Comparison of haplotype-based statistical tests for disease association with rare and common variants

Ananda S. Datta and Swati Biswas

Corresponding author: Swati Biswas, Department of Mathematical Sciences, University of Texas at Dallas, 800 W Campbell Road, FO 35 Richardson, TX 75080-3021, USA. Tel.: (972) 883-6686; Fax: (972) 883-6622; E-mail: swati.biswas@utdallas.edu

Abstract

Recent literature has highlighted the advantages of haplotype association methods for detecting rare variants associated with common diseases. As several new haplotype association methods have been proposed in the past few years, a comparison of new and standard methods is important and timely for guidance to the practitioners. We consider nine methods—Haplo.score, Haplo.glm, Hapassoc, Bayesian hierarchical Generalized Linear Model (BhGLM), Logistic Bayesian LASSO (LBL), regularized GLM (rGLM), Haplotype Kernel Association Test, wei-SIMc-matching and Weighted Haplotype and Imputation-based Tests. These can be divided into two types—individual haplotype-specific tests and global tests depending on whether there is just one overall test for a haplotype region (global) or there is an individual test for each haplotype in the region. Haplo.score is the only method that tests for both; Haplo.glm, Hapassoc, BhGLM and LBL are individual haplotype-specific, while the rest are global tests. For comparison, we also apply a popular collapsing method—Sequence Kernel Association Test (SKAT) and its two variants—SKAT-O (Optimal) and SKAT-C (Combined). We carry out an extensive comparison on our simulated data sets as well as on the Genetic Analysis Workshop (GAW) 18 simulated data. Further, we apply the methods to GAW18 real hypertension data and Dallas Heart Study sequence data. We find that LBL, Haplo.score (global test) and rGLM perform well over the scenarios considered here. Also, haplotype methods are more powerful (albeit more computationally intensive) than SKAT and its variants in scenarios where multiple causal variants act interactively to produce haplotype effects.

Key words: haplotype-based methods; collapsing methods; rare variants; case-control samples; Genetic Analysis Workshop; Dallas Heart Study

Introduction

Genome-wide association studies (GWAS) have implicated many common single nucleotide polymorphisms (SNPs) to be associated with various common diseases; however, collectively those variants explain only a small proportion of heritability of any disease/trait (a situation commonly referred as ‘missing heritability’) [1, 2]. This realization gave way to the search for missing heritability in the post-GWAS era wherein rare variants have been heralded to be an important key, as they were not genotyped in GWAS. This paradigm shift fueled with concurrent availability of the next-generation sequencing (NGS) data lead to burgeoning of methods for collectively analyzing rare single nucleotide variants (rSNV), typically referred as ‘collapsing methods’ [3, 4]. Nonetheless, there are still many limitations

associated with NGS data such as accuracy of genotype calling and cost, and as collapsing methods rely on these data solely, they inherit these limitations as well [5]. Moreover, collapsing methods only test for overall significance of the ‘collapsed’ variant and provide no information about which variant(s) may be responsible for significance if an overall significance is found, leaving open the question of prioritization of variants for biological validation [5].

Another class of rare variant approach is haplotype-based; however, it has received relatively lesser attention in this ongoing race to unlock the missing heritability. One major advantage of haplotype-based methods is that they can be applied to not only NGS data but also to existing GWAS data. As common SNPs can give rise to rare haplotype variants (rHTV), they allow

Ananda S. Datta is a PhD student in Statistics at the University of Texas at Dallas.

Swati Biswas is an Associate Professor of Statistics in the Department of Mathematical Sciences at the University of Texas at Dallas.

Submitted: 26 May 2015; Received (in revised form): 23 July 2015

© The Author 2015. Published by Oxford University Press. For Permissions, please email: journals.permissions@oup.com

testing of rare variant association, especially for following up on candidate genomic regions. Thus, the full richness of GWAS data can be explored for rHTV association, which has remained largely untapped. In fact, haplotype analysis is a natural next step once a specific region has been implicated through single-SNP approaches. Haplotype association has a rich literature with many earlier studies highlighting their importance including their increased power than single-SNP approaches when there are multiple causal variants acting in cis or if the causal variant(s) are not genotyped [6–8]. More recent works have shown that rHTVs formed by common SNPs are more likely to tag rSNVs and that rHTV association methods may be more powerful than popular collapsing methods for detecting rSNVs in some situations [9–12].

Motivated by these advantages, several haplotype association methods have been proposed in the past few years that are designed to handle or incorporate rHTVs in a manner that allows their detection unlike the earlier haplotype association methods that typically pool rHTVs into one ‘pooled’ group. These methods can be broadly divided into two types—(1) those testing for an overall association of a haplotype block as a whole with disease/trait under study (Global tests) and (2) those testing association of individual haplotypes (compared with a base haplotype) in a haplotype block (individual haplotype-specific tests). An advantage of the latter over the former is that they can identify the specific haplotypes that are significant and hence provide more specific information for future validation studies. With the availability of new rHTV association methods, there is a pressing need for a rigorous comparison of these methods to help guide the practitioners.

In this article, our goal is to compare several recently proposed haplotype-based rare variant association methods of both types. These include regularized Generalized Linear Model (rGLM) [13], Logistic Bayesian LASSO (LBL) [14–16], Bayesian hierarchical GLM (BhGLM) [17], wei-SIMc-matching [10], Weighted Haplotype and Imputation-based tests (WHaIT) [9] and Haplotype Kernel Association Test (HKAT) [18]. We will compare these with more standard haplotype association methods, in particular, Hapassoc [19, 20], Haplo.score [6] and Haplo.glm [21], which were mainly proposed for common haplotype association but may be helpful for rHTVs also. As our focus is on rare variant association, for comparison, we will also apply a popular collapsing method of Sequence Kernel Association Test (SKAT) [22] and its variants SKAT-Optimal (SKAT-O) [23] and SKAT-Combined (SKAT-C) [24]. However, we do not intend to provide an in-depth comparison of haplotyping versus collapsing approaches, which has been reported elsewhere [11]. We use our own simulated data as well as Genetic Analysis Workshop (GAW) 18 simulated data [25, 26]. We further apply the methods to two real data sets—hypertension data from GAW18 and Dallas Heart Study (DHS) case-control data [27, 28].

Methods

Here we briefly describe all haplotype-based and collapsing methods considered in this study. A more technical description of the methods can be found in the [Supplementary Material](#). The haplotype-based methods are divided into global and individual haplotype-specific tests. Haplo.score is the only method among the ones we consider that provides global as well as individual haplotype-specific tests and it is described in the first category. For a focused study, we consider case-control data with no covariates even though several of these methods can handle other types of data and/or covariates. Many of these

methods are based on fitting a GLM to the response variable Y denoting case-control status.

