Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2022;8(3):281-294

Original Article

Non-Parametric MCMC Gibbs Sampler Approach and Misclassification Assessment of Estimating Haplotype Frequencies among Related Statistical Approaches

Gie Ken-Dror[1*], Pankaj Sharma[1,2]

[1]Institute of Cardiovascular Research, Royal Holloway University of London (ICR2UL), UK.
[2]Department of Clinical Neuroscience, Imperial College Healthcare NHS Trust, London, UK.

ARTICLE INFO

Received 28.03.2022
Revised 01.06.2022
Accepted 25.06.2022
Published 15.10.2022

Key words:
Haplotype Reconstruction;
Single Nucleotide Polymorphisms;
Markov chain Monte Carlo;
Gibbs Sampler Algorithm;
Misclassification Bias

ABSTRACT

Introduction: Haplotype analysis allows higher resolution analysis in genetic association studies and is used as a reference panel for genotype imputation in genome-wide association studies. Haplotypes estimates from genotypes among unrelated individuals but misclassification of the haplotype reconstruction will directly affect the accuracy of the results.

Methods: This study proposes a novel statistical method Gibbs sampler algorithm to estimate haplotype frequency and quantify the influence of misclassification bias of the estimate haplotype. The performance of the algorithm is evaluated on simulated datasets assuming that linkage phase unknown. The simulation used different minor allele frequencies at each single nucleotide polymorphism (SNP) and different linkage-disequilibrium between the SNPs.

Results: The Gibbs sampler algorithm presents higher accuracy among over seven SNPs or less, validated, and deals with missing genotype compared to previous related statistical approaches. Misclassification of estimated haplotypes leads to non-differential bias in exposure and affects haplotype estimates in haplotype analysis. The observed odds ratio underestimates the association between haplotype and phenotype by 36% to 99%.

Conclusion: The Gibbs sampler algorithm provides higher accuracy and robust effectiveness performance, handles missing genotypes and provides uncertain probabilities of haplotype frequencies. The misclassification bias of the estimate haplotype underestimates the genetic association by more than forty percent.

*.Corresponding Author: Gie.KenDror@rhul.ac.uk

Introduction

Haplotype analysis allows higher resolution analysis among genetic association studies (GAS) compared to the genotype analysis.[1] In addition, the haplotype may be used as a reference panel for genotype imputation among genome wide association studies (GWAS).[2] However, the use of haplotypes is not straight forward because haplotypes are not directly measured in the laboratory and haplotypes are estimated based on the genotype of unrelated individuals. Misclassification of haplotype reconstruction will directly affect the accuracy of further analysis results such as linkage-disequilibrium (LD) across genomic regions, inferred population history, fine scale mapping correlation between alleles at closely link loci and multiple markers in candidate genes.[3]

The most difficult aspect of reconstructing haplotypes in unrelated individuals is haplotypic uncertainty which occurs when two or more markers are heterozygous and their genetic phase is unknown.[1] These factors can have a large impact causing haplotypes to be systematically missed. The parsimony and phylogeny methods are rule based approaches that seek for an optimal set of haplotypes that satisfy specific rules. The parsimony rules maximise the genotype resolution while reducing the number of haplotypes. The phylogeny rules give a set of genotypes and find a set of explaining haplotypes that defines a perfect phylogeny. The Expectation Maximization (EM) algorithm and the Markova Chine Monte Carlo (MCMC) algorithm are two probability-based approaches that based on calculate probability of haplotypes conditional on genotypes.[4]

Several statistical approaches to estimate haplotype reconstruction from unrelated individual have been proposed including expectation-maximization (EM)[5] using an efficient iterative maximum likelihood approach (haplo.em, R package and PLINK software),[6,7] Bayesian statistical approach estimation using a Metropolis-Hastings Markov chain Monte Carlo (PHASE)[8], hidden Markov model (HMM) based on cluster or templates (fastPHASE, MACH1, IMPUTE2, BEAGLE).[9-12]

The aim of this study is to propose a novel statistics algorithm approach for haplotype reconstruction and compare the results with those obtained from the most commonly used statistical approaches in the literature: R [7], PLINK [6], fastPHASE [9], PHASE [8], MACH1 [10], IMPUTE2 [11], and BEAGLE [12]. In addition, we examine the accuracy of the various methods for estimating population haplotype frequencies and quantify the influence of the misclassification bias of estimated haplotypes in genetic association analysis.

Methods

The simulated datasets, estimation algorithm and statistical analysis have been implemented in the R statistical software system version 4.0.2 [13], on a 32-bit computer with 2.00 GB of random access memory and an Intel(R) Core(TM)2 Duo central processing unit (CPU) with 2.00GHz processor.

