CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Kent Academic Repository

Screening tests for Disease Risk Haplotype Segments in Genome by Use of Permutation

Fadhaa Ali, Guohua Zou and Jian Zhang
University of Kent, Capitial Norm University, and University of Kent

Abstract

The haplotype 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. It starts with inferring haplotypes from genotypes followed by a haplotype co-classification and marginal screening for disease-associated haplotypes. Unfortunately, phasing uncertainty may have a strong effects on the haplotype co-classification and therefore on the accuracy of predicting risk haplotypes. Here, to address the issue, we propose an alternative approach: In Stage 1, we select potential risk genotypes instead of co-classification of the inferred haplotypes. In Stage 2, we infer risk haplotypes from the genotypes inferred from the previous stage. 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.

Some key words: Region-based association analysis; genotype mixture models; odds ratios; genome wide association studies; expectation-maximization algorithm.
Short title: Search for Disease Risk Haplotype Segments

1 Introduction

Advances in genotyping and sequencing technologies, coupled with the development of high-dimensional statistical methods, have provided investigators opportunities to reveal the role of sequence varia-

*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.

tion in the development of complex diseases. At the forefront of these investigations is genome-wide association studies (GWAS) by the use of dense maps of single-nucleotide polymorphisms (SNPs) and the haplotypes derived from these polymorphisms (Stranger et al., 2011). The early landmark study using the GWAS 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 (WTCCC, 2007). 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. In these studies, people identified variants that predispose to a disease by conducting association tests (i.e., marginal screening tests) on SNPs, one at a time. For the majority of complex diseases, it was found that single-SNP variants might explain only $< 10\%$ of disease variations as many variants showed only weak effects on the risk of disease and, therefore, a joint analysis of multiple SNPs might be necessary for understanding the etiology of complex diseases (Manolio et al., 2009). A popular strategy in the GWAS analysis, suggested by the block-like structure of the human genome, is to segment each chromosome into a list of genetically meaningful SNP 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 (Schaid et al., 2002; Zhang et al., 2003; van Greevenbroek et al., 2008; Li et al., 2011). A haplotype may affect phenotype directly through influencing promoter activity and protein formulation or indirectly through tagging nearby untyped causal variants. 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. Haplotype segments hold the promise of reducing the complexity of analyzing the human genome for association with disease. Identifying risk haplotype segments 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. Inferring haplotypes from observed genotypes by using the computational software such as PHASE is a popular strategy to overcome the uncertainty of genotype phases (Stephens et al., 2001; Scheet et al., 2006). In the PHASE, a coalescent model-based Gibbs sampling was employed to infer the most probable hyplotype pair for each individual in the sample, given all the possible haplotype pairs that are consistent with the observed genotypes. Existing haplotype methods improve the power of the association testing by grouping haplotypes before testing (Zöllner and Pritchard, 2005; Browning and Browning, 2007, and references therein). Zhu

et. al (2010) developed a two-stage screening procedure for GWAS data, which requires phasing to obtain haplotypes followed by grouping. Unfortunately, grouping inferred haplotypes may be affected by phasing uncertainty.

This paper aims to improve the above two-stage procedure by grouping genotypes (instead of haplotypes) before the association testing. For this purpose, we combine a genotype permutation technique with the PHASE procedure to form a basis for testing risk haplotypes. Our method relies on the observation that if a set of SNPs is not associated with disease (which is the null hypothesis), the permutated genotype frequencies can be employed to generate the null distributions of genotypes. Then, in Stage 1, for each genotype, we test its association with disease by checking whether the observed case-frequency is located in the tail areas of its null distribution. This provides a list of selected genotypes for further investigation in the next stage. In Stage 2, we calculate the corresponding PHASE-inferred haplotypes and their frequencies in cases and controls for the selected genotypes. The odds ratios (ORs) are calculated for these haplotypes. These haplotypes are further screened by the OR test. 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. (2010) in most cases. We apply the proposed method to the Coronary Artery Disease (CAD) and Hypertension (HT) data in the Wellcome Trust Case Control Consortium (WTCCC), identifying potential risk haplotypes for these diseases.

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 are given in the Appendix.

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. Let $N_0 = \sum_{j=1}^{J} N_{0j}$ and $N_1 = \sum_{j=1}^{J} N_{1j}$. We perform the PHASE on genotypes in controls and cases. Let (h_{j1}, h_{j2}) be the inferred haplotype pair for G_j. We also let $\mathbf{H} = \{h_k, 1 \leq k \leq K\}$ denote the distinct haplotypes inferred from $\mathbf{G}$, where $\mathbf{G} = \{G_j, 1 \leq j \leq J\}$ with haplotype counts $n_{0k}, n_{1k}, 1 \leq k \leq K$ in controls and cases respectively and with total counts n_0, n_1. Then, the respective frequencies of the genotype

G_j in the controls and cases can be estimated by $q_{0j} = N_{0j}/N_0$, $q_{1j} = N_{1j}/N_1$, respectively. The proposed method contains two stages, where we screen genotypes and halotypes respectively.

