

# HapBoost: A Fast Approach to Boosting Haplotype Association Analyses in Genome-Wide Association Studies

Xiang Wan, Can Yang, Qiang Yang,  
Hongyu Zhao, and Weichuan Yu

**Abstract**—Genome-wide association study (GWAS) has been successful in identifying genetic variants that are associated with complex human diseases. In GWAS, multilocus association analyses through linkage disequilibrium (LD), named haplotype-based analyses, may have greater power than single-locus analyses for detecting disease susceptibility loci. However, the large number of SNPs genotyped in GWAS poses great computational challenges in the detection of haplotype associations. We present a fast method named HapBoost for finding haplotype associations, which can be applied to quickly screen the whole genome. The effectiveness of HapBoost is demonstrated by using both synthetic and real data sets. The experimental results show that the proposed approach can achieve comparably accurate results while it performs much faster than existing methods.

**Index Terms**—SNP, haplotype, genome-wide association studies, linkage disequilibrium

## 1 INTRODUCTION

GENOME-WIDE association study (GWAS) has proven to be a valuable tool to unravel the etiology of complex diseases. Recent technological developments have enabled researchers to investigate the entire genome with relative low cost, allowing for the identification of many new and unexpected associations that relate regions of the genome to disease risk. Along with the development of GWAS, one major question emerges: How do we make the best usage of the GWAS data? To date, researchers have mostly used single-marker-based methods to analyze single nucleotide polymorphisms (SNPs) in GWAS, looking for statistical disease-related associations. Although some susceptibility SNPs have been identified, these SNPs can only explain a small portion of genetic contributions to complex diseases, which is known as the “missing heritability.” Multilocus association analyses through linkage disequilibrium (LD), named haplotype-based analyses, may be more informative for association analyses [1]. A haplotype is a combination of alleles at adjacent locations on a chromosome that are transmitted together. A haplotype includes several loci or an entire chromosome depending on the number of recombination events that have occurred for a given set of loci [2]. The reasons that haplotype association analysis is more informative are as follows:

- For genes that contain two or more functional mutations, haplotypes have potential impact on gene functions. Particular combinations of amino acids are responsible for protein folding, which is directly influenced by genetic sequence variations.
- Haplotype analyses can reveal risk factors that show little individual effects but jointly produce significant associations. Identifying these risk factors may provide more insights in understanding complex diseases.
- Haplotype analyses can have greater power than additive models in the analysis of joint effect of multiple loci because they incorporate LD among loci that can potentially reduce the degrees of freedom of test statistics.

Recently, haplotype association analyses have started to draw great attention [3]. However, the large number of SNPs genotyped in GWAS poses great computational challenges in the detection of haplotype associations.

There have been extensive studies [4], [5], [6], [7], [8], [9] on haplotype association analyses. One strategy adopted by most recent studies is to first partition the whole genome into small blocks and then focus on the analysis of major haplotypes within each block. Because haplotypes are not directly observable from genotype data in most cases, haplotypes need to be inferred in each block and associations will be assessed between the estimated haplotypes with a phenotype of interest.

The first well-known method to obtain haplotype information from genotype data was published by [10]. This algorithm requires that some individuals have unambiguous haplotypes (individuals with at most one heterozygous marker). However, with many markers, it often happens that all subjects are heterozygous at multiple loci and no individual has known haplotypes. Methods based on the Expectation-Maximization (EM) algorithm [11] were proposed to estimate haplotype frequencies for a small number of polymorphisms. But for larger numbers of markers, the EM method is computationally expensive and loses accuracy by using a suboptimal model for haplotype frequencies. Some work translate the haplotype inference into the task of missing haplotype allele imputation [12]. More accurate phasing (inferring haplotypes from genotype data) can be achieved with better a priori modeling of probabilities of haplotype configurations [13], as is done by the coalescent-based and hidden Markov model (HMM) methods [14], [15], [16], [17]. However, haplotype inference in GWAS may still take several days and remain a bottleneck. A recent work [18] used the EGEE computing grid (more than 40,000 CPUs) to analyze haplotype associations in the coronary artery disease (CAD) data from the Wellcome Trust Case Control Consortium (WTCCC) [19]. This analysis took around 45 days to finish.

In this work, we propose a new method to accelerate haplotype association analysis of GWAS data. Instead of working on haplotype estimation improvement to speed up the analysis of haplotype association, we use a different strategy. Intuitively, we know that among the large number of markers queried in GWAS, only a small portion of them may be relevant with the phenotype. It is computationally inefficient to estimate haplotypes in those irrelevant regions. If we could remove those irrelevant regions, then the entire process will be significantly faster. Therefore, our proposed method begins with an efficient screening process, which directly scans the genotype data to select possible association regions in a fast manner. Then the haplotype estimation and the association analysis are conducted only in the selected regions. The main components of the proposed method are summarized as follows:

- *X. Wan is with the Department of Computer Science and Institute of Computational and Theoretical Studies, Hong Kong Baptist University, Hong Kong. E-mail: eexiangwu@ust.hk.*
- *C. Yang and H. Zhao are with the Division of Biostatistics, School of Public Health, Yale University, New Haven, CT 06520.*
- *Q. Yang is with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong.*
- *W. Yu is with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong.*

Manuscript received 16 July 2012; revised 12 Nov. 2012; accepted 11 Jan. 2013; published online 22 Jan. 2013.

