Supplemental Material and Methods of "CSHAP:Efficient haplotype frequency estimation based on sparse representation"

Yinsheng Zhou, Han Zhang and Yaning Yang

2018-12-12

Contents

1 Supplemental Methods

1.1 Derivation of Variance-Covariance Matrix Σ_e

The unbiased estimator $\widehat{\mathbf{b}} = \widehat{\mathbf{b}}(\widehat{\omega}_0, \widehat{\eta}_0)$ can be expressed as $\widehat{\mathbf{b}} = (1, \widehat{\omega}, \widehat{\eta}_{ij})' = (\mathbf{b}_1, \widehat{\mathbf{b}}_2, \widehat{\mathbf{b}}_3)'$, where $\mathbf{b}_1 \equiv 1$, $\widehat{\mathbf{b}}_2 = (\widehat{\omega}_1, \ldots, \widehat{\omega}_q)$, $\widehat{\mathbf{b}}_3 = (\widehat{\eta}_{12}, \ldots, \widehat{\eta}_{1q}; \widehat{\eta}_{23}, \ldots, \widehat{\eta}_{2q}; \ldots; \widehat{\eta}_{q-1,q}) = (\widehat{\eta}_{i,j})$, $1 \le i \le q-1$, $i+1 \le j \le q$. All elements of Σ_e is given by $\mathrm{Cov}\,(\widehat{\omega}_i, \widehat{\omega}_j)$, $\mathrm{Cov}\,(\widehat{\eta}_{i,k}, \widehat{\omega}_l)$ and $\mathrm{Cov}\,(\widehat{\eta}_{i,k}, \widehat{\eta}_{j,l})$. Define $\mu^G = \left(\mu_i^G, \ldots, \mu_q^G\right)'$ as the mean vector of pooled genotypes G_i, $\bar{G} = \sum_{i=1}^T G_i/T = \left(\bar{G}_{(1)}, \ldots, \bar{G}_{(q)}\right)'$ as the sample mean vector, $\widehat{\mu}^G = \bar{G} = 2N\widehat{\omega}$. Let $\Sigma^G = (\sigma_{i,j}^G)_{q\times q}$ represents the variance-covariance matrix of pooled genotypes G_i, $S^G = (s_{i,j}^G)_{q\times q}$ represents the sample variance covariance matrix of genotype observations, or $S^G = \widehat{\Sigma}^G = \sum_{i=1}^T (G_i - \bar{G})(G_i - \bar{G})'/T$ equivalently. From multivariate sampling we see that

$$\begin{aligned} \mathrm{Var}(\widehat{\omega}) &= \frac{1}{4N^2T}\Sigma^G \\ \mathrm{Cov}\,(\widehat{\omega}_i, \widehat{\omega}_j) &= \sigma_{i,j}^G \end{aligned} \tag{1}$$

According to Chapter 6.2 in Bilodeau and Brenner [1], we have

$$\begin{aligned} (T-1)S^G &\xrightarrow{d} \mathcal{W}_q\left(T-1, \Sigma^G\right) \\ \mathrm{Var}(S^G) &= \frac{1}{T-1}(I_q + K_q)\left(\Sigma^G \otimes \Sigma^G\right) \end{aligned}$$

We can also write this expression componentwise as

$$\mathrm{Cov}\left(s_{i,k}^G, s_{j,l}^G\right) = \frac{1}{T-1}\left(\sigma_{i,j}^G\sigma_{k,l}^G + \sigma_{k,j}^G\sigma_{i,l}^G\right) \tag{2}$$

Now, rewrite $\widehat{\eta} = \widehat{\Sigma}_0 + \widehat{\omega}\widehat{\omega}'$ componentwise as

$$\widehat{\eta}_{i,j} = \widehat{\sigma}_{i,j}^0 + \widehat{\omega}_i\widehat{\omega}_j = \frac{1}{2N}s_{i,j}^G + \frac{1}{4N^2}\bar{G}_{(i)}\bar{G}_{(j)}$$

This lead to

$$\begin{aligned} \mathrm{Cov}\,(\widehat{\eta}_{i,k}, \widehat{\omega}_l) &= \mathrm{Cov}\left(\frac{1}{2N}s_{i,k}^G + \frac{1}{4N^2}\bar{G}_{(i)}\bar{G}_{(k)}, \bar{G}_{(l)}\right) \\ &= \mathrm{Cov}\left(\frac{1}{2N}s_{i,k}^G, \bar{G}_{(l)}\right) + \mathrm{Cov}\left(\frac{1}{4N^2}\bar{G}_{(i)}\bar{G}_{(k)}, \bar{G}_{(l)}\right) \\ &= 0 + \frac{1}{8N^3T}\left(\mu_i^G\sigma_{k,l}^G + \mu_k^G\sigma_{i,l}^G\right) \end{aligned} \tag{3}$$

