Statistics and Its Interface Volume 4 (2011) 273–283

Detecting association with rare variants for common diseases using haplotype-based methods

Yali Li*, Tao Feng* and Xiaofeng Zhu†

Current Genome-Wide Association Studies (GWAS) have successfully detected many genetic variants contributing to common diseases but not rare ones. Here two haplotype-based methods are proposed for detecting rare variants contributing a common disease. One method is a haplotype-based truncated product method (HTPM), for which we borrow a p-value combination method from testing for the multiple hypotheses, but use it for the purpose of clustering the information on rare risk haplotypes. The other method is the combined method, for which a set of risk haplotypes are chosen based on haplotype frequency comparison between cases and controls, and then testing for association using the same sample. Our simulation study demonstrates that both methods have improved power for detecting the association between rare variants and diseases, compared with other available methods. Both methods are applied to the Wellcome Trust Case Control Consortium (WTCCC) coronary artery disease and hypertension data and replicated the previous findings of genes associated with hypertension and coronary artery disease respectively at a genome-wide significance level of 5%. These results suggest that haplotype-based methods are powerful methods in searching for rare genetic variants and can be applicable to the data from current GWAS.

Keywords and phrases: Genome-wide association studies, Rare variants, Haplotype-based truncated product method, Combined method, Risk haplotypes.

1. INTRODUCTION

Genome-wide association studies (GWAS) have detected many common susceptibility genetic variants responsible for common diseases, with the underlying common disease common variant (CDCV) hypothesis. However, these common variants can only account for a limited fraction of the observed familial aggregation with modest odds ratios between 1.2 and 1.5 (Bodmer and Bonilla, 2008). There have been many cases that follow-up association studies failed to identify any associated single nucleotide polymorphisms (SNPs) in regions identified and confirmed by previous family-based linkage studies, although it has been argued that linkage analysis is less powerful than association analysis for identifying complex-disease genes (Risch and Merikangas, 1996). One emerging explanation for this deficit in follow-up association studies is the common disease multiple rare variants (CDMRV) hypothesis, which suggests the missing heritability for common diseases can be attributable to rare genetic variants with intermediate penetrance effects (Manolio et al., 2009).

Recent studies, mainly based on resequencing methods, have identified multiple rare variants for several common diseases. One remarkable finding is with breast cancer. There are ten genes accounting for inherited breast cancer and all those genes bear many rare mutations (Walsh and King, 2007). The accumulated evidence suggests that the high heterogeneity of inherited breast cancer can be at least partially explained by a CDMRV model. Other common diseases are also shown to have a similar pattern of inheritance. It has been reported that rare variants in three genes – SLC12A3, SLC12A1 and KCNJ1 – contribute to the reduction in blood pressure and protection from hypertension (Ji et al., 2008). Also, Cohen et al. (2004) have sequenced three genes – ABCA1, APOA1, and LCAT – and found that multiple rare genetic variants in the coding regions significantly contribute to low plasma HDL cholesterol level. In addition, multiple rare variants have been reported to be associated with metabolic phenotypes (Romeo et al., 2007) and plasma angiotensinogen level (Zhu et al., 2005).

Li and Leal (2008) have developed the Combined Multivariate and Collapsing (CMC) method, which first collapses genotypes across rare variants and then applies multivariate test, e.g., Hotelling's T^2 test, to the collapsed groups. This method can be applied to analysis of sequence data. However, under current GWAS design, CMC will be just similar to multiple-marker test since only common variants are directly genotyped and thus have a decreased power. Haplotype-based analysis may provide a better solution because it has been theoretically proven to be more powerful compared to single SNP analysis, based on accurate haplotype frequency estimates (Zaitlen et al., 2007). There have been some studies successfully detected rare variants using haplotype analyses, including a finding of two rare haplotypes having significant effects on the osteoporosis phenotype (Liu et al., 2005) and a report on detection of rare variants contributing to variation in angiotensinogen levels

*These authors contribute equally.
†Corresponding author.

(Zhu et al., 2005). However, methods based on individual haplotype analysis still face the problem of low power and thus require large sample sizes to detect rare variants. Zhu et al. (2010) have developed the two-stage method which co-classify rare risk haplotypes together and test the co-classified haplotypes as a set to improve the power to test association. However, the proposed methods are based on two stages, which raises the question of optimal sample sizes for each stage given a fixed total sample size and thus needs further investigation. In order to take the advantage of using haplotypes to capture rare variants and avoid allocating samples into two stages at the same time, two haplotype-based methods are developed: the Haplotype-based Truncated Product Method (HTPM) and the combined method. Both methods are shown to be efficient and powerful.

2. METHODS

2.1 Haplotype-based truncated product method (HTPM)

In a candidate gene or a genomic region, assume there are m different haplotypes $h_1, h_2, \ldots, h_m$ with corresponding haplotype frequencies $\boldsymbol{q} = (q_1, q_2, \ldots, q_m)'$ in the disease population and $\boldsymbol{q^0} = (q_1^0, q_2^0, \ldots, q_m^0)'$ in the control population, with $\sum_{i=1}^m q_i = 1$ and $\sum_{i=1}^m q_i^0 = 1$. Since we are detecting rare variants that increase the disease risk, we are interested in testing the one-sided hypothesis for each haplotype:

$$H_0 : q_i - q_i^0 = 0; \quad H_a : q_i - q_i^0 > 0 \quad (i = 1, 2, \ldots, m).$$

Assume a sample of N_1 cases and N_2 controls is considered and the observed counts of haplotypes in cases and controls are $\boldsymbol{X} = (x_1, x_2, \ldots, x_m)'$ and $\boldsymbol{Y} = (y_1, y_2, \ldots, y_m)'$, respectively. For each haplotype, a 2×2 contingency table can be constructed to compare haplotype frequencies between cases and controls. Usually a chi-square test is used in this situation. However, we decided to apply Fisher's exact test here because some expected values in the table may be small due to rare haplotype counts.

