Institute of Statistics Mimeo Series # 2611

Retrospective Haplotype Clustering Methods for Detecting Haplotypes Effects and Haplotype-Environment Interactions

Marti L. Jones
PRA International, Charlottesville, VA

Michael P. Epstein
Department of Human Genetics
Emory University School of Medicine, Atlanta, GA

Jau-Tsuen Kao
Department of Clinical Laboratory Sciences and Medical Biotechnology
National Taiwan University, Taipei, Taiwan

Andrew S. Allen
Department of Biostatistics and Bioinformatics
Duke University Medical Center, Durham, NC

Glen A. Satten
Centers for Disease Control and Prevention, Atlanta, GA

Jung-Ying Tzeng*
Department of Statistics and Bioinformatics Research Center
North Carolina State University, Raleigh, NC

* Address for correspondence: Jung-Ying Tzeng, Department of Statistics and Bioinformatics Research Center, North Carolina State University, Campus Box 7566, Raleigh NC, 27695. Tel: 919-513-2723. Fax: 919-515-7315. E-mail: jytzeng@stat.ncsu.edu.

RUNNING TITLE: Retrospective Haplotype Clustering Methods

Abstract

Haplotype based association analysis faces the issues of (a) a large number of degrees of freedom, (b) rare haplotypes, (c) phase missing, and (d) case-control sampling designs. These issues typically reduce power, destabilize inference algorithms, increase complexity and invalidate a number of common assumptions such as that prospective analyses of case-control data can achieve the same statistical efficiency as retrospective analyses. On one hand, previous work has shown that the strategy of clustering haplotypes under the prospective framework can increase the power of haplotype-based association analysis. On the other hand, both theoretical and simulation work have demonstrated that a retrospective framework can be more efficient than a prospective framework under certain conditions. Combining the merits of haplotype clustering and retrospective likelihood framework, this work aims to provide a unified solution to all issues described above. Specifically, we extend the concept of haplotype clustering to the retrospective framework, and generalize the coverage of the clustering approach to allow for global and haplotype-specific testing of both main effects and interaction effects with environment. We derive generalized score statistics to test for haplotype main effects and interaction effects at the global and individual levels. We apply our proposed method to real data from a genetic study of hypertriglyceridemia. Through simulation, we assess the validity of the proposed tests and, where appropriate, compare the power with the power of the retrospective full-dimensional and prospective analyses.

1 Introduction

Haplotype-based association analyses evaluate the association between trait values and the joint variants of several closely linked SNPs. Because haplotypes represent a unit of inheritance and preserve the linkage disequilibrium (LD) among loci, haplotype analyses are generally believed to be more effective in detecting gene-trait association than individual SNP analyses, especially in cases where the causal variants are not measured directly [de Bakker et al., 2005; Zaitlen et al., 2007], are in low frequency [Lin et al., 2004; Schaid, 2004; de Bakker et al., 2005], or exhibit cis-acting effect [Schaid, 2004; Clark, 2004]. In typical haplotype association analyses, all distinct haplotypes are considered as unrelated units and a degree of freedom is assigned to each haplotype except the baseline haplotype. However, because of the way haplotypes evolve, the majority of the polymorphism is concentrated on a relatively small number of haplotypes while the rest is sparsely spread over a fair number of categories. These rare and non-common haplotypes altogether consume valuable degrees of freedom in the analysis, and often lead to either power loss or unstable estimates in practice.

One solution that addresses these issues is haplotype clustering, a method that clusters evolutionarily-close haplotypes and evaluates association of groups of haplotypes. This approach eliminates the excessive degrees of freedom, but does not discard samples of rare haplotypes or arbitrarily lump them into a certain category. Various haplotype clustering methods have been developed, and they can be generally classified into three categories: (a) methods based on haplotype similarity [Molitor et al., 2003; Yu et al., 2004; Waldron and Whittaker, 2006; Li and Jiang, 2005; Li et al., 2006], (b) methods based on evolutionary relationships [Templeton et al., 1987; Templeton, 1995; Seltman et al., 2003; Durrant et al.,

2004; Tzeng, 2005; Tzeng et al., 2006; Liu et al., 2007; Tachmazidou et al., 2007], and (c) methods based on Markov modeling of the LD structure [Browning, 2006; Browning and Browning, 2007; Su et al., 2008].

While haplotype clustering methods overcome the issues of rare haplotypes and large degrees of freedom, existing methods tend to overlook other surrounding issues that arise in haplotype analysis. These issues include retrospectively-sampled data, missing haplotype information, and the needs to evaluate specific effects of haplotypes, environmental covariates, and their interactions. Abundant work has been carried out to address these issues [see Allen and Satten, 2008, and reference therein]. However, most of the approaches only focus on solutions to a subset of these issues, and conflicts often arise when one tries to combine these solutions together to address all issues simultaneously.

Aiming to provide an integrated solution that addresses all these difficulties collectively, in this paper we extend the haplotype clustering method of Tzeng et al. [2006] in several aspects. First, Tzeng et al. [2006] constructed a regression-based clustering approach that cladistically groups rare haplotypes with their corresponding ancestors, and tests for haplotype-trait association while adjusting for covariates and incorporating clustering uncertainty and phase uncertainty. However, the test can only be used for testing the main effects of haplotypes. In this work, we extend the test to assess both haplotype main effects and haplotype interaction effects with environment (H×E interaction hereafter) at both global level and haplotype-specific level. Second, the clustering method of Tzeng et al. [2006] is a score test constructed based on the prospective likelihood of observed genotypes. Motivated by the facts that (a) retrospective methods have been reported to have better power than prospective methods to detect H×E interactions as noted in Kwee et al. [2007] and that (b)

substantial bias can occur in assessing effects at haplotype-specific level if not accounting for the retrospective sampling scheme (Spinka et al. 2005), in this work, we extend the original clustering method from the prospective likelihood framework of Schaid et al. [2002] to a retrospective likelihood framework proposed by Kwee et al. [2007]. As a result, the retrospective clustering method improves power further by reducing the degrees of freedom and also accounts for the biased sampling of the case-control design. Third, the proposed method allows for comparisons among all categories of haplotypes and H×E interactions simultaneously. In other words, it does not require pre-specified haplotypes in order to test the main and H×E effects. This is an improvement over several proposed retrospective methods. Although the existing methods are capable of evaluating haplotype main effects and H×E interactions in theory, their practicality is limited due lack of power and unstable inference from the large degrees of freedom consumed in modeling all main and H×E effects. As a result, these approaches can only be implemented to evaluate effects of some pre-specified haplotypes and/or their interactions with covariates while treating the rest as baseline or assuming that they do not have effects. Our proposed method is not limited by such restrictions. Finally, following the case-only derivation by Kwee et al. [2007], we also construct the retrospective haplotype clustering method for case-only analysis to study H×E interactions.

The rest of the paper is organized as the following. We first introduce the retrospective haplotype clustering model and then derive the likelihood function following the work of Kwee et al. [2007]. We then construct the score tests for evaluating the main effect and H×E interactions globally or individually. We then demonstrate the utility of the proposed method through a data analysis example using the hypertriglyceridemia study and simulations. A discussion section is provided at the end of the paper on the pros and cons of the method.

2 Methods

2.1 Retrospective Clustering Test for Main and H×E Effects

2.1.1 Model

Assume there are totally $L+1$ distinct haplotypes observed in a population and they result in (L^*+1) haplotype clusters. Let $\tilde{\mathbf{X}}_{\mathbf{F}}$ be the over-parameterized $(L+1)^2 \times (L+1)$ design matrix of the full-dimensional data following a certain scoring rule. Each row of $\tilde{\mathbf{X}}_{\mathbf{F}}$ corresponds to a particular ordered haplotype pair and each column corresponds to a haplotype category. For example, consider an ordered haplotype pair $H = (h_1, h_2)$ (i.e., (h_2, h_1) is counted as a separate haplotype pair). The (H, h) element of $\tilde{\mathbf{X}}_{\mathbf{F}}$ is $I(h_1 = h) + I(h_2 = h)$ for multiplicative effect, $I(h_1 = h) \times I(h_2 = h)$ for recessive effect, and $I(h_1 = h) + I(h_2 = h) - I(h_1 = h) \times I(h_2 = h)$ for dominant effect. The function $I(\cdot)$ is an indicator function.

The over-parameterized design matrix of clustered haplotypes, denoted by $\tilde{\mathbf{X}}_{\mathbf{C}}$ (dimension $(L+1)^2 \times (L^*+1)$), is a product of the design matrix for the full-dimensional haplotypes and the clustering allocation matrix, i.e.,

$$\tilde{\mathbf{X}}_{\mathbf{C}} = \tilde{\mathbf{X}}_{\mathbf{F}} \times \tilde{\mathbf{B}}(\mathbf{p}),$$

where $\mathbf{p}$ is the vector of haplotype frequencies of the population, and the matrix $\tilde{\mathbf{B}}(\mathbf{p})$ contains the allocation probabilities that describe how the $(L+1)$ haplotypes are grouped into the (L^*+1) clusters. Note that the allocation probabilities are functions of the haplotype frequencies p_h's. See Appendix A for details on the structure of $\tilde{\mathbf{B}}(\mathbf{p})$.

Define $\mathbf{X}_{\mathbf{C}}$ (dimension $(L+1)^2 \times L^*$) the clustering design matrix with the baseline haplotype category removed, and let X_{C_H} be the row of $\mathbf{X}_{\mathbf{C}}$ that corresponds to a haplotype pair

H. Let vector X_{E_i} (dimension $1 \times k$) be the design vector for the environmental covariates of individual i, and vector X_{HE_i}=$X_{C_H} \bigotimes X_{E_i}$ (dimension $1 \times L^* (k-1)$) be the design matrix for the H×E interactions. The odds of disease for an individual with haplotype pair H and covariate E_i, denoted by $\theta_i (H, E_i)$, is defined as

$$\theta_i(H, E_i) = \frac{\Pr(D_i = 1|H, X_{E_i})}{\Pr(D_i = 0|H, X_{E_i})} = \exp(\beta_0 + X_{C_H}\beta + X_{E_i}\gamma + X_{HE_i}\nu)$$

where β_0 is the baseline effect, β is the vector of L^* haplotype cluster effects, γ is the vector of $k-1$ covariate effects, and ν is the vector of $L^*(k-1)$ interaction parameters.

