BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 4 2015, pages 515–522
doi:10.1093/bioinformatics/btu670

Sequence analysis

Advance Access publication October 9, 2014

Accurate estimation of haplotype frequency from pooled sequencing data and cost-effective identification of rare haplotype carriers by overlapping pool sequencing

Chang-Chang Cao and Xiao Sun*

State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China

Associate Editor: John Hancock

ABSTRACT

Motivation: A variety of hypotheses have been proposed for finding the missing heritability of complex diseases in genome-wide association studies. Studies have focused on the value of haplotype to improve the power of detecting associations with disease. To facilitate haplotype-based association analysis, it is necessary to accurately estimate haplotype frequencies of pooled samples.

Results: Taking advantage of databases that contain prior haplotypes, we present *Ehapp* based on the algorithm for solving the system of linear equations to estimate the frequencies of haplotypes from pooled sequencing data. Effects of various factors in sequencing on the performance are evaluated using simulated data. Our method could estimate the frequencies of haplotypes with only about 3% average relative difference for pooled sequencing of the mixture of 10 haplotypes with total coverage of 50×. When unknown haplotypes exist, our method maintains excellent performance for haplotypes with actual frequencies >0.05. Comparisons with present method on simulated data in conjunction with publicly available Illumina sequencing data indicate that our method is state of the art for many sequencing study designs. We also demonstrate the feasibility of applying overlapping pool sequencing to identify rare haplotype carriers cost-effectively.

Availability and implementation: Ehapp (in Perl) for the Linux platforms is available online (http://bioinfo.seu.edu.cn/Ehapp/).

Contact: xsun@seu.edu.cn

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

Received on June 22, 2014; revised on September 24, 2014; accepted on October 7, 2014

1 INTRODUCTION

Genome-wide association studies have identified numerous genetic variants associated with complex diseases and traits, and have provided valuable insights into their genetic architecture. However, most variants identified so far confer relatively small increments in risk, and explain only a small proportion of heritability (Manolio *et al.*, 2009). A variety of hypotheses have been proposed to explain the missing heritability, including rare variants, gene–gene and gene–environment interaction (Stranger *et al.*, 2011). In addition, studies have focused on the values of haplotypes and genome-wide haplotype association (GWHA) studies have been presented to find the association between haplotypes and diseases (Trégouët *et al.*, 2009). The combination of genetic marker alleles such as single nucleotide polymorphisms (SNPs) on a single chromosome is called a haplotype, which provides valuable information on evolutionary history and may lead to the development of more efficient strategies to identify genetic variants that increase susceptibility to diseases (Niu, 2004). Rather than examining variants independent of each other, simultaneously considering the values of multiple variants within haplotypes can improve the power of detecting associations with disease (Iliadis *et al.*, 2012). Up to now, numerous haplotypes have been found associated with several diseases (Chang *et al.*, 2010; Jin *et al.*, 2009; Lambert *et al.*, 2013; Martin *et al.*, 2010).

To facilitate GWHA studies, it is necessary to estimate haplotype frequencies accurately from deoxyribonucleic acid (DNA) pools of case and control samples. Long *et al.* (2011) presented *PoolHap* to estimate haplotype frequencies from pooled samples by next-generation sequencing. On the basis of an expectation–maximization algorithm, Kessner *et al.* (2013) put forward *Harp* to calculate the frequencies of haplotypes from pooled sequencing data. However, both *PoolHap* and *Harp* required that the haplotypes for pooled samples were known for inferring the frequency which may conflict with real situation. Taking advantage of haplotype database information has been proved feasible and helpful for haplotype frequency estimation (Gasbarra *et al.*, 2011; Pirinen, 2009). Under the plausible assumption that few haplotypes actually possess the vast majority of proportion or haplotype database is large enough, compressed sensing (CS) (Candes *et al.*, 2006; Donoho, 2006) could be used for the rapid and accurate reconstruction of the haplotype composition of pooled samples.

Koslicki *et al.* (2013) utilized techniques from CS to reconstruct the composition of a mixture of bacteria by utilizing databases that contain all the possible species. Jajamovich *et al.* (2013) translated the haplotype frequencies estimation as a joint constrained sparse optimization problem which has been studied in the CS literature. With the development of sequencing technology, a large number of haplotypes are rapidly gathered into public databases such as HapMap (Frazer *et al.*, 2007), DGRP (The *Drosophila* Genetic Reference Panel, Mackay *et al.*, 2012), and 1001 genomes (Weigel and Mott, 2009). The number of samples used in HapMap allows the project to find

*To whom correspondence should be addressed.

99% of haplotypes with frequencies of 5% or greater in a population (The International HapMap Consortium, 2005). Since CS was proposed to recover sparse signals from incomplete and inaccurate measurements (Candes *et al.*, 2006; Donoho, 2006), suppose the haplotypes for the samples in pooled sequencing are included in the database which is large enough, it is viable to infer the frequencies of haplotypes contained in the prior database by utilizing techniques from CS. However, for low-frequency haplotypes not contained in the database, it is hard to recover the frequency.

Once a haplotype is confirmed that it has association with diseases, similar to genomic variants, screening for the haplotype carriers is of great value in practical application. Borrowing ideas from group testing theory (Ding-Zhu and Hwang, 2000), overlapping pool sequencing was presented for the purpose of identifying rare variant carriers cost-effectively (Erlich *et al.*, 2009; Prabhu and Pe'er, 2009; Shental *et al.*, 2010). On the basis that the frequencies of haplotypes could be estimated precisely from pooled sequencing data, clearly, overlapping pool sequencing can also be applied in the identification of rare haplotype carriers.

