This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

Estimating haplotype frequencies by combining data from large DNA pools with database information

Dario Gasbarra*, Sangita Kulathinal*, Matti Pirinen* and Mikko J. Sillanpää

Abstract—We assume that allele frequency data have been extracted from several large DNA pools, each containing genetic material of up to hundreds of sampled individuals. Our goal is to estimate the haplotype frequencies among the sampled individuals by combining the pooled allele frequency data with prior knowledge about the set of possible haplotypes. Such prior information can be obtained, for example, from a database such as HapMap. We present a Bayesian haplotyping method for pooled DNA based on a continuous approximation of the multinomial distribution. The proposed method is applicable when the sizes of the DNA pools and/or the number of considered loci exceed the limits of several earlier methods. In the example analyses the proposed model clearly outperforms a deterministic greedy algorithm on real data from the HapMap database. With a small number of loci the performance of the proposed method is similar to that of an EM-algorithm which uses a multinormal approximation for the pooled allele frequencies, but which does not utilize prior information about the haplotypes. The method has been implemented using Matlab and the code is available upon request from the authors.

Index Terms—DNA pools, haplotype frequency estimation, HapMap database, multinomial distribution

1 Introduction

Current genetic studies consider thousands of markers simultaneously, and therefore the haplotype information is essential so that analyses can account for the linkage between the markers. Apart from laboratory techniques for experimentally-derived haplotypes [8] only genotypes can usually be determined for the sampled individuals resulting in missing haplotype information. With polymorphic (informative) markers haplotypes can be partially determined from genotype data on the consecutive generations of a pedigree [47], and the uncertain parts can then be estimated using statistical methods developed for pedigree-based haplotyping [39, 40, 44, 35, 12]. Methods are also available for haplotype frequency estimation directly from population samples of individuals [5, 28, 9, 15, 42, 32, 41]. Moreover, specific techniques like radiation hybrid mapping [2, 38], individual sperm typing [27, 30, 31, 7, 1] or utilization of DNA in megagametophytes [45, 51] have been proposed for haplotype determination.

Use of pooled DNA can substantially reduce the cost of genotyping and thus DNA pooling has been proposed for large-scale association studies [37, 3, 43, 54]. The idea is to combine equal amounts of DNA from several individuals and to analyze the allele frequencies of the whole pool in a single genotyping. Unfortunately, the reduction in genotyping costs that is achieved by pooling is accompanied by a decrease in haplotype information compared to individually genotyped data. Despite these difficulties, the use of pooled DNA has been suggested also for haplotype estimation [18, 50, 46, 13, 25, 34] and haplotype-association testing [52], as well as for fine scale mapping [22, 23]. However, these methods have strict limitations on the size of the pools and the number of loci that can be analyzed. This is because large pools and/or large numbers of loci increase the number of consistent haplotype configurations exponentially, if no prior knowledge is available to restrict this quantity. Recently, it has been illustrated that a multinormal approximation to the haplotype distribution provides a useful tool to tackle the problem of large pool size [53, 26], but the exponential increase of the possible haplotypes as a function of the number of loci still remains.

Here we introduce a way to decrease the number of haplotype configurations that are consistent with the pooled data by using external prior information about the existing haplotype structures. Currently, such information is being rapidly gathered into public databases such as HapMap [17]. The external database information is included in our model via a preprocessing step that solves the system of linear equations arising from the structures of the known haplotypes and the observed allele counts in the pools. It follows that our model is able to estimate the frequencies of only those haplotypes that are included in the database, and, in particular, is not able to identify novel haplotypes outside of the database. Thus a prerequisite for using our method is that the database provides a good coverage of the haplotype structures of the population. If, instead, the available database information were known to be incomplete, for example due to a small sample size, then it might still be necessary to do single-individual genotyping in order to gather more information about the population haplotype structures.

We work within the Bayesian framework where priors on the haplotype frequencies are chosen to represent the idea of sparsity, i.e. only a few haplotypes are assumed to have a considerable frequencies in the sample. As a computational tool, we use a Markov chain Monte Carlo algorithm to analyze the probability model that is constrained to the feasible set of haplotype configurations.

* *These authors contributed equally to this work.*

- *D. Gasbarra, S. Kulathinal, M. Pirinen, M. J. Sillanpää are with the Department of Mathematics and Statistics, University of Helsinki, FIN-00014 Helsinki, Finland. E-mail: matti.pirinen@iki.fi (Communicating author: M. Pirinen)*
- *S. Kulathinal is also with the Indic Society for Education and Development, Nashik, India*

Authorized licensed use limited to: Helsingin Yliopisto. Downloaded on May 11,2010 at 18:35:30 UTC from IEEE Xplore. Restrictions apply.
Digital Object Indentifier 10.1109/TCBB.2009.71 1545-5963/09/$26.00 © 2009 IEEE

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

It has been reported that the allele frequency measurement from the pooled data are quite accurate, and that smaller error levels of 1-2% are achievable by several technologies [37]. Other studies have also concluded that the errors in genotyping accuracy are not of major concern in small DNA pools [33, 3, 49, 36, 25]. This is because the possible allele frequencies are discrete multiples of the basic unit of $(1/2n)$, where n is the number of individuals in the pool. Thus, for small pools the possible frequency values are relatively far apart from each other. Based on these facts we assume that the accuracy of measuring allele frequencies from pooled DNA is high in the lab, but we review the possible sources of errors [19] in the Discussion.

We illustrate the proposed method using the publicly available data from the HapMap project [17]. These data provide a fine map of the single-nucleotide-polymorphisms present in the human genome, in terms of known haplotypes and their frequencies in several human populations. Although one may still question whether the information in HapMap is accurate enough, because some rare haplotypes may not have been reported at all due to the small number of genotyped individuals, the existence of such a database certainly opens up new perspectives for genetic studies (e.g. for genome-wide association analyses). Obviously, there will be more accurate and comprehensive data available in such databases in the near future, and this creates a need for methods that are able to utilize those data efficiently.

The proposed method is applicable, for example, in association studies and forensic genetics [6]. In particular, the method is well suited for case-control studies where separate pools of cases and controls are analyzed and their haplotype frequencies are compared.