For information on obtaining reprints of this article, please send e-mail to: [tcbb@computer.org](mailto:tcbb@computer.org), and reference IEEECS Log Number TCBB-2012-07-0170. Digital Object Identifier no. 10.1109/TCBB.2013.6.

- **Screening.** A sliding-window approach is applied to partition the whole genome into multiple overlapping short windows. Each window is evaluated by the two-locus haplotype-based testing and those windows passing a specified threshold are selected. The motivation behind this screening step is the assumption that any disease-associated haplotype should contain at least one two-locus haplotype segment.
- **Phasing.** The genotype data in the selected windows are analyzed to obtain the haplotype information.
- **Testing.** A conditional testing is first applied to remove the redundant SNP markers in the estimated haplotypes

TABLE 1  
Haplotype Frequency Table for Two Loci

Control	B	b	Case	B	b
A	$f_{AB,0}$	$f_{aB,0}$	A	$f_{AB,1}$	$f_{Ab,1}$
a	$f_{aB,0}$	$f_{ab,0}$	a	$f_{aB,1}$	$f_{ab,1}$

and then produce the core haplotypes. Next, the classical chi-square test is used to measure the association significance of these core haplotypes.

Experiments on the WTCCC data sets show that our method is much faster than current methods. Meanwhile, it can find the same results in the CAD data as those in [18]. With this computational efficiency, we also identify some interesting haplotype association patterns from the data sets of other diseases. The source code of our method is publicly available at: <http://bioinformatics.ust.hk/HapBoost>.

## 2 METHOD

We assume a case-control design in which  $n$  independent subjects are genotyped at  $\mathcal{L}$  SNPs. We denote the phenotype of each subject by  $Y$ , with  $Y = 0$  corresponding to control subjects and  $Y = 1$  to case subjects. Because SNPs are biallelic markers, we code each allele as 0 (major allele) or 1 (minor allele). We use capital letters (e.g.,  $A, B, \dots$ ) to denote major alleles and lowercase letters (e.g.,  $a, b, \dots$ ) to denote minor alleles. The three genotypes of each SNP, homozygous reference genotype ( $AA$ ), heterozygous genotype ( $Aa$ ), and homozygous variant genotype ( $aa$ ), are coded as 0, 1, 2, respectively.

### 2.1 Screening

The screening process is based on the two-locus haplotype association testing. Here, we first show how to efficiently test the haplotype association of two loci. Then we present the details of the screening process. Let  $\{AB, Ab, aB, ab\}$  denote four haplotypes at two loci and  $f_{h,0}$  and  $f_{h,1}$  denote the frequency of haplotype  $h \in \{AB, Ab, aB, ab\}$  in controls and cases (see Table 1). Given four haplotype frequencies of two loci in both control group and case group, we can test the haplotype association by calculating the following absolute  $z$  value [20]

$$z = \frac{|\log(R) - \log(S)|}{\sqrt{SE(R) + SE(S)}}, \quad (1)$$

where

$$R = \frac{f_{AB,1}f_{ab,1}}{f_{Ab,1}f_{aB,1}}, SE(R) = \frac{1}{f_{AB,1}} + \frac{1}{f_{Ab,1}} + \frac{1}{f_{aB,1}} + \frac{1}{f_{ab,1}},$$

$$S = \frac{f_{AB,0}f_{ab,0}}{f_{Ab,0}f_{aB,0}}, SE(S) = \frac{1}{f_{AB,0}} + \frac{1}{f_{Ab,0}} + \frac{1}{f_{aB,0}} + \frac{1}{f_{ab,0}}.$$

TABLE 2  
Haplotype Representation of Joint  
Genotype Distribution at Two Loci

Locus A	Locus B	Counts	Haplotype configuration
AA	BB	$n_{00,1}$	$\{AB, AB\}$
AA	Bb	$n_{01,1}$	$\{AB, Ab\}$
AA	bb	$n_{02,1}$	$\{Ab, Ab\}$
Aa	BB	$n_{10,1}$	$\{AB, aB\}$
Aa	Bb	$n_{11,1}$	$\{AB, ab\}$ or $\{Ab, aB\}$
Aa	bb	$n_{12,1}$	$\{Ab, ab\}$
aa	BB	$n_{20,1}$	$\{aB, aB\}$
aa	Bb	$n_{21,1}$	$\{aB, ab\}$
aa	bb	$n_{22,1}$	$\{ab, ab\}$

Only the genotype combination ( $Aa, Bb$ ) gives ambiguous haplotype configurations.

TABLE 3  
Estimating Haplotype Frequencies from  
the Genotype Data of Two Loci

Haplotype	Frequency
AB	$f_{AB,1} = 2 n_{00,1} + n_{01,1} + n_{10,1} + \theta n_{11,1}$
Ab	$f_{Ab,1} = 2 n_{02,1} + n_{01,1} + n_{12,1} + (1 - \theta) n_{11,1}$
aB	$f_{aB,1} = 2 n_{20,1} + n_{21,1} + n_{10,1} + (1 - \theta) n_{11,1}$
ab	$f_{ab,1} = 2 n_{22,1} + n_{21,1} + n_{12,1} + \theta n_{11,1}$

The haplotype frequencies ( $f_{h,0}, f_{h,1}$ ) are unknown but can be estimated from the genotype data. See [21] for details. Here, we briefly explain the estimation procedure for the sake of completeness. Table 2 explains the estimation of the haplotype frequencies from the genotype distribution of two loci in the case group (the estimation method in the control group is the same). Among nine genotypes, only one combination ( $Aa, Bb$ ) of the genotypes gives ambiguous haplotype configurations.