Here, by means of prior haplotype information contained in the database, we present a method to estimate the haplotype frequencies from pooled sequencing data. The proposed method uses as inputs a database of haplotypes and pooled sequencing results, and returns estimated frequency for each haplotype contained in the database. A mathematical framework is first presented to translate haplotype frequency estimation as solving a system of linear equations. *NNREG* algorithm (Foucart and Koslicki, 2014) is employed to calculate the frequencies of haplotypes. Effects of various factors in sequencing on the performance of our method were evaluated using simulated data. The effects of unknown haplotypes contained in the pooled samples were also demonstrated. Comparison with *Harp* on both simulated data and publicly available Illumina sequencing data showed that our method may be more preferable for current massive parallel sequencing. Finally, on the basis of accurate estimation of haplotype frequency from pooled sequencing data, we revealed that it is feasible to identify rare haplotype carriers cost-effectively by applying overlapping pool sequencing. We have implemented the method in an open-source software tool Ehapp (**e**stimate haplotype frequency from **p**ooled sequencing data, http://bioinfo.seu.edu.cn/Ehapp/).

2 METHODS

2.1 Encode database information and pooled sequencing results

For haplotype consisting of SNPs, we first design a binary matrix to encode the database information. A binary vector with length 4 is defined to denote each kind of base (Fig. 1a). A haplotype with L SNPs in the database can be converted into a vector with length $4L$. For instance, the haplotype 'TGCA' is denoted as $(0,1,0,0,0,0,1,0,0,0,0,1,1,0,0,0)^{T}$. Consequently, a database containing N haplotypes with L SNP loci can be translated into a $4L \times N$ binary matrix M, in which columns are indexed by haplotypes and every four rows are indexed by an SNP. For the pooled sequencing experiment, the results are transformed as a numeric vector Y with length $4L$, which consists of the proportions of A, T, C and G for each SNP in the sequencing data (Fig. 1b).

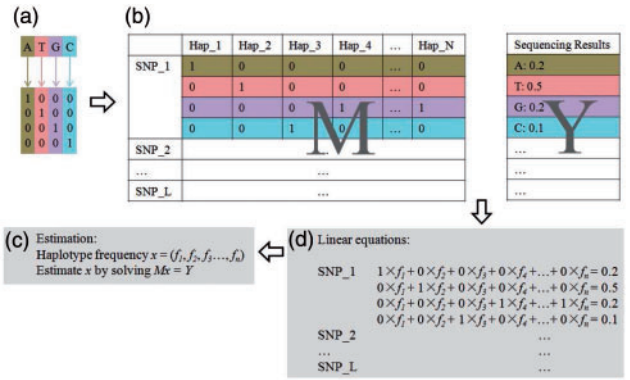

Fig. 1. Framework of our method. (**a**) Encode bases as binary vectors. (**b**) Encode the haplotype database and pooled sequencing results. (**c**) Four linear equations for the proportions of bases A, T, C and G can be constructed for each SNP. (**d**) Estimate haplotype frequency x by solving $Mx = Y$

Let $x = (f_1, f_2, \ldots, f_N)$ be the frequencies of all possible haplotypes, which are listed in the database, for the pooled samples. For each SNP, four equations for the proportions of bases A, T, C and G can be constructed (Fig. 1c). As a consequence, the problem to estimate frequency for each haplotype can be translated as solving Equation (1) (Fig. 1d).

$$Mx = Y \tag{1}$$

The optimal solution of (1) can be considered as the inferred haplotype frequencies. Before solving Equation (1), pre-processing could be carried out to simplify M and accelerate the decoding. Rows in M that consist of zeros, which result from the situation that less than four alleles exist for an SNP site, should be deleted. Furthermore, since any one row can be inferred from the other three rows easily for each SNP, i.e. the information is redundant, one row can be deleted for each SNP. In our method, rows with most '1' are deleted for each SNP.

2.2 Decoding

Under the assumption that few haplotypes actually possess the vast majority of proportion or haplotype database is large enough which indicates that x is a sparse vector where few elements are non-zero, Equation (1) could be solved by using techniques from CS theory. Once M and Y are given, CS aims to find the sparsest possible x. This can be written as the following optimization problem:

$$x* = \arg\min \|x\|_{l0} \ \ s.t.\ Mx = Y \tag{2}$$

Problem (2) is a non-deterministic polynomial complete problem. Under certain conditions, one can relax the constraint to l_1 norm and still get a solution that is identical to the solution of Problem (2) (Candes and Tao, 2005). Hence, Problem (2) is reformulated as (3) which can be solved by convex optimization techniques efficiently.

$$x* = \arg\min \|x\|_{l1} \ \ s.t.\ Mx = Y \tag{3}$$

Under the situation that measurements are corrupted by noise, CS aims to recover x by solving the optimization problem as follows:

$$x* = \arg\min \|x\|_{l1} \ \ s.t.\ \|Mx - Y\|_{l2} \le \varepsilon \tag{4}$$

where $\varepsilon > 0$ is set to be the maximal level of noise we are able to tolerate.