Simulation of genotype and haplotype datasets

The simulation starts to create random diallelic (A, major allele; B, minor allele) SNPs for 100 individuals with varying minor allele

frequencies (MAF) ranging from 1%, to 50% and pair-wise LD simulating the r^2 measure with values ranging from low (r^2=0.01) to high (r^2=0.9) correlation between the SNPs. To represent studies of unrelated individuals assumed that the haplotypic phase was unknown and the statistical approaches: R-EM [7], PLINK [6], fastPHASE [9], PHASE [8], MACH1 [10], IMPUTE2 [11], BEAGLE [12], and Gibbs were used to infer the haplotypes. The results obtained after the use of the statistical approaches were compared to results based on the known phase of the haplotypes. Thousand datasets were generated and analysed assuming differing number of SNPs. Finally, a reality check was run on the simulated blood dataset as would be done for real data, search for samples with observed one of the SNPs is monomorphic and identical among all samples. In this case, the SNPs is reset and simulate again to have heterozygous SNPs as would likely occur when processing clinical samples.

Novel haplotype reconstruction methods

The Markov chain Monte Carlo (MCMC) non-parametric Gibbs sampler

The Gibbs sampler also known as the Glauber dynamics or the heat-bath algorithm, is a leading MCMC method for obtaining a sequence of observations which are approximated from a specified multivariate probability distribution, when direct sampling is difficult.[14] The Gibbs sampling algorithm generates a new sample from the distribution of each variable based upon the conditional distribution among the current values of the other variable.[14-16] The Gibbs sampler is a popular MCMC algorithm and is widely used in phylogenetic analysis, sequence motif discovery and haplotype estimation. The non-parametric Gibbs sampler does not base the sampling of the artificial samples (Dirichlet distribution, Poisson distribution, or Gamma distribution) on the assumed model of the marginal distribution of genotype observations but uses instead a non-parametric description of the joint distribution, the empirical distribution. The statistical method is an optimization algorithm, iterative, starting with an initial value that belongs to the parameter space. At each step of the algorithm (Table 1), a new parameter value is selected, hopefully a value closer to the final target value than the previous one. The algorithm stops when it converges, namely when the new value is very close to the current value. The value of the parameter that was selected when the algorithm converged is declared to be the maximiser.

The algorithm consists of several steps:

The observed data consists of SNP genotypes (homozygote major/minor, heterozygote) at several genetic loci. *n* represent the number of blood sample (sample size), *i* is an index used to refer to an individual blood sample (i=1,…,n), *j* is an index used to refer to a unique haplotype combination within a blood sample i, $h_{(i,j)}$ is a set of haplotypes, s is the number of SNPs genotyped, *z* is the number of potential haplotypes in the population (2^s). k is the number of iterations. G is a vector of genotype group for each patient, $G = (g_1,\ldots,g_n)$. H is a vector of haplotype sets, $H = (H_1,\ldots,H^z)$ and θ is a vector of estimated haplotype frequencies, $\theta = (\theta_1,\ldots,\theta^z)$. The Gi is the number of patients with genotype group i, $H_{(i,j)}$ is the count of hapltypes in $h_{(i,j)}$, $\Sigma_{i\epsilon Gi}$ is summation of all individuals *i* that are in genotype group G_i.

Table 1. The algorithm step

Step	The Algorithm
1.	Assigning a sequence (k=1) of haplotypes ($h_{(i,j)}$) from a multinomial distribution (initial guess) that can give rise to the observed SNP genotype in each patient ($h^{(0)} = [h_{(1)}^{(0)}, \ldots, h_{(n)}^{(0)}]$).
2.	Choose an individual ($_i$) at random from among those individuals with ambiguous genotypes and proposed an update a new sequence of haplotypes ($h_i^{(k)} \sim Pr\{h_{(i)}|G\}$).
3.	An update is proposed by simulating a new sequence of haplotypes consistent with observed genotype base on the conditional distribution: $\Upsilon = \left(1/H_{(i,j)}/\sum_{i=1}(H_{(i,j)}\right) * H_{(i,j)} * \sum_{i\in} G_i\Big)^2 * \left(log\ (\prod_{i=1}(\theta_{(h_{i,j})})\right)^2$
4.	Repeat step 2 until chain converges.

The update is always accepted, the Υ is the probability of accepting the update. At the end of each iteration the current estimate of haplotype frequencies in θ is updated. This process of updates until the estimated haplotype frequencies converged stationary distribution. Iterations defined as being completed when every heterozygous individual has been selected and tested for an update. Patients are selected at random, but every patient can only be selected one time in each iteration. The algorithm makes at least 500 iteration and the trace and autocorrelation output as graphs (supplementary Figures S3 to S5). The algorithm tested with run 500 iteration simulates multiple chains with different starting values and reached the same results.

Existing statistical methods of haplotype reconstruction

There are seven other well-known published methods that are available to use: haplo. em software,[7] Maximum likelihood (ML) estimation using Expectation-maximization (EM) algorithm (hereon called "R-EM"). Expectation-maximization (EM) algorithm PLINK v1.07 [17] estimation using efficient iterative maximum likelihood approach (hereon called "PLINK"). Cluster into groups algorithm fastPHASE v1.4 [9] according to a hidden Markov model (hereon called "fastPHASE"). A Bayesian statistical approach PHASE v2.1.1 [8] estimation using a Metropolis-Hastings Markov chain Monte Carlo (hereon called "PHASE"). Another hidden Markov model algorithm MACH v1.0 [10] estimation using large number of templates (hereon called "MACH"). Another hidden Markov model algorithm IMPUTE v2.0 [11] estimation using forward algorithm (hereon called "IMPUTE"). And another hidden Markov model algorithm BEAGLE v.3.3.2 [12] estimation using localized haplotype-cluster model (hereon called "BEAGLE").

