Original Paper

Human Heredity

Hum Hered 2005;59:144–156
DOI: 10.1159/000085938

Received: June 1, 2004
Accepted: March 24, 2005
Published online: May 26, 2005

Haplotype Reconstruction for Diploid Populations

Jian Zhang[a, b, c] Martin Vingron[d] Margret R. Hoehe[d]

[a]Institute of Mathematics and Statistics, University of Kent, Canterbury, Kent, UK, [b]EURANDOM, Eindhoven, The Netherlands, [c]Chinese Academy of Sciences, and [d]Max Planck Institute for Molecular Genetics, Berlin, Germany

Key Words
Haplotype analysis · Haplotype likelihood · Multiple SNPs · Unlinked SNPs · Markov chain Monte Carlo · Empirical Bayes · Evolutionary tree

Abstract
The inference of haplotype pairs directly from unphased genotype data is a key step in the analysis of genetic variation in relation to disease and pharmacogenetically relevant traits. Most popular methods such as Phase and PL do require either the coalescence assumption or the assumption of linkage between the single-nucleotide polymorphisms (SNPs). We have now developed novel approaches that are independent of these assumptions. First, we introduce a new optimization criterion in combination with a block-wise evolutionary Monte Carlo algorithm. Based on this criterion, the 'haplotype likelihood', we develop two kinds of estimators, the maximum haplotype-likelihood (MHL) estimator and its empirical Bayesian (EB) version. Using both real and simulated data sets, we demonstrate that our proposed estimators allow substantial improvements over both the expectation-maximization (EM) algorithm and Clark's procedure in terms of capacity/scalability and error rate. Thus, hundreds and more ambiguous loci and potentially very large sample sizes can be processed. Moreover, applying our proposed EB estimator can result in significant reductions of error rate in the case of unlinked or only weakly linked SNPs.

Introduction

In all phases of disease gene discovery, it is of paramount importance to correctly determine the haplotypes, the specific combinations of given sequence variants for each of the two chromosomes of an individual [1–3]. These haplotypes may relate to ancestral chromosomal segments and/or defined candidate genes [1]. The currently used standard experimental approaches to the analysis of genetic variation, sequencing and genotyping, rely on the analysis of diploid (mixed) genomic DNA. Applying these technologies does not allow the direct determination of the underlying molecular haplotypes [4]. This means that we usually have to depend on unphased genotypes. In principle, phase information can be obtained by genotyping the family members of each individual; in many cases, however, their genotypes simply are not available or not sufficient to resolve haplotype ambiguity completely [5]. Existing methods for molecular haplotyping such as allele-specific long-range PCR [6, 7], carbon nanotube probing [8], or the construction of somatic cell hybrids [9] are to date too cost and labour intensive and not amenable to automation. Recently developed, novel and promising technologies such as polony haplotyping [10], single-copy DNA genotyping in conjunction with the MassARRAY system [11], or (fosmid/cosmid) clone-based systematic haplotyping [12] have not yet become available for efficient routine use.

Thus, in the past decade, several in silico methods have been developed that allow the prediction of haplo-

KARGER

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

0001–5652/05/0593–0144$22.00/0

Accessible online at:
www.karger.com/hhe

Dr. Margret R. Hoehe
Genetic Variation, Haplotypes & Genetics of Complex Disease, Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Ihnestrasse 73
DE–14195 Berlin (Germany)
Tel. +49 30 8413 1468, Fax +49 30 8413 1462, E-Mail hoehe@molgen.mpg.de

types from unphased genotype data. Two of them have become particularly popular. One is Clark's approach [6, 13], the other the EM approach based on the genotype likelihood [14–18]. The first approach is motivated by the fact that for any haplotype that is common enough such that homozygotes can be found in the sample, the sample is expected to have several heterozygotes bearing one copy of that haplotype [19]. This approach focuses on the prediction of haplotype pairs instead of estimating haplotype frequencies. However, it does not take full advantage of the count information (i.e. the information from the repeated observations) of some genotypes. This limitation can strongly affect the conclusions drawn by the subsequent haplotype-based statistical analyses (see Results). In contrast, the second approach has a clear statistical background and is not restricted to biallelic loci. EM is applied first to estimate haplotype frequencies and then to assign to each individual genotype the haplotype pair with the highest frequency among all possible pairs consistent with the genotype under consideration. Unfortunately, application of EM is limited by the size of the problem it can tackle [20].

In an effort to contend with these limitations, Stephens et al. [20] and Niu et al. [17] developed some new statistical methods, called Phase and PL, respectively. Both Phase and PL require the assumption that the SNPs are linked and consider the unknown haplotypes as unobserved random quantities, which avoids the requirement of storing estimated haplotype frequencies for every possible haplotype in the sample, a limitation of the EM approach. This allows both methods handling many more ambiguous loci than EM. Moreover, the error-rate improvement of Phase over Clark's and EM results from detailed model assumptions, specifically, the assumption of a coalescent model that has been made to infer the conditional distribution of the unknown haplotypes underlying the genotype data. In contrast, PL attempts to improve the error-rate performance of Clark's and EM by using partition ligation and prior annealing techniques in order to regularise the conditional distribution of the unknown haplotypes, as compared to imposing a specific model on this conditional distribution. This explains why the performance of Phase is better than PL under the coalescent assumption and worse than PL, when this assumption is invalid.

A major motivation of our work has been to develop an empirical Bayesian procedure, termed EB, in order to overcome the limitation of PL, i.e. deterioration of performance, when the SNPs under consideration become weakly linked or unlinked. The proposed EB procedure represents an extension of/improvement over PL in the sense that the same Bayes model is used, it is similarly based on the same haplotype likelihood framework, however, the requirement of tightly linked loci has been removed through introducing the empirical pseudo-counts in the prior. It has been demonstrated by Niu et al. [17] that PL may provide better results than Phase under the condition of an invalid coalescence assumption. Thus, this conclusion is supposed to hold for our EB. The above described improvement is achieved through the use of a new optimization criterion, an empirical posterior, as well as a block-wise evolutionary Markov chain Monte Carlo algorithm. The empirical posterior is based on the so-called profile haplotype likelihood, which usually assigns the relatively higher likelihood to that specific haplotype that occurs in both homozygotes and heterozygotes. Relying on this profile likelihood, we present a new haplotype estimation procedure as an alternative to EM, called the maximum haplotype likelihood (MHL). We show that both MHL and EB can be derived by optimization of the profile haplotype likelihood or of a posterior of the haplotypes only. This property helps also to contend the limitation of the EM approach, that is, the need to store the estimated haplotype frequencies for every possible haplotype in the sample. In contrast to PL, we address the issue of how to specify the prior by using the Excoffier-Slatkin assumption [14]. That is, we use those haplotype frequencies in the space of haplotypes that are compatible with the genotypes, in order to specify the pseudo-counts in the prior.

In order to evaluate the performance of our proposed solutions, we carried out at first several simulation studies considering two scenarios, one characterized by tightly linked SNPs, and the other by weakly linked or unlinked SNPs, respectively. While the first scenario has been well motivated in the literature [17, 20], the second one has attracted attention just recently [21]. The rationale of the second scenario may be outlined as follows: SNPs may, in practice, not reside on the same chromosome and therefore be physically unlinked. However, some of them may still interact with each other and in that, confer genetic risk to complex disease [21]. This implies that the two physically unlinked haplotypes may have the same parental identity if they are coupled with each other as the result of the potential interactions of disease genes in the gene network. It has been demonstrated by Zhang et al. [21] that phasing two unphased genotype blocks together allows prediction of these interactions. Moreover, considering the potentially interacting SNPs as one block, that means in practice, combining

all SNPs and phase them as a total, was shown to improve the accuracy of haplotype prediction.

In order to extend assessment of the performance of our proposed procedures, we applied them consecutively to two real data sets, the human μ opioid receptor gene (OPMR1) and the angiotensin converting enzyme (ACE) data sets [2, 22]. Referring to the first one, our motivation was to examine whether the risk pattern extracted by Hoehe et al. [2] could be derived by our proposed procedures as compared to the other methods outlined, Clark, EM, Phase and PL, respectively. The ACE data set, on the other side, was used, because it appeared to allow a much more detailed investigation of the specific mechanisms involved in the performance of the different methods under comparison.

Upon application to both simulated and real data sets, we conclude that our EB procedure not only can tackle a large size problem, but also can perform significantly better than the existing methods in terms of error rate in many instances. This improvement in performance might be due to using the efficient evolutionary Monte Carlo algorithm and adopting the above described empirical Bayesian method to regularize the model. Note that haplotyping is an ill-posed problem, in which the dimension of parameter space is usually much higher than the sample size. Based on our simulations we conclude further that compared to all other existing approaches, under a coalescent model, our method underperforms (slightly) Phase and is very similar on average to PL in the case of tightly linked SNPs. However, in the case where the SNPs are unlinked or only weakly linked, our EB method clearly yields better results than PL.