Furthermore, elements in x denote haplotype frequencies that are constrained to be non-negative; hence, we employ *NNREG* algorithm to recover x, which achieves sparse recovery by means of a conventional non-negative least squares algorithm (Foucart and Koslicki, 2014). *NNREG* recovers x by solving the optimization Problem (5) for some large $\lambda > 0$.

$$x* = \arg\min \ \|x\|_{l1}^2 + \lambda^2 \|Mx - Y\|_{l2}^2 \ \ s.t.\ x \geq 0 \quad (5)$$

NNREG allows for sparsity promotion and constrains x to be non-negative at the same time (Foucart and Koslicki, 2014). In general, *NNREG* reports a vector $x* \in R_0^{+\,4L}$. For estimating haplotype frequencies, a post-processing procedure should be performed to normalize x^* to make the sum equal to 1.

The relative difference (R_{diff}) between true frequency and estimated frequency is used to evaluate the accuracy.

$$R_{diff} = |x_i - x_i^*|/x_i \quad (6)$$

2.3 Overlapping pool sequencing

Borrowing ideas from group testing theory, overlapping pool sequencing was presented for the cost-effective identification of rare variant carriers among large-scale samples. Since a large number of samples are mixed into pools followed by a single sample preparation for each pool, overlapping pool sequencing could drastically reduce the cost for sample preparation as well as the cost for screening rare variant carriers.

The foundation for overlapping pool sequencing is the design for pooling samples. A design for n samples and t pools is associated with a $t \times n$ binary matrix $P = \{m_{ij}\}$, in which the rows are indexed by pools, the columns are indexed by samples and $m_{ij} = 1$ if and only if the jth sample is contained in the ith pool. For a simple instance, to find a variant carrier among seven individuals, samples could be pooled according to the pooling matrix below:

$$P = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

This pooling design matrix consists of three pools: {4, 5, 6, 7}, {2, 3, 6, 7} and {1, 3, 5, 7}. Suppose the variant was found only in the second and third pools, then we can infer that the third sample is a variant carrier.

To apply overlapping pool sequencing for the identification of rare haplotype carriers, pools containing rare haplotypes should be distinguished correctly from those that are not. Based on the estimation of haplotype frequencies from pooled sequencing data, a pool is suggested to contain rare haplotypes if the estimated frequency is higher than a threshold, which can be set as 0.5 divided by the pool size that denotes the lowest possible frequency of haplotypes in the mixture of diploid samples.

Currently, many pooling designs have been proposed for identifying rare variant carriers. Benefiting from the Chinese remainder theorem, Erlich *et al.* (2009) presented the DNA Sudoku design and a pattern consistency decoding algorithm to find the carriers. Thierry-Mieg (2006) put forward shifted transversal design (STD) and provided a freely available tool called *interpool* (Thierry-Mieg and Bailly, 2008) which can be utilized to choose candidate STD designs and identify variant carriers. The optimal design should be chosen depending on the cost of the whole experiment (Cao *et al.*, 2013).

As long as the haplotype frequencies can be estimated precisely, indicating that pools containing rare haplotypes can be recognized correctly, it is feasible to apply overlapping pool sequencing in screening for rare haplotype carriers cost-effectively.

2.4 Haplotype database and sequencing data

Two haplotype databases were used in our simulation experiment. The first one consists of 158 haplotypes of *Drosophila* from DGRP (Mackay *et al.*, 2012) and the second is made up of 50 haplotypes of *Arabidopsis thaliana* from 1001 Genomes Data Center (Weigel and Mott, 2009).

To evaluate the performance of our method, both simulated sequencing data and publicly available Illumina sequencing data are used. The simulated sequencing data were generated by using a program called *simreads* which was used in *Harp* and provided by Kessner *et al.* (2013). Real Illumina sequencing data for *A.thaliana* were downloaded from the GenBank Short Read Archive with accession number SRP012869 (Long *et al.*, 2013). Because of the limited computing resource, sequencing data of only five strains were downloaded for simulating pooled sequencing data *in silico*.

3 RESULTS

3.1 Effect of various factors in sequencing

We first evaluated the effect of various factors in sequencing on the performance of our method. In the simulation, we generated pooled sequencing data by using *simreads* which produced mapped reads (SAM format) along with sequencing qualities. The frequency for each haplotype was drawn from a symmetric Dirichlet distribution. The symmetric Dirichlet distribution is parameterized with a single parameter α that governs the uniformity of the randomly drawn frequency distributions. Following the study conducted by Kessner *et al.* (2013), α was set as 0.2 to produce haplotype frequency distributions similar to reality.

One hundred and fifty-eight haplotypes of *Drosophila* were downloaded from DGRP. SNPs that were identical among all the haplotypes were filtered first. 747 660 SNPs on chromosome X passed the filter procedure and were used in the following analysis. The SNP density was about 1/30 SNP/bp.

Since the choice of λ in *NNREG* is very critical for the accuracy of estimation (Foucart and Koslicki, 2014), we first evaluated the effect of λ on the performance of our method under different situations where non-varying parameters were held as default values (coverage $50\times$, read error rate 0.01, 10 haplotypes, read length 100 bp, region width 2 Mb). To guarantee the accuracy, each pair of haplotypes that have different alleles at less than 50 non-degenerate SNP sites (defined as those containing no degenerate bases) were not used to simulate pooled sequencing. Perl scripts were used to encode the haplotype database information and sequencing results, following by utilizing *NNREG* to calculate the frequency of each haplotype in the database. Later, relative differences (R_{diff}) between true frequency and estimated frequency were calculated for 10 replicates of each scenario. The results indicated that the optimal λ that resulted in lowest R_{diff} was closely related to coverage, number of haplotypes, and region width (Supplementary Fig. S1). Substantial simulations should be conducted for choosing the optimal λ.