Stage 1 (Genotype screening based on permutation):

To perform the permutation on individual disease statuses between cases and controls, we randomly swap a half of cases with the same number of controls. We then calculate the corresponding frequencies of the resulting permuted cases, denoted by $q^*_{i1j} = N^*_{1j}/N_1$, where $i = 1, 2, 3, ..., I$ with I being the total number of permutations we conducted, and $N^*_{i1j}, 1 \leq i \leq I$ represent the counts of the genotype j in the permuted cases for the permutation i. In the later simulation and real data analyses, we choose $I = 1000$. Let $q^*_{1j} = (\sum_{i=1}^{I} q^*_{i1j})/I$ denote the average frequencies of genotype G_j over I permutations. Consider the following statistic for genotype G_j:

$$T_j = \frac{q_{1j} - q^*_{1j}}{\delta_j},$$

where

$$\delta_j = \sqrt{\frac{\sum_{i=1}^{I}(q_{1j} - q^*_{i1j})^2}{I-1}}.$$

Under the null hypothesis that G_j is not associated with disease, the statistic T_j is asymptotically distributed as a standard normal. Therefore, $\{T_j\}$ can be used to test for disease associated genotypes, finding a set of potential risk haplotyes as follows:

$$S_r = \{h : h \in \{h_{0j}, h_{1j}\}, 1 \leq j \leq J,\ T_j > \gamma\},$$

where γ is a pre-defined critical value after adjusting multiple testing effects.

Stage 2 (Haplotype screening based on OR testing):

We examine the frequency differences of the haplotypes in the set S in controls and cases to find the potential risk group. Let $|S|$ be the number of all different haplotypes in S_r. To calculate their OR values, we let $n_{0\bar{r}} = \sum_{h_k \notin S} n_{0k}$, $n_{1\bar{r}} = \sum_{h_k \notin S} n_{1k}, 1 \leq k \leq K$ denote the cumulative frequencies of the haplotypes not in S for controls and cases respectively. Then, the corrected OR values for the haplotype $h_\nu, 1 \leq \nu \leq |S|$ is calculated by

$$\mathrm{OR}_\nu = \frac{(n_{1\nu} + 0.5)(n_{0\bar{r}} + 0.5)}{(n_{0\nu} + 0.5)(n_{1\bar{r}} + 0.5)}.$$

Then, the set of risk haplotypes S_r is updated by

$$\mathbf{H}_r = \{h_\nu \in S : \mathrm{OR}_\nu \geq \exp(c_1 \phi(n_{0\nu}, n_{1\nu}, n_{0\bar{r}}, n_{1\bar{r}}))\},$$

where

$$\phi(n_{0\nu}, n_{1\nu}, n_{0\bar{r}}, n_{1\bar{r}}) = \sqrt{1/(n_{0\nu}+0.5) + 1/(n_{1\nu}+0.5) + 1/(n_{0\bar{r}}+0.5) + 1/(n_{1\bar{r}}+0.5)},$$

adding 0.5 to the OR for the continuity correction was suggested by Agresti (1999) and c_1 is a pre-specified critical value after adjusting multiple testing effects.

2.1 Multiple testing method

To compare the proposed method to the multiple testing procedure of Zhu et al. (2010), 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 $(r_{0k}^{(a)}, r_{1k}^{(a)})$, $1 \le k \le K$ 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:

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

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 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:

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

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 $\mathbf{H}_r$ and $\mathbf{H}_{\bar{r}}$ the estimated true risk and non-risk

