Published in final edited form as:

Genet Epidemiol. 2007 January ; 31(1): 75–90. doi:10.1002/gepi.20192.

Simple Methods for Assessing Haplotype-Environment Interactions in Case-Only and Case-Control Studies

L.C. Kwee^{1,†}, M.P. Epstein^{2,†,*}, A.K. Manatunga¹, R. Duncan², A.S. Allen³, and G.A. Satten⁴

¹Department of Biostatistics, Emory University, Atlanta, Georgia

²Department of Human Genetics, Emory University, Atlanta, Georgia

³Department of Biostatistics and Bioinformatics and Duke Clinical Research Institute, Duke University, Durham, North Carolina

⁴Centers for Disease Control and Prevention, Atlanta, Georgia

Abstract

For investigating haplotype-environment interactions in case-control studies, one can implement statistical methods based either on a retrospective likelihood (modeling the probability of haplotype and environment conditional on disease status) or a prospective likelihood (modeling the probability of disease status conditional on haplotype and environment). Retrospective approaches are generally more powerful than prospective approaches, but require an explicit model of the joint distribution of haplotype and environmental factors in the sample with the latter being particularly unattractive to specify. To resolve this issue, we propose a number of simple retrospective procedures for haplotype-environment interaction analysis that do not require explicit modeling of environmental covariates in the sample. We first consider a cases-only procedure, followed by a simple likelihood for case-control data that is proportional to the full-retrospective likelihood. Finally, we consider a retrospective procedure for inference on haplotype-environment interaction effects in matched or finely-stratified case-control studies. Our methods are based on the assumptions that haplotypes and environmental covariates are independent in the target population and that disease is rare. We illustrate our approaches using case-control data from the Finland-United States Investigation of Non-Insulin Dependent Diabetes Mellitus (FUSION) genetic study and simulated data.

INTRODUCTION

The case-control study design is a popular and economical design for conducting genetic association studies of complex disease. The study design entails collecting genetic data, usually consisting of dense sets of biallelic single-nucleotide polymorphisms (SNPs), from a sample of persons with a disease or condition of interest (case participants) and a sample of persons who do not have the disease or condition of interest (control participants). Using these two samples, one can conduct χ^2 tests of association between a single SNP and disease status [Olson and Wijsman, 1994] that compare SNP allele frequencies in case and control participants. Rather than analyzing single SNPs, one can also base inference on SNP-based

*Correspondence to: Michael P. Epstein, Department of Human Genetics, Emory University School of Medicine 615 Michael Street, Suite 301, Atlanta, GA 30322. E-mail: mepstein@genetics.emory.edu.

†Contributed equally to this work

ELECTRONIC-DATABASE INFORMATION

The URL for methods presented within is as follows: CHAPLIN software, <http://www.genetics.emory.edu/labs/epstein/software>

haplotypes, which are a series of tightly linked SNP variants that are inherited as a unit from a parent. While some studies have debated the merit of haplotype analysis over single-SNP analysis [Chapman et al., 2003; Tzeng and Roeder, 2006], simulation studies [Akey et al., 2001; Morris and Kaplan, 2002; Rosenberg et al., 2006] have demonstrated haplotype procedures are more powerful than single-SNP procedures when either the causal polymorphism is not genotyped (which is highly likely) or when multiple causal polymorphisms act in cis fashion in the haplotype region. Studies provide empirical evidence of this latter event occurring in many genetic diseases, such as prostate cancer [Tavtigian et al., 2001].