Using the optimal value of λ for each scenario (Supplementary Table S1), we next evaluated the effect of sequencing coverage, read error rate, read length, sequencing region width and number of haplotypes on the performance of our method. Non-varying parameters were held as default values. The results showed that the performance of our method decreased with high sequencing error and low coverage, as both imported noises in the measurements of four base frequencies for each SNP (Fig. 2). And our method was not sensitive to read length, but became more robust for less haplotypes and longer sequencing region.

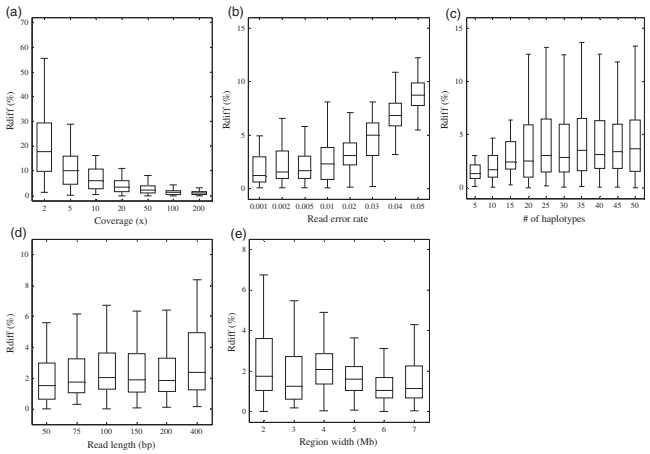

Fig. 2. The effects of various factors on the performance of our method. The algorithm performed better with lower read error rate (**a**) and higher sequencing coverage (**b**). Our method is not sensitive to read length (**c**), but more robust for longer region and less haplotypes (**d** and **e**). Ten replicates were conducted for each scenario and non-varying parameters were set as default values. Haplotype with true frequency lower than 0.02 is neglected in the statistics of R_{diff}

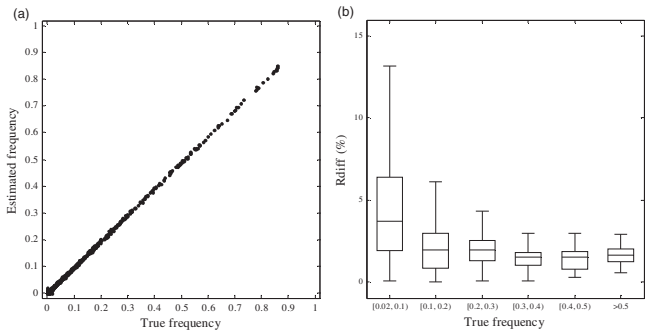

Fig. 3. Estimation accuracy of our method. (**a**) The correlation between true frequency and estimated frequency. (**b**) The relative difference for haplotypes with different frequencies

Next, we conducted 100 replicates for simulation experiment with all parameters set as default values and computed the relative differences between true and estimated frequencies (Fig. 3). The results showed that our method could estimate the haplotype frequency with only about 3% average relative difference for pooled sequencing of the mixture of 10 haplotypes with total coverage of 50×.

3.2 Effect of unknown haplotypes

Although the haplotype database increases rapidly with the development of sequencing technology, the available database information may be incomplete to provide a good coverage of the haplotypes of the population. Haplotypes not contained in the database (defined as unknown haplotypes) could introduce variations between Mx and Y, as the columns in M are indexed by known haplotypes in the database but reads from unknown haplotypes are taken into the calculation of Y. Therefore, we next evaluate the influence of unknown haplotypes on the estimation of frequencies for known haplotypes.

We suppose that 40 of the 158 haplotypes of *Drosophila* from DGRP are unknown and delete them from the database. For database consisting of 118 haplotypes, we first did simulation to find the optimal λ value when all the parameters were held as default values (Supplementary Fig. S2). The results showed that λ equaling 10^5 performed best. Next, we used all the 158 haplotypes to produce pooled sequencing reads to simulate a mixture of known and unknown haplotypes. Only these 118 known haplotypes were employed to infer the haplotype frequencies. Two hundred replicates were conducted to evaluate the impact of unknown haplotypes where 20 random selected haplotypes were mixed for each replicate, the estimated versus true frequency was drawn in Figure 4a. From the results, we can infer that our method is able to estimate the frequencies of known haplotypes and maintain excellent performance for haplotypes with actual frequencies higher than 0.05 when unknown haplotypes exist. The performance decreased rapidly as the proportion of unknown haplotypes increased which was inconsistent with expectation (Fig. 4b), since the unknown haplotypes imported variations between Mx and Y.

We counted the differences between Mx and Y when unknown haplotypes exist (Supplementary Fig. S3 and Supplementary Table S2). The results showed that the unknown haplotypes only introduced variations in very few rows. In theory, each unknown haplotype should import variations in one row (A/T/C/G) for each SNP. However, after the simplification of M, the number of rows that could be affected by unknown haplotypes decreased a lot. Because rows with the most 1 are deleted for each SNP, meaning that the allele with the highest frequency for an SNP is deleted in M, the differences between Mx and Y in these rows that are caused by unknown haplotypes will have no impacts on the accuracy of estimation for the frequency of known haplotypes. Accordingly, our method could maintain good performance when unknown haplotypes exist.

3.3 Comparisons with current methods

PoolHap and *Harp* were presented to infer haplotype frequencies from pooled samples by next-generation sequencing. Both required that the haplotypes for the pooled samples are known. Unfortunately, the current version of *PoolHap* relies on a Java library which is no longer publicly available, and could not run correctly on our workstation. Therefore, we compared only the performance of *Harp* with that of our method in inferring haplotype frequencies from pooled sequencing data.