We then apply the truncated product method (TPM) to combine p-values from individual haplotype tests. The test statistic is constructed by taking the product of all the p-values smaller than a fixed value τ:

$$W = \prod_{i=1}^{m} (p_i)^{I(p_i \le \tau)}, \text{ where } I(\cdot) \text{ is the indicator function.}$$

Under the null hypothesis, the distribution of W for $w < 1$ can be evaluated by conditioning on k - the number of the p-values that are less than τ:

$$\begin{aligned}\Pr(W \le w) &= \sum_{k=1}^{m} \Pr(W \le w | k) \Pr(k) \\ &= \sum_{k=1}^{m} \binom{m}{k} (1-\tau)^{m-k} \left(w \sum_{s=0}^{k-1} \frac{(kln\tau - lnw)^s}{s!} \right. \\ &\quad \left. \times I(w \le \tau^k) + \tau^k I(w > \tau^k) \right).\end{aligned}$$

The TPM method was originally developed for combining independent p-values. However the single haplotype tests described above are correlated. To deal with non-independent tests, the empirical distribution of HTPM statistic W was estimated by permutation tests.

2.2 The combined method

Similarly to the two-stage method, the combined method involves first co-classifying rare risk haplotypes and then detecting association by comparing the frequency of classified haplotypes between cases and controls. The difference is the combined method uses the same sample for both co-classification and association testing.

The rare risk haplotypes are co-classified by defining the rare risk haplotype set as $S = \{h_i | \text{Fisher exact test p-value of haplotype } h_i < \tau\}$, where τ is a predefined value. The frequency of co-classified risk haplotype set S is then compared between cases and controls. The empirical p-value is estimated by permutation tests.

2.3 Two-stage method

Two-stage method first co-classify rare risk haplotypes using cases or affected sibpairs and then test association by comparing the frequency of classified haplotypes between cases and controls (Zhu et al., 2010). For the first stage, the rare risk haplotypes are co-classified by defining the rare risk haplotype set as $S = \{h_i | q_i - q_i^0 > \mu \sqrt{\frac{q_i^0(1-q_i^0)}{2N}}\}$, where N is the number of a subgroup of cases used for co-classification; q_i and q_i^0 are the frequencies of rare risk haplotype h_i in cases and the population, respectively; μ is a predefined constant. Similarly, the rare risk haplotype set can be define by using affected sibpairs: $S = \{h_i | q_i - q_i^0 > \mu \sqrt{\frac{q_i^0(1-q_i^0)}{3N}}\}$, where N is the number of sibpairs used for co-classification; q_i and q_i^0 are the frequencies of rare risk haplotype h_i in sibpairs and the population, respectively; μ is a predefined constant. q_i^0 can be estimated from control samples under both situations. For stage 2, the association between haplotype subset S and disease is tested by comparing the frequency of S between cases and controls in the rest of sample.

2.4 CMC method

Variants with minor allele frequencies (MAFs) $\le$0.001 are collapsed in CMC method. An indicator variable X is defined for the j^{th} case as $X_j = \begin{cases} 1 \text{ rare variants present} \\ 0 \text{ otherwise} \end{cases}$,

Y_j is defined in a similar way for controls. Variants with MAFs >0.001 are not collapsed. Assume in one gene or genomic region, M variants are collapsed into K groups, a multivariate Hotelling's T^2 test is then applied for the analysis of the K groups of genotype data.

2.5 Simulation

We applied the same method as Zhu et al. (2010) to generate cases and controls. Briefly, the haplotype frequencies in the ACE gene from a previous hypertension study in African samples were obtained (Zhu et al., 2001). There were 13 polymorphisms genotyped in this gene, resulting in a total of 149 different haplotypes with frequencies ≥0.01%. A total of 8 rare haplotypes, with frequencies in the range 1.0%–1.5% and with a cumulative risk haplotype frequency of 10%, were set to be risk haplotypes with the assumption that their effects on the phenotype are the same, i.e. the penetrance is only dependent on how many risk haplotypes an individual carried. An individual's genotype was simulated by randomly drawing two haplotypes according to the haplotype frequencies and the disease status was simulated based on the three modes of inheritance (dominant, recessive, and multiplicative). There were 1900 cases and 3000 controls simulated and the total sample size was approximately equivalent to the WTCCC study.

To assess the type I error, a null model was simulated that no risk haplotype was assigned. To assess the power, we assigned 8 rare haplotypes as risk haplotypes, as described above. The type I error and power were calculated as the proportion of 1000 simulations that resulted in rejection of the null hypothesis. To evaluate the performance of the methods when haplotype phase is unknown, the haplotypes were inferred from genotypes by Beagle 3.1 (Browning and Browning, 2007), a software package that can efficiently infer haplotype phases for genome-wide SNP data sets with a reasonable accuracy based on a localized haplotype cluster model (http://faculty.washington.edu/browning/beagle/beagle.html).

To compare haplotype-based methods with CMC method, which is designed for resequencing data, HapMap ENCODE data of one genomic region for the four HapMap samples were downloaded. The HapMap ENCODE resequencing project was dedicated to provide dense genotypes which will result in the knowledge of a comprehensive catalogue of human genomic components. Ten genomic regions of 500 kilobases were resequenced in 48 unrelated individuals (16 Yoruba, 8 Japanese, 8 Han Chinese, and 16 CEPH). All newly identified SNPs and SNPs that previously existed in dbSNP were genotyped in the 269 HapMap DNA samples (90 Yoruba, 44 Japanese, 45 Han Chinese, and 90 CEPH). Region ENm010 has the shortest haplotype length and thus is used in the simulation study to compare haplotype-based methods and the CMC method. Haplotypes of the individuals were inferred by Beagle 3.1. A total of 55 rare haplotypes, with frequency <1% and with a cumulative risk haplotype frequency of 10%, are chosen from the 529 haplotypes to be risk haplotypes, with the assumption that their effects on the phenotype is the same. Similar simulation strategies as described in section 3.1 were applied. There were 1900 cases and 3000 controls simulated and 1000 simulations were used to compare type I error and power.

