Published in final edited form as:

Ann Hum Genet. 2009 September ; 73(Pt 5): 520–526. doi:10.1111/j.1469-1809.2009.00536.x.

A regression-based association test for case-control studies that uses inferred ancestral haplotype similarity

Youfang Liu¹, Yi-Ju Li^{2,3}, Glen A. Satten⁴, Andrew S. Allen³, and Jung-Ying Tzeng^{1,§}

¹Bioinformatics Research Center, North Carolina State University, Raleigh, NC

²Center for Human Genetics, Duke University Medical Center, Durham, NC

³Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC

⁴Centers for Disease Control and Prevention, Atlanta, GA

Summary

Association methods based on haplotype similarity (HS) can overcome power and stability issues encountered in standard haplotype analyses. Current HS methods can be generally classified into evolutionary and two-sample approaches. We propose a new regression-based HS association method for case-control studies that incorporates covariate information and combines the advantages of the two classes of approaches by using inferred ancestral haplotypes. We first estimate the ancestral haplotypes of case individuals and then, for each individual, an ancestral-haplotype-based similarity score is computed by comparing that individual's observed genotype with the estimated ancestral haplotypes. Trait values are then regressed on the similarity scores. Covariates can easily be incorporated into this regression framework. To account for the bias in the raw p-values due to the use of case data in constructing ancestral haplotypes, as well as to account for variation in ancestral haplotype estimation, a permutation procedure is adopted to obtain empirical p-values. Compared with the standard haplotype score test and the multilocus T^2 test, our method improves power when neither the allele frequency nor linkage disequilibrium between the disease locus and its neighboring SNPs is too low and is comparable in other scenarios. We applied our method to the Genetic Analysis Workshop 15 simulated SNP data and successfully pinpointed a stretch of SNPs that covers the fine-scale region where the causal locus is located.

Keywords

case-control studies; haplotype similarity; haplotype sharing; haplotype-based association test; covariates; regression-based association analysis

Introduction

Association analysis of case-control studies is becoming the primary approach for localizing disease loci, especially for detecting genes with modest effects on a disease. To assess the association between genetic variants and disease, one can either test individual SNPs or the haplotypes of closely linked SNPs. Although studies of their relative efficiency revealed varying conclusions, it is generally appreciated that haplotype-based analyses have greater

[§]**Corresponding author's contact information:** Jung-Ying Tzeng, Bioinformatics Research Center and Department of Statistics, Campus Box 7566, Raleigh, NC 27569-7566, jytzeng@stat.ncsu.edu Tel Number: (919) 513-2723, Fax Number: (919) 515-7315.

Disclaimer: The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

power when SNPs are in strong multilocus LD with the disease locus (Akey et al., 2001; Nielsen et al. 2004; Zaitlen et al. 2007), and are helpful in identifying rare causal variants (The International HapMap Consortium, 2003; de Bakker et al. 2005). However, the practical potential of haplotype-based analysis may not be fully realized due to difficulties balancing the dimensionality of the haplotypes with the amount of information they contain (e.g., Thomas et al. 2003; Browning, 2006). This high dimensionality can often lead to a loss of power.

Many approaches have been proposed to tackle the dimensionality problem implicit in haplotype-based methods. Most methods can be broadly grouped into three categories: haplotype clustering, haplotype smoothing, and haplotype similarity. Haplotype clustering (Seltman et al. 2003; Molitor et al. 2003a; Durrant et al. 2004; Tzeng, 2005; Tzeng et al. 2006; Browning, 2006) attempts to decrease the dimensionality of the problem by clustering evolutionarily close haplotypes into groups. Haplotype smoothing (Thomas et al., 2001; Thomas et al. 2003; Molitor et al. 2003b; Schaid, 2004; Tzeng and Zhang, 2007) pools information across haplotypes by introducing a correlation structure on the effect of similar haplotypes. Both haplotype clustering and smoothing use measures of haplotype similarity in forming clusters or pooling haplotype effects, but the focus of the analysis is on the clusters or pooled effects. In contrast, haplotype similarity methods (e.g., Houwen et al. 1994; McPeck and Strahs, 1999; Su et al. 2008) focus directly on the similarity metric and look for unusual sharing of chromosomal segments within homogeneous trait groups. In this study, we focus on the haplotype-similarity approach and introduce a method that aims to capture the merits of current similarity methods while also being able to incorporate covariate information.