We first compared the performance by using the 158 haplotypes of *Drosophila* and simulated sequencing data. The read length for simulating pooled sequencing data ranged from 50 to 150 bp and the error rate ranged from 0.01 to 0.05. The other parameters were all set as default values (coverage 50×, region width 2 Mb, 10 haplotypes). The SNP density for the selected region (first 2 Mb on chromosome X) was about 1/50 SNP/bp. Both our method and *Harp* were applied to calculate the frequency of haplotypes. Since *Harp* required that the haplotype for pooled samples were known, we ran *Harp* twice—all the haplotypes in the database (*Harp all*) or haplotypes of only pooled samples (*Harp pooled*) are taken as possible haplotypes

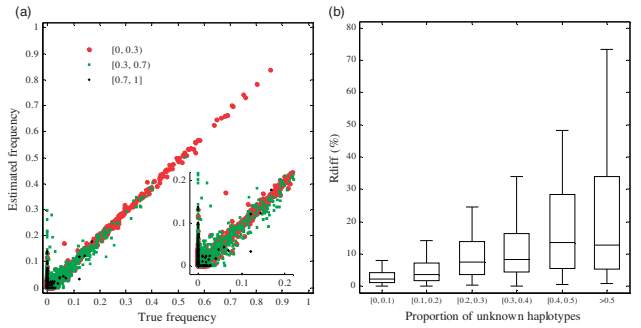

Fig. 4. The performance of our method when different proportions of unknown haplotypes exist. (**a**) The correlation between true and estimated frequency. Different colors denote different proportions of unknown haplotypes. Our method could maintain excellent performance for haplotypes with actual frequencies higher than 0.05. (**b**) The relative difference increased rapidly with the proportion of unknown haplotypes. Haplotype with true frequency lower than 0.05 is neglected in the statistics of R_{diff}

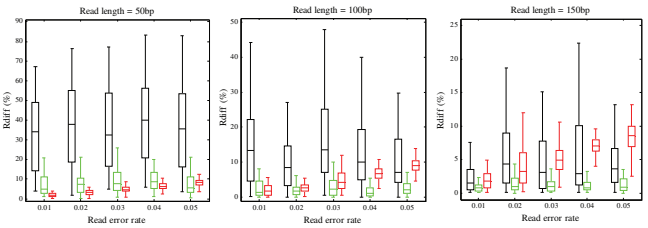

Fig. 5. Comparison of our method with *Harp* based on simulated sequencing data and haplotypes for *Drosophila*. Black box denotes the scenario that *Harp* is applied and all the haplotypes in the database are used to infer frequency (*Harp all*). Green box denotes the scenario that haplotypes for only pooled samples are used (*Harp pooled*). Red box denotes the performance of our method. Haplotype with true frequency lower than 0.02 is neglected in the statistics of R_{diff}

for estimating the frequency (Fig. 5; scatter plots for error rate 0.01 were shown in Supplementary Fig. S4).

When all the haplotypes in the database were used to infer the frequency, our method had better performance than *Harp* for short reads and low error rate. Since only the proportions of four bases for each SNP were used to recover haplotype frequencies, our method was not sensitive to reads length. However, *Harp* was better for long reads as haplotype information contained in single reads was employed to estimate haplotype frequency and long reads covered multiple SNPs which were sufficient for the estimation (Kessner *et al.*, 2013). Besides, *Harp* was more robust for high error rate due to the advantage that base qualities were taken into account in *Harp*, especially when haplotypes of exact pooled samples were known and taken as possible haplotypes for estimating the frequency, as *Harp* searched the optimal solution in a much smaller space. These results also indicated that which method performs better depends both on read error rate and read length with respect to SNP density. And our method may be more preferable for current massive parallel sequencing which has short reads and possesses error rate at the level of 0.01 (Shendure and Ji, 2008).

We also compared the performance of our method with that of *Harp* by using SNPs located in other two regions (5–7 Mb and 10–12 Mb) on chromosome X. The SNP density was about 1/27 (5–7 Mb) and 1/31 SNP/bp (10–12 Mb), respectively. Both verified that our method has better performance for short reads with respect to SNP density and *Harp* is preferable for long reads (Supplementary Fig. S5).

When unknown haplotypes exist, *Harp* filters out reads whose maximum haplotype likelihood falls outside a specified range to accurately estimate the frequency of known haplotypes (Kessner *et al.*, 2013). Hence, we also compared the performance of our method with *Harp* when unknown haplotypes exist. Forty of 158 haplotypes of *Drosophila* from DGRP were still supposed to be unknown and all the 158 haplotypes were utilized to produce pooled sequencing reads to simulate a mixture of known and unknown haplotypes. After using 118 known haplotypes to infer the frequency, results also showed that our method is better for short reads and *Harp* is superior for long reads (Supplementary Fig. S6). When known haplotypes for only pooled samples were employed, *Harp* yielded results with the sum of the inferred frequencies for these known haplotypes equal to 1 which was inconsistent with reality since unknown haplotypes existed, and this situation resulted in high R_{diff}.