2 METHODS

We are interested in estimating the frequencies of sampled haplotypes on a set of L loci residing on a narrow chromosomal interval. We assume that each locus is diallelic and denote the alleles by 0 and 1. Let the columns of an $L \times M$ matrix $H = (H_{lj})$ contain the existing M haplotypes in this region from an external haplotype database (e.g. HapMap). We assume that the database is comprehensive in the sense that it contains all the (relevant) haplotypes that are present in the studied population. Suppose that the sampled DNA originates from n individuals ($2n$ haplotypes) and is divided into O separate pools, the size of pool i being n_i individuals ($2n_i$ haplotypes). Thus $n = \sum_{i=1}^{O} n_i$. We denote by p_h^i the (unknown) proportion of haplotype $H_{\cdot h}$ (i.e. column h of matrix H) in pool i. It follows that if we let a_l^i be the observed relative frequency of the allele 1 at locus l in pool i, then $a^i = (a_1^i, \ldots, a_L^i)^\top = Hp^i$, where $p^i = (p_1^i, \ldots, p_M^i)^\top$. The problem now corresponds to solving the constrained linear system

$$Hp^i = a^i, \sum_{h=1}^{M} p_h^i = 1, \text{ for } i = 1, \ldots, O,$$

where the admissible values for the variables are

$$p_h^i \in \left\{0, \frac{1}{2n_i}, \frac{2}{2n_i}, \ldots, 1\right\}.$$

However, we relax the problem to continuous variables since it makes the linear equations more tractable and because in practice the allele frequency measurements for large DNA pools contain some errors, which makes them continuous rather than discrete. Thus, in the following we study the system

$$Hp^i = a^i, \sum_{h=1}^{M} p_h^i = 1, \; p_h^i \geq 0, i = 1, \ldots, O. \tag{1}$$

Typically the system (1) has infinitely many solutions and we are working with an ill-posed inverse problem (see [24]), where only a statistical regularization based on prior knowledge allows us to reach a reasonable solution. In our problem the key point is to incorporate into the statistical model the fact that the contents of all pools originate from the same population haplotype distribution.

2.1 Prior distribution

To derive an appropriate model we seek insight from the discrete case. If we let $m_h^i = 2n_i p_h^i$ denote the number of haplotype h in pool i then by assuming Hardy-Weinberg equilibrium among the sampled individuals we model the vectors of (integer-valued) haplotype counts $(m_1^i, \ldots, m_M^i)$ as multinomial samples from an underlying population frequency π. Furthermore, since the pools are conditionally independent of each other given π, the probability mass function for the discrete case is

$$f(p^1, ..., p^O \mid \pi) = \prod_{i=1}^{O} (2n_i)! \prod_{h=1}^{M} \frac{\pi_h^{m_h^i}}{m_h^i!}. \tag{2}$$

Because in the relaxed problem (1) the variables $m_h^i = 2n_i p_h^i$ need not be integers, we have to adjust the model accordingly. Firstly, we could extend the discrete probabilities (2) to a continuous density function by substituting the Gamma function for factorials and by introducing poolwise normalizing constants $A(2n_i, \pi)$:

$$f(p^1, ..., p^O \mid \pi) = \prod_{i=1}^{O} \frac{(2n_i)!}{A(2n_i, \pi)} \prod_{h=1}^{M} \frac{\pi_h^{2n_i p_h^i}}{\Gamma(2n_i p_h^i + 1)}. \tag{3}$$

However, it seems that the computation of normalizing constants $A(2n_i, \pi)$ is complicated and therefore we turn to another approach, but we will comment on the formula (3) in the Discussion.

The approach taken in this article is to substitute the Dirichlet distribution as an approximation of the multinomial model:

$$f(p^1, ..., p^O \mid \pi) = \prod_{i=1}^{O} \Gamma(2n_i - 1) \prod_{h=1}^{M} \frac{(p_h^i)^{(2n_i - 1)\pi_h - 1}}{\Gamma((2n_i - 1)\pi_h)}, \tag{4}$$

with the constraints $\sum_{h=1}^{M} p_h^i = 1, p_h^i \geq 0$. It has been shown [20] that (4) gives the same first and second moments and product moments of the p^i's as well as the same range of variation as (2). For details of this approximation, see pp. 285-288 in [21]. To avoid the singularities of (4) that occur when some $p_h^i \approx 0$, we introduce a threshold value ε and set all p_h^i's below ε to equal ε while evaluating (4).

Authorized licensed use limited to: Helsingin Yliopisto. Downloaded on May 11,2010 at 18:35:30 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

What remains to be specified is the prior for π. As a result of the common ancestry of the haplotypes and a possible ascertainment process, we expect that there are only a few haplotypes with considerable frequencies, while most haplotypes are rare. Thus we are seeking a sparse solution: a haplotype distribution π that has small entropy. This is reflected by a prior $\pi \sim \text{Dirichlet}(\alpha, \ldots, \alpha)$, where α itself is a hyperparameter with an improper prior $f(\alpha) \propto \alpha^{-1}$. We let the prior density of α be unbounded near the origin, since for small values of α the random distribution π is likely to have low entropy. When pool size $n_i = 1$ for all pools, we are back to the traditional haplotyping problem, i.e. reconstructing haplotypes from individually genotyped data. In such a case a similar model without a sparsity-producing prior on α, and without a continuous approximation to the multinomial distribution, has been used by Niu et al. [32].

2.2 Posterior distribution

For each pool i, consider the convex set

$$C^i = \left\{ p^i : Hp^i = a^i,\ \sum_{h=1}^{M} p_h^i = 1,\ p_h^i \geq 0 \right\}$$

containing the solutions to (1). The posterior distribution is obtained by constraining the joint prior distribution to the events $\{p^i \in C^i\}$. For example, using formula (4), the posterior density, up to a normalizing constant, is

$$\begin{aligned} &f(\alpha, \pi, p^1, ..., p^O \mid a^1, ..., a^O) \propto \\ &f(\alpha) f(\pi \mid \alpha) f(p^1, ..., p^O \mid \pi) f(a^1, ..., a^O \mid p^1, ..., p^O) \propto \\ &\alpha^{-1} \frac{\Gamma(M\alpha)}{\Gamma(\alpha)^M} \prod_{h=1}^{M} \pi_h^{\alpha-1} \ \times \\ &\prod_{i=1}^{O} \left[I(p^i \in C^i) \Gamma(2n_i - 1) \prod_{h=1}^{M} \frac{(p_h^i)^{(2n_i-1)\pi_h - 1}}{\Gamma((2n_i-1)\pi_h)} \right]. \end{aligned}$$

Here $I(\cdot)$ is the indicator function. To study this constrained distribution we use a Markov chain Monte Carlo sampling method.

2.3 Markov chain Monte Carlo sampling algorithm

Our goal is to use the geometry of the problem in constructing a well mixing proposal distribution for the Metropolis-Hastings algorithm (see [14, 4]). We assume that D, the common dimension of solution sets C^i, is at most about 20, and discuss later alternative ways to deal with higher dimensional spaces.

Initial configuration. When D is small enough ($D \leq 20$) it is possible to use a preprocessing step that identifies for each pool i set $E^i = \{q^{i,e}\}$ of all extremal points of convex set C^i using the **cdd+** algorithm [11]. Using these points we construct an initial value $\hat{\pi}$ for the underlying haplotype frequencies by taking a convex combination of all $q^{i,e},\ i = 1, \ldots, O, e \in E^i$ with weights proportional to $\exp(-b\eta(q^{i,e}))$, where $b \geq 0$ is a constant of our choice and $\eta(q) = -\sum_{h=1}^{M} q_h \log(q_h)$ denotes the entropy of haplotype distribution q. By setting b large enough we have an initial distribution with low entropy, which is in line with our prior knowledge of the haplotype frequencies.