Methods

Haplotype Likelihood

Consider a chromosomal region of u_0 loci specified by an allelic vector $(r_1, r_2, ..., r_{u_0})^T$, in the reference genomic sequence. Let $\boldsymbol{G} = (G_1, ..., G_n)$ denote the observed genotypes for the n individuals, where $G_i = (g_{i1}, ..., g_{iu_0})^T$, $()^T$ is the transpose, and g_{ij} is the genotype for individual i at locus j. Let g_{ij} take 0, 1, or 2 according to whether its genetic haplotype at the locus j is homozygous and identical with the reference sequence, or homozygous but different from the reference sequence, or heterozygous. A genotype is called ambiguous if it has at least 2 heterozygous sites. Let $H_i = (H_{i1}, H_{i2})$ denote the unobserved haplotype pair of G_i, and Θ_i the set of all possible haplotype pairs compatible with G_i (called the candidate haplotype set for G_i). Set $\mathbf{H} = (H_1, ..., H_n)$, one possible way to decompose $\boldsymbol{G}$ into haplotypes. Let $\mathbf{p} = (p_1, ..., p_{m0})$ denote population frequencies of all possible haplotypes compatible with $\boldsymbol{G}$, where m_0 is the number of these candidate haplotypes. Then, given $\boldsymbol{G}$, under the assumption of Hardy-Weinberg equilibrium, we derive the 'haplotype-likelihood'

$$L(\boldsymbol{G}|\,\mathbf{p},\,\mathbf{H}) = 2^c \prod_{i=1}^{n} p(H_{i1})\, p(H_{i2}), \tag{1}$$

as the function of unknown parameters $(\mathbf{p}, \mathbf{H})$, where c is the number of genotypes in $\boldsymbol{G}$ that have at least one heterozygous locus, and $p(H_{i1})$ and $p(H_{i2})$ are the population frequencies of haplotypes H_{i1} and H_{i2}, respectively.

Note that the constant 2^c does not affect the estimation of $(\mathbf{p}, \mathbf{H})$. This point can be shown in the following example:

Example 1. Suppose that $\boldsymbol{G} = \{(0,0,0,1)^T, (1,0,0,1)^T, (2,2,0,1)^T, (1,1,2,2)^T\}$ and that these genotypes have single count (i.e., multiplicity = 1). Then, the candidate haplotype sets of these genotypes are $\Theta_1 = \{h_1\}$, $\Theta_2 = \{h_2\}$, $\Theta_3 = \{h_1, h_2, h_3, h_4\}$, $\Theta_4 = \{h_5, h_7, h_3, h_6\}$, respectively, where $h_1 = (0,0,0,1)^T$, $h_2 = (1,0,0,1)^T$, $h_3 = (1,1,0,1)^T$, $h_4 = (0,1,0,1)^T$, $h_5 = (1,1,1,1)^T$, $h_6 = (1,1,1,0)^T$, and $h_7 = (1,1,0,0)^T$. So all the different haplotypes $\{h_1, h_2, h_3, h_4, h_5, h_6, h_7\}$ are compatible with $\boldsymbol{G}$. Denote their unknown population frequencies by p_1, p_2, p_3, p_4, p_5, p_6, p_7, respectively. Then, there are four ways (or so-called assignments) to decompose $\boldsymbol{G}$ into haplotypes, namely,

$\mathbf{H}_1 = \{(h_1,h_1), (h_2,h_2), (h_3,h_1)\ (h_5,h_7)\}$,
$\mathbf{H}_2 = \{(h_1,h_1), (h_2,h_2), (h_3,h_1)\ (h_3,h_6)\}$,
$\mathbf{H}_3 = \{(h_1,h_1), (h_2,h_2), (h_2,h_4)\ (h_5,h_7)\}$,
$\mathbf{H}_4 = \{(h_1,h_1), (h_2,h_2), (h_2,h_4)\ (h_3,h_6)\}$,

their likelihoods can be expressed as follows:

$L(\boldsymbol{G}|\mathbf{p},\mathbf{H}_1) = 4p(h_1)^2\, p(h_2)^2\, p(h_1)\, p(h_3)\, p(h_5)\, p(h_7)$,
$L(\boldsymbol{G}|\mathbf{p},\mathbf{H}_2) = 4p(h_1)^2\, p(h_2)^2\, p(h_1)\, p(h_3)\, p(h_3)\, p(h_6)$,
$L(\boldsymbol{G}|\mathbf{p},\mathbf{H}_3) = 4p(h_1)^2\, p(h_2)^2\, p(h_2)\, p(h_4)\, p(h_5)\, p(h_7)$,
$L(\boldsymbol{G}|\mathbf{p},\mathbf{H}_4) = 4p(h_1)^2\, p(h_2)^2\, p(h_2)\, p(h_4)\, p(h_3)\, p(h_6)$.

Here, we are mainly interested in the selection of the best among all possible ways $\mathbf{H}_k$, $1 \le k \le 4$ to decompose $\boldsymbol{G}$. The selection leads to an estimator of $\mathbf{p} = (p_1, ..., p_7)$. Note that the constant 4 does not affect the maximization of these haplotype likelihoods with respect to $(\mathbf{p}, \mathbf{H})$.

EM Approach

Note that the genotype likelihood in [14, 23] can be viewed as the marginal likelihood of $\mathbf{p}$ in $L(\boldsymbol{G}|\mathbf{p}, \mathbf{H})$, namely

$$L(\boldsymbol{G}|\boldsymbol{p}) = \prod_{i=1}^{n} \left\{ \sum_{(Z_1,Z_2)\in\Theta_i} p_{Z_1} p_{Z_2} \right\}.$$

Let $\log(L(\boldsymbol{G}|\mathbf{p}))/(2n)$ denote the log-genotype-likelihood. Within the EM approach, we attempt to find $\mathbf{p}$ that maximizes the above marginal likelihood. Haplotype pairs can be reconstructed by choosing the most probable ones, given the genotype data and the estimated population haplotype frequencies $\hat{\mathbf{p}}$.

Clark's Approach

Clark [6] proposed an algorithm for haplotype assignment, which consists of two steps: First, the initial set of the haplotypes is formed from the 'self-resolved' genotypes (i.e., those genotypes with one heterozygous position at the most). Second, a known haplotype is chosen in order to see whether any of the unresolved genotypes is the composite of a known haplotype and a complementary haplotype, and, if this is the case, the known haplotype set is updated by

adding the complementary haplotype. The second step is repeated until all unresolved genotypes are resolved or the remaining genotypes cannot be resolved any further. 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, is, the better the solution [13].

In order to find a potential approach to improve this method, let us consider the following example:

Example 2. Suppose that we have the same four different genotypes, $(0,0,0,1)^T$, $(1,0,0,1)^T$, $(2,2,0,1)^T$ and $(1,1,2,2)^T$ as in 'Example 1', but with counts (i.e., multiplicities) 1, n_2 (≥ 1), 1 and 1, respectively. That is, $\boldsymbol{G} = \{(0,0,0,1)^T, (1,0,0,1)^T, ..., (1,0,0,1)^T, (2,2,0,1)^T, (1,1,2,2)^T\}$. We adopt the same notations h_k, $1 \leq k \leq 7$ as defined in 'Example 1'. Then, similar to 'Example 1', we have four possible ways, $\mathbf{H}_1$, $\mathbf{H}_2$, $\mathbf{H}_3$, $\mathbf{H}_4$, to decompose the above described four different genotypes.

Using Clark's approach, we obtain a unique optimal solution, $\mathbf{H}_2 = \{(h_1,h_1), (h_2,h_2), (h_3,h_1) (h_3,h_6)\}$, in which h_2 has no heterozygous descendant. From there, we take first (h_1, h_2) as the initially known haplotype set, which we choose h_1 from, in order to resolve $(2,2,0,1)^T$. Then, the known haplotype set is updated by adding the complementary haplotype h_3. At last, $(1,1,2,2)^T$ is resolved by h_3. Note that the count information is completely ignored in this algorithm. This is in contradiction to the rationale of the approach as outlined in the 'Introduction'. For example, if we set $n_2 = 10$, the haplotype h_2 already has a much higher frequency than the other known haplotype in the initial haplotype set, h_1. 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 [24]. This implies that $(2,2,0,1)^T$ is more probably resolved by h_2 than by h_1 according to the rationale of Clark's method. However, choosing h_2, we fail to resolve all genotypes according to Clark's concept. This motivated us to develop a new procedure in the next section, taking into account the frequency information.