Evaluation of different statistical methods

Thousand datasets were simulated assuming that linkage phase unknown at 2 to 10 SNPs there are 4 to 1024 haplotypes, respectively. Each dataset is obtained by a process of four sequential steps: I. The population frequencies of haplotypes are defined by selecting a MAF for each locus at random and the number of population haplotype frequencies obtained

assuming linkage equilibrium between the alleles; II. A field survey of unrelated subjects with linkage phase unknown blood samples is simulated to obtain the genotypes observed; III. The estimated haplotype frequencies are obtained from each of the statistical approaches; IV. The estimated haplotype frequencies from dataset are used to evaluate the performance of the eight methods. Each of these four steps is repeated for each of the 1,000 datasets. The datasets and selected haplotype in each dataset are kept the same for each of the eight-analysis method, this allows a direct comparison between the different methodologies used to infer haplotype frequencies.

The performance and the accuracy of the different methods is measured as follows: 'P' is a vector whose number of elements equal h, the number of potential haplotypes in the population. The elements of the vector are indicated by the superscript *i*. The population values are compared with the estimated value as: I. The correlation coefficient (R^2) between population, and estimated haplotype frequencies; II. Similarity index (I_F) [5,18] to examine how close the computationally estimated haplotype frequencies are to the population haplotype frequencies as:

$$I_F = \sum_{i=1}^{h} \min\left(Pi_{estimated,} Pi_{population}\right) = 1 - {}^{1}/_{2} \sum_{i=1}^{h} |Pi_{estimated} - Pi_{population}|.$$

This measure incorporates all h haplotype frequencies and thus captures the overall difference between estimated and population frequencies. It varies between one, when population and estimated haplotypes frequencies are identical, and zero, when estimated haplotypes frequencies tending to zero; III. The mean squared error (MSE)[19,20] was calculated as:

$$MSE = {}^{[\sum_{i=1}^{h}(Pi_{estimated} - Pi_{population})^2]}\Big/_{h}$$

the estimated and the population haplotype frequency of i haplotype, h is the number of haplotype frequencies in the population; IV. Change coefficient *C*[21,22] assess the scaled change in haplotype frequencies and was calculated as:

$$C_i = {}^{(|Pi_{estimated} - Pi_{population}|)}\Big/_{Max[(Pi_{estimated}, Pi_{population}]}$$

The coefficients were computed for each possible haplotype across statistical methods and presented as difference of estimation (%). The value of the coefficient *C* ranges from 0 to 1, the value 0 indicating that the haplotype frequency estimated, and the haplotype frequency population are identical. Positive values indicate that haplotype frequency estimates tend to be larger than the population frequency; V. The IH index[18] to examine the number of different haplotypes detected in the population with the number of different haplotypes inferred by the statistical approaches and was calculated as:

$$I_H = {}^{2(k_{population} - k_{missed})}\Big/_{k_{population} + k_{estimated}} \cdot k_{population}$$

is the number of haplotypes in the population, $k_{estimated}$ is the number of estimated haplotypes with frequency above the threshold ${}^{1}/_{(2n)}$ in a population sample of n individuals, and $k_{estimated}$

is the number of population haplotypes not identified in the estimated haplotypes. The value of IH index ranges between one, when the statistical approach identified estimated haplotypes are the same as the population haplotypes, and zero, when none of the population haplotypes are identified by the statistical approach. In addition, the speed of the analyses which is self-explanatory recorded. The haplotype frequency estimation of the real data set uses the international collaboration Biorepository to Establish the Aetiology of Sinovenous Thrombosis (BEAST) study. [23,24]

To quantify the potential misclassification bias of estimated haplotype frequencies in genetic association analysis, probabilistic sensitivity analysis[25] was used to model the odds of phenotype that would have been observed among population haplotype frequencies without misclassification. Based on sensitivity and specificity values from these results, the sensitivity of the major allele frequency was modelled as 58%–84%, and the specificity of the major allele frequency was modelled as 71%–89%. The model is running 1,000 times for each statistical method, combining systematic and random errors in the simulation to produce probability estimates of the odds ratio (OR).

