Bioinformatics, 35(16), 2019, 2827–2833
doi: 10.1093/bioinformatics/bty1040
Advance Access Publication Date: 24 December 2018
Original Paper

OXFORD

Data and text mining

CSHAP: efficient haplotype frequency estimation based on sparse representation

Yinsheng Zhou[1], Han Zhang[2] and Yaning Yang[1,*]

[1]Department of Statistics and Finance, University of Science and Technology of China, Hefei, Anhui 230026, China and [2]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 20852, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on July 26, 2018; revised on December 14, 2018; editorial decision on December 17, 2018; accepted on December 20, 2018

Abstract

Motivation: Estimating haplotype frequencies from genotype data plays an important role in genetic analysis. *In silico* methods are usually computationally involved since phase information is not available. Due to tight linkage disequilibrium and low recombination rates, the number of haplotypes observed in human populations is far less than all the possibilities. This motivates us to solve the estimation problem by maximizing the sparsity of existing haplotypes. Here, we propose a new algorithm by applying the compressive sensing (CS) theory in the field of signal processing, compressive sensing haplotype inference (CSHAP), to solve the sparse representation of haplotype frequencies based on allele frequencies and between-allele co-variances.

Results: Our proposed approach can handle both individual genotype data and pooled DNA data with hundreds of loci. The CSHAP exhibits the same accuracy compared with the state-of-the-art methods, but runs several orders of magnitude faster. CSHAP can also handle with missing genotype data imputations efficiently.

Availability and implementation: The CSHAP is implemented in R, the source code and the testing datasets are available at http://home.ustc.edu.cn/~zhouys/CSHAP/.

Contact: ynyang@ustc.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Inferring haplotype pairs from unphased genotypes is of great importance in genetic analysis. The key point is to solve the phase ambiguity from genotype data, and there have been many methods for haplotype inference (Browning and Browning, 2011; Liu *et al.*, 2008; Niu *et al.*, 2002). Likelihood-based approaches using the expectation-maximization (EM) algorithm and related methods are proved to be statistically efficient under Hardy-Weinberg equilibrium (HWE) (Excoffier and Slatkin, 1995; Qin *et al.*, 2002). On the other hand, a number of Bayesian methods that incorporate complex biological knowledge have been widely studied in literature (Lin *et al.*, 2002; Niu *et al.*, 2002; Stephens *et al.*, 2001; Stephens and Donnelly, 2003; Stephens and Scheet, 2005; Xing *et al.*, 2007; Zhang *et al.*, 2006). Among those, the PHASE algorithm (Stephens and Scheet, 2005) had been considered as a gold standard, in which an approximate coalescent prior was used to model the clustering property of haplotypes. For modern phasing methods, hidden Markov models (HMMs) are often used. Under the assumption that haplotypes tend to cluster into groups over short regions, Scheet and Stephens (2006) proposed the fastPHASE algorithm based on HMM. This algorithm made it possible to a phase larger number of markers and samples. To reduce the computational time of PHASE, Delaneau *et al.* (2008) used the binary trees to represent the sets of possible haplotypes and proposed a linear complexity phasing method Shape-IT (Delaneau *et al.*, 2011). By combining the local windows approach of phasing in Impute2 (Howie *et al.*, 2011), Delaneau *et al.* (2013) proposed Shape-IT v2, which has been widely used in the 1000 Genomes Project (Delaneau *et al.*, 2014).

Due to low rates of mutations and re-combinations in genetic evolution and the resultant high linkage disequilibrium (LD) in the genome, only a few haplotypes out of the large number of possible haplotypes are present in population (Daly *et al.*, 2001; Patil *et al.*,

2001). Many algorithms have been developed to maximize different measurements of parsimony of population haplotypes.

Clark (1990) suggested a heuristic approach to reconstruct haplotypes that use the least number of distinct haplotypes. Gusfield (2001) reformulated Clark's algorithm into a *conceptual* integer linear programing (ILP) problem, then made a further modification (Gusfield, 2003) to make it practical. Computational results reported that the ILP method is very accurate (Gusfield and Orzack, 2005) for tightly linked polymorphisms. Moreover, Gusfield (2003) proposed an approach to the haplotype inference problem called *Pure-Parsimony* approach, which tries to find a solution that minimizes the total number of distinct haplotypes used, or equivalently, the ℓ_0 norm of haplotype frequency vector. Unfortunately, the Pure-Parsimony method is NP-hard, which cannot be computed efficiently. Jajamovich and Wang (2012) proposed to obtain a sparse solution by minimizing the Tsallis entropy of the frequency vector, which is still an NP-hard problem.

The most impressive conclusion in compressive sensing (CS) theory in the field of signal processing is that sparse or compressible signals could be accurately and efficiently recovered by minimizing the ℓ_1 norm of signals subject to the measurements from well-behaved under-determined linear sensing systems. Based on this, we propose a regularization algorithm, CSHAP, to reconstruct haplotypes by maximizing an ℓ_1 measure of parsimony defined as the weighted sum of haplotype frequencies, subject to constraints of a series of under-determined linear equations determined by estimated allele frequencies and pairwise covariance (or LD coefficients) of genotypes. Although the use of the first two moments as constraints may lose information, but as we will see later the loss is acceptable and exact or almost exact reconstruction is possible if haplotypes that present in population are sparse.

In this article, we propose a CS framework, CSHAP, for haplotype inference. We first introduce the approach based on the sparse representation of the exact allele frequencies and LD coefficients. Then we relax the constraints by allowing estimation errors in allele frequencies and LD coefficients. We will show that the method is applicable to both individual and pooling designs and works under both HWE and Hardy-Weinberg disequilibrium (HWD). Furthermore, by applying a modified EM algorithm, we can improve the accuracy of the CSHAP algorithm and make it possible to deal with missing data. Finally, based on the idea of partition-ligation (PL) first proposed by Niu *et al.* (2002), we further extend the algorithm to support long-range haplotypes. Performance of CSHAP is evaluated and compared with other methods by simulation studies and real data analysis.