Maximum Haplotype Likelihood Approach

In order to make these improvements, we first introduce the maximum likelihood (ML) estimator of ($\mathbf{p}$, $\mathbf{H}$), denoted by ($\hat{\mathbf{p}}$, $\hat{\mathbf{H}}$) by maximizing the haplotype likelihood $L(\boldsymbol{G}|\mathbf{p}, \mathbf{H})$ in (1). On the other hand, for each particular assignment $\mathbf{H}$, $L(\boldsymbol{G}|\mathbf{p}, \mathbf{H})$ is proportional to

$$\prod_{k=1}^{k_0} p(H_k)^{s_k},$$

where H_k, $1 \leq k \leq k_0$ are all the different haplotypes in $\mathbf{H}$, (s_1, ..., s_{k_0}) are the numbers of times that they appear in $\mathbf{H}$, and $p(H_k)$, $1 \leq k \leq k_0$, are the unknown population frequencies of these haplotypes. It is obvious that $\Sigma_k s_k = 2n$. Maximizing $L(\boldsymbol{G}|\mathbf{p}, \mathbf{H})$ with respect to these population frequencies under the constraints $\Sigma_k p_k = 1$, $p_k \geq 0$, $1 \leq k \leq m_0$, we have

$$L\left(\boldsymbol{G}\middle|\hat{\mathbf{p}}(\mathbf{H}), \mathbf{H}\right) = \prod_{k=1}^{k_0}\left(\frac{s_k}{2n}\right)^{s_k},$$

with $\hat{\mathbf{p}}(\mathbf{H})$ being the maximum estimator of $\hat{\mathbf{p}}$ given $\mathbf{H}$. This leads to a definition of profile log-haplotype-likelihood:

$$l(\boldsymbol{G}|\mathbf{H}) = \sum_{k=1}^{k_0} \frac{s_k}{2n} \log \frac{s_k}{2n}. \tag{2}$$

Then, it is directly shown that $\hat{\mathbf{H}}$ attains the maximum of the above described profile log-haplotype-likelihood. Note that $-l(\boldsymbol{G}|\mathbf{H})$ is the entropy of $\mathbf{p}$. Furthermore, for haplotype Z, the ML estimator of its population frequency is the frequency of Z in $\hat{\mathbf{H}}$. We use 'Example 2' to show the advantage of the MHL over Clark's method in that MHL chooses the haplotypes with higher frequencies in resolving the unresolved genotypes. 'Example 2' serves moreover to elaborate the mathematical details on derivation of the above described profile log-likelihood in a reader-friendly manner. Any additional mathematical details will be made available on request to J.Zhang@kent.ac.uk.

Example 2 (continued). We have

$$\begin{aligned}
L(\mathbf{G}|\mathbf{p},\mathbf{H}_1) &= 4p(h_1)^2\, p(h_2)^{2n_2}\, p(h_1)\, p(h_3)\, p(h_5)\, p(h_7),\\
L(\mathbf{G}|\mathbf{p},\mathbf{H}_2) &= 4p(h_1)^2\, p(h_2)^{2n_2}\, p(h_1)\, p(h_3)\, p(h_3)\, p(h_6),\\
L(\mathbf{G}|\mathbf{p},\mathbf{H}_3) &= 4p(h_1)^2\, p(h_2)^{2n_2}\, p(h_2)\, p(h_4)\, p(h_5)\, p(h_7),\\
L(\mathbf{G}|\mathbf{p},\mathbf{H}_4) &= 4p(h_1)^2\, p(h_2)^{2n_2}\, p(h_2)\, p(h_4)\, p(h_3)\, p(h_6).
\end{aligned}$$

To maximize the first likelihood with respect to $p(h_k)$, $1 \leq k \leq 7$, we consider the following Lagrange multiplier:

$$l(\mathbf{p}, \lambda) = \log\left(L\left(\boldsymbol{G}|\mathbf{p}, \mathbf{H}_1\right)\right) - \lambda\left(\sum_{k=1}^{7} p_k - 1\right),$$

where λ is the Lagrange coefficient. Setting all partial derivatives of $l(\mathbf{p}, \lambda)$ to be zero and solving the resultant simultaneous equations, we obtain the estimate $\hat{\mathbf{p}}(\mathbf{H}_1)$. Substituting this estimate into $\log(L(\boldsymbol{G}|\mathbf{p}, \mathbf{H}_1))$ and dropping out constants, we obtain the following profile log-haplotype likelihood at $\mathbf{H}_1$:

$$l\left(\boldsymbol{G}|\mathbf{H}_1\right) = \frac{1}{6+2n_2}\left(3\log\left(\frac{3}{6+2n_2}\right) + 2n_2\log\left(\frac{2n_2}{6+2n_2}\right) + \log\left(\frac{1}{6+2n_2}\right) + \log\left(\frac{1}{6+2n_2}\right) + \log\left(\frac{1}{6+2n_2}\right)\right).$$

Similarly, we can derive the profile log-haplotype likelihoods at $\mathbf{H}_2$, $\mathbf{H}_3$, and $\mathbf{H}_4$ as follows:

$$l\left(\boldsymbol{G}|\mathbf{H}_2\right) = \frac{1}{6+2n_2}\left(3\log\left(\frac{3}{6+2n_2}\right) + 2n_2\log\left(\frac{2n_2}{6+2n_2}\right) + 2\log\left(\frac{2}{6+2n_2}\right) + \log\left(\frac{1}{6+2n_2}\right)\right),$$

$$l\left(\boldsymbol{G}|\mathbf{H}_3\right) = \frac{1}{6+2n_2}\left(2\log\left(\frac{2}{6+2n_2}\right) + (2n_2+1)\log\left(\frac{2n_2+1}{6+2n_2}\right) + \log\left(\frac{1}{6+2n_2}\right) + \log\left(\frac{1}{6+2n_2}\right) + \log\left(\frac{1}{6+2n_2}\right)\right),$$

$$l\left(\boldsymbol{G}|\mathbf{H}_3\right) = \frac{1}{6+2n_2}\left(2\log\left(\frac{2}{6+2n_2}\right) + (2n_2+1)\log\left(\frac{2n_2+1}{6+2n_2}\right) + \log\left(\frac{1}{6+2n_2}\right) + \log\left(\frac{1}{6+2n_2}\right) + \log\left(\frac{1}{6+2n_2}\right)\right).$$

It is obvious that if $n_2 = 1$, $l(\mathbf{G}|\mathbf{H}_2) = \max\{l(\mathbf{G}|\mathbf{H}_k)$: $1 \leq k \leq 4\}$. Thus, the MHL assignment (solution) is $\mathbf{H}_2 = \{(h_1,h_1), (h_2,h_2), (h_3,h_1) (h_3,h_6)\}$, which also attains the maximum resolution. However, if $n_2 = 10$, then there are two MHL assignments, $\mathbf{H}_3 = \{(h_1,h_1), (h_2,h_2), (h_2,h_4), (h_5,h_7)\}$ and $\mathbf{H}_4 = \{(h_1,h_1), (h_2,h_2), (h_2,h_4), (h_6,h_3)\}$.

Moreover, the log-genotype-likelihood obtains the same value of –0.7083934 through these two solutions. However, both have not attained the maximum resolution. This is not surprising because the haplotype h_2 has a greater frequency than h_1. Note that EM provides the same solution as MHL in this toy example. However, MHL does produce a better result than EM for the OPMR1 data as shown in the section 'Results'.

Empirical Bayesian Approach

As pointed out in the 'Introduction', the dimension of space of the candidate haplotypes can be much larger than the sample size. This can cause some problems such as the bias due to over-fitting. Even if the space dimension is lower than the sample size, the MHL estimator may not be unique and then the surface of the haplotype likelihood could be flat around the assignments of some genotypes (for instance, genotype $(1,1,2,2)^T$ in 'Example 2'). We called these unidentified genotypes orphans. In this situation, we seek the following (empirical) Bayesian approach to the problem by adopting the Dirichlet distribution

$$\prod_{k=1}^{m_0} p_k^{\alpha_k - 1}$$

and 1 as the priors for the (**p**, **H**), where $\alpha_1, ..., \alpha_{m_0}$ will be determined by the observed genotypes later. Then the posterior distribution of (**p**, **H**), p (**p**, **H**|***G***) is proportional to

$$\prod_{k=1}^{m_0} p_k^{s_k + \alpha_k - 1},$$

$(s_1, ..., s_{m_0})$ represent the counts (in **H**) of the haplotypes from the candidate haplotype space. The marginal posterior distribution for assignment **H**, p (**H**|***G***) is proportional to

$$f(\mathrm{H}|G) = \frac{\prod_{k=1}^{m_0} \Gamma(s_k(\mathbf{H}) + \alpha_k)}{\Gamma\left(2n + \sum_{k=1}^{m_0} \alpha_k\right)}, \qquad (3)$$

where $\Gamma(\cdot)$ is the Gamma function. We take the mode of p (**H**|***G***), $\mathbf{H}_B$ as the estimated haplotype assignment to ***G***, which can be written as $\mathbf{H}_B = \arg\max_{\mathrm{H}} f(\mathbf{H}|\mathbf{G})$ because $p(\mathbf{H}|G)$ is proportional to $f(\mathbf{H}|G)$. The details of deriving the above marginal posterior will be made available on personal request to J.Zhang@kent.ac.uk.