2.6 WTCCC data analysis

The Wellcome Trust Case Control Consortium (WTCCC) coronary artery disease (CAD) and hypertension (HT) data (WTCCC, 2007) were downloaded from the WTCCC website. The individuals excluded in the WTCCC study were also excluded in our analysis, resulting in 1952 HT cases, 1926 CAD cases and 2838 controls, respectively. We applied the same criteria as the WTCCC study for SNP exclusion, except that we kept all the SNPs with minor allele frequencies <1%. The HTPM and combined methods were applied to WTCCC HT and CAD dataset for a subset of genes identified by using the 2-stage method previously.

3. RESULT

3.1 Evaluation of type I error

The type I error is evaluated for the HTPM method and combined method and compared to those of two-stage method (Zhu et al., 2010), at significance levels of 0.05 and 0.01, respectively (Table 1). The type I error is well controlled for HTPM. When the haplotype phase is known, the 95% confidence interval of type I error is (0.0347, 0.0608) at a significance level of 0.05 and (0.0047, 0.017) at a significance level of 0.01. The observed type I error rate for both HTPM method and combined method falls within the 95% confidence region. When the haplotype phase is unknown, the type I error of HTPM still falls in the 95% confidence interval, and so did that of the combined method. Two-stage method show reasonable type I errors as well, because the significance levels of 0.05 and 0.01 are within the 95% confidence intervals of (0.034, 0.072) and (0.0091, 0.0246), respectively.

3.2 Power of haplotype-based methods

The power of the HTPM and combined method is compared with the power of two-stage method, under different modes of inheritance and genotypic relative risk (Figure 1). Single SNP analysis is also performed, by comparing the allele frequencies between cases and controls. The minimum p-value for testing the set of markers was corrected by the estimated effective number of tests or the number of independent tests. For the two-stage method, two designs were used, affected sibpair and unrelated cases, in the co-classification stage. According to the power analysis comparing the different sample sizes used in the first stage (co-classification stage) and second stage (testing stage), designs with 800

Table 1. Type I error rate for simulation data analyzed as haplotype phase known and phase unknown. Genotypes of 13 polymorphisms in ACE gene were simulated for 1900 cases and 3000 controls. For each individual, the genotype was simulated by randomly drawing two haplotypes according to the haplotype frequencies of 149 haplotypes obtained from previous hypertension study in African samples (Zhu et al., 2001). To assess the type I error, a null disease model was simulated that no risk haplotype was assigned

	Two-stage method (Zhu et al., 2010)			
Type I error rate	Affected sibpairs	Unrelated cases	HTPM*	Combined method
Phase known				
0.05	0.048	0.047	0.046	0.053
0.01	0.009	0.011	0.009	0.012
Phase unknown				
0.05	0.056	0.046	0.048	0.046
0.01	0.015	0.012	0.009	0.009

*HTPM: haplotype-based truncated product method

cases or 400 affected sibpairs in first stage have the best power (Zhu et al., 2010). Therefore, 800 cases and 400 affected sibpairs are used in the first stage of the two–stage method in this simulation study.

In general, single SNP analysis has virtually no power, no matter which mode of inheritance is considered. Under dominant models, there is a large increase in the power for all the methods except single SNP analysis, when genotypic relative risk rises from 1.2 to 2. The power approaches to 1 when the genotypic relative risk rises above 2. For the HTPM method and the combined method, the power is greater than the two-stage method. The multiplicative model shows a similar pattern, only with a slower increase in power when genotypic relative risk rises. In general, the recessive model doesn't show much power except when genotypic relative risk is as large as 3.

Figure 1 shows the results when haplotype phase is known. However, it is difficult to acquire the phase information in practice. Therefore, the situation of unknown haplotype phase is considered and phase is inferred by software Beagle 3.1. The power of the HTPM and the combined method is then compared against the two-stage method (Figure 2). The power is slightly compromised when the haplotype phase is inferred, compared to the situation where we know the haplotype phase. However, the HTPM and the combined method still show reasonable power when genotypic relative risk is above 1.5 for the dominant model and above 2 for the multiplicative model. Overall, the power of the HTPM and the combined method is greater than that of the two-stage method. Under the multiplicative model, the two-stage method has substantially lower power than both the HTPM and the combined method, when the co-classification of rare haplotypes is performed with unrelated cases. Co-classifying rare haplotypes in affected sibpairs still have reasonable power, under both dominant and multiplicative modes. When the haplotype phase is unknown and inferred, the combined method shows substantially better power than HTPM, especially under the multiplicative model.

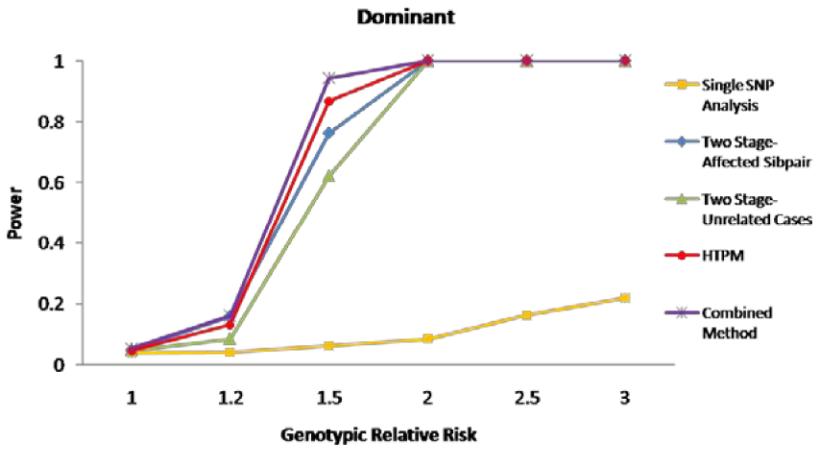
3.3 Comparison of CMC and haplotype-based methods

The CMC method was originally designed for application to analysis of resequencing data. To perform a fair comparison of haplotype-based methods and the CMC method in term of efficacy and power, the HapMap ENCODE resequencing data (ENCODE, 2004) is used in our simulation studies.