Results

The estimated haplotype frequencies with simulated population frequencies across eight statistical methods: R-EM, PLINK, fastPHASE, PHASE, MACH, IMPUTE, BEAGLE, Gibbs and the population haplotype frequencies among two to ten SNPs showed high concordance. Figure 1A shows the absolute deviation of the estimated haplotype

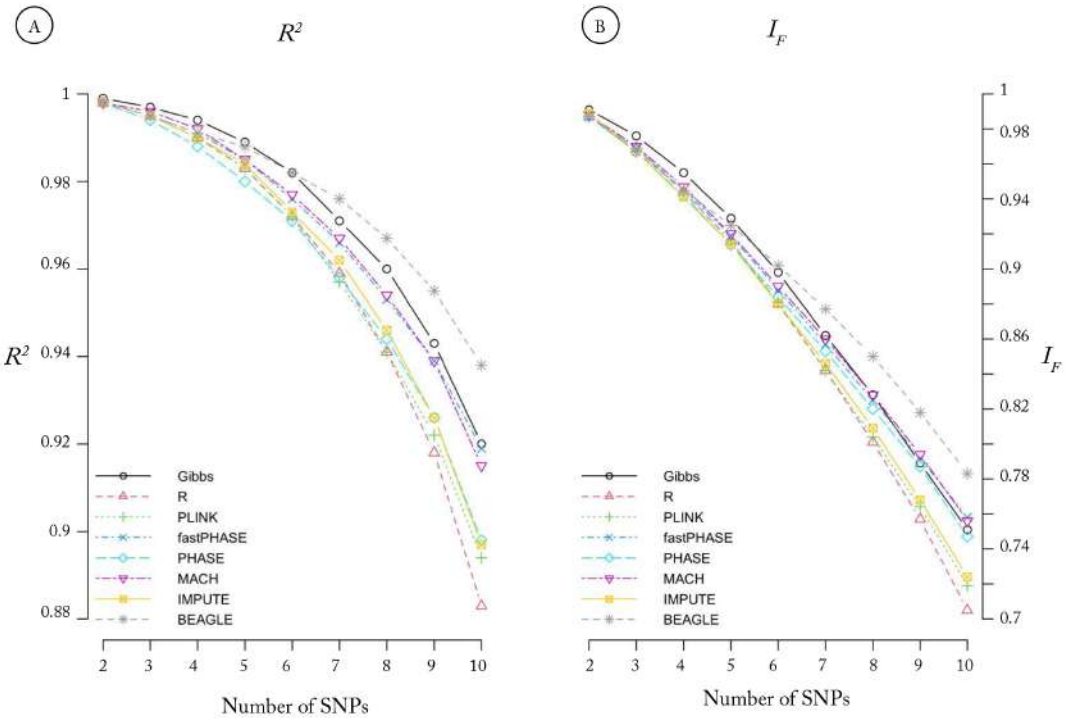

Figure 1. The correlation (R^2) and similarity index (I_F) of the estimated haplotype frequencies with simulated population haplotype frequencies across statistical methods and number of SNPs.
Gibbs, Gibbs sampler; SNPs, Single Nucleotide Polymorphisms

frequencies from simulated population haplotype frequencies, higher value represents higher accuracy. Increasing number of SNPs decreases the correlation coefficient by 0.06% to 0.12% among all statistical methods. The difference between correlation coefficients among statistical methods is less than 5.5% among population haplotype frequencies. The bias is slight and changes with increasing number of SNPs. Figure 1B shows the similarity index (I_F) of the estimated haplotype frequencies compared with simulated population haplotype frequencies, higher value represents higher accuracy. The eight statistical methods provided similarity index (I_F) values very close to each other. The difference between similarity indexes among statistical methods is less than 7.8% among population haplotype frequencies. Increasing number of SNPs decreases the similarity index between 0.20% to 0.28% among all statistical methods. This tendency is reflected in the mean squared error (MSE) statistics.

Figure 2A shows the average change coefficient *C* of the estimated haplotype frequencies compared simulated population haplotype frequencies for haplotype frequency >5%, lower value represents higher accuracy. The difference between change coefficient *C* among statistical methods is less than 11.3% among population haplotype frequencies. Increasing number of SNPs increases the change coefficient *C* between 13.3% to 24.7% among all statistical methods. There was a tendency for the estimates to cluster more closely to the population haplotype frequencies at high frequencies, showing that there is a tendency for high-frequency haplotypes to be more accurately estimated among all statistical methods. Figure 2B shows the I_H index of the estimated haplotype frequencies compared simulated population haplotype frequencies, higher value represents higher accuracy. The eight statistical methods provided I_H index values very close to each other. The difference between I_H indexes among statistical methods is less than 6.6% among population haplotype frequencies. Increasing number of SNPs decreases the I_H index between 0.29% to 0.35% among all statistical methods.

The estimated haplotype frequencies for real data among eight statistical methods are shown in Figure 3A. This was anonymized data from the international collaboration Biorepository to Establish the Aetiology of Sinovenous Thrombosis (BEAST) study. The aim of the study is to perform a genome-wide association analysis to assess the association and impact of common and low-frequency genetic variants on cerebral venous thrombosis (CVT) risk. The data set comes from nine countries, contains four biallelic SNPs from chromosome 4 is used for 1153 individuals, and is used to check the application of eight statistical methods (Gibbs, R-EM, PLINK, fastPHASE, PHASE, MACH, IMPUTE and BEAGLE). The results obtained from different methods were very similar with the mean difference of estimated haplotype frequencies between the statistical methods, is about 1%. The lowest difference of estimated probability is 0.5% present between R-EM and BEAGLE methods and the highest difference of estimated probability is 1.4% present between IMPUTE and BEAGLE methods.