To specify the pseudo-counts α_k, $k = 1, ..., m_0$, we first follow Excoffier and Slatkin [14], assuming that all genotypes are equally important and that for each genotype all possible haplotype pairs are equally likely. This assumption is based on the fact that the individuals are identically and independently distributed. Furthermore, we assume that the two haplotypes for each haplotype pair are equally important. Then, we take the total weights for different haplotypes in the space of all haplotypes compatible to ***G*** as the pseudo-counts. We call this assumption the structural information in the genotypes, which leads to the following scheme:

We first assign the same weight $2d$ to the observed different genotypes, where d is a constant with a default value of 1. For each of these genotypes, we distribute the weight $2d$ equally to all its candidate haplotypes. Then, for $k = 1, ..., m_0$, let α_k be the summation of all the weights we put on the kth haplotype, which appears in the n candidate haplotype sets, Θ_i, $i = 1, ..., n$. The resulting Bayesian estimator is called EB.

We shall now use the toy 'example 2' again in order to line out the mechanism behind EB step by step.

Example 2 (continued). To apply EB, we set $d = 1$ in the above scheme and adopt the notations in example 1. Then, each haplotype in the candidate haplotype sets, $\Theta_1 = \{h_1\}$ and $\Theta_2 = \{h_2\}$, receives the prior weight of 1, while each haplotype in $\Theta_4 = \{h_1,h_2,h_3,h_4\}$ and $\Theta_3 = \{h_5,h_7,h_3,h_6\}$ gets the prior weight of 0.5. This gives the empirical prior for all candidate haplotypes h_k, $1 \le k \le 7$:

$\alpha_1 = 2.5$, $\alpha_2 = 2\ n_2 + 0.5$, $\alpha_3 = 0.5 + 0.5 = 1$,
$\alpha_4 = \alpha_5 = \alpha_6 = \alpha_7 = 0.5$.

As pointed out before, there are four possible ways, $\mathbf{H}_1$, $\mathbf{H}_2$, $\mathbf{H}_3$, $\mathbf{H}_4$, to decompose ***G***, with the marginal posteriors

$$f(\mathbf{H}_1|G) = \frac{\Gamma(3+2.5)\Gamma(2n_2+2n_2+0.5)\Gamma(1+1)\Gamma(1+0.5)^2}{\Gamma(4(n_2+3))},$$

$$f(\mathbf{H}_2|G) = \frac{\Gamma(3+2.5)\Gamma(2n_2+2n_2+0.5)\Gamma(2+1)\Gamma(1+0.5)}{\Gamma(4(n_2+3))},$$

$$f(\mathbf{H}_3|G) = \frac{\Gamma(2+2.5)\Gamma(2n_2+1+2n_2+0.5)\Gamma(1+0.5)^3}{\Gamma(4(n_2+3))},$$

$$f(\mathbf{H}_4|G) = \frac{\Gamma(2+2.5)\Gamma(2n_2+1+2n_2+0.5)\Gamma(1+0.5)^2\Gamma(1+1)}{\Gamma(4(n_2+3))}.$$

Note that

$$\frac{f(\mathbf{H}_2|G)}{f(\mathbf{H}_1|G)} = \frac{\Gamma(3)}{\Gamma(2)\Gamma(1.5)} = \frac{4}{\sqrt{\pi}} > 1,$$

$$\frac{f(\mathbf{H}_3|G)}{f(\mathbf{H}_2|G)} = \frac{\Gamma(4.5)\Gamma(4n_2+1.5)\Gamma(1.5)^2}{\Gamma(5.5)\Gamma(4n_2+0.5)\Gamma(3)} = \frac{(4n_2+0.5)\pi}{36},$$

$$\frac{f(\mathbf{H}_4|G)}{f(\mathbf{H}_3|G)} = \frac{\Gamma(2)}{\Gamma(1.5)} = \frac{2}{\sqrt{\pi}} > 1,$$

$$\frac{f(\mathbf{H}_4|G)}{f(\mathbf{H}_2|G)} = \frac{\Gamma(4.5)\Gamma(4n_2+1.5)\Gamma(2)\Gamma(1.5)}{\Gamma(5.5)\Gamma(4n_2+0.5)\Gamma(3)} = \frac{(4n_2+0.5)\sqrt{\pi}}{18},$$

and that

$$\frac{f(\mathbf{H}_3|G)}{f(\mathbf{H}_2|G)} > 1, \quad \frac{f(\mathbf{H}_4|G)}{f(\mathbf{H}_2|G)} > 1 \quad \text{if} \quad n_2 \ge 3;$$

$$\frac{f(\mathbf{H}_3|G)}{f(\mathbf{H}_2|G)} < 1, \quad \frac{f(\mathbf{H}_4|G)}{f(\mathbf{H}_2|G)} < 1 \quad \text{if} \quad 1 \le n_2 \le 2.$$

Therefore, we obtain the unique empirical Bayes estimate $\mathbf{H}_B = \mathbf{H}_4$, if $n_2 \ge 3$; $\mathbf{H}_B = \mathbf{H}_2$, if $1 \le n_2 \le 2$. Note that in the case where $n_2 = 10$, both the MHL and EM procedures provide the non-unique solutions $\mathbf{H}_3$ or $\mathbf{H}_4$, since both haplotype and genotype likelihoods have the same value at $\mathbf{H}_3$ and $\mathbf{H}_4$. The calculation of

$$\frac{f(\mathbf{H}_4|G)}{f(\mathbf{H}_3|G)}$$

showed that the empirical prior frequency of each candidate haplotype has played the key role for differentiation of $\mathbf{H}_4$ from $\mathbf{H}_3$. This illustrates why EB could be superior to MHL and EM. Note

that PL and Phase result, similarly to Clark's algorithm, in the solution $\mathbf{H}_2$, which is inconsistent with the rationale of the coalescent theory, if a coalescent model is assumed for this data set. See also the discussion in the paragraph 'Clark's Approach'.

Evolutionary Tree

As pointed out before, the analysis of haplotypes can be useful in both population and disease studies. Here we focus on the second issue. In order to test for association with complex disease, we estimate the frequencies of given haplotypes in cases and controls and examine, whether significant differences between cases and controls are given. In practice, the number of different haplotypes is often unfeasibly large, resulting in a sparse contingency table so that ordinary tests of association like the χ^2 test do not have sufficient power to detect an association with any single haplotype. In order to cope with this problem, several methods have been suggested, which rely on the classification of haplotypes into evolutionarily related ones [25] or functionally related (ideally functionally equivalent) ones, based on sequence-structure-function similarity [2]. In the present work, we build such a tree via a modified UPGMA procedure, where two modifications have been made. First, we use an information distance instead of the traditional Hemming distance. For this purpose, we calculate the pairwise percentage identities for m_0 different haplotypes defined in (1), say q_{ij}, $1 \le i, j \le m_0$. Then, the information distance between the ith and jth haplotypes is defined by the Kullback-Leiber distance between the probability vectors $(q_{i1}, ..., q_{im})$ and $(q_{i1}, ..., q_{jm})^T$ namely

$$d(i,j) = \sum_{k=1}^{m} q_{ik} \log\left(\frac{q_{ik}}{q_{\{i,j\},k}}\right) + (1-q_{ik}) \log\left(\frac{1-q_{ik}}{1-q_{\{i,j\},k}}\right) + \log\left(\frac{q_{jk}}{q_{\{i,j\},k}}\right) + (1-q_{jk}) \log\left(\frac{1-q_{jk}}{1-q_{\{i,j\},k}}\right),$$

where

$$q_{\{i,j\},k} = \frac{q_{ik} + q_{jk}}{2}.$$

Secondly, we adopt the following weighted average distance for any cluster pair, C_i and C_j, namely

$$d(C_i, C_j) = \frac{1}{\sum_{k_1 \in C_1} s_{k_1} \sum_{k_2 \in C_1} s_{k_2}} \sum_{k_1 \in C_1, k_2 \in C_1} s_{k_1} s_{k_2} d(k_1, k_2).$$

Unlike the conventional Hemming distance, here, not only the similarities between two haplotypes, but also their similarities to the other haplotypes, are used in pairwise comparison. Compared with the UPGMA procedure used in [2], our modification explores the multiplicity information of each genotype, which can effectively reduce the potential random fluctuation in the distance calculations.

Computation

For the Bayesian estimator, we note that according to (3), for any two assignments **H** and **H***,

$$\log\left(\frac{p(\mathbf{H}|\boldsymbol{G})}{p(\mathbf{H}^*|\boldsymbol{G})}\right) = \sum_{s_k(\mathbf{H}) \ge 1} \log\left(\frac{\Gamma(s_k(\mathbf{H}) + \alpha_k)\alpha_k}{\Gamma(1+\alpha_k)}\right) - \sum_{s_k(\mathbf{H}^*) \ge 1} \log\left(\frac{\Gamma(s_k(\mathbf{H}^*) + \alpha_k)\alpha_k}{\Gamma(1+\alpha_k)}\right)$$

That is, it suffices to calculate the pseudo-counts of the haplotypes in the assignments **H** and **H***, when we compare **H** and **H*** in terms of their posteriors. This is very similar to the case of maximizing the profile log-haplotype likelihood in (2). Thus, we only need to cope with the problem of how to calculate the MHL estimator. This can be solved taking the following two steps: First, for a haplotype assignment, we calculate the objective function, that is, the haplotype likelihood in the MHL case. Then we optimize this function with respect to the assignment.

