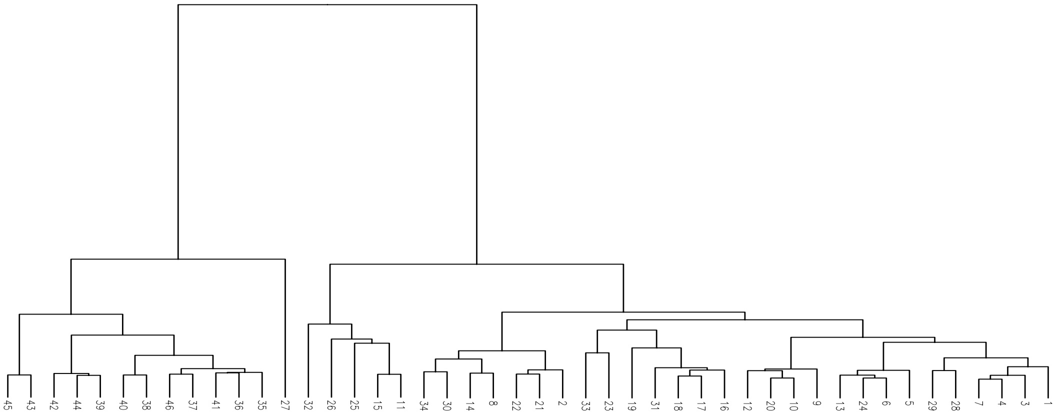
Zhang, J. and Vingron, M. (2001) An information distance with applications in sequence clustering. *manuscript.*

Table 3.1. Haplotypes reconstructed from African-American substance-dependent individuals and controls predicted by the PL approach

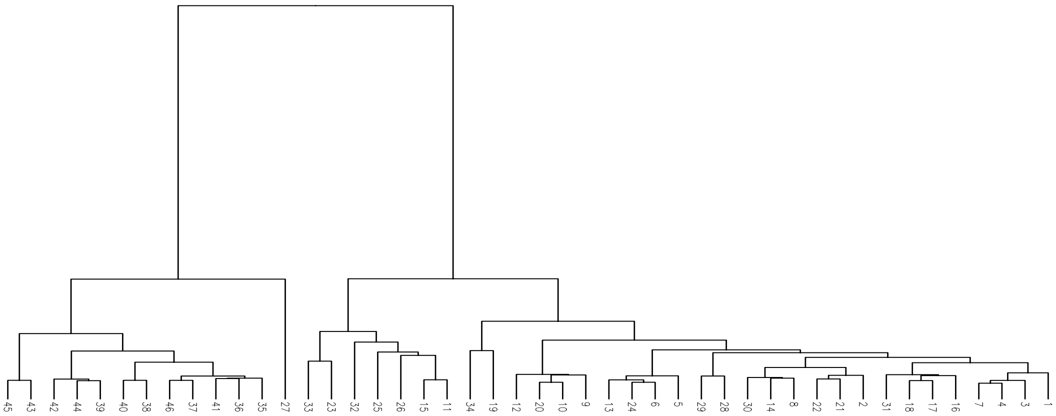
No.	Haplotypes	Counts		No.	Haplotypes	Counts	
		Cases	Controls			Cases	Controls
1	0000000000000000000000000	31	8	24	0001000000000000010000000	4	2
2	0000000000000000000000100	108	24	25	0001000000000010010010000	0	1
3	0000000000000000000100000	2	0	26	0010000100000010000010000	1	0
4	0000000000000000001000000	1	1	27	0010100000001000000000000	2	0
5	0000000000000000010000000	14	7	28	0100000000000000000000000	0	1
6	0000000000000000010000010	1	0	29	0100000000000000000100000	1	0
7	0000000000000000100000000	0	1	30	0100000000000001000000100	2	0
8	0000000000000001000000100	8	3	31	0100000000100000000000000	0	1
9	0000000000000010000000000	13	8	32	0100000000100010000010001	1	0
10	0000000000000010000000010	1	0	33	0100000010100010000000000	1	0
11	0000000000000010000010001	3	2	34	0100000100000000000000100	1	0
12	0000000000000010010000000	5	0	35	1010100000001010000000000	3	0
13	0000000000010000010000000	2	0	36	1010100000001010000000100	1	0
14	0000000000010001000000100	1	0	37	1010100000001010000001000	1	0
15	0000000000010010000010001	1	0	38	1010100000001010000100000	1	1
16	0000000000100000000000000	20	5	39	1010100000011010000001000	1	0
17	0000000000100000100000000	1	0	40	1010100001001010000100000	2	0
18	0000000000100100000000000	4	1	41	1010101000001010000000000	1	0
19	0000000000100100000000100	1	0	42	1010110000001010000001000	9	0
20	0000000010000010000000000	5	0	43	1010110000001010000010001	2	0
21	0000000100000000000000100	12	3	44	1010110000011010000001000	1	0
22	0000001000000000000000100	3	0	45	1010110100001010000010000	1	0
23	0000001000100010000000000	0	1	46	1010110000001010000000000	1	0