Given $\pi = \hat{\pi}$ we initialize haplotype frequencies p^i in each pool i by taking a convex combination of the extremal points $q^{i,e} \in E^i$ with weights proportional to

$$\left((2n_i)! \prod_{h=1}^{M} \frac{\pi_h^{2n_i\, q_h^{i,e}}}{\Gamma(2n_i\, q_h^{i,e} + 1)} \right)^{c_1} \times \left(\prod_{h=1}^{M} \left(q_h^{i,e} \right)^{q_h^{i,e}} \right)^{c_2}, \qquad (5)$$

where $c_1, c_2 \geq 0$ are parameters of our choice. Thus, we use weights that combine the multinomial likelihood and the exponential of the entropy of the extremal points. We initialize hyperparameter α by setting $\hat{\alpha} = 1/M$, where M is the number of haplotypes.

Updating π. To construct a proposal $\tilde{\pi}$ we sample without replacement coordinates h_1 and h_2 from the current distribution π. For $h \notin \{h_1, h_2\}$, we set $\tilde{\pi}_h = \pi_h$, and we redistribute the probabilities between h_1 and h_2 as

$$\tilde{\pi}_{h_1} = s(\pi_{h_1} + \pi_{h_2}), \quad \tilde{\pi}_{h_2} = (1-s)(\pi_{h_1} + \pi_{h_2}),$$

where s is sampled from $\text{Beta}(d_1 + d_2\, \pi_{h_1}, d_1 + d_2\, \pi_{h_2})$ distribution. Here $d_1, d_2 \geq 0$ are tuning constants. For large values of d_2 and small values of d_1 proposal $\tilde{\pi}$ will be very close to the current value of π. The proposal is accepted or rejected according to the Metropolis-Hastings rule [4].

Updating p^i. We update the haplotype frequencies of each pool separately by proposing $\tilde{p}^i$ as a convex combination of those $D+1$ extremal points $q^{i,e} \in E^i$ that have the highest scores according to formula (5). More precisely, we set

$$\tilde{p}^i = \sum_{e=1}^{d+1} w_e^i q^{i,e},$$

where random weights w^i are sampled from $\text{Dirichlet}(d_1 + \gamma_1^i, \ldots, d_1 + \gamma_{D+1}^i)$, where γ_e^i is the value of formula (5) for the corresponding $q^{i,e}$ and current π, and d_1 is the tuning parameter introduced earlier. Again acceptance or rejection of the proposed value is determined by the Metropolis-Hastings rule.

While updating p^i we consider only $D+1$ extremal points in order to have a unique convex representation of $\tilde{p}^i$ in terms of weights w_e^i. This means that the sampling density for proposing $\tilde{p}^i$ coincides with the sampling density of the weights.

Updating α. The hyperparameter α is updated by using a lognormal random walk proposal distribution $\log(\tilde{\alpha}) \sim \mathcal{N}(\log(\alpha), \sigma^2)$, where σ is a fixed parameter.

Parameter values. In all of the examples of this article we have used the following parameter values: $b = M, c_1 = c_2 = 0.05, d_1 = 0.01, d_2 = 10$ and $\sigma = 0.5$.

2.4 A Multinormal approximation

Recently, Zhang et al. suggested a multinormal approximation to the pooled allele frequencies [53]. To see how this is derived assume that for each pool i the vector of haplotype counts is a multinomial sample from the underlying population haplotype frequencies π (see equation (2)). Now the expected values (μ_i)

Authorized licensed use limited to: Helsingin Yliopisto. Downloaded on May 11,2010 at 18:35:30 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

and variance matrices (Σ_i) of the observed allele counts ($2n_i a^i$) can be represented as

$$\begin{aligned}\mu_i &= \mathrm{E}(2n_i a^i) = 2n_i H\pi \\ \Sigma_i &= \mathrm{Var}(2n_i a^i) = 2n_i H(\mathrm{diag}(\pi) - \pi\pi^T)H^T.\end{aligned}$$

The central limit theorem states that as n_i increases the multinormal approximation $2n_i a^i \sim \mathcal{N}(\mu_i, \Sigma_i)$ becomes more accurate, and therefore the corresponding likelihood function can be used for inferences about π.

Zhang et al. used this idea in their algorithm PoooL [53], which implements a constrained EM-algorithm attempting to find the value of π that maximizes the multinomial likelihood. Shortly after the publication of the PoooL algorithm Kuk et al. [26] introduced an approximate EM-algorithm for maximizing the multinomial likelihood, and implemented it in an R-code AEM (Approximate EM-algorithm). Multinormal approximation makes these approaches less accurate than the EM-algorithms that maximize the exact multinomial likelihood [18, 50], but an advantage of the approximation is that large pool sizes do not pose computational problems anymore. Both Zhang et al. and Kuk et al. formulated their models in such a way that the matrix H is supposed to contain all 2^L possible haplotypes, and therefore the algorithms cannot handle large numbers of loci (say over 15). In our examples we have run PoooL and AEM on a data set that contains 5 loci, but they were not applicable to our larger data set (21 loci). However, we note that the multinormal approximation could also be used with an adjusted haplotype matrix that contains only a subset of the haplotypes, but we leave the implementation of this idea as a topic for future studies (see the Discussion).

2.5 A Greedy algorithm that utilizes haplotype database

The proposed method utilizes additional information from a database, which makes it applicable to larger pools and/or to larger numbers of loci compared to several existing methods that estimate haplotypes from pooled data [18, 25, 34, 53]. These differences hinder relevant comparisons between the proposed and the existing methods. Thus, in order to have an idea of the complexity of the analyzed data sets and naive frequency estimates, we consider an appropriate version of a greedy algorithm (cf. [25, 34]). This algorithm takes the poolwise allele frequency data and the list of known haplotypes as input, and proceeds as follows until all alleles have been assigned to haplotypes:

- Step 1: For each haplotype in the list, calculate how many copies of it can be formed from the available alleles in each pool, and choose the haplotype h for which the sum of possible copies over all pools is maximal.
- Step 2: Take out as many copies of h from each pool as possible and reduce the allele frequencies of the pools accordingly.
- Step 3: If no listed haplotype can be formed in any pool go to step 4, otherwise return to step 1.
- Step 4: Choose pool i which has maximal number of unresolved haplotypes, and form haplotype h by choosing at each locus the allele whose frequency in pool i is largest. Return to step 2.

3 Results

The proposed method was tested on human data extracted from the HapMap database [17]. We considered region ENm010 on chromosome band 7p15.2 of the CEU population (Utah residents with ancestry from northern and western Europe). The loci were chosen by including the first 25 SNPs of the ENm010 region such that the distances between the adjacent loci is at least 100 base pairs. This resulted in an average distance of 1062 base pairs between neighboring loci.