We use  $\theta$  to denote the proportion of the haplotype configuration of  $\{AB, ab\}$  from the genotype combination ( $Aa, Bb$ ). We denote the probabilities of the four haplotypes in the case group as  $p_{AB,1}$ ,  $p_{Ab,1}$ ,  $p_{aB,1}$ , and  $p_{ab,1}$ , respectively. Then the frequencies of these four haplotypes can be calculated as in Table 3.

Assuming Hardy-Weinberg equilibrium (HWE), the probability of obtaining the genotype ( $Aa, Bb$ ) is  $p_{AB,1} p_{ab,1} + p_{aB,1} p_{Ab,1}$ . Then the proportion of the haplotype configuration of  $AB, ab$  is

$$\theta = \frac{p_{AB,1} p_{ab,1}}{p_{AB,1} p_{ab,1} + p_{aB,1} p_{Ab,1}}. \quad (2)$$

Here,  $p_h$  can be estimated with  $f_h/2n$ , where  $n$  is the number of samples. Notice that (2) is a cubic function of  $\theta$ , which has a closed-form solution. Correspondingly, the haplotype frequencies in Table 3 can be estimated straightforwardly and association effects of these haplotypes can be measured efficiently.

Based on the fast estimation of two-locus haplotype association, we design a computationally efficient screening method to select the candidate regions from the whole genome. We consider sliding windows of  $W$  adjacent SNPs. Within each window, we compute the  $z$  scores of all SNP pairs and use their average  $\bar{z}$  as an approximation of haplotype association of this window. The window width  $W$  is specified by users and the adjacent windows are overlapped by  $W - 1$  SNPs. All  $\bar{z}$  are collected and sorted. The windows with their  $\bar{z}$  in the top  $T$  percent will be selected. If two selected windows overlap with more than  $W/2$  SNPs, they will be merged into one short genomic region for further examination.

### 2.2 Phasing

After the screening stage, we expect only a limited number of genomic regions to remain. Therefore, the inference of haplotypes for selected regions can be done quickly. In our method, we use HaploRec [22], a statistical haplotype reconstruction algorithm developed for large-scale studies. It is particularly suitable for data sets with a large number of subjects. With sample sizes large enough, its results appeared to be the best compared to many other methods such as Phase [14] and fastPhase [16] in the simulation studies. It is several orders of magnitude faster than fastPhase and its running time is roughly linear with respect to the number of subjects and the number of markers. The reader is referred to [22] for details of HaploRec.

### 2.3 Testing

After the phasing stage, we obtain the haplotype information for all selected regions. Given one selected genomic region  $G_{1,\dots,k}$  ( $k$  is

1. The reason we use mean instead of median is due to the computing efficiency. The mean can be computed in one iteration, but the calculation of the median needs sorting. There are at least 105  $z$  values for each window. Thus, the outliers have negligible effect on the mean.

TABLE 4  
Observed Contingency Table for Haplotype  $h$

Groups	$h_1$	$\cdots$	$h_s$	Total
Controls ( $Y = 0$ )	$x_{01}$	$\cdots$	$x_{0s}$	$N_0$
Cases ( $Y = 1$ )	$x_{11}$	$\cdots$	$x_{1s}$	$N_1$
	$M_1$	$\cdots$	$M_s$	$N$

the number of SNPs in the region) with the estimated haplotype set  $H = \{h_i\}_{1 \leq i \leq s}$  ( $s$  is the number of estimated haplotypes), an intuitive method to test the haplotype association is to first collect a count table as Table 4 and then compute the test-statistic as described in (3)

$$X^2 = \sum_{i=0}^1 \sum_{j=1}^s \frac{(x_{ij} - u_{ij})^2}{u_{ij}}, u_{ij} = \frac{N_i M_j}{N}. \quad (3)$$

Under the null hypothesis that  $\{h_i\}_{1 \leq i \leq s}$  and the disease trait are independent from each other,  $X^2$  has an asymptotic  $\chi^2$  distribution with  $(s - 1)$  degrees of freedom. However, the power of this test may be compromised by the large number of rare haplotypes consisting of unimportant markers (unimportant in the sense of showing insignificant phenotype association) in the selected regions. Therefore, it may not be optimal to analyze all the haplotypes constructed from all markers in the region.

In our method, we design a sequential selection procedure to remove those markers uninformative for association signals. In this procedure, a SNP will be removed unless it contributes a significant amount of information to the phenotype association. The typical single-locus-based analysis without utilizing haplotype information cannot be applied here because it may remove those loci that individually display weak association but jointly show strong haplotype associations. Instead, we use the Mantel-Haenszel test statistic [23], which is often used in the analysis of stratified categorical data. In Mantel-Haenszel test, the data are arranged in a series of associated  $2 \times 2$  contingency tables. The null hypothesis is that the observed response is independent of the factor used in any  $2 \times 2$  contingency table.