Global haplotype tests

Haplo.score

It uses a GLM to regress Y on haplotype pairs for a person. As haplotypes cannot be usually inferred unambiguously from the unphased genotype data, the likelihood is expressed as a sum over all compatible haplotype pairs for all persons weighted by the frequencies of haplotype pairs. Haplotype pair frequency is calculated using estimated haplotype frequencies in the population and assuming Hardy Weinberg equilibrium (HWE). The haplotype frequencies are estimated using an Expectation Maximization (EM) algorithm independent of trait or covariates under the null hypothesis of no association. A score test statistic is constructed, which is asymptotically distributed as a χ^2 with degrees of freedom equal to the rank of the variance-covariance matrix of the score vector. Apart from the global test, Haplo.score also tests for individual haplotypes again based on a score test statistic. However, it does not provide magnitude of individual haplotype effects. The default minimum frequency for a haplotype to be included in the model is $1/(2n)$, and we used it for all our analyses except for GAW18 data. As GAW18 sample size is small, we set the minimum frequency to be 0.0001 to allow rare haplotypes to be scored.

rGLM

Similar to Haplo.score, it prospectively models the distribution of Y given haplotypes for each person and the distribution is averaged over all compatible haplotype pairs for a person weighted by the frequency of haplotype pair. However, here the population frequencies of all haplotypes are treated as unknown parameters and estimated along with regression parameters. The regression coefficients are penalized using LASSO. The maximum likelihood estimators are obtained using an EM algorithm. The tuning parameter for the penalty is chosen using Akaike Information Criterion. rGLM uses a likelihood ratio test to test whether there is any association between the haplotypes and trait. The significance is assessed using a permutation procedure; we used 200 permutations.

WHaIT

There are two tests in WHaIT—a haplotype test and an imputation-based test. We use the former in this study but call it WHaIT for ease of reference. The data set is split into testing and training sets by randomly choosing 30% of the sample to be training set. In the training set, haplotype frequencies are compared between cases and controls to define protective and risk haplotypes. Then in the testing set, a weighted haplotype score is defined for each individual that takes into account the direction of each haplotype effect found in the training set. The weight of a haplotype is inversely proportional to the frequency of haplotype, i.e. rHTVs are given more weight. The weighted scores are then used to calculate a standard Wilcoxon test statistic for comparing cases and controls, and significance is assessed using permutation. To infer the haplotype pair of a person, a standard phasing software such as MACH [29] is used.

wei-SIMc-matching

It is a haplotype similarity-based approach. A GLM is fitted with an explanatory variable x_i , which compares i th subject's haplotypes against the haplotypes of other subjects. Specifically, x_i is a product of a vector aggregating haplotype information of all subjects, a similarity matrix S and haplotype frequency vector of the i th subject. S is a diagonal matrix of dimension equal to the number of distinct haplotypes with diagonal elements being inverse of the estimated standard deviation of haplotype counts. It serves as the weight of each haplotype and upweights the rHTVs. A score test statistic is constructed and its asymptotic distribution is approximated by three-moment approximation method. An EM algorithm is used to infer the haplotype frequency vector.

HKAT

It uses a GLM to relate trait to haplotypes, which are assumed to have random effects. In particular, the regression coefficients for haplotypes follow an arbitrary distribution with mean zero and variance $w_j\tau$, where τ is variance component and w_j is a pre-specified weight of j th haplotype. The weight w_j is an inverse function of frequency of haplotype j . The haplotype frequencies are inferred using an EM algorithm. The null hypothesis of no association is framed as $H_0: \tau = 0$ and tested using a variance-component score test statistic. The asymptotic distribution of the test statistic is a mixture of χ^2 distributions.

Individual haplotype-specific tests

Haplo.glm

It is an extension of Haplo.score. It iteratively estimates haplotype frequencies conditional on all observed data and current estimates of regression parameters. The individual haplotype effects are tested using a Wald type test. Further, unlike Haplo.score, it provides estimates of the magnitude of individual haplotype effects. The software pools haplotypes below a certain frequency into a pooled group. As our interest is in rare haplotypes, we set the pooling tolerance to 0.001, the minimum allowed by Haplo.glm.

Hapassoc

It was originally proposed as an extension of Haplo.glm to accommodate missing genotype data at individual SNPs (however, Haplo.glm can now handle missing genotypes as well). It also uses an improved approximation for standard error estimation. Similar to Haplo.glm, it has an option of pooling tolerance, but unlike Haplo.glm, it allows a zero value for it. So, we used that in our analysis, i.e. no haplotypes are pooled. However, because of this, Hapassoc does not converge sometimes, especially for small n .

BhGLM

Similar to rGLM, BhGLM uses regularization and fits a prospective GLM model. However, instead of averaging over all compatible haplotype pairs for the individuals with ambiguous haplotypes, it uses haplotype dosage, which is the estimated number of copies of a specific haplotype for a subject. The regression coefficients are regularized by assigning independent Student t priors $t_{\nu_j}(0, s_j^2)$, with ν_j and s_j prespecified. An EM algorithm is used to find the posterior modes of regression coefficients with the M Step using standard iteratively reweighted

least squares for maximization. At the convergence of the algorithm, the estimates of coefficients, their standard errors and P -values are obtained, which are used for inference. Note that although the model is Bayesian, the inference is carried out in a frequentist manner.

LBL

All of the above GLM-based methods use prospective likelihood. In contrary, LBL uses retrospective likelihood, that is, models the probability of haplotype pairs conditional on disease status. It is much more complicated than its prospective counterpart but is more appropriate in haplotype association settings, as haplotypes are not observed directly [14]. LBL treats haplotype frequencies as unknown parameters and haplotype pairs of subjects as missing data, thus allowing for uncertainty in their values. It also allows for Hardy-Weinberg disequilibrium unlike the above approaches. The regularization of the regression parameters is carried out by Bayesian LASSO (a double exponential prior with mean 0). Posterior distributions of all parameters are obtained using Markov chain Monte Carlo (MCMC) methods. The inference for association is carried out by calculating Bayes Factor (BF). If $BF > 2$, it is considered as a significant evidence for association [14]. An estimate of odds ratio (OR) and its credible set is provided for each haplotype.

Only Haplo.glm, Hapassoc, rGLM and LBL treat haplotype frequencies as unknown parameters and estimate them along with regression parameters (rather than just plugging their estimates) and thereby account for their uncertainty. All methods except WHaIT, HKAT, wei-SIMc-matching and BhGLM average over all compatible haplotypes for each person and thereby incorporate the uncertainty in estimation of haplotype pair for each person. HKAT, wei-SIMc-matching and BhGLM use haplotype dosage (which could be non-integer) as an estimate of haplotype pair for a person but the uncertainty in such estimation is ignored. WHaIT simply uses the most probable haplotype returned by a phasing algorithm directly. All methods except LBL assume HWE. Some key features of each method are summarized in Table 1.

Collapsing tests

SKAT

This is an SNV analogue of HKAT; in fact, the latter was proposed as a haplotype version of SKAT. Thus, the modeling is same as in HKAT with haplotypes replaced by SNVs. The effect of all (collapsed) SNVs is tested with $H_0: \tau = 0$ using a variance-component score statistic as in HKAT.

SKAT-O

The so-called burden tests are more powerful than SKAT when there is a large proportion of causal rare variants with effects in the same direction. So, to optimize power, SKAT-O combines burden test and SKAT by an arbitrary linear combination $Q\rho$ of these two test statistics. Here ρ is a correlation parameter incorporating the correlation structure of effects of variants and is used as a weight in the linear combination. The linear combination is computed over a grid of ρ values and the corresponding P -values are obtained. The minimum of these P -values is the final test statistic, whose significance is assessed using a mixture of χ^2 distributions.