We start out with m different genotypes G_1, ..., G_m with ν_1, ..., ν_m ambiguous loci, respectively. As stated before, although the EM approach does use the count information, it is limited by the requirement of storing 2^{h-1} variables for each genotype with h ambiguous loci. In contrast, for the MHL and Bayesian approaches, we only need to optimize an objective function with $\sum_{i=1}^{m} (\nu_i - 1)$ variables by re-expressing $l(\boldsymbol{G}|\mathbf{H})$ as a function with the $\sum_{i=1}^{m} (\nu_i - 1)$ ambiguous loci as variables. For this purpose, let $\mathbf{z}$ denote a $\sum_{i=1}^{m} (\nu_i - 1)$ dimensional variable, in which each component takes the value of 0 or 1. From $\mathbf{z}$, we can construct the haplotype pair (H_{1i}, H_{2i}) for each G_i, $i = 1, ..., m$ 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 remaining ambiguous positions of G_i are set $z_{k_{1+1}}$, ..., $z_{k_{1+\nu_1-1}}$, respectively, in H_{1i}, where

$$k_i = \sum_{j=1}^{i-1} (\nu_j - 1).$$

Ambiguous positions of G_i are set in H_{2i} to the opposite of the entry in H_{1i}. This implies that there is a one-to-one correspondence between z and the assignment **H** in (2). Moreover, for each $\mathbf{z}$, we identify the corresponding haplotypes and calculate their counts. Then, the log-haplotype likelihood in (2) can be written in the form $l(\boldsymbol{G}|\mathbf{z})$, as a function of $\mathbf{z}$. Now, the optimal assignment can be obtained by maximizing $l(\boldsymbol{G}|\mathbf{z})$. Although the number of the operational variables is, compared to the original optimization problem, significantly reduced, the new problem posed remains still a very hard optimization problem in a highly dimensional space. In particular, the new objective function $l(\boldsymbol{G}|\mathbf{z})$ with many subtle local maxima is not a convex function. Here, we apply a recently developed MCMC algorithm, called the evolutionary Monte Carlo algorithm [26], in order 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 supplied with mutation, crossover and exchange operators, and the updates are accepted or rejected according to the Metropolis rule. More specifically, given the current population $\mathbf{Z} = \mathbf{z}_1, ..., \mathbf{z}_N\}$ and a temperature ladder $\mathbf{t} = \{t_1, ..., t_N\}$, we construct a Boltzmann distribution for the population $\mathbf{Z}$ by

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

We sample the next population by taking 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 $\{1, ..., N\}$ 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 $w(\mathbf{z}_2|\mathbf{z}_1) = w(\mathbf{z}_{N-1}|\mathbf{z}_N) = 1$. Note that in the mutation operator, 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 $\mathbf{Y} = \{\mathbf{z}_1, ..., \mathbf{y}, ..., \mathbf{z}_N\}$ 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 $\mathbf{z}_i$, $\mathbf{z}_j$ $(i = j)$ selected from $\mathbf{Z}$ according to a roulette wheel procedure. For details see [26]. $\mathbf{Z}$ is updated by the proposal population $\mathbf{Y} = \{\mathbf{z}_1, ..., \mathbf{y}_i, ..., \mathbf{y}_j, ..., \mathbf{z}_N\}$ with probability $\min\{1, r_m\}$, where is the Metropolis-Hastings ratio,

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

and

$$P((\mathbf{z}_i, \mathbf{z}_j) \mid \mathbf{Z}) \propto \exp\{l(G \mid \mathbf{z}_i)/t_i\} + \exp\{l(G \mid \mathbf{z}_j)/t_j\},$$
$$P((\mathbf{y}_i, \mathbf{y}_j) \mid \mathbf{Y}) \propto \exp\{l(G \mid \mathbf{y}_i)/t_i\} + \exp\{l(G \mid \mathbf{y}_j)/t_j\},$$

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

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

In this paper, we set $t_i = t_h - (t_h - t_l)i/N$, $i = 1, ..., N$, where t_h and t_l are the highest and lowest temperatures, respectively. As in [24], we choose t_h and t_l such that $Var(l(\boldsymbol{G}|\mathbf{z}_i))(t_h - t_l)^2 = O(1)$ or simply by checking whether the overall acceptance rates of mutation, crossover and exchange operations are around 0.50.

Our experiences show that the EMC algorithm is often efficient, when the number of ambiguous loci in each genotype is less than 20. However, when a large number of ambiguous loci are involved, the idea of segmentation [17] is very useful in speeding up the convergence of our EMC algorithm. Like Niu et al. [17], we first divide each genotype into several blocks, say m_p blocks. We run the EMC algorithm for each block to find a list of partial haplotype pairs with the first m_0 highest likelihood values. We slightly modify our EMC algorithm and apply it to the restricted haplotype space generated by all the possible combinations of these partial haplotypes. More specifically, in this space we need to consider the $n \times m_p$ table (S_{ij}), where S_{ij} (subsets of all integers) represents the set of different partial haplotypes for the jth segment in the ith genotype. The vectors $m_i = (m_{i1}, ..., m_{im_p})$ with $m_{ij} \in S_{ij}$, $1 \leq i \leq n$ give an assignment of haplotype pairs to the genotypes. We view the segment with the size of $S_{ij} \geq 2$ as an ambiguous segment. Assume that there are v_i ambiguous segments in the ith genotype. Then it is similar to the case of the unrestricted haplotype space such that $L(\boldsymbol{G}|\mathbf{H})$ can be rewritten as a function of $z^* = (z_k^*)$, where z^* stands for the $\Sigma\, v_i$ ambiguous segments and z_k^* belongs to some integer subset S_k^*. For the new objective function, we only need to change the mutation operator in the EMC algorithm by randomly mutating some components of z^*, say z_k^*, from the current value to $z_k^* + 1$ or $z_k^* - 1$. We replace $z_k^* + 1$ by the left boundary of S_k^*, when it exceeds the right boundary of S_k^*. Similarly, when $z_k^* - 1$ exceeds the left boundary of S_k^*, we replace it by the right boundary. This strategy can be repeatedly used until the size of the restricted haplotype space becomes moderate. The resulting algorithm is called block-wise EMC algorithm.

Results

Simulation Studies

Comparison of Procedures Operating under the Assumption of Linkage between SNPs. First, the performances of the proposed procedures were examined under the assumption of linkage between SNPs. Following Stephens et al. [20] and using a coalescent-based program kindly provided by R. Hudson, we simulated 100 independent data sets, each containing 100 haplotypes, for $(\theta,R) = (4,0),(4,4)$, and $(4,40)$, respectively. Here $\theta = 4N_e\mu$, $R = 4N_e r$, N_e was the effective population size, μ the total per-generation mutation rate across the region sequenced and r the length (in Morgans) of the region sequenced. The lengths of these haplotypes varied from 10 to 40 and depended on (θ,R). For each data set, the haplotypes were randomly paired to form 50 genotypes. Then we applied the MHL, EB, and PL methods, respectively, to reconstruct these haplotypes from the resulting genotypes in which phase information was ignored. In our algorithm, we set the population size $N = 20$, the highest and lowest temperatures $(t_h,t_l) = (0.02, 0.001)$ in MHL, and $(t_h,t_l) = (0.0, 0.003)$ in EB, and the mutation and crossover parameters, $p_m = 0.2$, $p_0 = 0.004$, $p_1 = 0.008$ and $p_2 = 0.01$. We set the round parameter equal to 20 in PL as suggested by Niu et al. [17]. The performance of each method on each data set was measured by the error rate, being the proportion of ambiguous genotypes, the haplotypes of which were incorrectly assigned. For $(\theta,R) = (4,0), (4,4), (4,40)$, we compared the Clark, EM, EB, MHL, Phase and PL methods in terms of estimated error rates in substantial numbers of simulated data sets (see table 1). Although the results of the EM, Clark, and Phase methods in table 1 were obtained from different data sets, the average behaviour of these methods should not change significantly, because these data were generated from the above coalescent model with the same setting of (θ,R). This means that the results are comparable with those of the MHL, EB, and PL methods. Table 1 demonstrates that Phase has the lowest error rate, while PL and EB performed similarly and rank second. More specifically, the first three plots in figure 1 suggest that, while performing slightly worse than PL, EB in fact has a better performance than PL when the recombination rate R is very large, and

Table 1. The mean error rates (ME) and standard errors (SE) of the different methods

(θ,R)	MHL ME (SE)	EM ME (SE)	Clark ME (SE)	Phase ME (SE)	PL ME (SE)	EB1 ME (SE)
(4,0)	0.170 (0.023)	0.167 (0.028)	0.343 (0.030)	0.084 (0.028)	0.088 (0.012)	0.103 (0.012)
(4,4)	0.166 (0.020)	0.167 (0.020)	0.343 (0.030)	0.064 (0.011)	0.078 (0.006)	0.086 (0.007)
(4,40)	0.387 (0.023)	0.390 (0.028)		0.2000 (0.018)	0.270 (0.014)	0.264 (0.014)

Both ME and SE represent the sample mean and standard error of the error rate, which is the proportion of ambiguous genotypes, haplotypes of which have been incorrectly assigned. For each combination of parameters subjected to analyses, the values of ME (SE) for the EB, MHL and PL methods have been derived from the 100 simulated data sets. The values of ME (SE) for the EM, Clark, and Phase methods, kindly provided by Stephens et al. [20] have been derived from 90 to 100 simulated data sets using the same coalescent model described in the Results section. The EM method requires a pre-treatment of the data sets, that is, discarding those data sets for which the total number of possible haplotypes was $>10^5$ (the limit of Stephens' implementation of the EM algorithm). Stephens et al. had not provided the result for the Clark method when $(\theta,R) = (4,40)$.

therefore the loci weakly linked. Table 1 also shows that MHL has almost the same error rate as EM, whereas, in contrast to EM, MHL can handle genotypes with a large number of ambiguous loci in these simulated data sets. Although the Clark method can cope with genotypes with large numbers of ambiguous loci, too, it has a considerably higher error rate than the MHL method developed by us.

