Statistics and Its Interface Volume 9 (2016) 267–280

Search for risk haplotype segments with GWAS data by use of finite mixture models[*]

Fadhaa Ali and Jian Zhang[†]

The region-based association analysis has been proposed to capture the collective behavior of sets of variants by testing the association of each set instead of individual variants with the disease. Such an analysis typically involves a list of unphased multiple-locus genotypes with potentially sparse frequencies in cases and controls. To tackle the problem of the sparse distribution, a two-stage approach was proposed in literature: In the first stage, haplotypes are computationally inferred from genotypes, followed by a haplotype coclassification. In the second stage, the association analysis is performed on the inferred haplotype groups. If a haplotype is unevenly distributed between the case and control samples, this haplotype is labeled as a risk haplotype. Unfortunately, the in-silico reconstruction of haplotypes might produce a proportion of false haplotypes which hamper the detection of rare but true haplotypes. Here, to address the issue, we propose an alternative approach: In Stage 1, we cluster genotypes instead of inferred haplotypes and estimate the risk genotypes based on a finite mixture model. In Stage 2, we infer risk haplotypes from risk genotypes inferred from the previous stage. To estimate the finite mixture model, we propose an EM algorithm with a novel data partition-based initialization. The performance of the proposed procedure is assessed by simulation studies and a real data analysis. Compared to the existing multiple Z-test procedure, we find that the power of genome-wide association studies can be increased by using the proposed procedure.

AMS 2000 subject classifications: Primary 62P10.
Keywords and phrases: Region-based association analysis, Genotype mixture models, Odds ratios, Genome wide association studies, Expectation-maximization algorithm.

1. INTRODUCTION

The advanced genotyping technology has made it possible to conduct genome-wide association studies (GWAS) on complex diseases in recent years [3, 19]. Genome-wide association studies systematically analyze genetic variation across the genome by its effects on phenotypic traits. The early landmark study using these technologies was the Wellcome Trust Case Control Consortium (WTCCC), which reported genetic association results for over 500,000 single nucleotide polymorphisms (SNPs) in seven disease sample sets of 2000 individuals each and 3000 control individuals [23]. Most of these studies were based on the so-called common-disease-common-variant hypothesis that the variants being sought are common to many individuals with the disease. To date, these studies have identified hundreds of signposts associated with disease. But the search for causative variants derived from them has been remarkably less successful, with only a handful of causative variants discovered in follow-up sequencing studies. The so-called winner's curse, where the detected effect is likely stronger in the GWAS sample than in the general population, is one of factors underpinning this phenomenon [28, 26]. On the other hand, many of the variants found have had only a weak effect on the risk of disease and therefore explained only a small proportion of the risk. Moreover, the signals in these studies might not always be pointing to a few common genetic variants but instead to many rare variants, each of which causes relatively few cases [14, 8]. The rapid increase in the number and the volume of GWAS provides an unprecedented opportunity to examine effects of rare variants on disease susceptibility. This also gives rise to a challenging problem of search for multiple variant sets in a high-dimensional genotype space. A popular strategy, suggested by the block-like structure of the human genome, is to segment each chromosome into a list of genetically meaningful regions. The multilocus haplotype, the ordered allele sequences on a chromosome, provides a unit of analysis for capturing linear and non-linear correlations among variants [15, 25, 22, 7]. In general, if a particular haplotype of a pre-specified group of SNPs is unevenly distributed between the case and control samples, this haplotype is highlighted as a risk haplotype. Identifying risk haplotypes is an important but hard task in genetics, because haplotypes are often unknown and sparsely distributed. In practice, what we can observe are genotypes not haplotypes. As each genotype is made up by two unknown haplotypes, the underlying haplotypes have to be inferred. Direct, laboratory-based haplotyping to infer the unknown phase are expensive ways to obtain haplotypes. So, people prefer to infer haplotypes from observed genotypes by us-

[*] We are grateful to the Wellcome Trust Case Control Consortium Data Access Committee for allowing us to use their WTCCC CAD and HT data. We also thank the Editor-in-Chief, an Associate Editor and three reviewers for their constructive comments that have substantially helped to improve the paper. The research of the first author was funded by the Ministry of Higher Education and Scientific Research, Iraq.

[†] Corresponding author.

ing the computational software such as PHASE [18, 16]. Many existing procedures suffers from the problem caused by sparsely distributed genotypes, where the resulting haplotype can also be sparsely distributed. To deal with the haplotype distribution sparsity, a number of haplotype clustering methods have been developed in literature [10, 20, 2, 27] and in references therein. However, computational inferred haplotypes may contain both true and false haplotypes, resulting in a high false discovery rate of risk haplotypes. This paper aims to improve over PHASE to achieve a more precise classification of haplotypes and subsequently improve the power of identifying risk haplotypes.

We first propose a finite mixture model for directly clustering genotypes on the basis of their prospective frequencies. The main advantage of the proposed model over the other existing methods is that it can reduce haplotyping-error effects on grouping rare haplotypes. Moreover, using the estimated prospective frequencies derived from a retrospective study to estimate genotype (and haplotype) disease odds ratio is known to be asymptotically consistent even though the prospective frequency estimators may not be [12]. The rationale behind the proposal is as follows. We assume that haplotypes of a specific chromosome segment can be classified as risk or non-risk (neutral and protective) and that the corresponding genotypes can be grouped into three categories $\nu = 0, 1, 2$, where in category ν, the genotypes contain ν risk haplotypes. Given the total number of individuals with genotype j and risk category ν, we further split the number into the accounts of individuals with disease or without disease. This gives rise to the genotype frequency contingency table, where rows stand for the disease status (case or control) and columns for genotypes. We can directly assess whether two genotypes belong to the same group by their column similarity in the table. Formally, given its risk category, we regard each genotype account in cases as a random variable following a binomial distribution. Then, integrating over its risk category, each genotype account in cases can be viewed as a random variable following a three-component binomial mixture model. So, we fit each column in the above contingency table by a binomial distribution with the disease-penetrance as the success probability and infer the grouping of these columns through use of three-component binomial mixtures. The fitted mixture model is then utilized to decide whether or not a specific genotype belong to a risk group. Consequently, the number of potential risk genotypes to be examined further is substantially reduced. This helps us reduce the error of identifying risk haplotypes in the haplotype thresholding stage.

We employ the expectation-maximization (EM) algorithm to calculate the maximum likelihood estimator for the proposed mixture model. The EM algorithm can guarantee monotone convergence to a local maximum. On the other hand, it needs to choose initial values in order to reach a local maximum which is close to the global maximum. The existing methods for initialization include: multiple random initializations, initially grouping the data and among others [6]. In this paper, we propose a new initialization procedure by grouping the estimated genotype frequencies. We conduct simulation studies on the proposed method in both prospective and retrospective design settings, showing that our method can outperform the approach of Zhu et al. [27] in most cases. We also apply both the proposed method and the method of Zhu et al. [27] to the Coronary Artery Disease (CAD) and Hypertension (HT) data in the Wellcome Trust Case Control Consortium (WTCCC), identifying potential risk haplotypes for each pre-specified chromosomal region.

The rest of the paper is organized as follows. The proposed methodology is introduced in Section 2. The simulation studies and real data applications are presented in Sections 3 and 4. Discussions and conclusion are made in Section 5. The details on the haplotype reconstruction software PHASE and the EM algorithm can be found in the Appendices.

2. METHODOLOGY

Consider a case-control sample with N_0 controls and N_1 cases, typed at m SNP markers in a candidate region, yielding unphased genotype set $\mathbf{G}$. Suppose that $\mathbf{G}$ contains distinct genotypes $G_j, 1 \leq j \leq J^*$ with counts N_{0j}, N_{1j} in controls and cases respectively. To tackle the issue of extremely rare genotypes, we first collapsed these genotypes by defining the set

$$G_c = \left\{ G_j | N_{0j} = 0 \text{ or } N_{1j} = 0 \text{ or } \frac{N_{0j} + N_{1j}}{N_0 + N_1} \leq 0.001, \right. \\ \left. j = 1, ..., J^* \right\},$$

where we say that G_j is extremely rare if its prospective frequency is less than 0.1%. A pilot simulation indicates that the collapsing of extremely rare genotypes can improve the accuracy of genotype co-classification (the data are not shown here). By the term “extremely rare genotype”, we imitate the similar concept in the literature [11], where an allele is called rare if its frequency is less than 1%. With a slight abuse of notation, we still denote these non-extreme genotypes as $G_1, ..., G_{J-1}$ with accounts $N_{0j}, N_{1j}, 1 \leq j \leq J - 1$, and the set G_c by G_J with the collapsed account N_{0J} and N_{1J} in controls and cases respectively. We write $\mathbf{N} = \{(N_{0j}, N_{1j}) : 1 \leq j \leq J\}$ and rewrite $\mathbf{G} = \{G_1, ..., G_J\}$. Let $\mathbf{H}^2$ denote all haplotype pairs reconstructed from $\mathbf{G}$ by using the software PHASE [18]. A brief introduction to PHASE can be found in the Appendix A.

2.1 Two-stage procedure

We introduce the following two-stage approach for finding risk haplotypes. In Stage 1, genotypes are clustered and risk genotypes are derived, whereas in Stage 2 the odds ratio thresholding is employed to infer risk-haplotypes. As the reconstructed haplotypes may contain errors, to reduce the effect of hapolotying errors on clustering, we co-classify genotypes instead of the inferred haplotypes in Stage 1. The details are given below.