The first part becomes 0 because $\bar{G} \perp S^G$, the second part is provided by multivariate cumulants. Also,

$$\begin{aligned} \mathrm{Cov}\,(\widehat{\eta}_{i,k}, \widehat{\eta}_{j,l}) &= \mathrm{Cov}\left(\frac{1}{2N}s_{i,k}^G + \frac{\bar{G}_{(i)}\bar{G}_{(k)}}{4N^2}, \frac{1}{2N}s_{j,l}^G + \frac{\bar{G}_{(j)}\bar{G}_{(l)}}{4N^2}\right) \\ &= \frac{1}{4N^2}\,\mathrm{Cov}\left(s_{i,k}^G, s_{j,l}^G\right) + \frac{1}{\left(4N^2\right)^2}\,\mathrm{Cov}\left(\,\bar{G}_{(k)}, \bar{G}_{(j)}\bar{G}_{(l)}\right) + \frac{1}{8N^3}\left[\mathrm{Cov}\left(s_{i,k}^G, \bar{G}_{(j)}\bar{G}_{(l)}\right) + \mathrm{Cov}\left(s_{j,l}^G, \bar{G}_{(i)}\bar{G}_{(k)}\right)\right] \\ &\triangleq \mathrm{I} + \mathrm{II} + \mathrm{III} \end{aligned} \tag{4}$$

Respectively, we have

$$\begin{aligned} \mathrm{I} &= \frac{1}{4N^2(T-1)}\left(\sigma_{i,j}^G\sigma_{k,l}^G + \sigma_{k,j}^G\sigma_{i,l}^G\right) \\ \mathrm{II} &= \frac{1}{(4N^2)^2}\left\{\frac{1}{T}\mu_i^G\mu_j^G\sigma_{k,l}^G + \frac{1}{T}\mu_i^G\mu_l^G\sigma_{k,j}^G + \frac{1}{T}\mu_k^G\mu_l^G\sigma_{i,j}^G + \frac{1}{T}\mu_k^G\mu_j^G\sigma_{i,l}^G + \frac{1}{T^2}\sigma_{i,j}^G\sigma_{k,l}^G + \frac{1}{T^2}\sigma_{i,l}^G\sigma_{k,j}^G\right\} \\ \mathrm{III} &= 0, \text{ because } \bar{G} \perp S^G. \end{aligned}$$

Hence, by (1)(3)(4) we can calculate the elements in the variance-covariance matrix $\mathrm{Var}(\widehat{\mathbf{b}})$ one by one to construct the entire Σ_e, or its estimator $\widehat{\Sigma}_e$ similarly.

1.2 Missing Data Imputation

For an individual, the genotype data can be represented as a vector $g = (g_1, g_2, \ldots, g_q) \in \{0, 1, 2, 3\}^q$. Here $0, 2$ stand for homozygous loci, 1 stands for heterozygous loci, and 3 stands for missing data, that is, no information is observed at some specific loci. $h_k \oplus h_j = g$ denotes that a haplotype pair $[h_k, h_l]$ is compatible with the obersved genotype data g, then $[h_k, h_l]$ is called a *resolution* of g. Define $\mathcal{S}(g) = \{\, [h_k, h_l] \mid h_k \oplus h_j = g\}$ as the set of g's all resolutions.

When g has no missing data, It is easy to enumerate the set $\mathcal{S}(g)$, then the EM alogrithmn can be used. Now we demonstrate how to get $\mathcal{S}(g)$ when g contains missing data. Suppose g has n missing locus located separately at $i_1, i_2, \ldots, i_n$. Denote a new vector $f \in \{0, 1, 2\}^q$, the element of which are assigned as

$$f_i = \begin{cases} g_i & \text{if } g_i \neq 3, \\ 0 & \text{if } g_i = 3. \end{cases}$$