2 Materials and methods

2.1 Notations

Consider q single nucleotide polymorphism (SNP) loci. Each locus is diallelic and thus theoretically there are $r = 2^q$ possible haplotypes. Denote the minor allele at each locus by 1 and major allele by 0. We index the haplotypes by the sequence of coefficients of binary expansion of $0, 1, \ldots, r$, i.e. let $h_j = (h_{1j}, \ldots, h_{qj})'$ represent the jth haplotype vector.

For a specific locus of an individual, if both alleles have a 0 (or 1), this site is called homozygous and is encoded with a 0 (or 2). Conversely, if the two alleles are different, the site is heterozygous and the genotype for the locus is 1. The above encoding also works for pooling designs. Suppose we have T pooled genotype observations at the q loci, each is from a pool of N individuals (or $2N$ chromosomes equivalently). Let $G_i = (G_{i1}, \ldots, G_{iq})'$ be the pooled genotypes of the ith pool. Since G_i is a sum of $2N$ haplotypes, we have $0 \le G_{ik} \le 2N$, $i = 1, \ldots, T, k = 1, 2, \ldots, q$. If $N = 1$, then the observations are individual genotypes and T stands for the number of individuals. Our objective is to estimate the haplotype frequencies $\mathbf{p}$ based on the genotype observations $G = (G_1, \ldots, G_T)'$.

2.2 Sparse representation of haplotype frequency

The minor allele frequency (MAF) for each locus and the joint probability of minor alleles at any two loci can be represented linearly using the haplotype frequencies $\mathbf{p}$. In fact, the MAFs have a linear form representation $\omega_0 = H\mathbf{p}$, and the joint probabilities of minor alleles can be expressed as $\eta_0 = H\Lambda H'$, where $H = (h_1, \ldots, h_r)$ is a $q \times r$ matrix and $\Lambda = \mathrm{diag}\{\mathbf{p}\}$ is a diagonal $r \times r$ matrix with diagonal elements $\mathbf{p}$ (Sham *et al.*, 2002). Although η_0 is a matrix consisting of all pairwise allele probabilities of any two loci, i.e. $\eta_{0i,j} = \mathbb{P}(\text{locus } i = 1, \text{locus } j = 1)$. These two representations are true whether the HWE holds or not. In addition, we have the natural constraint $1'\mathbf{p} = 1$ since $\mathbf{p}$ is a probability distribution. Now we write the above three equations into a single linear equation:

$$\Psi\mathbf{p} = \mathbf{b}, \tag{1}$$

with

$$\Psi = \begin{pmatrix} 1' \\ H \\ H \wedge H \end{pmatrix}$$

where the $\frac{q(q-1)}{2} \times r$ matrix $H \wedge H$ denotes the matrix whose rows are the logical product of all pairs of H, and $\mathbf{b}$ consists of the scalar 1, the vector ω_0 and vector of the off-diagonal entries of η_0, corresponding to the above three constraints, respectively.

Next, we normalize matrix Ψ such that its columns have unit ℓ_2 norms. Denote $\Lambda_\tau = \mathrm{diag}(\tau_1, \ldots, \tau_r)$, where τ_j is the ℓ_2 norm of the jth column of Ψ. We denote the normalized matrix as $\Phi = \Psi\Lambda_\tau^{-1}$ and will call it the sensing matrix. Define $\mathbf{p}^* = \Lambda_\tau\mathbf{p}$. Then (1) can be rewritten as

$$\Phi\mathbf{p}^* = \mathbf{b} \tag{2}$$

Note that the sensing matrix Φ has dimension $\frac{q^2+q+2}{2} \times r$ and the (2) is underdetermined when $\frac{q^2+q+2}{2} < r = 2^q$, or equivalently $q > 3$. Therefore, $\mathbf{b}$ can be regarded as the noiseless partial observations of $\mathbf{p}^*$ from underdetermined linear sensing system Φ.

Because only relatively few haplotypes can be present in a population when the number of possible haplotypes is large, the principle of parsimony has been emphasized in haplotype reconstruction (Clark, 1990; Gusfield, 2003). Noticing the equivalency between the sparsity of $\mathbf{p}$ and $\mathbf{p}^*$, the sparse signal $\mathbf{p}^*$ can be recovered by solving the following optimization problem:

$$\min_{\mathbf{p}^* \in \mathbb{R}_+^r} ||\mathbf{p}^*||_{\ell_0} \text{ subject to } \Phi\mathbf{p}^* = \mathbf{b}. \tag{3}$$

Solving (3) is equivalent to seeking the sparsest solution in the feasible space, which is a combinatorial optimization problem and is NP-hard.

The ℓ_1 norm is known as the convex envelope of the ℓ_0 norm over convex set $\Delta = \{x \in \mathbb{R}^r : ||x||_\infty \le 1\}$ when $||x||_\infty = \max x_i$. In other words, $||x||_{\ell_1}$ is the best pointwise approximation to $||x||_{\ell_0}$ among the set of convex function on Δ (Hiriart-Urruty and Lemaréchal, 1993; Recht *et al.*, 2010). Thus, instead of minimizing the ℓ_0 norm of $\mathbf{p}^*$ directly, we suggest the ℓ_1 minimization:

$$\min_{\mathbf{p}^* \in \mathbb{R}_+^r} ||\mathbf{p}^*||_{\ell_1} \text{ subject to } \Phi\mathbf{p}^* = \mathbf{b}. \tag{4}$$

which is the typical form in CS theory and can be solved efficiently by linear programing.

The above Equation (4) is based on known exact value of $\mathbf{b}$, however, $\mathbf{b}$ is unknown and needs to be estimated based on genotype observations. In the next subsection, we will make some modifications of (4) by allowing $\mathbf{b}$ with estimation errors.

