Mixture Model-Based Association Analysis with Case-Control Data in Genome Wide Association Studies

Fadhaa Ali and Jian Zhang
School of Mathematics, Statistics and Actuarial Science
University of Kent, Canterbury, UK

Short title: Model-based Association Analysis

[0]Address for correspondence: Professor Jian Zhang, School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, Kent CT2 7NF, United Kingdom. E-mail: jz79@kent.ac.uk. Tel: +44 1227823648

Abstract

Multilocus haplotype analysis of candidate variants with genome wide association studies (GWAS) data may provide evidence of association with disease, even when the individual loci themselves do not. Unfortunately, when a large number of candidate variants are investigated, identifying risk haplotypes can be very difficult. To meet the challenge, a number of approaches have been put forward in recent years. However, most of them are not directly linked to the disease-penetrances of haplotypes and thus may not be efficient. To fill this gap, we propose a mixture model-based approach for detecting risk haplotypes. Under the mixture model, haplotypes are clustered directly according to their estimated disease penetrances. A theoretical justification of the above model is provided. Furthermore, we introduce a hypothesis test for haplotype inheritance patterns which underpin this model. The performance of the proposed approach is evaluated by simulations and real data analysis. The simulation results show that the proposed approach outperforms an existing multiple testing method in terms of average specificity and sensitivity. We apply the proposed approach to analyzing two datasets on coronary artery disease and hypertension in the Wellcome Trust Case Control Consortium, identifying many more disease associated haplotype blocks than does the existing method.

KEY WORDS: Genome wide association studies; haplotype mixture model; testing for inheritance patterns; odds ratios.

1 Introduction

The advanced genotyping technology and the availability of a large number of dense single nucleotide polymorphisms (SNPs) across human genome have enabled the design of genome-wide association studies (GWAS) for complex diseases. These studies have progressed from genotyping the SNPs over thousands of case and control subjects [Hindorff et al., 2009], producing large, high-dimensional genotype datasets. The rapid increase in the number of GWAS provides an unprecedented opportunity to examine the effects of rare SNPs on disease susceptibility by the integrative analysis of these data under the assumption that both common and rare SNPs contribute to the underlying genetic mechanisms of complex diseases [Li et al., 2010; Zhu et al., 2010]. It is generally believed that jointly analyzing rare SNPs within a region of strong linkage disequilibrium can be more informative and effective than individual SNP analysis, as multiple SNPs influence the risk of complex diseases in aggregate [Schaid et al., 2002; Tzeng et al., 2005; Morris, 2006; Li et al., 2011;Stranger et al., 2011]. The multilocus haplotype, the ordered allele sequences on a chromosome, provides a nature unit of analysis for capturing linear and non-linear correlations in SNPs [Zhang et al., 2003]. Unfortunately, the multi-SNP analysis discussed above can suffer from high-dimensional problems that are associated with many predictors, some of which are highly correlated. A popular strategy, suggested by the block-like structure of the human genome, is to divide each chromosome into a list of genetically meaningful regions to reduce the dimensions of these genotype data. Direct, laboratory-based haplotyping to infer the unknown phase are expensive ways to obtain haplotypes. So, in a typical haplotype-based association analysis, people infer haplotypes together with their population frequencies in cases and controls from observed genotypes by using the software such as PHASE [Stephens et al., 2001; Scheet et al., 2006]. The empirical evidence suggests that the majority of the polymorphism is concentrated on a relatively small number of haplotypes while the rest is sparsely spread over a number of categories. These non-common haplotypes can be rare and thus hard to assess their disease-susceptibility [Schaid et al., 2002;Tzeng et al., 2005].

Haplotype clustering offers a promising avenue for addressing the above issue. Over the past decade, enormous progress has been made in this direction and various methods of clustering have been developed on the basis of haplotype similarity and evolution characteristics [Molitor et al., 2003; Tzeng et al., 2006; Browning and Browning, 2007; and references therein]. However, none of them except Zhu et al. [2010] has explored advantage of the haplotype similarity in terms of their contributions to disease risks. Zhu et al. [2010] implemented a method for clustering rare

risk haplotypes by performing multiple mariginal Z-tests for the significant differences between retrospective haplotype frequencies in cases and controls, on the basis that rare risk haplotypes can be enriched in cases. The method of Zhu et al. [2010] may be too naive to be efficient. Therefore, it is desirable to develop a model-guided approach for haplotype clustering. Here, we propose a prospective model for haplotype counts in cases and controls, where given the marginal counts of haplotypes, the disease status of each haplotype follows a binomial mixture distribution. The main advantage of the proposed model over the other existing methods is that it allows the clustering to be directly linked to the haplotype disease-penetrances. Our intuition is as follows. We arrange the haplotype frequencies derived from a case-control study by a contingency table, where rows stand for the disease status (case or control) and columns for haplotypes. Then, we can directly assess whether two haplotypes belong to the same group by their column similarity in the table. To do that, we fit each column by a binomial distribution with the disease-penetrance as the success probability and group these columns by use of a binomial mixture. To account for the variation of disease-penetrances of haplotypes within risk and non-risk groups, the disease-penetrances are assumed to be random factors following certain prior distributions. Note that using the estimated prospective haplotype frequencies derived from a retrospective study to estimate disease odds ratio is known to be asymptotically consistent even though the disease-penetrance estimators may not be [Prentice and Pyke, 1979].

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. In this paper, taking advantage of the fact that the disease-penetrance can be varying across different risk haplotypes, we propose a Bayesian regularization procedure to improve the proposed mixture model and the corresponding EM algorithm by posterior sampling. We show its superior performance over the existing EM algorithm by simulations. We also conduct a large scale simulation studies on the proposed clustering method in both prospective and retrospective design settings, showing that the proposed method can outperform the approach of Zhu et al. [2010] in most cases. We apply both the proposed method and the method of Zhu et al. [2010] 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 and some theory are 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. Some technical details can be found in the Online Supplementary Material.

2 Methodology

Consider a case-control sample with N_0 controls and N_1 cases, typed at a list of pre-specified SNP markers in a candidate region, yielding unphased genotype set $\mathbf{G}$. Let $H_j, 1 \leq j \leq J$ denote the distinct haplotypes inferred from $\mathbf{G}$ with haplotype counts $n_{0j}, 1 \leq j \leq J$ in controls summing to $2N_0$, and $n_{1j}, 1 \leq j \leq J$ in cases summing to $2N_1$ respectively. The respective frequencies of the jth haplotype in controls and cases can be estimated by $r_{0j} = n_{0j}/(2N_0)$ and $r_{1j} = n_{1j}/(2N_1)$ respectively. Similarly, letting $n_j = n_{0j} + n_{1j}$, the prospective frequencies of the jth haplotype in cases and controls can also be estimated by $p_{0j} = n_{0j}/n_j$ and $p_{1j} = n_{1j}/n_j$ respectively.

When a haplotype is unevenly distributed between cases and controls, its odds ratio (OR) will be deviated from one and it is likely to be a risk haplotype. Therefore, multiple OR tests can be used for detecting risk haplotypes. Here, we opt for multiple OR testing, because risk haplotypes can directly be assessed by using their disease-penetrances via the OR values [Jewell, 2004]. The main thrust of our proposal below is to perform a model-based clustering on haplotypes before the OR testing. This can help reduce the number of haplotypes to be tested and thus reduce the multiple OR testing error.

2.1 Two-stage standard mixture approach

Our standard two-stage approach is processed as follows.

Stage 1 (Model-based clustering): We hypothesize that haplotypes are either risk or non-risk, where non-risk means neutral or protective to the disease. Under this assumption, given the haplotypes $H_j, 1 \leq j \leq J$ and their the marginal counts $(n_1, ..., n_J)$, the conditional distribution of the counts $\mathbf{n} = \{(n_{0j}, n_{1j})^T : 1 \leq j \leq J\}$ are modeled by the two-component binomial mixture,

$$f((n_{0j}, n_{1j})^T | p_r, p_{\bar{r}}, \pi) = \pi f((n_{0j}, n_{1j})^T | p_r) + (1 - \pi) f((n_{0j}, n_{1j})^T | p_{\bar{r}}),$$

where $p_r = P(\text{affected}|H_r)$ and $p_{\bar{r}} = P(\text{affected}|H_{\bar{r}})$ are the disease-penetrances of risk haplotype H_r and non-risk haplotype $H_{\bar{r}}$ respectively, and

$$\begin{aligned} f((n_{0j}, n_{1j})^T | p_r) &= \binom{n_{0j} + n_{1j}}{n_{1j}} p_r^{n_{1j}} (1 - p_r)^{n_{0j}}, \\ f((n_{0j}, n_{1j})^T | p_{\bar{r}}) &= \binom{n_{0j} + n_{1j}}{n_{1j}} p_{\bar{r}}^{n_{1j}} (1 - p_{\bar{r}})^{n_{0j}}. \end{aligned}$$