Table 1. Some key features of haplotype association methods

Method	Test (global/ individual haplotype)	Trait (quantitative/ case-control)	Requires phased haplotypes as input data	Assumes HWE	Allows covariates	Run time for $n = 2000$ and Setting 3 ^a (in seconds)
Haplo.score	Both ^b	Both	No	Yes	Yes	1.47
rGLM	Global ^c	Case-control	No	Yes	Yes	901.01
WHaIT	Global	Case-control	Yes	Yes	No	0.93
wei-SIMc	Global	Both	No	Yes	Yes	6.91
HKAT	Global	Both	No	Yes	Yes	3.66
Haplo.glm	Individual	Both	No	Yes	Yes	1.40
Hapassoc	Individual	Both	No	Yes	Yes	7.20
BhGLM	Individual	Both	No	Yes	Yes	1.52
LBL	Individual	Both ^d	No	No	Yes	94.05

All methods except BhGLM were run on a 3.4 GHz Xeon processor under Linux operating system with 31.32 GB RAM. BhGLM was run on a 2.3 GHz Intel(R) Core i3-2350 M under Windows 7 operating system with 4 GB RAM; there is no Linux version available.

^aSetting 3 listed in Tables 2 and 3.

^bHaplo.score does not provide effect sizes of individual haplotypes.

^crGLM provides effect sizes but no test for individual haplotypes.

^dLBL for quantitative trait is proposed in [30]; software currently unavailable.

SKAT-C

It extends SKAT by combining the score test statistics for rare (Q_{rare}) and common (Q_{common}) variants as a weighted sum. So, it is useful when association is caused by a combination of rare and common variants. SKAT-C is, in fact, a member of class of tests referred as SKAT-RC and was shown to have similar or better performance than the other members in this class [24]. The weight parameter ϕ in SKAT-C is chosen in such a manner that $(1 - \phi)Q_{rare}$ and ϕQ_{common} have the same variance. The null distribution of the weighted sum is a mixture of χ^2 distributions.

Results

Comparison on simulated data sets

We will compare the methods on two types of simulated data that complement each other in terms of type of causal variants. We simulate on our own the first type of data wherein the causal variants are haplotypes. So, the underlying disease model has haplotype effects, that is, specific haplotypes formed by specific combinations of alleles of multiple causal SNPs increase the odds of disease. While in the second type of data, the causal variants are SNPs and for this we consider GAW18 simulated data, which has no haplotype effects. In GAW data, 1458 functional variants interact with each other and other factors in a complex manner to affect systolic and diastolic blood pressure levels [25, 31]. Thus, these data allow us to evaluate the approaches in scenarios where the causal variants are individual SNPs rather than haplotypes. Even though lack of any haplotype effect may not be fully biologically justifiable, its otherwise extremely complicated underlying disease model mimics the reality closely and thus serves as a useful test data.

Causal haplotype-based simulated data

We simulate these data in the same manner as in [14]. In particular, we consider three settings with 6, 9 and 12 haplotypes, each made up of five SNPs. For each of these settings, we generate the following association scenarios: (1) RR: two rare causal haplotypes, (2) RC: one rare and one common causal haplotypes and (3) Null: no causal haplotypes, i.e. odds ratio (OR) = 1 for all

haplotypes. The haplotype frequencies and OR for RR and RC are shown in the first three columns of Tables 2 and 3. The phased haplotype pairs are generated using these haplotype frequencies and assuming HWE. Then a logistic regression model with specified ORs and baseline prevalence of 10% is used to generate case/control status. The phase information is removed after a sample is generated. Samples of sizes 1000, 1500 and 2000 with equal number of cases and controls are generated. For each simulation setup, 500 samples are generated. More details about simulations can be found in [14].

For each method, we estimate power/type I error rate (overall and/or individual haplotype-specific) by noting the proportion of samples in which P-value is less than a cutoff (for all methods except LBL) or BF exceeds a cutoff for LBL.

We present the results of individual haplotype-specific tests for RR ($n = 1500$) and RC ($n = 1000$) in Tables 2 and 3. The most frequent haplotype is used as the baseline in all analyses reported here. A nominal 5% significance level is used as cutoff for P-value, while for BF the suggested cutoff of 2 is used [14]. We note that Haplo.score, Haplo.GLM and Hapassoc have high powers but some individual type I error rates are higher than 5%, especially for Haplo.score. On the other hand, BhGLM is highly conservative with almost all type I error rates close to 0. LBL, with the cutoff of 2, has controlled type I error rates, each below 5%, consistent with previous studies [12, 14, 15, 26]. Similar trend is seen for other sample sizes (see Supplementary Tables S1 and S2).

To be able to compare powers at similar type I error rates, we plot the powers of the two non-null haplotypes against the average type I error rates (averaged over three, six and nine null haplotypes in Settings 1, 2 and 3, respectively) as receiver operating characteristic (ROC) curves in Figures 1 and 2. For BhGLM, huge cutoff adjustments had to be made, e.g. it achieves 5% average type I error rate at adjusted cutoffs of 22.08%, 25.82% and 29.49% for Settings 1, 2 and 3 in Figure 1. After adjustments, we see in Figure 1 for RR scenario that BhGLM and LBL have similar powers for small average type I error rates (with BhGLM performing slightly better in detecting Rare 1 in Settings 2 and 3 and much worse in detecting Rare 2 in Setting 1), while in the rest, LBL performs best. However, in Figure 2 for RC scenario, LBL outperforms BhGLM in the whole range for both rare and

Table 2. Simulation results for individual haplotype-specific methods in RR scenario with $n = 1500$ showing percentage of replicates in which a haplotype is found to be significant

Haplo names	True OR	True freq	LBL	Hapassoc	Haplo.glm	Haplo.score	BhGLM
01100	1	0.3	0.6	5.2	5.2	8	0
10100	3	0.005	56.6	68.4	68.6	71	22.6
11011	2	0.01	38	51.6	51.8	52.2	7.6
11100	1	0.155	0.2	3.2	3.2	3.8	0
11111	1	0.11	0.6	5.2	5.2	7	0
01010	1	0.06	1	5.4	5.4	4.2	0
01100	1	0.25	0.6	4.4	4.6	7.6	0
10000	1	0.08	2.4	7	7	7	0
10100	3	0.005	58.4	66.6	67.8	71	36.8
11011	2	0.01	37	45.8	47.4	48.8	12.8
11100	1	0.09	1.8	7.6	7.6	8	0
11101	1	0.085	0.8	5	5	4.8	0.2
11111	1	0.1	1.2	5	5.2	6.2	0
00111	1	0.07	1.2	5	5	6.4	0
01000	1	0.02	3	5.6	5.6	5.2	1.2
01011	1	0.05	2	5.2	5.2	6.2	0.4
01101	1	0.06	1.6	5.8	5.8	7.4	0.4
01110	1	0.14	1	5	5	6.8	0
10010	1	0.08	2.2	4.8	4.8	6	0
10100	3	0.005	56.2	62.8	62.8	67.4	42.8
11011	2	0.01	37.8	48.8	49	49.8	22
11101	1	0.09	2.2	6.4	6.4	5.8	0
11110	1	0.13	0.8	6.4	6.4	7.2	0
11111	1	0.1	1.8	6.2	6	6.6	0

The three parts are for Settings 1, 2 and 3; the baseline (most frequent) haplotype in each setting is not listed. For declaring significance, a nominal cutoff of 5% is used for all methods except LBL for which significance is declared if $BF > 2$.