2.1.2 Retrospective Clustering Likelihood

One challenge in retrospective likelihood approaches is to incorporate environmental covariates, as this requires the specification of the conditional covariates distribution given disease status. Several approaches, including the work of Chatterjee and Carroll [2005], Spinka et al. [2005], Lin and Zeng [2006], Chen and Kao [2006], Kwee et al. [2007] and Chen et al. [2008], have recently been developed to address this difficulty. These approaches differ in how the distribution of the covariate is handled and what assumptions are imposed on the haplotype-environment relationship, Hardy-Weinberg Equilibrium (HWE), and the prevalence of the disease being investigated. Here we focus on the work of Kwee et al. [2007], which assumes haplotype-environment independence and HWE in the target population, and rare disease. The resulting retrospective likelihood can then be factorized into two components: one being the retrospective likelihood of haplotypes given covariates and disease status, and the other being the prospective likelihood of disease given the covariates. This factorization bypasses the needs to specify the distribution of the covariates in

a retrospective analysis. The observed retrospective likelihood can be written as $L_{obs} = \prod_{i=1}^{n} P(G_i, E_i|D_i) \quad = \prod_{i=1}^{n} \sum_{H \in S(G_i)} P(H, E_i|D_i) = \prod_{i=1}^{n} \sum_{H \in S(G_i)} P(H|E_i, D_i)P(E_i|D_i)$, where n is the total number of subjects, G_i is the multilocus genotype of subject i, $S(G_i)$ represents the set of ordered haplotype pairs (h, h') consistent with G_i, E_i is the environmental covariate, and D_i is the disease status for subject i. Kwee et al. [2007] show that the likelihood function is proportional to

$$L_{obs} \propto \prod_{i=1}^{n} \sum_{H \in S(G_i)} P(H|E_i, D_i)P(D_i|E_i). \tag{1}$$

This comes from the result of Prentice and Pyke [1979] that says the retrospective likelihood $P(D_i|E_i)$ is proportional to the prospective likelihood $P(E_i|D_i)$ if we assume a saturated distribution for E. For control subjects, the first term in equation (1) is $P(H = (h, h')\,|E_i, D_i = 0) = P(H = (h, h')\,|D_i = 0) = p_h p_{h'}$.This follows from the previous assumptions of HWE and haplotype-covariate independence in the target population under a rare-disease assumption. For case subjects, it can be shown that $P(H|E_i, D_i = 1) = \theta(H, E_i)P(H|D_i = 0)/\sum_{H'} \theta(H', E_i)P(H'|D_i = 0)$. Finally, the second term $P(D_i|E_i)$ of the likelihood function in equation (1) is $P(D_i = d|E_i) = \theta(E_i)^d/1 + \theta\,(E_i)$, where $\theta\,(E_i) = \sum_H \theta(H, E_i)P(H|D = 0)$.

With these results and following a similar argument of Kwee et al. [2007], the retrospective clustering likelihood can be shown as

$$L_{obs} \propto \prod_{i=1}^{n} \left[\frac{\sum_{H \in S(G_i)} \exp(\beta_0 + X_{C_H}\beta + X_{E_i}\gamma + X_{HE_i}\nu)p_h p_{h'}}{1 + \theta^*(E_i)}\right]^{d_i} \left[\frac{\sum_{H \in S(G_i)} p_h p_{h'}}{1 + \theta^*(E_i)}\right]^{1-d_i}, \tag{2}$$

where $\theta^*(E_i) = \sum_H \exp(\beta_0^* + X_{C_H}\beta + X_{E_i}\gamma + X_{HE_i}\nu)p_h p_{h'}$, i.e., replacing the true intercept β_0

in $\theta(E_i)$ by a modified intercept $\beta_0^* = \beta_0 - \Pr(\text{a case being sampled}) / \Pr(\text{a control being sampled})$, a quantity that can be estimated from the case-control data.

2.1.3 Retrospective Clustering Score Test

Denote η the vector of parameters of interest and ξ the vector of nuisance parameters. For example, to evaluate the effect of haplotype category h, we set $\eta = \beta_h$ and $\xi = (\beta_0^*, \gamma, \mathbf{p}, \beta_{-h})$. The score statistic for testing $H_0 : \eta = 0$ is $S_\eta = U_\eta^T V_\eta^{-1} U_\eta \Big|_{\eta=0, \xi=\hat{\xi}}$, where $U_\eta = \partial \log L_{obs} / \partial \eta$ is the score function, and V_η is the generalized variance function for U_η with $V_\eta = D_{\eta\eta} - I_{\eta\xi} I_{\xi\xi}^{-1} D_{\eta\xi}^T - D_{\eta\xi}^T I_{\eta\xi}^{-1} I_{\eta\xi}^T + I_{\eta\xi} I_{\xi\xi}^{-1} D_{\xi\xi} I_{\xi\xi}^{-1} I_{\eta\xi}^T$ [Boos, 1992]. Matrix D is the variance-covariance matrix of the score function $U = (U_\eta, U_\xi)^T$ and matrix I is the observed Fisher information matrix that is constructed by taking the derivatives of U. See Appendix B for detailed derivation of U_η and V_η, and Appendix C for parameter estimation for the global test and individual test of the main and H×E effects. Under the null hypothesis, the score statistic S_η follows asymptotically a χ^2_κ-distribution, with κ being the length of η. One key feature of the score test derived here is that it evaluates a specific effect (e.g. β_h) by only testing the hypothesis of zero coefficients of the target parameters (e.g., $\mathrm{H}_0 : \beta_h = 0$), and leaving all other effects unconstrained (e.g. not setting $\beta_{-h} = 0$).

2.2 Case-only Methods for H×E Interactions

Interactions between haplotypes and environmental covariates can also be assessed using only case samples. The case-only interaction studies have several practical applications, more commonly seen in cancer therapy efficacy or drug adverse event studies. In these cases, researchers may be interested in studying how genetic factors modify the effect of treatment,

but may only have data from those who responded to the treatment.

2.2.1 Case-only Retrospective Likelihood for H× E Interactions

In the full-dimensional analysis, Kwee et al. [2007] showed that the retrospective likelihood function factors into $P(G \mid E, D=1)$, $P(G \mid E, D=0)$ and $P(E \mid D)$, where the latter two terms contain no information about β and ν by assuming haplotype-environment independence and a saturated distribution for the covariate. They also showed that, under a multiplicative genetic effect, $P(G \mid E, D=1)$ depends on both the parameter ν and a transformed parameter involving β and $\mathbf{p}$. This result makes it possible to make inference on ν by maximizing the case-only likelihood component $P(G \mid E, D=1)$.

Here we show that a similar property of $P(G \mid E, D=1)$ also exists for retrospective clustering likelihood under a multiplicative effect. Consider

$$P(G|E,D=1) = \sum_{H \in S(G)} P(H|E,D=1) = \frac{\sum_{H \in S(G)} \theta(H,E) P(H|D=0)}{\sum_{H'} \theta(H',E) P(H'|D=0)}$$
$$= \frac{\sum_{H=(h,h') \in S(G)} \exp(X_{C_H}\beta + X_{HE}\nu) p_h p_{h'}}{\sum_{H^*=(h^*,h'^*)} \exp(X_{C_{H'}}\beta + X_{H'E}\nu) p_{h^*} p_{h'^*}}. \quad (3)$$

We can write $\exp(X_{C_H}\beta)$ as $\exp((\mathbf{B}(\mathbf{p})[h,] + \mathbf{B}(\mathbf{p})[h',])\beta)$, where $\mathbf{B}(\mathbf{p})_{(L+1)\times L^*}$ is the clustering allocation matrix $\tilde{\mathbf{B}}(\mathbf{p})_{(L+1)\times(L^*+1)}$ with the first column (i.e., the baseline haplotype group) removed, and $\mathbf{B}(\mathbf{p})[h,]$ (dimension $1 \times L^*$) is the hth row of $\mathbf{B}(\mathbf{p})$. Incorporating this into (3), we rewrite $P(G|E,D=1)$ as

$$\frac{\sum_{H=(h,h') \in S(G)} \exp(\sum_{c=1}^{L^*} \mathbf{B}(\mathbf{p})[h,c]\beta_c) \exp(\sum_{c=1}^{L^*} \mathbf{B}(\mathbf{p})[h',c]\beta_c) \exp(X_{HE}\nu) p_h p_{h'}}{\sum_{H^*=(h^*,h'^*)} \exp(\sum_{c=1}^{L^*} \mathbf{B}(\mathbf{p})[h^*,c]\beta_c) \exp(\sum_{c=1}^{L^*} \mathbf{B}(\mathbf{p})[h'^*,c]\beta_c) \exp(X_{H'E}\nu) p_{h^*} p_{h'^*}}.$$

Define the quantity $\tilde{p}_h = \exp(\sum_{c=1}^{L^*} B(p)[h,c]\beta_c) p_h / \sum_{h^*} \exp(\sum_{c=1}^{L^*} B(p)[h^*,c]\beta_{c^*}) p_{h^*}$, the

case-only likelihood function becomes

$$L_{obs-\text{case}} \propto \prod_{i=1}^{d} \frac{\sum_{H=(h,h')\in S(G_i)} \exp(X_{HE_i}\nu)\tilde{p}_h\tilde{p}_{h'}}{\sum_{H'=(h^*,h'^*)} \exp(X_{H'E_i}\nu)\tilde{p}_{h^*}\tilde{p}_{h'^*}} \tag{4}$$

where d is the number of cases. Therefore just like the full-dimensional analysis, the inference of interaction parameters ν under the clustering analysis can also be performed using the case-only likelihood function in equation (4).

3 Data Application to Hypertriglyceridemia Study

We applied the proposed retrospective clustering method to the hypertriglyceridemia study conducted by Kao et al. [2003]. The study consisted of 303 controls and 290 cases (defined as individuals having serum triglycerides >400mg/dl). For each subject, demographic information and genotypes information of the five SNPs in APOA5 gene (IVS3+476, c.457, c.553, c.1177, and c.1259) were collected. This dataset was previously analyzed by Tzeng et al. [2006] and Chen and Kao [2006]. Tzeng et al. [2006] reported that there are 12 haplotypes with frequency $>1 \times 10^{-5}$ in the APOA5 region based on EM estimates of haplotype frequencies, among which the four most frequent haplotypes, GGGCT, GGTCT, AGGCC, and GAGTT, explained 95.8% of the total variation. Using haplotype GGGCT as the baseline, both Tzeng et al. [2006] and Chen and Kao [2006] reported that haplotypes GGTCT, AGGCC, and GAGTT were significantly associated with the disease risk. Among these haplotypes, GGTCT and AGGCC were shown elsewhere to be associated with increased plasma triglyceride concentration (Pennacchio et al. 2001; Kao et al. 2003). Chen and Kao [2006] further identified the interacting effect between GGTCT and age.

Motivated by these results, we conducted a series of main-effect and haplotype-age interaction analyses to further explore the specific effects of haplotypes. To better illustrate the potential H×E effects, we dichotomized age into old and young using the control mean age of 49 as the cutoff. We also excluded participants with no covariate information or with no genetic information at all 5 SNPs, resulting a sample of 210 cases and 287 controls for further analysis. The clustering algorithm created 4 haplotype clusters represented by the 4 most frequent haplotypes, and the results of the retrospective clustering tests are shown in Table 1.

For main-effect analysis, the global test is highly significant (p-value $< 1 \times 10^{-6}$). To identify the cause for global significance, we investigated the relative effects of all haplotype pairs via a series of haplotype specific tests. The reasoning is that the haplotype polymorphism is, in essence, one "factor" with L^* "levels"; hence ideally a pair-wise analysis similar to the post-hoc analysis in ANOVA should be conducted to reveal the pattern of effects. To do so, each haplotype group was taken to be the reference in a different regression model and the significance of the coefficients in the models was examined. There were totally 6 comparisons and all of them were highly significant except one (i.e., AGGCC vs. GAGTT) using a Bonferroni corrected α level of $0.05/6 = 8.3 \times 10^{-3}$. This finding suggests that, in addition to previous findings in the literature, AGGCC and GAGTT should be categorized into the same level as they have the same effect on disease risk.