In other words, we first regard all the missing locus in g as 0, and constitute $\mathcal{S}(f)$. Next we consider g's first missing loci i_1. For every element $[h_k, h_l]$ in $\mathcal{S}(f)$, define a function `IMPUTE` as

$$\texttt{IMPUTE}([h_k, h_l]) = \{[h_k, h_l],\, [h_k^1, h_l],\, [h_k, h_l^1],\, [h_k^1, h_l^1]\}$$

where $h_k^1 \in \{0, 1\}^q$ is a haplotype vector which differs from h_k only at locus i_1. Formally, the i-th element of h_k^1 is defined as

$$h_k^1(i) = \begin{cases} h_k(i) & \text{if } i \neq i_1, \\ 1 & \text{if } i = i_1. \end{cases}$$

When we apply the function `IMPUTE` to the set $\mathcal{S}$, it means

$$\texttt{IMPUTE}(\mathcal{S}) = \bigcup_{[h_k, h_l] \in \mathcal{S}} \texttt{IMPUTE}([h_k, h_l])$$

Let $\mathcal{S}^1(f) = \texttt{IMPUTE}(\mathcal{S}(f))$, $\mathcal{S}^2(f) = \texttt{IMPUTE}(\mathcal{S}^1(f))$, $\ldots$, $\mathcal{S}^n(f) = \texttt{IMPUTE}(\mathcal{S}^{n-1}(f))$. Obviously the $\mathcal{S}(g) = \mathcal{S}^n(f)$ is actually what we want, then the EM algorithm can be implemented on $\mathcal{S}^n(f)$. However, we notice that $|\mathcal{S}^n(f)| = 4^n |\mathcal{S}(f)|$, for every missing site, the number of all possible resolutions will increase by 3 times, which leads to an increasement in computational complexity. In CSHAP, we can overcome this problem by limiting the number of elements in the $\mathcal{S}(g)$ as

$$\mathcal{S}_{\mathcal{H}}(g) = \{[h_k, h_l] \mid\ [h_k, h_l] \in \mathcal{S}^n(f), h_k \in \mathcal{H}\}$$

Notice that the missing data in samples has no impact on the estimation of sample moments, since we only need to estimate sample moments based on samples which are complete on the locus or loci involved. So that the main procedure still works, we can always get $\mathcal{H}$ whether there is missing data.

1.3 Brief Introduction of EM Algorithm

The expectation-maximization (EM) algorithm is a likelihood-based method which regards the phase as unobserved latent variables[2–5]. The algorithm starts with an initial haplotype frequencies, e.g. $p_1^{(0)} = p_2^{(0)} = \ldots = p_r^{(0)} = 1/r, r = 2^q$. In step t ($t = 0, 1, 2, ...$), the current value of haplotype frequencies are denoted as a vector $\mathbf{p}^{(t)}$, the posterior of each resolution (the possible phase in $\mathcal{S}(g)$) of each individual are calculated based on HWE (E-Step). Then the new frequencies $\mathbf{p}^{(t+1)}$ are estimated based on the posterior probabilities (M-Step). The E-Step and M-Step are repeated alternately until convergence is reached (i.e. $|\mathbf{p}^{(t+1)} - \mathbf{p}^{(t)}| < \epsilon$). Formally, the EM algorithm for haplotype estimation is described as follows.

E-Step

$$P^{(t)}([h_k, h_l]) = \begin{cases} p_k^{(t)} p_l^{(t)} & \text{if } k \neq l, \\ p_k^{(t)} p_l^{(t)}/2 & \text{if } k = l. \end{cases}$$

Denote G_i as the genotype of the i-th individual, then the posterior probability of each resolution $[h_k, h_l] \in \mathcal{S}(G_i)$ is estimated as

$$P^{(t)}(G_i) = \sum_{[h_k,h_l]\in\mathcal{S}(G_i)} P^{(t)}([h_k, h_l])$$

$$P^{(t)}([h_k, h_l]|G_i) = \frac{P^{(t)}([h_k, h_l])}{P^{(t)}(G_i)}$$

M-Step For each $j = 1, 2, \cdots, r$, update haplotype frequencies as

$$p_j^{(t+1)} = \frac{1}{2} \sum_i \sum_{[h_k,h_l]\in\mathcal{S}(G_i)} (\mathbb{I}_{j=k} + \mathbb{I}_{j=l})\, P^{(t)}([h_k, h_l]|G_i)$$

Phasing For an individual, denote his/her genotype as g, the haplotype pair with the highest posterior probability is used as an estimate of phase.

$$[\hat{h}_k, \hat{h}_l] = \underset{k,l}{\operatorname{argmax}}\, P(h_k, h_l|g)$$

Note that the standard EM algorithm does not use any prior biological knowledge, and does not make any assumptions about sparsity property. Since the total number of all haplotypes r increases exponentially with the number of sites q, the maximum number of q that EM method can handle is usually very small. On the other hand, the EM method requires to enumerate all possible resolutions $\mathcal{S}(g)$ from all haplotypes for all individuals, this leads to a heavy computational burden. When there are some missing sites in the genotype data, this defect is more severe since every missing site will quadruple $|\mathcal{S}(g)|$.