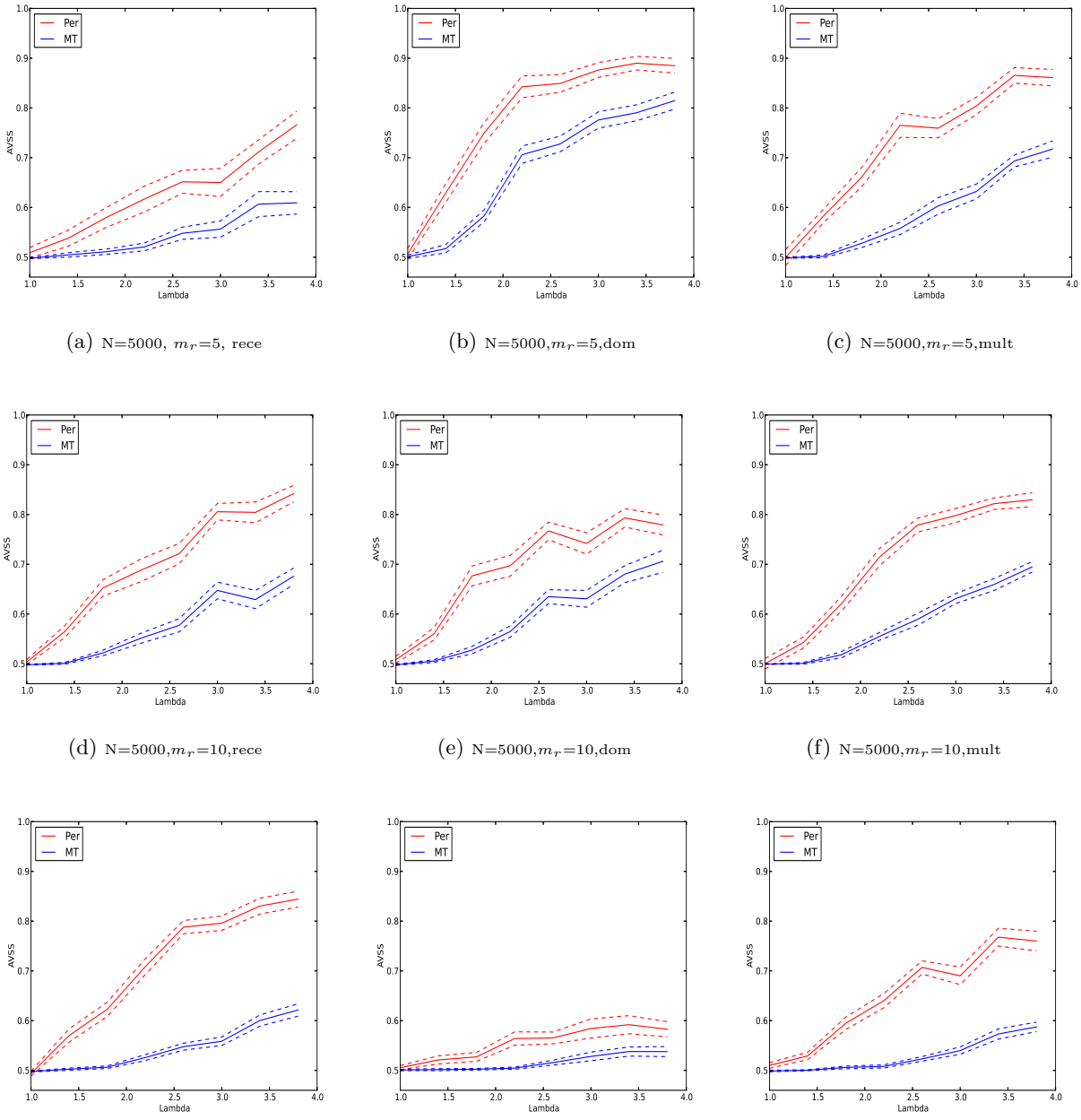
haplotype sets respectively. Let $\mathbf{T}_r$ and $\mathbf{T}_{\bar{r}}$ be the true risk and non-risk haplotype sets. Then, by the sensitivity and specificity of $\mathbf{H}_r$ and $\mathbf{H}_{\bar{r}}$, we mean the positive discovery rate and the negative discovery rate:

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

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

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. We used the software MS (Hudson, 2002) 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. We then randomly paired $2(N_0 + N_1)$ haplotypes, producing $N_0 + N_1$ genotype which contained m_r risk haplotypes. 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 $N_0 + N_1 = 3000, 5000$, $m_r = 5, 10, 20$, 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 figures 1 and 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 permutation-based genotype screening.