Comparison of Procedures Operating under the Assumption of Weak Linkage or No Linkage between SNPs. We partitioned each genotype in the previously simulated 100 datasets into two blocks with approximately equal lengths,

$$\boldsymbol{G}_i = \begin{pmatrix} \boldsymbol{G}_i^{(1)} \\ \boldsymbol{G}_i^{(2)} \end{pmatrix} = \left(\begin{pmatrix} G_{i,1}^{(1)} \\ G_{i,1}^{(2)} \end{pmatrix}, \ldots, \begin{pmatrix} G_{i,50}^{(1)} \\ G_{i,50}^{(2)} \end{pmatrix} \right),$$

where $i = 1, \ldots, 100$. A random sample $\{\boldsymbol{G}_i^*,\ i = 1,2,\ldots,30\}$ was then formed by pairing $G_{i,1}^{(1)}, \ldots, G_{i,50}^{(1)}$ with $G_{i+50,1}^{(2)}, \ldots, G_{i+50,50}^{(2)}$ for $i = 1, \ldots, 30$, where the underlying haplotype pairs were $(H_{i,k,1}^{(1)}, H_{i,k,2}^{(1)})$ for $G_{i,k}^{(1)}$, $1 \le k \le 50$ and $(H_{i+50,k,1}^{(2)}, H_{i+50,k,2}^{(2)})$ for $G_{i+50,k}^{(2)}$, $1 \le k \le 50$. Here, without loss of generality, the underlying haplotype blocks were assumed to have parental connections: $H_{i,k,1}^{(1)}$ was assumed being coupled with $H_{i+50,k,1}^{(2)}$ for $1 \le i \le 30$, $1 \le k \le 50$. Note that these genotypes consist of two independent parts, which, however, have the same population parameters (θ,R). We phased $\{\boldsymbol{G}_i^*,\ i = 1,2,\ldots,30\}$ directly by using EB, MHL and PL. The differences between resulting error rates of the three methods are represented by the last three plots in figure 1. This figure clearly demonstrates that EB performs better than PL in general. In particular, reductions of error rate up to 50% in the case of unlinked or only weakly linked SNPs can be obtained.

OPRM1 Data Set

The OPRM1 data set used included 172 African-American cases and controls, 25 of the identified variants were non-unique [2]. We first applied the MHL approach to this data set, making $M = 10^6$ iterations, a long run to obtain sufficient samples for inferring the maximum value of the log-haplotype likelihood $l(\boldsymbol{G}|\mathbf{z})$. 46 different haplotypes, numbered 1 to 46, were inferred. These haplotypes were either consistent with or slightly different from the 52 different haplotypes predicted by Hoehe et al. [2] by use of the EM method. In a second step, an evolutionary tree was constructed from these haplotypes by use of the modified UPGMA described in 'Methods'.

Based on this tree, the inferred haplotypes could be classified into two groups. Group one was almost the same as the group found associated with substance-dependence by Hoehe et al. [2] by means of similarity clustering: it included 13 haplotypes, 11 of which were common, significantly more frequent in the substance dependent individuals and shared the same pattern of sequence variants characterized by a unique combination of 5 polymorphic sites [2]. These 11 haplotypes occurred in 22 genotypes. Notably, the above described risk pattern has been experimentally validated by Hoehe et al. [2]. Thus, the phases of this subset of the SNPs are known.

EB yielded altogether 41 different haplotypes; Phase with the settings (burn-in, iteration) = $(10^4,10^4)$ and with (burn-in, iteration) = $(10^4,10^5)$, yielded 35 and 36 haplotypes, respectively; PL with two settings, round = 20 and round = 1000, resulted in 39 and 37 haplotypes, respectively. Although these seven sets of haplotypes may be quite different with respect to size, they do contain, however, a similar subset, which features the above described risk pattern.

It is obvious that both our solutions and that of Hoehe et al. [2] attain maximum resolution. In fact, Clark's method identifies a solution, which can be divided into two groups. But none of these groups carried the above described risk pattern. The relatively weaker performance

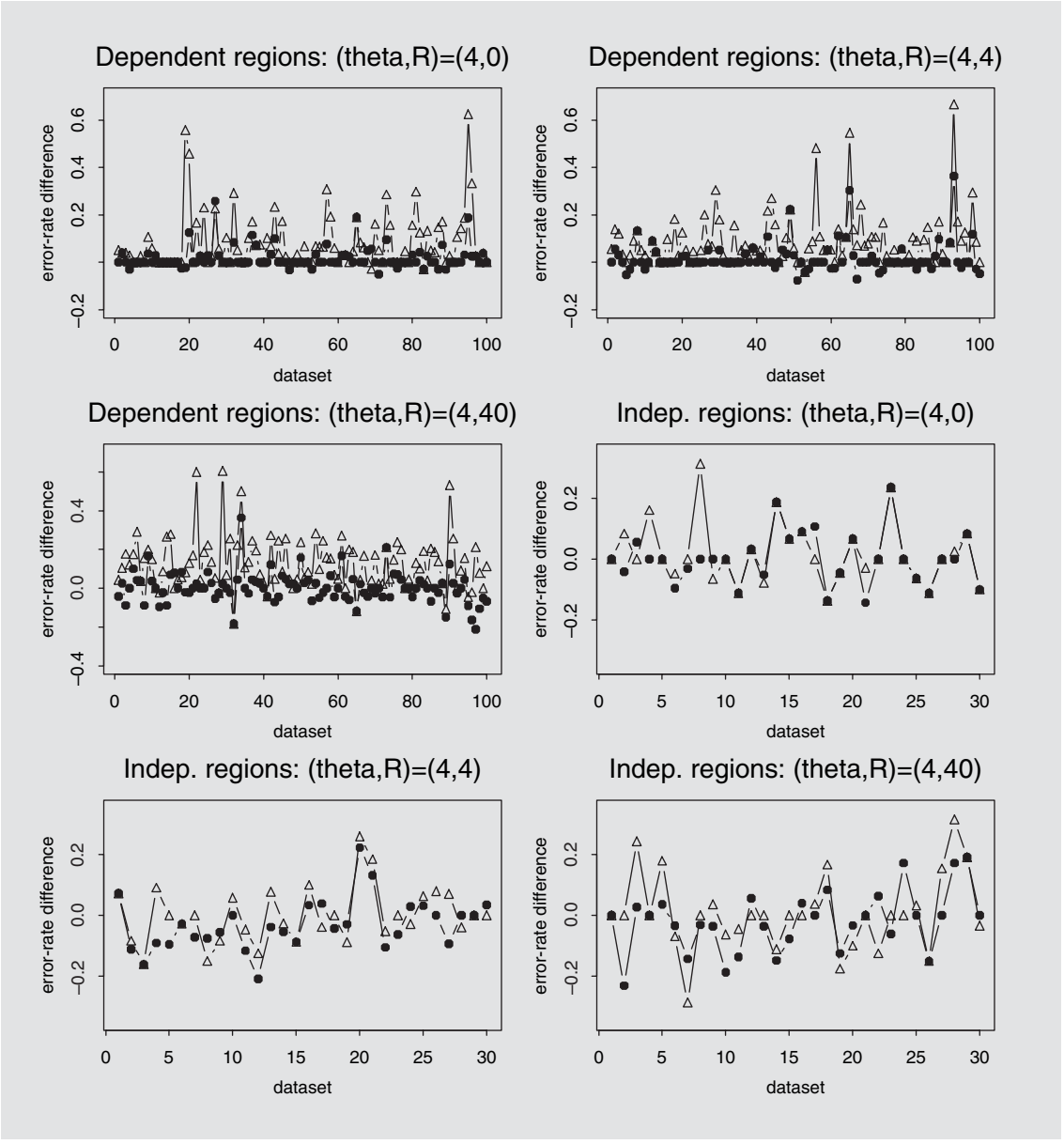

Fig. 1. The differences in error rates $\delta_A = E_A - E_{PL}$ are plotted for each data set; and denote the error rates generated by method A (either EB or MHL) and the PL method, respectively. The lower the value, the better the performance of method A compared to PL. The dotted lines represent the results obtained by application of EB (δ_{EB}), the lines marked by triangles represent the results obtained by application of MHL (δ_{MHL}). The first three plots refer to examples where the SNPs are linked. The remaining three plots represent examples where the SNPs result from two independent genetic regions, show the same mutation and recombination rates, however.

of Clark's method in this example is due to the fact that it has not taken into consideration the count information.