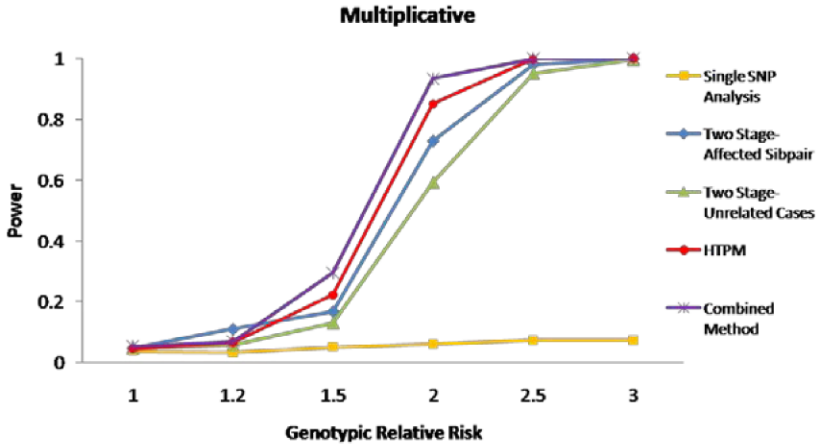
The CMC method is based on multivariate tests of collapsed groups. However, only rare variants are collapsed, while common variants are still included as one variant per group in the multivariate analysis. In this particular simulation study, the total number of variants is 808 within a 500 kb region, with the number of rare variants varying from 180 to 220 for different repeat of simulation. Therefore, the number of variables involved in multivariate analysis can be as large as 600, with many of them highly correlated. A large degree of freedom for the test statistic of Hotelling's T^2 test can thus be expected. In an initial analysis including all the common variants, the type I error of the CMC methods shows an unreasonable value of 0. The explanation for the abnormal value of type I error exhibited by CMC method lies in the large number of variables included in multivariate analysis, which may affect the validity of the null distribution assumption of test statistics. A similar CMC test has been applied with only 1 common variant and 1 group of collapsed rare variants included. The results showed a reasonable type I error of 0.048 at a significance level of 0.05. Above evidence suggested that the CMC method fails when a large number of non causal high-frequency variants are included in the analysis.

To make a fair comparison with the haplotype-based methods, a CMC test with 30 randomly chosen common variants included was performed. Table 2 compares the type I error of haplotype-based methods with the CMC method. The two haplotype methods – HTPM and the combined methods – can control type I error reasonably well. The type I error of the CMC methods is reasonable for both significance levels of 0.05 and 0.01.

a.



b.



c.

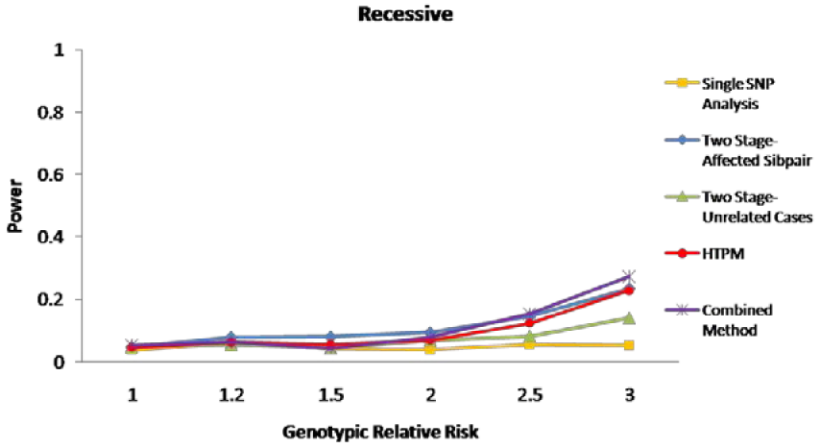
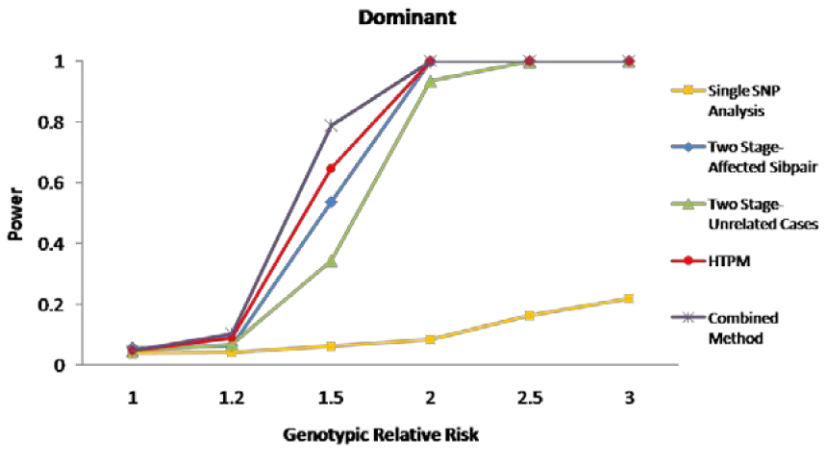