Table 3.2. Haplotypes reconstructed from African-American substance-dependent individuals and controls predicted by the PL approach

No.	Haplotypes	Counts		No.	Hyplotypes	Counts	
		Cases	Controls			Cases	Controls
1	0000000000000000000000000	32	10	27	1000000000001010000000100	1	0
2	0000000000000000000000100	81	21	28	1000000000000010000001100	2	0
3	0000000000000000000100000	3	0	29	1000000000001010000100000	1	1
4	0000000000000000001000000	1	1	30	1000000000000010010000000	1	0
5	0000000000000000010000000	14	5	31	1000001000001010000000000	1	0
6	0000000000000000100000100	1	1	32	1010100000000010000001000	2	0
7	0000000000000001000000100	10	3	33	1010000000001000000000001	1	0
8	0000000000000010000000000	12	6	34	1000110000000000000001000	1	0
9	0000000000000010000010001	5	2	35	1000100000001001000000000	1	0
10	0000000000000010010000000	6	2	36	1010010000000000000000100	1	0
11	0000000000010000010000000	1	0	37	1010110000001010000001000	4	0
12	0000000000010000000000100	3	0	38	1010110000001010000010001	1	0
13	0000000000100000000000000	21	7	39	0000000000000000000000110	2	0
14	0000000000100100000000100	5	0	40	0000000000000100010000000	0	1
15	0000000010000010000000000	6	0	41	0010000000000010000000000	2	0
16	0000000100000000000000100	13	3	42	0010100000000000000000100	3	1
17	0000001000000000000000100	3	0	43	0010100000011000000000000	1	0
18	0000001000100010000000000	0	1	44	0010000001000010000100000	2	0
19	0001000000000000010000000	4	2	45	0000010000001000000000000	1	0
20	0001000000000010000010000	0	1	46	0010010000000010000001000	2	0
21	0010000100000010000010000	1	0	47	0000110000000010000010100	1	0
22	0010100000001000000000000	4	0	48	0010000000001011000000100	1	0
23	0100000000000000000000000	1	1	49	0000010000001010000000000	1	0
24	0100000000000000000000100	3	1	50	0010010000001000000001100	1	0
25	0100000000100000000000000	2	0	51	0000100100001010000010000	1	0
26	1000100000001000000000100	6	0	52	1010110000001010000000000	1	0


45
43
42
44
39
40
38
46
37
41
36
35
27
32
26
25
15
11
34
30
14
8
22
21
2
33
23
19
31
18
17
16
12
20
10
9
13
24
6
5
29
28
7
4
3
1

(a)



(b)

Figure 3.1. Dendrogram of the haplotypes in Table 3.1 based on the UPGMA clustering procedure according to the information distance (a) without weighting and (b) with weighting. Tips 1, 2, ..., 46 represent 46 haplotypes in Table 3.1. The lengths of trees (a) and (b) are 1.4262205 and 1.706099 respectively.