Given complex diseases originate from the interplay among both genetic and environmental factors, interest will increasingly focus on identifying interactions between genetic and environmental factors that increase risk of disease. When interest focuses on identifying interactions between a given SNP and environment, one can apply a variety of existing approaches for inference including case-only procedures, logistic regression (for unmatched data), and conditional logistic regression (for matched data). However, when interest centers on identifying haplotype-environment interactions, the number of available procedures for analysis is far more limited. Lake et al. [2003] developed a logistic-regression approach based on a prospective likelihood that modeled the probability of disease conditional on haplotype and environment. While the approach is easy to implement, its power may suffer with respect to an approach based on a retrospective likelihood that models the probability of haplotype and environment conditional on disease [Satten and Epstein, 2004]. This claim is in contrast to the well-known result of Prentice and Pyke [1979], who showed that the analysis of retrospective case-control data using a prospective likelihood should yield no loss of efficiency relative to the (more appropriate) retrospective likelihood. However, their result assumed that the distribution of the exposure (here, the haplotype pair) was saturated. This assumption does not hold in haplotype analysis, since one must impose some assumption on the distribution of haplotype pairs to resolve haplotype ambiguity in the genotype data.

Under a non-saturated exposure distribution, Carroll et al. [1995] showed that retrospective analysis of case-control data can be more efficient than prospective analysis. This finding suggests that a retrospective approach for detecting haplotype-environment interactions should have greater or equal power to the prospective approach of Lake et al. [2003]. However, while retrospective approaches are more powerful, they are also more inconvenient since they require specification of the joint distribution of haplotype and environmental factors in the sample with the latter being particularly tedious to model. This inconvenience is the main reason investigators do not use retrospective approaches in case-control studies.

In this paper, we propose a variety of retrospective approaches for conducting inference on haplotype-environment interaction effects. We base all our retrospective approaches on two key assumptions: (a) rare disease and (b) haplotype-environment independence in the target population. By implementing these assumptions, we show that we can conduct retrospective likelihood-based inference on haplotype and environmental effects without specifying the distribution of environmental covariates in the sample. Our approaches differ from those of Lin et al. [2005] and Spinka et al. [2005], both of whom proposed profile-likelihood approaches that circumvented the modeling of environmental factors in a retrospective likelihood using the assumption of haplotype-environment independence. While these approaches are elegant and should yield improved efficiency with respect to prospective approaches, they have the unusual feature of estimating the (identifiable) absolute risk of disease from case-control data. This parameter will become more difficult and computationally challenging to estimate as the absolute risk of disease decreases. Although Spinka et al. give a rare-disease approximation to their profile likelihood, we show here that making the rare-disease assumption at the outset

leads to a factorization of the likelihood that yields a number of novel methods for the analysis of haplotype data in retrospective studies.

In the subsequent section, we first consider the full retrospective likelihood of genetic and environmental data conditional on disease data. We factor this likelihood into different components and explore the information that each component provides to the haplotype and environmental parameters of interest. We then use this factorization to motivate cases-only analysis of haplotype-environment interactions and show that, under certain circumstances, this analysis procedure is equivalent to an analysis using both cases and controls. Next, we use our likelihood factorization and the famous result of Prentice and Pyke [1979] to show that the full retrospective likelihood is proportional to the retrospective likelihood of haplotypes conditional on environmental factors and disease multiplied by the prospective likelihood of disease given environmental factors. As a result, we can conduct a retrospective likelihood analysis without having to specify the distribution of environmental factors. We note that the function that we maximize is equivalent to the rare-disease approximation of the profile likelihood proposed by Spinka et al. [2005].

Finally, we consider a retrospective procedure for haplotype analysis of matched or finely-stratified case-control data. Our procedure differs from that of Kraft et al. [2005], who developed a series of haplotype methods for matched data but ignored both the case-control sampling design and the uncertainty in haplotype-frequency estimation. We accommodate both of these factors within our likelihood. We propose a novel conditioning scheme that conditions only on case or control status and the set of environmental covariates within each stratum. Within each stratum, the resulting likelihood is the product of the retrospective likelihood of genetic effects conditional on disease status and environmental covariates for each stratum member multiplied by the likelihood of the environmental covariates conditional on disease status and the stratum set of environmental covariates. We will show it is unnecessary to condition further on the stratum set of genotype information since our retrospective likelihood of genotype data conditional on environmental and disease data already eliminates stratum-specific intercepts and is thus appropriate for matched analyses.