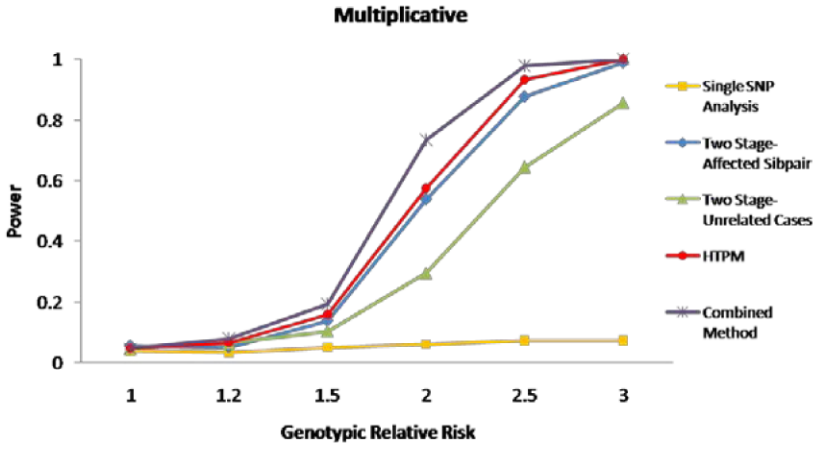
Figure 1. Power comparison of single-SNP analysis, two-stage method, HTPM and combined method, when haplotype phase is known. Power is plotted against genotypic relative risk at 1, 1.2, 1.5, 2, 2.5, and 3. Penetrance is simulated as 10%.

The power of haplotype-based methods has also been compared to the CMC method (figure 3). Under each genetic model, HTPM and the combined method have very similar power. Both HTPM and combined methods showed reasonable power under additive model. Lower power is observed in the multiplicative model compared to the additive

a.



b.



c.

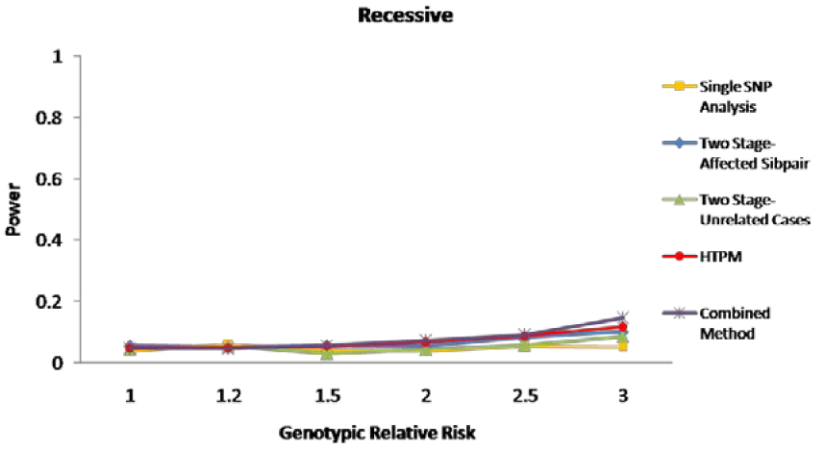


Figure 2. Power comparison of single-SNP analysis, two-stage method, HTPM and combined method, when haplotype phase is unknown. Power is plotted against genotypic relative risk at 1, 1.2, 1.5, 2, 2.5, and 3. Penetrance is simulated as 10%.

Table 2. Type I error rate comparison of haplotype-based methods and the Combined Multivariate and Collapsing (CMC) method using simulations based on ENCODE re-sequencing data. Genotypes of 808 polymorphisms in a genomic region of 500 kb were simulated for 1900 cases and 3000 controls. For each individual, the genotype was simulated by randomly drawing two haplotypes according to the haplotype frequencies of 529 haplotypes obtained from the HapMap ENCODE resequencing project (ENCODE, 2004). To assess the type I error, a null disease model was simulated that no risk haplotype was assigned

Type I error rate	CMC (Li and Leal, 2008)	HTPM*	Combined method
0.05	0.043	0.057	0.057
0.01	0.012	0.014	0.016