2.3 Presence of noise

The MAFs can be readily estimated by sample means $\hat{\omega}_0 = \sum_i G_i/(2NT)$. Under the assumption of HWE, the variance–covariance matrix of the q alleles can be estimated by sample variance–covariance matrix $\hat{\Sigma}_0 = \sum_i G_i G_i'/(2NT) - 2N\hat{\omega}_0\hat{\omega}_0'$ (Zhang *et al.*, 2008). Then the two-locus joint probabilities $\eta_0 = \Sigma_0 + \omega_0\omega_0'$ can be estimated by $\hat{\eta}_0 = \hat{\Sigma}_0 + \hat{\omega}_0\hat{\omega}_0'$. Then a natural estimate of $\mathbf{b}$ is $\hat{\mathbf{b}}$ consisting of 1, $\hat{\omega}$ and the off-diagonal entries of $\hat{\eta}$.

Now allowing error of $\hat{\mathbf{b}}$ in measuring $\mathbf{\Phi p}^*$, we transform the optimization Equation (4) into the following robust problem:

$$\min_{\mathbf{p}^* \in \mathbb{R}_+^r} ||\mathbf{p}^*||_{\ell_1} \text{ subject to } ||\mathbf{\Phi p}^* - \hat{\mathbf{b}}||_{\ell_2} \leq \epsilon \tag{5}$$

where ϵ is a tuning parameter controlling the error of estimates. This problem is a special case of second order cone programing (Boyd and Vandenberghe, 2004), which can be solved efficiently by many standard approaches such as iteratively reweighted least squares or interior-point algorithms.

The haplotype frequency p_j is then estimated by $\hat{p}_j = \hat{p}_j^*/\tau_j, j = 1, \ldots, r$, $\mathbf{p}^*$ is the solution to (5). Necessary normalization is carried out to ensure $\hat{\mathbf{p}}$ is a proper probability distribution.

2.4 Selection of tuning parameter ϵ

To solve the Equation (5), we need to select a proper tuning parameter ϵ. Given the fact that $\epsilon = 0$ when $\hat{\mathbf{b}} = \mathbf{b}$, the ϵ should be as small as possible as the ℓ_2 estimation error of the first two moments should not be too large. However, we found that there is usually no solution for (5) if $\epsilon = 0$ as typically $\hat{\mathbf{b}} \neq \mathbf{b}$. On the other hand, the estimated frequencies are degenerate to $(1, 0, \ldots, 0)'$ as $\epsilon \to \infty$.

It is reasonable to believe that, for sufficiently large T and N, the estimates $\hat{\mathbf{b}}$ is very close to $\mathbf{b}$ with certain probability, thereby the level of noise (ℓ_2 error of estimates) is bounded. The ϵ should be related to the signal-to-noise ratio, which is directly determined by T, N and $\mathbf{b}$.

From the Central Limit Theorem in statistics, we know that if the pool size N is large, then G_i will be asymptotically normally distributed with mean $\mu_G = 2N\omega_0$ and covariance matrix $\Sigma_G = 2N\Sigma_0$ (Kuk *et al.*, 2008). Using this fact we obtain (Bilodeau and Brenner, 2008)

$$\begin{aligned} 2N\sqrt{T}(\hat{\omega}_0 - \omega) &\xrightarrow{d} \mathcal{N}(0, \Sigma_G) \\ 2N(T-1)\hat{\Sigma}_0 &\xrightarrow{d} \mathcal{W}_q(T-1, \Sigma_G) \end{aligned} \tag{6}$$

Here $\xrightarrow{d}$ represents convergence in distribution and $\mathcal{W}_q(T-1, \Sigma_G)$ represent the Wishart distribution with $T-1$ degrees of freedom and scale matrix Σ_G.

Denote $e = \hat{\mathbf{b}} - \mathbf{\Phi p}^*$, approximately we have $e \sim \mathcal{N}(0, \Sigma_e)$, where $\Sigma_e = \text{Var}(\hat{\mathbf{b}})$. Consider the squared norm of the error $||e||_{\ell_2}^2 = e'e$, then:

$$\begin{aligned} \text{E}(e'e) &= \text{tr } (\Sigma_e) \\ \text{Var}(e'e) &= 2 \text{ tr } (\Sigma_e\Sigma_e) \end{aligned}$$

holds asymptotically. Based on the well-known concentration inequalities, the probability that $||e||_{\ell_2}^2$ exceeds its mean plus 2 or 3 standard deviations is small (<5 or 2%, respectively). We then solve (5) with

$$\epsilon_\lambda^2 = \text{tr } (\Sigma_e) + \lambda\sqrt{2 \text{ tr } (\Sigma_e\Sigma_e)}$$

and simply select $\lambda = 2$.

Since all the components of the vector $\hat{\mathbf{b}}$ are made up of the elements in $\hat{\omega}_0$ and $\hat{\Sigma}_0$, and we have obtained the asymptotic distribution of them in (6). The variance–covariance matrix Σ_e of $\hat{\mathbf{b}}$, can be derived or estimated from $\hat{\omega}_0$ and $\hat{\Sigma}_0$ naturally (details can be found in the Supplementary Section 1.1).