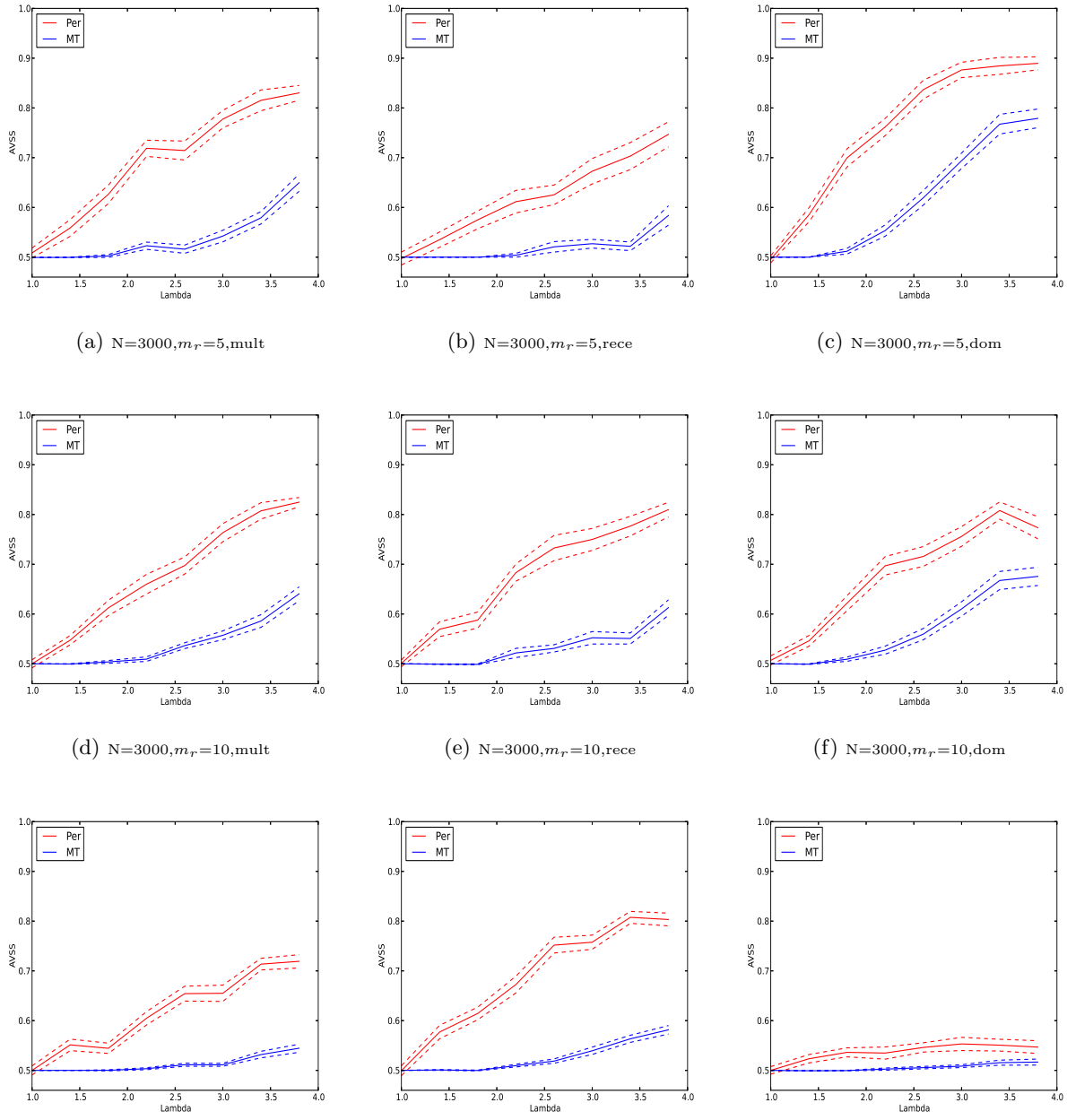


(a) N=5000, m_r=5, rece (b) N=5000,m_r=5,dom (c) N=5000,m_r=5,mult

(d) N=5000,m_r=10,rece (e) N=5000,m_r=10,dom (f) N=5000,m_r=10,mult

(g) N=5000,m_r=20,rece (h) N=5000,m_r=20,dom (i) N=5000,m_r=20,mult

Figure 1: Performances of the proposed permutation method and the multiple testing method on the cohort-design data with multiplicative or dominant or recessive inheritance models based on sample sizes of 5000.



(a) N=3000,m_r=5,mult (b) N=3000,m_r=5,rece (c) N=3000,m_r=5,dom

(d) N=3000,m_r=10,mult (e) N=3000,m_r=10,rece (f) N=3000,m_r=10,dom

(g) N=3000,m_r=20,mult (h) N=3000,m_r=20,rece (i) N=3000,m_r=20,dom

Figure 2: Performances of the proposed permutation method and the multiple testing method on the cohort-design data with multiplicative or dominant or recessive inheritance models based on sample sizes of 3000.

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

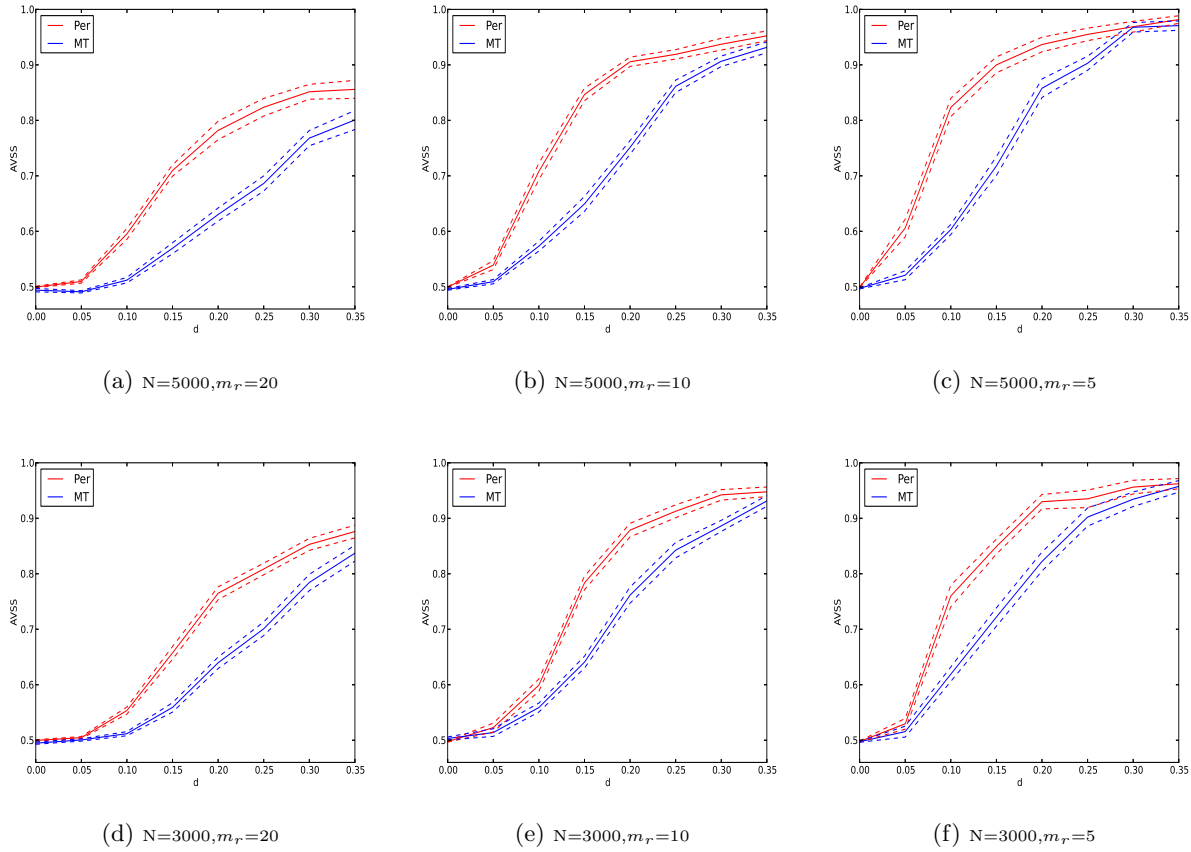
haplotypes. We then randomly paired these haplotypes to form N_1 case-genotypes. Let G_j, $1 \leq j \leq J$ be all the different genotypes contained in the N_1 cases and $r_{1j}, 1 \leq j \leq J$ be the retrospective frequencies. These case-genotypes formed three groups according to the number of risk 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 calculated 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.