The unknown parameter $\theta = (p_r, p_{\bar{r}}, \pi)^T$ can be estimated by maximizing the log-likelihood

$$l(\theta|\mathbf{n}) = \sum_{j=1}^{J} \log\left(\pi f((n_{0j}, n_{1j})^T|p_r) + (1-\pi) f((n_{0j}, n_{1j})^T|p_{\bar{r}})\right).$$

Note that the direct calculation of the above maximum likelihood estimator (MLE) is difficult. Instead, we calculate it indirectly by the EM algorithm [McLachlan and Basford, 1988]. For this purpose, we introduce the following group membership indicators I_{jr} and $I_{j\bar{r}}$,

$$I_{jr} = \begin{cases} 1, & H_j \text{ in the risk group} \\ 0, & \text{otherwise} \end{cases}, \quad I_{j\bar{r}} = 1 - I_{jr}$$

for $1 \le j \le J$. Set $\mathbf{I} = \{(I_{jr}, I_{j\bar{r}})^T : 1 \le j \le J\}$. Then, the so-called complete-data log-likelihood can be written as

$$l(\theta|\mathbf{n}, \mathbf{I}) = \sum_{j=1}^{J} \left\{ I_{jr} \log(\pi f((n_{0j}, n_{1j})^T|p_r)) + I_{j\bar{r}} \log((1-\pi) f((n_{0j}, n_{1j})^T|p_{\bar{r}}))\right\}.$$

Given the current value $\theta^{(t)} = (p_r^{(t)}, p_{\bar{r}}^{(t)}, \pi^{(t)})^T$ and the data $\mathbf{n}$, we first calculate the current log-likelihood $l(\theta^{(t)}|\mathbf{n})$. Then, in the E-step, we calculate the expectation of the complete-data log-likelihood with respect to $\mathbf{I}$,

$$\begin{aligned} Q(\theta, \theta^{(t)}) &= E[l(\theta|\mathbf{n}, \mathbf{I})|\mathbf{n}, \theta^{(t)}] \\ &= \sum_{j=1}^{J} (\tau_{jr}^{(t)} \log(\pi) + \tau_{j\bar{r}}^{(t)} \log(1-\pi)) \\ &\quad + \sum_{j=1}^{J} (\tau_{jr}^{(t)} \log(f((n_{0j}, n_{1j})^T|p_r)) + \tau_{j\bar{r}}^{(t)} \log(f((n_{0j}, n_{1j})^T|p_{\bar{r}}))), \end{aligned}$$

where

$$\begin{aligned} \tau_{jr}^{(t)} &= \frac{\pi^{(t)} f((n_{0j}, n_{1j})^T|p_r^{(t)})}{\pi^{(t)} f((n_{0j}, n_{1j})^T|p_r^{(t)}) + (1-\pi^{(t)}) f((n_{0j}, n_{1j})^T|p_{\bar{r}}^{(t)})}, \\ \tau_{j\bar{r}}^{(t)} &= \frac{\pi^{(t)} f((n_{0j}, n_{1j})^T|p_{\bar{r}}^{(t)})}{\pi^{(t)} f((n_{0j}, n_{1j})^T|p_r^{(t)}) + (1-\pi^{(t)}) f((n_{0j}, n_{1j})^T|p_{\bar{r}}^{(t)})}. \end{aligned}$$

In the M-step, we update $\theta^{(t)}$ by solving the partial derivatives equations

$$\frac{\partial Q}{\partial \pi} = 0, \quad \frac{\partial Q}{\partial p_r} = 0, \quad \frac{\partial Q}{\partial p_{\bar{r}}} = 0.$$

We obtain

$$\pi^{(t+1)} = \frac{\sum_{j=1}^{J} \tau_{jr}^{(t)}}{J}, \quad p_r^{(t+1)} = \frac{\sum_{j=1}^{J} \tau_{jr}^{(t)} n_{1j}}{\sum_{j=1}^{J} \tau_{jr}^{(t)} (n_{1j} + n_{0j})}, \quad p_{\bar{r}}^{(t+1)} = \frac{\sum_{j=1}^{J} \tau_{j\bar{r}}^{(t)} n_{1j}}{\sum_{j=1}^{J} \tau_{j\bar{r}}^{(t)} (n_{1j} + n_{0j})}.$$

We calculate the updated log-likelihood $l(\theta^{(t+1)}|\mathbf{n})$ and its absolute distance to the previous $l(\theta^{(t)}|\mathbf{n})$ and $\text{err}^{(t+1)}$.

Start with the initial value $\theta^{(0)}$, we alternatively run the E-step and the M-step for $t = 0, 1, ...,$ till $\text{err}^{(t+1)}$ is less than a pre-specified value d_0 (we set $d_0 = 0.0001$ in our codes). Suppose that the algorithm stops at $(t+1)$th iteration. Note that $\tau_{jr}^{(t+1)}$ and $\tau_{j\bar{r}}^{(t+1)}$ are the posterior probabilities of the j-th haplotype being in risk and non-risk haplotype clusters respectively. So, based on these quantities, the estimated risk and non-risk haplotype clusters can be defined by

$$S_r^{(t+1)} = \{H_j : \tau_{jr}^{(t+1)} > \tau_{j\bar{r}}^{(t+1)}\}, \quad S_{\bar{r}}^{(t+1)} = \{H_j : \tau_{jr}^{(t+1)} \leq \tau_{j\bar{r}}^{(t+1)}\}.$$

We consider the two methods to choose the initial values for the EM algorithm: random initialization and data initial partition. See the Online Supplemental Material (Appendix I) for the details.

Stage 2 (Multiple OR testing): We are going to refine the above selected risk haplotype set on the basis of their odds ratios. Let n_{0H} and n_{1H} be control- and case-counts of the haplotype H. Let $n_{0\bar{r}} = \sum_{H_* \in S_{\bar{r}}^{(t+1)}} n_{0H_*}$ and $n_{1\bar{r}} = \sum_{H_* \in S_{\bar{r}}^{(t+1)}} n_{1H_*}$. The corrected OR statistic is defined by

$$\text{OR}_H = \frac{(n_{1H} + 0.5)(n_{0\bar{r}} + 0.5)}{(n_{0H} + 0.5)(n_{1\bar{r}} + 0.5)},$$

where adding 0.5 to the counts before computing the odds ratio was suggested by Agresti [1999] for continuity correction. Note that under the null hypothesis that the haplotype is evenly distributed between cases and controls,

$$\log(\text{OR}_H) \sim N(0, \phi(n_{0H}, n_{1H}, n_{0\bar{r}}, n_{1\bar{r}})^2),$$

where

$$\phi(n_{0H}, n_{1H}, n_{0\bar{r}}, n_{1\bar{r}}) = \sqrt{\frac{1}{n_{0H} + 0.5} + \frac{1}{n_{1H} + 0.5} + \frac{1}{n_{0\bar{r}} + 0.5} + \frac{1}{n_{1\bar{r}} + 0.5}}.$$

Then, the risk haplotype set $S_r^{(t+1)}$ (which are significant in the OR test) is updated by

$$\hat{S}_r = \left\{ H \in S_r^{(t+1)} : \text{OR}_H \geq \exp(c_1 \phi(n_{0H}, n_{1H}, n_{0\bar{r}}, n_{1\bar{r}})) \right\}$$

where c_1 is a pre-specified critical value for testing (invoking the Bonferroni adjustment, we set $c_1 = 2.6$ in the later simulations and $c_1 = 5.3$ in the real data analysis). The non-risk haplotype set is updated by

$$\hat{S}_{\bar{r}} = S_{\bar{r}} \cup (S_r - \hat{S}_r).$$

Given the clusters $\hat{S}_r$ and $\hat{S}_{\bar{r}}$, the estimators of π, p_r, and $p_{\bar{r}}$ are updated by

$$\hat{\pi} = \frac{|\hat{S}_r|}{|\hat{S}_r| + |\hat{S}_{\bar{r}}|}, \quad \hat{p}_r = \frac{\sum_{H \in \hat{S}_r} n_{1H}}{\sum_{H \in \hat{S}_r}(n_{1H} + n_{0H})}, \quad \hat{p}_{\bar{r}} = \frac{\sum_{H \in \hat{S}_{\bar{r}}} n_{1H}}{\sum_{H \in \hat{S}_{\bar{r}}}(n_{1H} + n_{0H})}.$$

The population frequencies of $\hat{S}_r$ and $\hat{S}_{\bar{r}}$ (i.e., $P(H \in \hat{S}_r)$ and $P(H \in \hat{S}_{\bar{r}})$ are estimated by their retrospective frequencies in controls,

$$\hat{P}(\hat{S}_r) = \frac{\sum_{H \in \hat{S}_r} n_{0H}}{\sum_{H \in \hat{S}_r \cup \hat{S}_{\bar{r}}} n_{0H}}, \quad \hat{P}(\hat{S}_{\bar{r}}) = 1 - \hat{P}(\hat{S}_r).$$

Note that according to the theory of Prentice and Pyke [1979], the OR based on $\hat{p}_r$ and $\hat{p}_{\bar{r}}$ above is asymptotically insensitive to the case-control sample ratio although $\hat{p}_r$ and $\hat{p}_{\bar{r}}$ can be affected by the ratio.