2.5 The deviation of HWE

Likelihood-based methods for estimating haplotype frequencies often require the assumption of HWE, but which may not be true if there exists population substructure. Here we introduce the inbreeding coefficient ρ to describe the possible dependence of the two haplotypes within a subject (Zeng and Lin, 2005). The distribution of diplotype $H = [h_k, h_l]$ is assumed as:

$$\mathbb{P}(H|\mathbf{p}, \rho) = (1-\rho)p_k p_l + \rho p_k \mathbb{I}_{k=l}, \tag{7}$$

where $\mathbb{I}_{k=l}$ is the indicator of whether $k = l$. Larger ρ means more similar between H_1 and H_2. Note that if one takes $\rho = 0$, then this is equivalent to the HWE assumption. Under model (7), the variance–covariance matrix of the genotype data is $\Sigma_\rho = (1+\rho)\Sigma_0 = (1+\rho)(\eta_0 - \omega_0\omega_0')$ which can be estimated by $\hat{\Sigma}_\rho = \frac{1}{2NT}\sum_i G_i G_i' - 2N\hat{\omega}_\rho\hat{\omega}_\rho'$ and $\hat{\omega}_\rho = \hat{\omega}_0 = \frac{1}{2NT}\sum_i G_i$. The joint probabilities η_ρ therefore are estimated by $\hat{\eta}_\rho = \hat{\Sigma}_\rho/(1+\hat{\rho}) + \hat{\omega}_\rho\hat{\omega}_\rho'$, where $\hat{\rho}$ is an estimate of ρ. In this article, unless when HWE is explicitly assumed, we adopt the following estimate

$$\hat{\rho} = \frac{\sum_{i=1}^q V_i}{\sum_{i=1}^q \bar{G}_i(1-\bar{G}_i/2N)} - 1,$$

where $\bar{G}_i$, V_i is the sample mean and variance of genotypes of ith locus, respectively.

Then $\mathbf{b}$ can be readily estimated by the estimator $\hat{\mathbf{b}}_\rho = \hat{\mathbf{b}}(\hat{\omega}_\rho, \hat{\eta}_\rho)$. Now problem (5) is then modified to

$$\min_{\mathbf{p}^* \in \mathbb{R}_+^r} ||\mathbf{p}^*||_{\ell_1} \text{ subject to } ||\mathbf{\Phi p}^* - \hat{\mathbf{b}}_\rho||_{\ell_2} \leq \epsilon \tag{8}$$

where ϵ is the ℓ_2 error bound.

On the other hand, introducing a new estimator $\hat{\rho}$ will increase the variance, especially when ρ gets close to zero. In this case, the departure from HWE is negligible and it is not worth considering an extra inbreeding parameter, since our EM algorithm is relatively robust to deviations from HWE. An adaptive way to test the hypothesis

$$H_0 : \rho = 0 \leftrightarrow H_1 : \rho > 0 \tag{9}$$

In the simulation experiment, we compute $\hat{\rho}_L = \hat{\rho} - 1.645\sqrt{\hat{v}_\rho}$, the 95% lower one-sided confidence bound, where the $\sqrt{\hat{v}_\rho}$ is the variance of $\hat{\rho}$ obtained by bootstrap. If $\hat{\rho}_L > 0$, we will use the modified Equation (8). If $\hat{\rho}_L \leq 0$, we will ignore the inbreeding parameter and still use (5).

2.6 CSHAP for individual DNA data

For individual genotyping design ($N = 1$), the Equation (5) can accurately identify the major haplotypes which really exist in the population, but the values may be biased.

Notice that the Equation (5) does not guarantee the compatibility of haplotypes with all the observed genotypes, especially when ϵ is relatively large, while too small ϵ may leads to no solution. Therefore, we first find the smallest ϵ_0 such that solution exists. Then we calculate the largest ϵ_λ described in Section 2.4. Now we have a range of ϵ values $[\epsilon_0, \epsilon_\lambda]$. For every ϵ that falls within this interval, our equation will return a series of solutions with different

compatibility and sparsity. In all these compatible solutions, we choose the sparsest one as $\hat{\mathbf{p}}$.

Once we get the above solution $\hat{\mathbf{p}}$, we obtain a set of haplotypes $\mathcal{H}$ based on the non-zero entries of $\hat{\mathbf{p}}$. Now we can use $\hat{\mathbf{p}}$ and $\mathcal{H}$ as the input of the standard EM algorithm. In each iteration, we will only consider the haplotypes contained in $\mathcal{H}$, and only update their corresponding frequencies in $\hat{\mathbf{p}}$.

In fact, the $\hat{\mathbf{p}}$ is quite close to the true haplotype frequency, especially for common-frequency ($p_i \geq 5\%$) variants. Correspondingly, $\mathcal{H}$ is very similar to the haplotypes that really present in sample genotypes, so that we can regard $\mathcal{H}$ as a haplotype reference panel. For most individuals, we can find only one compatible diplotype configurations in the reference panel; then, we regard these solutions as 'solved' and exclude them in the following iterations. Although most of the phasing algorithm needs to update the estimated haplotypes for every individuals in each iteration, we only consider the remaining ones. Furthermore, when the total number of individuals T is large, we only use the distinctive genotypes that appear in the remaining individuals, since the number of distinctive genotypes is very limited compared with T. These modifications will significantly reduce the number of individuals that need to be considered in each iteration.

Another important feature of our algorithm is that it can deal with missing data easily. We only need to estimate sample moments in Equation (5) based on samples which are complete on the locus or loci involved. The missing locus is then imputed with $\mathcal{H}$ in the subsequent EM procedures (details can be found in the Supplementary Sections 1.2 and 1.3).

The resulting hybrid algorithm, referred as the CSHAP algorithm, incorporates the advantages of both CS theory and EM algorithm. Our algorithm can be more accurate than purely using Equation (5) and more computationally efficient than the current state-of-the-art algorithms.

2.7 CSHAP for large DNA pools

The haplotype estimation methods of pooled DNA data can be divided into two categories: To focus on a small number of pool sizes N (each pool contains about two to three individuals) with large number of markers q, or to consider on a small number of markers q but lager pool sizes N. The former can be solved with EM algorithm (Yang *et al.*, 2003), however, for the other case, the EM algorithm is computationally involved and is not feasible when pool size $N \geq 10$. Kuk *et al.* (2008) proposed an approximate EM (AEM) algorithm based on Central Limit Theorem, which makes the method based on EM algorithm computationally feasible for large N, with substantial improvement in accuracy simultaneously. However, the time consumption of AEM grows exponentially with q, which limits the performance of AEM when q is larger.

In this case, we substitute the EM procedure in CSHAP with the AEM algorithm. The resulting hybrid algorithm runs nearly two orders of magnitude faster than AEM while exhibiting almost same accuracy.