For the CEU population the HapMap database includes the estimated haplotype data for 30 trios (mother, father and their child) whose 60 parents were used here. The HapMap project has estimated the haplotypes by using the available pedigree information together with the state-of-the-art software PHASE [41] for haplotyping individually genotyped population samples. In the examples those 120 haplotypes represented the population whose haplotype frequencies were estimated by applying varying pooling schemes.

HapMap also contains the haplotypes of 45 Han Chinese in Beijing, China (CHB) and 45 Japanese in Tokyo, Japan (JPT). In all of our examples the haplotype list H that was used to represent our prior knowledge of the available haplotypes in the population combined the lists of the haplotypes from CEU, CHB and JPT populations. Although we sampled the pooled data only from the CEU population, the Asian samples were included in H in order to avoid unrealistically accurate prior information that would match exactly with the haplotypes that are actually present in the data.

The performance of the proposed algorithm was tested using prior (4) with the adjustment to avoid the singularities as described in section 2.1. In the examples we experimented with several values of the threshold ε and found that this had almost no effect on the estimated frequencies. Hence, we report here only the results obtained with $\varepsilon = 0.01$.

3.1 Example 1: 120 haplotypes on 21 loci

In this example we pooled the genotype data of 60 CEU parents to either 10 pools (6 individuals, 12 haplotypes per pool) or to 6 pools (10 individuals, 20 haplotypes per pool). Given the combined list of 17 existing haplotypes from CEU, CHB and JPT populations from HapMap we then estimated the underlying population frequencies of haplotypes and compared them with the original frequencies available in the HapMap database. Out of the 17 listed haplotypes 12 were actually present in the CEU population.

Of the 25 loci, 4 turned out to be redundant, i.e. produced exactly the same information as some other locus, and therefore the data were reduced to 21 loci. If no prior knowledge were available, such data could have $2^{21} = 2097152$ different haplotypes, which together with pool sizes of 6 and 10 is too large for any existing method to handle simultaneously. Thus in this example the comparisons were done only with the greedy algorithm.

Figure 1 (a) shows the true and estimated haplotype frequencies of the 12 haplotypes that were present in the CEU population arranged in the descending order of the true frequencies. Our method correctly estimated zero frequencies for those 5

Authorized licensed use limited to: Helsingin Yliopisto. Downloaded on May 11,2010 at 18:35:30 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

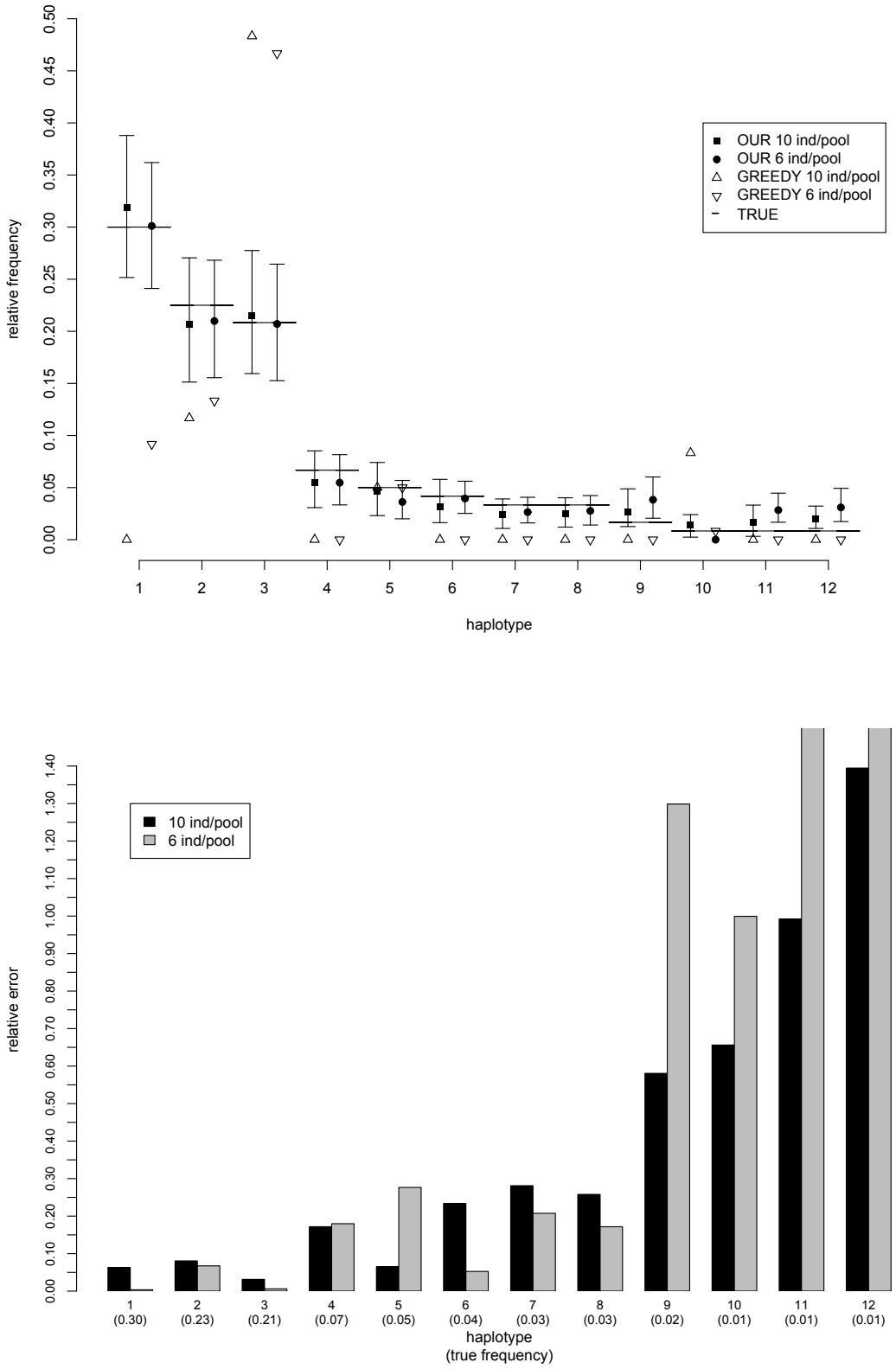

Fig. 1. HapMap data with 21 loci and 120 haplotypes. (a) Upper panel. The true haplotype frequencies (black horizontal lines) for 12 haplotypes of the CEU population and the estimates of OUR method and the greedy algorithm using two different pooling schemes. Vertical lines cover 90 % probability regions of our estimates. (b) Lower panel. Relative errors of frequency estimates. The y-axis is cut at 1.50.