Probabilistic analysis correcting exposure misclassification leading to non-differential misclassification of the exposure (Figure 3B). The analysis hypothesis that the haplotype carrier differs from the highest-frequency haplotype, or the wild-type haplotype is

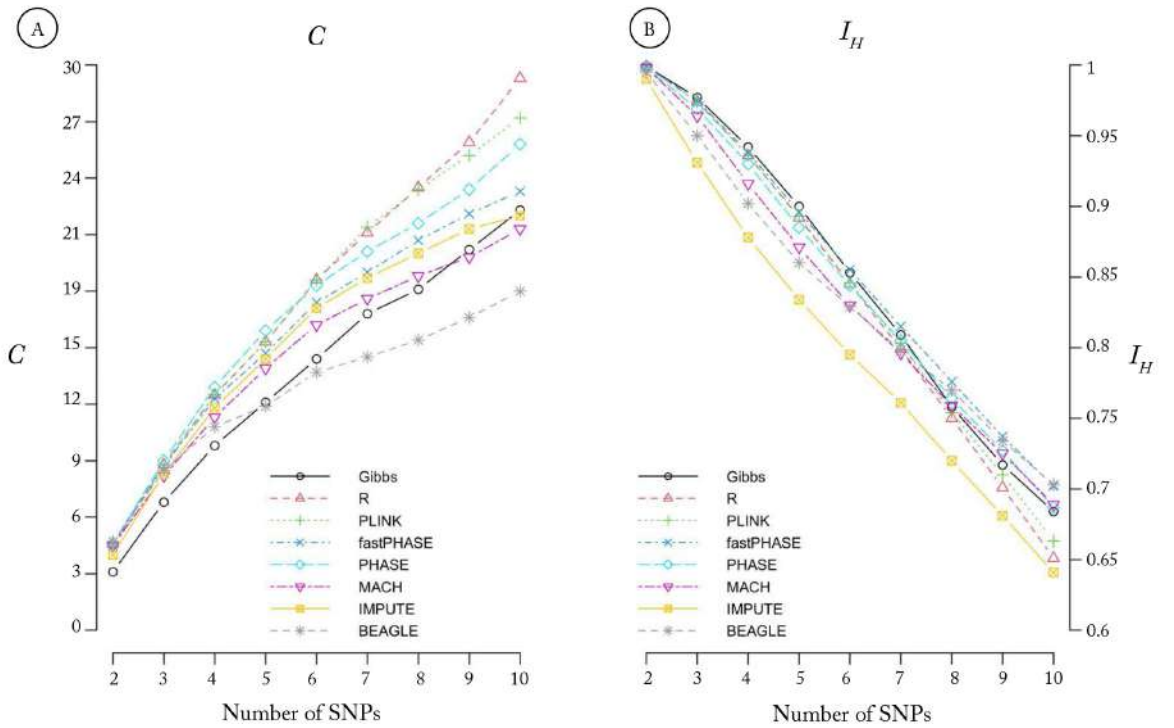

Figure 2. The average change coefficient (C) and haplotype identification (I_H) of the estimated haplotype frequencies with simulated population haplotype frequencies across statistical methods and number of SNPs.
Change coefficient (C) for haplotype frequency >5%; Gibbs, Gibbs sampler; SNPs, Single Nucleotide Polymorphisms.

associate with an increasing risk of the phenotype by observed OR of 1.50 (95% CI: 1.01-2.25, P<0.05). The OR observed is underestimate the association between the estimated haplotype and the phenotype between 36% to 99%. In addition, suggesting misclassification may have driven some of the null findings.

Supplementary Material Figures S6, to S8 shows the performance of the statistical methods when sample size is 1000 individuals. Similar patterns were observed when sample size was defined at 1000 individuals. Table S1 shows the computational time for the statistical methods, lower value represents faster calculation. There is a large difference between the statistical methods of almost 209 seconds. Increasing number of SNPs increased the time of the analysis between 95% to 10,178% among all statistical methods.

Discussion

This study proposed the Gibbs sampler statistical method for haplotype reconstruction of human blood samples. This method presents a greater accuracy, robust validity performance, the ability to deal with missing genotypes and return the probability of possible haplotype combinations in each individual and the uncertain probability of haplotype frequencies. However, all statistical methods give similar results.

There are differences between the statistical methods, especially as the number of SNPs increases. The number of haplotypes increases exponentially with the number of SNPs.[1] When the number of SNPs is small (four or

less), the EM methods (R-EM and PLINK) is accurate and valid, but the increase in the number of SNPs will reduce the accuracy and validity of the results. The Markov model methods (Gibbs sampler, fast PHASE, PHASE, MACH1, IMPUTE2 and BEAGLE) presents the same pattern, but when the number of SNPs increases the accuracy and validity of the results are higher than that of the EM method. All statistical methods have dimensionality problems, and the results obtained are very sensitive to the increase in the number of SNPs, their accuracy and validity decrease with the increase in the number of SNPs. Although, in the case of seven SNPs or less, the new method still provides better results than the existing methods. The computational complexity increases exponentially with the number of SNPs, and when investigating genetic association studies, it is rarely necessary to analyse more than seven SNPs at the same time in practice. While all statistical methods have similar accuracy, R-EM[7] is the fastest method for estimating haplotype frequencies. The PLINK v1.07[6] is the only version with the phasing and haplotype testing algorithms. Future versions of PLINK recommend using BEAGLE[12] to estimate haplotype frequencies and haplotype associations. The PHASE v1.0[26,27] uses Gibbs sampling method to estimate haplotype frequencies base on mutation rate. The second version 8 uses the Metropolis-Hastings (MH) method and is considered more accurate.