Table 3. Simulation results for individual haplotype-specific methods in RC scenario with $n = 1000$ showing percentage of replicates in which a haplotype is found to be significant

Haplo names	True OR	True freq	LBL	Hapassoc	Haplo.glm	Haplo.score	BhGLM
01100	1	0.3	0.4	4	4	33.6	0
10100	3	0.005	36.8	46.8	47.8	44	11.4
11011	1	0.01	3.6	3.6	3.6	5	1
11100	1	0.155	1.2	4.4	4.2	16.6	0
11111	2	0.11	99.6	98.6	100	100	4.4
01010	1	0.06	1.8	3.8	3.8	5.2	0.2
01100	1	0.25	0.8	4.8	5	17.4	0
10000	2	0.08	93	93.2	97	99.2	40.4
10100	3	0.005	38.2	42.4	44.6	42.6	21.4
11011	1	0.01	3	2.2	2.6	4.6	1
11100	1	0.09	1.2	4.4	4.4	5.2	0.2
11101	1	0.085	1.6	5.2	5.4	7.2	0.2
11111	1	0.1	2	4.6	4.8	16.8	0
00111	1	0.07	0.8	3.2	3.2	7	0.2
01000	1	0.02	4	5.2	5.4	7.4	1.8
01011	1	0.05	2	4.6	4.6	8.2	1
01101	1	0.06	1.8	5.6	5.6	14.2	0.4
01110	1	0.14	0.6	4	4	7.8	0.2
10010	2	0.08	90	94.2	95.2	98.6	62
10100	3	0.005	42.2	41.6	42.2	44.8	24.4
11011	1	0.01	4.2	4.6	4.8	6.4	2.2
11101	1	0.09	1.2	4	3.8	17.2	0.2
11110	1	0.13	1.6	6.8	6.8	8.4	0
11111	1	0.1	1.4	4.8	4.6	8.2	0

The three parts are for Settings 1, 2 and 3; the baseline (most frequent) haplotype in each setting is not listed. For declaring significance, a nominal cutoff of 5% is used for all methods except LBL for which significance is declared if $BF > 2$.

Figure 1. Simulation results for individual haplotype-specific methods in RR scenario with $n = 1500$. The haplotype settings are described in Table 2.

common haplotypes. In fact, the power of BhGLM for detecting the common haplotype is lower than all other methods, most notably in Setting 1. In this same setting, BhGLM's power for even the rare haplotype is smaller than those of Hapassoc and Haplo.glm. Similar results are seen in Supplementary Figures S1 and S2 for different sample sizes in RR scenario.

For comparison of global haplotype tests, we plot their powers versus type I error rates (estimated from the Null scenario) in Figures 3 (for RR with $n = 1500$) and 4 (for RC with $n = 1000$). We also plot SKAT and its variants in these plots. We see that rGLM and Haplo.score perform best with HKAT performing close to them in RR and similar to them in RC. Further, in the RC scenario, wei-SIMc-matching has comparable performance and WHaIT performs close to these methods in Setting 1. Similar results can be found in Supplementary Figures S3 and S4 for RR scenario. To investigate if global approaches maintain their nominal type I error rates, we plotted the empirical type I error rates for RR scenario ($n = 1000$ and 2000) against the respective nominal rates in Supplementary Figure S5. We find that WHaIT does not maintain its nominal error rate in Settings 2 and 3.

To explore how the above results may be affected when there are a larger number of causal rHTVs or a larger number of interacting SNPs that produce haplotype effects, we performed two additional simulations. In the first one, we consider a variation of Setting 3 (12 haplotypes) by letting six rHTVs to be causal. For each of the first four haplotypes listed in Table 2 under Setting 3, we changed its frequency to be 0.005 and OR to be 1.5. As the frequencies must add up to 1, we modified the frequencies of haplotypes 01110, 11101, 11110 and 11111 to be 0.18, 0.15, 0.16 and 0.15, respectively. We kept the other information same as listed in the table. The results for $n = 1500$ are summarized and plotted in the same manner as described above. Figure 5 shows the results for the individual haplotype-specific methods for Rare 1 and Rare 2 and the global methods. Supplementary Figure S6 shows the results for the other four causal rHTVs. The results are mostly the same as before except BhGLM's drastic drop in power. A closer look reveals that its powers for Rare 1 and Rare 2 at 5% cutoff drop to 28% and 10% from 42.8% and 22% in the original Setting 3 (in Table 2). At the same time, its type I error rates increase (although it is still conservative), leading to relatively lower power boost when

Figure 2. Simulation results for individual haplotype-specific methods in RC scenario with $n = 1000$. The haplotype settings are described in Table 3.

cutoffs are adjusted in Figure 5. Thus, it seems BhGLM's performance gets notably worse in presence of larger number of rHTVs.

For the other simulation, we consider a setting described in Table 3 of [11]. There are 10 SNPs with relatively weak individual effects that combine to give eight haplotypes. The haplotype frequencies and OR are listed in Table 4; the corresponding information on individual SNPs can be found in [11]. Figure 6 shows the results with $n = 1500$. Here BhGLM performs best among the individual haplotype-specific methods; however, it achieves an average type I error rate of 5% at a cutoff of 30.44%. It is closely followed by LBL, Haplo.score and Haplo.glm. Among the global methods, Haplo.score and rGLM perform best followed by HKAT as before.

Causal SNP-based simulated data

Next we investigate the methods under a non-haplotype-based disease model using GAW18 simulated data. These were simulated to mimic the GAW18 real data (analyzed in the next section). There are 200 replicates, each consisting of the same

genotypes as in the real data. The phenotypes of systolic and diastolic blood pressures were generated [25, 31]. We used a binary hypertension status derived from these phenotypes and medication status to define case/control status as in [26]. The original data are on 20 extended families, so founders were extracted from these families to get a sample of 142 independent individuals. Approximately half of them are cases but that number varied across replicates.

As this is a small sample size especially for detecting rare variants, following [26], we focus on analysis of two genomic regions, each of 8000 base pairs (bp) and centered at the following two functional (causal) variants: (1) the most frequent variant at location 47 957 741 bp with minor allele frequency (MAF) = 0.378 (Region 1) and (2) the variant with the strongest effect size at location 48 040 283 bp with MAF = 0.032 (Region 2). Both of these regions are in MAP4 gene, which has a total of 15 functional variants, with MAF ranging from 0.0016 to 0.378. Out of them, five are in Region 1 and two are in Region 2. As in [26], we included SNPs with MAF > 0.01 in forming the two haplotype blocks in these regions, thus requiring at least two copies of minor allele of a SNP to be present in a sample of size 142. In

Figure 3. Simulation results for global haplotype methods in RR scenario with $n = 1500$. The haplotype settings are described in Table 2.

each haplotype block, after filtering, there were a total of nine SNPs. The nine SNPs in each block form 9 and 12 haplotypes in Regions 1 and 2, respectively, with several of them being rare. These haplotypes, their estimated frequencies and the ones containing the functional SNPs can be found in [26] in addition to other details such as MAF of SNPs. In particular, Region 1 has seven causal haplotypes (containing minor allele of one or more functional SNPs), while Region 2 has two causal haplotypes.