Authorized licensed use limited to: Helsingin Yliopisto. Downloaded on May 11,2010 at 18:35:30 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

haplotypes that were included in the haplotype list but were not present in the data, and therefore they are not included in Figure 1. The vertical lines around the point estimates (here posterior means) describe 90% posterior probability regions, and thus reflect the uncertainty related to the estimates. The results were confirmed by executing several different MCMC runs, each of length 20,000 iterations. All runs produced similar results suggesting that the chains had converged. It seems that in this setting the pool size does not affect the estimates considerably. However, a careful examination verifies what one would expect: the smaller pool size results in a slightly improved accuracy, both in terms of the point estimates and the probability intervals. This can be seen also in Figure 1 (b) that shows the absolute values of the relative errors in estimating the haplotype frequencies using different pool sizes. The relative errors are calculated as the ratio of the absolute difference between the true and the estimated frequency to the true frequency. For the common haplotypes (frequency $\geq 10\%$) the relative errors of the proposed method stays below 10%. Note that the y-axis in Figure 1 (b) has been restricted to extend only to 150% in order to see the details of the relative errors of the major haplotypes. In any case it is likely that by using large DNA pools we cannot estimate accurately the rare haplotypes whose relative frequencies in the population are only a few per cent or less.

The greedy algorithm was able to identify 4 and 5 out of the 12 correct haplotypes for pool sizes of 10 and 6 individuals, respectively, but its frequency estimates are unacceptably different from the true frequencies for all practical purposes (Figure 1 (a)). Hence, we conclude that it is not trivial to resolve these data accurately into haplotypes.

3.2 Example 2: 1000 haplotypes on 5 loci

Our method is able to handle also much larger DNA pools than were considered in Example 1. In this example we sampled 1000 haplotypes (with replacement) from the above-described HapMap data set of 120 haplotypes of the CEU population. We divided the sampled haplotypes to 10 pools of 50 individuals (100 haplotypes) in each, and to 20 pools of 25 individuals (50 haplotypes) in each. We estimated the population haplotype frequencies by our method using the combined list of haplotypes from CEU, CHB and JPT populations as our H matrix. In addition, we ran the same data sets with the programs PoooL [53] and AEM [26]. Because PoooL was not able to handle more than 5 loci on these data, we restricted the considerations to the first 5 loci of the region. As a result list H contained 7 haplotypes of which 6 were present in the CEU population. Note that AEM and PoooL are not able to use the predefined haplotype list, but instead assume that all 32 haplotypes may exist in the population.

The frequency estimates are given in Figure 2 (a). Our method and AEM performs quite similarly, except for haplotype 4 for which our estimates are less accurate than those of AEM. The reason may be that our method estimated that the additional haplotype, that was included in H from the Asian populations, had the relative frequency of 1.8% and 2.3% in the pooling schemes of 50 and 25 individuals per pool, respectively. Also our 90% probability region (vertical lines) is relatively far from the true value for haplotype 4, whereas for haplotypes 1, 2, 5 and 6 the corresponding regions cover the true value.

These results also suggest that PoooL does not achieve as high accuracy as the other two methods. This is in line with the comparisons between PoooL and AEM that were carried out in [26].

4 Discussion

We have presented a Bayesian approach to estimate haplotype frequencies from large pools of DNA samples jointly for each pool and for the underlying population. The novelty of our approach lies in the incorporation of the database information, which can significantly reduce the number of relevant haplotypes. This becomes evident from our two examples where the numbers of all possible haplotypes, $2^5 = 32$ and $2^{21} = 2097152$, were reduced to 7 and 17, respectively. On the other hand, a successful application of our method requires that the database provides accurate information about the haplotype structures that are present in the population. This is because our method can be used only for estimating the frequencies of the haplotypes in the database, but not for identifying novel haplotype structures outside of the database.

A very useful property of the Bayesian model is that it can quantify the uncertainty related to the estimates by using the posterior samples of the frequencies generated by the MCMC algorithm. For example, the 90% probability regions in Example 1 are wider than those of Example 2, reflecting the differences in the sizes of the data sets and the numbers of putative haplotypes.

Pool size. Several existing methods for haplotype inference from pooled DNA data have strict limitations on the sizes of pools that can be analyzed. For example, the maximum possible pool size in the program HaploPool [25] is 3 and the program LDPooled [18] failed to work (with 2 GB of memory) on the HapMap data set used in this article even for 10 loci when the size of the pools was larger than 2. An extension of the PHASE algorithm to the pooled data is technically applicable to larger pools, but the accuracy seems to decrease considerably as the pool size increases [34]. These limitations are unfortunate as significant savings in genotyping could be achieved only when one were able to extract accurately haplotype information from larger DNA pools (say, pools with over 10 individuals). In our framework the sizes of pools do not pose problems because we consider a system of linear equations whose size is determined by the number of loci. Thus, we need not enumerate all possible haplotype combinations in the pools as is required, for example, by LDPooled [18]. In this article we have tested our algorithm with pool sizes of up to 50 individuals (100 haplotypes). Recently, another way to circumvent the problem of large pool sizes was introduced by Zhang et al. in the program PoooL [53] and modified by Kuk et al. in the program AEM [26]. A topic for future studies is to combine such a multinormal approximation with external database information and to compare the resulting method with the one presented here.

Number of loci. The number of possible haplotypes grows exponentially with respect to the number of loci if no prior

Authorized licensed use limited to: Helsingin Yliopisto. Downloaded on May 11,2010 at 18:35:30 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