The asymptotic statistical performance of all methods is similar in same situations, the differences in the statistical performance among methods are due to the different priors they use for modelling population haplotype frequencies. It is not possible to calculate the population haplotype frequencies directly. The non-parametric Gibbs sampler does not base the sampling of the artificial samples (Dirichlet, Poisson, or Gamma distribution) on the assumed model of the marginal distribution of genotype observations but uses instead a non-parametric description of the joint distribution, the empirical distribution. The differences in performance among the methods are relatively minor. The accuracy increases due to the prior used in the inference methods is more like the population haplotype frequencies. The Gibbs sampler method draws iteratively from conditional distributions particularly useful and lower in dimension rather than drawing directly from the joint distribution with which it may not always be easy to work. While the Gibbs sampler relies on conditional distributions, the Markov model methods bases on Metropolis-Hastings sampler uses a full joint density distribution to generate a candidate draws. The candidate draws are not automatically added to the chain but rather an acceptance probability distribution is used to accept or reject candidate draws. These methods are sensitive to the step size between draws. Either too large or too small of a step size can have a negative impact on convergence. The Gibbs algorithm sampling likely haplotypes for all subjects does not need to consider every possible haplotype unlike the EM-algorithm which must sum over every possible haplotype during the E-step. This property of the Gibbs sampler makes it better suited to deal with situations where there are many possible haplotypes, and many markers. While the EM-algorithm will converge to a maximum, it may be only a local maximum. However, the Gibbs sampler may get trapped in a local mode but it does have a chance of escaping such a mode

and finding the true regions of parameter space with high probability.[28]

The convergence diagnostics plots present in the Supplementary Material Figures S3 to S8 demonstrates clear convergence. The algorithm successfully converged to reach a stationary distribution after a few runs. The chain did not get stuck in certain areas of the parameter space, indicating poor mixing. The median of the shrink factor does not increase above 1.1 among all haplotype frequency groups. The Gelman and Rubin's convergence diagnostic the scale reduction factors for each parameter is 1.09 maximum value at each parameter for 500 iterations and it decreases to 1.05 maximum value at each parameter for 1000 iterations. A factor of 1 implies that between and within chain variances are equal, larger values suggest that there is still a notable difference between chains. Shrink values below 1.1 or 1.05 acceptable for practical purposes.[29] Gelman and Rubin[30] and Brooks and Gelman[31] suggest that the maximum Gelman–Rubin diagnostic across all model parameters values greater than 1.2 for any of the model parameters should indicate non-convergence. In addition, the mean plots present how well the chains are mixing and how the two chains go in the same direction. More iteration cause further decreases in the scale reduction factor however, Raftery and Lewis[32,33] test the number of iterations and suggest a minimum of 300 iterations. These diagnostics tend to be conservative so that more iterations may be necessary. Heidelberg and Welch diagnostic[34,35] calculates a test statistic to change the null hypothesis that the Markov chain is from a stationary distribution, the chain passes the test, so the chain does not need to run longer. In addition, the chain