In Figure 7, we present ROC curves for all methods, including an ‘overall’ curve for LBL and BhGLM. For obtaining these ‘overall’ values, detection of any truly associated haplotype is considered as an overall true positive, while detection of any haplotype under the null model is considered as an overall false positive. The data under the null model were obtained by randomly permuting the affection status of individuals [26]. There are no curves for Hapassoc and Haplo.glm because Hapassoc did not converge and Haplo.glm gave at least one P-value of 0 in all replicates under the null model, leaving no room for cutoff adjustment to get a curve (i.e. its overall type I error rate is 1 at

all cutoffs). The curve shown for Haplo.score is for its global test. In both regions, LBL performs the best followed by BhGLM and SKAT. In Region 2, these are closely followed by rGLM, Haplo.score and SKAT-O, especially at 5% type I error rate. The conservative nature of BhGLM is observed in this analysis also, as it achieves 5% overall type I error rate at cutoffs of 29.68% and 35.70% in the two regions. SKAT and SKAT-O perform much better in these data compared with the earlier simulations because the GAW18 data simulation model was based on effects of individual SNPs and no haplotype effect (e.g. multiple causal variants acting in cis) was modeled as such [25].

Analysis of GAW18 real hypertension data

As these are real data and causal variants are unknown, here we analyze the whole MAP4 gene using sliding overlapping windows of five SNPs as haplotype blocks (i.e. windows with SNP# 1–5, 2–6, and so on). We used SNPs with MAF > 0.1 for this analysis and that gave 113 SNPs and $113 - 5 + 1 = 109$ sliding

Figure 4. Simulation results for global haplotype methods in RC scenario with $n = 1000$. The haplotype settings are described in Table 3.

windows. The phenotype is hypertension as in the simulated data. There are 27 cases and 115 controls. To declare significance, we used 5% cutoff for P -value and 2 for BF.

LBL found significant association in nine windows, Haplo.score (individual haplotype test) in 44, Haplo.score (overall test) in 4, Haplo.glm in 70, Hapassoc in 1 and rGLM in 1. Other methods did not find association in any window. Hapassoc did not converge in most windows owing to presence of rare haplotypes and small sample size. Table 5 shows results for the windows where at least three methods detected significance. The windows where only two methods found significance are reported in Supplementary Table S3. SKAT and its variants did not detect any significance even when all SNPs in the gene were analyzed together at MAF cutoff of 0.1 or 0.01.

To get an idea about whether the association signals found by various methods are likely to be true or false signals, we also analyzed the same windows by permuting the affection status and averaging the number of significant results over all 109 windows (as in [12]). The type I error rates were estimated to be

1.35%, 34.98%, 3.59%, 3.67% and 0.92% for LBL, Haplo.glm, Haplo.score individual haplotype-specific and global tests and rGLM, respectively, while for the rest it was 0. As Hapassoc did not converge in most windows, we do not report its rate. Thus, we may conclude that the significant results found by LBL, Haplo.score and rGLM are likely to be true positive.

Analysis of DHS data

The sequence data from DHS have been analyzed in many studies focusing on methods for rare variants. Thus, these data serve as an excellent benchmark for comparison of methods considered here. Following [11], we study association of variants in gene *ANGPTL5* with a dichotomized version of trait serum triglyceride (TG). The dichotomization is achieved by assigning the subjects with the top 20% TG values as cases and the bottom 20% TG values as controls [11, 32]. The sample includes African Americans, European Americans and Hispanics, and excludes subjects treated with statin [32]. There are 628 cases and 621

Figure 5. Simulation results for (A) individual haplotype-specific and (B) global haplotype methods for a variation of Setting 3 with six causal rHTVs as described in the text and $n = 1500$. Rare 1 and Rare 2 refer to the haplotypes with OR of 3 and 2, respectively. The plots for the other four causal rHTVs are shown in [Supplementary Figure S6](#).

Table 4. Simulation setting with 10 SNPs (from [11])

Haplo names	True OR	True freq
0001000000	0.3028	1
0100000000	0.4123	1
1010010000	0.1623	1
1010010001	0.0546	1
1010111100	0.0246	1
1010010010	0.0050	1
0000000000	0.0163	2
1010011100	0.0221	1.5

controls. After including missense substitutions, nonsense substitutions and frameshift insertions and deletions, we obtained 15 SNVs on *ANGPTL5* of which only one has $MAF > 1\%$ [32]. These are the following (in order): 11727_L98P, 14885_S175P, 17006_W192R, 22472_L225V, 22487_I230L, 22496_I233V, 22602_T268M, 22604_R269G, 22623_L275X, 25956_D293H, 25986_S303P, 26004_G309R, 26065_N329S, 26164_T362M and 26238_Y386X.

As in the above GAW18 analysis, we consider sliding overlapping windows of 5 variants, thus analyzing 11 windows. Most methods found significance in one or more windows, with LBL finding the most number of significant results. The number of windows found to be significant by each method is LBL: 8, Haplo.score (individual haplotype-specific tests): 5, Haplo.glm: 5, rGLM: 5, HKAT: 6, wei-SIMc-matching: 3, SKAT: 1, SKAT-O: 3, SKAT-C: 5. Hapassoc did not converge in any window. Haplo.score's global test, BhGLM and WHaIT did not find any significance at 5% level. However, Haplo.score's global P -values were in the range of 0.06 to 0.08 in the windows where its individual haplotype-specific test was significant. All the windows found to be significant are shown in [Table 6](#). We see that several methods detected a common set of five windows (covering SNVs 4 through 12). BhGLM's P -values were in the range of 0.1 to 0.24 for the specific haplotypes that were found to be significant by LBL, Haplo.score and Haplo.glm in these five windows so this may be again a manifestation of the conservative nature of BhGLM seen in the simulation studies. All haplotypes found to be significant by LBL are protective with estimated OR < 1 ; in particular, the four haplotypes in the windows encompassing

Figure 6. Simulation results for (A) individual haplotype-specific and (B) global haplotype methods with 10 SNPs as described in Table 4 and $n = 1500$. Rare 1 and Rare 2 refer to the haplotypes with OR of 2 and 1.5, respectively.

SNVs 4–11 have their whole 95% credible sets for OR lying below 1 (results not shown). The protective direction of effects is also confirmed by the estimates from Haplo.glm and BhGLM, the other two methods that provide this type of information. We also applied SKAT and its variants with all 15 SNVs together and found the P -values to be 0.03 for both SKAT and SKAT-C, and 0.02 for SKAT-O. Overall, the results are consistent with those found in [11].

Similar to the GAW18 real data analysis, we permuted the case-control status and reanalyzed the same windows. However, as there are only 11 sliding windows in DHS as opposed to 109 in GAW18 analysis, we replicated the permutation process 10 times to obtain 110 null windows. The estimated type I error rates are as follows: LBL: 2.55%, Haplo.glm: 6.54%, Haplo.score (individual haplotype-specific): 0.91%, rGLM: 10%, HKAT: 0.91%, wei-SIMc-matching: 3.64%, SKAT-C: 2.73% and 0% for the other methods. Thus, type I error rates appear to be controlled for all methods except rGLM and Haplo.glm.