3.4 Simulation with Illumina sequencing data

We investigated the performance of our method and *Harp* on real Illumina sequencing data by using haplotypes for 50 strains of *A.thaliana*. Sequencing data for five strains were downloaded from GenBank Short Read Archive with accession number SRP012869 and mixed *in silico* to simulate pooled sequencing. Details about the haplotypes and sequencing data are contained in Supplementary Tables S3 and S4. Quality control was first done to obtain haplotypes with high qualities. First, SNPs with intermediate coverage were selected by filtering SNPs with coverage lower than 0.5 times or higher than 1.5 times the average sequencing depth (Long *et al.*, 2011). Because these *Arabidopsis* strains are inbred lines, most of the SNPs are homozygous. Therefore, SNPs with major base proportion lower than 0.8 were also filtered. Finally, 131 110 SNPs on chromosome 1 passed the filter procedure and were used in the following analysis and the SNP density is about 1/232 SNP/bp.

Haplotype frequencies for five strains were drawn from a symmetric Dirichlet distribution, followed by taking reads randomly from the data set and mixing *in silico*. The coverage for pooled sequencing was set as 25×. Bowtie 0.12.9 (Langmead *et al.*, 2009) was used to map pooled reads back to *A.thaliana* genome. On the basis of the mapping results (SAM format), both our method and *Harp* were utilized to infer haplotype frequencies. Experiments based on simulated reads showed that the optimal λ is 10^3 for the database consisting of 50 haplotypes of *A.thaliana* (Supplementary Fig. S7).

After estimating haplotype frequency for 100 replicates, the results showed that our method is superior to *Harp* for this scenario no matter whether all the 50 haplotypes or these 5 haplotypes for pooled strains were employed in *Harp* to compute the frequency (Fig. 6a and Supplementary Fig. S8a). This situation may partly result from the low SNP density with respect to

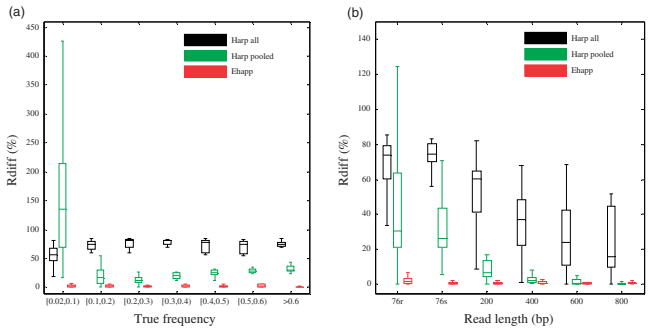

Fig. 6. Comparison of our method with *Harp* based on the haplotypes for *A.thaliana*. (**a**) Comparison on real Illumina sequencing data. (**b**) Comparison on simulated sequencing data with read length ranging from 76 to 800 bp. '76r' denotes the results for real Illumina sequencing reads with length 76 bp and '76s' denotes the results for simulated sequencing reads with length 76 bp. Outliners for box plots are not shown. Haplotype with true frequency lower than 0.02 is neglected in the statistics of R_{diff}

reads length (76 bp). Besides, inconformity between the sequencing quality for real sequencing data and quality model used in *Harp* could also introduce errors.

To find the key factor that is responsible for the performance of *Harp* on real Illumina sequencing reads, we investigated the performance on simulated sequencing data where the sequencing coverage and read error rate were set as 25× and 0.01, respectively (Fig. 6b and Supplementary Fig. S8b–f). *Harp* showed similar performance for real and simulated sequencing data when the read lengths for both were 76 bp, which revealed that the mismatch between the true errors and the *Harp* error model had little effects. Accordingly, we can infer that *Harp* error model is consistent with real sequencing data. However, the accuracy of *Harp* increased rapidly with the reads length which could prove that the short read length for real sequencing data should be responsible for the performance of *Harp*. These also verified that our method is more preferable for current massive parallel sequencing which has short reads.

We also conducted two experiments by using haplotypes for chromosomes 2 and 3; the SNP density was about 1/215 and 1/210 SNP/bp, respectively. Both showed similar results to previous simulations (Supplementary Fig. S9). When only five haplotypes for the pooled samples were used in *Harp*, these results verified the underestimation for haplotypes with high frequency (>0.2) and overestimation for haplotypes with low frequency (<0.2). This situation may result from the low SNP density with respect to short reads where substantial reads containing no SNPs will be fractionally assigned to all the haplotypes, leading to more uniform distribution of haplotype frequency, i.e. underestimation for high frequency and overestimation for low frequency. This could also explain the performance of *Harp* when all the haplotypes in the database were used.

Furthermore, there were lots of points for *Harp* where the true frequency was 0 but the estimated frequency was non-zero (Supplementary Figs S8 and S9). When haplotypes for chromosome 1 were used, the vast majority of these points represented the estimated frequency for haplotype *Had-2*. After counting the different alleles between each pair of haplotypes, we found that haplotypes *Had-1* and *Had-2* were very similar which had different alleles at only 264 SNPs on chromosome 1 (131 110 SNPs in total). This pair of nearly identical haplotypes could hardly be separated by *Harp*. However, this was not an issue in our method.

3.5 Overlapping pool sequencing to identify rare haplotype carriers

Overlapping pool sequencing was presented for the cost-effective identification of rare variant carriers. Since numerous rare haplotypes have been identified to have strong correlations with several diseases (Chang *et al.*, 2010; Jin *et al.*, 2009; Lambert *et al.*, 2013; Martin *et al.*, 2010), screening large-scale samples to find rare haplotype carriers becomes increasingly important. On the basis that the frequencies of haplotypes could be estimated precisely from pooled sequencing data, overlapping pool sequencing can be applied to identify rare haplotype carriers.