The rationale behind similarity methods is that, around a causative disease locus, affected individuals are more likely sharing the same haplotype descended from a common ancestor and, hence, will tend to have more similar haplotypes than randomly sampled individuals. Depending on how this concept is implemented, similarity methods can be divided into two categories: evolutionary based approaches and two-sample based approaches. Evolutionary based approaches tend to use cases only and identify regions of increased similarity by comparing to levels expected from a genealogical process under the null (Durham and Feingold, 1997; Service et al. 1999). This approach takes direct advantage of the decay process of haplotype sharing--the driving mechanism for haplotype similarity, but its assumptions concerning evolutionary history can be hard to verify (McPeck and Strahs, 1999; Morris et al. 2002; Morris et al. 2003; Morris, 2005; Su et al. 2008). In contrast, the two-sample based methods use empirical sharing comparisons among case and control haplotypes (Van der Meulen and te Meerman, 1997; Tzeng et al. 2003a; Tzeng et al. 2003b; Allen and Satten 2009). Although the use these empirical sharing comparisons among case and control samples bypasses the need to model the evolutionary process, many of these methods limit similarity calculations to concordant samples (i.e., case-case similarity and control-control similarity) and do not use information obtained from measures of dissimilarity between cases and controls. Using the discordant information has been shown to significantly improve the power of similarity-based tests (Sha et al. 2007; Nolte et al. 2007). Finally, the existing haplotype-similarity methods do not incorporate covariate information. This makes them less attractive for studying complex traits where covariate adjustments can be crucial.

In this article, we address these issues by proposing a new method for analyzing data from case-control studies that combines aspects of both evolutionary and two-sample approaches to haplotype similarity. Our method follows the two-sample framework, while quantifying similarity in a spirit consistent with evolutionary-based approaches. Specifically, we estimate 'ancestral haplotypes' of cases using the multilocus decay-of-haplotype-sharing

(DHS) model of McPeck and Strahs (1999). We then define the ‘*ancestral haplotype similarity (AHS) scores*’ as the degree of similarity between these ancestral haplotypes of cases and the haplotypes of an individual. The use of AHS scores allows us to utilize all sample information in association testing. It also makes it straightforward to extend similarity methods from the 2-sample to a regression framework, potentially allowing for both covariates and quantitative traits to be incorporated in the haplotype similarity analysis. Finally, though we refer to ‘ancestral haplotypes’, our method does not rely on the actual reconstruction of the evolutionary history of the sample – we merely use the estimated sequences as summaries of the case haplotypes and use them to calculate individual similarity scores. Thus the validity of our approach is not affected by our ability to reconstruct the actual ancestral haplotypes of the sample.

In the following sections, we illustrate the proposed AHS method, and present simulation results of type I errors and power. The results are compared to the standard haplotype score test of Schaid et al. (2002) and the multilocus T^2 test of Xiong et al. (2002). We also applied our method to the simulated SNP data for Rheumatoid Arthritis (RA) from the Genetic Analysis Workshop (GAW) 15 and examined whether this proposed method can detect the DR and D loci that affect the risk of RA.

Methods

Model and approach

Our proposed method can be described in the following three major steps. First, we estimate the ancestral haplotypes of cases from the genotypes of case participants using the DHS method of McPeck and Strahs (McPeck and Strahs, 1999; Strahs and McPeck, 2003). Next, the inferred ancestral haplotypes are used as reference haplotypes to compute the AHS scores. The AHS score for an individual is obtained by directly comparing an individual's unphased genotypes to the inferred ancestral haplotypes. Finally, the phenotype of interest is regressed onto the AHS scores and covariates. A significant coefficient of the AHS scores indicates genetic association, as there is a different amount of sharing among the control haplotypes compared to the case haplotypes. The declaration of significance is based on the empirical p-values generated from permutation. Using permutation testing allows us to account for the estimation variation that is induced by the estimated ancestral haplotypes, and to correct the bias introduced by using only the case sample in determining the ancestral haplotypes. Below we describe each of the steps in detail.