For interaction analysis, we did not observe a significant effect. The global test for haplotype and age interaction is close to significant at the nominal level of 0.05 (p-value 0.08). For illustrational purposes, we carried out the specific tests (see Table 1). Relatively speaking, the interaction effect of age×GGTCT is more significantly different from that

between age×GGGCT/AGGCC, whereas the rest interaction effects are the same. However these results do not have statistical significance. Our results are different from what was found by Chen and Kao [2006], who reported a significant difference between age× GGTCT and age×GGGCT. However, Chen and Kao [2006] used a model that assumed no interactions between age and other haplotypes (i.e., $\nu_{GAGTT} = \nu_{AGGCC} = 0$). When we systematically explored the global interaction effects with all potential interaction terms unconstrained, we did not replicate their finding.

4 Simulations

4.1 Design

We conducted simulations to assess the performance of the proposed retrospective clustering methods for testing the main and H×E effects at the global and the individual levels. The details of the simulation scheme are given below. In each simulation scenario, we evaluated type I error and power of the reduced-dimension analysis by retrospective clustering method (referred to as the retro-RD method). When applicable, we compare the RD performance to the full-dimension analysis of Kwee et al. [2007] (referred to as the retro-FD method). The simulated data set contains 500 cases and 500 controls.

Global-level Test. We first generated 100 haplotypes under the coalescent model using the program by Wall and Prichard [2003]. The SNP sequences were generated with an effective population size of 10^4, a scaled mutation rate of 5.6×10^{-4} (per bp), a scaled recombination rate around 6×10^{-3} (per bp) for the cold spots and 45 times greater for the hot spots. These parameters are chosen to produce a similar number of common SNPs to the

European American sample in the SeattleSNP database and to mimic the LD pattern of the SELP gene observed there. We then selected 6 different disease loci that represented different level of allele frequency (0.1 or 0.3) and haplotype diversity around the chosen loci (high, moderate, or low). Given a disease locus, we defined a 6-SNP haplotype region consisting of the 3 adjacent SNPs directly to the left and right of the disease locus. The disease locus was excluded from the haplotype sequence. The numbers of distinct haplotypes in the high, moderate and low diversity regions were about 10–16, 9–12, and 5–8, respectively.

For each individual, we obtained the haplotype pair by randomly selected 2 haplotypes with replacement from the 100 haplotypes, and generated the binary covariate E from Bernoulli(0.5). The disease status D was then determined according to $\text{logit}\Pr(D = 1 \mid G^*, E) = \beta_0 + \beta G^* + \gamma E + \nu G^* \times E$, where G^* is the number of causal alleles at the disease locus. In all scenarios, we set $\gamma = 0.3$ and determined β_0 according to the effect size and the allele frequency so that the disease prevalence was approximately 5%. We set $\beta = 0.5$ (i.e., OR = 1.65) and $\nu = 0$ for power analysis of the global main effect test, and $\beta = 0.5$ and $\nu = 0.7$ for power analysis of the global interaction test. With this set of simulation, we also compared the performance of the retrospective approaches (retro-RD and retro-FD) to the performance of the full-dimensional prospective approaches (referred to the prosp-FD method) as implemented in Schaid et al. [2002] and the reduced-dimensional prospective method (referred to the prosp-RD method) by Tzeng et al. [2006].

Individual-level Test. The simulation scheme remains the same in general as previously described with the following modifications. First, instead of considering the 6 scenarios, we focused on an arbitrarily chosen region that contained 10 haplotypes. Second, we determined the disease status using the haplotype information instead of the genotype at

the disease locus. We assumed 2 causal haplotypes H_1 and H_2 with frequencies 0.12 and 0.08, respectively, and the disease probability was determined by $\text{logit}\Pr(D = 1 \mid G, E) = \beta_0 + \beta_1 H_1 + \beta_2 H_2 + \gamma E + \nu_1 H_1 \times E + \nu_2 H_2 \times E$. For the main effect test, we considered $(\beta_1, \beta_2) = (0.5, 0.5)$, $(0.7, 0.5)$, and $(0.7, 0.3)$. For the interaction test, we considered $(\nu_1, \nu_2) = (0.5, 0.5)$ and $(0.7, 0.5)$ and set $(\beta_1, \beta_2) = (0.5, 0.5)$. Following the same setup as in the global test, we set $\gamma = 0.3$, and varied β_0 to maintain a 5 % disease prevalence.

4.2 Results

Global-level Test. Tables 2 lists the type I error rate for the global main-effect tests and the global H ×E tests based on 2000 replications. The type I error rate for the case-only H×E tests are shown in the parenthesis. Overall speaking, the observed type I error rates are around the nominal level for each of the analyses, suggesting that the χ^2 distribution approximates the asymptotic distribution of the score statistics reasonably well.

Results of the power analysis (based on 1000 replications) are shown in Table 3 (for main effects) and Table 4 (for interactions). The retro-FD method is used as the benchmark except in the cases of global H×E interaction analyses with high/moderate diversity, where the retro-FD method failed to converge in estimating the high-dimensional β (i.e., all haplotype main effect parameters). In scenarios where we were able to conduct both the retro-RD and retro-FD analyses, we observed that the retro-RD method produced higher power than the retro-FD method. The power improvement was consistent for the main-effect tests and for the H×E tests (low diversity). As expected, the degrees of power improvement depended on the haplotype diversity. In the low diversity regions, the power difference between retro-RD and retro-FD was the smallest since the total number of the full-dimensional haplotypes

was not large and the reduction in the degrees of freedom was small. Nevertheless we still observed similar or slightly improved power for both the main-effect and H×E analyses. The overall pattern suggests that the clustering strategy could further improve power when the haplotype diversity is moderate or high.

Table 4 displays the power of retro-RD test for global interactions using case samples only (see figures in the parenthesis). The power of the case-only test is very similar to the case-control test (under multiplicative effect). This result provides validation to the case-only retro-RD test, and it matches with what Kwee et al. [2007] reported in full-dimensional analysis case.

With global main-effect analysis, we also explored the relative power between the prospective and retrospective approaches (Table 3). In both analyses, the RD tests were more powerful than the FD tests. However, it is interesting to see that there is little difference between the retrospective methods and the prospective methods. For the FD analysis, Satten and Epstein [2004] found the retrospective and the prospective likelihood approaches have similar power under multiplicative effects. The simulation results indicate that the same situation holds for the RD analysis.

Individual-level Test. The simulation results for the haplotype specific tests are presented in Table 5 for the main-effect tests and 6 for the H×E tests. The observed type I error rates suggest that the haplotype specific test is slightly conservative for both the main-effect tests and H×E tests (first column of both tables). This is likely caused by the fact that a precise evaluation of a specific effect would require a more accurate estimate for the nuisance effects. However, with the parameter space being relatively large, a larger sample size would be needed to obtain a more reliable estimate and to achieve the asymptote. Also

because a large number of nuisance parameters (all β_{-h} for main-effect analysis and all β_h's and ν_{-h} for interaction analysis) need to be estimated, here we encountered the same estimation difficulty as in the global tests. As a result, the benchmark method, retro-FD, was not applicable and only the power of the retro-RD methods was reported. Nevertheless, we observed that the haplotype specific tests had acceptable power for both the main-effect and interaction analyses, and as expected, the power tend to be higher when the causal variants had higher frequencies or larger effect.

5 Discussion

We have proposed a method that addresses one of the major limitations to the usefulness of haplotype analysis for detecting genetic associations in complex diseases. Our method reduces the degrees of freedom by clustering haplotypes and carrying out inference based on a core set of haplotypes. The method uses unphased genotype data and can incorporate environmental covariates, which is important when studying complex diseases. The proposed method has greater power than the retrospective full-dimensional approach, evidence that reducing the degrees of freedom through clustering improves the performance of haplotype analysis. The greater the dimension reduction due to clustering, the larger the difference in power between the clustering and full-dimensional approaches.

One significant contribution of this work is the ability to test for global interaction effects. As the study of complex diseases becomes an important focus of scientific research, it is critical to develop methods for genetic analysis that can incorporate environmental covariates and interactions between genetic factors and these covariates. Including interaction

terms involving the full-dimensional space of haplotypes limits power and causes estimation difficulties for low frequency haplotypes. As a result, many full-dimensional methods are restricted to only including interaction terms involving certain pre-determined haplotypes of interest. Our method works with a reduced parameter space, thereby making a global test more practical. Thus we can employ the testing strategy of first testing for a global interaction effect and then testing for specific haplotype-environment effects. We have presented score statistics to test for both the global and specific haplotype-environment interaction effects.

Established on the likelihood framework of Kwee et al. [2007], the proposed retrospective clustering method also inherits its three major assumptions: (a) rare disease, (b) HWE in the target population, and (c) the haplotype-environment independence. For (a), while a rare disease may be often caused by rare variants, the proposed clustering methods only evaluate the effect of common haplotype variants. We do not expect too much contradiction here as that rare causal variants do not necessarily correspond to rare haplotypes [e.g., Hugot et al., 2001], and that rare causal variants tend to be oversampled in a case-control study. In addition, Kwee et al. [2007] found that the retrospective method assuming a multiplicative disease model is robust to the assumption of a rare disease. Case-control studies are usually used when the disease prevalence is less than 10 %, and they found the method has similar size and power for disease prevalence of 5 % and 10 %.

For (b) the HWE assumption, Satten and Epstein [2004] showed that the retrospective approach with a multiplicative model is robust to the assumption of HWE in the target population. When needed, they also propose a method to model departure from HWE due to inbreeding and population stratification by incorporating a fixation index. The extension

of our proposed method to this approach is relatively straightforward.

Finally, Spinka et al. [2005] and Kwee et al. [2007] found that the retrospective methods are sensitive to this assumption of haplotype-environment independence. We expect this assumption to be valid in many cases, but if there is evidence that the environmental covariate is also influenced by genetic factors, other methods should be used. Spinka et al. [2005] propose a modified prospective approach that is similar to the estimating equation method of Zhao et al. [2000]. This method is more robust to the haplotype-environment independence assumption. Chen et al. [2008] develop a method that allows a direct relationship between haplotypes and environmental covariates. It may be interesting to see if these methods can incorporate haplotype clustering, resulting in a method that can be used if there is a known relationship between haplotypes and covariates.

Acknowledgement

This work was sponsored by NSF grant DMS 0504726 (to M.L.J and J.Y.T), NIH grant MH074027-01A1 (to J.Y.T.), and NIH grant HG003618 (to M.P.E).