2.2 Two-stage hybrid mixture approach

In the previous mixture model, risk haplotypes are assumed to have the same disease-penetrance and so are non-risk haplotypes. Such a homogenous assumption may not hold in reality. To allow for the disease-penetrance variations within each group, we take p_r and $p_{\bar{r}}$ as random factors by imposing prior distributions on them. The resulting model is called a Bayesian regularized mixture model. The details are as follows.

Bayesian regularization. We first randomly generate i_0 (say $i_0 = 100$) initial values at which we calculate the log-likelihoods, and take the one, which attains the maximum and is denoted by $\theta^{(0)} = (p_r^{(0)}, p_{\bar{r}}^{(0)}, \pi^{(0)})^T$, as the initial value for the posterior sampling. Motivated by the Gibbs sampling, we employ the posterior of θ to improve each iteration of the EM. Here, we draw $q_r^{(t)}$ and $q_{\bar{r}}^{(t)}$ from the posteriors of p_r and $p_{\bar{r}}$ at the iteration t. Start with the initial $\theta^{(0)}$ and set $q_r^{(0)} = p_r^{(0)}$ and $q_{\bar{r}}^{(0)} = p_{\bar{r}}^{(0)}$. At the iteration $t+1$, given $\theta^{(t)} = (p_r^{(t)}, p_{\bar{r}}^{(t)}, \pi^{(t)})^T$, we have the expected values of I_{jr} and $I_{j\bar{r}}$, say $\tau_{jr}^{(t)}$ and $\tau_{j\bar{r}}^{(t)}$. Haplotype clusters can be defined by

$$S_r^{(t)} = \{H_j : \tau_{jr}^{(t)} > \tau_{j\bar{r}}^{(t)}\}, \quad S_{\bar{r}}^{(t)} = \{H_j : \tau_{jr}^{(t)} \leq \tau_{j\bar{r}}^{(t)}\}.$$

Collapse haplotypes in S_r and calculate the counts of the collapsed S_r in controls and cases, s_{0r} and s_{1r}. Similarly, collapse $S_{\bar{r}}$ and calculate the counts of the collapsed $S_{\bar{r}}$ in controls and cases, $s_{0\bar{r}}$ and $s_{1\bar{r}}$. Based on these counts, the likelihood functions of p_r and $p_{\bar{r}}$ can be written as

$$l(p_r|(s_{0r}, s_{1r})^T) \propto p_r^{s_{1r}}(1-p_r)^{s_{0r}}, \quad l(p_{\bar{r}}|(s_{0\bar{r}}, s_{1\bar{r}})^T) \propto p_{\bar{r}}^{s_{1\bar{r}}}(1-p_{\bar{r}})^{s_{0\bar{r}}}.$$

Let $p_r^{\delta_1}(1-p_r)^{\delta_0}$ and $p_{\bar{r}}^{\delta_0}(1-p_{\bar{r}})^{\delta_1}$ denote the conjugate priors for p_r and $p_{\bar{r}}$ respectively, with the pre-specified pseudo-counts δ_0 and δ_1. We expect that a risk haplotype appears more frequently in

cases than does any non risk haplotype. So, the pseudo-counts should satisfy the constrain $\delta_1 > \delta_0$. They should also be small compared to the number of cases. In this paper, we set $\delta_1 = 8$ and $\delta_0 = 2$. In our simulations, we found the results are not very sensitive to the choice of these constants.

After setting the above priors, we then derive the posteriors,

$$p(p_r|(s_{0r}, s_{1r})^T) \propto \text{Beta}(\delta_1 + s_{1r}, \delta_0 + s_{0r}), \quad p(p_{\bar{r}}|(s_{0\bar{r}}, s_{1\bar{r}})^T) \propto \text{Beta}(\delta_0 + s_{0\bar{r}}, \delta_1 + s_{1\bar{r}})$$

We draw $q_r^{(t+1)}$ from $p(p_r|(s_{0r}, s_{1r})^T)$ and $q_{\bar{r}}^{(t+1)}$ from $p(p_{\bar{r}}|(s_{0\bar{r}}, s_{1\bar{r}})^T)$. We update the estimates of p_r, $p_{\bar{r}}$ and π by posterior averaging,

$$p_r^{(t+1)} = \frac{1}{t+2}\sum_{k=0}^{t+1} q_r^{(k)}, \quad p_{\bar{r}}^{(t+1)} = \frac{1}{t+2}\sum_{k=0}^{t+1} q_{\bar{r}}^{(k)}, \quad \pi^{(t+1)} = \frac{|S_r^{(t)}|}{|S_r^{(t)}| + |S_{\bar{r}}^{(t)}|}.$$

Finally, we repeat the above procedure until the absolute difference between the estimates of θ in two consecutive iterations is less than a pre-specified value, say 0.0001.

In the Online Supplementary Material (Appendix II), we show the superiority of the Bayesian regularized M-step over the standard M-step by simulations. In light of this, we replace the M-step in the EM by the Bayesian regularized M-step to form a hybrid EM algorithm. In summary, we opt for the following *two-stage hybrid mixture approach* for association analysis in the remaining paper:

Stage 1 (Clustering): Use the hybrid EM algorithm to estimate the two-component binomial mixture model.

Stage 2 (OR testing): Use the OR statistic to test for risk haplotypes further as before.

2.3 Model justification

To make the proposed model identifiable, we need to assume that the disease-penetrance ratio $p_r/p_{\bar{r}} > 1$, that is, risk haplotypes are more enriched in cases than non-risk haplotypes. In this section, under the commonly used inheritance models, we prove the above hypothesis holds when the so-called relative risk measure is larger than one.

For this purpose, let S_r and $S_{\bar{r}}$ denote the risk and non-risk haplotype sets in the population. Suppose that the disease-penetrance of a genotype depends only on the number of risk haplotypes contained in that genotype. Then, we have three types of penetrance:

$$f_0 = P(\text{affected}|H_{\bar{r}}H_{\bar{r}}), \quad f_1 = P(\text{affected}|H_rH_{\bar{r}}), \quad f_2 = P(\text{affected}|H_rH_r),$$

where $H_r \in S_r$ and $H_{\bar{r}} \in S_{\bar{r}}$. Denote the relative risk measures $\lambda_1 = f_1/f_0$ and $\lambda = f_2/f_0$. In the Online Supplementary Material (Appendix III), we show that the haplotype disease-penetrances,

$P(\text{affected}|H_r)$ and $P(\text{affected}|H_{\bar{r}})$ are linear functions of the relative risk measures of genotypes and the population haplotype frequencies, namely

$$\begin{aligned} P(\text{affected}, H_r) &= f_0 \{\lambda P(H \in S_r) + \lambda_1 P(H \in S_{\bar{r}})\}, \\ P(\text{affected}|H_{\bar{r}}) &= f_0 \{\lambda_1 P(H \in S_r) + P(H \in S_{\bar{r}})\}, \end{aligned}$$

where $P(H_r)$, $P(H \in S_r)$ and $P(H \in S_{\bar{r}})$ are the population frequencies of H_r, S_r and $S_{\bar{r}}$.

The disease-penetrance ratio between risk and non-risk haplotypes,

$$\frac{P(\text{affected}|H_r)}{P(\text{affected}|H_{\bar{r}})} = \frac{\lambda_1\{\lambda P(H \in S_r)/\lambda_1 + P(H \in S_{\bar{r}})\}}{\lambda_1 P(H \in S_r) + P(H \in S_{\bar{r}})}.$$

We can further show that under the commonly used models of inheritance (multiplicative, dominant, and recessive), the haplotype relatively risk (i.e., the the disease-penetrance ratio between the risk and non-risk haplotypes) is larger than one if and only if the corresponding genotype relative risk is larger than one.

The above results imply that when the genotype relative risk $\lambda > 1$, the individuals carrying the risk haplotype H_r will have more chance of getting the disease than do non-risk haplotype carriers; when $\lambda < 1$, the individuals carrying H_r have the less chance of getting the disease than do non-risk haplotype carriers and thus H_r plays a disease-protective role.