Stage 1 (genotype clustering): We assume that haplotypes can be annotated by two categories: risk and non-risk, where non-risk category include both neutral and protective risk haplotypes. As each genotype consists of a haplotype pair, the observed genotypes can be clustered into three categories according to the numbers of risk haplotypes which they have. In light of the above fact, given genotype counts $(N_{0j}, N_{1j}) : 1 \leq j \leq J$, we consider the following three-component binomial mixture model:

$$\begin{aligned} f((N_{0j}, N_{1j})^T|\theta) &= \pi_0 f((N_{0j}, N_{1j})^T|q_0) \\ &\quad +\pi_1 f((N_{0j}, N_{1j})^T|q_1) \\ &\quad +\pi_2 f((N_{0j}, N_{1j})^T|q_2), \end{aligned}$$

where $\theta = (q_0, q_1, q_2, \pi_0, \pi_1, \pi_2)^T$ with $0 \leq q_\nu \leq 1, 0 \leq \pi_\nu \leq 1, \nu = 0, 1, 2, \pi_0 + \pi_1 + \pi_2 = 1$, and

$$f((N_{0j}, N_{1j})^T|q_\nu) = \binom{N_j}{N_{1j}} q_\nu^{N_{1j}} (1 - q_\nu)^{N_{0j}}, \nu = 0, 1, 2$$

with $N_j = N_{0j} + N_{1j}$. Note that q_0, q_1 and q_2 are the unknown disease penetrances for genotypes which contain 0, 1, and 2 risk haplotypes respectively.

The (incomplete) likelihood of θ given data $\mathbf{N}$ can be calculated by

$$L(\theta|\mathbf{N}) = \prod_{j=1}^{J} f((N_{0j}, N_{1j})^T|\theta).$$

We take the maximum likelihood estimator (MLE) $\hat{\theta}$ to estimate the unknown parameter θ. We employe the so-called expectation-maximization (EM) algorithm [9] to calculate $\hat{\theta}$. To this end, we introduce the following complete log-likelihood

$$l(\theta|\mathbf{N}, \mathbf{I}) = \sum_{j=1}^{J} \sum_{\nu=0}^{2} I_{\nu j} \log \left[\pi_\nu\ f((N_{0j}, N_{1j})^T|q_\nu) \right],$$

where $\mathbf{I} = \{(I_{0j}, I_{1j}, I_{2j})^T : 1 \leq j \leq J\}$and $(I_{0j}, I_{1j}, I_{2j})^T$ are unknown group membership indicators defined by

$$I_{\nu j} = \begin{cases} 1, & \text{if } G_j \text{ in the group } \nu \\ 0, & \text{otherwise} \end{cases} \qquad \nu = 0, 1, 2.$$

The further details on the EM algorithm can be found in the Appendix B.

Let the prospective frequencies of G_j in the controls and cases be estimated by

$$\hat{p}_{0j} = \frac{N_{0j}}{N_{0j} + N_{1j}}, \quad \hat{p}_{1j} = \frac{N_{1j}}{N_{0j} + N_{1j}}$$

respectively. Note that under the null hypothesis that the j-th genotype is not risk to the disease, then $X = (N_{0j} + N_{1j})\hat{p}_{1j}$ approximately follows a binomial distribution $f((N_{0j}, N_{1j})^T|\hat{q}_0)$ which can be further approximated by the Normal distribution with mean $\hat{q}_0$ and variance $\hat{q}_0(1 - \hat{q}_0)/(N_{0j} + N_{1j})$. In light of this fact, we can determine the risk status of genotype G_j by checking whether the value of the following Z-test statistic is larger than the critical value μ_j, i.e.,

$$(\hat{p}_{1j} - \hat{q}_0)/\sqrt{\hat{q}_0(1 - \hat{q}_0)/(N_{0j} + N_{1j})} > \mu_j.$$

Therefore, the risk-genotype group (which consists of genotypes with at least one risk haplotype) can be estimated by

$$\mathbf{G}_r = \{G_j : \hat{p}_{1j} > w_j, j = 1, ..., J\},$$

where

$$w_j = \hat{q}_0 + \mu_j \sqrt{\hat{q}_0(1 - \hat{q}_0)/(N_{0j} + N_{1j})}$$

and μ_j is determined by

$$(1) \qquad P\left(X \geq (N_{0j} + N_{1j})w_j\right) < \varepsilon,$$

with ε being a pre-specified constant. In the simulation studies later, around 100 different genotypes will be involved in each dataset. Using the Bonferroni correction, we set $\varepsilon = 0.05/100$ so that the total probability of type I errors involved in the thresholding is less than 0.05. Similarly, in the real data analysis section below, we will use the Bonferroni correction to set a different value of ε.

Stage 2 (haplotype thresholding): We introduce the following approach for identifying risk haplotypes, where only genotypes identified as in risk groups in Stage 1 are subject to further analysis. Let $\mathbf{H}_a^2$ be all haplotype pairs corresponding to $\mathbf{G}_r$, which are derived from $\mathbf{H}^2$ directly by taking advantage that $\mathbf{G}_r$ is a subset of $\mathbf{G}$. Let $\mathbf{H}_a = (h_1, ..., h_K)^T$ be all the distinct haplotypes in $\mathbf{H}_a^2$ with accounts n_{0k} and n_{1k}, $k = 1, ..., K$ in controls and cases respectively. For each k, we define

$$n_{0\bar{k}} = \sum_{t \neq k} n_{0t}, \quad n_{1\bar{k}} = \sum_{t \neq k} n_{1t}.$$

Note that $\mathbf{H}_a$ may contain non-risk haplotypes when $\mathbf{G}_a$ carries genotypes of a risk haplotype paired with a non-risk haplotype. For example, in the so-called dominant inheritance mode, risk haplotypes are often paired with non-risk haplotypes in producing genotypes. Therefore, to find risk haplotypes, we need to further threshold $\mathbf{H}_a$. It is well-known that the prospective frequency-based penetrance estimators with case-control data can be biased. However, the odds ratio estimator based on the prospective frequencies is asymptotically unbiased [12]. So, we use the odds ratio to judge whether a haplotype is risk or not. Here, non-risk haplotypes are defined as haplotypes which are neutral or protective to the disease. The technical details are described as follows.

We first calculate the odds ratio between h_k and $\mathbf{H}_a - \{h_k\}$ by

$$\mathrm{OR}_k = \frac{(n_{1k}+0.5)(n_{0\bar{k}}+0.5)}{(n_{0k}+0.5)(n_{1\bar{k}}+0.5)},$$

where adding 0.5 to the OR for the continuity correction was suggested by Agresti [1]. By simulations, Agresti [1] showed that in finite sample settings, the above estimator performed much better than the estimator without continuity correction. Note that under the null hypothesis in which the underlying odds ratio is one, the distribution of the estimated odds-ratio OR_k is asymptotically Normal distributed as

$$(2) \qquad \log(\mathrm{OR}_k) \sim N(0, \phi(n_{0k}, n_{1k}, n_{0\bar{k}}, n_{1\bar{k}})^2),$$

where

$$(3) \qquad \phi(n_{0k}, n_{1k}, n_{0\bar{k}}, n_{1\bar{k}})^2 = U_k,$$

with

$$U_k = \frac{1}{n_{0k}+0.5} + \frac{1}{n_{1k}+0.5} + \frac{1}{n_{0\bar{k}}+0.5} + \frac{1}{n_{1\bar{k}}+0.5}.$$

See [1]. Then, based on the above asymptotic distribution, we calculate the risk haplotype set $\mathbf{H}_r$ by

$$(4) \qquad \mathbf{H}_r = \left\{ h_k \in \mathbf{H}_a : \mathrm{OR}_k \geq \exp(c_1 U_k^{1/2}) \right\},$$

where c_1 is a pre-specified critical value.

2.2 Multiple testing method

To compare the proposed method to the multiple testing procedure of Zhu et al. [27], we briefly describe their procedure as follows. In their procedure, a subsample A containing $N_0^{(a)}$ and $N_1^{(a)}$ individuals are randomly chosen from the controls and cases respectively. These individuals are used in the screening stage and the remaining forms a validation subsample B to be used in the validation stage. Suppose that there are K different haplotypes inferred from A by using the PHASE. Let $(r_{0k}^{(a)}, r_{1k}^{(a)})$, $1 \leq k \leq K$ be their retrospective frequencies in controls and cases respectively.

Screening stage: We perform a respective frequencies-based screening by calculating an estimated risk haplotype set as follows:

$$S^{(a)} = \{h_k : z_k^{(a)} > c_0, 1 \leq k \leq K\},$$

where c_0 is a pre-specified constant ($c_0 = 1$ in our later simulations) and

$$z_k^{(a)} = \frac{r_{1k}^{(a)} - r_{0k}^{(a)}}{\sqrt{r_{0k}^{(a)}(1 - r_{0k}^{(a)})/(2N_1^{(a)})}}.$$

Validation stage: The $S^{(a)}$ is refined by performing Fisher's exact test based on subsample B for each haplotype in $S^{(a)}$. This gives a final risk haplotype set denoted by $S^{(b)}$.