Appendix A: Structure of the $\tilde{\mathbf{B}}(\mathbf{p})$ matrix

The clustering algorithm of Tzeng [2005] and Tzeng et al. [2006] sequentially combines "rare" haplotypes into their one-step neighboring haplotypes, from the tips of the (unobserved) evolutionary tree toward the major nodes. Each of the resulting cluster is represented by the most common haplotype, and haplotypes within a cluster are assumed to have the same effect on the disease trait. Denote H_F as the full set of observed haplotypes, and

H_C as the set of clustered haplotypes. The algorithm first partitions the list H_F into $H^{(0)} = H_C$ (which is determined using the penalized cumulative Shannon's information content), and $H^{(j)}$(a set of haplotypes that are different from the $H^{(0)}$by j steps of mutation) with $j = 1, 2, ...J$. Starting from $j = J$ to $j = 1$, group each element of $H^{(j)}$ to its one-step ancestor in $H^{(j-1)}$ and combine the frequencies. The grouping rule is specified according to the branch probabilities that are stored in the allocation matrix $\tilde{\mathbf{B}}(\mathbf{p})^{(j)}$; each row of $\tilde{\mathbf{B}}(\mathbf{p})^{(j)}$ describes to whom and how a certain haplotype of $H^{(j)}$ is allocated among $H^{(j-1)}$. The branch probability is a function of haplotype similarity (reflecting the relatedness) and haplotype frequency (reflecting the age of haplotype). Let $\tilde{\mathbf{X}}_\mathbf{F}$ and $\tilde{\mathbf{X}}_\mathbf{C}$ as defined in the text; corresponding to $H^{(j)}$ we can decomposed $\tilde{\mathbf{X}}_\mathbf{F} = \left[\tilde{\mathbf{X}}_\mathbf{F}^{(0)}, \tilde{\mathbf{X}}_\mathbf{F}^{(1)}, \ldots, \tilde{\mathbf{X}}_\mathbf{F}^{(j)}, \ldots, \tilde{\mathbf{X}}_\mathbf{F}^{(J)}\right]$. The stepwise grouping process can be shown equivalent to the matrix operation $X^{(j)}\tilde{\mathbf{B}}(\mathbf{p})^{(j)}$, and the overall process can be described as $\tilde{\mathbf{X}}_\mathbf{C} = \tilde{\mathbf{X}}_\mathbf{F}{}^{(0)} + \tilde{\mathbf{X}}_\mathbf{F}{}^{(1)}\tilde{\mathbf{B}}(\mathbf{p})^{(1)} + \tilde{\mathbf{X}}_\mathbf{F}{}^{(2)}\tilde{\mathbf{B}}(\mathbf{p})^{(2)}\tilde{\mathbf{B}}(\mathbf{p})^{(1)} + \cdots + \tilde{\mathbf{X}}_\mathbf{F}^{(J)}\tilde{\mathbf{B}}(\mathbf{p})^{(J)}\tilde{\mathbf{B}}(\mathbf{p})^{(J-1)} \cdots \tilde{\mathbf{B}}(\mathbf{p})^{(2)}\tilde{\mathbf{B}}(\mathbf{p})^{(1)}$. Or, equivalently, $\tilde{\mathbf{X}}_\mathbf{F} = \tilde{\mathbf{X}}_\mathbf{F}\tilde{\mathbf{B}}(\mathbf{p})$,where $\tilde{\mathbf{B}}(\mathbf{p}) = \begin{bmatrix} \mathbf{I} \\ \hline \tilde{\mathbf{B}}(\mathbf{p})^{(1)} \\ \hline \tilde{\mathbf{B}}(\mathbf{p})^{(2)}\tilde{\mathbf{B}}(\mathbf{p})^{(1)} \\ \hline \vdots \end{bmatrix}$. For further detail please see Tzeng [2005].

Appendix B: Score function U_η and variance function V_η

Haplotype main-effect tests

For main-effect tests, $\eta = \beta$ and $\xi = (\alpha^*, \gamma, \mathbf{p})$. Let L_i be the observed likelihood function for individual i. The goal is to calculate the score function $U_\beta = \sum_{i=1}^n \frac{\partial}{\partial \beta} \log L_i$ and its variance $V_\beta = D_{\beta\beta} - I_{\beta\xi} I_{\xi\xi}^{-1} D_{\beta\xi}^T - D_{\beta\xi} I_{\xi\xi}^{-1} I_{\beta\xi}^T + I_{\beta\xi} I_{\xi\xi}^{-1} D_{\xi\xi} I_{\xi\xi}^{-1} I_{\beta\xi}^T$, where $D_{\theta_1\theta_2} = \sum_{i=1}^n \frac{\partial}{\partial \theta_1} \log L_i \frac{\partial}{\partial \theta_2} \log L_i$

and $I_{\theta_1\theta_2} = -\frac{\partial}{\partial\theta_1}\frac{\partial}{\partial\theta_2}\log L_i$ with $\theta_\ell \in \{\beta, \alpha^*, \gamma, \mathbf{p}\}$. Define the quantities $v = \sum_{H\in S(G_i)} \exp(\alpha^* + X_{C_H}\beta + X_{E_i}\gamma)p_h p_{h'}$, $u = \sum_H \exp(\alpha^* + X_{C_H}\beta + X_{E_i}\gamma)p_h p_{h'}$, and $w = \sum_{H\in S(G_i)} p_h p_{h'}$. Then the individual observed log likelihood can be simplified as $\log L_i = \text{c} + d_i \log v + (1 - d_i)\log w - \log(1 + u)$, where c is a constant. Hence the components of U_β and V_β can be obtained by noting that

$$\frac{\partial}{\partial\beta}\log L_i = d_i\frac{\frac{\partial}{\partial\beta}v}{v} - \frac{\frac{\partial}{\partial\beta}u}{u}$$

$$\frac{\partial}{\partial\alpha^*}\log L_i = d_i\frac{\frac{\partial}{\partial\alpha^*}v}{v} - \frac{\frac{\partial}{\partial\alpha^*}u}{u}$$

$$\frac{\partial}{\partial\gamma}\log L_i = d_i\frac{\frac{\partial}{\partial\gamma}v}{v} - \frac{\frac{\partial}{\partial\gamma}u}{u}$$

$$\frac{\partial}{\partial p_\tau}\log L_i = (1 - d_i)\frac{\frac{\partial}{\partial p_\tau}w}{w} + d_i\frac{\frac{\partial}{\partial p_\tau}v}{v} - \frac{\frac{\partial}{\partial p_\tau}u}{u}.$$

And also that

$$\frac{\partial}{\partial\alpha^*}\frac{\partial}{\partial\alpha^*}\log L_i = d_i\frac{v(\frac{\partial}{\partial\alpha^*}\frac{\partial}{\partial\alpha^*}v) - (\frac{\partial}{\partial\alpha^*}v)(\frac{\partial}{\partial\alpha^*}v)}{v^2} - \frac{u(\frac{\partial}{\partial\alpha^*}\frac{\partial}{\partial\alpha^*}u) - (\frac{\partial}{\partial\alpha^*}u)(\frac{\partial}{\partial\alpha^*}u)}{u^2}$$

$$\frac{\partial}{\partial\beta}\frac{\partial}{\partial\alpha^*}\log L_i = d_i\frac{v(\frac{\partial}{\partial\beta}\frac{\partial}{\partial\alpha^*}v) - (\frac{\partial}{\partial\alpha^*}v)(\frac{\partial}{\partial\beta}v)}{v^2} - \frac{u(\frac{\partial}{\partial\beta}\frac{\partial}{\partial\alpha^*}u) - (\frac{\partial}{\partial\alpha^*}u)(\frac{\partial}{\partial\beta}u)}{u^2}$$

$$\frac{\partial}{\partial\gamma}\frac{\partial}{\partial\alpha^*}\log L_i = d_i\frac{v(\frac{\partial}{\partial\gamma}\frac{\partial}{\partial\alpha^*}v) - (\frac{\partial}{\partial\alpha^*}v)(\frac{\partial}{\partial\gamma}v)}{v^2} - \frac{u(\frac{\partial}{\partial\gamma}\frac{\partial}{\partial\alpha^*}u) - (\frac{\partial}{\partial\alpha^*}u)(\frac{\partial}{\partial\gamma}u)}{u^2}$$

$$\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\alpha^*}\log L_i = d_i\frac{v(\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\alpha^*}v) - (\frac{\partial}{\partial\alpha^*}v)(\frac{\partial}{\partial p_\tau}v)}{v^2} - \frac{u(\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\alpha^*}u) - (\frac{\partial}{\partial p_\tau}u)(\frac{\partial}{\partial\alpha^*}u)}{u^2}$$

$$\frac{\partial}{\partial\beta}\frac{\partial}{\partial\beta}\log L_i = d_i\frac{v(\frac{\partial}{\partial\beta}\frac{\partial}{\partial\beta}v) - (\frac{\partial}{\partial\beta}v)(\frac{\partial}{\partial\beta}v)}{v^2} - \frac{u(\frac{\partial}{\partial\beta}\frac{\partial}{\partial\beta}u) - (\frac{\partial}{\partial\beta}u)(\frac{\partial}{\partial\beta}u)}{u^2}$$

$$\frac{\partial}{\partial\gamma}\frac{\partial}{\partial\beta}\log L_i = d_i\frac{v(\frac{\partial}{\partial\gamma}\frac{\partial}{\partial\beta}v) - (\frac{\partial}{\partial\beta}v)(\frac{\partial}{\partial\gamma}v)}{v^2} - \frac{u(\frac{\partial}{\partial\gamma}\frac{\partial}{\partial\beta}u) - (\frac{\partial}{\partial\beta}u)(\frac{\partial}{\partial\gamma}u)}{u^2}$$

$$\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\beta}\log L_i = d_i\frac{v(\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\beta}v) - (\frac{\partial}{\partial\beta}v)(\frac{\partial}{\partial p_\tau}v)}{v^2} - \frac{u(\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\beta}u) - (\frac{\partial}{\partial\beta}u)(\frac{\partial}{\partial p_\tau}u)}{u^2}$$

$$\frac{\partial}{\partial\gamma}\frac{\partial}{\partial\gamma}\log L_i = d_i\frac{v(\frac{\partial}{\partial\gamma}\frac{\partial}{\partial\gamma}v) - (\frac{\partial}{\partial\gamma}v)(\frac{\partial}{\partial\gamma}v)}{v^2} - \frac{u(\frac{\partial}{\partial\gamma}\frac{\partial}{\partial\gamma}u) - (\frac{\partial}{\partial\gamma}u)(\frac{\partial}{\partial\gamma}u)}{u^2}$$

$$\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\gamma}\log L_i = d_i\frac{v(\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\gamma}v) - (\frac{\partial}{\partial\gamma}v)(\frac{\partial}{\partial p_\tau}v)}{v^2} - \frac{u(\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\gamma}u) - (\frac{\partial}{\partial\gamma}u)(\frac{\partial}{\partial p_\tau}u)}{u^2}$$

$$\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial p_\tau}\log L_i = (1 - d_i)\frac{w(\frac{\partial^2}{\partial p_\tau^2}w) - (\frac{\partial}{\partial p_\tau}w)(\frac{\partial}{\partial p_\tau}w)}{w^2} + d_i\frac{v(\frac{\partial^2}{\partial p_\tau^2}v) - (\frac{\partial}{\partial p_\tau}v)(\frac{\partial}{\partial p_\tau}v)}{v^2} - \frac{u(\frac{\partial^2}{\partial p_\tau^2}u) - (\frac{\partial}{\partial p_\tau}u)(\frac{\partial}{\partial p_\tau}u)}{u^2}$$

$$\frac{\partial}{\partial p_\theta}\frac{\partial}{\partial p_\tau}\log L_i = (1 - d_i)\frac{w(\frac{\partial}{\partial p_\theta}\frac{\partial}{\partial p_\tau}w) - (\frac{\partial}{\partial p_\tau}w)(\frac{\partial}{\partial p_\theta}w)}{w^2} + d_i\frac{v(\frac{\partial}{\partial p_\theta}\frac{\partial}{\partial p_\tau}v) - (\frac{\partial}{\partial p_\tau}v)(\frac{\partial}{\partial p_\theta}v)}{v^2} - \frac{u(\frac{\partial}{\partial p_\theta}\frac{\partial}{\partial p_\tau}u) - (\frac{\partial}{\partial p_\tau}u)(\frac{\partial}{\partial p_\theta}u)}{u^2}$$