Our CSHAP methods use the sparse solution $\widehat{\mathbf{p}}$ of equation (5) and corresponding haplotype reference set $\mathcal{H}$ as the initial haplotype guess. Most individuals have only one possible resolution in $\mathcal{S}_{\mathcal{H}}(g)$. This can greatly reduce the number of haplotypes that need to be considered, and substantially reduce the number of iterations needed for a convergence, especially so when missing data are present in the genotype data.

1.4 Extra Long-range haplotype estimation

1.4.1 Methods

When the number of loci q increases to the scale of 1000s, the frequency estimation is meaningless, since almost every haplotypes are rare. What we actually need to do is "phasing", which is a totally different question. At this point, we don't need to connect all the blocks as one. Instead, we only need to determine how the haplotypes are "transferred" between blocks, since the linkage disequilibrium (LD) is decay with distance of genotyped markers.

First, We carry out $1 \sim 2$ ligation steps, to make "atomistic" blocks be connected into longer blocks. For each individual and each block, we select best phase estimate based on estimated frequencies. Note that most individuals have only 1 major phase estimation at this time.

Suppose we have two adjacent blocks called A and B. Denote $h_i^A h_j^B$ as the concatenation of i-th haplotype on the block A and j-th haplotype on the block B. p_i^A as the haplotype frequency of i-th haplotype h_i^A on the block A, p_j^B as the haplotype frequency of j-th haplotype h_j^B on the block B. Denote $P(i,j)$ as the haplotype frequency of $h_i^A h_j^B$ on the new block formed by the ligation of A and B. Let $P(i \to j) = P(i,j)/p_i^A$ as the "transition probability" of $h_i^A \to h_j^B$.

For a specific individual, suppose the partial genotype data on A and B are heterozygous (i.e. have at least one or more heterozygous sites on both A and B). And the best phase guess are $h_i^A \oplus h_k^A$ and $h_j^B \oplus h_l^B$ respectively. Then we calculate a statistic called "LOR" as

$$LOR = \log \frac{P(i \to j)P(k \to l)}{P(i \to l)P(k \to j)}$$

We choose $h_i^A h_j^B \oplus h_k^A h_l^B$ as the phase estimation upon AB if $LOR > 0$, and $h_i^A h_l^B \oplus h_k^A h_j^B$ on the contrary. Then we consider the transition of B and C (the next adjacent block of B), and so on. For some individuals there may exist some homozygous blocks, we skip these homozygous blocks and only consider the transition between the heterozygous blocks, since the phase on homozygous blocks are unambiguous. (i.e. If the block B is homozygous for this individual, we ligate A and C instead of A and B).

The idea we used here is similar to the 2SNP method in Brinza and Zelikovsky [6], but we only need one ligation step to determine the transition of all individuals on AB.

1.4.2 Results

We use a larger dataset from the HapMap database [7] to compare the performances. We choose the genotype data in ENr113 region of chromosome band 4q26 of USA Utah residents. The original data includes 30 phased mother, father, child trios genotypes with the number of locus $q = 1393$. We discard the child's genotype data, randomly combine the rest 60 parental haplotypes into 30 new diplotypes, and regard them as unrelated. This step is repeated 100 times to generate 100 sets of unrelated genotype data. Upon these trials, we use various phasing methods and compare the average Switch Error Rate [8] (SER). The results are shown in Table S1 below.

Table S1: Mean SER and runtime of phasing algorithms on the HapMap dataset

Methods	SER	SD	Time (s)
PHASE	—	—	—
fastPHASE	3.38%	1.02%	288.9
CSHAP	2.20%	0.51%	7.7
Shape-IT	1.49%	0.61%	18.5
PL-EM	—	—	—

Note: All the simulations are repeated 100 times.
The PHASE and PL-EM failed to gain a solution in a week.