*HTPM: haplotype-based truncated product method

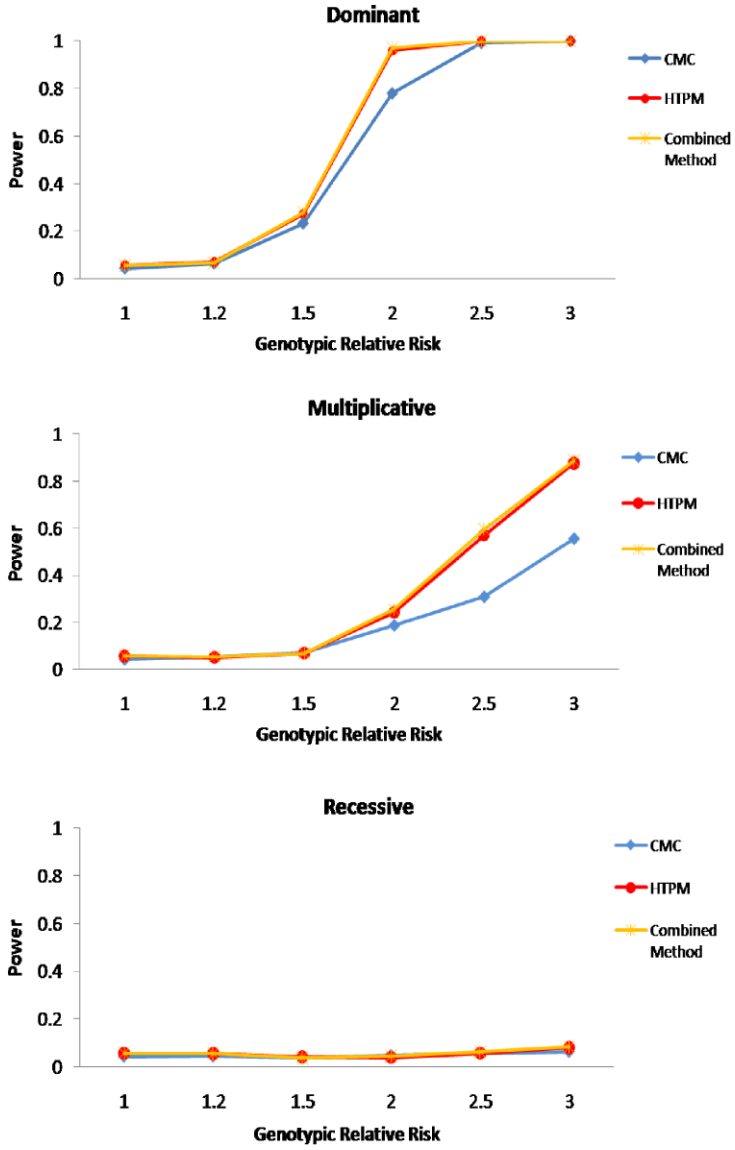


Figure 3. Power comparisons of CMC method, HTPM and combined method using simulations based on ENCODE re-sequencing data.

model, and the recessive model has virtually no power, as expected.

Under the additive model, the power of haplotype-based methods is lower for the ENCODE data based simulation study when genotypic relative risk equals to 1.5 compared to the previous simulation (shown in figure 2), but reaches over 90% when the genotypic relative risk rises to 2. ENCODE-based simulation considered much larger numbers of SNPs

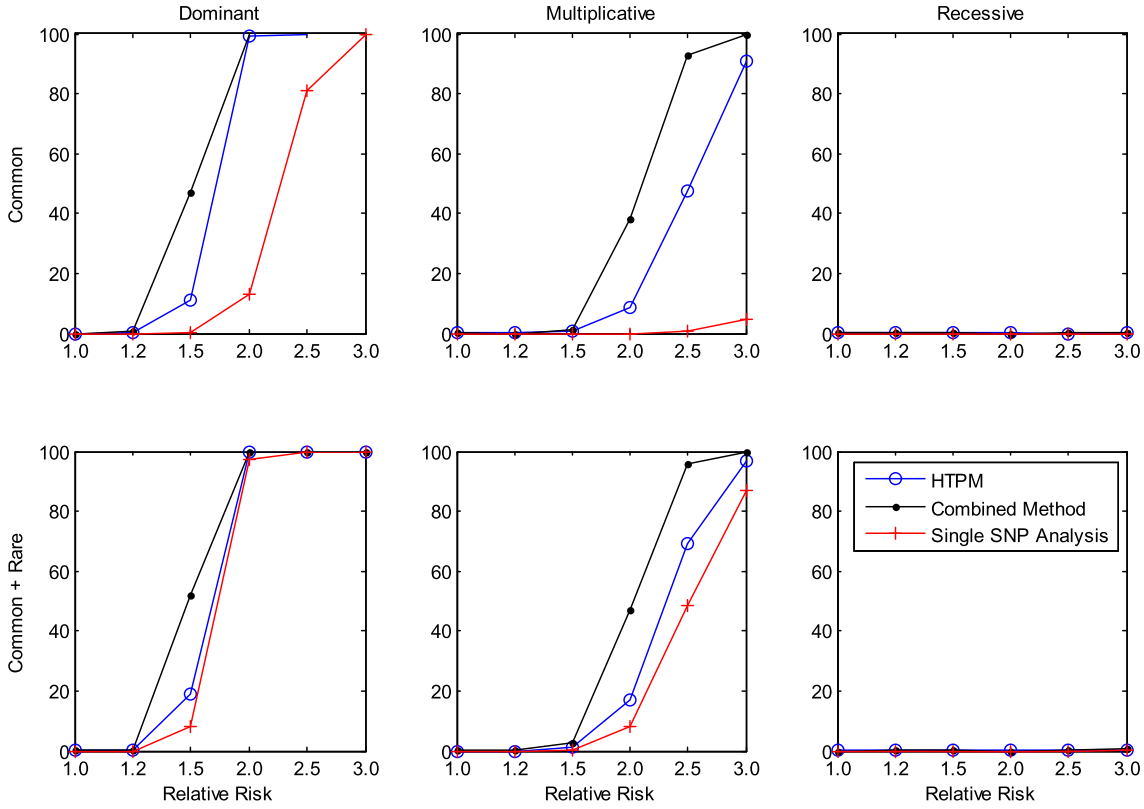


Figure 4. Power comparisons of single SNP analysis, HTPM and combined method using simulations based on ACE gene data when using common haplotype only (up panel) or using common and rare haplotype together (down panel) as risk haplotype.

(808 SNPs) than the previous simulation (13 SNPs). The accuracy of haplotype inference may decrease when the number of SNPs increases greatly, which may be a reason for the observed reduced power.

The power of the CMC method is consistently lower than the power of haplotype-based methods. The difference between CMC method and haplotype-based methods increases as the genotypic relative risk increases, under both the additive model and the multiplicative model.

3.4 Power when common variants only, or a mixture of common and rare variants as the risk variants

Previous we only simulated the disease models with multiple rare risk haplotypes. We now considered the situations when common haplotypes, or a mixture of common and rare haplotypes as the risk haplotypes contributing a disease. We simulated the following two scenarios: 1) only a common haplotype with its frequency 0.088 as the risk haplotype; 2) a common haplotype with its frequency 0.061 together with rare haplotypes as the disease risk haplotypes and the accumulated frequency is 0.117. The power of single SNP test, HTPM method and combined method was presented in figure 4. In general, we observed that HTPM method and combined method have better power than the single SNP analysis, which is consistent with we observed above.