Where the derivatives of v with respect to the parameters α^*, γ, β, and a specific p_τ are

$\frac{\partial}{\partial \alpha^*} v = \sum_{H \in S(G_i)} \exp(\alpha^* + X_{C_H}\beta + X_{E_i}\gamma) p_h p_{h'}$

$\frac{\partial}{\partial \beta} v = \sum_{H \in S(G_i)} \exp(\alpha^* + X_{C_H}\beta + X_{E_i}\gamma) X_{C_H}^T p_h p_{h'}$

$\frac{\partial}{\partial \gamma} v = \sum_{H \in S(G_i)} \exp(\alpha^* + X_{C_H}\beta + X_{E_i}\gamma) X_{E_i}^T p_h p_{h'}$

$\frac{\partial}{\partial p_\tau} v = \sum_{H \in S(G_i)} 2I(h = \tau) \left\{ p_{h'} exp(\alpha^* + X_{C_H}\beta + X_{E_i}\gamma) \right\}$

$\frac{\partial}{\partial \alpha^*} \frac{\partial}{\partial \alpha^*} v = v$

$\frac{\partial}{\partial \beta} \frac{\partial}{\partial \alpha^*} v = \frac{\partial}{\partial \beta} v$

$\frac{\partial}{\partial \gamma} \frac{\partial}{\partial \alpha^*} v = \frac{\partial}{\partial \gamma} v$

$\frac{\partial}{\partial p_\tau} \frac{\partial}{\partial \alpha^*} v = \frac{\partial}{\partial p_\tau} v$

$\frac{\partial}{\partial \beta} \frac{\partial}{\partial \beta} v = \sum_{H \in S(G_i)} \exp(\alpha^* + X_{C_H}\beta + X_{E_i}\gamma) X_{C_H}^T X_{C_H} p_h p_{h'}$

$\frac{\partial}{\partial \gamma} \frac{\partial}{\partial \beta} v = \sum_{H \in S(G_i)} \exp(\alpha^* + X_{C_H}\beta + X_{E_i}\gamma) X_{C_H}^T X_{E_i} p_h p_{h'}$

$\frac{\partial}{\partial p_\tau} \frac{\partial}{\partial \beta} v = \sum_{H \in S(G_i)} 2I(h = \tau) \left\{ \exp(\alpha^* + X_{C_H}\beta + X_{E_i}\gamma) X_{C_H}^T p_{h'} \right\}$

$\frac{\partial}{\partial \gamma} \frac{\partial}{\partial \gamma} v = \sum_{H \in S(G_i)} \exp(\alpha^* + X_{C_H}\beta + X_{E_i}\gamma) X_{E_i}^T X_{E_i} p_h p_{h'}$

$\frac{\partial}{\partial p_\tau} \frac{\partial}{\partial \gamma} v = \sum_{H \in S(G_i)} 2I(h = \tau) \left\{ \exp(\alpha^* + X_{C_H}\beta + X_{E_i}\gamma) X_{E_i}^T p_{h'} \right\}$

$\frac{\partial}{\partial p_\tau} \frac{\partial}{\partial p_\tau} v = I(h = h' = \tau) \left\{ 2 \exp(\alpha^* + X_{C_H}\beta + X_{E_i}\gamma) \right\}$

$\frac{\partial}{\partial p_\theta} \frac{\partial}{\partial p_\tau} v = \sum_{H \in S(G_i)} 2I(h_j = \tau, h'_j = \theta) \left\{ \exp(\alpha^* + X_{C_H}\beta + X_{E_i}\gamma) \right\}$

The derivatives of u are the same as those for v, except they are summed over all haplotype pairs H instead of $H \in S(G_i)$. The derivates of w are

$\frac{\partial}{\partial p_\tau} w = \sum_{H \in S(G_i)} 2I(h = \tau) p_{h'}$

$\frac{\partial}{\partial p_\tau} \frac{\partial}{\partial p_\tau} w = 2I(h = h' = \tau)$

$\frac{\partial}{\partial p_\theta} \frac{\partial}{\partial p_\tau} w = \sum_{H \in S(G_i)} 2I(h_j = \tau, h'_j = \theta)$

H×E tests

For H×E tests, $\eta = \nu$ and $\xi = (\alpha^*, \gamma, \beta, \mathbf{p})$. The definitions v, u, and w are same to the case of main-effect tests, except that for v and u there is an extra quantity $X_{HE_i}\nu$ involved in the exponential term, e.g., $v = \sum_{H \in S(G_i)} \exp(\alpha^* + X_{C_H}\beta + X_{E_i}\gamma + X_{HE_i}\nu) p_h p_{h'}$. The derivatives of v, u and w with respect to ξ are similar to the main-effect cases. Their derivatives involving the interaction parameter ν are listed below:

For v, $\frac{\partial}{\partial \nu} v = \sum_{H \in S(G_i)} \exp(\alpha^* + X_{C_H}\beta + X_{E_i}\gamma + X_{HE_i}\nu) X_{HE_i}^T p_h p_{h'}$

$\frac{\partial}{\partial \alpha^*}\frac{\partial}{\partial \nu} v = \sum_{H \in S(G_i)} \exp(\alpha^* + X_{C_H}\beta + X_{E_i}\gamma + X_{HE_i}\nu) X_{HE_i}^T p_h p_{h'}$

$\frac{\partial}{\partial \nu}\frac{\partial}{\partial \beta} v = \sum_{H \in S(G_i)} \exp(\alpha^* + X_{C_H}\beta + X_{E_i}\gamma + X_{HE_i}\nu) X_{C_H}^T X_{HE_i} p_h p_{h'}$

$\frac{\partial}{\partial \nu}\frac{\partial}{\partial \gamma} v = \sum_{H \in S(G_i)} \exp(\alpha^* + X_{C_H}\beta + X_{E_i}\gamma + X_{HE_i}\nu) X_{E_i}^T X_{HE_i} p_h p_{h'}$

$\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial \nu} v = \sum_{H \in S(G_i)} 2I(h = \tau \neq h') \Big\{ \exp(\alpha^* + X_{C_H}\beta + X_{E_i}\gamma + X_{HE_i}\nu) p_{h'} X_{HE_i} \Big\} + 2I(h = h' = \tau) \Big\{ \exp(\alpha^* + X_{C_H}\beta + X_{E_i}\gamma + X_{HE_i}\nu) p_\tau X_{HE_i} \Big\}$

$\frac{\partial}{\partial \nu}\frac{\partial}{\partial \nu} v = \sum_{H \in S(G_i)} \exp(\alpha^* + X_{C_H}\beta + X_{E_i}\gamma + X_{HE_i}\nu) X_{HE_i}^T X_{HE_i} p_h p_{h'}$

The derivatives of u with respect to ν are the same as those for v, except they are summed over all haplotype pairs H instead of $H \in S(G_i)$. The derivatives of w are the same as the main-effect results. Then we can obtain U_ν and V_ν in terms of derivatives of v, u, and w following the same spirit as described in the main-effect section.

Case-only tests

For case-only H×E tests, $\eta = \nu$ and $\xi = \tilde{p}$. We rewrite the case-only likelihood of a subject as $\log L_i = \text{c} + \log v - \log u$, where $v = \sum_{H \in S(G_i)} \exp(X_{HE}\nu) \tilde{p_h} \tilde{p_{h'}}$, and $u = \sum_H \exp(X_{HE}\nu) \tilde{p_h} \tilde{p_{h'}}$. Now we can write the components of U_ν and V_ν in terms of derivatives of v and u as follows:

$\frac{\partial}{\partial \nu} \log L_i = \frac{\frac{\partial}{\partial \nu} v}{v} - \frac{\frac{\partial}{\partial \nu} u}{u}$

$\frac{\partial}{\partial \tilde{p}_\tau} \log L_i = \frac{\frac{\partial}{\partial \tilde{p}_\tau} v}{v} - \frac{\frac{\partial}{\partial \tilde{p}_\tau} u}{u}$,

and

$\frac{\partial}{\partial \tilde{p}_\tau}\frac{\partial}{\partial \tilde{p}_\tau} \log L_i = \frac{v(\frac{\partial^2}{\partial \tilde{p}_\tau^2} v) - (\frac{\partial}{\partial \tilde{p}_\tau} v)(\frac{\partial}{\partial \tilde{p}_\tau} v)}{v^2} - \frac{u(\frac{\partial^2}{\partial \tilde{p}_\tau^2} u) - (\frac{\partial}{\partial \tilde{p}_\tau} u)(\frac{\partial}{\partial \tilde{p}_\tau} u)}{u^2}$

$\frac{\partial}{\partial \tilde{p}_\theta}\frac{\partial}{\partial \tilde{p}_\tau} \log L_i = \frac{v(\frac{\partial}{\partial \tilde{p}_\theta}\frac{\partial}{\partial \tilde{p}_\tau} v) - (\frac{\partial}{\partial \tilde{p}_\tau} v)(\frac{\partial}{\partial \tilde{p}_\theta} v)}{v^2} - \frac{u(\frac{\partial}{\partial \tilde{p}_\theta}\frac{\partial}{\partial \tilde{p}_\tau} u) - (\frac{\partial}{\partial \tilde{p}_\tau} u)(\frac{\partial}{\partial \tilde{p}_\theta} u)}{u^2}$

$\frac{\partial}{\partial \tilde{p}_\tau}\frac{\partial}{\partial \nu} \log L_i = \frac{v(\frac{\partial}{\partial \tilde{p}_\tau}\frac{\partial}{\partial \nu} v) - (\frac{\partial}{\partial \nu} v)(\frac{\partial}{\partial \tilde{p}_\tau} v)}{v^2} - \frac{u(\frac{\partial}{\partial \tilde{p}_\tau}\frac{\partial}{\partial \nu} u) - (\frac{\partial}{\partial \nu} u)(\frac{\partial}{\partial \tilde{p}_\tau} u)}{u^2}$

Where the derivatives of v with respect to the parameters ν and a specific $\tilde{p}_\tau$ are

$\frac{\partial}{\partial \nu} v = \sum_{H \in S(G_i)} \exp(X_{HE}\nu)\tilde{p_h}\tilde{p_{h'}} X_{HE}^T$

$\frac{\partial}{\partial \tilde{p_\tau}} v = \sum_{H \in S(G_i)} \exp(X_{HE}\nu) 2I(h = \tau)\tilde{p}_{h'}$

$\frac{\partial}{\partial \tilde{p_\theta}}\frac{\partial}{\partial \tilde{p_\tau}} v = \sum_{H \in S(G_i)} 2I(h = \tau, h' = \theta)\exp(X_{HE}\nu)$

$\frac{\partial}{\partial \tilde{p_\tau}}\frac{\partial}{\partial \nu} v = \sum_{H \in S(G_i)} \exp(X_{HE}\nu) 2I(h = \tau)\tilde{p_{h'}} X_{HE}^T$

The derivatives of u are the same as those for v, except they are summed over all haplotype pairs H instead of $H \in S(G_i)$.