From the result, we can see that although the switch error rate of CSHAP is not the best, it's better than fastPHASE, and the running time of CSHAP is lower than others'. The Shape-IT-gets the lowest switch error rate, since it was designed primarily for phasing instead of frequency estimation.

2 Supplemental Figures and Tables

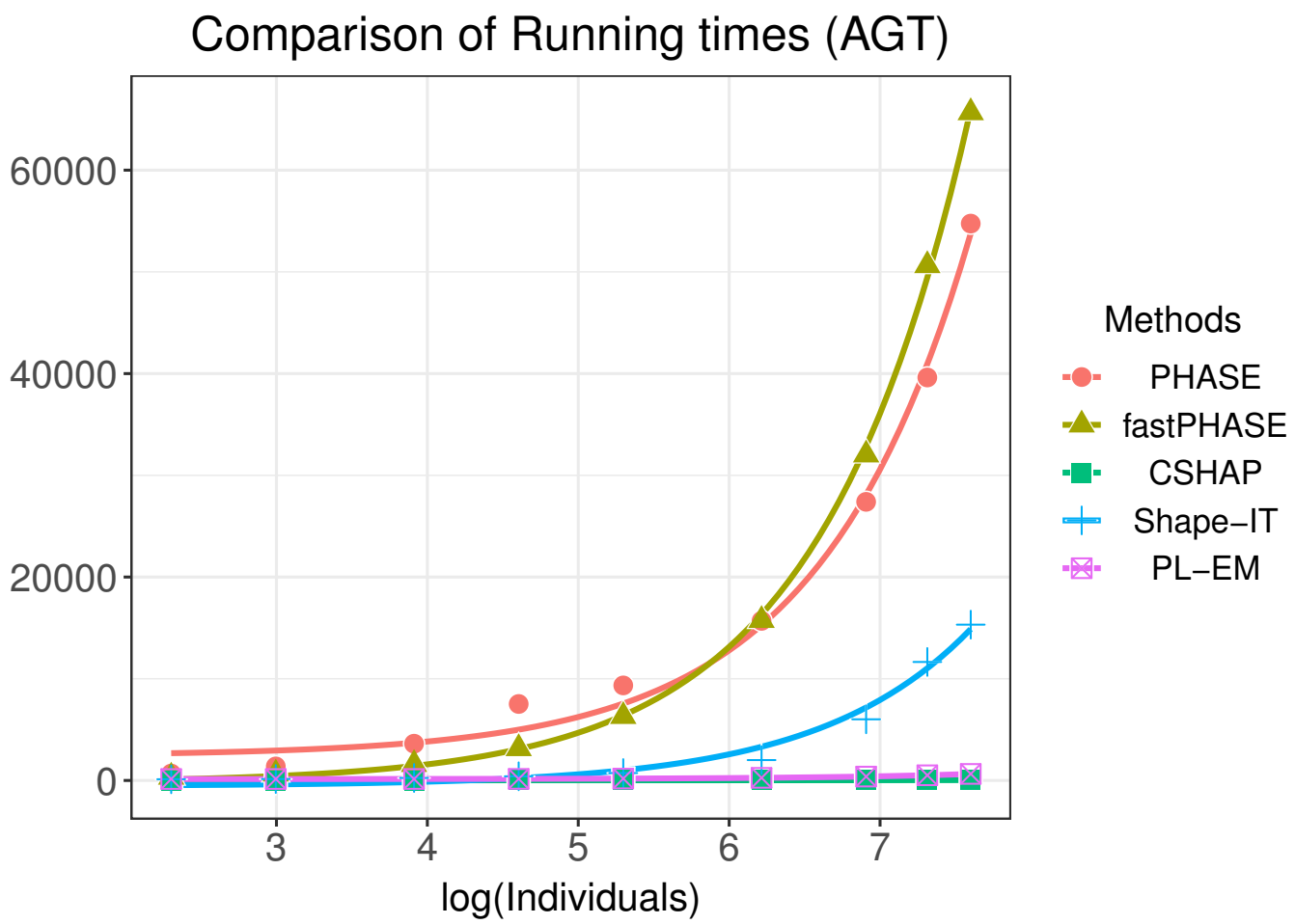

Figure S1: Running times of CSHAP, PHASE, fastPHASE, Shape-IT and PL-EM algorithm for 1,000 replicates of simulation on AGT dataset for individual data. The number of individuals T ranges from 10 to 2000. The time costs of PHASE, fastPHASE and Shape-IT are represented as exponential increase with $\log T$, while CSHAP just scales linear.