3. SIMULATION STUDIES

In this section, via simulations we will examine the performance of the proposed methods in terms of the estimated L_1 bias and the average of sensitivity and specificity under various scenarios. Here, we suppose that the disease-penetrance of a genotype depends only on the number of risk haplotypes contained in that genotype. As each genotype consists of two haplotypes, we have three types of penetrance:

$$\begin{aligned} f_0 &= P(\text{disease}|H_{\bar{r}}H_{\bar{r}}), \quad f_1 = P(\text{disease}|H_r H_{\bar{r}}), \\ f_2 &= P(\text{disease}|H_r H_r), \end{aligned}$$

where H_r and $H_{\bar{r}}$ stand for risk and non-risk haplotypes respectively. Denote the relative risk measures by $\lambda_1 = f_1/f_0$ and $\lambda = f_2/f_0$. Let $\hat{\theta}$ be the estimator of θ, and $\mathbf{H}_r$ and $\mathbf{H}_{\bar{r}}$ the estimated true risk and non-risk haplotype sets respectively. Let $\mathbf{T}_r$ and $\mathbf{T}_{\bar{r}}$ be the true risk and non-risk haplotype sets. Then, by the L_1 bias we mean the L_1 distance between $\hat{\theta}$ and θ. By the sensitivity and specificity of $\mathbf{H}_r$ and $\mathbf{H}_{\bar{r}}$, we mean the positive discovery rate and the negative discovery rate:

$$\text{sen} = \frac{|\mathbf{H}_r \cap \mathbf{T}_r|}{|\mathbf{T}_r|} \text{ and spe} = \frac{|\mathbf{H}_{\bar{r}} \cap \mathbf{T}_{\bar{r}}|}{|\mathbf{T}_{\bar{r}}|}.$$

We take the average AVSS = (sen + spe)/2 to assess the performance of a haplotype classification procedure.

3.1 Performance of the proposed data partition-based initialization

To compare the proposed data partition-based initialization (Method 2) to the random initialization (Method 1) in the Appendix B, we generated 30 genotype datasets on 10 single nucleotide polymorphisms (SNPs), each dataset, containing N_0 controls and N_1 cases, was obtained by the following two steps: In the step 1, we used the software MS [5] to simulate $2(N_0 + N_1)$ haplotypes with a mutation rate of 2. We randomly chose m_r of these haplotypes and labeled them as risk haplotypes. To save the space, we considered only $N_0 + N_1 = 5000$ and $m_r = 10$. The results for other values of $N_0 + N_1$ and m_r were similar. In the step 2, the disease states of the above genotypes were simulated from the multiplicative inheritance model with $q_0 = 0.1$ and $\lambda = 3$. Note that the number of genotypes depends on the mutation rate and was varying across 30 datasets.

The comparison was based on the log-likelihood, the run time, estimated bias and classification error rate (CER). The estimated bias can be calculated by sum all the absolute values of the differences between $\hat{\theta}$ and the true θ. Note that genotypes in each dataset could be divided into three (true) groups, say $\mathbf{G}_\nu$, $\nu = 0, 1, 2$ as we knew the number of risk haplotypes which each genotype contained in the simulation. We pretended that we did not know which haployes were risk (therefore, we did not know the group memberships of these genotypes). We then inferred their memberships by fitting a three-component binomial mixture model to each of

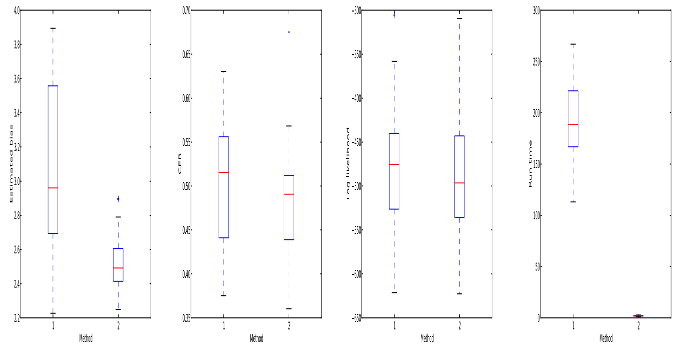

Figure 1. Performance of two initialization methods. Methods 1 and 2 denote the random initialization and the data partition-based methods. From the left to the right, the panels show the box-whisker plots of the estimated biases in estimating θ, the CERs, the attained log-likelihoods, and the time-costs for Methods 1 and 2 respectively.

30 datasets. By using the estimated posterior probabilities, $\tau_{\nu j}, \nu = 0, 1, 2$, of group memberships derived from the EM algorithm, we assigned the j-th genotype to the group $\hat{\mathbf{G}}_\nu$, $\nu = 0, 1, 2$ if $\tau_{\nu j} = \max_t \tau_{tj}$. Here, we labeled three estimated groups according to the ordered penetrances $\hat{q}_0 \leq \hat{q}_1 \leq \hat{q}_2$. This is a computationally simple approach to solving the so-called label switching problem in finite mixture models [13]. Our experience indicates it is effective for estimating our binomial mixture model. More advanced but time-consuming approach can be found in [17]. The accuracy of three estimated groups was evaluated by the CER defined as

$$\text{CER} = \sum_\nu \left(1 - \frac{|\mathbf{G}_\nu \cap \hat{\mathbf{G}}_\nu|}{|\mathbf{G}_\nu|}\right),$$

where we counted the total number of misclassified genotypes divided by the total number of the genotypes. The results were summarized in Figure 1 in terms of the box-whisker plots of the estimated biases, the CERs, likelihood values, and time-costs over 30 datasets for Methods 1 and 2 respectively. The result shows that Method 2 substantially outperformed Method 1. Therefore, we decided to initialize the EM algorithm by use of Method 2 in the remaining simulations as well as the real data analysis below.

3.2 Performance of the proposed two-stage method

Note that the proposed two-stage method is based on the prospective likelihood model although real data were obtained from retrospective studies. By the simulations below, we addressed whether the proposed method could outperform the multiple-testing procedure of Zhu et al. [27] in both prospective (i.e., cohort) and retrospective (i.e., case-control) studies.

Setting 1 (cohort design): We generated 30 datasets, each with N_1 case-genotypes and N_0 control-genotypes. They were obtained by the following steps. In the first two steps, we adopted the same approach for generating $N_0 + N_1$ genotypes which contained m_r risk haplotypes as we did before. In the third step, we simulated the disease status of each genotype by sampling from a Bernoulli distribution. The Bernoulli distribution took q_0, or $\lambda_1 q_0$, or λq_0 as a success probability according to whether the genotype contained zero, one or two risk haplotypes, where the relative risk measure λ_1 is specified as follows. For the recessive inheritance mode, $\lambda_1 = 1$. For the multiplicative inheritance mode, $\lambda_1 = \sqrt{\lambda}$. For the dominant inheritance mode, $\lambda_1 = \lambda$. We coded the inheritance modes by IM $= 1, 2, 3$ respectively for the multiplicative, the dominant, and the recessive. Note that the values of (N_0, N_1) may vary across different datasets. We considered various combinations of $(N_0 + N_1, m_r, \text{IM}, q_0, \lambda)$, where $N_0 + N_1 = 3000, 5000$, $m_r = 5, 10, 20$, $\text{IM} = 1, 2, 3$, $q_0 = 0.1$, $\lambda = 1, 1.4, 1.8, 2.2, 2.6, 3, 3.4$, and 3.8 respectively.

For each scenario, we applied both the proposed method and the multiple testing method to 30 datasets and calculated their AVSS values respectively. For each of the three inheritance modes, we plotted the means of these AVSS values over 30 datasets against λ. The results displayed in Figure 2 show that on the cohort data, the proposed two stage method performed substantially better than the multiple testing method in all the scenarios defined above. The improvement was achieved by using model-based genotype clustering. This is not surprising, because Yeung et al. [21] has already showed that the model-based clustering is often superior over non-model based clustering.

Setting 2 (case-control design): We generated 30 datasets, each of which were simulated by the following two steps. In Step 1, to generate N_1 case-genotypes, we first drew $2N_1$ haplotypes by using the software MS with mutation rate of 2, of which m_r haplotypes were labeled as risk haplotypes. We then randomly paired these haplotypes to form N_1 case-genotypes. Let G_j, $1 \leq j \leq J$ be all the different genotypes contained in the N_1 cases and $r_{1j}, 1 \leq j \leq J$ be the retrospective frequencies. These case-genotypes formed three groups according to the number of risk haplotypes which each genotype contained: Each genotype in Groups 0, 1 and 2 contained two non-risk haplotypes, only one risk-haplotype, and two risk haplotypes respectively. In Step 2, we generated N_0 control-genotypes, which also had genotypes G_j, $1 \leq j \leq J$ but with population retrospective frequencies $q_{0j}, 1 \leq j \leq J$. We first let $q_{0j}, 1 \leq j \leq J$ depend on the pre-specified constant d by

$$q_{0j} = \begin{cases} r_{1j}(1 - d/r_{1g_2}), & G_j \text{ belongs to Group 2} \\ r_{1j}(1 - 0.5d/r_{1g_1}), & G_j \text{ belongs to Group 1} \\ r_{1j}(1 + 1.5d/r_{1g_0}), & G_j \text{ belongs to Group 0} \end{cases}$$

where $r_{1g_k} = \sum_{G_j \in \text{ Group}_k} r_{1j}$ for $k = 0, 1, 2$, and d is a parameter to reflect the effects of risk haplotypes on genotype frequencies. We simulated N_0 control-genotype counts from the multinomial model $\text{MN}(N_0, (q_{01}, ..., q_{0J})^T)$ and cal-