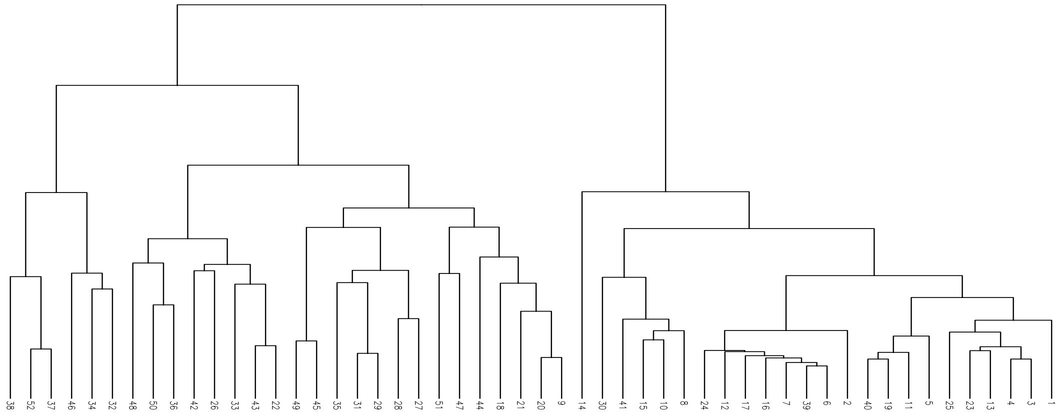


Figure 3.2. Dendrogram of the haplotypes in Table 3.2 based on the UPGMA clustering procedure according to the information distance with weighting. Tips 1, 2, ..., 52 represent 52 haplotypes in Table 3.2. The length of the tree is 1.382733.

Table 3.3. Simulated genotypes

No.	Genotypes	Counts		No.	Genotypes	Counts	
		Cases	Controls			Cases	Controls
1	100011000010110	1	0	26	221012002222210	2	2
2	000010101010010	1	0	27	212110111201110	3	0
3	010000011000011	16	2	28	220121211211222	1	0
4	001001110111010	1	0	29	211212000120222	1	0
5	000001110010101	4	2	30	122121222002222	1	0
6	011011101111001	0	1	31	222000100122021	14	3
7	010011101011000	2	2	32	022022002201102	0	1
8	110100111011010	3	1	33	202221022220020	1	0
9	101101010101001	1	0	34	212200222001021	0	1
10	110001011100101	1	0	35	222112022222212	0	1
11	020021212012122	1	0	36	121202222102110	1	0
12	221022002222202	1	0	37	201002222222000	1	0
13	212210202112022	1	0	38	222100222201211	1	0
14	201202222012021	2	1	39	202112122202220	2	0
15	220202122212121	7	2	40	222212022021222	0	1
16	110220111102212	1	1	41	112120202212020	1	0
17	022212202212022	1	0	42	222221001022001	1	0
18	012210020210220	0	1	43	102222202220200	2	0
19	022112010022201	4	0	44	221222020020220	1	0
20	111002020210210	2	0	45	212212111101020	3	0
21	021212102001221	3	0	46	122120202112021	10	1
22	222112122210011	1	0	47	021121122011012	1	0
23	221121112222122	1	0	48	022010202222222	1	0
24	220222022202022	2	1	49	100211222220221	1	0
25	202211212222000	3	0	50	212220122120012	1	0

Table 3.4. Haplotypes reconstructed from the genotypes in Table 3.2
when all the counts of the cases are set 1 and all the counts
of the controls are set 0

No.	Haplotypes	Counts		No.	Haplotypes	Counts	
		Cases	Controls			Cases	Controls
1	100011000010110	2	0	33	111001000101110	1	0
2	000010101010010	3	0	34	111110000010010	2	0
3	010000011000011	3	0	35	110111111101000	1	0
4	001001110111010	2	0	36	011111110011010	1	0
5	000001110010101	4	0	37	110000111100011	2	0
6	011011101111001	2	0	38	001101101011011	2	0
7	010011101011000	2	0	39	000111101011001	1	0
8	110100111011010	2	0	40	010110010010100	1	0
9	101101010101001	2	0	41	001111010000101	1	0
10	110001011100101	2	0	42	111000000010010	1	0
11	010011011011110	1	0	43	001010101001101	1	0
12	101001001010000	4	0	44	011110101010011	1	0
13	110110101111001	2	0	45	011101111111101	1	0
14	101000010010001	1	0	46	000111111011000	2	0
15	110100101111111	1	0	47	011110111001110	1	0
16	110110111101110	2	0	48	001000100100011	1	0
17	011010000110010	5	0	49	000001001001100	1	0
18	010110010011001	2	0	50	000111010100010	2	0
19	111001010110110	1	0	51	011000010001001	1	0
20	011111100001011	1	0	52	101100111100110	1	0
21	100111110110011	2	0	53	001000110101000	1	0
22	101111110000110	3	0	54	001100010101111	1	0
23	100111000101000	1	0	55	000110101101000	1	0
24	101011010100000	1	0	56	110100101111000	1	0
25	101011001001110	2	0	57	001011001010001	1	0
26	110101011111111	1	0	58	100110100100100	1	0
27	111111000100101	1	0	59	001001010000100	1	0
28	110101001001001	2	0	60	011010111101010	1	0
29	110000100111001	1	0	61	101100000110011	1	0
30	011010000101101	3	0	62	100011001000101	1	0
31	110100101001011	2	0	63	011110100110010	1	0
32	111110001011111	1	0				