2.4 Testing for haplotype inheritance modes

In the previous subsection, we develop a theory on the identification of the proposed model under certain inheritance assumption on hyplotypes. However, the biological justification for the choice of an inheritance model is seldom available and lack of a statistical justification for the specific genetic model is customary practice. To address the issue, we introduce a statistical test as follows.

We begin with deriving non-parametric estimators of the genotype disease-penetrances. Suppose that we have obtained $\hat{S}_r$ and $\hat{S}_{\bar{r}}$, the estimated risk and non-risk haplotype sets from our hybrid mixture approach. Let $\mathbf{G}_0$ be the set containing the observed genotypes which consist of two haplotypes in $\hat{S}_{\bar{r}}$, $\mathbf{G}_1$ the set containing the observed genotypes which consist of one haplotype in $\hat{S}_r$ and one in $\hat{S}_{\bar{r}}$, and $\mathbf{G}_2$ containing the observed genotypes which consist of two haplotypes in $\hat{S}_r$. For $k = 0, 1, 2$, we then calculate the total haplotype frequencies of $\mathbf{G}_k$ in controls and cases, denoted by $(n_{02}, n_{12}), (n_{01}, n_{11}), (n_{00}, n_{00})$ respectively. Then the disease-penetrances of genotypes can be estimated non-parametrically by

$$\hat{f}_0 = \frac{n_{10}}{n_{10} + n_{00}}, \quad \hat{f}_1 = \frac{n_{11}}{n_{01} + n_{11}}, \quad \hat{f}_2 = \frac{n_{12}}{n_{02} + n_{12}}.$$

Let A denote the set of the above three inheritance modes: the multiplicative, the dominant, and the recessive. We assume that genotypes are linked their underlying haplotype pairs via the Hardy-Weinberg equilibrium. To test for an inheritance mode, for $a \in A$ and $k = 0, 1, 2$, we first derive a parametric estimator of f_k, say $\hat{f}_k^{(a)}$ by using the estimators $\hat{p}_r$, $\hat{p}_{\bar{r}}$, $\hat{P}(\hat{S}_r)$ obtained in the previous subsection. We then calculate the statistic

$$D_a = |\hat{f}_0 - \hat{f}_0^{(a)}| + |\hat{f}_1 - \hat{f}_1^{(a)}| + |\hat{f}_2 - \hat{f}_2^{(a)}|.$$

We calculate the minimum $D_A = \min_{a \in A} D_a$ and record $\hat{a}$ at which D_a attains the minimum. We expect that D_A takes small values when one of modes in A is true. We can quantitatively justify the significance by use of the following parametric bootstrap test: We re-sampling genotypes M times on the basis of the estimated mode $\hat{a}$ with the estimated penetrances $\hat{f}_k^{(\hat{a})}$, $k = 0, 1, 2$. We set $M = 100$ in our simulation. Each bootstrap dataset contains the original genotypes (and their haplotype pairs) but with new sets of case and control counts. We apply the two-stage hybrid mixture approach to these datasets respectively, obtaining M bootstrap values $D_{Am}, m = 1, ..., M$. The empirical p-value $\sum_{m=1}^{M} I(D_A > D_{Am})/M$ can be used to judge the significance of the test.

To conclude this section, we now state the formulas for estimating the relative risk measures under the three inheritance models. The proofs are straightforward and thus omitted. We use the notations $\lambda = f_2/f_0$ and $\lambda_1 = f_1/f_0$ introduced before.

- *Multiplicative model, where* $\lambda = \lambda_1^2$. We have

$$\hat{\lambda} = \left(\frac{\hat{p}_r}{\hat{p}_{\bar{r}}}\right)^2, \quad \hat{f}_0 = \frac{\hat{p}_{\bar{r}}}{(\sqrt{\hat{\lambda}} - 1)\hat{P}(\hat{S}_r) + 1}, \quad \hat{f}_2 = \hat{\lambda}\hat{f}_0, \quad \hat{f}_1 = \sqrt{\hat{\lambda}}\hat{f}_0.$$

- *Dominant model, where* $\lambda = \lambda_1$. We have

$$\hat{\lambda} = \frac{\hat{P}(\hat{S}_{\bar{r}})}{\hat{p}_{\bar{r}}/\hat{p}_r - \hat{P}(\hat{S}_r)}, \quad \hat{f}_0 = \frac{\hat{p}_{\bar{r}}}{(\hat{\lambda} - 1)\hat{P}(\hat{S}_r) + 1}, \quad \hat{f}_1 = \hat{f}_2 = \hat{\lambda}\hat{f}_0.$$

- *Recessive model, where* $\lambda_1 = 1$. We have

$$\hat{\lambda} = (\hat{p}_r/\hat{p}_{\bar{r}} - \hat{P}(\hat{S}_{\bar{r}}))/P(\hat{S}_r), \quad \hat{f}_1 = \hat{f}_0 = \hat{p}_{\bar{r}}, \quad \hat{f}_2 = \hat{\lambda}\hat{f}_0.$$

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. Let

$\hat{\theta}$ be the estimator of θ, and $\hat{S}_r$ and $\hat{S}_{\bar{r}}$ the estimators of the true risk and non-risk haplotype sets S_r and $S_{\bar{r}}$ respectively. Then, by the L_1 bias we mean the L_1 distance between $\hat{\theta}$ and θ. The sensitivity and specificity of $\hat{S}_r$ and $\hat{S}_{\bar{r}}$ are defined as sen $= \frac{|\hat{S}_r \cap S_r|}{|S_r|}$ and spe $= \frac{|\hat{S}_{\bar{r}} \cap S_{\bar{r}}|}{|S_{\bar{r}}|}$. We take the average AVSS $=$ (sen $+$ spe)$/2$ to assess the performance of the haplotype classification above. As pointed out before, in light of our simulation study reported in the Online Supplementary Material (Appendix II), we adopted the two-stage hybrid mixture approach in the simulations and real data analysis below.

3.1 Performance of the proposed hybrid mixture approach

Note that the proposed hybrid mixture method is based on the prospective likelihood model although real data can be from retrospective studies. By the simulations below, we addressed whether the proposed hybrid mixture approach could outperform the multiple-testing procedure of Zhu et al. [2010] in both prospective (i.e., cohort) and retrospective (i.e., case-control) studies. See the Online Supplementary Material (Appendix IV) for the details of the procedure of Zhu et al. [2010].

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 in the Online Supplementary Material (Appendix II). In the third step, we simulated the disease status of each genotype by sampling from a Bernoulli distribution. The Bernoulli distribution took f_0, or $\lambda_1 f_0$, or λf_0 as a success probability according to whether the genotype contained zero, one or two risk haplotypes. We considered the three inheritance models coded by IM: the multiplicative (IM $= 1$), the dominant (IM $= 2$) and the recessive (IM $= 3$). Note that the values of (N_0, N_1) may vary across different datasets. We considered the scenarios with various combinations of $(N_0 + N_1, m_r, \text{IM}, f_0, \lambda)$, where $N_0 + N_1 = 3000, 5000$, $m_r = 5, 10, 20$, IM $= 1, 2, 3$, $f_0 = 0.1$, $\lambda = 1, 1.4, 1.8, 2.2, 2.6, 3, 3.4$, and 3.8 respectively.

For each scenario, we applied both the hybrid mixture method and the multiple testing method to 30 datasets and calculated their AVSS values respectively. For each of the three inheritance models, we plotted the means of these AVSS values over 30 datasets against λ. The results in Figure 1 show that on the cohort data, the hybrid mixture method performed substantially better than the multiple testing method in all the scenarios defined above.

[Put Figure 1 here.]

Setting 2 (case-control design): We generated 30 datasets, each of which were simulated

by the following two steps. Step 1, to generate N_1 case-genotypes, we first drew $2N_1$ haplotypes by 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, $1 \le j \le J$ be all the different genotypes contained in the N_1 cases and $r_{1j}, 1 \le j \le J$ 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. Step 2, we generated N_0 control-genotypes, which also had genotypes G_j, $1 \le j \le J$ but with population retrospective frequencies $q_{0j}, 1 \le j \le J$. We first let $q_{0j}, 1 \le j \le J$ 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 $d \ge 0$ 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 calculated the corresponding retrospective frequencies $r_{0j}, 1 \le j \le J$. We considered the cases where $m_r = 5, 10, 20$, and $d = 0, 0.05, 0.1, 0.1, 0.15, 0.2, 0.25, 0.3$, and 0.35 respectively.