2.8 CSHAP for long-range haplotypes

Due to the computer memory constraint, our algorithm supports at most 24–27 loci for the length of each block. When the number of loci q exceeds this limit, we apply a special PL strategy (Niu *et al.*, 2002).

Assuming the genotype data consists of $L = K \times M$ loci, where M is the number of blocks and K is the length of each block. Without loss of generality, the genotype data can be divided into M contiguous 'atomistic' blocks. In most of the previously published articles (Delaneau *et al.*, 2011; Niu *et al.*, 2002; Qin *et al.*, 2002; Stephens and Scheet, 2005;), K was usually set between 5 and 8 due to the limitation of algorithms. In contrast, our method allows setting $K = 16$–20 while maintaining efficiency. Obviously, larger K and smaller M can help avoid the local-mode problem (Qin *et al.*, 2002) of EM.

Once we have conducted the CSHAP procedure for each of M atomistic blocks, we iteratively combine the $(2i-1)$th block with the subsequent $(2i)$th block until all blocks are ligated as one. Most of the PL methods usually take D haplotypes with top estimated frequencies in each block, where D can be specified between 40 and 50 (Niu *et al.*, 2002) or be determined separately by some frequency thresholds (Stephens and Donnelly, 2003). This 'thrown away' procedure can limit the number of possible concatenated haplotypes to D^2, but may lead the algorithm to trap into a local mode simultaneously. Here we use all estimated haplotypes in each block instead. However, the total number of all possible concatenated haplotypes may be significantly larger than D^2. Notice that only relatively few concatenated haplotypes are true among all those possible haplotypes, we use CS again in this ligation step.

Next we demonstrate how to combine the block A with the adjacent block B. From the estimated haplotype frequencies $\hat{\mathbf{p}}_A$ we obtain a haplotype set $\mathcal{H}_A$ with n_A haplotypes. Denote $H_A = (h_{n_1}, \ldots, h_{n_A})$ is a $q_A \times n_A$ matrix obtained by combining the haplotype vectors in $\mathcal{H}_A$ by column. Similarly, we have $\mathcal{H}_B$ and H_B for block B. Denote the set of all possible concatenated haplotypes as $\mathcal{H}_{A\times B}$. Define H_{AB} is a $(q_A + q_B) \times (n_A n_B)$ matrix which can be regard as the Cartesian product of H_A and H_B. Similar to the structure of Ψ in the (1), we have:

$$\Psi_{AB} = \begin{pmatrix} 1' \\ H_{AB} \\ H_{AB} \wedge H_{AB} \end{pmatrix}$$

and $\Phi_{AB}, \Lambda_{AB}, \mathbf{p}^*_{AB}, \mathbf{b}_{AB}$ correspondingly.

Now we transform the ligation problem into this following problem, which has been already resolved in the above.

$$\min_{\mathbf{p}^*_{AB} \in \mathbb{R}_+} ||\mathbf{p}^*_{AB}||_{\ell_1} \text{ subject to } ||\Phi_{AB}\mathbf{p}^*_{AB} - \hat{\mathbf{b}}_{AB}||_{\ell_2} \leq \epsilon$$

The non-zero entries of $\hat{\mathbf{p}}_{AB}$ correspond to the estimated haplotypes and frequencies in $\mathcal{H}_{A\times B}$. We can also use the EM algorithm to improve the accuracy of $\hat{\mathbf{p}}_{AB}$. This process is repeatable until all blocks are connected as one.

When the number of loci increases to, say, more than 1000s, the frequency estimation is meaningless, since almost every haplotypes are rare. What we actually do at this time is 'phasing', which is a totally different task. However, with a few changes, our algorithm can be used as a phasing tool under this situation. We describe the details and results in the Supplementary Section 1.4.

3 Results

We demonstrate the performances of our algorithm by simulations. First we compare the estimation accuracies of CSHAP, PHASE v2.1.1, fastPHASE v1.4, Shape-IT v2.17 and PL-EM v1.0 (with default settings) using individual genotyping data under the assumption of HWE as well as HWD. In addition, we illustrate the computational efficiency of CSHAP in comparison with the others.

For pooling design, we compare the estimation accuracies of CSHAP, PoooL and AEM with various pool sizes and sample sizes. Meanwhile, we show the running time of CSHAP and AEM. All the

simulations are repeated 10 000 times if not specified explicitly. Our platform is a desktop computer with Ubuntu 16.04 64-bit, Intel Core i5-7400 CPU@3.00 GHz and 8.0 GB RAM.

To summarize the accuracies of our methods, and to compare the accuracies between the different algorithms, we use two measures: the absolute discrepancy (Excoffier and Slatkin, 1995) and the χ^2 distance. The asbolute discrepancy between the estimated haplotype frequencies and their true values is defined as $\sum_{i=1}^{r} |\hat{p}_i - p_i|/2$, and the χ^2 distance is defined as $\sum_{i=1,\ p_i \neq 0}^{r} (\hat{p}_i - p_i)^2/p_i$.

3.1 Individual design

We generate T unrelated individual genotypes according to the commonly used 10-locus haplotype frequencies of angiotensinogen (AGT) gene considered in Yang *et al.* (2003) by assuming HWE, $T = 10, 20, 50, 100, 200, 500, 1000, 2000$. Part of the results is demonstrated in Supplementary Table S2. We can see that for small sample size, the performances of CSHAP are better than the others in the sense of bias and the effective accumulated probability (EAP) of **p**. Also CSHAP has smaller estimation bias of almost every haplotype than PHASE's and Shape-IT's. For larger sample sizes ($T \geq 100$), the precision of CSHAP is very close to the PHASE and PL-EM, and slightly better than Shape-IT.

Simulation results are summarized in Figure 1. We can see that, for small sample size ($T < 100$), the absolute discrepancy of CSHAP is smaller than those of the others, and the solution of CSHAP is much sparser than those of PHASE and Shape-IT. When T exceeds about 100, CSHAP behaves as accurate as PHASE, while slightly better than fastPHASE and Shape-IT.