(1) Inferring ancestral haplotypes—We use the DHS method (McPeck and Strahs, 1999; Strahs and McPeck, 2003) to infer the ancestral case haplotypes. The DHS method models the decay process of haplotype sharing for haplotypes descended from certain common ancestors, taking into account recombination, mutation, and background LD. The model allows for multiple origins of case haplotypes, and incorporates haplotype correlation due to shared ancestry by applying a correction factor to the star-shaped genealogy of case haplotypes. The method has been implemented in the software of DHSMAP (Decay of Haplotype Sharing Mapping), and can use phased (haplotype) (McPeck and Strahs, 1999) or unphased (genotype) data (Strahs and McPeck, 2003).

(2) Calculating the AHS score—A similarity score for each individual is computed based on the amount of similarity between that individual's haplotypes and the estimated ancestral haplotypes. To tackle the phase unknown problem, we adopt the count statistic of Tzeng et al. (2003a), and measure the level of similarity using the number of matching alleles between an individual's haplotypes and ancestral haplotype. As shown in Schaid (2004) and Tzeng et al. (2009), such quantity can be obtained by counting the matching

alleles between the observed genotypes and the ancestral haplotypes. In the case of multiple estimated ancestral haplotypes, we calculate the overall AHS score by taking a weighted average of the ancestral specific AHS score, with weights dependent on the likelihood of the ancestral haplotypes. Specifically, define L_j the likelihood value obtained from DHSMAP for the ancestral haplotype j , and s_{ij} the count of the matching alleles between the genotype of individual i and ancestral haplotype j . The weight for ancestral haplotype j is

$$w_j = \frac{L_j}{\sum_{\substack{k=\text{all inferred} \\ \text{ancestral} \\ \text{haplotypes}}} L_k}, \text{ and the overall AHS score for person } i \text{ is } S_i = \sum_j s_{ij} w_j.$$

(3) Testing for association—To test association between haplotypes and disease status, we use logistic regression and write

$$\ln \frac{\Pr[D_i=1 | X_i, S_i]}{\Pr[D_i=0 | X_i, S_i]} = \gamma^T X_i + \beta S_i, \quad (1)$$

where D_i denotes disease status (=1 for case participants and 0 otherwise), X_i is the design matrix of covariates including an intercept, γ is a vector of parameters, and β is a parameter that governs the effect of AHS on disease status. The null hypothesis of no association corresponds to $\beta=0$. To address the bias and extra variability introduced into the procedure by using estimated ancestral haplotypes, we use permutation to obtain empirical p-values.

When covariates divide the sample into a relatively small number of categories, permutation datasets can be obtained by randomly shuffling individuals' phenotypes within each covariate category, so as to maintain the potential association between the phenotype and the covariates. When the categories formed in this way are sparsely populated, or when covariates are continuous, the stratification score approach of Epstein et al. (2007) can be used to form (typically five) strata based on the quantiles of $\hat{\gamma}_0^T X_i$, where $\hat{\gamma}_0$ is the estimator of γ obtained by fitting (1) with $\beta = 0$. Permutation can then be carried out with these strata. Alternatively, a parametric bootstrap procedure can be used to generate data under the null hypothesis. First, obtain $\hat{\gamma}_0$ as before. Next, define G_i the genotype of subject i , and sample (X_i, G_i) with replacement from the original data. Finally, use (1) with $\beta = 0$ and $\gamma = \hat{\gamma}_0$ to generate D_i . Repeat this sampling process until the required number of cases and controls is generated.

For each permuted dataset, we repeat the entire procedure of estimating the ancestral haplotypes, computing similarity scores and fitting the regression model. Empirical p-values are computed as the proportion that the absolute values of β estimates in the permutation datasets are greater than the absolute value of β estimate in the original dataset. All analysis procedures are implemented using R software (<http://www.r-project.org/>).

Simulation Study

Simulations are performed to evaluate the power and type I error of the AHS method. Haplotype data are generated based on the HapMap 2 CEU population. There are in total thirty trio families in the original CEU data set. To create an unrelated random population, we select two parents from each family to form an unrelated sample pool and randomly draw cases and controls from this pool with replacement. We focus on the shortest chromosome 22 to facilitate data processing.