In subsequent sections, we develop these various retrospective approaches and describe estimation procedures and statistical tests for inference using case-only, unmatched case-control, and matched case-control datasets. We illustrate the use of our methods by applying them to unphased genotype and environment data from the Finland-United States Investigation of Non-Insulin-Dependent Diabetes Mellitus (FUSION) Genetics study [Valle et al., 1998]. Using interesting FUSION results as a starting point, we conduct additional simulations to investigate the power and type-I error of the proposed retrospective approaches and compare results among one another, as well as to the prospective approach of Lake et al. [2003].

MATERIALS AND METHODS

NOTATION

Let D be a random variable indicating the disease status of a subject, where $D = 1$ for a case participant and $D = 0$ for a control participant. Let H denote the haplotype pair of a study participant, where $H = (h, h')$ indicates that a subject has haplotypes h and h' . H is not generally observable unless a subject is heterozygous at fewer than 2 SNP loci. Let G be a random variable denoting an individual's observed multilocus genotype, and let $S(G)$ be the set of haplotypes H consistent with the individual's genotype G . If the genotype G includes missing data at a particular SNP, this can be accommodated by including all haplotypes in $S(G)$ which are consistent with the known genotype information. Finally, we let E denote a vector of environmental covariates.

For the methods considered in this paper, we focus inference on parameters modeled within the odds of disease given haplotype and environment. We can express the odds of disease as

$$\theta(H, E) \equiv \frac{P[D=1|H, E]}{P[D=0|H, E]} = \exp(\alpha + X_H \cdot \beta + X_E \cdot \gamma + X_{H \cdot E} \cdot \nu) \quad (1)$$

where α is an intercept, β is a vector of main haplotype effects, γ is a vector of main environmental effects, and ν is a vector of haplotype-environment interaction effects. For each possible haplotype pair H , X_H is a design vector that codes the main haplotype effects [examples are provided in Epstein and Satten, 2003]; X_E is a design vector that codes the main environmental effects as a function of covariate E , and $X_{H \cdot E}$ is a design vector that codes the haplotype-environment interaction. Using the model in (1), we can assess main haplotype effects, main environmental effects, and haplotype-environment interaction effects by appropriate consideration of β , γ , and ν , respectively. Note that the “main” effect of haplotypes X_H can include any identifiable haplotype $\times$ haplotype interaction terms.

RETROSPECTIVE LIKELIHOOD FACTORIZATION

To conduct inference on the parameters in (1), we consider the observed retrospective likelihood of genotype and environmental data conditional on disease data. We can write this likelihood as a function of the underlying haplotype data as

$$L = P[G, E|D] = \sum_{H \in S(G)} P[H, E|D].$$

All our proposed methods are a consequence of the following factorization for L :

$$L = P[G|E, D] P[E|D] = \sum_{H \in S(G)} P[H|E, D] P[E|D] \quad (2)$$

and the following two assumptions: the rare-disease approximation and the assumption of haplotype-environment independence in the target population. We describe how we incorporate these assumptions within our framework in the following sections.

We first explore the information that each term in the factorization (2) contributes to the parameters of interest shown in model (1), starting with the term $P[G|E, D]$. For control subjects ($D = 0$), haplotype-environment independence in the target population and the rare-disease approximation together imply that

$$\begin{aligned} P[G|E, D=0] &= \sum_{H \in S(G)} P[H|E, D=0] \\ &= \sum_{H \in S(G)} P[H|D=0]. \end{aligned} \quad (3)$$

By further assuming Hardy-Weinberg Equilibrium (HWE) in the target population, it follows that control subjects will reflect a random sample from a population that is in HWE, i.e.

$$P[H = (h, h') | D=0] = 2^{1-\delta_{hh'}} p_h p_{h'} \quad (4)$$

where p_h is the frequency of haplotype h in the target population and $\delta_{hh'}$ denotes an indicator function that equals 1 when $h = h'$ and 0 otherwise.