3.5 Application to WTCCC HT and CAD data

Using two-stage method, Zhu et al. (2010) observed 3 genes associated with CAD (EIF4H, HFE2 and CDKN2B) and 1 gene (ZFAT1) associated with HT at a genome-wide significance level (p-value $\leq 10^{-6}$). In addition, PSRC1 was identified to be associated with CAD with a moderate signal (p-value $\leq 10^{-4}$). Those results provided the rationale that multiple rare variants may contribute to the variation of hypertension and CAD. Therefore, haplotype-based methods are now applied to the WTCCC CAD and HT data to confirm the results of the two-stage method.

Table 3 summarizes the HTPM and combined method test results of genes identified previously using a two-stage method in the WTCCC HT and CAD data. For CAD data, the two-stage method identified HFE2 on chromosome 1, EIF4H on chromosome 7 and CDKN2B on chromosome 9, with p-value $\leq 10^{-6}$. The results of both the HTPM and combined methods have replicated the findings of those three genes with p-value smaller than 10^{-6}, except that HTPM has a p-value of 3×10^{-5} for the CDKN2B gene. The two-stage method also identified PSRC1 on chromosome 1

Table 3. List of genes showing association to hypertension (HT) or Coronary artery disease (CAD) in WTCCC data

Disease	Chr	Gene	Range (MB)	Two-stage p-value	HTPM* p-value	Combined method p-value
CAD	1	HFE2	144.11–144.15	<1E-06	<1E-06	<1E-06
CAD	1	PSRC1	109.62–109.62	1.60E-05	1.30E-05	1.40E-05
CAD	7	EIF4H	73.20–73.25	<1E-06	<1E-06	<1E-06
CAD	9	CDKN2B	21.99–22.12	1.00E-06	3.00E-05	<1E-06
HT	8	ZFAT1	135.57–135.67	<1E-06	<1E-06	<1E-06

*HTPM: haplotype-based truncated product method

of a moderate effect with p-value 1.60×10^{-5}, which was also identified by HTPM and combined methods with p-values of 1.30×10^{-5} and 1.40×10^{-5}, respectively.

For the HT data, the two-stage method has identified ZFAT1 on chromosome 8 with p-value $\leq10^{-6}$. The results of both the HTPM and combined methods replicated this finding with p-values smaller than 10^{-6}.

4. DISCUSSION

Two haplotype-based methods have been developed to test association of rare variants with common diseases. Both simulation studies and the WTCCC data application demonstrated that the haplotype-based methods have reasonable power to detect rare variants, with well-controlled type I error. The single SNP analysis generally shows no power in detecting rare variants.

The methods are developed based on haplotypes rather than genotypes because haplotypes may capture ungenotyped rare variants in current genome-wide studies. Ideally, we wish the haplotype phase is known. However, in practice haplotype phase has to be inferred most of the time. As shown in the results, power decreases when phase is inferred by software Beagle 3.1, comparing to the situation where phase is known. However the loss in the power is not substantial. Beagle 3.1 is based on the localized haplotype-cluster model and was used in this study because it is fairly accurate in inferring haplotypes at reasonable computation cost. As suggested by Browning and Browning (2007), Beagle 3.1 outperforms the other existing software such as HaploRec, 2SNP and HAP. We simulated two data set using HapMap haplotype frequencies in CEU and YRI samples. The dataset using CEU data includes 18 haplotypes and the other using YRI data includes 59 haplotypes. In each dataset we simulated 2,000 unrelated individuals. We next evaluated the performances of Beagle and fastPHASE. For the dataset using CEU, the performances of Beagle and fastPhase are similar, with 98.6% and 98.5% haplotypes being inferred correctly by Beagle and fastPhase, respectively. For the dataset using YRI, the accuracy remains for Beagle (94.5%) but becomes worse for fastPhase (88.6%). We thus used Beagle 3.1 to infer haplotypes in this study.

The power of the HTPM and the combined methods is superior to that of the two-stage method, no matter whether the haplotype phase is known or unknown. The two-stage method allocates the sample into two independent parts, for a co-classification stage and an association testing stage, respectively. For the co-classification stage, we applied one-side test because our methods try to classify the risk haplotypes together. Similarly, we can classify all protective haplotypes together when we are interested in searching protective haplotypes. The power of detecting association may be decreased due to a smaller sample size used in the association test. With a fixed total sample size, the optimal sample size for each stage needs to be determined. Therefore, the HTPM and the combined methods have the unified advantage of an increased power by using the entire sample in the association test.

The combined method has a greater power than the HTPM when the haplotype phase is unknown and inferred using Beagle 3.1. For the same dataset, the HTPM and the combined methods identify the same set of risk haplotypes. The difference between the two methods is that HTPM uses the product of p-values while the combined method combines the frequencies of haplotypes and then conducts the association test by comparing the combined haplotype frequencies between cases and controls. As shown in the simulation study based on ENCODE data with many more SNPs, the difference in the power between the HTPM and the combined methods is negligible. Since the combined method has less computation burden, it is preferred in practice.

Here the haplotype-based methods are developed for current GWAS design. However, the methods can be applied to sequence data as well. Long-range haplotype information provided by next-generation sequencing data and the 1000 genomes project will offer a significant advantage over SNP data in detecting rare variants, especially for accurately inferring haplotypes (The 1000 Genomes Project Consortium, 2010). In the simulation study based on ENCODE re-sequencing data, the HTPM and the combined methods both show greater power than the CMC method. The simulation study is based on a region of 808 variants and the number of collapsed rare variants with an allele frequency <1% is around 200. The rest large number of variants is not collapsed and thus contributes to a large degree of freedom for the CMC method. The power of CMC method is compromised when many variants are not collapsed and thus are included individually in the multi-marker test. However, the power of the HTPM and combined methods is still reasonable with long-range haplotypes. Another advantage of

haplotype-based method over CMC method is that CMC method can't be applied to current GWAS design, where rare variants are not genotyped.