Given the observed region  $G_{1 \dots k}$  with the estimated haplotype set  $H = \{h_i\}_{1 \leq i \leq s}$ , we sequentially check every SNP. Suppose  $G_l (1 \leq l \leq k)$  is the one to be evaluated. We first remove  $G_l$  from  $H$  to form a subhaplotypes  $H_{\setminus G_l}$  (i.e.,  $H$  without  $G_l$ ). For each  $h_i \in H_{\setminus G_l}$ , we collect the allelic contingency table (shown in Table 5) between  $G_l$  and the phenotype  $Y$ .

The Mantel-Haenszel test statistic for  $G_l$  is computed using

$$MH_{G_l} = \frac{\left[ \sum_{h_i} \left( x_{11}^{h_i} - \frac{N_1^{h_i} \cdot M_1^{h_i}}{N^{h_i}} \right) \right]^2}{\sum_{h_i} \frac{M_0^{h_i} \cdot M_1^{h_i} \cdot N^{h_i} \cdot N^{h_i}}{N^{h_i} \cdot N^{h_i} \cdot (N^{h_i} - 1)}}. \quad (4)$$

The null hypothesis in this test is that  $G_l$  and  $Y$  are not associated in samples with any haplotype  $h_i$ . Under the null hypothesis,  $MH_{G_l}$  has an asymptotic  $\chi^2$  distribution with one degree of freedom. Given a confidence interval (0.90 for a typical setting) of  $\chi^2(1)$ , we can decide if we should keep  $G_l$  in the haplotype association analysis or not. This process will be repeated until there is no SNP to be removed. Then we build the core haplotypes and conduct the association test as described above (3). The  $P$ -value is adjusted using the Bonferroni correction. In our method, one test corresponds to the evaluation of one genomic region. Since the screening process in our method checks every SNP as a starting point of a window, the number of SNPs is used in the Bonferroni correction.

## 2.4 Algorithm

To summarize, Algorithm 1 describes three stages of haplotype association detection method.

TABLE 5  
Observed Contingency Table  
between  $G_l$  and  $Y$  for Haplotype  $h_i$

Group	$Allele_{G_l} = 0$	$Allele_{G_l} = 1$	Total
Controls ( $Y = 0$ )	$x_{00}^{h_i}$	$x_{01}^{h_i}$	$N_0^{h_i}$
Cases ( $Y = 1$ )	$x_{10}^{h_i}$	$x_{11}^{h_i}$	$N_1^{h_i}$
	$M_0^{h_i}$	$M_1^{h_i}$	$N^{h_i}$

## Algorithm 1. HapBoost

**Given:**  $G_{1 \dots n}$  (genotype data with  $n$  SNPs),  $W$  (window width),  $T$  (selection threshold)

**Screening:**

**for**  $0 \leq i < n - W$  **do**

**for**  $i \leq s, t < i + W$  **do**

        Estimate  $\theta$  for  $G_s$  and  $G_t$  using Eq. (2)

        Compute  $z_{s,t}$  using Eq. (1)

**end for**

$\bar{z}_i = \text{Average}(z_{s,t})$

**end for**

Sort  $\bar{z}_i$  and select the top  $T$  percent and denote their indices as  $\{I_1 \cdots I_p\}$

**Phasing:**

**for**  $1 \leq s, t \leq p$  **do**

**if**  $|I_s - I_t| < W$  **then**

        Merge  $I_s$  and  $I_t$

**end if**

**end for**

Call HaploRec to estimate haplotypes on the merged genomic regions.

**Testing:**

**for each**  $G_{1 \dots k}^i$  with the estimated haplotypes  $H^i$  **do**

**for**  $1 \leq l \leq k$  **do**

        Remove  $G_l^i$  from  $H^i$  and build  $H_{\setminus G_l^i}^i$

        Compute  $MH_{G_l^i}$  using Eq. (4)

**if**  $MH_{G_l^i} \leq 2.70$  **then**

            /\*2.70 corresponds the 0.90 percentile of  $\chi^2(1)$ \*/

            Remove  $G_l^i$  from  $G_{1 \dots k}^i$

$H^i = H_{\setminus G_l^i}^i$

**end if**

**end for**

    Compute the test statistic  $X^2$  using Eq. (4) and output the  $P$ -value.

**end for**

Adjust  $P$ -values using the Bonferroni correction with the number of SNPs  $n$

## 3 RESULTS

### 3.1 Simulation Studies

In this section, we compare our method HapBoost with BEAGLE [17], HapMiner [24], and the *locfdr* program [25] using the synthetic data:

- BEAGLE is a powerful and popular method for the analysis of large-scale genetic data sets with hundreds of thousands of markers genotyped on thousands of samples. It uses the localized haplotype clustering and fits the data using an EM method. It outperforms many methods in terms of both computational speed and measures of accuracy for large whole-genome data sets.