Appendix C: Estimation of the score statistic

Haplotype main-effect tests

I. Estimation for global-level main-effect analysis

We evaluate the score statistic using estimates of the nuisance parameters under the null hypothesis that $\beta = 0$. Under the null hypothesis, the likelihood becomes

$L_{obs} = \prod_{i=1}^n \left[\frac{\exp(\alpha^* + X_{E_i}\gamma)\sum_{H \in S(G_i)} p_h p_{h'}}{1+\exp(\alpha^* + X_{E_i}\gamma)}\right]^{d_i} \left[\frac{\sum_{H \in S(G_i)} p_h p_{h'}}{1+\exp(\alpha^* + X_{E_i}\gamma)}\right]^{1-d_i}$. We see that the observed likelihood factors into terms only involving the haplotype frequencies $\mathbf{p}$ and terms only involving the regression parameters α^* and γ. For estimating $\mathbf{p}$, we consider the terms of

the observed likelihood that contain $\mathbf{p}$:

$$L_{obs} \propto \prod_{i=1}^{n} \sum_{H \in S(G_i)} p_h p_{h'}.$$

If we assume haplotype phase is known, we can write the full data likelihood involving $\mathbf{p}$ as

$$L_{full} \propto \prod_{(h,h')} (p_h p_{h'})^{(c_{hh'}+d_{hh'})},$$

where $c_{hh'}$ is the number of controls with haplotype pair (h, h') and $d_{hh'}$ is the number of cases with haplotype pair(h, h'). Therefore, the full data likelihood has a multinomial distribution and we can use the EM algorithm implemented in the `haplo.em` function in R to estimate $\mathbf{p}$.

For estimating α^* and γ, we write the terms of the observed likelihood that contain the regression parameters:

$$L_{obs} \propto \prod_{i=1}^{n} \frac{\exp(\alpha^* + \mathbf{X}_{E_i}\gamma)^{d_i}}{1 + \exp(\alpha^* + \mathbf{X}_{E_i}\gamma)}.$$

α^* and γ are the parameter estimates from regressing the response d_i on the environmental covariate E. Therefore, we can use the `glm` function in R to obtain estimates for these parameters.

II. Estimation for individual-level main-effect analysis

Because we must now estimate all but one of the haplotype cluster parameters under H_0, we use the expectation-conditional-maximization (ECM) algorithm [Meng, 1993] to iteratively estimate the regression parameters and haplotype frequencies. Under the null hypothesis

that $\beta = 0$, the full data likelihood assuming the missing data (haplotype phase) is known is

$$L_{full} = \prod_{k=1}^{e} \prod_{(h,h')} \left[\frac{p_h p_{h'}}{1 + \sum_H \exp(\alpha^* + X_{C_H}\beta + X_{E_k}\gamma) p_h p_{h'}} \right]^{c_{hh',k}} \left[\frac{\exp(\alpha^* + X_{C_H}\beta + X_{E_k}\gamma) p_h p_{h'}}{1 + \sum_H \exp(\alpha^* + X_{C_H}\beta + X_{E_k}\gamma) p_h p_h} \right]^{d_{hh',k}}.$$

The steps for estimating the nuisance parameters are

1. Obtain initial estimates of $\mathbf{p}$, α^*, γ, and β.
2. For step s, use E Step to estimate $c^{(s)}_{hh',k}$ and $d^{(s)}_{hh',k}$.
3. Use M step 1 to find $\mathbf{p}^{(s+1)}$ that maximizes L_{full}, using $\beta^{(s)}$, $\gamma^{(s)}$, $\alpha^{*(s)}$, $\mathbf{p}^{(s)}$, $c^{(s)}_{hh',k}$ and $d^{(s)}_{hh',k}$.
4. Use M step 2 to find $\alpha^{*(s+1)}$, $\gamma^{(s+1)}$ and $\beta^{(s+1)}$ that maximize L_{full}, using $\mathbf{p}^{(s+1)}$, $c^{(s)}_{hh',k}$, and $d^{(s)}_{hh',k}$.
5. Check differences between parameters at steps $(s+1)$ and s.
6. If difference for at least one of the parameters is greater than a specified limit, start over with step 2 to estimate $c^{(s+1)}_{hh',k}$, and $d^{(s+1)}_{hh',k}$ using $\mathbf{p}^{(s+1)}$, $\beta^{(s+1)}$, $\gamma^{(s+1)}$, and $\alpha^{*(s+1)}$.

E step

The E step estimates the number of controls with haplotype pair (h, h') and covariate level k as

$$\begin{aligned} E(c_{hh',k}) &= \sum_g c_{g,k} I(H \in S(g)) P(H|g) \\ &= \sum_g c_{g,k} I(H \in S(g)) \frac{p_h p_{h'}}{\sum_{H \in S(g)} p_h p_h} \end{aligned}$$

and the number of cases with haplotype pair (h, h') and covariate level k as

$$\begin{aligned} E(d_{hh',k}) &= \sum_g d_{g,k} I(H \in S(g)) P(H|g) \\ &= \sum_g d_{g,k} I(H \in S(g)) \frac{\exp(\alpha^* + X_{C_H}\beta + X_{E_k}\gamma) p_h p_{h'}}{\sum_{H \in S(g)} \exp(\alpha^* + X_{C_H}\beta + X_{E_k}\gamma) p_h p_h}, \end{aligned}$$

where $c_{g,k}$ and $d_{g,k}$ are the number of controls and cases, respectively, with genotype g and covariate level k.

M step 1: Estimating p

To estimate $\mathbf{p}$, we maximize the log of the full data likelihood subject to the constraint $\sum_h p_h = 1$. We introduce a Lagrange multiplier λ and call the new likelihood L_M:

$$\begin{aligned} L_M &= \log L_{full} + \lambda(\sum_h p_h - 1) \\ &= \sum_{k=1}^{e} \sum_{(h,h')} \Big[c_{hh',k} \log(p_h p_{h'}) + d_{hh',k}[(\alpha^* + X_{C_{hh'}}\beta + X_{E_k}\gamma) + \log(p_h p_{h'})] - \\ &\quad \log(1 + \sum_H \exp(\alpha^* + X_{C_H}\beta + X_{E_k}\gamma) p_h p_h)(c_{hh',k} + d_{hh',k}) \Big] + \lambda(\sum_h p_h - 1) \\ &= \sum_{k=1}^{e} \sum_{(h,h')} \Big[c_{hh',k} \log(p_h p_{h'}) + d_{hh',k}[(\alpha^* + X_{C_{hh'}}\beta + X_{E_k}\gamma) + \log(p_h p_{h'})] - \\ &\quad \log(1 + \sum_H \exp(\alpha^* + X_{C_H}\beta + X_{E_k}\gamma) p_h p_h)(c_{hh',k} + d_{hh',k}) \Big] + \lambda(\sum_h p_h - 1) \\ &= \sum_{k=1}^{e} \sum_{(h,h')} \Big[\log(p_h p_{h'})(c_{hh',k} + d_{hh',k}) + d_{hh',k}(\alpha^* + X_{C_{hh'}}\beta + X_{E_k}\gamma) - \\ &\quad n_{hh',k} \log(1 + \sum_H \exp(\alpha^* + X_{C_H}\beta + X_{E_k}\gamma) p_h p_h) \Big] + \lambda(\sum_h p_h - 1). \end{aligned}$$

The portion of L_M that depends on $\mathbf{p}$ is

$$\begin{aligned}
L_M \propto & \sum_{k=1}^{e} \sum_{(h,h')} \Big[\log(p_h p_{h'}) n_{hh',k} - \\
& n_{hh',k} \log(1 + \sum_{H} \exp(\alpha^* + X_H\beta + X_{E_k}\gamma) p_h p_{h'}) \Big] + \lambda(\sum_h p_h - 1) \\
\propto & \sum_{k=1}^{e} \big(\sum_h m_{h,k} \log(p_h) \big) - \\
& \sum_{k=1}^{e} \sum_{(h,h')} \big(n_{hh',k} \log(1 + \sum_{H} \exp(\alpha^* + X_H\beta + X_{E_k}\gamma) p_h p_{h'}) \big) + \lambda(\sum_h p_h - 1),
\end{aligned}$$

where $m_{h,k}$ is the number of haplotypes of type h with covariate level k. The derivative of this expression with respect to a specific p_τ is

$$\begin{aligned}
\frac{\partial}{\partial p_\tau} L_M \propto & \sum_{k=1}^{e} \frac{m_{\tau,k}}{p_\tau} - \\
& \sum_{k=1}^{e} \sum_{(h,h')} \left\{ n_{hh',k} \frac{\sum_H \exp(\alpha^* + X_{C_H}\beta + X_{E_k}\gamma) I(h=\tau) 2 p_{h'}}{1 + \sum_H \exp(\alpha^* + X_{C_H}\beta + X_{E_k}\gamma) p_h p_{h'}} \right\} + \lambda \\
\propto & \sum_{k=1}^{e} \frac{m_{\tau,k}}{p_\tau} - \\
& \sum_{k=1}^{e} \left\{ n_k \frac{\sum_{h'} \exp(\alpha^* + X_{C_{\tau,h'}}\beta + X_{E_k}\gamma) 2 p_{h'}}{1 + \sum_H \exp(\alpha^* + X_{C_H}\beta + X_{E_k}\gamma) p_h p_{h'}} \right\} + \lambda
\end{aligned}$$

This expression is still difficult to solve for an analytical expression for p_τ, therefore we define the quantity $u(\mathbf{p})_\tau$ as

$$u(\mathbf{p})_\tau = \sum_{k=1}^{e} n_k \frac{\sum_{h'} \exp(\alpha^* + X_{C_{\tau,h'}}\beta + X_{E_k}\gamma) 2 p_{h'}}{1 + \sum_H \exp(\alpha^* + X_{C_H}\beta + X_{E_k}\gamma) p_h p_{h'}}$$

Then we can write the derivative with respect to p_τ as

$$\frac{\partial}{\partial p_\tau} L_m \propto \sum_{k=1}^{e} \frac{m_{\tau,k}}{p_\tau} - u(\mathbf{p})_\tau + \lambda$$

By setting the above equal to 0 and solving for p_τ we obtain the updating equation for p_τ

$$p_\tau = \frac{\sum_{k=1}^{e} m_{\tau,k}}{(u(\mathbf{p})_\tau - \lambda)}. \tag{5}$$

We carry out another iteration within the M step for estimating $\mathbf{p}$ by estimating $u(\mathbf{p})^{(s,k)}$ based on $\mathbf{p}^{(s)}$, then estimating $\mathbf{p}^{(s,k+1)}$ based on $u(\mathbf{p})^{(s,k)}$. This iteration continues until the difference between $\mathbf{p}^{(s,k)}$ and $\mathbf{p}^{(s,k+1)}$ is less than a specified limit. Then $\mathbf{p}^{(s,k+1)}$ becomes $\mathbf{p}^{(s+1)}$.

M step 2: Estimating α^*, β, and γ

The log of the full likelihood that depends on the regression parameters is

$$\begin{aligned}
\log L_{full} &\propto \sum_{k=1}^{e} \sum_{(h,h')} -n_{hh',k} \log(1 + \sum_{H} \exp(\alpha^* + X_{C_H}\beta + X_{E_k}\gamma) p_h p_{h'}) \\
&\quad + d_{hh',k}\big[\log(\exp(\alpha^* + X_{C_H}\beta + X_{E_k}\gamma)) + \log p_h p_{h'}\big] \\
&\propto \sum_{k=1}^{e} \sum_{(h,h')} -n_{hh',k} \log(1 + \sum_{H} \exp(\alpha^* + X_{C_H}\beta + X_{E_k}\gamma) p_h p_{h'}) \\
&\quad + d_{hh',k}(\alpha^* + X_{C_H}\beta + X_{E_k}\gamma)
\end{aligned}$$

We obtain maximum likelihood estimates for α^*, β, and γ using the optimization function `nlminb` in R.