We conducted a simulation experiment to identify 2 heterozygous carriers for an assigned haplotype (*Aedal-1*) among 100 simulated diploid individuals (Table 1). Pool size was constrained to be smaller than 10 individual samples to make sure the lowest frequency for a haplotype is higher than 0.05 (0.5/10) to guarantee accurate estimation of frequency. STD and DNA Sudoku required 33 and 34 pools to find 2 specified haplotype carriers among 100 samples, respectively (pooling matrixes are detailed in Supplementary Table S5). Pooled sequencing was simulated *in silico* by taking reads randomly from data sets and mixing them where the coverage was set as 25×. Equal amount of reads from two strains were mixed together to simulate heterozygous haplotype carriers. With the same procedure described previously to infer haplotype frequency for pooled sequencing of *A.thaliana*, the proportion of the rare haplotype was then extracted (Fig. 7 and Supplementary Fig. S10). Based on the inferred frequencies of haplotypes, pools containing rare haplotypes can be recognized clearly, and the rare haplotype carriers could be identified precisely by using corresponding decoding algorithms for STD and DNA Sudoku designs.

Next, we calculated the least pools required for identifying haplotype carriers with various frequencies successfully when applying group testing algorithms STD and DNA Sudoku to design overlapping pool sequencing (Supplementary Table S6). Pool size was also constrained to be smaller than 10 individual samples to guarantee accuracy in frequency estimation. Ignoring the cost for sequencing data production, overlapping pool sequencing showed great potentiality in reducing the cost for screening rare haplotype carriers.

4 DISCUSSION

Taking advantage of prior haplotype information contained in the database, we present Ehapp for inferring haplotype frequencies from pooled sequencing data. *NNREG* algorithm is employed to calculate the frequency, which achieves sparse recovery by means of a conventional non-negative least squares algorithm. And the optimal λ for *NNREG* could be obtained by conducting substantial simulations.

Table 1. Genotypes for 100 diploid individuals with each having 2 haplotypes from parents

Genotypes	No. of samples
Aedal-1|*Had-1*	2 (40th, 60th)
Ale-Stenar-44-4|*Ale-Stenar-44-4*	11
App1-12|*App1-12*	20
Eden-2|*Eden-2*	29
Had-1|*Had-1*	38
Total	100

Note: Two haplotypes are identical for homozygous individuals. *Aedal-1* is assigned as the rare haplotype. The 40th and 60th samples are simulated as heterozygous carriers for haplotype *Aedal-1*.

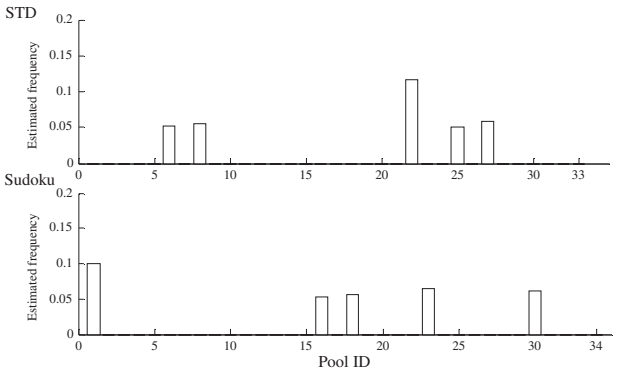

Fig. 7. The estimated frequency for haplotype *Aedal-1* in each pool. (**a**) Thirty-three pools in STD design. (**b**) Thirty-four pools in DNA Sudoku design

Results of experiments on simulated sequencing data showed that higher depth of coverage and lower error rate could improve the performance of our method. With 0.01 sequencing error rate, our method could estimate the frequencies of haplotypes with only about 3% average relative errors for pooled sequencing of the mixture of 10 haplotypes with total coverage of 50×. Thanks to the simplification for *M*; our method could maintain excellent performance for haplotypes with actual frequencies higher than 0.05 when unknown haplotypes exist. Comparisons with *Harp* on simulated data in conjunction with real sequencing data showed that our method is more outstanding for short reads with respect to SNP density and preferable for current massive parallel sequencing. Analysis for *Arabidopsis* data also revealed that our method performed better in distinguishing similar haplotypes. Since only the proportions of A, T, C and G for each SNP are used to recover haplotype frequencies, our method is much faster and requires fewer memories compared with *Harp* which takes all the reads into calculation. At last, we proved that overlapping pool sequencing could be applied in screening for rare haplotype carriers as long as haplotype frequencies can be inferred accurately from pooled sequencing data.

The number of samples used in HapMap allows the project to find 99% of haplotypes with frequencies of 5% or greater in a population (The International HapMap Consortium, 2005). However, haplotypes with low frequencies (<5%) can also be observed for many populations (Chattopadhyay *et al.*, 2003). Although substantial errors happened for estimating the frequency of haplotype with proportion lower than 0.05 when unknown haplotype exists, this drawback could be addressed by completing the database to include haplotypes as much as possible. Alternatively, the number of pooled individuals could be constrained to be smaller than a threshold to make the lowest haplotype frequency higher than 0.05; therefore, the frequency for known haplotype could still be inferred accurately even when unknown haplotypes exist.

One major application of our method is in case–control studies for finding diseases associated with haplotypes. After estimating frequencies for case and control samples independently, haplotype with frequencies that differ statistically significantly could be ascertained as disease-associated haplotypes. If the disease-associated haplotype is rare, overlapping pool sequencing could be applied to identify the carriers among large-scale samples precisely with much lower cost. We hope that our methods could be applied in the GWHA and screening for rare haplotype carriers. Besides, our method could also be applicable in non-pool situations, such as inferring the frequency of variants of duplicated regions (e.g. transposable elements) (Fiston-Lavier *et al.*, 2011).