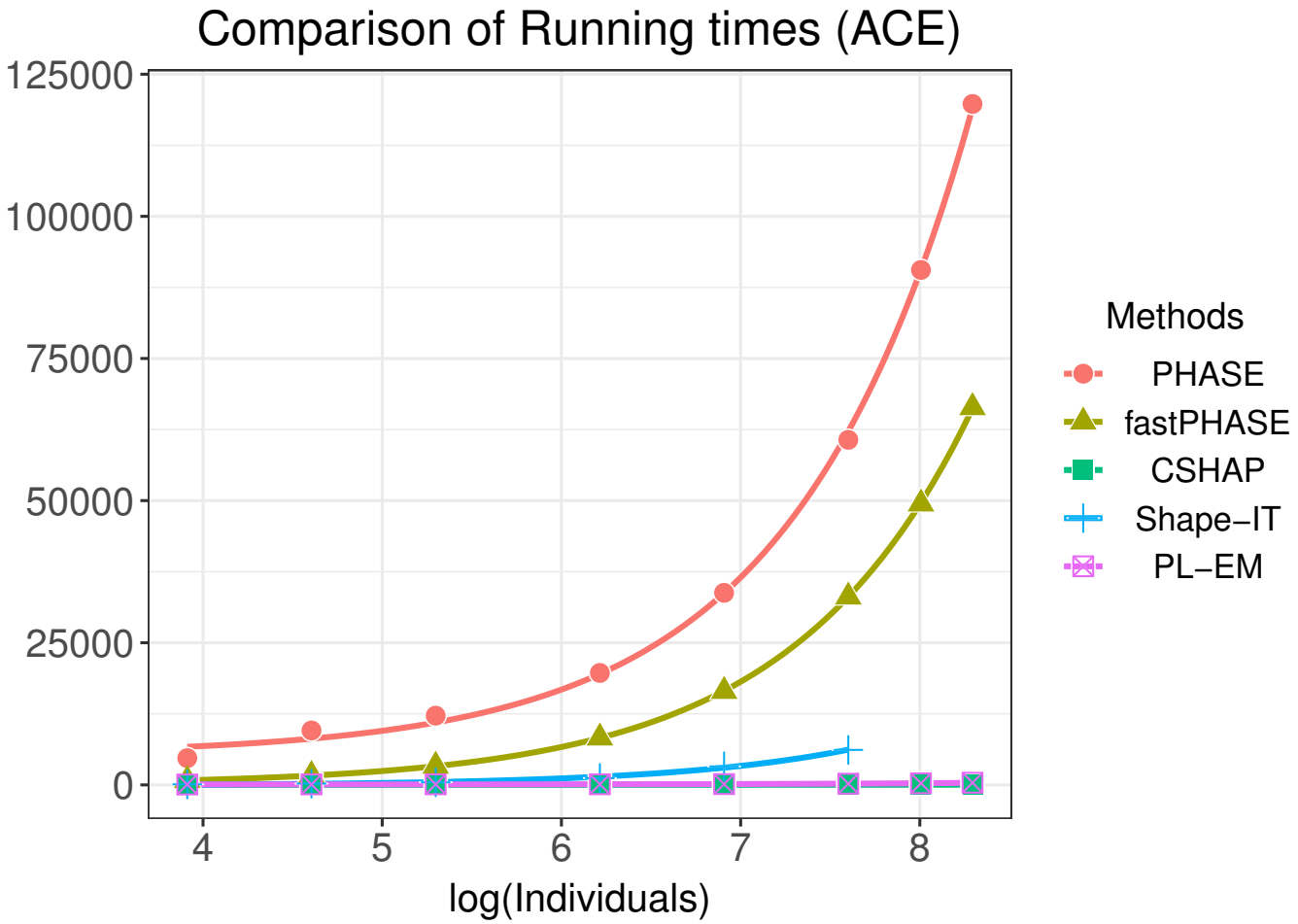

Figure S2: Running times of CSHAP, PHASE, fastPHASE, Shape-IT and PL-EM algorithm for 100 replicates of simulation on ACE dataset for individual data. The number of individuals T ranges from 50 to 4000.