For each dataset, the cumulative genotype 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. This implies that due to the impacts of risk haplotypes, the cumulative frequencies of Groups 2 and 1 in cases have been increased compared to those in controls. The odds ratios between Groups 2 and 0 and between Group 1 and Group 0, $(1 + 1.5d/r_{g_0})/(1 - d/r_{g_2})$ and $(1 + 1.5d/r_{g_0})/(1 - 0.5d/r_{g_1})$, are larger than one. Similarly, the odds ratio between the risk haplotype group and the non-risk haplotype group can be expressed as $(1 + 2.5d/(2r_{g_0} + r_{g_1}))/(1 - 2.5d/(r_{g_1} + 2r_{g_2}))$. All these ratios are increasing in d.

We applied the hybrid mixture 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 2. The results again demonstrate that the hybrid mixture method can be more powerful than the multiple testing method in detecting risk haplotypes.

[Put Figure 2 here.]

3.2 Performance of the proposed inheritance mode test

For each of the three inheritance models, we generated 30 datasets. Each dataset was simulated as follows. Following the cohort design, we first simulated $N_0 + N_1$ genotypes, where the underlying haplotypes contained $m_r = 10$ risk-haplotypes and followed the Hardy-Weinberg equilibrium. We then simulated their disease status by use of the inheritance models with $f_0 = 0.1$ and $\lambda = 1, 1.4, 1.8, 2.2, 2.6, 3, 3.4$, and 3.8 respectively as we did in the previous subsection.

For each dataset, we calculated D_A and the optimal mode $\hat{a}$. We generated 100 parametric bootstrap samples of the genotype frequencies based on the mode $\hat{a}$ and calculated the corresponding values of the inheritance testing statistic, $D_A^{(k)}$, $k = 1, ..., 100$. Based on these values, we obtained the empirical p-value.

We calculated the success rates by counting how many times that $\hat{a}$ is the true mode over the 30 datasets for each λ. These success rates and the empirical p-values are displayed in Figure 3. The results indicate that the success rates are increasing as λ is increasing. The box-whisker plots in Figure 3 show that almost all the empirical p-values are above 0.20, suggesting that almost all the tests are not significant. Therefore, the bootstrap test has a very high power in finding the true inheritance modes in the data.

[Put Figure 3 here.]

4 Real data analysis

We applied the proposed hybrid mixture approach to the GWAS genotype datasets on coronary artery disease (CAD) and hypertension (HT) obtained by Affymetrix 500K SNP chips in the WTCCC study [WTCCC, 2007]. 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. These data were downloaded from the WTCCC website. We first pre-processed the data by excluding the SNPs which meet one of the following criteria: (1) the HWE Fisher test p-value is less than 10^{-8} in controls; (2) the chi-square test p-value 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. We divided the genome into regions (or blocks) of around 8 SNPs according to their positions on the chromosomes, obtaining 61218 regions. Note that the long block will dilute the effects of risk SNPs whereas the

short block will miss interactions between SNPs. The block length of 8 was chosen to achieve a compromise between the above aspects. Also note that as we excluded the SNPs with bad callings, the numbers of cases and controls are varying across the different regions.

For all regions, we first reconstructed the haplotype pairs of genotypes by use of the software PHASE, to which we applied Stage 1 of the hybrid procedure. It led to 902909 haplotypes and 961942 haplotypes to be declared as risk haplotypes at Stage 1 for the CAD and the HT respectively. We then calculated the OR tests on these haplotypes at Stage 2. At Stage 2, According to the Bonferroni adjustment, the individual significance level was set at the levels of $0.05/902909 = 5.5 \times 10^{-8}$ and $0.05/961942 = 5.2 \times 10^{-8}$ for the CAD and the HT respectively. These individual significance levels were then used to determine the thresholding level c_1 in the multiple OR thresholding, which is $c_1 = 5.3$.

After performing the proposed hybrid mixture procedure on the datasets, we obtained the estimated risk and non-risk haplotype sets, $\hat{S}_r$ and $\hat{S}_{\bar{r}}$), for the CAD and the HT respectively.

Note that there were two sub-populations in controls. Any estimated risk haplotype which is significant in differing two control sub-populations should be viewed as an artifact. By using this, we made further quality control on the selected haplotypes by running the chi-square tests on the association of two control sub-populations with each selected risk haplotype. We eliminated these risk haplotypes whose p-values for the above chi-square tests were $< 30\%$. Here, 30% was chosen by the simulations in the Supplementary Web Material, aiming to filter out these artificial risk haplotypes with parameter $d \geq 0.05$. From the simulations, we can see that when $d = 0.05$, these p-values would be less than or equal to 0.30 most times.

Finally, we calculated the ORs for all the estimated haplotypes and thresholded them by using the bound

$$\exp(c_1\sqrt{1/(n_{0H}+0.5)+1/(n_{1H}+0.5)+1/(n_{0\bar{r}}+0.5)+1/(n_{1\bar{r}}+0.5)})$$

with $c_1 = 5.3$. This 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 Welter et al. (2014) and the web page at http://www2.le.ac.uk/projects /birth-cohort/1958bc. In the CAD case, we did rediscover the CAD risk gene CDKN2B and the risk haplotype "GGTGCCAG" found by the previous study (WTCCC, 2007; Zhu et al., 2010). We also tested the inheritance modes for these risk haplotypes. Taking the gene CDKN2B as an example,

we obtained $D_A = 0.4087$ with $\hat{a} =$ "dominant mode" and the empirical p-value of 0.97, suggesting that the hyplotype "GGTGCCAG" in the gene followed the dominant inheritance mode.

[Put Tables 1, 2, 3 and 4 here.]

5 Discussion and conclusion

The GWAS and sequencing studies have produced a huge amount of high-dimensional data. Analyzing these data offers many challenges to statistical inference. Several empirical studies have demonstrated the superiority of SNP region-based association analysis over single-SNP strategy [see Zakharov et al., 2013 and reference therein]. However, even restricted to a region, we may still obtain many sparsely distributed haplotypes derived from phasing the genotypes. In this case, the traditional generalized linear model-based approach [Schaid et al., 2002] may not be effective in detecting rare disease-associated haplotypes. In the presence of sparsely distributed haplotypes, haplotype clustering is very useful for performing statistical analysis on such kinds of data. Most of the existing methods of haplotype clustering are heuristic and not disease-penetrance based. To overcome this drawback, we have proposed a hybrid mixture model-based approach for grouping and identifying risk haplotypes. The key ingredient of the approach is a prospective mixture model with priors. The proposal includes two stages: in the stage 1, one groups haplotypes and therefore reduce the haplotype sparsity, while in the second stage, one conducts a two-sample Z-test based screening on the haplotypes derived from the previous stage. We have also provided a test for genetic inheritance modes. We have hypothesized that haplotypes are either risk or non-risk, where non-risk means neutral or protective to the disease. However, if we are also interested in identifying protective haplotypes, we can easily extend the current framework to address the issue by use of a three-component binomial mixture model.

We have examined the performance of the proposed procedure by a theoretical analysis, simulations and a real data analysis. We have showed that under the Hardy-Weinberg equilibrium, the risk haplotype group is identifiable if genotype relative risk is not equal to one. Compared to the standard multiple Z-testing method, the proposed procedure is more efficient in terms of sensitivity and specificity. We applied our procedures to the WTCCC CAD and hypertension data, rediscovering 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 often performs better than does the multiple Z-test

approach.

We note that the proposed mixture model can be combined with haplotype-based logistic regression to account for covariates. However, further studies are beyond the scope of the paper.

Description of online supplementary materials

Online supplementary material contains further information about the initialization of the EM algorithm, the performance of the proposed Bayesian regularization, the proof of the disease-penetrance formulas, and the multiple Z-testing method.

Acknowledgements

We thank the WTCCC for sharing their data with us. The project was partially done when the second author was visiting Chinese Academy of Sciences, Beijing. The second author thanks Professor Guohua Zou for his hospitality. The research of the first author was funded by the Ministry of Higher Education and Scientific Research, Iraq. The authors have no conflict of interest to declare.

References

Agresti, A. 1999. On logit confidence intervals for the odds ratio with small samples. *Biometrics* **55**: 597-602.

Browning, S. R. & 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

Hindorff, L.A., Sethupathy, P., Junkins,H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. & 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**: 93629367.

Hudson, R. R. 2002 Generating samples under a Wright-Fisher neutral model. *Bioinformatics*, **18**: 337-8.

Karlis, D. and Xekalaki, E. 2003. Choosing initial values for the EM algorithm for finite mixtures. *Comput. Stat. & Data Ana.* , **41**: 577-590.

Jewell, N.P. 2004. Statistics for Epidemiology. New York: Chapman & Hall/CRC.

Li, M., Ye, C., Fu, W., Elston, R.C., & Lu, Q. 2011. Detecting Genetic Interactions for Quantitative Traits with U-Statistics. *Genet. Epidemiol.* **35**: 457-468.

Li Y., Byrnes, A.E. & Li, M. 2010. To identify associations with rare variants, Just WHalt: Weighted haplotype and imputation-based tests. *Ameri. Jour. Hum. Genet.* **87**: 728-735.