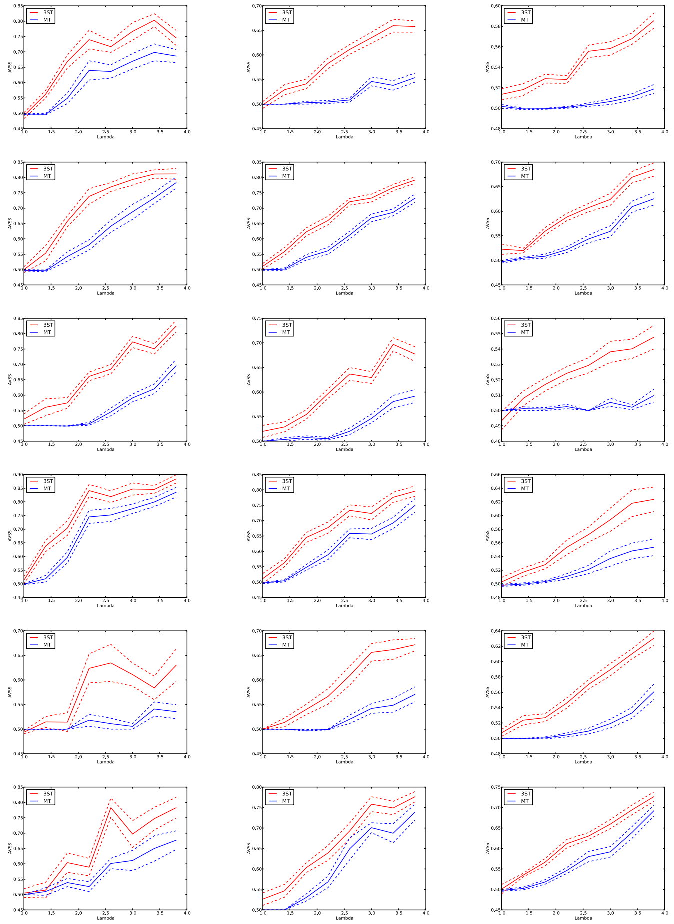

Figure 2. Performances of the proposed two-stage method with Bonferroni adjustment and the multiple testing method on the cohort-design data with multiplicative or dominant or recessive inheritance modes. In these plots, the red and the blue solid curves show means of the AVSS values (i.e., the values of $(\text{specificity and sensitivity})/2$) *over 30 datasets are plotted against the values of* λ *for the proposed method and the multiple testing method respectively. The two red dash curves are one standard deviation up and down from the red mean curves. Similarly, the two blue dash curves are one standard deviation up and down for blue mean curves. The plots in the columns from the left to the right are for the cases where there were* 5, 10, *and* 20 *risk haplotypes in the underlying haplotypes. The top two rows, the middle two rows and the bottom two rows are the results for* $(N_0, N_1) = (2000, 1000)$ *and* $(3000, 2000)$ *under the multiplicative, the dominant and the recessive inheritance modes respectively. (Color figure online)*

culated the corresponding retrospective frequencies $r_{0j}, 1 \leq j \leq J$. We considered the cases where $d = 0, 0.05, 0.1, 0.1, 0.15, 0.2, 0.25, 0.3,$ and 0.35 respectively.

For each dataset, the cumulative frequencies of Groups 0, 1, and 2 in controls are $r_{g_0} + 1.5d$, $r_{g_1} - 0.5d$, and $r_{g_2} - d$ respectively, whereas the corresponding frequencies in cases are r_{g_0}, r_{g_1} and r_{g_2} respectively. It can be proved that the odds ratios of Groups 1 and 2 to Group 0 are increasing in the value of d.

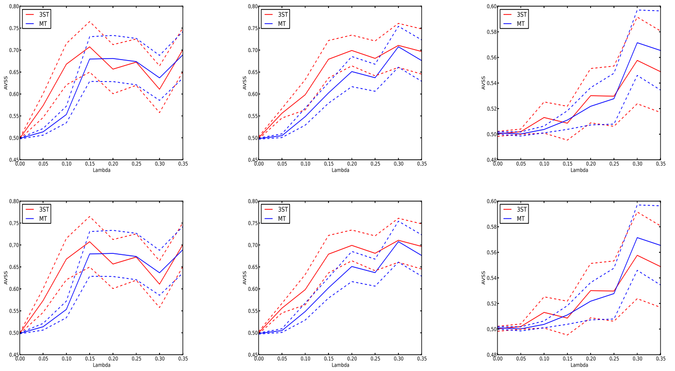

Figure 3. Performances of the proposed two-stage method with Bonferroni adjustment and the multiple testing method on the case-control data. The plots in the columns from the left to the right are for the scenarios, where the underlying number of risk haplotypes $m_r = 5, 10,$ *and* 20. *The top row stands for the cases, where* $(N_0, N_1) = (2000, 1000)$, *while the bottom row stands for the cases, where* $(N_0, N_1) = (3000, 2000)$. *In these plots, the red and the blue solid curves show mean curves of the AVSS values over 30 datasets as functions of* $d = 0, 0.05, 0.1, 0.1, 0.15, 0.2, 0.25, 0.3,$ *and* 0.35 *for the proposed method and the multiple testing method respectively. The dash curves are one standard error up or down from the mean curves. (Color figure online)*

We applied the proposed two-stage method and the multiple testing method to these case-control data. The mean curves of the AVSS values with one standard error up and down were plotted against the d values in Figure 3. The results again demonstrate that the proposed two-stage method can be more powerful than the multiple testing method in detecting risk haplotypes. However, the AVSS gain was decreasing in the number of risk haplotypes, m_r, as well as the underlying odds ratios in Groups 1 and 2. In particular, the AVSS gain can be negative when there were many risk-haplotypes presented in the data. This is due to the effect of unbalanced case and control sample sizes in the finite sample size setting, because our model in Stage 1 is a prospective model.

4. REAL DATA ANALYSIS

We applied the proposed two-stage procedure to the GWAS genotype datasets on coronary artery disease (CAD) and hypertension (HT) obtained by Affymetrix 500K SNP chips in the WTCCC study [23]. The data were downloaded from the European Genotype Archive (EGA) with formal data access permission of the WTCCC Data Access Committee. Each dataset contained 2000 unrelated cases as well as 3000 unrelated controls. The controls came from two sources: 1500 from the 1958 British Birth Cohort (58C) and 1500 from the three National UK Blood Services (NBS). There were about 500600 SNPs across the human genome, which are genotyped. We first pre-processed the data by excluding the SNPs which meet one of the following criteria: (1) the p-value of Fisher test for Hardy-Weinberg equilibrium is less than 10^{-8} in controls; (2) the p-value of the chi-square test between 58C and NBS is less than 10^{-8}; (3) the minor allele frequency is less than 1%; (4) the calling score is less than 95%. After the exclusion, around 4897746 SNPs remained for the analysis. To reduce the dimension of the genotypes, we segmented the genome into regions of 8 SNPs according to their positions on the chromosomes, obtaining 61218 regions and the corresponding genotype datasets $\mathbf{G}_k, k = 1, 2, ..., 61218$. Note that the long region will dilute the effects of risk SNPs and can result in many rare genotypes, whereas the short region will miss interactions between SNPs. The region length of 8 was chosen to achieve a compromise between the above aspects by using a pilot study. Also note that as we excluded the SNPs with bad callings, the numbers of cases and controls are varying across the different regions.

Note that $\{\mathbf{G}_k : k = 1, ..., 61218\}$ contained 1983537 genotypes in total for the CAD data and 2097111 genotypes in total for the HT data respectively. The proposed procedure includes two stages. In Stage 1, we obtained the estimated risk genotypes, while in Stage 2, we further inferred haplotype pairs from the estimated risk genotypes. In Stage 1, we first fitted a three-component binomial mixture model to each $\mathbf{G}_k$ and then thresholded the genotypes based on the smallest penetrance in the three components. The thresholding would involve 1983537 tests for the CAD data and 2097111 tests for the HT data. So in equation (1), we set $\varepsilon = 0.05/1983537 = 2.52 \times 10^{-8}$ for the CAD data and $\varepsilon = 0.05/2097111 = 2.38 \times 10^{-8}$ for the HT data. In Stage 2, we employed the PHASE to infer the haplotypes from the risk genotypes derived from the previous stage. This gave rise to 201528 potential risk haplotypes out of 1448586 in CAD data and 213578 potential risk haplotypes out of 1463838 in HT data. We further conducted the OR thresholding for these haplotypes. There would involve 201528 tests in the CAD case and 213578 tests in the HT case. By using the Bonferroni adjustment, we set the corresponding individual test level at $0.05/201528 = 2.48 \times 10^{-7}$ and $0.05/213578 = 2.34 \times 10^{-7}$ for the CAD and the HT respectively. These individual test levels were then used to determine the tuning constant c_1 in equation (4). This yielded $c_1 \approx 5$. After performing the proposed two-stage method on the datasets, we obtained the estimated risk and nonrisk haplotype sets, $\hat{\mathbf{H}}_r$ and $\hat{\mathbf{H}}_{\bar{r}}$), for the CAD and the HT respectively.

Finally, we carried out a genomic control on the above results by taking advantage of the fact that there were two sub-populations in controls. The genomic control can eliminate these false haplotypes generated by the PHASE and population substructures from the selected list of risk haplotypes. In the genomic control, we run the chi-square tests on the association of two control sub-populations with each estimated risk haplotype. We eliminated these estimated risk haplotypes with p-values for the above chi-square tests less than $< 30\%$. Here, 30% was chosen by the simulations, aiming to filter out false risk haplotypes. The details are omitted but can be obtained from the authors.

The genomic control gave the final risk-haplotype set as displayed in Tables 1, 2, 3, and 4 below. In the tables, each haplotype has been assigned to a physically closest gene on the basis of the information provided the GWAS catalog and the genetic information from the British 1958 Birth cohort. See [24] and the web page at http://www2.le.ac.uk/projects/birthcohort/1958bc. In the CAD case, we did rediscover the CAD risk genes TNIK in chromosome 3, CDKN2B in chromosome 9, BTG1 in chromosome 12, and A2BP1 in chromosome 16, which were found by the previous study [24]. Among these genes, Zhu et al. [27] identified only CDKN2B. In the HT case, we also identified a number of variants which were potentially associated with hypertension. Compared to the multiple testing approach of Zhu et al. [27], where 7 CAD-associated genes and 2 HT-associated genes were declared, our approach was much powerful by finding more than 80 CAD-associated haplotypes and 11 HT-associated haplotypes. However, we were not able to confirm other existing discoveries in the literature [24]. A possible reason is that we set a very stringent level for the odds ratio thresholding based on the Bonferroni adjustment for multiple testing. It is well-known that the Bonferroni adjustment is very conservative.