H×E tests

Estimation for global-level and individual-level H×E analysis

We use the expectation-conditional maximization (ECM) algorithm to estimate all of the nuisance parameters ξ under the null hypothesis. The estimation scheme is similar to that above for the score test for a specific haplotype main effect. The difference for the global interaction test is that we must estimate the haplotype main effect parameters β for all clusters. For the specific H×E effect, we must now also estimate the interaction effects ν, excluding the parameter of interest ν_t. The E step is similar, with the appropriate terms involving ν incorporated into the derivatives of the likelihood. In the first M step for estimating $\mathbf{p}$, the quantity $u(\mathbf{p})_\tau$ is now

$$u(\mathbf{p})_\tau = \sum_{k=1}^{e} n_k \frac{\sum_{h'} \exp(\alpha^* + X_{\tau,h'}\beta + X_{E_k}\gamma + X_{(\tau,h')E_k}\nu) 2p_{h'}}{1 + \sum_H \exp(\alpha^* + X_H\beta + X_{E_k}\gamma + X_{HE_k}\nu) p_{h_1} p_{h_2}}$$

In the second M step, we obtain maximum likelihood estimates for α^*, β, γ, and ν using the optimization function `nlminb` in `R`.

Case-only tests

I. Estimation for global-level analysis

We use the EM algorithm to maximize the case-only likelihood with respect to $\tilde{\mathbf{p}}$, while using a Lagrange multiplier to constrain $\sum_h \tilde{p}_h = 1$. We first assume that haplotype phase

is known and construct the full data likelihood:

$$L_{full} = \prod_{k=1}^{e} \prod_{(h,h')} \left[\frac{\exp(X_{HE_k}\nu)\tilde{p}_h\tilde{p}_{h'}}{\sum_{H'} \exp(X_{H'E_k}\nu)\tilde{p}_h\tilde{p}_{h'}} \right]^{d_{hh',k}},$$

where $d_{hh',k}$ is the number of cases with haplotype pair (h, h') and covariate level k. Under the null hypothesis that $\nu = 0$, the full data likelihood becomes

$$L_{full} = \prod_{k=1}^{e} \prod_{(h,h')} \left[\frac{\tilde{p}_h\tilde{p}_{h'}}{\sum_{H'} \tilde{p}_h\tilde{p}_{h'}} \right]^{d_{hh',k}}.$$

Taking the log of the full likelihood and incorporating the Lagrange multiplier, we maximize the new likelihood, L_M:

$$\begin{aligned}
L_M &= \log L_{full} + \lambda(\sum_h \tilde{p}_h - 1) \\
&= \sum_{k=1}^{e} \sum_{(h,h')} \left[d_{hh',k} \log(\tilde{p}_h\tilde{p}_{h'}) - d_{hh',k} \log(\sum_{H'} \tilde{p}_h\tilde{p}_{h'}) \right] + \lambda(\sum_h \tilde{p}_h - 1) \\
&= \sum_{k=1}^{e} \sum_{(h,h')} d_{hh',k} \log(\tilde{p}_h\tilde{p}_{h'}) + \lambda(\sum_h \tilde{p}_h - 1) \\
&= \sum_{k=1}^{e} (\sum_h d_{h,k} \log(\tilde{p}_h)) + \lambda(\sum_h \tilde{p}_h - 1),
\end{aligned}$$

where $d_{h,k}$ is the number of cases with haplotype h and covariate level k. We estimate $d_{hh',k}$ in the E step:

$$\begin{aligned}
E(d_{hh',k}) &= \sum_g d_{g,k} I(H \in S(g)) P(H|g) \\
&= \sum_g d_{g,k} I(H \in S(g)) \frac{\tilde{p}_h\tilde{p}_{h'}}{\sum_{H \in S(g)} \tilde{p}_h\tilde{p}_{h'}}
\end{aligned}$$

where $d_{g,k}$ is the number of cases with genotype g and covariate level k. The derivative of L_M with respect to a specific $\tilde{p}_\tau$ is

$$\frac{\partial}{\partial \tilde{p_\tau}} L_M \propto \sum_{k=1}^{e} \frac{d_{\tau,k}}{\tilde{p}_\tau} + \lambda.$$

By setting the above equal to 0 and solving for $\tilde{p}_\tau$, we obtain the updating equation for $\tilde{p}_\tau$

$$\tilde{p}_\tau = \frac{\sum_{k=1}^{e} d_{\tau,k}}{\lambda}.$$

II. Estimation for individual-level analysis

We use the ECM algorithm to estimate the nuisance parameters ξ under the null hypothesis. Assuming that haplotype phase is known, the full data likelihood is

$$L_{full} = \prod_{k=1}^{e} \prod_{(h,h')} \left[\frac{\exp(X_{HE_k}\nu)\tilde{p}_h\tilde{p}_{h'}}{\sum_{H'} \exp(X_{H'E_k}\nu)\tilde{p}_h\tilde{p}_{h'}} \right]^{d_{hh',k}}.$$

The steps for estimating the nuisance parameters are

1. Obtain initial estimates of $\tilde{\mathbf{p}}$ and ν.
2. For step s, use E Step to estimate $d^{(s)}_{hh',k}$.
3. Use M step 1 to find $\tilde{\mathbf{p}}^{(s+1)}$ that maximizes L_{full}, using $\nu^{(s)}$ and $d^{(s)}_{hh',k}$.
4. Use M step 2 to find $\nu^{(s)}$ that maximizes L_{full}, using $\tilde{\mathbf{p}}^{(s+1)}$ and $d^{(s)}_{hh',k}$.
5. Check differences between parameters at steps $(s+1)$ and s.
6. If difference for at least one of the parameters is greater than a specified limit, start over with step 2 to estimate $d^{(s+1)}_{hh',k}$ using $\tilde{\mathbf{p}}^{(s+1)}$ and $\nu^{(s+1)}$.

E step

The E step estimates the number of cases with haplotype pair (h, h') and covariate level k as

$$\begin{aligned} E(d_{hh',k}) &= \sum_g d_{g,k} I(H \in S(g)) P(H|g) \\ &= \sum_g d_{g,k} I(H \in S(g)) \frac{\exp(X_{HE_k}\nu)\tilde{p}_h\tilde{p}_{h'}}{\sum_{H \in S(g)} \exp(X_{HE_k}\nu)\tilde{p}_h\tilde{p}_{h'}} \end{aligned}$$

where $d_{g,k}$ is the number of cases with genotype g and covariate level k.

M step 1

Similar to the estimation of $\tilde{\mathbf{p}}$ for the global interaction test, we maximize the full data likelihood subject to the constraint that $\sum_h \tilde{p}_h = 1$. Incorporating the Lagrange multiplier, the new likelihood L_M is

$$\begin{aligned} L_M &= \log L_{full} + \lambda(\sum_h \tilde{p}_h - 1) \\ &= \sum_{k=1}^{e} \sum_{(h,h')} \Big[d_{hh',k}[(X_{HE_k}\nu) + \log(\tilde{p}_h\tilde{p}_{h'})] - \\ &\quad d_{hh',k} \log(\sum_H \exp(X_{HE_k}\nu)\tilde{p}_h\tilde{p}_{h'}) \Big] + \lambda(\sum_h \tilde{p}_h - 1) \end{aligned}$$

The portion of L_M that depends on $\tilde{\mathbf{p}}$ is

$$\begin{aligned} L_M \propto & \sum_{k=1}^{e} \sum_{(h,h')} \Big[d_{hh',k} \log(\tilde{p}_h \tilde{p}_{h'}) - \\ & d_{hh',k} \log(\sum_H \exp(X_{HE_k}\nu)\tilde{p}_h\tilde{p}_{h'}) \Big] + \lambda(\sum_h \tilde{p}_h - 1) \\ \propto & \sum_{k=1}^{e}(\sum_h d_{h,k}\log(\tilde{p})) + \sum_{k=1}^{e}\sum_{(h,h')} d_{hh',k}\log(\sum_H \exp(X_{HE_k}\nu)\tilde{p}_h\tilde{p}_{h'}) + \lambda(\sum_h \tilde{p}_h - 1) \end{aligned}$$

The derivative of this expression with respect to a specific p_τ is

$$\begin{aligned} \frac{\partial}{\partial p_\tau} L_M & \propto \sum_{k=1}^{e} \frac{d_{\tau,k}}{\tilde{p}_\tau} - \sum_{k=1}^{e}\sum_{(h,h')} d_{hh',k} \frac{\sum_H \exp(X_{HE_k}\nu) I(h=\tau) 2\tilde{p}_{h'}}{\sum_H \exp(X_{HE_k}\nu)\tilde{p}_h\tilde{p}_{h'}} + \lambda \\ & \propto \sum_{k=1}^{e} \frac{d_{\tau,k}}{\tilde{p}_\tau} - \sum_{k=1}^{e} d_k \frac{\sum_{h'} \exp(X_{HE_k}\nu) 2\tilde{p}_{h'}}{\sum_H \exp(X_{HE_k}\nu)\tilde{p}_h\tilde{p}_{h'}} + \lambda \end{aligned}$$

As in Chapters 2 and 3, we define a quantity $u(\tilde{\mathbf{p}})_\tau$ as

$$u(\tilde{\mathbf{p}})_\tau = \sum_{k=1}^{e} d_k \frac{\sum_{h'} \exp(X_{HE_k}\nu) 2\tilde{p}_{h'}}{\sum_H \exp(X_{HE_k}\nu)\tilde{p}_h\tilde{p}_{h'}}.$$

Then we write the derivative with respect to p_τ as

$$\frac{\partial}{\partial p_\tau} L_M \propto \sum_{k=1}^{e} \frac{d_{\tau,k}}{\tilde{p}_\tau} - u(\tilde{\mathbf{p}})_\tau + \lambda,$$

and by setting the above equal to 0 and solving for p_τ we obtain the updating equation for p_τ:

$$p_\tau = \frac{\sum_{k=1}^{e} d_{\tau,k}}{u(\tilde{\mathbf{p}})_\tau - \lambda}.$$

We carry out another iteration within the M step for estimating $\tilde{\mathbf{p}}$ by estimating $u(\tilde{\mathbf{p}})^{(s,k)}$ based on $\tilde{\mathbf{p}}^{(s)}$, then estimating $\tilde{\mathbf{p}}^{(s,k+1)}$ based on $u(\tilde{\mathbf{p}})^{(s,k)}$. This iteration continues until the difference between $\tilde{\mathbf{p}}^{(s,k)}$ and $\tilde{\mathbf{p}}^{(s,k+1)}$ is less than a specified limit. Then $\tilde{\mathbf{p}}^{(s,k+1)}$ becomes

$\tilde{\mathbf{p}}^{(s+1)}$.

M step 2

The log of the full data likelihood that depends on the interaction parameters ν is

$$\log(L_{full}) \propto \sum_{k=1}^{e} \sum_{(h,h')} \Big[d_{hh',k} \log(\exp(X_{HE_k}\nu)\tilde{p}_h\tilde{p}_{h'}) - d_{hh',k} \log(\sum_{H} \exp(X_{HE_k}\nu)\tilde{p}_h\tilde{p}_{h'}).$$

We use the optimization function `nlminb` in `R` to obtain maximum likelihood estimates of ν.