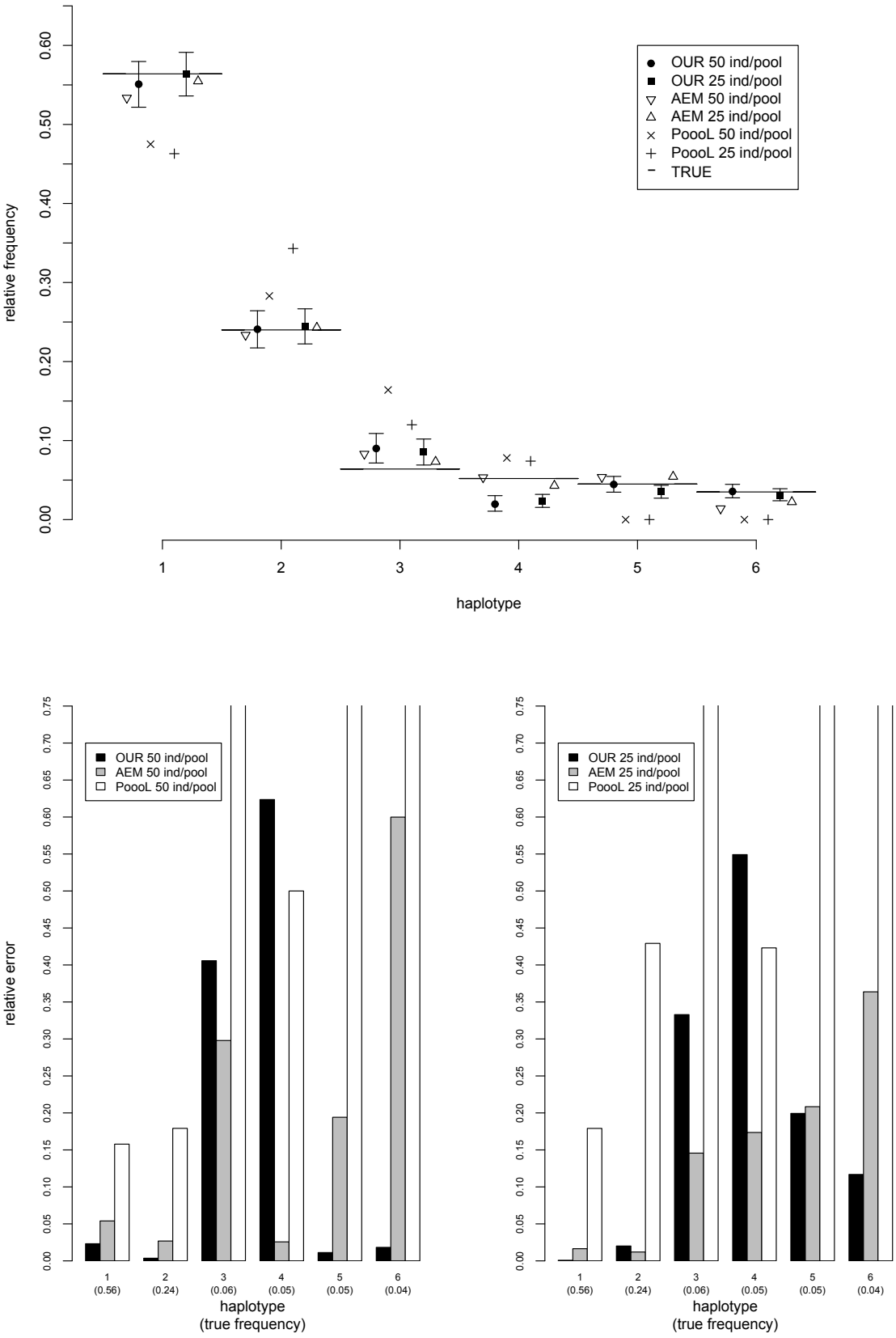

Fig. 2. HapMap data with 5 loci and 1000 haplotypes. (a) Upper panel. The true haplotype frequencies (black horizontal lines) for 6 haplotypes of the CEU population and the estimates of OUR method, PoooL and AEM using two different pooling schemes. Vertical lines cover 90% probability regions of our estimates. (b) Lower panel. Relative errors of frequency estimates. The y-axis is cut at 0.75.

Authorized licensed use limited to: Helsingin Yliopisto. Downloaded on May 11,2010 at 18:35:30 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

information is available to restrict the set of potential haplotypes. This is one of the reasons why the previous methods for haplotyping pooled DNA data can handle only a few loci at a time (at most about 15 loci). Such local haplotypes would then need to be combined to span longer chromosomic regions by some external procedure. Of the available methods the HaploPool program [25] includes such a procedure. It estimates the haplotype frequencies of up to four-loci haplotypes at a time by using an EM-algorithm, and then combines the results using linear regression. In our approach the number of loci does not pose direct limitations for the applicability of the method, since the number of possible haplotypes is determined by a list which is extracted from an external database. However, there are reasons why we will anyway restrict the considerations to relatively short chromosomal segments. Firstly, dimension D of the convex set of solutions of linear system (1), should be at most about 20 in order that the extremal solutions could be enumerated using the current approach (cdd+ program). This means that haplotype matrix H must satisfy the condition: $M-\text{rank}(H) \leq 20$, where M is the number of columns of H (i.e. the number of listed haplotypes). If the external list contains too many haplotypes to be considered simultaneously one could adopt a stepwise approach, where at each step the haplotype frequencies are estimated only for a short region, and after each step all the haplotypes whose partial content is assigned only a small frequency are deleted from the list. By controlling the size of the blocks, the length of the haplotype list can be restricted to satisfy the given constraints. A similar idea called partition-ligation has been used earlier in individual-level haplotyping [32].

Another reason for considering only narrow segments is that we assume *a priori* that the haplotype distribution has a low entropy. This assumption is more likely to hold for relatively short haplotype segments than for longer regions that contain more variability. It may also be that the haplotype information stored in databases provides a good representation of the population haplotype frequencies only locally, for short segments between recombination hotspots, whereas on a larger scale, recombinations constantly shuffle the genetic material in the population. Thus, it may not be justified to use the database information to build systems of linear equations on a large number of loci spanning long chromosomic regions. However, if the available marker map is tightly linked, then a relatively narrow region may already contain many SNPs, and therefore the number of loci that can be considered simultaneously by our method can be substantially higher than in the previous methods. For example, in this article we considered a chromosomal segment that contained 21 loci and spanned over 25 kbp of the human genome.

Errors in pooling designs. In our model we assumed that there are no errors related to the allele frequency measurements from DNA pools. Even though it has been reported that allele frequencies can be measured quite accurately from DNA pools [37], in reality at least three types of errors may occur. There may be errors in pool formation (e.g. unequal amounts of DNA from different individuals), in allele amplification (one allele amplifies more efficiently than the other), or in allele frequency measurements (instrumental measurement errors). In a recent study Jawaid and Sham concluded that the differential allelic amplification is the most important contributor to the error variance in absolute allele frequency estimation, but that it can be handled efficiently by adjusting for the measurements for differential amplification [19]. Whether these sources of errors have a notable effect on haplotype estimation from pooled DNA data is left as a topic for future studies.

Model and algorithm. If a comprehensive list of haplotypes is available then the frequencies of haplotypes in the pools satisfy the set of linear equations (1). In some cases it might be possible to use the available database information to identify such a set of loci that would yield a system of linear equations with a unique solution, but we leave this as a topic for further studies. In the examples that we have described, the corresponding linear systems did not yield unique solutions and a statistical model was needed to extract reasonable frequency estimates.

Our approach utilizes a preprocessing step where the extremal points of the solution sets are computed. This has an advantage that the systems of equations (1) need to be solved only once. Because of the convexity of the feasible sets, the extremal points can be used to characterize all the solutions. On the other hand, this approach takes us to the continuous domain even though the original problem is about discrete haplotype counts.

Passing from discrete variables to continuous ones also poses some challenges from the modeling point of view. The appropriate model for the integral haplotype counts would be a multinomial one, and here we have made efforts to extend this idea into the continuous domain. We introduced two candidates to carry out the task: a straightforward generalization (3) and a continuous approximation (4). Unfortunately, both of them have their own shortcomings. For (3) the computation of normalizing constants $A(2n_i, \pi)$ would be needed in the MCMC algorithm when calculating the Metropolis-Hastings ratios. However, we do not know any effective way to carry out that calculation. Completely ignoring A's could technically be justified by introducing an extra term into the prior of π. It seems that such a shortcut has the largest effect when some frequency π_h is small. Because we expect a sparse solution for π, it seems that it would be beneficial to be able to account for A's.

