HaploPool: Improving Haplotype Frequency Estimation through DNA Pools and Phylogenetic Modeling

Bonnie Kirkpatrick* Carlos Santos Armendariz† Richard M. Karp‡
Eran Halperin§

Abstract

In the context of disease association studies, haplotype frequencies are usually estimated from genotype data. In order to reduce genotyping costs, one can estimate the haplotype frequencies from DNA pools. Here we present a method, HaploPool, which estimates the haplotype frequencies in a population from a set of DNA pools of l individuals each, where l is a small number (typically two or three). HaploPool is based on a combination of a perfect phylogeny model, together with the EM algorithm on subsets of the SNPs. Using HaploPool, we study the trade-off between DNA pooling and the accuracy of haplotype frequency estimation. We show that the accuracy of frequency estimates obtained by HaploPool on a set of n DNA pools of two individuals each, is roughly the same as the accuracy of the frequency estimates reported by PHASE when applied to a set of $1.45n$ genotypes from the same population.

We compared our algorithm to three state of the art haplotype frequency estimation programs. HaploPool is consistently more efficient and more accurate than previous methods on pooled data, and its accuracy on approaches the accuracy of PHASE on genotype data.

Human genetic variation is key to understanding complex hereditary diseases. Most of this variation can be characterized by single nucleotide polymorphisms (SNPs), which are evidence of mutations that occurred once in history and then were passed on through heredity. Recent progress in technology for high-throughput SNP genotyping provides an opportunity to understand the genetic basis of complex disease through whole genome association studies. In these studies, hundreds of thousands of SNPs are genotyped for two sets of individuals (cases and controls), and discrepancies between the SNP-allele distributions serve as evidence for the association of a genetic region with the disease.

Unfortunately, the advance in high-throughput genotyping is a mixed blessing. Since more SNPs can be genotyped for a given cost, more individuals are needed for genotyping in order to overcome the loss of power due to multiple hypothesis testing. In particular, to find the etiology of complex disease, association studies need thousands of individuals. With today's genotyping costs, a well-powered whole genome association study may cost millions of dollars. Finding new ways of reducing the burden of genotyping is critical for larger and more rigorous association studies.

One step in this direction is the use of haplotypes. The rationale for this strategy is based on the fact that SNPs in close physical proximity to each other are often correlated (in *Linkage Disequilibrium*), and the variation of the *haplotype* (sequence of alleles in contiguous SNP sites along a chromosomal region) is known to be of limited diversity. Therefore, haplotypes give evidence for the presence of SNPs that have not been genotyped in the study [3]. In particular, the identification and analysis of haplotypes [2], is currently playing a key role in trait and disease associations studies [6].

Another natural strategy for the reduction of genotyping costs is the use of pooled DNA. Some technologies are able to determine the SNP-allele frequencies with high precision in pooled samples, thereby replacing many individual genotype measurements with one consolidated analysis. These technologies extract the DNA from a pool of individuals using approximately equal amounts of DNA from each individual.

*Computer Science Department, UC Berkeley
†International Computer Science Institute (ICSI), Berkeley, CA
‡International Computer Science Institute (ICSI), Berkeley, CA
§Corresponding author. International Computer Science Institute, 1947 Center St., Suite 600, Berkeley, CA, USA. Email: heran@icsi.berkeley.edu. Phone: +(510)-666-2952. Fax: (510)-666-2956.

This bulked DNA is then genotyped and the frequency of an allele in each position is given. Therefore, for pools of size k, the cost of genotyping is reduced by a factor of k.

DNA pools are not without caveats. First, DNA pools lose the information of the individual genotypes, and hence they lose the haplotype information. Second, the error rate for DNA pools is quite high, and in practice the exact determination of the allele frequency of a large DNA pool is infeasible. For this reason, DNA pools are currently used in association studies as a screening procedure. The cases and the controls are pooled separately. SNPs for which there is a large discrepancy between the pool allele frequencies in the cases and the controls are individually genotyped in a validation stage. Unfortunately, the screening procedure has the inherent problem of losing the haplotype information. Furthermore, inaccuracies in the allele frequency estimations from the DNA pools result in a large number of false positives carried along to the validation stage.

Here, we suggest a middle ground, where pools of a small number of individuals are used (two or three individuals per pool). We show that a careful analysis of small DNA pools reveals information about haplotypes. We leverage on the fact that the error rate for DNA pools of two individuals is comparable to the individual genotyping error rate [1]. We introduce a method, HaploPool, which estimates the haplotype frequencies from a set of n DNA pools of k samples each (a total of nk individuals). We compare these estimates to the haplotype frequencies estimated by the widely used phasing method PHASE [7] on a set of genotypes. We demonstrate that the accuracy of HaploPool's estimates from n genotyping experiments (i.e., n DNA pools), is similar to the accuracy achieved by PHASE when estimating the frequencies from $1.45n$ individual genotypes. Thus, we effectively save 45% of the budget to get the same accuracy of haplotype frequency estimates.

The use of DNA pools of small numbers of individuals has already been suggested in two previous works [4, 5]. In both cases, the expectation-maximization (EM) algorithm was suggested to infer haplotype frequencies from DNA pools of a small number of individuals. Our method is different from previous methods in a number of points. First, as opposed to previous methods, our algorithm incorporates a perfect phylogeny model, which helps to increase the accuracy and efficiency at the same time. Second, we make use of an EM algorithm for small subsets of the SNPs, thus getting partial solutions, which we combine into one global solution using linear regression. Finally, we compared HaploPool to previous methods and found that our method is more accurate, much more efficient, and capable of dealing well with missing data and genotyping errors.

References

[1] K.B. Beckman, K.J. Abel, A. Braun, and E. Halperin. Using dna pools for genotyping trios. *Nucleic Acid Research*, 2006. doi: 10.1093/nar/gkl700.

[2] The International HapMap Consortium. The international hapmap project. *Nature*, 426:789–796, 2003.

[3] P.I.W. de Bakker, R.R. Graham, D. Altshuler, B.E. Henderson, and C.A. Haiman. Transferability of tag snps to capture common genetic variation in dna repair genes across multiple populations. In *Pacific Symposium on Biocomputing*, 2006.

[4] J. Hoh, F. Matsuda, X. Peng, D. Markovic, M.G. Lathrop, and J. Ott. Snp haplotype tagging from dna pools of two individuals. *BMC Bioinformatics*, 4(14), 2003.

[5] T. Ito, S. Chiku, E. Inoue, M. Tomita, T. Morisaki, H. Morisaki, and N. Kamatani. Estimation of haplotype frequencies, linkage-disequilibrium measures, and combination of haplotype copies in each pool by use of pooled dna data. *American Journal of Human Genetics*, 72:384–398, 2003.

[6] R.W. Morris and N.L. Kaplan. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genetic Epidemiology*, 23:221–223, 2002.

[7] M. Stephens, N. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, 68:978–989, 2001.