Table S2: Comparison of PHASE, fastPHASE, CSHAP, Shape-IT and PL-EM algorithms for frequency estimation in individual design under HWE (SDs are in parentheses)

		$T = 10$				
SNP	**p**	PHASE	fastPHASE	CSHAP	Shape-IT	PL-EM
0101001111	0.100	0.071 (0.069)	0.086 (0.069)	0.087 (0.071)	0.070 (0.067)	0.085 (0.071)
0101100111	0.017	0.006 (0.018)	0.012 (0.025)	0.013 (0.026)	0.005 (0.015)	0.013 (0.026)
1001011001	0.017	0.007 (0.018)	0.012 (0.026)	0.013 (0.026)	0.006 (0.015)	0.013 (0.026)
1100001111	0.033	0.017 (0.031)	0.027 (0.039)	0.027 (0.038)	0.017 (0.030)	0.026 (0.038)
1101001001	0.017	0.008 (0.018)	0.013 (0.025)	0.015 (0.027)	0.008 (0.017)	0.014 (0.027)
1101011001	0.017	0.017 (0.025)	0.015 (0.028)	0.015 (0.028)	0.019 (0.027)	0.014 (0.028)
1101011110	0.016	0.009 (0.019)	0.011 (0.023)	0.012 (0.025)	0.010 (0.019)	0.012 (0.025)
1101011111	0.193	0.178 (0.090)	0.201 (0.100)	0.203 (0.102)	0.170 (0.084)	0.196 (0.106)
1111011000	0.033	0.018 (0.031)	0.021 (0.034)	0.021 (0.035)	0.015 (0.025)	0.021 (0.035)
1111011101	0.507	0.437 (0.132)	0.500 (0.127)	0.513 (0.123)	0.422 (0.113)	0.507 (0.136)
1111111111	0.050	0.033 (0.042)	0.041 (0.047)	0.041 (0.047)	0.029 (0.038)	0.039 (0.047)
Total	1.000	0.801	0.939	0.960	0.771	0.940

		$T = 100$				
SNP	**p**	PHASE	fastPHASE	CSHAP	Shape-IT	PL-EM
0101001111	0.100	0.100 (0.021)	0.098 (0.022)	0.099 (0.022)	0.101 (0.021)	0.099 (0.022)
0101100111	0.017	0.016 (0.010)	0.015 (0.009)	0.016 (0.009)	0.013 (0.011)	0.016 (0.009)
1001011001	0.017	0.017 (0.010)	0.015 (0.009)	0.016 (0.009)	0.015 (0.010)	0.016 (0.009)
1100001111	0.033	0.033 (0.013)	0.031 (0.013)	0.031 (0.013)	0.033 (0.014)	0.031 (0.013)
1101001001	0.017	0.016 (0.010)	0.015 (0.009)	0.016 (0.009)	0.016 (0.010)	0.016 (0.009)
1101011001	0.017	0.018 (0.010)	0.017 (0.010)	0.016 (0.009)	0.020 (0.010)	0.016 (0.009)
1101011110	0.016	0.015 (0.009)	0.014 (0.009)	0.015 (0.009)	0.015 (0.010)	0.015 (0.009)
1101011111	0.193	0.193 (0.028)	0.192 (0.029)	0.195 (0.029)	0.192 (0.028)	0.195 (0.029)
1111011000	0.033	0.033 (0.013)	0.030 (0.013)	0.030 (0.013)	0.031 (0.013)	0.029 (0.013)
1111011101	0.507	0.505 (0.035)	0.503 (0.036)	0.511 (0.036)	0.499 (0.034)	0.512 (0.036)
1111111111	0.050	0.050 (0.016)	0.047 (0.016)	0.047 (0.016)	0.049 (0.016)	0.047 (0.016)
Total	1.000	0.996	0.977	0.992	0.984	0.992

Using Yang et al.[9]'s 10-locus haplotype frequencies **p** as gold-standard. For each simulation, T unrelated individual genotypes are randomly generated by assuming HWE. All the simulations are repeated $10,000$ times.

Table S3: Haplotypes frequencies for the G6PD gene in six populations

SNP	Afri.Am.	Asian	Beni	Euro.Am.	Shona	Yoruba
00000000000	0.133	0.067	0.150	0.031	0.133	0.207
00000000001	0.067	0.800	—	0.569	—	—
00000000010	0.189	0.067	0.233	0.169	0.446	0.241
00000001000	0.011	—	0.017	—	0.036	—
00000001100	0.144	—	0.217	—	0.060	0.115
00000100000	0.222	—	0.167	—	0.108	0.103
00110110000	0.033	—	0.050	—	0.060	0.069
01111110000	0.133	—	0.167	—	0.133	0.230
10000000001	0.067	0.067	—	0.231	0.024	0.034

From published haplotype data in Sabeti et al. [10].