References

Allen A, Satten G. 2008. Robust estimation and testing of haplotype effects in case-control studies. Genetic Epidemiology 32:29–40.

Boos D. 1992., On generalized score tests. The American Statistician 46:327–333.

Browning, B. Browning, S. 2007. , Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. Genetic Epidemiology 31:365–375.

Browning, S. 2006. , Multilocus association mapping using variable-length markov chains. American Journal of Human Genetics 78, 903–913.

Chatterjee, N. Carroll, R. 2005. , Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika 92, 399–418.

Chen, Y.-H., Chatterjee, N. Carroll, R. 2008. , Retrospective analysis of haplotype-based case-control studies under a flexible model for gene-environment association. Biostatistics 9:81–99.

Chen, Y.-H. Kao, J.-T. 2006. , Multinomial logistic regression approach to haplotype association analysis in population-based case-control studies. BMC Genetics 7:43.

Clark, A. 2004. , The role of haplotypes in candidate gene studies. Genetic Epidemiology 27:321–333.

de Bakker, P., R., Y., Pe'er, I., Gabriel, S., Daly, M. Altshuler, D. 2005. , Efficiency and power in genetic association studies. Nature Genetics 37:1217–1223.

Durrant, C., Zondervan, K., Cardon, L., Hunt, S., Deloukas, P. Morris, A. 2004. , Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. American Journal of Human Genetics 75:35–43.

Epstein, M. Satten, G. 2003. , Inference on haplotype effects in case-control studies using unphased genotype data. American Journal of Human Genetics 73:1316–1329.

Kao, J.-T., Wen, H.-C., Chien, K.-L., Hsu, H.-C. Lin, S.-W. 2003. , A novel genetic variant in the apolipoprotein a5 gene is associated with hypertriglyceridemia. Human Molecular Genetics 12:2533–2539.

Kwee, L., Epstein, M., Manatunga, A., Duncan, R., Allen, A. Satten, G. 2007., Simple methods for assessing haplotype-environment interactions in case-only and case-control studies. Genetic Epidemiology 31, 75–90.

Hugot, J.P., Chamaillard, M., Zouali, H., Lesage, S., Cezard, J.P., Belaiche, J., Almer, S., Tysk, C., O'Morain, C.A., Gassull, M., Binder, V., Finkel, Y., Cortot, A., Modigliani, R., Laurent-Puig, P., Gower-Rousseau, C., Macry, J., Colombel, J.F., Sahbatou, M., Thomas, G. 2001., Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature 411, 599–603.

Li, J. Jiang, T. 2005. , Haplotype-based linkage disequilibrium mapping via direct data mining. Bioinformatics 21:4384–4394.

Li, J., Zhou, Y. Elston, R. 2006., Haplotype-based quantitative trait mapping using a clustering method. BMC Bioinformatics 7:258.

Lin, D. Zeng, D. 2006. , Likelihood-based inference on haplotype effects in genetic association studies. Journal of the American Statistical Association 101:89–104.

Lin, S., Chakravarti, A. Cutler, D. 2004. , Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. Nature Genetics 36, 1181–1188.

Liu, J., Papasian, C. Deng, H. 2007. , Incorporating single-locus tests into haplotype cladistic analysis in case-control studies. PLoS Genetics 3 :e46.

Meng, X.L. andRubin, D.B. (1993), 'Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika, 80, 267–278

Molitor, J., Marjoram, P. Thomas, D. 2003., Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. American Journal of Human Genetics 73,1368–1384.

Pennacchio, L.A., Olivier, M., Hubacek, J.A., Cohen, J.C., Cox, D.R., Fruchart, J.C., Krauss, R.M., Rubin, E.M. (2001) An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. Science 294, 169–173.

Prentice, R. and Pyke, R. 1979. , Logistic disease incidence models and case-control studies. Biometrika 66:403–411.

Satten, G. A. Epstein, M. P. 2004. , Comparison of prospective and retrospective methods for haplotype inference in case-control studies. Genetic Epidemiology 27:192–201.

Schaid, D. 2004. , Evaluating associations of haplotypes with traits. Genetic Epidemiology 27:348–364.

Seltman, H., Roeder, K. Devlin, B. 2003. , Evolutionary-based association analysis using haplotype data. Genetic Epidemiology 25, 48–58.

Spinka, C., Carroll, R. Chatterjee, N. 2005. , Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. Genetic Epidemiology 29:108–127.

Su, S., Balding, D. Coin, L. 2008. , Disease association tests by inferring ancestral haplotypes using a hidden markov model. Bioinformatics 24:972–978.

Tachmazidou, I., Verzilli, C. De Iorio, M. 2007. , Genetic association mapping via evolution-based clustering of haplotypes. PLoS Genetics 3:e111.

Templeton, A. 1995. , A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping or dna sequencing. v. analysis of case/control sampling designs: Alzheimer's disease and the apoprotein e locus. Genetics 140:403–409.

Templeton, A., Boerwinkle, E. Sing, C. 1987. , A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. i. basic theory and an analysis of alcohol dehydrogenase activity in drosophila. Genetics 117, 343–351.

Tzeng, J. 2005. , Evolutionary-based grouping of haplotypes in association analysis. Genetic Epidemiology 28:220–231.

Tzeng, J.-Y., Wang, C.-H., Kao, J.-T. Hsiao, C. 2006. , Regression-based assocation analysis

with clustered haplotypes through use of genotypes. American Journal of Human Genetics 78:231–242.

Waldron, E. Whittaker, J.and Balding, D. 2006. , Fine mapping of disease genes via haplotype clustering. Genetic Epidemiology 30:170–179.

Yu, K., Gu, C., Province, M., Xiong, C. Rao, D. 2004. , Genetic association mapping under founder heterogeneity via weighted haplotype similarity analysis in candidate genes. Genetic Epidemiology 27:182–191.

Zaitlen, N., Kang, H., Eskin, E. Halperin, E. 2007. , Leveraging the hapmap correlation structure in association studies. American Journal of Human Genetics 80:683–691.

Zhao, J., Curtis, D. Sham, P. 2000. , Model-free analysis and permutation tests for allelic associations. Human Heredity 50:133–139.

		Main Effect		H×Age Interaction	
		Score Statistic	P-value	Score Statistic	P-value
Global Test		133.71	$< 1 \times 10^{-6}$	6.62	0.080
Specific Test					
Reference	Haplotype				
GGTCT	GAGTT	20.50	5.97×10^{-6}	1.59	0.208
	GGGCT	59.79	$< 1 \times 10^{-6}$	3.42	0.065
	AGGCC	8.94	2.79×10^{-3}	3.55	0.060
GGGCT	GAGTT	12.65	3.75×10^{-4}	0.21	0.643
	AGGCC	58.88	$< 1 \times 10^{-6}$	0.03	0.867
AGGCC	GAGTT	5.92	1.50×10^{-2}	0.11	0.739

Table 1: Test statistics and p-values for the hypertriglyceridemia study.

	Global Main Effect		Global H×E Interaction	
Hap Diversity	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
High:				
$q^{\dagger} = 0.1$	0.048	0.011	0.054 (0.054)	0.015 (0.009)
$q = 0.3$	0.043	0.007	0.048 (0.047)	0.010 (0.010)
Moderate:				
$q = 0.1$	0.047	0.009	0.056 (0.054)	0.010 (0.019)
$q = 0.3$	0.040	0.007	0.067 (0.046)	0.014 (0.010)
Low:				
$q = 0.1$	0.054	0.011	0.057 (0.053)	0.009 (0.010)
$q = 0.3$	0.046	0.012	0.047 (0.050)	0.014 (0.013)

Table 2: Type I error of the global test using the retrospective clustering method. Quantity q denote the allele frequency of the causal allele. Results in the parenthesis are based on case-only analysis.

	$\alpha = 0.05$				$\alpha = 0.01$			
Hap Diversity	retro-RD	retro-FD	prosp-RD	prosp-FD	retro-RD	retro-FD	prosp-RD	prosp-FD
High:								
$q^{\dagger} = 0.1$	0.849	0.775	0.843	0.772	0.642	0.541	0.640	0.549
$q = 0.3$	0.646	0.639	0.644	0.639	0.401	0.377	0.403	0.384
Moderate:								
$q = 0.1$	0.562	0.479	0.560	0.504	0.319	0.257	0.314	0.269
$q = 0.3$	0.911	0.844	0.907	0.868	0.769	0.643	0.763	0.685
Low:								
$q = 0.1$	0.839	0.827	0.835	0.826	0.660	0.645	0.663	0.644
$q = 0.3$	0.796	0.779	0.798	0.778	0.598	0.573	0.597	0.565

Table 3: Power of the global test for main effects. Quantity q denote the allele frequency of the causal allele.

	$\alpha = 0.05$		$\alpha = 0.01$	
Hap Diversity	retro-RD	retro-FD	retro-RD	retro-FD
High:				
$q^{\dagger} = 0.1$	0.717 (0.721)	NA	0.469 (0.481)	NA
$q = 0.3$	0.310 (0.328)	NA	0.112 (0.116)	NA
Moderate:				
$q = 0.1$	0.452 (0.452)	NA	0.244 (0.244)	NA
$q = 0.3$	0.436 (0.494)	NA	0.204 (0.266)	NA
Low:				
$q = 0.1$	0.790 (0.802)	0.752	0.586 (0.574)	0.560
$q = 0.3$	0.422 (0.430)	0.388	0.196 (0.204)	0.178

Table 4: Power of the global test for H×E interaction. Quantity q is allele frequency of the causal allele. Results in the parenthesis are based on case-only analysis.

	$\alpha = 0.05$			$\alpha = 0.01$		
True Effect	Global	Hap1	Hap2	Global	Hap1	Hap2
$(\beta_1, \beta_2) = (0, 0)$	0.039	0.025	0.029	0.007	0.005	0.003
$(\beta_1, \beta_2) = (0.5, 0.5)$	0.958	0.892	0.794	0.854	0.716	0.556
$(\beta_1, \beta_2) = (0.7, 0.5)$	0.998	1.000	0.798	0.990	0.980	0.570
$(\beta_1, \beta_2) = (0.7, 0.3)$	0.994	0.998	0.312	0.976	0.988	0.120

Table 5: Type I error and power of the haplotype-specific test for main-effect analysis.

	$\alpha = 0.05$			$\alpha = 0.01$		
True Effect	Global	Hap1	Hap2	Global	Hap1	Hap2
$(\nu_1, \nu_2) = (0, 0)$	0.051 (0.054)	0.022 (0.024)	0.027 (0.024)	0.007 (0.008)	0.004 (0.004)	0.004 (0.004)
$(\nu_1, \nu_2) = (0.5, 0.5)$	0.550 (0.566)	0.456 (0.460)	0.366 (0.370)	0.324 (0.328)	0.206 (0.210)	0.162 (0.162)
$(\nu_1, \nu_2) = (0.7, 0.5)$	0.782 (0.786)	0.738 (0.746)	0.340 (0.358)	0.536 (0.544)	0.470 (0.482)	0.130 (0.144)

Table 6: Type I error and power of the haplotype-specific test for H×E analysis. Results in the parenthesis are based on case-only analysis.