To overcome the problem with the normalizing constant, we turned to a continuous Dirichlet approximation (4) given by [20]. It seems that the largest discrepancies between (4) and (3) appear when some frequencies p_h^i are small. In particular, when some p_h^i is zero the density (4) either vanishes or goes to infinity. To smooth the behavior of (4) for small frequencies we used a positive threshold value below which we did not allow the frequencies to fall, when calculating the Metropolis-Hastings ratio.

The correlated recombination history and the ascertainment process both reduce the number of distinct haplotypes in the sample. To partly account for this, we have used a sparsity-producing prior for Dirichlet hyperparameters to reduce the number of haplotypes with non-zero estimated frequencies. Similar sparsity-producing priors have earlier been used e.g. with the normal distribution (see [48, 10, 16]). Another possi-

Authorized licensed use limited to: Helsingin Yliopisto. Downloaded on May 11,2010 at 18:35:30 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

bility is to adopt a more appropriate prior for the haplotype frequencies based on population genetics (e.g. the one used in [41]). In fact, for the case where no haplotype list exists, we have presented an extension of the PHASE haplotyping algorithm [41] to pooled DNA data [34].

In this article the database information was used only for identifying the set of possible haplotypes, but not for defining the prior distribution for the haplotype frequencies. This is because we expect that the haplotype frequencies in the ascertained sample of individuals may vary significantly from the haplotype frequencies of the general population that is included in the database. If more information about the haplotype frequencies among the sampled group were available then we could include that into the prior of π by using an asymmetric Dirichlet-distribution that better reflects the available prior knowledge.

Future perspectives. A problem for future research is to develop an algorithm to deal with higher dimensional situations, where we may have hundreds of haplotypes in the list. In such cases it is not feasible to enumerate all extremal points of the convex sets C^i, and it is necessary to use some other strategy for building a proposal distribution. The hit-and-run proposal (see for example [29]) might be a good candidate.

Another way to extend this work is to combine the external database information with a computationally efficient multinormal approximation [53, 26]. Because of computational advantages provided by such a model it might be possible to take into account uncertainty related to the haplotype list by allowing non-zero frequencies also for the haplotypes that are not included in the database. As long as the available database information is based on a relatively small number of sampled individuals such an extension would be an important contribution to the field of haplotype estimation.

One possible application of this methodology is in case-control studies, where cases and controls would be analyzed in separate pools. When the population haplotype frequencies are known, the haplotype frequencies among cases and controls can be estimated jointly by using the fact that the whole population is a mixture of cases and controls with the disease proportion as the mixing parameter. Naturally, once the frequencies are estimated for different groups, the existing methods for analyzing individual-level data can be applied. These include, for example, haplotype-based association analysis methods and epistatic models.

FUNDING

Academy of Finland (114786, 122883, 202324) and ComBi graduate school.

ACKNOWLEDGMENTS

We are grateful to Komei Fukuda for distributing the program **cdd+**, to Han Zhang for distributing the program PoooL and to Anthony Kuk for distributing the R-codes of the AEM algorithm. We would like to thank the three anonymous reviewers for their constructive comments which resulted in considerable improvements to the article. We are also thankful to Dr. Bijoy Joseph from the Indic Society for Education and Development, India for his help with the English language.

REFERENCES

[1] N Arnheim, P Calabrese, and M Nordborg. Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved. *Am J Hum Genet*, 73:5–16, 2003.

[2] M Boehnke, K Lange, and D Cox. Statistical methods for multipoint radiation hybrid mapping. *Am J Hum Genet*, 49:1174–1188, 1991.

[3] L M Butcher, E Meaburn, L Liu, C Fernandez, L Hill, A AL-Chalabi, R Plomin, L Schalkwyk, and I W Craig. Genotyping pooled DNA on microarrays: a systematic genome screen of thousands of SNPs in large samples to detect QTLs for complex traits. *Behav Genet*, 34:549–555, 2004.

[4] S Chib and E Greenberg. Understanding the Metropolis-Hastings algorithm. *Am Stat*, 49:327–335, 1995.

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

[6] R G Cowell, S L Lauritzen, and J Mortera. Identification and separation of DNA mixtures using peak area information. *Forensic Science International*, 166:28–34, 2007.

[7] M Cullen, S P Perfetto, W Klitz, G Nelson, and M Garrington. High-resolution patterns of meiotic recombination across the human major histocompatibility complex. *Am J Hum Genet*, 71:759–776, 2002.

[8] J A Douglas, M Boehnke, E Gillanders, J M Trent, and S B Gruber. Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibirium studies. *Nat Genet*, 28:361–364, 2001.

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

[10] M A T Figueiredo. Adaptive sparseness for supervised learning. *IEEE Trans Pattern Anal Mach Intell*, 25:1150–1159, 2003.

[11] K Fukuda. cdd/cdd+ Reference manual. *Available at ftp://ftp.ifor.math.ethz.ch/pub/fukuda/cdd/cddman/cddman.html*, 1999.

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

[13] D Gasbarra and M J Sillanpää. Constructing parental linkage phase and genetic map over distances < 1 cM using pooled haploid DNA. *Genetics*, 172:1325–1335, 2006.

[14] W K Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[15] M E Hawley and K K Kidd. Haplo: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J Hered*, 86:409–411, 1995.

[16] F Hoti and M J Sillanpää. Bayesian mapping of genotype x expression interactions in quantitative and qualitative traits. *Heredity*, 97:4–18, 2006.

[17] International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449:851–861, 2007.

Authorized licensed use limited to: Helsingin Yliopisto. Downloaded on May 11,2010 at 18:35:30 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

[18] T Ito, S Chiku, E Inoue, M Tomita, T Morisaki, H Morisaki, and N Kamatani. Estimation of haplotype frequencies, linkage-disequilibrium measures, and combination of haplotype copies in each pool by use of pooled DNA data. *Am J Hum Genet*, 72:384–398, 2003.

[19] A Jawaid and P Sham. Impact and quantification of the sources of error in DNA pooling designs. *Ann Hum Genet*, 73:118–124, 2009.

[20] N L Johnson. An approximation to the multinomial distribution: some properties and applications. *Biometrika*, 47:93–102, 1960.

[21] N L Johnson and S Kotz. *Distributions in Statistics - Discrete distributions.* Houghton Mifflin Company, U.S.A., 1969.

[22] T Johnson. Multipoint linkage disequilibrium mapping using multilocus allele frequency data. *Ann Hum Genet*, 69:474–497, 2005.

[23] T Johnson. Bayesian method for gene detection and mapping using case and control design and DNA pooling. *Biostatistics*, 8:546–565, 2007.

[24] J Kaipio and E Somersalo. *Statistical and Computational Inverse Problems, Applied Mathematical Series, vol. 160.* Springer, Berlin, 2004.

[25] B Kirkpatrick, C S Armendariz, R M Karp, and E Halperin. Haplopool: improving haplotype frequency estimation through DNA pools and phylogenetic modeling. *Bioinformatics*, 23:3048–3055, 2007.