McLachlan, G.J. & Basford, K.E. 1988. *Mixture models: Inference and applications to clustering.* New York: Marcel Dekker.

Molitor, J., Marjoram, P.& Thomas, D. 2003. Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. *Am J Hum Genet.* **73**: 13681384.

Morris, A.P. 2006. A flexible Bayesian framework for modeling haplotype association with disease, allowing for dominance effects of the underlying causative variants. *Am J Hum Genet.* **79**: 679694.

Prentice, R.L. & Pyke, R. 1979. Logistic disease incidence models and casecontrol studies *Biometrika* **66**: 403411.

Schaid, D.J., Rowland, C.M., Tines, D. E., Jacobson, R. M., Poland, G.A. 2002. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Ameri. Jour. Hum. Genet.* **70**: 425- 434.

Scheet,P. & Stephens, M. 2006. A fast and flexible method for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Ameri. Jour. Hum. Genet.* **78**: 629-644.

Stephens, P., Smith, N.J. & Donnelly, P. 2001. A new statistical method for haplotype reconstruction from population data. *Ameri. Jour. Hum. Genet.* **68**: 978-989.

Stranger, B.E., Stahl, E.A. & Raj, T. 2011. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* **187**: 367-383.

Tzeng, J.Y., Wang, C. H., Kao, J.T. & Hsiao, C.K. 2006. Regression-based association analysis with clustered haplotypes through use of genotypes. *Ameri. Jour. Hum. Genet.* **78**: 23142.

WTCCC 2007. The Wellcome Trust Case Control Consortium Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**: 661668.

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.

Zakharov, S., Wong, T.Y., Aung, T., Vithana, E.N., Khor, C.C., Salim, A., Thalamuthu, A. 2013. Combined genotype and haplotype tests for region-based association studies. *BMC Genomics* **14**: 569.

Zhang, J., Liang, F., Dassen, W.R., Veldman, B.A., Doevendans, P.A., De Gunst, M. 2003. Search for haplotype interactions that influence susceptibility to type 1 diabetes, through use of unphased genotype data. *Ameri. Jour. Hum. Genet.* **73**: 1385401.

Zhu, X., Feng, T., Li, Y., Lu, Q. & Elston, R.C. 2010. Detecting rare variants for complex traits using family and unrelated data. *Genet. Epidemiol.* **34**: 171-187.

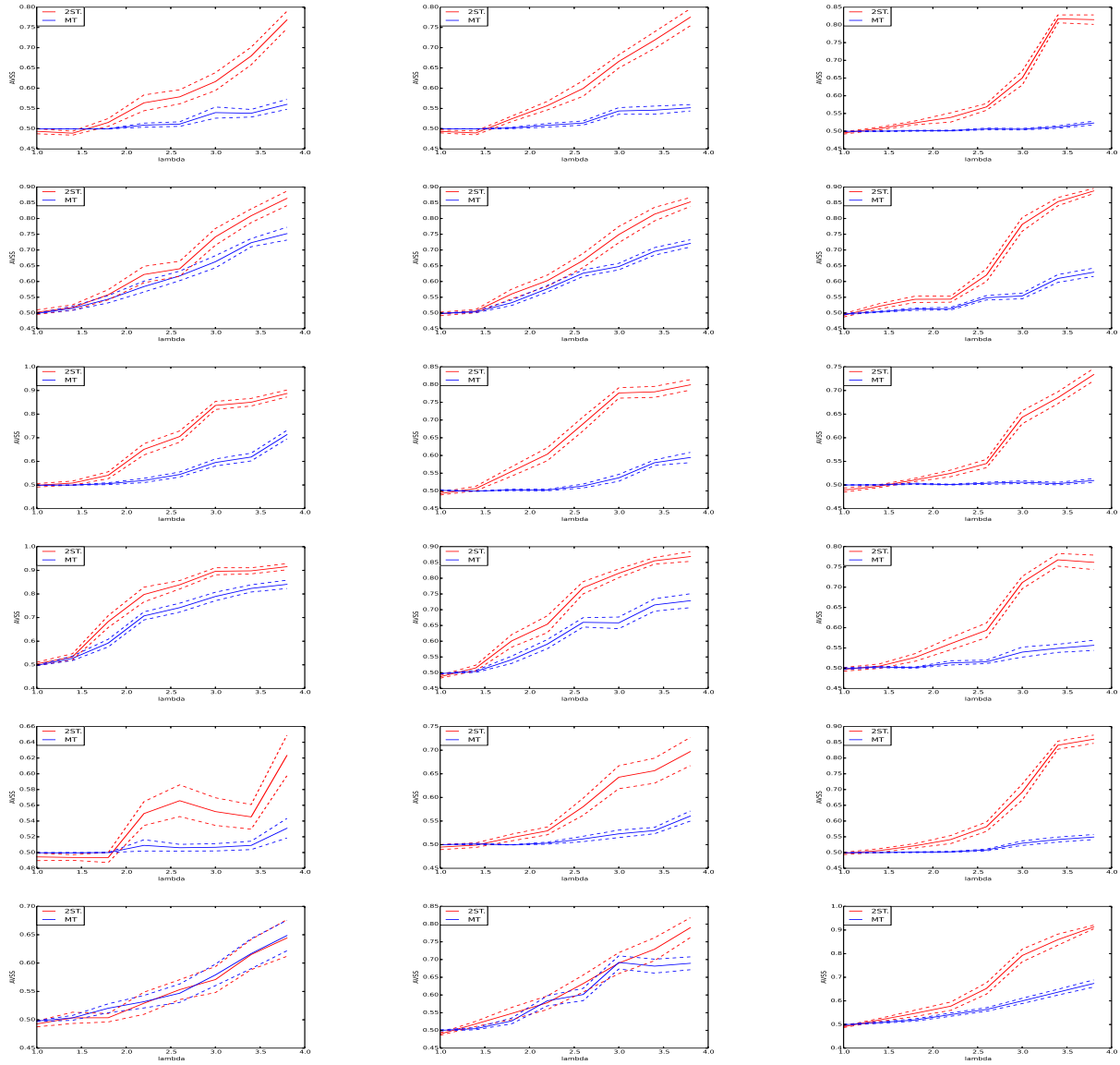

Figure 1: Performances of the proposed hybrid mixture method and the multiple testing method on the cohort-design data with multiplicative or dominant or recessive inheritance models. In these plots, the red and the blue solid curves, showing means of the AVSS values (i.e., the values of (specificity and sensitivity)/2)) over 30 datasets, were plotted against the values of λ for the hybrid mixture 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 models respectively.

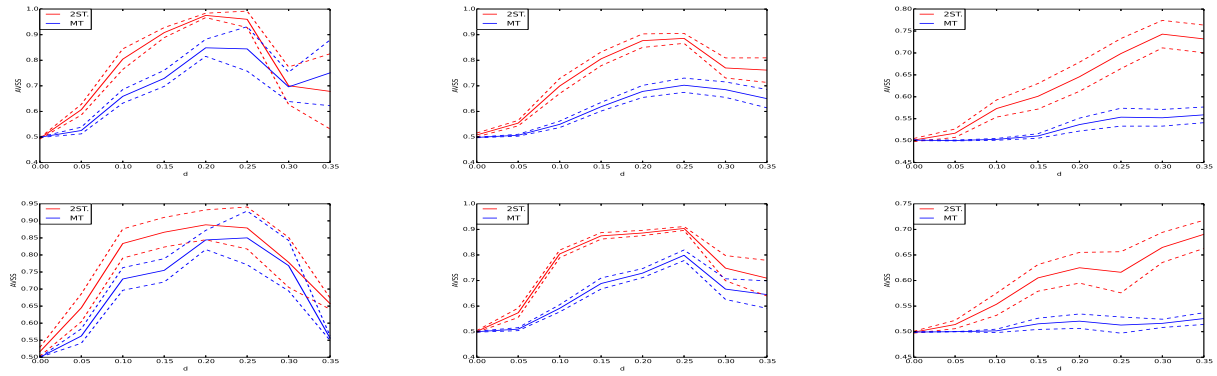

Figure 2: Performances of the proposed hybrid mixture and the multiple testing method on the case-control data. The plots in the columns from the left to the right are for the scenarios, where the underlying number of risk haplotypes $m_r = 5, 10$, and 20. The top row stands for the cases, where $(N_0, N_1) = (2000, 1000)$, 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 hybrid mixture method and the multiple testing method respectively. The dash curves are one standard error up or down from the mean curves.