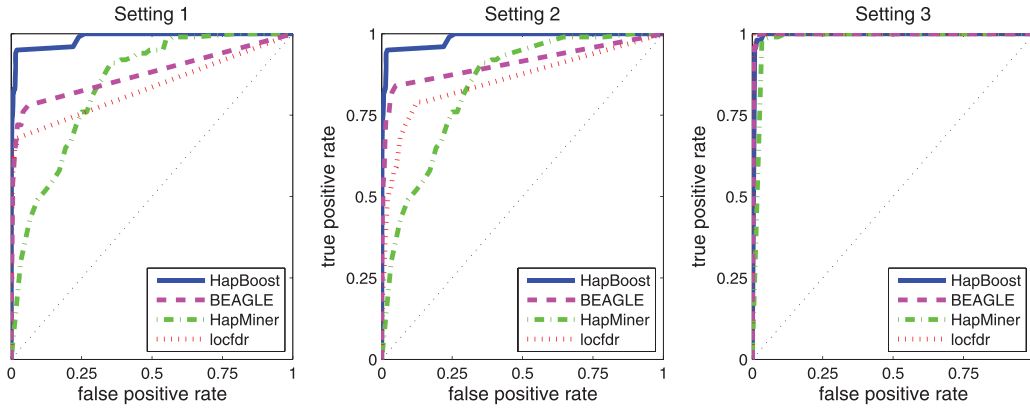


Fig. 1. The performance comparison of HapBoost, BEAGLE, HapMiner, and *locfdr* using ROC curve. From left to right, the  $z$  score of the causal marker is 2.0, 3.0, and 4.0, respectively.

- HapMiner is a data mining approach that utilizes a density-based clustering algorithm to find haplotype associations. The approach considers haplotype segments as data points in a high-dimensional space. Clusters are then identified based on a similarity measure using the density-based clustering algorithm. The association significance of each cluster is then evaluated. It has been shown that HapMiner can effectively obtain meaningful information from noisy data sets.
- The *locfdr* program extends the original false discovery rate (FDR) [26] to the local FDR using the two-group model. The  $z$  value of each SNP can be computed using the Cochran-Armitage trend test. It assumes that  $z_i$  with  $i \in \mathcal{G}_0$  comes from the null distribution  $f_0(z|\theta_0)$  with probability  $p_0$  and others come from the alternative distribution  $f_1(z|\theta_1)$  with probability  $p_1 = 1 - p_0$ , where  $\theta_0, \theta_1$  are parameters of the distributions  $f_0$  and  $f_1$ , respectively. Under mild assumptions,  $p_0, \theta_0, p_1$ , and  $\theta_1$  can be accurately estimated from data. We note that the *locfdr* program was developed for broader applications and it is used here to illustrate the power of single marker analysis.

### 3.1.1 Data Generation

To simulate realistic LD patterns of SNP data, we use the Hapmap configuration to generate the synthetic data. We download a genomic region of chromosome 10, which contains 2,000 SNPs in 280 haplotype blocks. The number of SNPs in each block ranges from 2 to 35. We use the configuration of this region to generate a Markov chain, in which each block is a state that consists of several common haplotypes controlled by an emission distribution. The connections between adjacent haplotype blocks are specified by a transition probability matrix. In more details, each block consists of a number of markers, the common haplotypes with their population frequencies, and the transition probabilities of each common haplotype connecting the common haplotypes in the next block. A random walk in this chain will generate one chromosome of one sample. Repeating this process will generate the sample pool. Fig. 1 in the supplementary document, which can be found on the Computer Society Digital Library at <http://doi.ieeecomputersociety.org/10.1109/TCBB.2013.6>, displays the LD patterns in the generated sample pool. It matches the LD patterns in the same genomic region of Hapmap data.

To generate the case-control data, we first randomly select one haplotype block (containing 7 major haplotypes) as the block containing the causal variant, denoted as  $H_d$ , and then use the following logistic regression model to assign class labels to samples

$$Pr(Y = 1|X_{H_d} = \{h_i, h_j\}) = \frac{\exp(\beta_0 + \beta_i + \beta_j)}{1 + \exp(\beta_0 + \beta_i + \beta_j)}, \quad (5)$$

where  $Y$  is the case-control label,  $\{h_i, h_j\}$  are the two haplotypes of the sample  $X$  in the block  $H_d$ ,  $\beta_i$ , and  $\beta_j$  indicate the association effect of  $h_i$  and  $h_j$ , respectively. Here, we assume the additive model of the disease risk on the log scale.

This simulation study aims to compare our method with other methods in the situation that the marginal effect of single SNP is weak, but the haplotype effect (joint effect of multiple SNPs) is significant. This situation poses a challenging issue in the analysis of GWAS data. Given the association effects  $\beta_{1,\dots,s}$  for all haplotypes in  $H_d$ , we can compute the expectation of  $z$  score for every marker in  $H_d$ . However, given the expectation of  $z$  score for every marker in  $H_d$ , there is no closed-form solution to compute the association effects  $\beta_{1,\dots,s}$ . Thus, in the simulation, we design a brute-force strategy to estimate  $\beta_{1,\dots,s}$ . In this strategy, a threshold of  $z$  score is first given. Next, the association effects  $\beta_{1,\dots,s}$  for all haplotypes are randomly assigned and the absolute  $z$  score for every marker is computed. The process repeats until the maximum absolute  $z$  score of markers is less than the given threshold.