Comparisons of computational efficiencies are given in Table 1. The CSHAP is at least two to three orders of magnitude faster than PHASE and fastPHASE. For $T = 100$ in our simulation, the CSHAP is 365 times faster more efficient than PHASE while maintaining the same accuracy. For $T = 2000$, the CSHAP is 1775 times faster than PHASE, 2010 times faster than fastPHASE, 468 times faster than Shape-IT and 19 times faster than PL-EM. Notice that the running time of most other methods shows approximately linear trends with sample size T, and in contrast, our method only in the order of a logarithmic scale. More details are in Supplementary Figure S1.

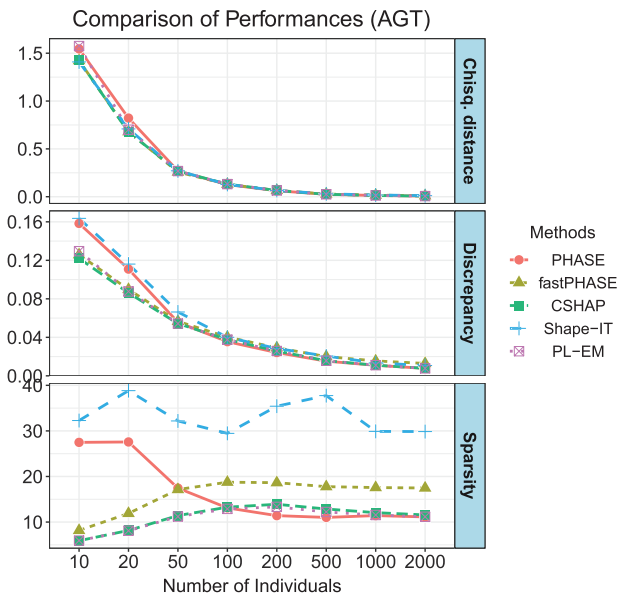

Fig. 1. Average accuracy comparison in 10 000 trials with increasing numbers of sample size T. The top panel shows the χ^2 distance between the true frequency and estimated frequencies. The second panel shows the absolute discrepancy. The bottom panel shows the sparsity (the number of existing haplotypes in solution), respectively. The true haplotype dataset has 11 unique existing haplotypes

3.2 Haplotype diversity and missing data

To measure the performances under different haplotype diversities, we generate $T = 100$ individual genotypes according to the 11-locus G6PD haplotypes in Sabeti *et al.* (2002). The G6PD haplotypes are different among the following six ethnic populations: African American, Asian, Beni (Nigeria), European American, Shona (Zimbabwe) and Yoruba (Nigeria). The published haplotypes and frequencies are given in Supplementary Table S3. In addition, we randomly mask about 5% of the data as missing sites.

As Figure 2 shows, our CSHAP method performs well under varying degrees of diversities, as well as missing data imputation. In fact, the χ^2 distances and the absolute discrepancies of PHASE, fastPHASE, CSHAP are similar. The Shape-IT identified too many non-existent haplotypes, which leads to higher discrepancies. PL-EM seems to have problems accurately imputing missing data in the Asian population. Our CSHAP method has almost the same precision as PHASE, while fastPHASE is only a little less accurate than PHASE, but our method gives a sparser solution with the same precision.

3.3 The case of HWD

We generate $T = 100$ haplotype pairs from (7) using the 10-locus AGT gene with a series of inbreeding coefficients $\rho = 0.05, 0.1, 0.15, 0.2, 0.3$. For each of 10 000 simulation trials, we first construct a bootstrap sample by resampling with replacement from these sample genotypes 1000 times to estimate the variance of $\hat{\rho}$ in (9). The proportion of times out of the simulations that $\hat{\rho}_L \leq 0$ is $0.888, 0.750, 0.502, 0.303, 0.065$, for the above mentioned different values of ρ, respectively.

Results are summarized in the Supplementary Table S4. We can see that introducing a $\hat{\rho}$ helps to reduce the impact of deviations from HWE, but the SDs are slightly larger. Our adaptive estimator $\hat{\rho}_L$ can make this tradeoff conveniently.

3.4 Pooling design

Next we consider large pooling design. The simulations are based on pooled 10-locus AGT gene data with $N = 50$, 100 and $T = 50$, 100 where HWE is assumed. We investigate the performance of CSHAP with PoooL and AEM. To compare with previous published

Table 1. Running times of CSHAP, PHASE, fastPHASE, Shape-IT and PL-EM algorithm for 1000 replicates of simulation

T	CSHAP	PHASE	fastPHASE	Shape-IT	PL-EM
10	10.2	741	338	120	140
20	12.9	1459	651	139	142
50	17.5	3726	1589	246	146
100	21.1	7727	3145	410	156
200	26.0	9851	6309	731	188
500	29.1	16 619	15 746	2007	258
1000	30.8	28 471	32 019	6014	384
2000	32.6	57 990	65 662	15 318	645

Note: Using Yang *et al.* (2003)s 10-locus haplotype frequencies of AGT dataset for individual data. The unit of running time is second. T stands for the sample size.