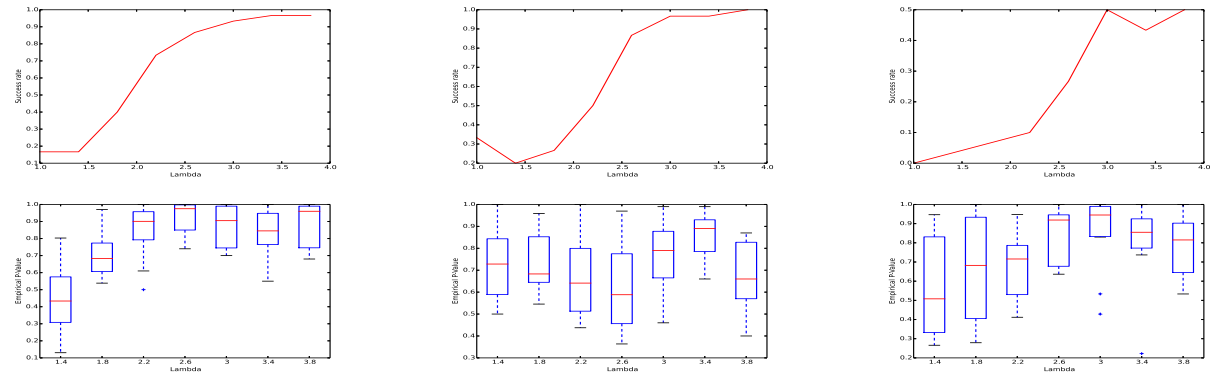

Figure 3: Performances of the proposed test for inheritance patterns. The plots in the columns from the left to the right are for the dominant, the multiplicative and the recessive models respectively. The top row show the success rate of identifying the true inheritance mode against λ over 30 datasets, while the bottom row show the box-whisker plots of the empirical p-values (based on 100 bootstrap samples) against λ for the inheritance test statistic $D_{\min}$ over 30 datasets.

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-square 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	$202166400 - 202187685$	$rs6692041 - rs1041311$	$AAATGGGA$	0.07815	0.05083	1.95856	2.8×10^{-13}	LOC284577
1	$237650028 - 237672617$	$rs6683639 - rs10802930$	$TCAAATGC$	0.05256	0.02763	2.57538	6.1×10^{-13}	RGS7
3	$102073696 - 102093722$	$rs973309 - rs4928094$	$TAACCTTT$	0.07591	0.06898	7.73184	5.6×10^{-15}	ABI3BP
3	$142488272 - 142537277$	$rs7643346 - rs2871887$	$CGCCCATC$	0.05008	0.03809	11.90617	2.0×10^{-15}	ACPL2
3	$147806667 - 147828893$	$rs17433833 - rs17434589$	$CCGGGGGC$	0.03363	0.01291	3.14753	3.2×10^{-15}	PLSCR5
4	$132550 - 344051$	$rs11735742 - rs17719492$	$TGGCACTC$	0.05993	0.04793	1.9902	7.6×10^{-11}	LOC654254
4	$4464610 - 4499426$	$rs16835627 - rs4234727$	$TCGAGCAT$	0.04072	0.0251	3.92994	1.1×10^{-14}	ZNF509
			$CTAAGCAT$	0.09413	0.07343	3.10696	1.2×10^{-13}	
4	$180659963 - 180699763$	$rs6811556 - rs17090633$	$CCCCCACT$	0.01782	0.00755	7.33583	5.5×10^{-15}	LOC391719
5	$157267571 - 157303032$	$rs10071157 - rs17055168$	$GTGAGCAA$	0.02135	0.00701	3.93074	4.0×10^{-13}	CLINT1
7	$77725471 - 77739291$	$rs10485891 - rs7803705$	$AACATGCG$	0.03652	0.04027	3.67364	6.8×10^{-13}	MAGI2
			$AACATGTA$	0.01312	0.01117	4.76163	2.1×10^{-11}	
			$AGTGCACA$	0.01312	0.00846	6.27027	1.2×10^{-14}	
7	$130749877 - 130784667$	$rs4728224 - rs4728225$	$AGAACCGG$	0.14061	0.13197	4.05796	1.0×10^{-12}	LOC647030
8	$104190450 - 104202402$	$rs2515173 - rs3019159$	$GGCCATCT$	0.14195	0.08768	2.20746	2.5×10^{-27}	BAALC
9	$22088619 - 22120515$	$rs2891168 - rs10965245$	$GGTGCCAG$	0.34939	0.29298	1.90115	2.7×10^{-11}	CDKN2B
9	$77341767 - 77366988$	$rs2889774 - rs3780296$	$ATGAGAGT$	0.01936	0.01072	5.31687	5.0×10^{-18}	GNA14
			$ATGAAGAC$	0.03898	0.03923	2.93116	4.5×10^{-13}	
			$ATGGAAAT$	0.06672	0.042	4.68028	4.9×10^{-30}	
			$GCGAAGAT$	0.14207	0.14656	2.85712	4.9×10^{-19}	
9	$131714465 - 131751663$	$rs3012758 - rs11243551$	$CGAATTGC$	0.06641	0.04652	2.41478	6.2×10^{-13}	RAPGEF1
			$CGAACTGC$	0.02448	0.01227	3.36929	4.4×10^{-12}	
10	$64409674 - 64442476$	$rs1509952 - rs2842286$	$TTTCTTAC$	0.02299	0.0073	9.37291	1.6×10^{-16}	NRBF2
10	$112527724 - 112597595$	$rs17763100 - rs1341055$	$GCCTCCCG$	0.07752	0.07383	1.85031	6.2×10^{-11}	RBM20
			$ACCTCCCG$	0.24688	0.21703	2.00368	6.7×10^{-23}	
10	$129835144 - 129894934$	$rs11016102 - rs1335014$	$AAGAACTT$	0.02987	0.01529	4.40461	6.2×10^{-14}	MKI67
11	$36361306 - 36410807$	$rs330255 - rs331485$	$GCGATTAA$	0.0309	0.00779	4.87953	1.5×10^{-21}	FLJ14213
11	$133079508 - 133113640$	$rs4937817 - rs4937826$	$GTAGTGCC$	0.04216	0.02425	2.69929	5.9×10^{-17}	LOC646522
			$CCGGCCCG$	0.05747	0.04018	2.22186	1.4×10^{-15}	
			$GTAGCCCG$	0.04001	0.02779	2.23683	8.3×10^{-12}	

Table 2: The continuation of Table 1.

Chr	Region	SNP range	Haplotype	$\hat{P}(H_i\|case)$	$\hat{P}(H_i\|control)$	OR	P-Value	Gene
11	133914862 – 133953680	$rs12417998 - rs10894845$	$GTTAGCCC$	0.12907	0.13389	3.70503	1.4×10^{-12}	IQSEC3
			$GTTAATCC$	0.09778	0.09576	3.92451	3.3×10^{-13}	
			$GTCAGCTC$	0.06932	0.07079	3.76457	7.6×10^{-12}	
12	24250132 – 24288211	$rs3922562 - rs17412555$	$CTGTGCCT$	0.07253	0.06027	5.51363	6.0×10^{-15}	SOX5
			$TCGCGCCC$	0.05454	0.03857	6.47638	9.9×10^{-17}	
			$TCGCGTCC$	0.02399	0.01788	6.14651	1.0×10^{-12}	
12	51469295 – 51501190	$rs17738862 - rs876407$	$CACCCTCG$	0.14455	0.13704	2.25981	2.6×10^{-13}	KRT3
12	127083338 – 127105747	$rs7960047 - rs9668398$	$GTGCGTCT$	0.06573	0.06076	3.67491	2.7×10^{-15}	TMEM132C
15	37962389 – 38014169	$rs11633436 - rs534757$	$TTACAACC$	0.07798	0.03763	2.66998	3.9×10^{-26}	EIF2AK4
16	79852394 – 79892297	$rs6564863 - rs11639552$	$TTCGTTAT$	0.02663	0.01053	5.1576	7.7×10^{-16}	BCMO1
17	29052246 – 29089136	$rs2046899 - rs17783280$	$AGTCAATC$	0.11305	0.0966	2.10899	5.7×10^{-14}	LOC646202
17	52973696 – 53057256	$rs17834557 - rs3744089$	$TGGTTAAC$	0.05825	0.03915	2.15515	8.7×10^{-14}	MSI2
18	9649377 – 9700554	$rs1965881 - rs1455587$	$TCACATGT$	0.06243	0.04149	2.15776	6.3×10^{-13}	RAB31
18	60647495 – 60688045	$rs1595904 - rs17678507$	$CAGTATAT$	0.09403	0.0848	2.55691	1.2×10^{-11}	C18orf20
18	72313651 – 72356779	$rs17059443 - rs8084536$	$GCGAGACC$	0.08958	0.08373	2.43635	1.0×10^{-11}	FLJ44313
19	4625799 – 4746342	$rs11670570 - rs1044409$	$AGCAACCG$	0.05419	0.02332	3.3426	6.7×10^{-25}	DPP9
19	56075162 – 56127664	$rs187930 - rs1654545$	$ACATGTGA$	0.03532	0.02898	7.24575	3.3×10^{-13}	KLK2
19	58460745 – 58519652	$rs1978611 - rs7408137$	$AGGTAGTG$	0.05628	0.042	1.99812	4.0×10^{-12}	VN1R4
22	35324014 – 35335429	$rs7410412 - rs12160203$	$TCCTAGGG$	0.44488	0.50199	3.09116	1.6×10^{-21}	CACNG2
			$GCCTAGAG$	0.03358	0.02891	4.05372	6.2×10^{-17}	