Finally, we also use DHS data to explore whether causal rSNVs can be tagged by rHTVs formed by common variants. For this, we examine the performance of haplotype methods when five most common SNVs (out of 15) are used in forming the

haplotype block. The five SNVs (with position numbers and MAF) are: 11727_L98P (SNV# 1; MAF = 0.0024), 17006_W192R (SNV # 3; MAF = 0.004), 22602_T268M (SNV # 7; MAF = 0.0264), 26004_G309R (SNV# 12; MAF = 0.004) and 26238_Y386X (SNV# 15; MAF = 0.0028). LBL gave two significant rare haplotypes: 00001 (BF = 2.32) and 00010 (BF = 2.2) with estimated frequencies of 0.0028 and 0.004, respectively. No other individual haplotype-specific method showed significance. The P -values of global methods are: Haplo.score: 0.031, rGLM: 0.04, HKAT: 0.03, wei-SIMc-matching: 0.054 and WHaIT: 0.9. The estimated type I error rates obtained by permutation of affection status are LBL: 4.17%, Haplo.glm: 3.8%, Haplo.score: 2% (global) and 2.5% (individual haplotype-specific), BhGLM: 0.5%, rGLM: 4%, wei-SIMc-matching: 7%, HKAT: 5% and WHaIT: 4%. Thus, it seems rHTVs obtained by combining common variants can tag causal rSNVs, consistent with [9, 10, 11, 18].

Discussion

Recent literature has highlighted the importance of haplotype association analysis as a promising vehicle for exploring rare variants irrespective of whether interest lies in detecting rHTV or

Figure 7. Results of GAW18 simulated data analysis.

Table 5. GAW18 real data results for MAP4 gene

SNP # in hap block	SNP location	Hap name	Hap freq	LBL (BF)	Haplo.glm (P-val)	Haplo.score (P-val)	Haplo.score overall (P-val)	rGLM (P-val)
30–34	47939626–	00001	0.004	2.105*	0.000*	0.038*	0.202	0.256
	47953405	11011	0.004	0.987	0.000*	0.627		
31–35	47943861–	00010	0.004	2.027*	0.000*	0.038*	0.202	0.286
	47953733	10111	0.004	0.977	0.000*	0.627		
70–74	48022573–	00001	0.025	3.977*	0.023*	0.008*	0.045*	0.090
	48027851							
71–75	48022707–	00010	0.021	2.096*	0.112	0.048*	0.051	0.020*
	48028144	00011	0.004	2.122*	0.000*	0.038*		
72–76	48024629–	00100	0.021	1.821	0.107	0.048*	0.033*	0.092
	48029118	00110	0.004	2.342*	0.000*	0.038*		
100–104		10111	0.046	1.016	0.000*	0.275	0.109	0.340
	48092017–	00100	0.039	2.423*	0.000*	0.094		
	48105528	00110	0.011	1.380	0.126	0.034*		

The blocks shown below were significant (denoted by *) by at least three methods. Major allele is coded as 0. SNP # corresponds to the order of SNP in the gene. Hap: haplotype, Freq: frequency, P-val: P-value.

rSNV. A major advantage of working with haplotypes is that the already amassed GWAS data can be used for finding rare variants without any additional costs associated with gathering sequence data. Thus, focusing on haplotypes allows one to use the ‘freely’ available data as well as the new sequence data that are currently being generated. This is especially helpful, as the current sequence data have still several limitations. Recognizing this advantage, several haplotype-based approaches have been proposed recently, and so to guide the practitioners, it is a prime time for comparison of these methods.

We have compared nine haplotype-based and three collapsing methods. Our overall findings may be summarized as follows. For the scenarios considered here: (1) of all individual haplotype-specific tests, LBL performs the best in general with BhGLM performing comparably or slightly better in some scenarios with a

properly calibrated cutoff, (2) of all overall haplotype tests, rGLM and Haplo.score perform best and (3) haplotype-based methods are more powerful than collapsing methods. Each of these points deserve a more thorough discussion.

Of the individual haplotype-specific tests, Haplo.score, Haplo.GLM and Hapassoc give some individual type I error rates to be higher than the nominal value. While BhGLM is extremely conservative, requiring the cutoffs to be increased to as much as 20–30% to achieve a 5% type I error rate; this trend was seen in GAW18 simulated data also. The behavior might be different for other scenarios not considered here; however, users need to be aware of these pitfalls while using these methods at a specific nominal level. With cutoffs adjusted to match the type I error rates for all methods, LBL has highest power in general. BhGLM performs as good or even slightly better for small type I

Table 6. Results from the DHS for ANGPTL5 gene

SNV # in hap block	Hap name	Hap freq	LBL (BF)	Haplo.glm (P)	Haplo.score (P ^{**})	rGLM (P)	HKAT (P)	wei-SIMc (P)	SKAT (P)	SKAT-C (P)	SKAT-O (P)
4–8	00001	0.002	6.330*	0.000*	0.024*	0.005*	0.025*	0.039*	0.064	0.018*	0.016*
5–9	00010	0.002	6.872*	0.000*	0.024*	0.015*	0.034*	0.054	0.064	0.018*	0.052
6–10	00100	0.002	6.065*	0.000*	0.024*	0.010*	0.045*	0.071	0.064	0.019*	0.091
7–11	01000	0.002	5.617*	0.000*	0.024*	0.045*	0.048*	0.060	0.075	0.049*	0.076
8–12	00001	0.004	2.449*	0.074	0.054	0.020*	0.022*	0.021*	0.023*	0.023*	0.030*
	10000	0.002	5.247*	0.000*	0.024*						
9–13	00010	0.004	2.085*	0.076	0.055	0.270	0.277	0.228	0.158	0.158	0.243
10–14	00100	0.004	2.207*	0.076	0.055	0.300	0.200	0.146	0.158	0.158	0.207
11–15	00001	0.003	2.589*	0.092	0.056	0.140	0.031*	0.018*	0.158	0.158	0.024*
	01000	0.004	2.462*	0.074	0.055						

Significant results are marked by *. Major allele is coded as 0. SNV# corresponds to the order of SNV in the gene. Hap: haplotype, Freq: frequency, P: P-value.

**Haplo.score P-value is individual haplotype-specific.

error rates in detecting a rare causal haplotype in some settings in RR scenario. However, when the number of causal rHTVs increase, its performance seems to drop considerably. Moreover, in RC scenario, the power of BhGLM for detecting the common haplotype is greatly reduced and is less than all other methods. In GAW18 simulated data also, LBL performs better than BhGLM presumably because of presence of both rare and common causal variants. As there is now strong evidence that both rare and common variants contribute to disease susceptibility [33, 34], LBL may be the preferred tool if such a scenario is anticipated.

Although Haplo.score's individual haplotype-specific test does not perform well, its global test is one of the best overall tests and is comparable with rGLM. This is a somewhat unexpected result, as Haplo.score was not proposed for detecting rHTV association as such; however, the result is consistent with [10, 18]. So, it is remarkable that it performs much better than rHTV-focused methods such as wei-SIMc-matching and WHaIT. HKAT is the only other method that comes close to Haplo.score and rGLM in power comparison.

SKAT performed close to some haplotype-based methods in GAW18 simulated data, in which the true data generating model does not include any haplotype effect (e.g. no interacting SNPs in a block to produce larger haplotype effects). From a biological point of view, haplotype effects can be expected in many real data sets, and in such cases, our work shows that just collapsing SNPs together does not retain same power as analyzing them as a haplotype block. Similar results have been recently shown elsewhere [11, 12]; however, we add a disclaimer that comparison of haplotyping versus collapsing approaches is not the focus here and thus has not been investigated in a more rigorous manner. We would like to refer readers interested in this comparison to [9, 10, 11, 18] wherein some other collapsing approaches are also considered. SKAT, SKAT-C and SKAT-O performed similarly in RR, while in RC, SKAT-C performed better as expected. In GAW18 simulated data, however, SKAT had the highest power among the three.