[26] A Y C Kuk, H Zhang, and Y Yang. Computationally feasible estimation of haplotype frequencies from grouped DNA with and without Hardy-Weinberg equilibrium. *Bioinformatics*, 25:379–386, 2009.

[27] L C Lazzeroni, N Arnheim, K Schmitt, and K Lange. Multipoint mapping calculations for sperm-typing data. *Am J Hum Genet*, 55:431–436, 1994.

[28] J C Long, R C Williams, and M Urbanek. An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am J Hum Genet*, 56:799–810, 1995.

[29] L Lovász and S Vempala. Hit-and-run from a corner. *Siam J Comput*, 35:985–1005, 2006.

[30] W Navidi and N Arnheim. Analysis of genetic data from the polymerase chain reaction. *Stat Sci*, 9:320–333, 1994.

[31] W Navidi and N Arnheim. Combining data from polymerase chain reaction DNA typing experiments: application to sperm typing data. *J Am Stat Assoc*, 94:726–733, 1999.

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

[33] N Norton, N M Williams, H J Williams, G Spurlock, G Kirov, D W Morris, B Hoogendoorn, M J Owen, and M C O'Donovan. Universal, robust, highly quantitative SNP allele frequency measurement in DNA pools. *Hum Genet*, 110:471–478, 2002.

[34] M Pirinen, S Kulathinal, D Gasbarra, and M J Sillanpää. Estimating population haplotype frequencies from pooled DNA samples using PHASE algorithm. *Genet Res*, 90:509–524, 2008.

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

[36] S R E Quade, R C Elston, and K A B Goddard. Estimating haplotype frequencies in pooled DNA samples when there is genotyping error. *BMC Genetics*, 6:25, 2005.

[37] P Sham, J S Bader, I Craig, M O'Donovan, and M Owen. DNA pooling: a tool for large-scale association studies. *Nat Rev Genet*, 3:862–871, 2002.

[38] D Slonim, L Kruglyak, L Stein, and E Lander. Building human genome maps with radiation hybrids. *J Comput Biol*, 4:487–504, 1997.

[39] E Sobel and K Lange. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am J Hum Genet*, 58:1323–1337, 1996.

[40] E Sobel, K Lange, J R O'Connell, and D E Weeks. Haplotyping algorithms. In T P Speed and M S Waterman, editors, *Genetic Mapping and DNA Sequencing, IMA Volume 81 in Mathematics and its applications*, pages 89–110. Springer-Verlag, New York, 1996.

[41] M Stephens and P Scheet. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet*, 76:449–462, 2005.

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

[43] G Tamiya, M Shinya, T Imanishi, T Ikuta, S Makino, and et al. Whole genome association study of rheumatoid arthritis using 27 039 microsatellites. *Hum Mol Genet*, 14:2305–2321, 2005.

[44] Ghosh S Tapadar, P and P P Majumder. Haplotyping in pedigrees via a genetic algorithm. *Hum Hered*, 50:43–56, 2000.

[45] L K Tulsieram, J C Glaubitz, G Kiss, and J E Carlson. Single tree genetic linkage mapping in conifers using haploid DNA from megagametophytes. *Bio/Technology*, 10:686–690, 1992.

[46] S Wang, KK Kidd, and H Zhao. On the use of DNA pooling to estimate haplotype frequencies. *Genet Epidemiol*, 24:74–82, 2003.

[47] E Wijsman. A deductive method of haplotype analysis in pedigrees. *Am J Hum Genet*, 41:356–373, 1987.

[48] S Xu. Estimating polygenic effects using markers of the entire genome. *Genetics*, 163:789–801, 2003.

[49] H-C Yang, C-C Pan, R C Y Lu, and C S J Fann. New adjustment factors and sample size calculation in DNA-pooling experiment with preferential amplification. *Genetics*, 169:399–410, 2005.

[50] Y Yang, J Zhang, J Hoh, F Matsuda, P Xu, M Lathrop, and J Ott. Efficiency of single-nucleotide polymorphism haplotype estimation from pooled DNA. *Proc Natl Acad Sci USA*, 100:7225–7230, 2003.

[51] R F Yazdani, C Yeh, and J Rimsha. Genomic mapping of *pinus sylvestris* (L.) using random amplified polymorphic DNA markers. *For Genet*, 4:209–215, 1995.

[52] D Zeng and D Y Lin. Estimating haplotype-disease associations with pooled genotype data. *Genet Epidemiol*, 28:70–82, 2005.

[53] H Zhang, H C Yang, and Y Yang. Poool: an efficient method for estimating haplotype frequencies from large DNA pools. *Bioinformatics*, 24:1942–1948, 2008.

Authorized licensed use limited to: Helsingin Yliopisto. Downloaded on May 11,2010 at 18:35:30 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

[54] Y Zhao and S Wang. Optimal DNA pooling-based two-stage designs in case-control association studies. *Hum Hered*, 67:46–56, 2009.

AUTHORS

Dario Gasbarra received his M.Sc. degree in mathematics from the University of Rome "La Sapienza" in 1991, and his Ph.D. degree in applied mathematics from the Univerisity of Oulu, Finland, in 1998. Currently he works as a university lecturer in stochastics at the Department of Mathematics and Statistics, University of Helsinki, Finland. His research interests are bioinformatics, mathematical and computational statistics and stochastic analysis.

Sangita Kulathinal received her Ph.D. degree in statistics from the University of Pune, India, in 1996 and then worked as a postdoctoral researcher at the University of Helsinki, Finland. At present she is working at the National Institute for Health and Welfare, Finland, and has an honorary position at the Indic Society for Education and Development, India. Her research interests include parametric and non-parametric statistical inference, and design and analysis of case-cohort studies and related designs.

Matti Pirinen received his M.Sc. degree in mathematics from the University of Helsinki, Finland, in 2004 and his Ph.D. degree in statistics from the same university in 2009. Currently he works as a postdoctoral researcher at the Wellcome Trust Centre for Human Genetics, Oxford, UK. His research interests include Bayesian statistics, mathematical and statistical genetics and related computational aspects.

Mikko J. Sillanpää received his M.Sc. degree in applied mathematics from the University of Jyväskylä, Finland, in 1992 and his Ph.D. degree in biometry from the University of Helsinki, Finland, in 2000. Currently he works as a university lecturer at the Department of Animal Science and as an adjunct professor in the Department of Mathematics and Statistics, University of Helsinki. He is also the responsible researcher of project "Computationally practical statistical methods for genomewide association studies, genetic biomarker identification, and data integration" funded by the Academy of Finland (2009-2012) and University of Helsinki's Research Funds (2009-2011). His current research areas include statistical genomics, bioinformatics and population genetics.

Authorized licensed use limited to: Helsingin Yliopisto. Downloaded on May 11,2010 at 18:35:30 UTC from IEEE Xplore. Restrictions apply.