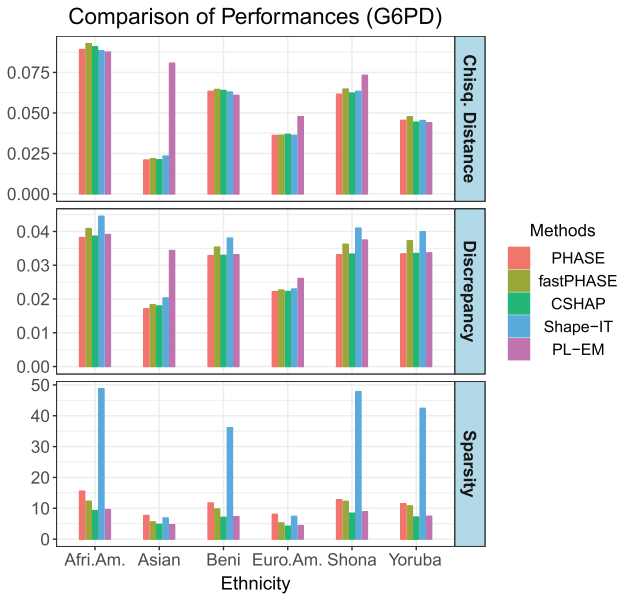

Fig. 2. Measures of performance of PHASE, fastPHASE, CSHAP, Shape-IT and PL-EM based on the simulated G6PD gene datasets among six different populations. Sample size $T = 100$, all the simulations are repeated 10 000 times

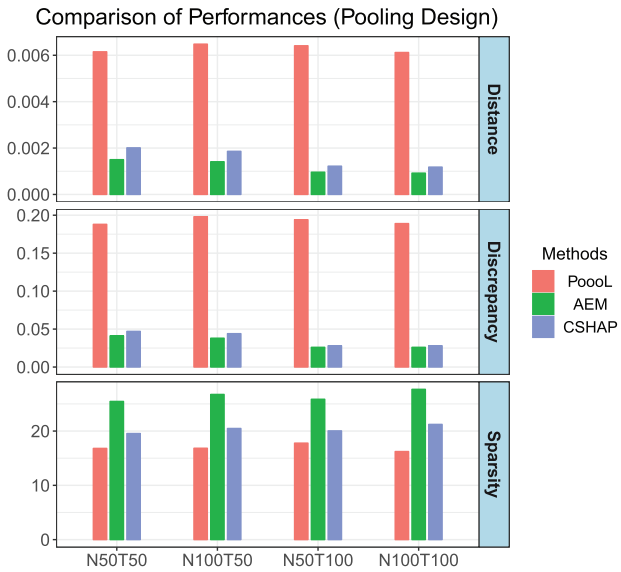

Fig. 3. Measures of performance of PoooL, AEM and CSHAP. *T* stands for the number of pools, and each pool contains *N* individuals. For each pooling design and method, simulation was repeated 10 000 times

results by Kuk *et al.* (2008) and Zhang *et al.* (2008), we use the averaged Euclidean distance $d = \sqrt{\sum_{i=1}^{r} (\hat{p}_i - p_i)^2 / r}$ instead of χ^2 distance.

Figure 3 shows summarized results. The performance of PoooL is not optimal, since the distance and discrepancy are many times higher than those of AEM and CSHAP. Meanwhile, CSHAP shows comparable performance with AEM, the distance and discrepancy are slightly higher, but CSHAP can obtain sparser results than AEM. However, the computational efficiency of CSHAP is about two orders of magnitude faster than that of AEM, Supplementary Table S5 demonstrates that CSHAP runs about 120 times faster than AEM in our simulations.

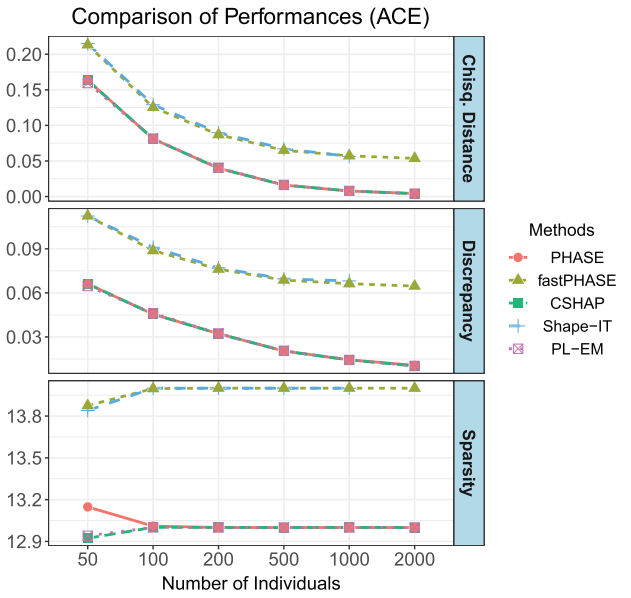

Fig. 4. Measures of performance of PHASE, fastPHASE, CSHAP, Shape-IT and PL-EM. For each method and sample size *T*, simulation was repeated 1000 times. The Shape-IT failed to estimate haplotype frequencies when $T = 2000$, since the program aborted with an error

3.5 Long-range haplotypes

In this section, we use the human angiotensin converting enzyme (ACE) dataset. This dataset contains $q = 52$ SNPs and 11 individuals, the genotypes have up to 37 heterozygous sites. For all individuals, 13 unique haplotypes from the 22 chromosomes are resolved in Rieder *et al.* (1999). We generate $T = 50, 100, 200, 500, 1000, 2000$ unrelated individual genotypes by assuming HWE and test all the methods in Section 3.1. The result is shown in Figure 4. Our method, PHASE and PL-EM produce the best solution, but CSHAP costs much less time. The biases of Shape-IT and fastPHASE are relatively slightly larger.

The running time of all methods are described in Table 2. CSHAP is about two to three orders of magnitude faster than PHASE and fastPHASE, while providing the most accurate results. For example, when $T = 2000$, the CSHAP is 3000 times faster than PHASE, 1634 times faster than fastPHASE, 303 times faster than Shape-IT and 10 times faster than PL-EM. More details are in Supplementary Figure S2.

4 Conclusion

In this study, we propose an efficient algorithm, CSHAP, for estimating haplotype frequencies from individual or pooled DNA data, under the HWE assumption or not. The CSHAP algorithm minimizes the weighted sum of haplotype frequencies under constraints on the allele frequencies and covariances (i.e. LD coefficients). This method is based on the maximum parsimony principle of Gusfield (2003), which was to minimize the total number of distinct haplotypes, subject to the condition that the solutions are consistent with genotype observations. In our approach, we substitute the ℓ_0 norm by the ℓ_1 norm and reduce the consistency condition to a system of linear constraints on the first two moments of genotype observations. Besides, we use a modified EM algorithm to boost accuracy efficiently.