In GAW18 real data analysis, several regions were found to be significant. In particular, the window spanning SNP 70–74 was found to be significant by LBL, Haplo.score (individual as well as overall tests), Haplo.GLM and Hapassoc. Similarly, the adjacent overlapping window of SNP 71–75 was also found to be significant by LBL, Haplo.score and rGLM. Several GAW18 participants found signals on MAP4 gene in the real data (e.g. see [35]); however, most of the rare variant methods used by investigators (mostly collapsing) cannot pinpoint the significant locations, a limitation

that can be overcome with use of haplotype-based approaches as demonstrated here. In fact, MAP4 gene is one of the easy ones to detect in GAW18 data [35] and so it serves as a good benchmark for investigation of methods even with limited sample size. Thus, the fact that some of the methods that we considered here could not detect any signal on MAP4 gene at 5% level is little concerning. All those methods had estimated type I error rates of 0 at 5% level when phenotypes were permuted while keeping the genotypes same. Thus, with small sample sizes, these methods appear to behave far from what is expected asymptotically and may not work reliably.

In DHS analysis, LBL found the most number of significant windows, further illustrating its superior performance seen in simulations and GAW18 analyses. In addition to LBL, several other haplotype-based methods were able to uncover association, illustrating aptly that haplotype-based methods work well with not only GWAS data but also sequence data, contrary to popular belief. This was also demonstrated in GAW19 sequence data analysis recently [12]. Moreover, individual haplotype-specific methods provide additional information about the specific haplotypes that are associated and the direction of their effects. This is a useful piece of information for follow-up studies to understand the causal mechanism and is a feat unmatched by collapsing methods or even the global haplotype methods. In particular, in DHS analysis, we found interestingly that all significant haplotypes are protective, consistent with [11].

One possible reason for low power and/or high type I error rates of some haplotype methods could be that they ignore uncertainty in estimation of phased haplotype data and/or population haplotype frequencies, which gets magnified in presence of rare haplotypes. For example, wei-SIMc-matching had a much worse performance in RR setting compared with RC setting. For WHaIT, another factor might be the fact that the uncertainty in defining risk and protective haplotypes in the training data set is also not taken into account, and that might contribute to its inflated type I error rates seen in Settings 2 and 3. It is well known in the literature that use of inferred haplotypes in downstream analysis without accounting for its uncertainty may give biased estimates, low power and/or high type I error rates [36].

Our findings confirm with the well-known result that haplotype methods are useful when multiple causal variants act in cis. Some haplotype methods such as LBL and Haplo.score also seem to be able to tag rare single causal variants through rHTVs, as indicated by GAW18 and DHS results. Nonetheless, when there are many extremely rare SNVs, the haplotypes tagging them will be even more rarer and haplotype methods,

especially the individual specific ones, may not work well owing to large variability in effect size estimates. Another situation where haplotype methods work well as documented in earlier literature is when causal (typically common) variants have not been genotyped. With the availability of reference panels such as HapMap and 1000 Genomes databases and sophisticated imputation approaches, this advantage of haplotype testing methods may be relatively less relevant now than before. However, a comparison of haplotypic and imputation methods focused to this aspect will be interesting especially when uncertainty in imputation is ignored.

We did not adjust for multiple testing in our real data analyses even though there were many sliding windows especially in the GAW18 real data analysis. The multiplicity adjustment issue is nontrivial in this context owing to dependence across windows so to keep the study goals focused, we simply used 5% cutoff for P-value and BF cutoff of 2. However, to get an idea about whether the detected signals are likely to be true, we permuted affection status to simulate a null scenario and reported the percentage of significant results. If the results reported here are to be used in follow-up studies, we recommend assessing their significance by applying an appropriate multiplicity adjustment to the reported P-values and BF. Also, for the sliding window size in the real data analyses, we used an arbitrarily chosen size of five SNPs. There is no single recommended window size as the optimum choice depends on the underlying linkage disequilibrium (LD) structure and that varies across the genome. One option is to use sliding windows of varying sizes and that may perform better than fixed sized windows [37, 38]; however, it is much more computationally intensive. Alternatively, haplotype blocks could be detected in a region and analyses done in those blocks. Although there are a number of algorithms for detecting blocks of high LD, the block structures and boundaries may differ across different methods.

An important issue is computation time, as haplotype association methods are typically more computationally intensive than their single variant counterparts. Table 1 lists the computing times needed by the haplotype methods to run one sample of size 2000 in Setting 3. SKAT and its variants are faster than haplotype methods, as expected, taking about 0.47 s. Among the haplotype methods, the methods using asymptotic null distribution are fastest, while rGLM takes longest time, as it uses permutation to compute P-value. It is followed by LBL as it is MCMC based. Even though BhGLM is based on a Bayesian model, it circumvents the use of MCMC and uses asymptotic normality of posterior mode to make inference, making it computationally appealing. However, it is not clear whether its highly conservative behavior follows partly from its reliance on asymptotic distribution. In general, owing to computational intensity, most haplotype methods are not suitable for genome-wide analysis (especially rGLM and LBL even though they are among the best methods) rather are more suited for gene-wide analysis [12] or zooming into regions deemed to be of interest by collapsing or other fast screening methods.

A limitation of this work is that we ran each method at its default setting as provided by the associated software (unless noted otherwise in the Methods section), and they may not be optimum for the data sets being analyzed. Thus, it is possible that by using different options, different results may be obtained for some methods. However, our goal with this work is to provide guidance to a typical practitioner and thus we mimicked one who lacks expert knowledge about these methods and hence would use the default settings. Further, we stress that the results are valid for the scenarios and settings similar to those considered here; different data generating models may

give different results. We note that several of the methods also allow covariates to be included as well as gene–environment and/or gene–gene interactions (as indicated in Table 1). Thus, it will be interesting to compare the methods in such settings. Finally, to keep the study focused, we included only some but not all haplotype-based methods in this comparison; the choices were partly driven by whether a method was specifically developed for detecting rHTV as well as by the availability of working software and its support.

Web resources

The URLs for software packages are as follows:

- BhGLM: <http://www.ssg.uab.edu/bhglm/>
- Hapassoc: <http://stat-db.stat.sfu.ca:8080/~statgen/research/hapassoc>
- Haplo.score and Haplo.glm: <http://www.mayo.edu/research/labs/statistical-genetics-genetic-epidemiology/software>
- LBL: <http://www.utdallas.edu/~swati.biswas>
- rGLM: <http://www.stat.osu.edu/~statgen/SOFTWARE/rGLM/>
- SKAT and its variants: <http://www.hsph.harvard.edu/skat/>
- wei-SIMc-matching and HKAT: <http://homepage.ntu.edu.tw/~linwy/>
- WHaIT: <http://csg.sph.umich.edu/yli/whait/>

Key Points

- Several haplotype association methods have been proposed recently for uncovering rare variant association.
- We compared nine haplotype association methods—Haplo.score, Haplo.glm, Hapassoc, BhGLM, LBL, rGLM, HKAT, wei-SIMc-matching and WhaIT, and a popular collapsing method—SKAT and its two variants (SKAT-O and SKAT-C).
- The methods that seem to work best over the range of scenarios considered here include LBL, Haplo.score and rGLM.
- Haplotype methods, in general, outperform the collapsing methods when there are multiple causal variants acting in cis.