We consider six simulation scenarios, as listed in Table 1, with different minor allele frequency (MAF) of the causal variant (0.1 and 0.4) and different LD between the causal variant and its surrounding SNPs (high for average $R^2 > 0.8$, moderate for average $R^2 \approx 0.5$, and low for average $R^2 < 0.2$). Given a causal locus, the three neighboring markers (one marker to its left and two to its right) form a haplotype of size 3 (Figure 1). We then randomly draw two haplotypes with replacement and convert it to the unphased genotype for an individual. The observed genotype of an individual does not include the causal locus, though the trait value is determined by the genotype of the causal marker using the logistic model: $\text{Logit} [\Pr (D=1 \mid G, E)] = \gamma_0 + \gamma_1 \times E + \beta \times G$, where D is the disease status coded as 0 and 1; E is for a binary covariate; G is the genotype of the causal marker coded as 0, 1 and 2. We randomly generate the value for covariate E from Bernoulli distribution with probability 0.5. Then, conditional on G and E , we use the logistic formula given above to determine the disease status of an individual. We repeat this process until obtain the desired sample size (100 cases and 100 controls or 200 cases and 200 controls). The logistic model parameters are set as $\gamma_1 = 1.0$ and $\gamma_0 = -4$ which corresponds to a population disease prevalence of 0.03. For type I error analysis, β is set to be 0 = $\ln(1.0)$, and for power analysis, $\beta = \ln(1.5)$. For each simulated dataset, we obtain the empirical p-value by analyzing 10^4 permutation datasets. In the power analysis, we compare the AHS method to the standard haplotype score test of Schaid et al. (2002) and the multilocus T^2 test of Xiong et al. (2002). The McNemar's test (McNemar, 1947) is used to examine whether the power difference is significant between the proposed method and the two benchmark methods.

Data Application to GAW 15 data

The Genetic Analysis Workshop (GAW) has provided several data sets biannually since 1982. In 2006, GAW15 provided the simulated Rheumatoid Arthritis (RA) SNP data which includes the HLA DR locus and the D locus on chromosome 6. Both loci directly affect the risk of RA but Locus D has a low minor allele frequency of 0.0083. According to the simulation information provided by GAW, smoking status is an important covariate which affects RA risk.

We select one child from each affected sib-pair family to create a case pool. We randomly select N individuals from this case pool and N controls from all the controls in the simulated data. We consider $N = 100$ and 200 in the analysis. Among all the SNPs located on chromosome 6, we select an 8cM region containing 103 SNPs covering the HLA DR locus located at 49.45cM and D locus located at 54.57cM. We use three-SNP sliding windows of haplotypes to scan this region. For each sliding window, we estimate ancestral haplotypes of the cases using DHSMAP and calculate the AHS score for each individual. We then regress the disease status on the AHS score, with smoking status included as a covariate in the regression. We obtain the empirical p-value by 10^6 permutations.

Results

Simulation Study

Table 2 displays the type I error rates of the AHS tests which are calculated on the basis of 1000 simulation replicates having either 100 cases and 100 controls or 200 cases and 200 controls. The values in Table 2 are close to the nominal level 0.05, indicating that our permutation procedure controls type I error. The power analysis for the nominal level 0.05 is shown in Table 3 and Table 4; the results are obtained on the basis of 500 simulation replicates each having either 100 cases and 100 controls, or 200 cases and 200 controls. As expected, if the disease locus is in low LD with its nearby markers and the sample size is small, there is no statistical power to detect the association for each method, and hence the power is around the nominal level. If the disease locus is in moderate or high LD with

surrounding SNPs, we notice that the power performance depends on the disease allele frequency. When compared to the standard regression analysis, the AHS method exhibits no significant power difference if the disease allele frequency is low, although the power figures of the standard regression are often slightly higher. If the disease allele frequency is high, the power of the AHS method is significantly higher. A similar pattern is observed when compare the AHS method to the T^2 test, except that the AHS method has a significantly or slightly higher power across different LD levels and disease allele frequencies. As the T^2 test can not adjust for covariates, such power improvement of the AHS method may be from its use of a smaller degrees of freedom and its incorporation of the covariates in the analysis.

Data Application to the GAW15 data

Using a dataset having 100 cases and 100 controls, the AHS method identifies a fine-scale region of 1cM containing 6 SNPs that are significantly associated with RA disease status. Figure 2 shows the profile of empirical p-values over chromosomal locations. The empirical p-values peak near 49.45cM where the true causal HLA DR locus lies. We also plot p-values from the score test in Figure 2, showing that both methods are capable of detecting the HLA DR signal.

Using a dataset having 200 cases and 200 controls, the AHS method identifies two very close regions. One region is exactly the same region as described above. The other one is a 1cM region containing 8SNPs that is highly linked with the first region. However, for the sample sizes we consider, neither methods detect the second disease loci, the D locus, located 5.12cM downstream from the DR locus on chromosome 6. This may be due to the extremely low minor allele frequency of the D locus (0.0083).