The one with the maximum absolute  $z$  score will be considered as the causal variant. We use three thresholds (2.0, 3.0, 4.0) to control the  $z$  score of the causal variant in  $H_d$  to generate different data sets. For each threshold, we generate 100 data sets. Each data set contains 2,000 samples and 2,000 markers.

### 3.1.2 Performance Comparison

The performance of these methods using ROC curve (a plot of true positive rate versus false positive rate) is shown in Fig. 1. When the effect of the causal variant is weak, HapBoost performs the best. As the effect of the causal variant increases, all the methods shows similar performance. It may not be surprising that HapBoost and BEAGLE outperform the *locfdr* method. The main advantage of *locfdr* is to detect SNPs with differentiable marginal effects from a large number of makers without any effects. It cannot find SNPs with mild or weak marginal effects while their residing haplotypes show strong associations with phenotype. However, it seems surprising that HapMiner does not perform as well as *locfdr* even though it also uses the haplotype information to find associations. It displays a high-false positive rate. The possible explanation is maybe the parameter sensitivity of HapMiner [27]. HapMiner is a density-based clustering method that needs five parameters to specify the clustering process. It is usually difficult to automatically determine the parameter values and thus the user input is required. A large number of trials are needed to obtain the best combination of five parameters in HapMiner. Using the default setting may underestimate the power of HapMiner.



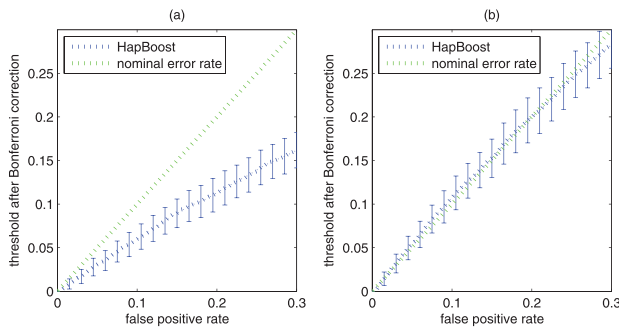


Fig. 2. The type I error rates in null simulation. (a) Null simulation with LD. (b) Null simulation without LD.

### 3.1.3 Null Simulation for Testing Type I Errors

To show the type I errors of our method, we conduct null simulation in two scenarios:

- *Scenario 1—without LD.* The generation of the haplotype pool is the same as the process described in the comparison experiment. The sample is generated by merging two haplotypes randomly selected from the haplotype pool. We generate 5,000 null data sets, each of which contains 2,000 SNPs and 2,000 samples.
- *Scenario 2—without LD.* We generate 5,000 null data sets. Each data set contains 2,000 SNPs and 2,000 samples. All SNPs are generated independently with major allele frequencies (MAFs) uniformly distributed in  $[0.05, 0.5]$ .

The result is shown in Fig. 2. It can be seen that the type I errors of our method in both settings agree with the nominal error rates. Due to the LD pattern that reduces the degrees of freedom, the error rates in Scenario 2 are lower than those in Scenario 1.

## 3.2 Experiments on WTCCC Data Sets

We have applied our method to analyze the data (14,000 cases in total and 3,000 shared controls) from the WTCCC on seven common human diseases: bipolar disorder (BD), CAD, Crohn's disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D) and type 2 diabetes (T2D). The quality control procedure is the same as that used in [19], including missing value process, HWE test and removing SNPs with minor allele frequencies less than 0.05. The number of SNPs for the haplotype analysis and the number of identified genomic regions for seven diseases are reported in Table 6. The details about each region are provided in the online supplementary document.

### 3.2.1 CAD

In [18], the authors identified the *SLC22A3-LPAL2-LPA* gene cluster in chromosome 6 as a strong disease-associated genetic region. Our method exactly repeats this finding (see Table 1 in the online supplementary document). The uncorrected  $P$  value is  $4.59 \times 10^{-8}$ , which is very close to  $4.34 \times 10^{-8}$  reported in [18]. The minor difference is due to the small difference of estimated haplotypes. Please note that the minimum uncorrected  $P$  value of individual marker in this region is 0.001, which is very weak in the large-scale hypothesis testing. Thus, the single-locus-based method, such as the *locfdr* program, cannot identify this region.

BEAGLE and HapMiner cannot be directly applied to analyze the data set because they need to estimate the haplotypes from the data before conducting the analysis.

### 3.2.2 Other Diseases

In other diseases, we have identified some interesting susceptibility regions (see Tables 2, 3, 4, 5, and 6 in the online supplementary document) besides the well-known major histocompatibility complex (MHC) region. Many of these regions only contain markers with weak marginal effects. In BD, we report a region (rs1635003-rs12704342) in chromosome 7, which contains gene GRM3. A similar result has been reported in [28] that one SNP rs6465084 in gene GRM3 was associated with a fourfold increased risk of lifetime history of psychotic symptoms, and thus the authors confirmed that psychosis and relapse in BD are related to gene GRM3. Another reported region (rs2893863-rs10994594) in chromosome 10 contains gene CDK1, whose function is closely related to gene GSK3 [29]. The pharmacological inhibition of GSK3 activity lithium is the most common treatment for BD [30].