5. DISCUSSION AND CONCLUSION

We are currently at an era of extraordinary growth in the data describing human genetic variation and its correlation with complex traits. The recent development of biotechnologies allows an international consortium of geneticists to revolutionize genetic research through large scale genome wide association studies (GWAS). Although these studies have identified hundreds of loci at very stringent levels of statistical significance across many different human traits, these loci are only able to explain a small fraction of the population risk. To address the issue, new models and new hypotheses have been proposed, which pose challenges to conventional statistics underlying much of our genetic analysis. For example, GWAS analyses are most commonly performed by testing the association of individual variants with the disease, ignoring the potential interactions between the variants. It is believed that the region or gene-based analysis is more powerful in capturing the collective activity of sets of variants by testing the association of the group instead of each component individually with the disease.

In this paper, we have adopted the region-based strategy that segments the genome into 61218 regions with around 8 SNPs each. For each region, a list of distinct genotypes with

Table 1. The predicted risk haplotypes for CAD by use of the WTCCC data. In the table, the p-values were derived from the chi-squared test of the frequencies of H_i against the collapsed frequencies of the estimated non-risk haplotypes

Chr	Region	SNP range	Haplotype	$\hat{P}(H_i\|case)$	$\hat{P}(H_i\|control)$	OR	p-Value	Gene
1	$17921479 - 17955334$	$rs11203219 - rs638425$	$AATGCCGC$	0.04602	0.01388	3.05038	4.1×10^{-12}	ACTL8
1	$75974016 - 76018681$	$rs3806162 - rs5745391$	$TCTATCAA$	0.05105	0.01954	3.18049	1.2×10^{-12}	MSH4
2	$49934439 - 50000082$	$rs6736617 - rs17039375$	$CCAAAGGT$	0.02347	0.00757	3.08898	6.6×10^{-10}	NRXN1
2	$81387425 - 81525659$	$rs4401229 - rs2862499$	$TTGCTCCA$	0.0451	0.02468	2.54951	1.8×10^{-12}	LOC442021
2	$222486954 - 222527591$	$rs16863087 - rs2392937$	$CCAAACGG$	0.04059	0.02497	2.09348	4.3×10^{-08}	LOC402120
2	$230201571 - 230228527$	$rs6755403 - rs13391903$	$AGTTTGCC$	0.1132	0.04164	2.78377	2.3×10^{-08}	DNER
2	$239420300 - 239491966$	$rs4545955 - rs13008279$	$TTCCAGGA$	0.05558	0.02584	2.17494	1.3×10^{-12}	FLJ43879
2	$241821720 - 241873661$	$rs4675991 - rs935262$	$CGGGGTTT$	0.03735	0.01659	2.32538	1.4×10^{-10}	PPP1R7
3	$4927181 - 5001898$	$rs17041733 - rs11925620$	$CCTCCTCC$	0.04287	0.01795	2.16999	1.2×10^{-07}	BHLHB2
3	$14422977 - 14471151$	$rs4684216 - rs9834629$	$GATGATGC$	0.01815	0.00509	3.63785	1.7×10^{-09}	SLC6A6
3	$60586653 - 60641652$	$rs7432576 - rs1716739$	$CTATAAGC$	0.15989	0.11374	1.55681	9.4×10^{-12}	FHIT
3	$63365648 - 63390235$	$rs17068494 - rs1403700$	$TCCTTCGG$	0.08979	0.04741	2.04072	7.1×10^{-09}	SYNPR
3	$67509601 - 67525645$	$rs9867659 - rs17046411$	$ACGATGTT$	0.05192	0.03019	1.95683	5.1×10^{-09}	SUCLG2
3	$103285842 - 103325614$	$rs7623627 - rs9844712$	$GTCCCTAT$	0.02744	0.00999	3.15138	1.6×10^{-09}	NFKBIZ
3	$106353367 - 106411138$	$rs16850901 - rs9846852$	$TATCGAGA$	0.02931	0.0065	4.87306	7.5×10^{-18}	ALCAM
3	$144925558 - 144993828$	$rs4330252 - rs12233446$	$TGGGATAC$	0.02976	0.00733	5.71824	1.8×10^{-16}	SLC9A9
3	$145364476 - 145471873$	$rs9854202 - rs3925560$	$AACGGACT$	0.37409	0.29638	2.25725	5.5×10^{-34}	C3orf58
3	$172422863 - 172457251$	$rs954749 - rs16856054$	$TTCTTACT$	0.12948	0.08707	1.50219	2.2×10^{-08}	TNIK
3	$192463499 - 192526004$	$rs7644510 - rs293871$	$GACGCGTA$	0.04375	0.01075	3.69505	1.3×10^{-18}	UTS2D
3	$197256495 - 197339533$	$rs6583286 - rs9834962$	$TAGACTTA$	0.0498	0.02364	2.27577	2.7×10^{-10}	TFRC
4	$3636361 - 3700212$	$rs10025237 - rs16844722$	$GGGGAGGG$	0.22491	0.15492	1.65607	1.9×10^{-07}	FLJ35424
5	$120487082 - 120547238$	$rs11956204 - rs17514347$	$ATTGGGAG$	0.02739	0.00735	3.8359	1.5×10^{-13}	LOC728682
5	$166764561 - 166801933$	$rs6863935 - rs7724862$	$CTATGTGT$	0.09145	0.05448	1.69398	8.7×10^{-09}	ODZ2
7	$4779368 - 4930112$	$rs2942566 - rs4320451$	$CGGGTCAT$	0.10433	0.06243	1.66428	5.5×10^{-10}	RBAK
7	$10052046 - 10079446$	$rs10225194 - rs11768931$	$GGTTCGCT$	0.04951	0.0245	2.64149	9.4×10^{-15}	LOC340268
7	$34178282 - 34260002$	$rs17169771 - rs16878925$	$AGGTTGCG$	0.05229	0.02631	2.71386	3.3×10^{-13}	AAA1
7	$42931717 - 42940671$	$rs2024125 - rs2330742$	$AGTGTAGA$	0.09745	0.0513	1.90132	2.0×10^{-10}	HECW1
7	$153564509 - 153621369$	$rs869490 - rs6953905$	$TCGTATCG$	0.0667	0.03524	1.93779	6.6×10^{-11}	LOC653748
8	$5482876 - 5498858$	$rs2189889 - rs4875607$	$CGGACCGA$	0.07873	0.0533	1.64615	2.4×10^{-08}	LOC648237
8	$17486464 - 17509327$	$rs2705093 - rs2588121$	$CCTGCGAG$	0.05925	0.02338	2.67404	1.6×10^{-15}	PDGFRL
8	$38345434 - 38449100$	$rs16887343 - rs12677355$	$ACGTACCT$	0.09472	0.05661	1.82381	7.0×10^{-13}	WHSC1L1
8	$104190450 - 104202402$	$rs2515173 - rs3019159$	$GGCCATCT$	0.14195	0.08768	1.62006	1.5×10^{-08}	BAALC

their frequencies in cases and controls have been worked out. The problem facing us is of the sparse distribution of these genotypes. To circumvent it, people often first infer haplotypes from the genotypes and then cluster the haplotypes into a number of groups. The association analysis is conducted on the basis of the inferred groups, for example,