For case subjects ($D = 1$), we can write $P[G|E, D = 1]$ as

$$\begin{aligned}
P[G|E, D=1] &= \sum_{H \in S(G)} P[H|E, D=1] \\
&= \frac{\sum_{H \in S(G)} \theta(H, E) \cdot P[H|D=0]}{\sum_{H' \in S(G)} e^{X_{H'} \cdot \beta + X_{H'} \cdot \gamma} \cdot P[H'|D=0]} \\
&= \frac{\sum_{H \in S(G)} \theta(H, E) \cdot P[H|D=0]}{\sum_{H' \in S(G)} e^{X_{H'} \cdot \beta + X_{H'} \cdot \gamma} \cdot P[H'|D=0]}.
\end{aligned} \tag{5}$$

Note that $P[E|D=0]$ as well as terms involving α and γ cancel from the numerator and denominator of (5). Taking the results of (5) and (3) together, we have shown that $P[G|E, D]$ contains no information on the intercept α or main environmental effects γ .

Within the factorization of (2), we next consider the term $P[E|D]$. If we assume a saturated model for the main effects of the environmental covariate (e.g., if E is categorical with K levels and we allow for $K-1$ parameters $\gamma_1, \gamma_2, \dots, \gamma_{K-1}$ in the vector γ), we now show that the term $P[E|D]$ contains no information on the main haplotype effects β and the haplotype-environment effects v . For control subjects, $P[E|D=0]$ is independent of β and v since we assume E and H are independent among such subjects. For case subjects,

$$P[E|D=1] = \frac{\tilde{\theta}(E) P[E|D=0]}{\sum \tilde{\theta}(E') P[E'|D=0]}$$

where

$$\tilde{\theta}(E) = \frac{P[D=1|E]}{P[D=0|E]} = \sum_H \theta(H, E) \cdot P[H|E, D=0].$$

Using our model in (1), we obtain

$$\tilde{\theta}(E) = \exp(\alpha + \gamma_E + \phi_E) \tag{6}$$

where γ_E is the appropriate component of γ (with the convention that $\gamma_K = 0$) and $e^{\phi_E} = \sum_H e^{X_H \cdot \beta + X_H \cdot \gamma} P[H|E, D=0]$. Thus, $P[E|D]$ depends only on $\tilde{\gamma}_E := \gamma_E + \phi_E - \phi_K$. Because the model for γ is already saturated, $P[E|D]$ contains no information on β or v .

From these results, we see that the likelihood in (2) factors into a term that has information on the haplotype parameters β and haplotype-environment interaction parameters v (namely, $P[G|E, D]$) and a term that involves only parameters γ_E and ϕ_E (namely, $P[E|D]$). We now explore the consequences of this factorization in conducting haplotype-environment interaction analyses.

CASES-ONLY ANALYSES

Our factorization of the retrospective likelihood in the previous section demonstrated that all information on main haplotype effects and haplotype-environment interactions are contained within $P[G|E, D]$ when we assume a saturated model for main environmental effects. We can further show that in certain circumstances all information on the haplotype-environment interaction parameters v are contained within only $P[G|E, D=1]$. Therefore, in these instances, we can conduct inference on haplotype-environment interaction parameters using a cases-only analysis. Such a cases-only analysis for haplotype analysis is interesting to investigate, since other investigators [Piegorisch et al., 1994; Yang et al., 1997] found this procedure to be as efficient for detecting interactions between single SNPs and environment as a full-retrospective analysis of case-control data.

To demonstrate situations where $P[G|E, D=1]$ contains all information on v , we let $H = (h, h')$ and incorporate (4) into (5), such that

$$P[G|E, D=1] = \frac{\sum_{h, h' \in S(G)} e^{X_{hh'} \cdot \beta + X_{hh'} \cdot E \cdot \gamma} \cdot 2^{1-\delta_{hh'}} p_h p_{h'}}{\sum_{h_1, h'_1} e^{X_{h_1 h'_1} \cdot \beta + X_{h_1 h'_1} \cdot E \cdot \gamma} \cdot 2^{1-\delta_{h_1 h'_1}} p_{h_1} p_{h'_1}}. \quad (7)$$