## 3.3 Computation Time

From a practical point of view, a key issue of finding haplotype associations in GWAS is the computational efficiency. As we mentioned above, [18] spent 45 days to finish the analysis of one data set from WTCCC using the EGEE computing grid. For the same data set, our method was able to finish the analysis of the haplotype association within one day using a standard 3.0-GHz desktop with 4 G memory running Windows XP system. The computational efficiency is significantly improved.

## 3.4 Parameter Sensitivity

There are two parameters to be specified in HapBoost. One is the selection threshold  $T$  for selecting candidate regions and the other is the window width  $W$  in the sliding window screening. The default settings are  $T = 0.01$  and  $W = 15$ . HapBoost is robust to  $T$  because we find increasing  $T$  does not change the results but it will increase the running time linearly. The window width  $W$  is a key parameter of HapBoost. If the distance between the causal variant and the closest linkage marker is bigger than  $W$  and the LD between them is weak, then HapBoost may not identify the genomic region covering the causal variants. In this regards, a large  $W$  is desired. However,  $W$  cannot be too large since the running time spent on phasing will be significantly increased. More importantly, a large  $W$  will give rise to a large number of haplotypes. With a limited number of samples, the power, and the stability will be significantly reduced. We have conducted one experiment to show the performance of our method under different settings of  $W$ . Please see Fig. 2 in the online supplementary document for details. As the experimental result displays, our method performs slightly different for different setting of  $W$ .

## 4 CONCLUSION

During the last few years, there have been growing interests in developing and applying computational and statistical approaches to finding haplotype associations. However, this task is computationally challenging in GWAS due to the large number of SNPs. Presently, the majority of existing methods have been focusing on detecting signals from a small number of loci. In this paper, we

TABLE 6  
The Numbers of Diseased-Associated Genomic Regions Identified from Seven Diseases Data Sets

Disease	BD	CAD	CD	HT	RA	T1D	T2D
SNP Number	351,545	352,068	354,181	352,924	352,536	352,538	351,542
Identified genomic regions	38	28	55	35	50	40	34

have presented HapBoost to analyze seven data sets from WTCCC. Our experimental results have demonstrated that HapBoost is both computationally efficient and statistically powerful in GWAS.

There are some limitations of HapBoost. HapBoost has mainly focused on the genome-wide case-control studies, i.e., the disease phenotype can be represented as a binary variable. In the current stage, our method cannot be applied to genome-wide association studies involving continuous phenotypes unless those continuous phenotypes can be discretized. Another limitation is that HapBoost uses the two-locus haplotype testing to preselect candidate regions. In the situation that the LD of any two loci in the region containing the causal variant is weak, HapBoost may not find this region and the multilocus haplotype testing has to be applied. At present, multilocus (more than 2) haplotype testing involves inferring haplotypes from samples, which is not fast enough to be applied at the genome-wide scale. We will investigate this issue in the future work.