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.



(a) N=5000,m_r=20 (b) N=5000,m_r=10 (c) N=5000,m_r=5

(d) N=3000,m_r=20 (e) N=3000,m_r=10 (f) N=3000,m_r=5

Figure 3: Performances of the proposed permutation method and the multiple testing method on the case-control data.

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 (WTCCC, 2007). 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. We used the total number of the genotypes to set the Bonferroni correction to the critical value in the permutation test. To achieve a significance level of 0.05 for all genotypes, the adjusted significance level was set to be $0.05/1983537 = 2.52 \times 10^{-8}$ and $0.05/2097111 = 2.38 \times 10^{-8}$ for the CAD data and the HT data respectively. This resulted in an approximate critical value of 5.5 for both the CAD and HT data. The genotype screening in Stage 1 resulted in 1433 potential risk haplotypes in the CAD data and 430 potential risk haplotypes in the HT data.

Note that there were two sub-populations in controls. We applied further filtering on the regions to exclude the ones that have significant differences in the haplotypes frequencies within the two sub-control samples. The exclusion criterion was based on calculating chi-square p-value. Any region resulted in p-value less than 0.30 was excluded from the suspicious regions. This criterion was concluded from the simulated case-control samples when the risk factor d is less than 0.15 as we found out that the p-values for most of the 30 datasets are greater than 0.30. The numerical details were omitted. We applied the above criterion on the above potential risk haplotypes and eliminated these haplotypes with the chi-square p-value being less than 0.30. In Stage 2, we calculated the OR values of the selected haplotypes and thresholded them by using the bounds

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

defined in the methodology section with $c_1 = 4$ and 3.6 for the CAD data and the HT data respectively. Note that the values of c_1 were determined by the Bonferroni correction according to the corrected significance levels of 0.05/1433 and 0.05/430 for the CAD and the HT respectively.

This gave the final risk-haplotype sets as displayed in Tables 1, 2, and 3 below. In these tables, each haplotype has been assigned to a physically closest gene on the basis of the information provided in the GWAS catalog (Welter et al., 2014) and the genetic information from the British 1958 Birth cohort *http://www2.le.ac.uk/projects /birthcohort/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). Note that by use of the multiple testing method, Zhu et al. (2010) identified the following genes (ZFAT1 and MACROD2 for HT; EIF4H, CDKN2B, HFE2, ZBTB43 and LDHA for CAD) reaching genome-wide significance. Therefore, the proposed method can be much more powerful than the multiple testing method in the identification significant genes (and SNPs) for association studies.

Table 1: The risk haplotypes for coronary artery disease of WTCCC data detected by permutation method.