Table S4: Estimates of haplotype frequencies with variant inbreeding coefficients for individual design (SDs are in parentheses)

SNP	$\mathbf{p}$	$\widehat{\mathbf{p}}$ (SD)				
		$\rho = 0.05$	$\rho = 0.1$	$\rho = 0.15$	$\rho = 0.2$	$\rho = 0.3$
0101001111	0.100	0.099 (0.022)	0.099 (0.023)	0.099 (0.023)	0.099 (0.023)	0.099 (0.025)
0101100111	0.017	0.016 (0.009)	0.016 (0.010)	0.016 (0.010)	0.016 (0.010)	0.016 (0.010)
1001011001	0.017	0.016 (0.009)	0.016 (0.010)	0.016 (0.010)	0.016 (0.010)	0.016 (0.011)
1100001111	0.033	0.031 (0.013)	0.032 (0.014)	0.032 (0.014)	0.032 (0.014)	0.032 (0.015)
1101001001	0.017	0.016 (0.009)	0.016 (0.010)	0.016 (0.010)	0.016 (0.010)	0.016 (0.011)
1101011001	0.017	0.016 (0.010)	0.016 (0.010)	0.016 (0.010)	0.016 (0.010)	0.016 (0.011)
1101011110	0.016	0.015 (0.009)	0.015 (0.010)	0.015 (0.010)	0.015 (0.010)	0.015 (0.010)
1101011111	0.193	0.195 (0.030)	0.195 (0.030)	0.195 (0.031)	0.196 (0.031)	0.195 (0.033)
1111011000	0.033	0.029 (0.014)	0.030 (0.014)	0.030 (0.015)	0.030 (0.015)	0.031 (0.015)
1111011101	0.507	0.513 (0.037)	0.512 (0.038)	0.511 (0.038)	0.511 (0.039)	0.511 (0.041)
1111111111	0.050	0.047 (0.016)	0.047 (0.016)	0.048 (0.017)	0.048 (0.017)	0.048 (0.018)
Total	1.000	0.992	0.992	0.993	0.993	0.994

Using Yang et al.[9]'s 10-locus haplotype frequencies. For each simulation, $T = 100$ unrelated individual genotypes are randomly generated with respective inbreeding coefficients ρ. All the simulations are repeated $10,000$ times.

Table S5: Running times of CSHAP and AEM algorithm for 1,000 replicates of simulation.

N	T	CSHAP	AEM
50	50	108	13350
50	100	115	14644
100	50	184	21709
100	100	192	24031

Using Yang et al.[9]'s 10-locus haplotype frequencies of AGT dataset for pooled data.
The unit of running time is second. T: pool number, N: pool size.

References

[1] Bilodeau, M. and Brenner, D. (2008). *Theory of multivariate statistics*. Springer Science & Business Media.

[2] Hawley, M. and Kidd, K. (1995). Haplo: a program using the em algorithm to estimate the frequencies of multi-site haplotypes. Journal of Heredity *86*, 409–411.

[3] Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Molecular biology and evolution *12*, 921–927.

[4] Fallin, D. and Schork, N. J. (2000). Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. The American Journal of Human Genetics *67*, 947–959.

[5] Polańska, J. (2003). The em algorithm and its implementation for the estimation of frequencies of snp-haplotypes. International Journal of Applied Mathematics and Computer Science *3*, 419–429.

[6] Brinza, D. and Zelikovsky, A. (2008). 2snp: Scalable phasing method for trios and unrelated individuals. IEEE/ACM Transactions on Computational Biology and Bioinformatics *5*, 313–318.

[7] Consortium, T. I. H., Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B. *et al.* (2003). The international hapmap project. Nature *426*, 789.

[8] Lin, S., Cutler, D. J., Zwick, M. E., and Chakravarti, A. (2002). Haplotype inference in random population samples. The American Journal of Human Genetics *71*, 1129–1137.

[9] Yang, Y., Zhang, J., Hoh, J., Matsuda, F., Xu, P., Lathrop, M., and Ott, J. (2003). Efficiency of single-nucleotide polymorphism haplotype estimation from pooled dna. Proceedings of the National Academy of Sciences *100*, 7225–7230.

[10] Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J. *et al.* (2002). Detecting recent positive selection in the human genome from haplotype structure. Nature *419*, 832–837.