Table 2. The continuation of Table 1

Chr	Region	SNP range	Haplotype	$\hat{P}(H_i \mid case)$	$\hat{P}(H_i \mid control)$	OR	p-Value	Gene
9	22088619 − 22120515	$rs2891168 - rs10965245$	$GGTGCCAG$	0.34939	0.29298	1.52609	1.0×10^{-07}	CDKN2B
9	74180343 − 74241329	$rs10114124 - rs17081046$	$GTATTTAT$	0.21608	0.13046	1.61055	1.2×10^{-07}	RORB
9	114777214 − 114805868	$rs1322060 - rs10121268$	$GAGCCTAA$	0.09498	0.06007	1.56664	2.3×10^{-08}	TNFSF8
9	119506057 − 119537035	$rs2191675 - rs10984648$	$GTTGGCTA$	0.08762	0.03361	2.41642	3.0×10^{-16}	CDK5RAP2
10	11879196 − 11924252	$rs6602535 - rs11257355$	$TCTGCCGG$	0.1694	0.12811	1.41273	6.4×10^{-08}	C10orf47
10	64409674 − 64442476	$rs1509952 - rs2842286$	$TTTCTTAC$	0.02299	0.0073	4.03039	1.6×10^{-09}	NRBF2
11	8165969 − 8200374	$rs4758310 - rs11041816$	$ATAATGGG$	0.36298	0.3164	1.3306	1.1×10^{-08}	LOC644497
11	21323965 − 21363331	$rs17233214 - rs1945444$	$GGTAACAT$	0.08147	0.04232	1.98043	8.6×10^{-12}	NELL1
11	69213458 − 69295251	$rs1192923 - rs3168175$	$TCGTGGCA$ $TTGTGGCA$	0.10225 0.05213	0.05587 0.02803	1.98038 2.01202	8.9×10^{-14} 5.6×10^{-09}	FGF4
11	83230307 − 83256927	$rs1878266 - rs1878264$	$TATATTCA$	0.03571	0.01807	2.11905	2.5×10^{-07}	CCDC90B
12	90721177 − 90758721	$rs10745571 - rs17193868$	$GGGCTATA$	0.0351	0.00949	3.88035	1.7×10^{-16}	BTG1
12	114038450 − 114074493	$rs1828384 - rs35346$	$TGTACCCT$	0.03245	0.01341	2.52817	2.3×10^{-07}	TBX3
12	127146384 − 127182360	$rs10847535 - rs10773498$	$TTGTCGCG$	0.10562	0.07049	1.50842	1.3×10^{-07}	TMEM132C
12	129086441 − 129129809	$rs713149 - rs1027557$	$AAAGCGGT$	0.18839	0.11206	1.74867	4.4×10^{-14}	FLJ31485
13	26845975 − 26875430	$rs11616513 - rs17085553$	$TACGCACA$	0.04431	0.02025	2.30656	7.1×10^{-10}	MTIF3
13	31414174 − 31438047	$rs17076954 - rs169410$	$CCTCCCGT$	0.30306	0.29469	2.6188	6.9×10^{-08}	LOC196549
13	48154476 − 48209065	$rs7330127 - rs9562843$	$ACGATAGA$	0.02762	0.0048	5.63922	2.7×10^{-10}	RCBTB2
14	25140850 − 25159405	$rs8020556 - rs1951062$	$AGTACATA$ $AGTAAACT$ $GCTACATA$	0.24934 0.09084 0.04608	0.2259 0.02999 0.01682	1.41488 3.87615 3.50368	3.5×10^{-08} 1.0×10^{-41} 3.4×10^{-22}	LOC401767
14	32591680 − 32606647	$rs12883961 - rs10140504$	$CATGGGAG$	0.03736	0.01879	2.21665	1.1×10^{-08}	NPAS3
14	65343491 − 65401760	$rs3924222 - rs12896836$	$TATAACTC$	0.0462	0.01904	2.55404	5.2×10^{-14}	FUT8
15	20592297 − 20610835	$rs4778334 - rs1991922$	$TAGCCCAT$	0.04494	0.01488	2.75061	1.1×10^{-12}	NIPA1
15	20624103 − 21246055	$rs7166056 - rs8024346$	$GTGACGTG$	0.08093	0.04109	2.10848	2.4×10^{-13}	NIPA1
15	21610088 − 21670901	$rs824163 - rs7181211$	$TTTTCAAC$	0.22034	0.15435	1.43864	4.9×10^{-09}	MAGEL2
15	37962389 − 38014169	$rs11633436 - rs534757$	$TTACAACC$	0.07798	0.03763	1.99235	2.7×10^{-11}	GPR176
15	64637416 − 64669062	$rs1030986 - rs4776800$	$CACGTCGT$	0.04575	0.01594	2.65924	2.2×10^{-09}	LCTL
15	79193543 − 79223619	$rs1317059 - rs6495541$	$CTCGGACC$	0.02813	0.00459	6.34974	2.2×10^{-15}	C15orf26
15	90365510 − 90400043	$rs12906289 - rs992838$	$ACGTAAGG$	0.07777	0.02342	3.50153	1.1×10^{-26}	SLCO3A1
15	91435452 − 91473401	$rs4778099 - rs17526830$	$GATCCCTA$	0.07536	0.04084	1.94917	1.7×10^{-09}	RGMA

by using multiple Z-tests [27]. There is a drawback of the above approach: The in-silico reconstruction of haplotypes can generate a proportion of false haplotypes which may hamper the finding of rare but true haplotypes. We have proposed an alternative two-stage approach to the association analysis with GWAS data. Our major contribution is to develop a method for co-classifying genotypes in terms of their penetrances to the disease. In Stage 1, we cluster the genotypes through a finite mixture model, followed by estimating the risk genotypes. In Stage 2, we infer the risk haplotypes from the estimated risk genotypes by using the software PHASE and the odds ratio thresholding. We have

Table 3. The continuation of Table 2

16	6155489 − 6181184	$rs11642397 - rs1946127$	*TTGGGTTG*	0.02433	0.00883	2.92587	1.7×10^{-07}	A2BP1
16	46937666 − 47050362	$rs11076564 - rs8054696$	*AACGGGCC* *TGAAGGCT*	0.18717 0.04224	0.15302 0.02781	1.62027 2.01195	1.1×10^{-07} 2.3×10^{-07}	LONP2
16	51239337 − 51264345	$rs3112587 - rs4386133$	*CCTATGAG*	0.07702	0.0442	1.68656	7.3×10^{-08}	LOC643714
16	55207138 − 55253047	$rs8055724 - rs12447986$	*TTCTCCTC*	0.03044	0.01113	2.65805	9.0×10^{-09}	MT1L
17	73602775 − 73670122	$rs16970811 - rs9909570$	*CCCACTAG*	0.02022	0.00446	4.82821	3.1×10^{-13}	TNRC6C
17	74629176 − 74682195	$rs2612793 - rs8072667$	*CGAGGTTG*	0.06276	0.03471	1.95026	6.7×10^{-09}	FLJ21865
18	8212591 − 8279839	$rs10468776 - rs11876033$	*GGGACAAG*	0.02689	0.00982	2.86846	1.7×10^{-10}	PTPRM
18	8772147 − 8782163	$rs12606001 - rs8084401$	*TCAGTGAC*	0.09539	0.03649	2.66938	1.3×10^{-17}	KIAA0802
18	60647495 − 60688045	$rs1595904 - rs17678507$	*CAGCGTGC*	0.08119	0.04205	2.1482	6.5×10^{-16}	C18orf20
19	50064169 − 50153836	$rs17561351 - rs204907$	*AGGCAGAA*	0.05937	0.02583	2.35486	5.1×10^{-14}	PVRL2
19	52946204 − 53026777	$rs10402957 - rs4427918$	*CATTCAGC*	0.0741	0.04321	1.87681	1.7×10^{-11}	GLTSCR2
19	59113663 − 59296006	$rs7257613 - rs3760698$	*CCGGCCGC* *CCGGCCAC*	0.06977 0.12473	0.0159 0.08441	5.01246 1.69429	2.7×10^{-43} 6.7×10^{-13}	CACNG7
20	5265473 − 5327486	$rs6085111 - rs6085143$	*ACCAATCC*	0.04815	0.02744	1.83971	1.3×10^{-07}	FLJ33544
20	42465269 − 42498442	$rs3181206 - rs6017342$	*GGCTTCCA*	0.12685	0.06245	2.08814	3.0×10^{-14}	HNF4A
20	44639977 − 44681497	$rs376438 - rs847096$	*AAGTCTGC*	0.09805	0.04784	1.90457	8.8×10^{-12}	SLC13A3
20	49937544 − 50006641	$rs6067996 - rs6021570$	*ATTGGACA*	0.03133	0.01165	2.82133	2.6×10^{-11}	SALL4
20	51762764 − 51798874	$rs4811452 - rs4811457$	*GATGTTCA*	0.05611	0.03099	1.87441	1.7×10^{-08}	ZNF217
20	57707915 − 57741702	$rs12481511 - rs16984986$	*TGTACCAG*	0.0773	0.0427	1.95199	1.2×10^{-07}	PHACTR3
21	2015127 − 13517135	$rs2847443 - SNP_A$	*TACAAGAT*	0.10999	0.09446	1.65501	2.4×10^{-08}	TPTE
22	16871076 − 16895136	$rs8142200 - rs975826$	*TCGGGAGG*	0.03219	0.00253	10.88401	1.8×10^{-19}	LOC729269
22	31354524 − 31372260	$rs8139704 - rs5749480$	*CGCTAGGG*	0.02584	0.00524	5.07641	3.4×10^{-16}	SYN3
22	35324014 − 35335429	$rs7410412 - rs12160203$	*TTTCAAGG*	0.17403	0.10746	1.67423	1.3×10^{-10}	CACNG2

proposed a novel data-partition-based initialization for the associated EM algorithm.

We have examined the performance of the proposed procedure by simulations and applications to the CAD and HT data generated from the WTCCC. Compared to the standard multiple Z-testing method, the proposed procedure has been shown to be more powerful in terms of sensitivity and specificity for detecting the true risk haplotypes. In the real data analysis, we have rediscovered some existing risk gene and haplotypes and identifying many more risk haplotypes than did the multiple Z-test based approach. This is not surprising as the simulations have already demonstrated that the model-based clustering can perform better than the multiple Z-test. The Bonferroni adjustment for multiple testing has been applied when multiple tests or thresholding are involved. We note that the results may be further improved if we use advanced multiple testing adjustment methods in Stage 2, although this may not be possible for Stage 1 as the computation is too time-consuming to run on a PC. For example, in Stage 2, we can apply Hochberg's procedure to adjusting and thresholding the individual p-values in two steps as follows [4].

Step 1: We calculate the p-value for each haplotype $h_k \in \mathbf{H}_a$. Note that under the null hypothesis in which the underlying odds ratio is one, the distribution of the estimated odds-ratio OR_k is asymptotically Normal distributed as stated in the equation (2). Then, the p-value can be approximated by

$$p_k = 1 - \Phi\left(\log(\mathrm{OR}_k^{(0)}/\phi(n_{0k}, n_{1k}, n_{0\bar{k}}, n_{1\bar{k}}))\right),$$

where $\Phi(\cdot)$ is the standard normal distribution function and $\phi(n_{0k}, n_{1k}, n_{0\bar{k}}, n_{1\bar{k}})$ is defined in the equation (3).