As a local optimisation algorithm, EM can lead to a local maximum of genotype likelihood. To demonstrate this for the OPRM1 data set, we report the log-haplotype likelihood and log-genotype likelihood values of the solutions derived from the five different haplotyping methods in table 2. These calculations show that among the five procedures, the MHL provides the maximum value of the genotype likelihood. This implies that the EM algorithm fails to give a global maximum of its objective function.

Table 2. The performances of the different methods (analysis of the OPRM1 data set)

Methods	MHL	EM	Phase	PL	EB1
HL	–2.518	–2.547	–2.518	–2.599	–2.553
GL	–2.512	–2.532	–2.590	–2.562	–2.543
Haplotypes	46	52	36	37	41

Phase is with burn-in = 10000, iteration = 100000; MHL has been combined with the block-wise EMC; EM is based on the implementation described by Hoehe et al. [2]; PL is with round = 1000. HL represents the log-haplotype-likelihood, while GL represents the log-genotype-likelihood.

In contrast, the evolutionary Markov chain Monte Carlo used in MHL tends to provide a solution for the problem of maximizing the genotype likelihood, which is closer to unknown globally optimal solutions. Therefore the application of the EM algorithm by Hoehe et al. [2] seems less favourable than the use of MHL.

ACE Data Set

This data set is composed of 11 genotypes, each with 52 non-unique varying sites [22]. The 11 genotypes defined by these bi-allelic loci were resolved into combinations of 13 distinct haplotypes, which had been determined by application of molecular genetic techniques [22]: $G_1 = (H_1, H_6)$, $G_2 = (H_1, H_1)$, $G_3 = (H_6, H_7)$, $G_4 = (H_6, H_7)$, $G_5 = (H_6, H_9)$, $G_6 = (H_1, H_7)$, $G_7 = (H_2, H_3)$, $G_8 = (H_4, H_5)$, $G_9 = (H_{10}, H_{11})$, $G_{10} = (H_1, H_8)$, and $G_{11} = (H_{12}, H_{13})$, where G_i, $1 \le i \le 11$ are the genotypes and H_j, $1 \le j \le 13$ the haplotypes. The log-haplotype likelihood of this assignment is –2.323397. This assignment could only partially be recovered by the Clark method, because, under this assignment, there were three orphan genotypes, G_8, G_9, G_{11}. Moreover, the haplotype likelihood reached the same value when we changed the true assignment by assigning any compatible haplotype pairs to these orphans. Where we randomly assigned compatible haplotype pairs to these orphans, the average number of erroneous phase calls was about 2.7, when running the Clark algorithm.

In order to apply MHL, we divided the genotype table into 5 blocks, the first four with the same width 10 and the last one with the width 12. In our EMC algorithm, we empirically set the population size $N = 20$, the highest and lowest temperatures $t_h = 0.5$ and $t_l = 0.01$, and the mutation and crossover parameters $p_m = 0.2$, $p_0 = 0.004$, $p_1 = 0.008$ and $p_2 = 0.01$. Then, we performed such a two-stage block-wise EMC algorithm 100 times. The average number of erroneous phase calls was 2.4, with a standard error of 0.19. Applying this procedure, G_1, ..., G_7, G_{10} were always correctly resolved. That is, all the erroneous phase calls resulted from G_8, G_9, and G_{11}.

Note that our empirical Bayesian method contains two components: haplotype likelihood and empirical prior. We have chosen to use an ad-hoc approach in order to demonstrate why our EB improved the accuracy of phase prediction. Naturally, we could also apply EB directly. Then, however, we would not be able to extract the specific mechanism underlying the described improvement. The ad-hoc approach involved two steps: First, dividing the genotype table into two blocks, one with width 30 and the other with width 22, and performing a two-stage MHL analysis, in which all the genotypes were resolved except for the orphan genotypes G_8, G_9 and G_{11}. In the second step, the haplotypes for these orphan genotypes were reconstructed using the empirical prior.

Note that the surface of the haplotype likelihood around the assignments for these orphan genotypes is flat. Therefore, the haplotype likelihood failed to gather information for resolving these orphan genotypes unambiguously. In this case, the following empirical structural information on these genotypes could be useful: unlike G_9 the segments of G_8 and G_{11} in two blocks are not orphans according to the haplotype likelihood. To demonstrate this, let $(H^{(1)}_{11,1}, H^{(2)}_{11,1})$ and $(H^{(1)}_{11,2}, H^{(2)}_{11,2})$ be the two partial haplotype pairs assigned by MHL to the two segments of G_{11}. Assembling these partial haplotypes, we obtained two options for assigning haplotype pairs to G_{11}: $(H^{(1)}_{11,1} \cup H^{(1)}_{11,2}, \mathrm{H}_{C1})$ and $(H^{(1)}_{11,1} \cup H^{(2)}_{11,2}, \mathrm{H}_{C2})$, where $\cup$ stands for the concatenating operator, and H_{C1} and H_{C2} denote the haplotypes complementary to $H^{(1)}_{11,1} \cup H^{(1)}_{11,2}$ and $H^{(1)}_{11,1} \cup H^{(2)}_{11,2}$, respectively. According to the scheme for specifying the prior in the Methods, we obtained the pseudocounts of $H^{(1)}_{11,1} \cup H^{(1)}_{11,2}$, H_{C1}, $H^{(1)}_{11,1} \cup H^{(2)}_{11,2}$, and H_{C2}, which were $\alpha_1 = d \times (1/2^4 + 1/2^{36} + 1/2^{35})$, $\alpha_2 = d/2^4$, $\alpha_3 = d/2^4$, and $\alpha_4 = d/2^4$, respectively. Consider any two assignments $\mathbf{H}_1$ and $\mathbf{H}_2$ that are different only on G_{11}, which $\mathbf{H}_1$ assigns $(H^{(1)}_{11,1} \cup H^{(1)}_{11,2}, H_{C1})$ to, whereas $\mathbf{H}_2$ assigns $(H^{(1)}_{11,1} \cup H^{(2)}_{11,2}, H_{C2})$. Invoking Equation (3), we have

$$\frac{p(\mathbf{H}_1|\boldsymbol{G})}{p(\mathbf{H}_2|\boldsymbol{G})} = \frac{G(1+\alpha_1)G(1+\alpha_2)G(\alpha_2)^2}{G(1+\alpha_2)^2 G(\alpha_1)G(\alpha_2)} = 1 + 1/2^{32} + 1/2^{30} > 1,$$

as being independent of the constant d. We prefer to assign $(H^{(1)}_{11,1} \cup H^{(1)}_{11,2}, H_{C1})$ to G_{11}. Note that our Bayesian approach does not help us to resolve G_8, because the pseu-

docounts of the two options mentioned in the last paragraph are the same. In summary, combination of the haplotype likelihood with the empirical Bayesian prior could resolve all genotypes in the ACE data set with a number of erroneous phase calls being less than or equal to 2. Note that the average numbers of erroneous phase calls by PL and Phase were shown by Niu et al. [17] to be 2.09 and 3.96 with standard errors of 0.033 and 0.044, respectively. EM is excluded from the comparison because it is limited with regard to the number of heterozygous loci allowable for each genotype [17]. Thus, the combination of our MHL and Bayesian approaches seems to perform slightly better than the other methods at least in this example.

Discussion

Reconstructing haplotypes from genotypes is an essential first step in the analysis of genetic variation in relation to gene function and phenotype. It will represent a particular challenge, where multiple DNA polymorphisms have been identified as the result of high-resolution resequencing studies in order to assess complete variation in genes, or genomic regions of interest, in defined populations [1, 4]. Thus, powerful methods suitable for application to various scenarios regarding the nature, pattern and organization of genetic variation (in relation to disease) will be required. Whereas scenarios characterized by strongly linked SNPs have been intensively covered in the field, scenarios where SNPs are weakly linked or physically unlinked, respectively, seem yet to require more elaboration. In our work, we have in particular focused on the second scenario, both evaluating the performances of existing methods in this context and developing approaches to cope with such a scenario in generating better results. This may be of value for the analysis of so-called 'gene-based', or candidate gene-related, haplotype structures, where we cannot necessarily rely on the existence of linkage disequilibrium between given SNPs as a precondition [1], in contrast to the decomposition into haplotype block structures [1]. This may also, less obvious at first sight, apply to the analysis of gene-gene interactions, where SNPs do not reside on the same chromosome, i.e. are evidently physically unlinked, do interact, however, with each other to confer genetic risk to complex disease. In order to elaborate on this latter issue, let us for instance assume two unphased genotype blocks, A and B, formed by two physically unlinked chromosomal segments. Let us further assume that A has two bi-allelic loci with alleles $\{a_1, A_1\}$ and $\{a_2, A_2\}$, and B two bi-allelic loci with alleles $\{b_1, B_1\}$ and $\{b_2, B_2\}$, respectively. If we phase these two blocks separately, then we obtain unrelated decompositions of the two blocks only, for example, (a_1A_2, A_1a_2) and (b_1b_2, B_1B_2), not knowing whether a_1A_2 may be coupled with b_1b_2 or B_1B_2. If the underlying haplotype a_1A_2 does in fact interact with b_1b_2, then these haplotypes should co-occur more frequently than other combinations in the haplotype space. However, testing such an interaction using unphased genotypes is difficult, particularly when several SNPs are weakly linked and when we do not know, how many blocks may exist in the data. For this reason, it does not seem very practical to identify the unknown blocks first and then phase these blocks separately. A reasonable solution seems to combine all SNPs and phase them as a total. Of course, we are at the risk of loss of efficiency, when several independent sites in the genotypic data do exist. Therefore, it is important to evaluate the robustness of both existing and proposed procedures for such a scenario.