CAD								
Chr	Region	SNP range	Haplotype	$P(H_i\|case)$	$P(H_i\|control)$	OR	P-Value	Gene
1	3910010 – 3932838	$rs4654522 - rs10915469$	$CGACGGCC$	0.04238	0.01861	3.09933	4.5×10^{-16}	$hCG_2036596$
1	1902751 – 37450147	$rs6673253 - SNP_A$	$CAACGGAT$	0.05116	0.03019	2.33902	3.0×10^{-14}	LOC728431
1	202166400 – 202187685	$rs6692041 - rs1041311$	$AAATGGGA$	0.07815	0.05083	1.72409	4.3×10^{-09}	LOC284577
1	225406446 – 225425470	$rs4654697 - rs10916399$	$TTGTAAAA$	0.06155	0.03524	1.85056	8.1×10^{-10}	RHOU
1	227569611 – 227620956	$rs7514972 - rs9431663$	$CGCGTAGG$	0.05807	0.0297	2.06768	2.2×10^{-12}	TRIM67
1	239380743 – 239454253	$rs2491826 - rs7533316$	$AGCTCACG$	0.09857	0.07858	1.63864	7.4×10^{-08}	CEP170
1	240360846 – 240438647	$rs12083813 - rs472276$	$CAACATAG$	0.01905	0.00712	2.94026	2.1×10^{-08}	AKT3
2	3789586 – 3821960	$rs7576476 - rs12618184$	$GCTTACAG$	0.03451	0.01119	3.14706	3.1×10^{-15}	LOC442006
2	$rs2314703$ – 3942429	SNP_A – 1841609	$CACGCCGT$	0.02055	0.00552	3.78775	3.3×10^{-11}	LOC442006
2	49934439 – 50000082	$rs6736617 - rs17039375$	$CCAAAGGT$	0.02347	0.00757	3.09136	2.7×10^{-10}	NRXN1
2	81525887 – 81577090	$rs1011364 - rs17020239$	$GGATGTGC$	0.03758	0.0202	1.96428	1.3×10^{-07}	LOC442021
3	2557255 – 2599938	$rs6787604 - rs2619566$	$AAGGACGA$	0.07666	0.04763	1.64989	3.1×10^{-08}	CNTN4
3	14422977 – 14471151	$rs4684216 - rs9834629$	$GATGATGC$	0.01815	0.00509	3.67773	8.7×10^{-10}	SLC6A6
3	73461569 – 73510299	$rs7647311 - rs3845868$	$AGGCGCGG$	0.03876	0.01161	3.98169	6.9×10^{-23}	PDZRN3
3	197256495 – 197339533	$rs6583286 - rs9834962$	$TAGACTTA$	0.0498	0.02364	2.17213	2.5×10^{-11}	TFRC
4	3636361 – 3700212	$rs10025237 - rs16844722$	$GGGGAGGG$	0.22491	0.15492	1.62473	6.4×10^{-15}	FLJ35424
4	167440772 – 167457521	$rs9995087 - rs17047336$	$GGACGCAG$	0.03434	0.01139	3.12327	8.2×10^{-14}	TLL1
5	124765522 – 124843518	$rs4836190 - rs13187198$	$TGAAGGCA$	0.04275	0.02795	2.02205	2.0×10^{-09}	LOC644659
5	157267571 – 157303032	$rs10071157 - rs17055168$	$GTGAGCAA$	0.02135	0.00701	3.09771	9.0×10^{-10}	CLINT1
5	166764561 – 166801933	$rs6863935 - rs7724862$	$CTATGTGT$	0.09145	0.05448	1.63602	8.8×10^{-09}	ODZ2
7	77695246 – 77717237	$rs2215379 - rs4515471$	$TCTAAAAA$	0.03291	0.01786	2.04961	1.7×10^{-07}	MAGI2
			$CTTGGAAA$	0.03609	0.01061	3.77003	7.3×10^{-19}	
7	153371858 – 153449397	$rs6464391 - rs1861139$	$CGGGTAGA$	0.04119	0.02159	2.31998	1.7×10^{-11}	LOC653748
8	71022178 – 71086937	$rs7836791 - rs388511$	$TACAGAAG$	0.02204	0.00555	3.68611	4.1×10^{-11}	SLCO5A1
9	22088619 – 22120515	$rs2891168 - rs10965245$	$GGTGCCAG$	0.34939	0.29298	1.40724	3.2×10^{-13}	CDKN2B
9	74180343 – 74241329	$rs10114124 - rs17081046$	$GTATTTAT$	0.21608	0.13046	1.66562	4.0×10^{-17}	RORB
9	77341767 – 77366988	$rs2889774 - rs3780296$	$ATGGAAAT$	0.06672	0.042	1.69537	1.2×10^{-07}	GNA14
9	119506057 – 119537035	$rs2191675 - rs10984648$	$GTTGGCTA$	0.08762	0.03361	2.8056	1.8×10^{-28}	CDK5RAP2
9	135269746 – 135320703	$rs731533 - rs7870302$	$TGTCTCCC$	0.03175	0.01296	2.57076	9.3×10^{-11}	OLFM1
10	11879196 – 11924252	$rs6602535 - rs11257355$	$TCTGCCGG$	0.1694	0.12811	1.57916	1.3×10^{-12}	C10orf47
10	14795325 – 14817082	$rs2688827 - rs12246518$	$ATGACCGC$	0.34815	0.32333	1.71018	4.1×10^{-09}	FAM107B
11	8165969 – 8200374	$rs4758310 - rs11041816$	$ATAATGGG$	0.36298	0.3164	1.34831	2.8×10^{-08}	LOC644497
			$GCTGTAGA$	0.05243	0.02741	2.24619	7.5×10^{-12}	
11	36361306 – 36410807	$rs330255 - rs331485$	$GCGATTAA$	0.0309	0.00779	4.20172	5.6×10^{-18}	FLJ14213
11	69213458 – 69295251	$rs1192923 - rs3168175$	$TCGTGGCA$	0.10225	0.05587	2.24141	5.7×10^{-21}	FGF4
11	83230307 – 83256927	$rs1878266 - rs1878264$	$TATATTCA$	0.03571	0.01807	2.24283	6.3×10^{-09}	CCDC90B
11	99383206 – 99391536	$rs3911286 - rs10501939$	$TTAGATAT$	0.03303	0.01472	2.21561	9.3×10^{-09}	CNTN5
11	112952870 – 113015533	$rs4936278 - rs12577253$	$CCTCGTGC$	0.05824	0.03474	1.75496	1.9×10^{-08}	DRD2
11	129102667 – 129124330	$rs532427 - rs691197$	$ACCGCGGA$	0.08519	0.05612	1.73953	2.1×10^{-11}	TMEM45B
11	133079508 – 133113640	$rs4937817 - rs4937826$	$CCGGCCCG$	0.05747	0.04018	1.89429	5.6×10^{-10}	LOC646522
			$GTAGCCCG$	0.04001	0.02779	1.90705	9.3×10^{-08}	
			$GTAGTGCC$	0.04216	0.02425	2.30133	8.2×10^{-12}	
12	5619429 – 5628923	$rs11063791 - rs454704$	$TACATAAA$	0.02897	0.0124	2.50152	8.0×10^{-10}	TMEM16B
12	112703139 – 112738033	$rs11066758 - rs7137339$	$ACGGTCAC$	0.02681	0.01286	3.14709	1.5×10^{-12}	RBM19
12	116500495 – 116514298	$rs10850852 - rs1400593$	$CTCTCTTT$	0.14523	0.12089	3.21401	8.3×10^{-21}	NOS1