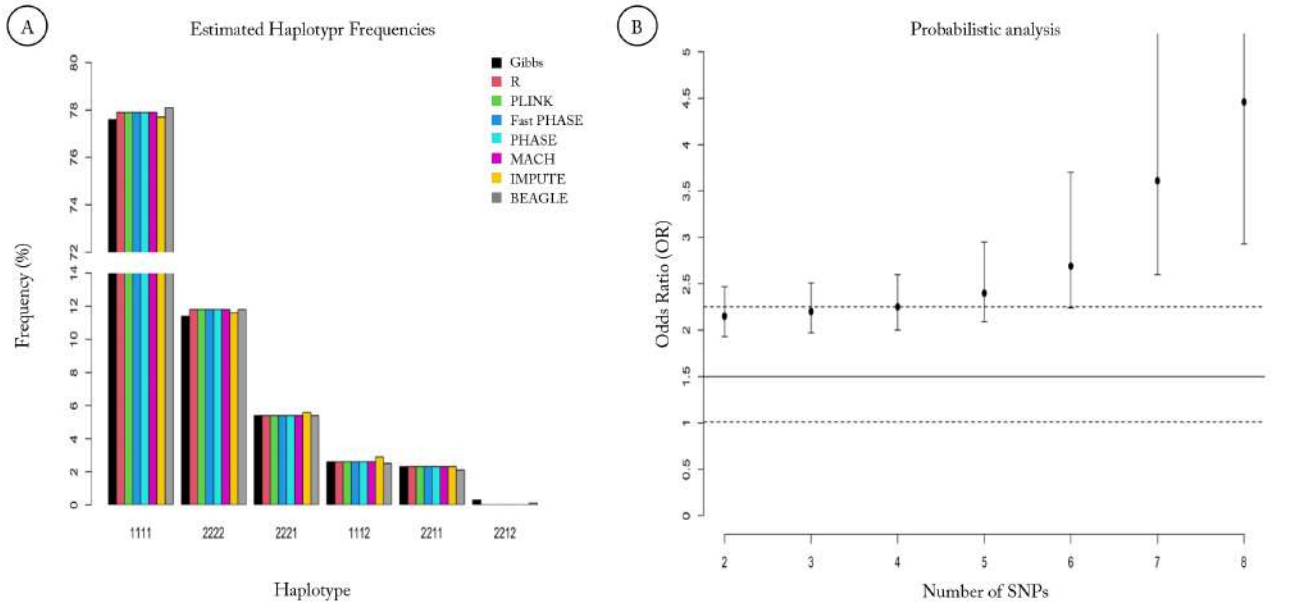

Figure 3. The estimated haplotype frequencies for real data set (BEAST study), n=1153 individuals. In addition, the probabilistic analysis correcting the OR of estimated haplotype frequencies across number of SNPs. Solid line represents observe OR=1.50; dashed line represents observe 95% CI=1.01-2.25; solid circle and dashed line represents misclassification bias corrected OR.

Haplotype 1, major allele; Haplotype 2, minor allele; Gibbs, Gibbs sampler; OR, odds ratio; SNPs, Single Nucleotide Polymorphisms.

passes the Geweke diagnostic test[36] that takes two non-over lapping parts the first 0.1 and last 0.5 proportions of the Markov chain and compares the means of both parts, using a difference of means test to see if the two parts of the chain are from the same distribution (null hypothesis). Accordingly, the current algorithm runs 500 iterations and 200 burn ins. The algorithm allows the user to choose the number of iterations, number of chains and number of burn in.

The Gibbs sampler methods examines possible haplotype combination in an individual that could plausibly give rise to the observed genotype and obtain the probability that any given subject carrier disease susceptibility haplotype. The presence of a disease susceptibility haplotype can be inferred in individual patients and the probability of their presence used as a weighting in a logistic regression predicting the risk of the haplotype. A positive impact of the disease susceptibility haplotype on the phenotype risk would indicate it truly affects susceptibility levels.

It is estimated that the misclassification of the haplotypes leads to non-differential bias in the exposure. The different result patterns produced by the probabilistic sensitivity analysis (Figure 3B) indicate that the misclassification of the exposure may have affected haplotype estimates in haplotype analysis, although it can explain all invalid associations reported in the literature. Due to phase uncertainty, haplotype analysis of unrelated individuals underestimated the estimated association between haplotype and phenotype by more than 40%. However, only a subset of SNPs contains information about haplotype effects. The inclusion of non-informative SNPs will effectively divide the sample into multiple haplotypic groups, consequently decreasing the statistical power of the study, while increasing the degrees of freedom of the test.[1] Variable selection techniques can be used to identify the most parsimonious haplotype responsible for the association with the phenotype. These methods can be used to reduce the number of SNPs considering a set of markers that can independently contribute to the association.[1]

The simulations were limited to ten SNPs and hundred individuals to simplify the comparison. Results from haplotypes defined at thousand individuals are presented in Supplementary Material Figures S1 and S2, the same pattern irrespective of whether haplotypes are defined at hundred or thousand loci. The novel approach described above did not have to limit the number of SNPs analysed. However, the large number of SNPs may be too many possible haplotype configurations to make estimation computationally practical. The best solution to analyse large number of SNPs is partition ligation (PL). Regions of interest are divided into short, non-overlapping SNPs, usually consisting of fewer than 10 SNPs. Haplotype reconstruction can then be applied within each short number of SNPs to obtain accurate population frequency estimates.[37] In addition, the examples were limited to ten SNPs because the complexity of calculations rises exponentially with the number of SNPs and it is recommending to drop the non-informative SNPs when investigating the association with the phenotype.[1] However, calculating large number of SNPs increases the calculation time and depends on available computer memory.[38]

Conclusion

The proposed Gibbs sampler method provides

greater accuracy, robust validity performance, the ability to handle missing genotypes and return the probability of possible haplotype combinations in each individual and the uncertain probability of haplotype frequencies. Misclassification of estimated haplotype frequencies leads to non-differential bias in exposure and affects haplotype analysis. Due to phase uncertainty, it underestimates the estimated association between haplotype and phenotype by more than 40%. The R code used for these simulations and analyses are freely available on request to GKD.

Abbreviations

GAS, Genetic association studies;
GWAS, Genome wide association studies;
LD, Linkage-disequilibrium;
EM, Expectation-Maximization;
MCMC, Markov chain Monte Carlo;
HMM, Hidden Markov model;
MAF, Minor allele frequencies;
ML, Maximum Likelihood;
SNPs, Single Nucleotide Polymorphisms;
IF, Similarity index;
MSE, Mean squared error;
CI, Confidence Intervals;
BEAST, Biorepository to Establish the Aetiology of Sinovenous Thrombosis;
CVT, cerebral venous thrombosis

References

1. Ken-Dror G, Humphries SE, Drenos F. The use of haplotypes in the identification of interaction between SNPs. Hum Hered 2013; 75(1): 44-51.

2. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet 2010; 11(7): 499-511.

3. Slatkin M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. Nat Rev Genet 2008; 9(6): 477-85.

4. Niu T. Algorithms for inferring haplotypes. Genet Epidemiol 2004; 27(4): 334-47.

5. Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 1995; 12(5): 921-7.

6. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007; 81(3): 559-75.

7. Sinnwell JP, Schaid, D.J. haplo. stats: Statistical Analysis of Haplotypes with Traits and Covariates when Linkage Phase is Ambiguous. 2020.

8. Stephens M, Scheet P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 2005; 76(3): 449-62.

9. Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 2006; 78(4): 629-44.

10. Li Y, Willer CJ, Ding J, Scheet P,

Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 2010; 34(8): 816-34.

11. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 2009; 5(6): e1000529.

12. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 2007; 81(5): 1084-97.

13. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. 2020.

14. Brooks S, Brooks S, Gelman A, Jones G, Meng X-L, Brooks S. Handbook of Markov chain Monte Carlo. Boca Raton, Fl: CRC Press; 2011.

15. Gilks WR, Richardson S, Spiegelhalter DJ. Markov chain Monte Carlo in practice. London ; New York: Chapman & Hall; 1996.

16. Roberts GO, Sahu SK. Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler. Journal of the Royal Statistical Society Series B 1997; 59: 291-317.

17. Li X, Foulkes AS, Yucel RM, Rich SM. An expectation maximization approach to estimate malaria haplotype frequencies in multiply infected children. Stat Appl Genet Mol Biol 2007; 6: Article33.

18. Adkins RM. Comparison of the accuracy of methods of computational haplotype inference using a large empirical dataset. BMC Genet 2004; 5: 22.

19. Fallin D, Schork NJ. Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am J Hum Genet 2000; 67(4): 947-59.

20. Istrail S, Waterman MS, Clark AG. Computational methods for SNPs and Haplotype inference : DIMACS/RECOMB satellite workshop, Piscataway, NJ, USA, November 2002 revised papers / Sorin Istrail, Michael Waterman, Andrew Clark, (eds.). Berlin ; New York: Springer-Verlag; 2004.

21. Tishkoff SA, Pakstis AJ, Ruano G, Kidd KK. The accuracy of statistical methods for estimation of haplotype frequencies: an example from the CD4 locus. Am J Hum Genet 2000; 67(2): 518-22.

22. Sabbagh A, Darlu P. Inferring haplotypes at the NAT2 locus: the computational approach. BMC Genet 2005; 6: 30.

23. Cotlarciuc I, Marjot T, Khan MS, et al. Towards the genetic basis of cerebral venous thrombosis-the BEAST Consortium: a study protocol. BMJ Open 2016; 6(11): e012351.

24. Ken-Dror G, Cotlarciuc I, Martinelli I, et al. Genome-wide association study identifies

first locus associated with susceptibility to cerebral venous thrombosis. Ann Neurol 2021.

25. Lash TL, Fox MP, Fink AK, SpringerLink. Applying Quantitative Bias Analysis to Epidemiologic Data. New York, NY: Springer New York : Imprint: Springer; 2009.

26. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 2001; 68(4): 978-89.

27. Stephens M, Donnelly P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 2003; 73(5): 1162-9.

28. Ken-Dror G, Sharma P. Markov chain Monte Carlo Gibbs sampler approach for estimating haplotype frequencies among multiple malaria infected human blood samples. Malar J 2021; 20(1): 311.

29. Lunn D, Lunn D. The BUGS book : a practical introduction to Bayesian analysis. Boca Raton, FL London: CRC Press Chapman & Hall; 2013.

30. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Statistical Science 1992; 7: 457–72.

31. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 1998; 7: 434–55.

32. Raftery AE, Lewis SM. One long run with diagnostics: Implementation strategiesfor Markov chain Monte Carlo. Statistical Science 1992; 7: 493-7.

33. Spiegelhalter WR, Gilks WR, Richardson S, Spiegelhalter DJ. Markov chain Monte Carlo in practice. Boca Raton, Fla: Chapman & Hall; 1996.

34. Heidelberger P, Welch PD. A spectral method for confidence interval generation and run lengthcontrol in simulations. Communications of the ACM 1981; 24(4): 233-45.

35. Heidelberger P, Welch PD. Simulation run length control in the presence of an initial transient. Operations Research 1983; 31(6): 1109-44.

36. Bernardo JM. Bayesian Statistics 4 : proceedings of the 4th Valencia International Meeting, April 15-20, 1991. Oxford: O.U.P; 1992.

37. Zeggini E, Morris A, ScienceDirect. Analysis of complex disease association studies: a practical guide. Amsterdam: Elsevier; 2011.

38. Ken-Dror G, Hastings IM. Markov chain Monte Carlo and expectation maximization approaches for estimation of haplotype frequencies for multiply infected human blood samples. Malar J 2016; 15(1): 430.