Discussion

In this work, we propose a new haplotype similarity method that combines the evolutionary and two-sample (statistical) approaches to association analysis using haplotypes. Our analysis is based on reconstruction of ancestral haplotypes of cases. For this reason, our approach utilizes the population genetic insight that, if a susceptibility locus is recent, then case participants should share a more recent common ancestor than the control participants. We then use the similarity between an individual's haplotypes and the case ancestral haplotypes in a regression formulation, which gives a simple testing scheme and also allows us to account for covariates that may affect disease status. Because of the bias introduced by constructing ancestral haplotypes among case participants only, and to account for the variability in the reconstruction of the ancestral haplotypes, a permutation procedure is used to evaluate significance. We have shown that our procedure has the proper size, and also has improved power compared with the standard haplotype analysis and the multilocus T^2 test. We define the similarity score based on the number of matching alleles between compared haplotypes, which can be calculated using unphased genotype data, thereby avoiding the estimation of haplotype phase for each individual.

Unlike many haplotype sharing approaches that use a two-sample testing framework with similarity scores of pairwise comparisons, we use a regression model with similarity scores of single individuals. We note here that a regression approach based on pairwise comparison data is also possible (Tzeng et al. 2009), in which the focus is the study of the correlation between the trait similarity and haplotype similarity. This approach also allows for covariates. It may be of interest to understand the relative efficacy of method of this type and the AHS methods. These different features also suggest how to choose a similarity based method in practice. Most of the two-sample based approaches are easy to compute, suitable for binary traits, but often impractical to adjust for covariates. The proposed AHS

method is suitable for binary traits, can incorporate covariates in the analysis, but is relatively more computational intensive. Finally the gene-trait similarity regression of Tzeng et al. (2009) is applicable to quantitative traits or binary traits, can model covariates, but also has higher computational cost than the two-sample based approaches.

In our proposed method we use the decay of haplotype sharing method (DHSMAP) to infer the ancestral haplotypes of cases. However, concerns may arise as to whether the estimated ancestral haplotypes are actually the true ancestral disease haplotypes, especially for complex diseases whose etiology tends to involve complicated interplay between genetic and environmental factors. However, what the AHS method requires is that the estimated ancestral haplotypes be representative of the case haplotypes, so that the similarity scores of each individual reflect similarity to case haplotypes. Further, because we use permutation to ensure proper size, any changes in choice of ancestral haplotypes will only affect the power of the test, not its size. We have been able to observe power gain using the ancestral disease haplotypes obtained from DHS, and further power improvement is possible by using more precisely reconstructed ancestral disease haplotypes.

Acknowledgments

The authors thank the GAW grant, R01-GM031575, for providing the GAW 15 simulated data to be used in this study. Support for generation of the GAW 15 simulated data is provided from NIH grants 5R01-HL049609-14, 1R01-AG021917-01A1, the University of Minnesota, and the Minnesota Supercomputing Institute. Y.L. is partially supported by NSF DMS 0504726. J.Y.T is partially supported by NSF DMS 0504726 and NIH R01MH084022-01A1.