Table 2: Continuation of Table 1.

CAD								
Chr	Region	SNP range	Haplotype	$P(H_i\|case)$	$P(H_i\|control)$	OR	P-Value	Gene
			$CTCTCTTC$	0.28034	0.26232	2.85847	1.5×10^{-19}	
13	108372995 – 108432811	$rs4773010 - rs3842945$	$AGAGACCC$	0.27486	0.19222	1.59282	1.3×10^{-21}	MYO16
14	25140850 – 25159405	$rs8020556 - rs1951062$	$AGTAAACT$	0.09084	0.02999	3.36068	1.2×10^{-37}	LOC401767
14	53221435 – 53244046	$rs1563719 - rs210351$	$AGATAGGT$	0.15385	0.10566	1.56278	1.2×10^{-12}	BMP4
14	65343491 – 65401760	$rs3924222 - rs12896836$	$TATAACTC$	0.0462	0.01904	2.70766	1.1×10^{-16}	FUT8
15	20624103 – 21246055	$rs7166056 - rs8024346$	$GTGACGTG$	0.08093	0.04109	1.90364	2.7×10^{-12}	NIPA1
15	21729952 – 21760003	$rs4778264 - rs9796712$	$TGATAGGG$	0.03064	0.00783	3.91789	2.2×10^{-16}	MAGEL2
15	37962389 – 38014169	$rs11633436 - rs534757$	$TTACAACC$	0.07798	0.03763	2.31448	1.1×10^{-18}	GPR176
16	55207138 – 55253047	$rs8055724 - rs12447986$	$TTCTCCTC$	0.03044	0.01113	2.89551	1.5×10^{-09}	MT1L
16	79852394 – 79892297	$rs6564863 - rs11639552$	$TTCGTTAT$	0.02663	0.01053	3.15992	2.7×10^{-10}	BCMO1
17	27921023 – 27963104	$rs225215 - rs17780520$	$GGGTTAAC$	0.0205	0.00465	4.05617	2.7×10^{-11}	MYO1D
17	74629176 – 74682195	$rs2612793 - rs8072667$	$CGAGGTTG$	0.06276	0.03471	1.82966	4.4×10^{-09}	FLJ21865
18	8212591 – 8279839	$rs10468776 - rs11876033$	$GGGACAAG$	0.02689	0.00982	2.94852	1.7×10^{-11}	PTPRM
18	2291328 – 22715430	$rs3974646 - SNP_A$	$TGCGGAGT$	0.05382	0.02751	1.98739	2.3×10^{-10}	AQP4
18	32033296 – 32083366	$rs8095718 - rs8082899$	$CAAAACCA$	0.0592	0.04484	1.65827	1.7×10^{-07}	MOCOS
19	6641966 – 6717213	$rs3745566 - rs7248911$	$TAAGCTAC$	0.02312	0.00521	4.97801	1.0×10^{-14}	C3
19	15365766 – 15477256	$rs7257156 - rs6512039$	$AAGCGCGG$	0.08169	0.05278	1.69741	1.1×10^{-09}	AKAP8L
19	17595848 – 17649789	$rs10419511 - rs7252308$	$TTGGTATG$	0.04657	0.01971	2.8095	1.1×10^{-17}	UNC13A
19	18225800 – 18277972	$rs10417536 - rs4808781$	$CTCCGCAA$	0.04034	0.02211	1.94095	6.7×10^{-08}	LOC729966
19	52946204 – 53026777	$rs10402957 - rs4427918$	$CATTCAGC$	0.0741	0.04321	1.81613	4.1×10^{-10}	GLTSCR2
20	5604763 – 5643174	$rs8118780 - rs805726$	$CCGTAGTA$	0.05455	0.03836	1.76976	1.3×10^{-08}	C20orf196
			$CTTTAGTA$	0.01801	0.00794	2.81211	2.7×10^{-08}	
			$CTTTAGTG$	0.01698	0.00777	2.7096	1.6×10^{-07}	
20	6055964 – 6078025	$rs6117090 - rs3897509$	$AGGCCGCA$	0.09945	0.05857	1.89101	9.9×10^{-13}	C20orf42
			$AAGCCGAA$	0.03039	0.01269	2.66015	1.2×10^{-09}	
20	51996013 – 52017348	$rs12480336 - rs6013853$	$CACCGATC$	0.02844	0.01511	2.17303	1.5×10^{-07}	BCAS1
20	55607831 – 55637003	$rs17498081 - rs17414380$	$CAATGTCC$	0.02768	0.01127	2.6821	1.2×10^{-09}	TMEPAI
22	16871076 – 16895136	$rs8142200 - rs975826$	$TCGGGAGG$	0.03219	0.00253	12.43113	5.4×10^{-28}	LOC729269
22	35324014 – 35335429	$rs7410412 - rs12160203$	$GCCTAGGG$	0.1967	0.14314	1.46774	4.7×10^{-11}	CACNG2