Table 3: The predicted risk haplotypes of hypertension by use of WTCCC data. In the table, the P-values were derived from the chi-square 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	1586208 – 1753641	$rs6603791 - rs2272908$	$AACCCATC$	0.03406	0.01973	2.45812	2.7×10^{-12}	SSU72
1	227569611 – 227620956	$rs7514972 - rs9431663$	$CGTATAGG$	0.03377	0.00926	7.08695	2.8×10^{-32}	TRIM67
1	227914995 – 228040530	$rs16854388 - rs1655296$	$CAAGGTAG$	0.04372	0.04622	2.90643	1.9×10^{-13}	TSNAX
1	236986859 – 237020204	$rs12137158 - rs16840310$	$GCTGTGGG$	0.02424	0.01534	2.95857	1.7×10^{-11}	GREM2
			$ATTTAGGG$	0.08733	0.05437	3.00646	3.0×10^{-26}	
			$GCTTTGAG$	0.0756	0.06745	2.09936	1.1×10^{-12}	
3	101569551 – 101696774	$rs277640 - rs4928098$	$CCCAGGCG$	0.02137	0.00908	6.27332	1.9×10^{-13}	TOMM70A
3	142488272 – 142537277	$rs7643346 - rs2871887$	$AGCTCATC$	0.17323	0.17868	2.2344	4.4×10^{-11}	ACPL2
3	142878508 – 142912781	$rs12485838 - rs16851691$	$GCATAGAG$	0.02089	0.00902	5.09818	1.3×10^{-13}	LOC646730
4	21080985 – 21131665	$rs1495517 - rs358574$	$GTCGCACG$	0.05716	0.04649	7.36766	4.2×10^{-13}	KCNIP4
			$GTTGCACG$	0.06033	0.04669	7.74264	7.1×10^{-14}	
4	23359572 – 23389742	$rs10008808 - rs1976201$	$AGTTCTTA$	0.03874	0.01347	3.68417	1.5×10^{-20}	PPARGC1A
5	10695437 – 10746687	$rs2062200 - rs6891527$	$GTCACACG$	0.16002	0.14866	6.18799	2.8×10^{-23}	LOC651746
5	32084851 – 32103155	$rs438834 - rs10065850$	$TGCTCCCA$	0.02254	0.01065	14.40157	1.6×10^{-24}	PDZD2
6	139560239 – 139612833	$rs7765885 - rs9495394$	$GCGCAACG$	0.0487	0.01774	4.54602	2.2×10^{-35}	HECA
			$ACGAAATG$	0.01641	0.00709	3.82046	6.4×10^{-12}	
			$GTACAATA$	0.14141	0.13391	1.75292	4.9×10^{-16}	
6	139693238 – 139758634	$rs11155050 - rs9373237$	$TTGCGGCT$	0.01924	0.00686	5.16669	1.1×10^{-14}	TXLNB
			$CTAAGATT$	0.25795	0.24508	1.9524	6.6×10^{-11}	
7	48232027 – 48237897	$rs17729647 - rs2362301$	$AGACTGGT$	0.07901	0.07156	3.41729	4.7×10^{-15}	ABCA13
			$AGATTGAC$	0.03345	0.02897	3.57621	3.7×10^{-12}	
			$AGATTGGC$	0.35755	0.38319	2.88725	3.0×10^{-14}	
7	77695246 – 77717237	$rs2215379 - rs4515471$	$CTTAAAAA$	0.03102	0.01998	4.32524	2.1×10^{-21}	MAGI2
			$TCTAAAAA$	0.02943	0.01786	4.58962	5.0×10^{-22}	
			$CTTGGAAA$	0.02094	0.01061	5.49009	8.4×10^{-21}	
			$CCTAGAAA$	0.05541	0.05534	2.79199	2.4×10^{-16}	
			$CCGAAAAA$	0.13203	0.13667	2.6926	4.0×10^{-21}	
9	77269212 – 77301387	$rs17063627 - rs7032444$	$GCGGACAG$	0.03393	0.01858	3.58867	1.6×10^{-12}	GNA14
10	119535731 – 119568729	$rs4752106 - rs10787797$	$TATTCACA$	0.09968	0.06304	2.91842	4.8×10^{-19}	RAB11FIP2

Table 4: The continuation of Table 3.

Chr	Region	SNP range	Haplotype	$\hat{P}(H_i\|case)$	$\hat{P}(H_i\|control)$	OR	P-Value	Gene
11	125683058 – 125763272	$rs2096915 - rs7118117$	$CACACGAG$	0.07736	0.04727	2.42988	4.9×10^{-12}	ST3GAL4
12	27155055 – 27179334	$rs841636 - rs841613$	$TAAAGGGT$	0.05414	0.04075	2.81343	7.5×10^{-15}	LOC729222
12	112703139 – 112738033	$rs11066758 - rs7137339$	$GGGGTCCC$	0.06128	0.04048	2.52574	2.3×10^{-18}	RBM19
12	114038450 – 114074493	$rs1828384 - rs35346$	$TGTACCTG$	0.09952	0.10526	3.07564	1.5×10^{-11}	TBX3
			$TCCAATTG$	0.04718	0.03821	4.01761	2.2×10^{-13}	
13	23708179 – 23726596	$rs881428 - rs2760374$	$AGAAGTTT$	0.12142	0.07922	1.89748	5.7×10^{-19}	SPATA13
			$GAAAGCTT$	0.2454	0.19993	1.51979	6.8×10^{-15}	
13	70170848 – 70209722	$rs17087430 - rs12876111$	$CGGGTTAT$	0.13996	0.13226	3.39099	3.0×10^{-14}	ATXN8OS
			$CGGGTCCT$	0.02217	0.01356	5.23473	1.9×10^{-14}	
			$CGGGTCAT$	0.13141	0.13526	3.11367	2.3×10^{-12}	
			$CGGACTCT$	0.04728	0.0398	3.8075	1.9×10^{-13}	
14	21674996 – 21704333	$rs12050442 - rs1894369$	$GGGGTTAC$	0.03075	0.00968	6.13598	8.7×10^{-19}	TRA@
14	36411583 – 36421982	$rs10872897 - rs2564848$	$ATCCACTT$	0.02299	0.00637	4.45891	8.9×10^{-16}	SLC25A21
			$TACCTCCC$	0.02712	0.01101	3.05584	1.8×10^{-12}	
16	4881048 – 4960784	$rs760117 - rs9937749$	$CTTCCCCA$	0.0847	0.08126	4.18237	1.8×10^{-12}	SEC14L5
16	17231173 – 17272606	$rs754067 - rs17277691$	$CGGACCCT$	0.02658	0.02179	3.37015	1.1×10^{-11}	XYLT1
17	69565860 – 69595387	$rs7406930 - rs8080915$	$CTGTACGC$	0.0413	0.02484	2.54279	8.3×10^{-14}	RPL38
19	3315188 – 3432578	$rs758257 - rs1860192$	$GTTTGATT$	0.27769	0.23516	1.99935	3.4×10^{-28}	NFIC
19	8475735 – 8540766	$rs2967603 - rs11259990$	$CCGCTCTT$	0.06824	0.04351	3.23984	2.1×10^{-17}	ZNF414
19	17595848 – 17649789	$rs10419511 - rs7252308$	$TTGGTGTG$	0.07791	0.05267	2.40001	1.9×10^{-23}	UNC13A
			$TTGGTATG$	0.04536	0.01971	3.72872	2.3×10^{-28}	
19	38822176 – 38857206	$rs2059876 - rs16968366$	$CAAATGCG$	0.06455	0.05252	2.83486	6.2×10^{-20}	CHST8
20	10019135 – 10038764	$rs552048 - rs670562$	$TATGAGGG$	0.04043	0.02307	7.32891	4.5×10^{-21}	ANKRD5
			$TATAAGAA$	0.03726	0.03549	4.39728	2.7×10^{-12}	
			$TGTGAGGG$	0.27299	0.294	3.88507	1.0×10^{-13}	
			$TGTATGGG$	0.19239	0.1808	4.4523	3.1×10^{-16}	