Step 2: We calculate the adjusted p-values by ordering

Table 4. The predicted risk haplotypes of hypertension by use of WTCCC data. In the table, the p-values were derived from the chi-squared test of the frequencies of H_i against the collapsed frequencies of the estimated non-risk haplotypes

Chr	Region	SNP range	Haplotype	$\hat{P}(H_i\|case)$	$\hat{P}(H_i\|control)$	OR	p-Value	Gene
1	$236986859 - 237020204$	$rs12137158 - rs16840310$	$ATTTAGGG$	0.08733	0.05437	1.69625	3.4×10^{-10}	GREM2
4	$3700382 - 3734797$	$rs177772 - rs12641338$	$TACCGATT$	0.12978	0.08988	1.59997	7.7×10^{-12}	FLJ35424
4	$170032303 - 170061525$	$rs6822949 - rs17614553$	$GAACGGAA$	0.0425	0.01579	2.86663	4.8×10^{-10}	PALLD
6	$152700181 - 152736079$	$rs7747166 - rs7776399$	$CGGCTCCC$ $CGGGTCCT$	0.52639 0.04238	0.49931 0.03768	3.36065 3.58962	2.7×10^{-23} 5.7×10^{-14}	SYNE1
11	$69213458 - 69295251$	$rs1192923 - rs3168175$	$TTGTGGCA$	0.05532	0.02803	2.12665	3.4×10^{-10}	FGF4
12	$116500495 - 116514298$	$rs10850852 - rs1400593$	$CTCTCTTC$	0.28748	0.26232	2.46528	5.2×10^{-17}	NOS1
14	$21674996 - 21704333$	$rs12050442 - rs1894369$	$GGGGTTAC$	0.03075	0.00968	3.28277	1.8×10^{-11}	TRA@
14	$25140850 - 25159405$	$rs8020556 - rs1951062$	$AGTAAACT$	0.08475	0.02999	2.94949	6.6×10^{-27}	LOC401767
14	$36411583 - 36421982$	$rs10872897 - rs2564848$	$TACCTCCC$ $ATCCACTT$	0.02712 0.02299	0.01101 0.00637	2.63669 3.84732	1.4×10^{-08} 1.3×10^{-11}	SLC25A21
14	$36969639 - 37032855$	$rs10132119 - rs17106785$	$CTATGACA$	0.01914	0.00402	5.57575	6.1×10^{-10}	MIPOL1
19	$17595848 - 17649789$	$rs10419511 - rs7252308$	$TTGGTATG$	0.04536	0.01971	2.16516	1.7×10^{-10}	UNC13A

the p-values as $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m_a)}$, where m_a is the size of $\mathbf{H}_a$. The adjusted p-values are then defined by

$$\tilde{p}_{(j)} = \min_{j \leq k \leq m_a} \min(m_a p_{(k)}/k, 1).$$

We assign the corresponding haplotype to the risk group if $\tilde{p}_{(j)} \leq 0.05$.

We have applied the above modified procedure to both simulated and real data, obtaining improved simulation results displayed in Figures 4–5 as well as the additional risk-haplotypes identified from the real data analysis in Table 5.

APPENDIX A. PHASE

PHASE is a Bayesian haplotype reconstruction method developed by Stephens et al. [18] to tackle the problem of statistically inferring haplotypes from unphased genotype data for a sample of unrelated individuals from a population. Based on the so-called coalescent model, it treats the unknown haplotypes as random quantities and combine prior information on haplotypes with the data likelihood to calculate the posterior distribution of the unobserved haplotypes (or haplotype frequencies) given the observed genotype data. The haplotypes themselves can then be reconstructed from this posterior distribution: for example, by choosing the most likely haplotype reconstruction for each individual.

APPENDIX B. EM ALGORITHM

The EM algorithm consists of two steps.

E-Step: Given the current estimator $\theta^{(t)}$ and the data, the conditional expectation of the complete log-likelihood can be calculated by

$$\begin{aligned} Q(\theta, \theta^{(t)}) &= E\left[l(\theta|\mathbf{N}, \mathbf{I})|\mathbf{N}, \theta^{(t)}\}\right] \\ &= \sum_{j=1}^{J} \sum_{\nu=0}^{2} \tau_{\nu j}^{(t)} \log\left[\pi_\nu^{(t)} f((N_{0j}, N_{1j})^T | q_\nu^{(t)})\right], \end{aligned}$$

where the expectation is taken with respect to the distribution of $\mathbf{I}$ and the estimated posterior probability of the j-th genotype being in the group ν, $\tau_{\nu j}^{(t)}$ admits

$$\begin{aligned} \tau_{\nu j}^{(t)} &= P(I_{\nu j} = 1|(N_{0j}, N_{1j})^T, \theta^{(t)}) \\ &= \frac{\pi_\nu^{(t)}\ f((N_{0j}, N_{1j})^T | q_\nu^{(t)})}{\sum_{\nu=0}^{2} \pi_\nu^{(t)} f((N_{0j}, N_{1j})^T | q_\nu^{(t)})}. \end{aligned}$$

M-Step: We update the current estimate $\theta^{(t)}$ by maximizing Q with respect to θ. This is equivalent to solving the following equations

$$\frac{\partial Q}{\partial \pi_\nu} = 0, \quad \frac{\partial Q}{\partial q_\nu} = 0, \ \nu = 0, 1, 2,$$

subject to $\pi_0 + \pi_1 + \pi_2 = 1$. For $\nu = 0, 1, 2$, we obtain the updated estimate $\theta^{(t+1)}$ via

$$\pi_\nu^{(t+1)} = \sum_{j=1}^{J} \tau_{\nu j}^{(t)}/J, \quad q_\nu^{(t+1)} = \frac{\sum_{j=1}^{J} \tau_{\nu j}^{(t)} N_{1j}}{\sum_{i=1}^{J} \tau_{\nu j}^{(t)} (N_{0j} + N_{1j})}.$$

The existing EM theory suggests that the value of the log-likelihood function at the updated estimate is not decreasing in the sense that $l(\theta^{(t+1)}|\mathbf{N}) \geq l(\theta^{(t)}|\mathbf{N})$. We alternatively

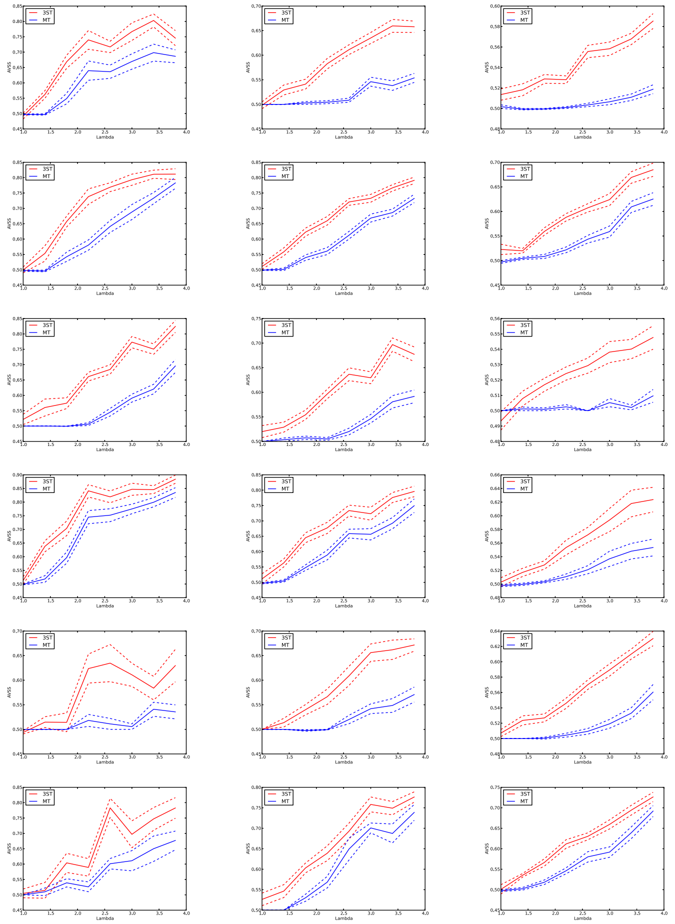

Figure 4. Performances of the proposed two-stage method with Hochberg's multiple testing adjustment and the multiple testing method on the cohort-design data with multiplicative or dominant or recessive inheritance modes. In these plots, the red and the blue solid curves show means of the AVSS values (i.e., the values of $(\text{specificity and sensitivity})/2$) *over 30 datasets are plotted against the values of* λ *for the proposed method and the multiple testing method respectively. The two red dash curves are one standard deviation up and down from the red mean curves. Similarly, the two blue dash curves are one standard deviation up and down for blue mean curves. The plots in the columns from the left to the right are for the cases where there were* 5, 10, *and* 20 *risk haplotypes in the underlying haplotypes. The top two rows, the middle two rows and the bottom two rows are the results for* $(N_0, N_1) = (2000, 1000)$ *and* $(3000, 2000)$ *under the multiplicative, the dominant and the recessive inheritance modes respectively. (Color figure online)*

repeat the E- and M-steps until $l(\theta^{(t+1)}|\mathbf{N}) - l(\theta^{(t)}|\mathbf{N})$ is less than a pre-specified number η, say $\eta = 0.0001$.

Choosing initial values for the EM algorithm is an important step in finding a maximum of the likelihood. There are various ways to do that such as random initialization and data partition. See [6] for a review. Here, we consider the following two methods to initialize the EM algorithm.

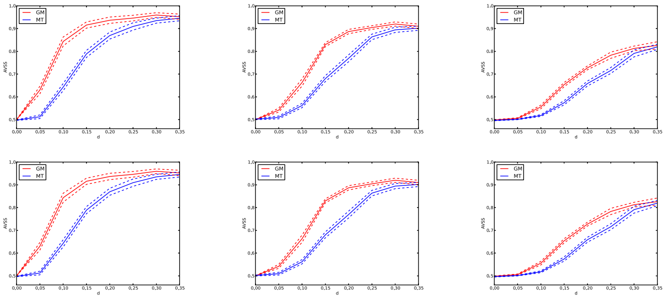

Figure 5. Performances of the proposed two-stage method with Hochberg's multiple testing adjustment and the multiple testing method on the case-control data. The plots in the columns from the left to the right are for the scenarios, where the underlying number of risk haplotypes $m_r = 5, 10,$ *and* 20. *The top row stands for the cases, where* $(N_0, N_1) = (2000, 1000)$, *while the bottom row stands for the cases, where* $(N_0, N_1) = (3000, 2000)$. *In these plots, the red and the blue solid curves show mean curves of the AVSS values over 30 datasets as functions of* $d = 0, 0.05, 0.1, 0.1,$ $0.15, 0.2, 0.25, 0.3,$ *and* 0.35 *for the proposed method and the multiple testing method respectively. The dash curves are one standard error up or down from the mean curves. (Color figure online)*

Method 1 (random initialization): We randomly choose i_0 initial values (say $i_0 = 100$) of θ and run the EM algorithm with each chosen initial value. We take the best one among these runs in terms of maximizing the log-likelihood.