Table 3: The risk haplotypes for hypertension of WTCCC data detected by permutation method.

HT								
Chr	Region	SNP range	Haplotype	$P(H_i\|case)$	$P(H_i\|control)$	OR	P-Value	Gene
2	39199834 – 39248354	$rs6758330 - rs10184046$	$CGCCAAAA$	0.03665	0.00147	26.83195	1.3×10^{-31}	SOS1
4	17856580 – 17878437	$rs11941617 - rs1503880$	$GTATTTGT$	0.0584	0.00019	236.45945	1.2×10^{-73}	LCORL
6	107236669 – 107248636	$rs3121432 - rs2354550$	$TGATTGTC$	0.07759	0.00247	35.82646	6.5×10^{-82}	QRSL1
10	30990752 – 31024312	$rs16931828 - rs7078126$	$AGTGTTGC$	0.47318	0.47676	1.45455	1.0×10^{-08}	LOC645954
			$AACGTTGT$	0.06589	0.00314	29.93248	3.1×10^{-79}	
			$AGCTCTGC$	0.24167	0.24983	1.41785	1.2×10^{-06}	
			$GGCGCCGC$	0.10573	0.10377	1.49364	4.1×10^{-06}	
11	55290776 – 55324792	$rs11825590 - rs17501618$	$GCCTGTGT$	0.04351	0.00947	4.47895	4.1×10^{-22}	OR5D14
11	121093256 – 121139818	$rs92061 - rs4936651$	$AATGCTGG$	0.86672	0.79508	2.49843	1.4×10^{-30}	SORL1
18	73486971 – 73493301	$rs1553419 - rs4890980$	$TTGGGTTC$	0.03825	0.00893	4.49948	2.9×10^{-21}	LOC728864

5 Discussion and conclusion

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 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, by using multiple Z-tests (Zhu et al., 2010). 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 by use of permutation. In Stage 1, we selected the potential risk genotypes through a permutation technique, followed by estimating the potential risk haplotypes by using the software PHASE. In Stage 2, we refine the above selected risk haplotypes from the estimated risk genotypes by using the odds ratio thresholding.

We have conducted a wide range of simulations to compare our method to the multiple Z-test approach, demonstrating a substantial improvement can be achieved by use of the proposed method in terms of average sensitivity and specificity. We have also examined the performance of the proposed procedure by applying it to the CAD data and HT data in 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 the proposed method 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.

Appendix: PHASE

PHASE is a Bayesian haplotype reconstruction method developed by Stephens et al. (2001) 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.

Acknowledgments

We are grateful to the Wellcome Trust Case Control Consortium Data Access Committee for allowing us to use their WTCCC CAD and HT data. The research of the first author was funded by the Ministry of Higher Education and Scientific Research, Iraq.

References

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

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

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

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.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B. and at al. (2009). Finding the missing heritability of complex diseases. *Nature* , **461**, 747-753.

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.

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.

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.

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.

Van Greevenbroek1, 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.

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

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.*. **73**, 1385401.

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.

Zöllner, S. and Pritchard, J.K. (2005). Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics*, **169**, 1071-1082.