References

- Akey J, Jin L, Xiong M. Haplotypes vs. single marker linkage disequilibrium tests: What do we gain? *Eur. J. Hum. Genet.* 2001; 9:291–300. [PubMed: 11313774]
- Allen AS, Satten GA. A novel haplotype-sharing approach for genome-wide case-control association studies implicates the calpastatin gene in Parkinson's disease. *Genet. Epidemiol.* 2009 To appear in.
- Browning SR. Multilocus association mapping using variable-length Markov chains. *Am. J. Hum. Genet.* 2006; 78:903–913. [PubMed: 16685642]
- de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. Efficiency and power in genetic association studies. *Nat. Genet.* 2005; 37:1217–1223. [PubMed: 16244653]
- Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P, Morris AP. Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *Am. J. Hum. Genet.* 2004; 75:35–43. [PubMed: 15148658]
- Durham LK, Feingold E. Genome scanning for segments shared identical by descent among distant relatives in isolated populations. *Am. J. Hum. Genet.* 1997; 61:830–842. [PubMed: 9382093]
- Epstein MP, Allen AS, Satten GA. A Simple and Improved Correction for Population Stratification in Case-Control Studies. *Am. J. Hum. Genet.* 2007; 80:921–930. [PubMed: 17436246]
- Houwen RHJ, Baharloo S, Blankenship K, Raeymaekers P, Juyn J, Sandkuijl LA, Freimer NB. Genome screening by searching for shared segments: Mapping a gene for benign recurrent intrahepatic cholestasis. *Nat. Genet.* 1994; 8:380–386. [PubMed: 7894490]
- McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika.* 1947; 12:153–157. [PubMed: 20254758]
- McPeck M, Strahs A. Assessment of linkage disequilibrium by the decay of haplotype sharing with application to fine-scale genetic mapping. *Am. J. Hum. Genet.* 1999; 65:858–875. [PubMed: 10445904]
- Molitor J, Marjoram P, Thomas D. Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. *Am. J. Hum. Genet.* 2003a; 73:1368–1384. [PubMed: 14631555]
- Molitor J, Marjoram P, Thomas D. Application of Bayesian spatial statistical methods to analysis of haplotypes effects and gene mapping. *Genet. Epidemiol.* 2003b; 25:95–105. [PubMed: 12916018]

- Morris AP, Whittaker JC, Balding DJ. Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *Am. J. Hum. Genet.* 2002; 70:686–707. [PubMed: 11836651]
- Morris AP, Whittaker JC, Xu CF, Hosking LK, Balding DJ. Multipoint linkage-disequilibrium mapping narrows location interval and identifies mutation heterogeneity. *Proc. Natl. Acad. Sci. U S A.* 2003; 100:13442–13446. [PubMed: 14597696]
- Morris AP. Direct analysis of unphased SNP genotype data in population-based association studies via Bayesian partition modeling of haplotypes. *Genet. Epidemiol.* 2005; 29:91–107. [PubMed: 15940704]
- Nielsen DM, Ehm MG, Zaykin DV, Weir BS. Effect of two- and three-locus linkage disequilibrium on the power to detect marker/phenotype associations. *Genetics.* 2004; 168:1029–1040. [PubMed: 15514073]
- Nolte IM, de Vries AR, Spijker GT, Jansen RC, Brinza D, Zelikovsky A, Meerman GJ. Association testing by haplotype-sharing methods applicable to whole-genome analysis. *BMC Proc.* 2007; 1 Suppl 1:S129. [PubMed: 18466471]
- Schaid DJ. Evaluating associations of haplotypes with traits. *Genet. Epidemiol.* 2004; 27:348–364. [PubMed: 15543638]
- Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.* 2002; 70:425–434. [PubMed: 11791212]
- Seltman H, Roeder K, Devlin B. Evolutionary-based association analysis using haplotype data. *Genet. Epidemiol.* 2003; 25:48–58. [PubMed: 12813726]
- Service SK, Lange DW, Freimer NB, Sandkuijl LA. Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. *Am. J. Hum. Genet.* 1999; 64:1728–1738. [PubMed: 10330361]
- Sha Q, Chen HS, Zhang S. A new association test using haplotype similarity. *Genet. Epidemiol.* 2007; 31:577–593. [PubMed: 17443704]
- Strahs, A.; McPeck, M. Multipoint fine-scale linkage disequilibrium mapping: importance of modeling background LD. In: Goldstein, DR., editor. *Statistics and science: a Festschrift for Terry Speed.* Vol. vol 40. Beachwood: Institute of Mathematical Statistics; 2003. p. 343–366.
- Su SY, Balding DJ, Coin JM. Disease association tests by inferring ancestral haplotypes using a hidden markov model. *Bioinformatics.* 2008; 24:972–978. [PubMed: 18296746]
- The International HapMap Consortium. The International HapMap Project. *Nature.* 2003; 426:789–796. [PubMed: 14685227]
- Thomas DC, Morrison JL, Clayton DG. Bayes estimates of haplotype effects. *Genet. Epidemiol.* 2001 Suppl. 21:S712–S717.
- Thomas DC, Stram DO, Conti D, Molitor J, Marjoram P. Bayesian spatial modeling of haplotype associations. *Hum. Hered.* 2003; 56:32–40. [PubMed: 14614236]
- Tzeng J, Devlin B, Wasserman L, Roeder K. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am. J. Hum. Genet.* 2003a; 72:891–902. [PubMed: 12610778]
- Tzeng JY, Byerley W, Devlin B, Roeder K, Wasserman L. Outlier detection and false discovery rates for whole-genome DNA matching. *J. Am. Stat. Assoc.* 2003b; 98:236–246.
- Tzeng JY. Evolutionary-based grouping of haplotypes in association analysis. *Genet. Epidemiol.* 2005; 28:220–231. [PubMed: 15726584]
- Tzeng JY, Wand CH, Kao JT, Hsiao CK. Regression-based association analysis with clustered haplotypes through use of genotype. *Am. J. Hum. Genet.* 2006; 78:231–242. [PubMed: 16365833]
- Tzeng JY, Zhang D. Haplotype-based association analysis via variance component score test. *Am. J. Hum. Genet.* 2007; 81:927–938. [PubMed: 17924336]
- Tzeng JY, Zhang D, Chang SM, Thomas DC, Davidian M. Gene-trait similarity regression for multimarker-based association analysis. *Biometrics* 2009. 2009 in press.
- Van der Meulen M, te Meerman G. Haplotype sharing analysis in affected individuals from nuclear families with at least one affected offspring. *Genet. Epidemiol.* 1997; 14:915–920. [PubMed: 9433600]