Method 2 (data partition): Note that as pointed out before, the prospective frequencies of G_j in the controls and cases can be estimated by

$$\hat{p}_{0j} = \frac{N_{0j}}{N_{0j} + N_{1j}}, \quad \hat{p}_{1j} = \frac{N_{1j}}{N_{0j} + N_{1j}}$$

respectively. We first exclude the outlying frequencies in $\{\hat{p}_{1j}, \hat{p}_{0j}\}$, which have values of 0 or 1, to obtain robust means of a partition. Then, letting $c = (\max_j \hat{p}_{1j} - \min_j \hat{p}_{1j})/3$, we partition the frequencies into three sets as follows:

$$S_0 = \{\hat{p}_{1k} : \ \hat{p}_{1k} \le \min_j \hat{p}_{1j} + c\},$$
$$S_1 = \{\hat{p}_{1k} : \ \min_j \hat{p}_{1j} + c \le \hat{p}_{1k} < \ \min_j \hat{p}_{1j} + 2c\},$$

and

$$S_2 = \{\hat{p}_{1k} : \ \hat{p}_{1k} > \min_j \hat{p}_{1j} + 2c\}.$$

Note that the prospective frequency is increasing in the number of risk haplotypes which it carries. So, we expect that S_2, S_1 and S_0 mainly contain the frequencies corresponding the sets of genotypes with two risk haplotypes,

Table 5. The additional predicted risk haplotypes derived from our modified two-stage approached for CAD and HT by use of the WTCCC data. In the table, the p-values were derived from the chi-squared test of the frequencies of H_i against the collapsed frequencies of the estimated non-risk haplotypes

Chr	Region	SNP range	Haplotype	$\hat{P}(H_i\|case)$	$\hat{P}(H_i\|control)$	OR	p-Value	Gene
				CAD				
1	$SNP_A - 1786647$	$rs1180966 - rs54908760$	$GCGTCGAC$	0.0108	0.00089	15.54545	9.7×10^{-06}	C1orf175
1	$SNP_A - 4238771$	$rs10789042 - rs56980458$	$GGTTCGTC$	0.23665	0.17127	1.52975	2.7×10^{-05}	C1orf168
4	$SNP_A - 2043443$	$2352223 - 114179750$	$TATCGCCC$	0.01136	0.00108	10.38462	3.9×10^{-06}	LOC91431
4	$130622122 - 130672763$	$rs4975216 - rs17014667$	$GCATCGGC$	0.00756	0.00104	8.81729	4.5×10^{-05}	LOC391697
4	$143705041 - 143731526$	$rs17715707 - rs9308152$	$AAATGGGG$ $AAACGGAA$	0.09662 0.0887	0.07696 0.07524	2.30005 2.16017	4.4×10^{-06} 3.6×10^{-05}	INPP4B
9	$16944279 - 16951911$	$rs7021242 - rs16935195$	$GCGACCGA$	0.02571	0.01502	3.57182	2.5×10^{-07}	BNC2
10	$119397605 - 119419979$	$rs855994 - rs12572201$	$AATATCTG$	0.03346	0.01532	2.13367	3.7×10^{-05}	EMX2OS
12	$116500495 - 116514298$	$rs10850852 - rs1400593$	$CTCTTTTC$ $CTCTCTTC$	0.51578 0.28034	0.5 0.26232	3.79043 3.92754	1.8×10^{-55} 2.4×10^{-85}	NOS1
13	$108372995 - 108432811$	$rs4773010 - rs3842945$	$AGAGACCC$	0.27486	0.19222	1.40317	3.0×10^{-05}	MYO16
16	$63792132 - 63847234$	$rs1862709 - rs1423798$	$CGGATACT$	0.21037	0.19685	2.2091	2.3×10^{-05}	LOC283867
17	$13110258 - 13147203$	$rs17565276 - rs17572446$	$GGGTTTGA$	0.0807	0.05399	1.53479	2.8×10^{-05}	HS3ST3A1
19	$58535811 - 58602417$	$rs10405660 - rs2061772$	$ACAGCTGA$	0.04005	0.01282	2.67636	1.8×10^{-05}	ZNF765
20	$45769577 - 45836335$	$rs4407304 - rs2840278$	$GTGTCTAC$	0.01675	0.00479	3.59601	1.3×10^{-06}	SULF2
				HT				
17	$6992193 - 7158208$	$rs4558460 - rs6503013$	$TCGCGTCG$	0.14256	0.10161	1.46353	1.6×10^{-05}	LLGL1

with one risk haplotype, and with no risk haplotypes respectively. We choose the following initial values for estimating q_ν and π_ν, $\nu = 0, 1$:

$$q_\nu^0 = \frac{\sum_{p_{1j} \in S_\nu} p_{1j}}{|S_\nu|}, \text{ and } \pi_\nu^0 = \frac{|S_\nu|}{m},$$

where $|S_\nu|$ denotes the cardinality of S_ν.

Received 23 November 2014

REFERENCES

[1] Agresti, A. (1999). On logit confidence intervals for the odds ratio with small samples. *Biometrics* **55** 597–602.

[2] Browning, S. R. and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. *American Journal of Human Genetics* **81** 1084–1097.

[3] Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106** 9362–9367.

[4] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75** 800–802. MR0995126

[5] Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model. *Bioinformatics* **18** 337–338.

[6] Karlis, D. and Xekalaki, E. (2003). Choosing initial values for the EM algorithm for finite mixtures. *Comput. Stat. & Data Ana.* **41** 577–590. MR1968070

[7] Li, M., Ye, C., Fu, W., Elston, R. C., and Lu, Q. (2011). Detecting genetic interactions for quantitative traits with U-statistics. *Genet. Epidemiol.* **35** 457–468.

[8] Li Y., Byrnes, A. E., and Li, M. (2010). To identify associations with rare variants, just WHaIT: Weighted haplotype and imputation-based tests. *Ameri. Jour. Hum. Genet.* **87** 728–735.

[9] McLachlan, G. J. and Basford, K. E. (1988). *Mixture models: Inference and applications to clustering.* Marcel Dekker, New York. MR0926484

[10] Molitor, J., Marjoram, P., and Thomas, D. (2003). Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. *Am J Hum Genet.* **73** 1368–1384.

[11] Panoutsopoulou, K., Tachmazidou, I., and Zeggin, E. (2013). In search of low-frequency and rare variants affecting complex traits. *Hum. Mol. Genet.* **22** (R1) R16–R21.

[12] Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case–control studies. *Biometrika* **66** 403–411. MR0556730

[13] Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussions). *Jour. Roy. Stat. Soc. B* **59** 731–792. MR1483213

[14] Robinson, R. (2010). Common disease, multiple rare (and distant) variants. *PLoS Biol* **8** e1000293.

[15] Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M., and Poland, G. A. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.* **70** 425–434.

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

[17] Stephens, M. (2000) Dealing with label switching in mixture models. *Jour. Roy. Stat. Soc. B* **62** 795–809. MR1796293

[18] Stephens, P., Smith, N. J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68** 978–989.

[19] Stranger, B. E., Stahl, E. A., and Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* **187** 367–383.

[20] Tzeng, J. Y., Wang, C. H., Kao, J. T., and Hsiao, C. K. (2006). Regression-based association analysis with clustered haplotypes through use of genotypes. *Am. J. Hum. Genet.* **78** 231–242.

[21] Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17** 977–987.

[22] Van Greevenbroek, M., Zhang, J., van der Kallen, C., Schiffers, P., Feskens, E., and de Bruin, T. (2008). Effects of interacting networks of cardiovascular risk genes on the risk of type 2 diabetes mellitus (the CODAM study). *BMC Medical Genetics* **9** Article 36.

[23] The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls (WTCCC). *Nature* **447** 661–668.

[24] Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkinson, H. (2014). The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research* **42** (Database issue): D1001–D1006.

[25] Zhang, J., Liang, F., Dassen, W. R., Veldman, B. A., Doevendans, P. A., and De Gunst, M. (2003). Search for haplotype interactions that influence susceptibility to type 1 diabetes, through use of unphased genotype data. *Am. J. Hum. Genet.* **34** 171–187.

[26] Zhong, H. and Prentice, R. (2008). Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics* **9** 621–634. MR2712234

[27] Zhu X., Feng, T., Li, Y., Lu, Q., and Elston, R. C. (2010). Detecting rare variants for complex traits using family and unrelated data. *Genet. Epidemiol.* **34** 171–187.

[28] Zöllner, S. and Pritchard, J. K. (2007). Overcoming the winner's curse: Estimating penetrance parameters from case-control data. Convergence of random processes and limit theorems in probability theory. *Am. J. Hum. Genet.* **80** 605–615.

Fadhaa Ali
School of Mathematics
Statistics and Actuarial Science
University of Kent
Canterbury, Kent CT2 7NF
UK
E-mail address: fmha2@kent.ac.uk

Jian Zhang
School of Mathematics
Statistics and Actuarial Science
University of Kent
Canterbury, Kent CT2 7NF
UK
E-mail address: jz79@kent.ac.uk