A drawback of the HTPM and combined methods is that they are both computationally intensive, since permutation tests are required to determine empirical p-values. Limited by computation speed, currently it is not practical to apply these haplotype-based methods on a genome-wide scale. However, they can still be used in the situation where candidate genes are identified by the two-stage method. A possible solution to the computation time problem is to allow the number of permutations to change dynamically when applied to GWAS data, in a way that the number of permutation tests varies depending on how many rejections have been observed. It should be noted that the co-classification of very rare haplotypes in our methods will suffer sampling error, winner's curse and false negative when the sample size is limited. These issues are common for statistical methods detecting rare variants. We suggest our methods are only suitable for detecting rare haplotypes with MAFs at least 0.5% in order to have a reasonable power for the typical sample sizes of current GWAS, i.e 2,000 cases and 2000 controls, respectively. When the sample size is large, we expect rare variants with MAFs >0.5% can be reasonably well represented by haplotypes (The 1000 Genomes Project Consortium, 2010).

Currently the haplotype-based methods are developed for a dichotomous trait. However, the methods can be easily adapted to be applied to continuous traits. A thought is to apply ANOVA test to each haplotype and then combine individual p-values together in a similar way as described in previous sections. The individual haplotypes can be grouped into 3 groups (risk, protective, non-risk/non-protective) at the minimum within-group difference and the maximum between-group difference of the trait. Our methods have not considered incorporating any covariates, which is important in practice. However, the combined method can be extended to incorporate covariates. For example, we can code individual's haplotypes as 0, 1 or 2 according to how many risk haplotypes he/she carries. In this way, we can apply the logistic regression adjusting for covariates. This method warrants further studies.

In summary, we have developed two haplotype-based methods that are more powerful than the two-stage method we developed before. The methods we developed can be useful to identify rare variants underlying complex traits.

ACKNOWLEDGMENTS

We thank two anonymous reviewers for their constructed comments. We thank Dr. RC Elston for his constructed reading and comments of the manuscript. The work was supported by the National Institutes of Health, grant numbers HL086718 from National Heart, Lung and Blood Institute, HG003054 from the National Human Genome Research Institute. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk/info/participants.shtml. Funding for that project was provided by the Wellcome Trust under award 076113. The authors declare that they have no competing financial interests.

Received 10 June 2010

REFERENCES

1000 Genomes Project Consortium, Durbin, R. M., Abecasis, G. R., Altshuler, D. L., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E. and McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature* **467**(7319), 1061–1073.

Bodmer, W. and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics* **40**(6) 695–701.

Browning, S. R. and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics* **81**(5) 1084–1097.

Cohen, J. C., Kiss, R. S., Pertsemlidis, A., Marcel, Y. L., McPherson, R. and Hobbs, H. H. (2004). Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**(5685) 869–872.

The ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**(5696) 636–640.

Ji, W., Foo, J. N., O'Roak, B. J., Zhao, H., Larson, M. G., Simon, D. B., Newton-Cheh, C., State, M. W., Levy, D. and Lifton, R. P. (2008). Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nature Genetics* **40**(5) 592–599.

Li, B. and Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics* **83**(3) 311–321.

Liu, P. Y., Zhang, Y. Y., Lu, Y., Long, J. R., Shen, H., Zhao, L. J., Xu, F. H., Xiao, P., Xiong, D. H., Liu, Y. J., Recker, R. R. and Deng, H. W. (2005). A survey of haplotype variants at several disease candidate genes: the importance of rare variants for complex diseases. *Journal of Medical Genetics* **42**(3), 221–227.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F., McCarroll, S. A. and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature* **461** 747–753.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273**(5281) 1516–1517.

Romeo, S., Pennacchio, L. A., Fu, Y., Boerwinkle, E., Tybjaerg-Hansen, A., Hobbs, H. H. and Cohen, J. C. (2007). Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nature Genetics* **39**(4) 513–516.

Walsh, T. and King, M. C. (2007). Ten genes for inherited breast cancer. *Cancer Cell* **11**(2) 103–105.

The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**(7145) 661–678.

Zaitlen, N., Kang, H. M., Eskin, E. and Halperin, E. (2007). Leveraging the HapMap correlation structure in association studies. *The American Journal of Human Genetics* **80**(4) 683–691.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H. and Weir, B. S. (2002). Truncated product method for combining P-values. *Genetic Epidemiology* **22**(2) 170–185.

Zhu, X., Bouzekri, N., Southam, L., Cooper, R. S., Adeyemo, A., McKenzie, C. A., Luke, A., Chen, G., Elston, R. C. and Ward, R. (2001). Linkage and association analysis of angiotensin I-converting enzyme (ACE)-gene polymorphisms with ACE concentration and blood pressure. *The American Journal of Human Genetics* **68**(5) 1139–1148.

Zhu, X., Fejerman, L., Luke, A., Adeyemo, A. and Cooper, R. S. (2005). Haplotypes produced from rare variants in the promoter and coding regions of angiotensinogen contribute to variation in angiotensinogen levels. *Human Molecular Genetics* **14**(5) 639–643.

Zhu, X., Feng, T., Li, Y., Lu, Q. and Elston, R. C. (2010). Detecting rare variants for complex traits using family and unrelated data. *Genetic Epidemiology* **34**(2) 171–187.

Yali Li
Department of Epidemiology and Biostatistics
Case Western Reserve University
Cleveland, Ohio 44106
USA

Tao Feng
Department of Epidemiology and Biostatistics
Case Western Reserve University
Cleveland, Ohio 44106
USA

Department of Mathematics
Heilongjiang University
Harbin, 150086
China

Xiaofeng Zhu
Department of Epidemiology and Biostatistics
Case Western Reserve University
Wolstein Research Building
2103 Cornell Road
Cleveland, Ohio 44106
USA
Tel.: (216) 368 0201
Fax: (216) 368 4880
E-mail address: xzhu1@darwin.case.edu