Supplementary Data

Supplementary data are available online at <http://bib.oxfordjournals.org/>.

Acknowledgements

We thank Drs Jonathan Cohen and Helen Hobbs for providing the DHS data and Dr. Asuman Turkmen for helping with preprocessing of DHS data. We are also thankful to the three anonymous reviewers for helpful comments. The GAW18 whole genome sequence data were provided by the T2D-GENES Consortium, which is supported by NIH grants U01 DK085524, U01 DK085584, U01 DK085501, U01 DK085526 and U01 DK085545. The other genetic and phenotypic data for GAW18 were provided by the San Antonio Family Heart Study and San Antonio Family Diabetes/Gallbladder Study, which are supported by NIH grants P01 HL045222, R01 DK047482 and R01 DK053889. The Genetic Analysis Workshop is supported by NIH grant R01 GM031575.

Funding

This work was partially supported by the National Cancer Institute (grant number R03CA171011-02).

References

1. Maher B. Personal genomes: the case of the missing heritability. *Nature* 2008;**456**:18–21.
2. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature* 2009;**461**:747–53.
3. Dering C, Hemmelmann C, Pugh E, et al. Statistical analysis of rare sequence variants: an overview of collapsing methods. *Genet Epidemiol* 2011;**35**(Suppl 1):S12–17.
4. Basu S, Pan W. Comparison of statistical tests for disease association with rare variants. *Genet Epidemiol* 2011;**35**:606–19.
5. Goldstein DB, Andrew A, Jonathan K, et al. Sequencing studies in human genetics: design and interpretation. *Nat Rev Genet* 2013;**14**:460–70.
6. Schaid DJ, Rowland CM, Tines DE, et al. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* 2002;**70**:425–34.
7. Schaid DJ. Genetic epidemiology and haplotypes. *Genet Epidemiol* 2004;**27**:317–20.
8. Morris RW, Kaplan NL. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet Epidemiol* 2002;**23**:221–33.
9. Li Y, Byrnes AE, Li M. To identify associations with rare variants, just WHaIT: weighted haplotype and imputation-based tests. *Am J Hum Genet* 2010;**87**:728–35.
10. Lin WY, Yi N, Zhi D, et al. Haplotype-based methods for detecting uncommon causal variants with common SNPs. *Genet Epidemiol* 2012;**36**:572–82.
11. Wang M, Lin S. Detecting associations of rare variants with common diseases: collapsing or haplotyping? *Brief Bioinform* 2015;**16**:759–68.
12. Datta AS, Zhang Y, Zhang L, et al. Association of rare haplotypes on ULK4 and MAP4 genes with hypertension. *BMC Proc* 2015, in press.
13. Guo W, Lin S. Generalized linear modeling with regularization for detecting common disease rare haplotype association. *Genet Epidemiol* 2009;**33**:308–16.
14. Biswas S, Lin S. Logistic Bayesian LASSO for identifying association with rare haplotypes and application to age-related macular degeneration. *Biometrics* 2012;**68**:587–97.
15. Biswas S, Xia S, Lin S. Detecting rare Haplotype-environment interaction with logistic bayesian LASSO. *Genet Epidemiol* 2014;**38**:31–41.
16. Zhang Y, Biswas S. An improved version of logistic Bayesian LASSO for detecting rare Haplotype-environment interactions with application to lung cancer. *Cancer Inform* 2015;**14**(Suppl 2):11–16.
17. Li J, Zhang K, Yi N. A Bayesian hierarchical model for detecting haplotype-haplotype and haplotype-environment interactions in genetic association studies. *Hum Hered* 2011;**71**:148–60.
18. Lin WY, Yi N, Lou XY, et al. Haplotype kernel association test as a powerful method to identify chromosomal regions harboring uncommon causal variants. *Genet Epidemiol* 2013;**37**:560–70.
19. Burkett K, McNeney B, Graham J. A note on inference of trait associations with SNP haplotypes and other attributes in generalized linear models. *Hum Hered* 2004;**57**:200–6.
20. Burkett K, Graham J, McNeney B. Hapassoc: software for likelihood inference of trait associations with SNP haplotypes and other attributes. *J Stat Softw* 2006;**16**:1–19.
21. Lake SL, Lyon H, Tantisira K, et al. Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. *Hum Hered* 2003;**55**:56–65.
22. Wu MC, Lee S, Cai T, et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 2011;**89**:82–93.
23. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 2012;**13**:762–75.
24. Ionita-Laza I, Lee S, Makarov V, et al. Sequence kernel association tests for the combined effect of rare and common variants. *Am J Hum Genet* 2013;**92**:841–53.
25. Almasy L, Dyer TD, Peralta JM, et al. Data for Genetic Analysis Workshop 18: human whole genome sequence, blood pressure, and simulated phenotypes in extended pedigrees. *BMC Proc* 2014;**8**(Suppl 1):S2.
26. Biswas S, Papachristou C. Evaluation of logistic Bayesian LASSO for identifying association with rare haplotypes. *BMC Proc* 2014;**8**(Suppl 1):S54.
27. Victor RG, Haley RW, Willett DL, et al. The Dallas Heart Study: a population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. *Am J Cardiol* 2004;**93**:1473–80.
28. Romeo S, Yin W, Kozlitina J, et al. Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. *J Clin Invest* 2009;**119**:70–79.
29. Li Y, Willer CJ, Ding J, et al. MACH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* 2010;**34**:816–34.
30. Zhang H, Lin S. Quantitative Bayesian lasso for detecting effects of rare haplotype variants and environmental factors on complex diseases. 2015. Abstract. *Joint Statistical Meetings*, Seattle, WA. <http://www.amstat.org/meetings/jsm/2015/onlineprogram/AbstractDetails.cfm?abstractid=315376>
31. Satten GA, Biswas S, Papachristou C, et al. Population-based association and gene by environment interactions in Genetic Analysis Workshop 18. *Genet Epidemiol* 2014;**38**(Suppl 1):S49–56.
32. Turkmen A, Yan Z, Hu YQ, et al. Kullback-Leibler distance methods for detecting disease association with rare variants from sequencing data. *Ann Hum Genet* 2015;**79**:199–208.
33. Iyengar SK, Elston RC. The genetic basis of complex traits: rare variants or “common gene, common disease”? *Methods Mol Biol* 2007;**376**:71–84.
34. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* 2008;**40**:695–701.
35. Cordell HJ. Summary of results and discussions from the gene-based tests group at Genetic Analysis Workshop 18. *Genet Epidemiol* 2014;**38**(Suppl 1):S44–48.
36. Lin DY, Huang BE. The use of inferred haplotypes in downstream analyses. *Am J Hum Genet* 2007;**80**:577–9.
37. Li Y, Sung W-K, Liu JJ. Association mapping via regularized regression analysis of single-nucleotide polymorphism haplotypes in variable-sized sliding windows. *Am J Hum Genet* 2007;**80**:705–15.
38. Guo Y, Li J, Bonham AJ, et al. Gains in power for exhaustive analyses of haplotypes using variable-sized sliding window strategy: a comparison of association-mapping strategies. *Eur J Hum Genet* 2009;**17**:785–92.