Extensive simulation studies show that our method is comparable to or better than the existing methods but has significant

Table 2. Running times of CSHAP, PHASE, fastPHASE, Shape-IT and PL-EM algorithm for 100 replicates of simulation

T	CSHAP	PHASE	fastPHASE	Shape-IT	PL-EM
50	9.5	4706	821	98	63
100	9.7	9571	1684	228	84
200	10.9	12 176	3324	457	93
500	12.0	19 687	8283	1196	107
1000	14.1	33 783	16 477	3250	140
2000	20.2	60 690	33 083	6153	211

Note: Using Rieder *et al.* (1999)s 52-locus haplotype frequencies on ACE dataset. The unit of running time is second. *T* stands for the sample size.

computational advantage. An outstanding feature of our method is its computing efficiency for both individual and pooled DNA data. We showed that if the number of existing haplotypes in population is small, our method can recover the haplotypes with high accuracy. In addition, by introducing the inbreeding coefficients the HWE assumption is not required in our approach and this quantity can also capture the inflation of variance in genotype observations caused by genotype errors, population substructures etc.

Furthermore, by introducing the divide-and-conquer idea of PL, our method is able to handle long-range haplotypes. The specially designed PL algorithm can help us to overcome the local-mode problem while maintaining efficiency. In conclusion, our method can be a powerful and efficient approach in genome-wide association studies.

Acknowledgements

We appreciate the insightful comments and helps from Dr Xing Hua and Dr Yifang Yang during preparation of this article.

Funding

This work has been supported by the National Science Foundation of China [NSFC 11671375 and 11271346].

Conflict of Interest: none declared.

References

Bilodeau,M. and Brenner,D. (2008) *Theory of Multivariate Statistics*. Springer, New York.

Boyd,S. and Vandenberghe,L. (2004) *Convex Optimization*. Cambridge University Press, Cambridge.

Browning,S.R. and Browning,B.L. (2011) Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.*, **12**, 703–714.

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

Daly,M.J. *et al.* (2001) High-resolution haplotype structure in the human genome. *Nat. Genet.*, **29**, 229–232.

Delaneau,O. *et al.* (2008) Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics*, **9**, 540.

Delaneau,O. *et al.* (2011) A linear complexity phasing method for thousands of genomes. *Nat. Methods*, **9**, 179–181.

Delaneau,O. *et al.* (2013) Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods*, **10**, 5–6.

Delaneau,O. *et al.* (2014) Integrating sequence and array data to create an improved 1000 genomes project haplotype reference panel. *Nat. Commun.*, **5**, 3934.

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

Gusfield,D. (2001) Inference of haplotypes from samples of diploid populations: complexity and algorithms. *J. Comput. Biol.*, **8**, 305–323.

Gusfield,D. (2003) Haplotype inference by pure parsimony. In: *Proceedings of the 14th Annual Conference on Combinatorial Pattern Matching, CPM'03*, Springer-Verlag, Berlin, Heidelberg, pp. 144–155.

Gusfield,D. and Orzack,S. (2005) *Haplotype Inference. CRC Handbook on Bioinformatics, Chapter 1*. CRC Press, Boca Raton, USA.

Hiriart-Urruty,J.-B. and Lemaréchal,C. (1993) *Convex Analysis and Minimization Algorithms II*. Springer, Berlin, Heidelberg.

Howie,B. *et al.* (2011) Genotype imputation with thousands of genomes. *G3*, **1**, 457–470.

Jajamovich,G.H. and Wang,X. (2012) Maximum-parsimony haplotype inference based on sparse representations of genotypes. *IEEE Trans. Sig. Process.*, **60**, 2013–2023.

Kuk,A.Y. *et al.* (2009) Computationally feasible estimation of haplotype frequencies from pooled DNA with and without Hardy-Weinberg equilibrium. *Bioinformatics*, **25**, 379–386.

Lin,S. *et al.* (2002) Haplotype inference in random population samples. *Am. J. Hum. Genet.*, **71**, 1129–1137.

Liu,N. *et al.* (2008) Haplotype-association analysis. *Adv. Genet.*, **60**, 335–405.

Niu,T. *et al.* (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am. J. Hum. Genet.*, **70**, 157–169.

Patil,N. *et al.* (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294**, 1719–1723.

Qin,Z.S. *et al.* (2002) Partition-ligation–expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am. J. Hum. Genet.*, **71**, 1242–1247.

Recht,B. *et al.* (2010) Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, **52**, 471–501.

Rieder,M.J. *et al.* (1999) Sequence variation in the human angiotensin converting enzyme. *Nat. Genet.*, **22**, 59.

Sabeti,P.C. *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.

Scheet,P. and Stephens,M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.*, **78**, 629–644.

Sham,P. *et al.* (2002) Dna pooling: a tool for large-scale association studies. *Nat. Rev. Genet.*, **3**, 862–871.

Stephens,M. and Donnelly,P. (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.*, **73**, 1162–1169.

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

Stephens,M. *et al.* (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, **68**, 978–989.

Xing,E.P. *et al.* (2007) Bayesian haplotype inference via the Dirichlet process. *J. Comput. Biol.*, **14**, 267–284.

Yang,Y. *et al.* (2003) Efficiency of single-nucleotide polymorphism haplotype estimation from pooled dna. *Proc. Natl. Acad. Sci. USA*, **100**, 7225–7230.

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

Zhang,H. *et al.* (2008) PoooL: an efficient method for estimating haplotype frequencies from large DNA pools. *Bioinformatics*, **24**, 1942–1948.

Zhang,Y. *et al.* (2006) A coalescence-guided hierarchical bayesian method for haplotype inference. *Am. J. Hum. Genet.*, **79**, 313–322.