The software implementing the method, *Ehapp*, is open source and available for download and can be easily integrated into existing analysis pipelines.

ACKNOWLEDGEMENT

The authors thank Darren Kessner for providing *simreads* to generate simulated sequencing reads.

*Fundi*ng: This work was supported by the National Basic Research Program of China [grant number 2012CB316501] and the National Natural Science Foundation of China [grant numbers 61472078, 61073141].

Conflict of Interest: none declared.

REFERENCES

Candes,E.J. *et al.* (2006) Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pure Appl. Math.*, **59**, 1207–1223.

Candes,E.J. and Tao,T. (2005) Decoding by linear programming. *IEEE Trans. Inf. Theory*, **51**, 4203–4215.

Cao,C.C. *et al.* (2013) Identifying rare variants with optimal depth of coverage and cost-effective overlapping pool sequencing. *Genet. Epidemiol.*, **37**, 820–830.

Chang,Y.C. *et al.* (2010) The associations of LPIN1 gene expression in adipose tissue with metabolic phenotypes in the Chinese population. *Obesity*, **18**, 7–12.

Chattopadhyay,P. *et al.* (2003) Global survey of haplotype frequencies and linkage disequilibrium at the RET locus. *Eur. J. Hum. Genet.*, **11**, 760–769.

Ding-Zhu,D. and Hwang,F.K. (2000) *Combinatorial Group Testing and Its Applications (Series on Applied Mathematics)*. Vol. 12. World Scientific Publishing Co. Inc, Singapore.

Donoho,D.L. (2006) Compressed sensing. *IEEE Trans. Inf. Theory*, **52**, 1289–1306.

Erlich,Y. *et al.* (2009) DNA Sudoku—harnessing high-throughput sequencing for multiplexed specimen analysis. *Genome Res.*, **19**, 1243–1253.

Fiston-Lavier,A.S. *et al.* (2011) T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data. *Nucleic Acids Res.*, **39**, e36.

Foucart,S. and Koslicki,D. (2014) Sparse recovery by means of nonnegative least squares. *IEEE Signal Process. Lett.*, **21**, 498–502.

Frazer,K.A. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.

Gasbarra,D. *et al.* (2011) Estimating haplotype frequencies by combining data from large DNA pools with database information. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **8**, 36–44.
Iliadis,A. *et al.* (2012) Fast and accurate haplotype frequency estimation for large haplotype vectors from pooled DNA data. *BMC Genet.*, **13**, 94.
Jajamovich,G.H. *et al.* (2013) Maximum-parsimony haplotype frequencies inference based on a joint constrained sparse representation of pooled DNA. *BMC Bioinformatics*, **14**, 270.
Jin,H. *et al.* (2009) A rare haplotype in the upstream regulatory region of COL1A1 is associated with reduced bone quality and hip fracture. *J. Bone Miner. Res.*, **24**, 448–454.
Kessner,D. *et al.* (2013) Maximum likelihood estimation of frequencies of known haplotypes from pooled sequence data. *Mol. Biol. Evol.*, **30**, 1145–1158.
Koslicki,D. *et al.* (2013) Quikr: a method for rapid reconstruction of bacterial communities via compressive sensing. *Bioinformatics*, **29**, 2096–2102.
Lambert,J.C. *et al.* (2013) Genome-wide haplotype association study identifies the FRMD4A gene as a risk locus for Alzheimer's disease. *Mol. Psychiatry*, **18**, 461–470.
Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
Long,Q. *et al.* (2011) PoolHap: inferring haplotype frequencies from pooled samples by next generation sequencing. *PLoS One*, **6**, e15292.
Long,Q. *et al.* (2013) Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nat. Genet.*, **45**, 884–890.
Mackay,T.F. *et al.* (2012) The *Drosophila melanogaster* genetic reference panel. *Nature*, **482**, 173–178.
Manolio,T.A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
Martin,R.J. *et al.* (2010) A rare haplotype of the vitamin D receptor gene is protective against diabetic nephropathy. *Nephrol. Dial. Transplant.*, **25**, 497–503.
Niu,T. (2004) Algorithms for inferring haplotypes. *Genet. Epidemiol.*, **27**, 334–347.
Pirinen,M. (2009) Estimating population haplotype frequencies from pooled SNP data using incomplete database information. *Bioinformatics*, **25**, 3296–3302.
Prabhu,S. and Pe'er,I. (2009) Overlapping pools for high-throughput targeted resequencing. *Genome Res.*, **19**, 1254–1261.
Shendure,J. and Ji,H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**, 1135–1145.
Shental,N. *et al.* (2010) Identification of rare alleles and their carriers using compressed se(que)nsing. *Nucleic Acids Res.*, **38**, e179.
Stranger,B.E. *et al.* (2011) Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, **187**, 367–383.
The International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
Thierry-Mieg,N. (2006) A new pooling strategy for high-throughput screening: the Shifted Transversal Design. *BMC Bioinformatics*, **7**, 28.
Thierry-Mieg,N. and Bailly,G. (2008) Interpool: interpreting smart-pooling results. *Bioinformatics*, **24**, 696–703.
Trégouët,D.A. *et al.* (2009) Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. *Nat. Genet.*, **41**, 283–285.
Weigel,D. and Mott,R. (2009) The 1001 genomes project for *Arabidopsis thaliana*. *Genome Biol.*, **10**, 107.