Table 3.5. Haplotypes reconstructed from the genotypes with counts in Table 3.2

No.	Haplotypes	Counts		No.	Haplotypes	Counts	
		Cases	Controls			Cases	Controls
1	100011000010110	2	0	33	111110001010010	1	1
2	000010101010010	4	0	34	111000000101110	1	0
3	010000011000011	34	5	35	100110100100100	4	0
4	001001110111010	2	0	36	100111001000001	3	0
5	000001110010101	16	6	37	111000000010010	3	0
6	011011101111001	0	2	38	111110111101000	3	0
7	010011101011000	4	4	39	011111111011010	1	0
8	110100111011010	6	2	40	111110111100010	1	0
9	101101010101001	2	0	41	001101100011011	3	1
10	110001011100101	2	0	42	110000111100011	1	1
11	010011011011110	1	0	43	010110010010100	0	1
12	101001001010000	4	0	44	001111010000101	4	0
13	110110101111001	11	1	45	001111100001001	3	0
14	101000011010001	2	1	46	010111101110011	1	0
15	110100101111111	7	2	47	011111110011101	1	0
16	110110111101110	4	1	48	000111111011000	4	0
17	011111000111001	1	0	49	011110111001110	3	0
18	011010000110010	4	3	50	010000100110011	15	3
19	010110010011001	4	1	51	000001001001100	0	1
20	111001010110110	2	0	52	000101001010010	1	0
21	011010101001111	3	0	53	011000111001001	0	1
22	101110110010011	1	0	54	000111010101111	0	1
23	101101111100110	2	0	55	001000110101000	1	0
24	100111000101000	2	1	56	001100111101111	1	0
25	101011010100000	4	0	57	001111111001010	2	0
26	101011001001110	2	3	58	110100100111000	1	0
27	110101011111111	1	0	59	011001001011001	1	0
28	111111000100101	1	0	60	001111010000100	1	0
29	111101110001110	1	0	61	010011111101010	3	0
30	101000100101001	14	3	62	101100000110011	10	1
31	011010000101101	2	1	63	100011110110111	1	0
32	110100000001011	1	1				

Table 3.6. A simulated genotype data set and the haplotypes reconstructed from this data set by the tree approach

No.	Genotypes	Haplotype pairs	
		Haplotype 1	Haplotype 2
1	020021212012122	010001011011100	000011110010111
2	221022002222202	111001000011000	001010001100101
3	220202122212121	100000110111101	010101101010111
4	022212202212022	011010100010000	000111001111011
5	220222022202022	100111000001001	010000011100010
6	122121222002222	110101010000001	101111101001110
7	022022002201102	011011000101101	000000001001100
8	202221022220020	100111011110000	001001000000010
9	212200222001021	111000011001011	010100100001001
10	222112022222212	111110010000110	000111001111011
11	201002222222000	101000001010000	001001110101000
12	222100222201211	110100101101011	001100010001111
13	202112122202220	101111101001110	000110110100000
14	222212022021222	110011000001110	001110011011001
15	102222202220200	101100000110100	100011101000000
16	022010202222222	010010100011010	001010001100101

Table 3.7. The haplotypes reconstructed from the same data set as in Table 3.6 by the PL approach

No.	Genotypes	Haplotype pairs	
		Haplotype 1	Haplotype 2
1	020021212012122	010001111010111	000011010011100
2	221022002222202	101001001110000	011010000001101
3	220202122212121	100100100111101	010001111010111
4	022212202212022	011111000011001	000010101110010
5	220222022202022	100011001100000	010100010001011
6	122121222002222	110111100001110	101101011000001
7	022022002201102	011010000001101	000001001101100
8	202221022220020	101001001110000	000111010000010
9	212200222001021	111000101001001	010100010001011
10	222112022222212	111110001111111	000111010000010
11	201002222222000	101001001110000	001000110001000
12	222100222201211	101100101101111	010100010001011
13	202112122202220	100110100000100	001111111101010
14	222212022021222	100111011011010	011010000001101
15	102222202220200	101001001110000	100110100000100
16	022010202222222	011010000001101	000010101110010