## REFERENCES

- [1] A. Clark, "The Role of Haplotypes in Candidate Gene Studies," *Genetic Epidemiology*, vol. 27, pp. 321-333, 2004.
- [2] Wikipedia, "Haplotype—Wikipedia, the Free Encyclopedia," <http://en.wikipedia.org/wiki/Haplotype>, 2004.
- [3] J. Kang, S. Kugathasan, M. Georges, H. Zhao, and J. Cho, "Improved Risk Prediction for Crohn's Disease with a Multilocus Approach," *Human Molecular Genetics*, vol. 20, no. 12, pp. 2435-2442, 2011.
- [4] D. Schaid, C. Rowland, D. Tines, R. Jacobson, and G. Poland, "Score Tests for Association between Traits and Haplotypes When Linkage Phase Is Ambiguous," *The Am. J. Human Genetics*, vol. 70, no. 2, pp. 425-434, 2002.
- [5] D.O. Stram, C. Leigh Pearce, P. Bretsky, M. Freedman, J.N. Hirschhorn, D. Altshuler, L.N. Kolonel, B.E. Henderson, D.C. Thomas, "Modeling and E-M Estimation of Haplotype-Specific Relative Risks from Genotype Data for a Case-Control Study of Unrelated Individuals," *Human Heredity*, vol. 55, pp. 179-190, 2003.
- [6] L. Zhao, S. Li, and N. Khalid, "A Method for the Assessment of Disease Associations with Single-Nucleotide Polymorphism Haplotypes and Environmental Variables in Case-Control Studies," *Am. J. Human Genetics*, vol. 72, no. 5, pp. 1231-1250, 2003.
- [7] D. Lin, "An Efficient Monte Carlo Approach to Assessing Statistical Significance in Genomic Studies," *Bioinformatics*, vol. 21, no. 6, pp. 781-787, 2005.
- [8] A. Morris, "A Flexible Bayesian Framework for Modeling Haplotype Association with Disease, Allowing for Dominance Effects of the Underlying Causative Variants," *Am. J. Human Genetics*, vol. 79, no. 4, pp. 679-694, 2006.
- [9] T. Druet and M. Georges, "A Hidden Markov Model Combining Linkage and Linkage Disequilibrium Information for Haplotype Reconstruction and Quantitative Trait Locus Fine Mapping," *Genetics*, vol. 184, no. 3, pp. 789-798, 2010.
- [10] A. Clark, "Inference of Haplotypes from PCR-Amplified Samples of Diploid Populations," *Molecular Biology and Evolution*, vol. 7, no. 2, pp. 111-122, 1990.
- [11] L. Excoffier and M. Slatkin, "Maximum-Likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population," *Molecular Biology and Evolution*, vol. 12, no. 5, pp. 921-927, 1995.
- [12] Y. Wang, Z. Cai, P. Stothard, S. Moore, R. Goebel, L. Wang, and G. Lin, "Fast Accurate Missing SNP Genotype Local Imputation," *BMC Research Notes*, vol. 5, no. 1, article 404, 2012.
- [13] S. Browning and B. Browning, "Haplotype Phasing: Existing Methods and New Developments," *Nature Rev. Genetics*, vol. 12, no. 10, pp. 703-714, 2011.
- [14] M. Epstein and G. Satten, "Inference on Haplotype Effects in Case-Control Studies Using Unphased Genotype Data," *Am. J. Human Genetics*, vol. 73, no. 6, pp. 1316-1329, 2003.
- [15] M. Stephens and P. Donnelly, "A Comparison of Bayesian Methods for Haplotype Reconstruction from Population Genotype Data," *Am. J. Human Genetics*, vol. 73, no. 5, pp. 1162-1169, 2003.
- [16] P. Scheet and M. Stephens, "A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase," *Am. J. Human Genetics*, vol. 78, no. 4, pp. 629-644, 2006.
- [17] S. Browning and B. Browning, "Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies by Use of Localized Haplotype Clustering," *Am. J. Human Genetics*, vol. 81, no. 5, pp. 1084-1097, 2007.
- [18] D. Trégouët et al., "Genome-Wide Haplotype Association Study Identifies the SLC22A3-LPAL2-LPA Gene Cluster as a Risk Locus for Coronary Artery Disease," *Nature Genetics*, vol. 41, no. 3, pp. 283-285, 2009.
- [19] The Wellcome Trust Case Control Consortium, "Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls," *Nature*, vol. 447, no. 7145, pp. 661-678, 2007.
- [20] J. Morris and M. Gardner, "Statistics in Medicine: Calculating Confidence Intervals for Relative Risks (Odds Ratios) and Standardised Ratios and Rates," *British Medical J. (Clinical Research ed.)*, vol. 296, no. 6632, pp. 1313-1316, 1988.
- [21] D. Siegmund and B. Yakir, *The Statistics of Gene Mapping (Statistics for Biology and Health)*. Springer, 2007.
- [22] L. Eronen, F. Geerts, and H. Toivonen, "HaploRec: Efficient and Accurate Large-Scale Reconstruction of Haplotypes," *BMC Bioinformatics*, vol. 7, no. 1, article 542, 2006.
- [23] T. O'Gorman, R. Woolson, M. Jones, and J. Lemke, "Statistical Analysis of  $K \times 2 \times 2$  Tables: A Comparative Study of Estimators/Test Statistics for Association and Homogeneity," *Environmental Health Perspectives*, vol. 87, pp. 103-107, 1990.
- [24] J. Li and T. Jiang, "Haplotype-Based Linkage Disequilibrium Mapping via Direct Data Mining," *Bioinformatics*, vol. 21, no. 24, pp. 4384-4393, 2005.
- [25] B. Efron, "Large-Scale Simultaneous Hypothesis Testing," *J. Am. Statistical Assoc.*, vol. 99, no. 465, pp. 96-104, 2004.
- [26] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *J. Royal Statistical Soc., Series B (Methodological)*, pp. 289-300, 1995.
- [27] L. Lin, L. Wong, T. Leong, and P.S. Lai, "Efficient Mining of Haplotype Patterns for Linkage Disequilibrium Mapping," *J. Bioinformatics and Computational Biology*, vol. 8, no. 1, pp. 127-146, 2010.
- [28] S. Dalvie, N. Horn, C. Nossek, L. van der Merwe, D. Stein, and R. Ramesar, "Psychosis and Relapse in Bipolar Disorder are Related to GRM3, DAOA, and GRIN2B Genotype," *African J. Psychiatry*, vol. 13, no. 4, pp. 297-301, 2010.
- [29] M. Leost, C. Schultz, A. Link, Y. Wu, J. Biernat, E. Mandelkow, J. Bibb, G. Snyder, P. Greengard, D. Zaharevitz, R. Gussio, A. Senderowicz, E. Sausville, C. Kunick, and L. Meijer, "Paullones are Potent Inhibitors of Glycogen Synthase Kinase-3beta and Cyclin-Dependent Kinase 5/p25," *European J. Biochemistry*, vol. 267, pp. 5983-5983, 2000.
- [30] S. Kaladchibachi, B. Doble, N. Anthopoulos, J. Woodgett, and A. Manoukian, "Glycogen Synthase Kinase 3, Circadian Rhythms, and Bipolar Disorder: A Molecular Link in the Therapeutic Action of Lithium," *J. Circadian Rhythms*, vol. 5, no. 1, article 3, 2007.

► For more information on this or any other computing topic, please visit our Digital Library at [www.computer.org/publications/dlib](http://www.computer.org/publications/dlib).