In this work, we have shown that the robustness of one of the best currently existing haplotyping methods, PL, in fact does dissolve under such a scenario. In order to overcome this limitation, we have developed an empirical Bayesian procedure (EB) that can make a better performance under these conditions. This should also apply to EB's performance relative to Phase specifically under the condition of an invalid coalescence assumption, extrapolating from the results presented by Niu et al. [17]. We have elaborated the haplotype-likelihood framework in relation to existing methods in the literature. Consequently, we have proposed a maximum haplotype-likelihood procedure for haplotype reconstruction, termed MHL.

The MHL and Clark's methods are related in the sense that the MHL estimator can be shown to be consistent with that derived from Clark's method under the condition that all genotypes have a single count (not shown). According to our yet limited experiences, this could hold true even for the more general condition, where all the genotypes can be resolved completely and an unambiguous solution derived by Clark's method.

The procedures proposed in this paper have two advantages over Clark's and the EM methods: (1) Unlike the Clark method, we take both the multiple count information and certain structural information on genotypes into account by introduction of the haplotype-likelihood and by the specification of an empirical prior. (2) In contrast to the EM method but similar to Phase and PL, the proposed procedures allow reconstruction of haplotypes

from ambiguous genotypes constituted by many heterozygous sites by application of the block-wise evolutionary Monte Carlo algorithm. On the other hand, for ambiguous genotypes comprised of a lower number of heterozygous sites, MHL has almost the same mean error rates as EM under a coalescent model. Furthermore, we have shown that among the existing methods, Phase is the best one when the population conforms to the coalescence assumption, whereas PL and EB perform very similarly under these conditions. EB is, in contrast, more accurate than PL when some SNPs are unlinked or weakly linked. The EMC algorithm can improve the EM algorithm and Gibbs sampling because it combines the features of three different algorithms: the simulated annealing, genetic algorithm, and Metropolis algorithms. These three algorithms are designed to address the problem of approximately global optimization. Compared with the objective function used in Phase, ours are simple but still powerful. Thus, our approach might also provide a novel basis for haplotype block decomposition using unphased genotype data, where the coalescent, model-based method is not easily extended.

At last, we would like to point out that our algorithms have also been generalized to allow handling missing genotype data in some individuals at some loci. Like Niu et al. [17], we have considered three types of missing data: both alleles are missing; only the allele 0 is missing, and only the allele 1 is missing (details available on request). Although in this paper we have discussed the haplotypes consisting of bi-allelic loci only, both our methods and algorithms can easily be modified in order to cope with other types of loci, such as for instance microsatellite loci. Like Rohde and Fuerst [18], we are also able to include nuclear family information in our procedures.

To summarize, the main advantage of the proposed procedure (EB) is that it allows analysis of weakly linked or physically unlinked SNPs, respectively, and in this, improved analysis of certain candidate gene data sets or of SNP interactions that may confer genetic risk to complex disease. Extension and solidification of performance results in this additional aspect would require extensive, physically validated data sets from the same sets of individuals for a number of loci resolved at high depth, however. Such data sets are hardly accessible at present. Thus, the ultimate standard that will allow informed, sensible statistical comparisons of different methods is not yet given. Because the underlying model can greatly affect the conclusions drawn from different in silico haplotyping methods in a simulation study [17], it seems at this point preferable to apply all currently available methods. This will help to avoid the possible bias introduced by any single method when we tackle real data, as has been illustrated at the examples of some concrete genotypic data sets.

Acknowledgements

We would like to acknowledge Drs K. Rohde, T. Müller and J. Stoye, as well as Drs. W. van Zwet, M. de Gunst and F. Liang for their very helpful comments. We are grateful to Dr. M. Stephens for providing some values related to the EM and Clark methods in table 1, and to Drs. J.S. Liu, Z.S. Qin and F. Liang for sharing the software of the EM and PL methods and the code of the EMC algorithm. We are moreover grateful to the Managing Editor and three anonymous reviewers for their very constructive comments that contributed significantly to improving the presentation of material. This work has been supported in part by the Research Programme for Computational Molecular Biology, EURANDOM, and by the Federal Ministry for Education and Research (BMBF) as part of the German Human Genome Project. MRH was supported by grants 01GR0155 and 01GR0414 from the BMBF as part of the German National Genome Research Network (NGFN) Core, and by a grant from the BMBF-BioProfile Program.

Electronic-Database Information

URLs for data in this article: Nickerson Lab, http://droog.mbt.washington.edu/cvgenesnp.html (for the ACE data).

Hudson Lab, http://home.uchicago.edu/ rhudson1/source.html (for the simulation of coalescent-processes).

References

1 Hoehe MR: Haplotypes and the systematic analysis of genetic variation in genes and genomes. Pharmacogenomics 2003;4:547–570. http://www.molgen.mpg.de/~mhoehe/Pharmacogenomics_Review.pdf.

2 Hoehe MR, Köpke K, Wendel B, Rohde K, Flachmeier C, Kidd KK, Berrettini WH, Church GM: Sequence variability and candidate gene analysis in complex disease: association of μ opioid receptor gene variation with substance dependence. Hum Mol Genet 2000; 9:2895–2908.

3 Taton TA, Mirkin CA: Haplotyping by force. Nat Biotechnol 2000;18:713.

4 Hoehe MR, Timmermann B, Lehrach H: Human inter-individual DNA sequence variation in candidate genes, drug targets, the importance of haplotypes and pharmacogenomics. Curr Pharm Biotechnol 2003;4:351–378.

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

6 Clark AG: Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 1990;7:111–122.

7 Michalatos-Beloin S, Tishkoff SA, Bentley KL, Kidd KK, Ruano G: Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. Nucleic Acids Res 1996;24: 4841–4843.

8 Woolley AT, Guillemette C, Li CC, Housman DE, Lieber CM: Direct haplotyping of kilobase-size DNA using carbon nanotube probes. Nat Biotechnol 2000;18:760–763.

9 Douglas JA, Boehnke M, Gillanders E, Trent JM, Gruber TS: High-resolution haplotype structure in the human genome. Nat Genet 2001;28:361–364.

10 Mitra RD, Butty VL, Shendure J, Williams BR, Housman DE, Church GM: Digital genotyping and haplotyping with polymerase colonies. Proc Natl Acad Sci USA 2003;100:5926–5931.

11 Ding C, Cantor CR: Direct molecular haplotyping of long-range genomic DNA with M1-PCR. Proc Natl Acad Sci USA 2003;100: 7449–7453.

12 Burgtorf C, Kepper P, Hoehe M, Schmitt C, Reinhardt R, Lehrach H, Sauer S: Clone-based systematic haplotyping (CSH): A procedure for physical haplotyping of whole genomes. Genome Res 2003;13:2717–2724.

13 Gusfield D: A practical algorithm for optimal inference of haplotypes from diploid populations; in Altman R, Bailey TL, et al (eds): ISMB 2000 Proceedings, Eighth International Conference on Intelligent Systems for Molecular Biology. AAAI Press, 2000, pp 183–189.

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

15 Hawley ME, Kidd ME: HAPLO: A program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered 1995;86:409–411.

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

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

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

19 Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF: Haplotype structure and population genetic inference from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Genet 1998;63:595–612.

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

21 Zhang J, Liang F, Dassen WRM, Doevendans PA, de Gunst M: Search for haplotype interactions that influence susceptibility to type 1 diabetes, through use of unphased genotype data. Am J Hum Genet 2003;73:1385–1401.

22 Rieder MJ, Taylor SL, Clark AG, Nickerson DA: Sequence variation in the human angiotensin converting enzyme. Am J Hum Genet 1999;22:59–62.

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

24 Posada D, Crandall KA: Intraspecific gene genealogies: Trees grafting into networks. Trends Ecol Evol 2001;16:37–45.

25 Templeton AR, Boerwinkle E, Sing CF: A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in Drosophila. Genetics 1987;117:343–351.

26 Liang F, Wong WH: Evolutionary Monte Carlo: Applications to model sampling and change point problem. Statist Sinica 2000;10:317–342.