- Xiong M, Zhao J, Eric Boerwinkle E. Generalized T^2 test for genome association studies. *Am. J. Hum. Genet.* 2002; 70:1257–1268. [PubMed: 11923914]
- Zaitlen N, Kang HM, Eskin E, Halperin E. Leveraging the HapMap correlation structure in association studies. *Am. J. Hum. Genet.* 2007; 80:683–691. [PubMed: 17357074]

Figure 1.
D is the causal marker and used for determining traits. M1, M2 and M3 are the markers highly linked with D and used for association test.

Figure 2.

The negative base 10 logarithm of empirical p-values of the proposed AHS method (solid line with filled circles) and those of the standard haplotype score test (dashed line with open circles) around the DR locus. The p-values of the AHS method that are smaller than 10^{-6} are replaced with 10^{-6} in this figure. The p-values of the standard haplotype score test that are smaller than 10^{-10} are replaced with 10^{-10} . The dotted horizontal line is the Bonferroni corrected threshold. The solid vertical line is the location of the DR locus. The dashed vertical line is the location of the D locus. Figure 2(a) shows the results for 100 cases and 100 controls and Figure 2(b) is for 200 cases and 200 controls.

Table 1

Lists of causal loci under the six simulation scenarios

	Low LD ($R^2 < 0.2$)	Moderate LD (R^2 around 0.5)	High LD ($R^2 > 0.8$)
MAF ≈ 0.1	rs134220	rs2858522	rs9614393
MAF ≈ 0.4	rs133914	rs5768355	rs5750870

Table 2

Type I error of the proposed method at nominal level 0.05

	Low LD	Moderate LD	High LD
100 cases and 100 controls			
MAF $\approx$ 0.1	0.055	0.056	0.053
MAF $\approx$ 0.4	0.050	0.045	0.040
200 cases and 200 controls			
MAF $\approx$ 0.1	0.050	0.048	0.054
MAF $\approx$ 0.4	0.052	0.042	0.044

Table 3

Power comparison at nominal level 0.05 with 100 cases and 100 controls

		Low LD	Med LD	High LD
MAF $\approx$ 0.1	T ² Test	0.049	0.112	0.154
	haplo.score	0.055	0.134	0.175
	AHS method	0.058	0.130	0.170
MAF $\approx$ 0.4	T ² Test	0.039	0.147*	0.426*
	haplo.score	0.040	0.160*	0.439*
	AHS method	0.044	0.212	0.530

* Sign '**' indicates the cases where the power of the method is significantly different from the AHS method at the level of 0.05 by McNemar test.

Table 4

Power comparison at nominal level 0.05 with 200 cases and 200 controls

		Low LD	Med LD	High LD
MAF $\approx$ 0.1	T ² Test	0.068	0.219	0.364
	haplo.score	0.088	0.240	0.387
	AHS method	0.091	0.254	0.371
MAF $\approx$ 0.4	T ² Test	0.118	0.337*	0.677*
	haplo.score	0.129	0.350*	0.696*
	AHS method	0.133	0.405	0.771

Sign “*” indicates the cases where the power of the method is significantly different from the power of the AHS method at the level of 0.05 by McNemar test.