Suppose, for the main haplotype effects on disease, we assume the model $X_{hh'} \cdot \beta = \beta_h + \beta_{h'} + Z_{hh'} \cdot \zeta$ where we include a separate parameter β_h for each haplotype h and additionally allow for a limited number of haplotype-haplotype interactions coded by $Z_{hh'}$. In this model, ζ denotes the parameter vector for the haplotype-haplotype interactions. Then, we can rewrite (7) as

$$P[G|E, D=1] = \frac{\sum_{h, h' \in S(G)} e^{Z_{hh'} \cdot \zeta + X_{hh'} \cdot E \cdot \gamma} \cdot 2^{1-\delta_{hh'}} \tilde{p}_h \tilde{p}_{h'}}{\sum_{h_1, h'_1} e^{Z_{h_1 h'_1} \cdot \zeta + X_{h_1 h'_1} \cdot E \cdot \gamma} \cdot 2^{1-\delta_{h_1 h'_1}} \tilde{p}_{h_1} \tilde{p}_{h'_1}} \quad (8)$$

where $\tilde{p}_h = e^{\beta_h p_h / \Sigma_h} \cdot e^{\beta_h p_h'}$. Under this parameterization in (8), we can see that $P[G|E, D=1]$ is a function of haplotype parameters γ , $\tilde{p}_h$, and ζ . This is in contrast to control participants, who only contribute information on p_h (from equation (3)). Therefore, $P[G|E, D=1]$ in equation (8) contains all information on γ so we can conduct cases-only analyses of haplotype-environment interaction effects by maximizing $P[G|E, D=1]$ with respect to parameters γ , $\tilde{p}_h$, and ζ . We provide details of an efficient Expected-Conditional-Maximization (ECM) algorithm to maximize (8) in the Appendix.

While the interpretation of γ and ζ is straightforward in cases-only analyses, the values of $\tilde{p}_h$ in (8) are more difficult to interpret as they represent a combination of the haplotype frequency and the main haplotype effect on disease. These parameters can correspond to the haplotype frequencies among cases having a particular exposure level E^* , but only under a restrictive situation of no haplotype-haplotype interaction and $X_{hh'} \cdot E^*$ being constant across all possible $H = (h, h')$.

If we require a haplotype-environment analysis involving a dominant or recessive model, then it is necessary to fit $P[G|E, D]$ to data from both case and control participants. In addition, power gains obtained by making model assumptions about multiplicative effects β_h (e.g., assuming that $\beta_h = 0$ for some haplotypes) can only be made when one includes data from control participants. We return to this point in both the Results and Discussion sections.

FULL RETROSPECTIVE LIKELIHOOD ANALYSES

If we wish to conduct simultaneous inference on all model parameters in (1), we must apply a full retrospective likelihood similar in form to that shown in (2). Such application requires modeling of both $P[G|E, D]$ and $P[E|D]$ in the likelihood. However, as discussed earlier, it is inconvenient to use $P[E|D]$ as it requires specification of both environmental main effects and $P[E|D=0]$. This is the main reason investigators do not use the retrospective likelihood in case-control studies of disease. Fortunately, if we assume a saturated model for $P[E]$ in the target population, we can use the results of Prentice and Pyke [1979] to write $P[E|D] \propto P[D|E]$, with the latter probability being far easier to specify. In particular, we can write $P[D|E]$ simply using a modified version of the disease odds $\tilde{\theta}(E)$ provided in equation (6) with the original intercept α replaced by a modified intercept α^* . We denote this modified odds by $\tilde{\theta}^*(E)$.

We can conduct inference on model parameters using this modified retrospective likelihood. For disease-status indicator $d(d=0,1)$, the likelihood takes the form

$$\begin{aligned}
L &\propto P[G|D=d, E] P[D=d|E] \\
&= \frac{\sum_{h, h' \in S(G)} e^{\{\alpha + X_{hh'} \beta + X_E \gamma + X_{hh'} E \nu\}^d} \cdot 2^{1-\delta_{hh'}} p_h p_{h'}}{\bar{\theta}(E)^d} \\
&\times \frac{\theta^*(E)^d}{1+\theta^*(E)} \\
&= \frac{\sum_{h, h' \in S(G)} e^{\{\alpha^* + X_{hh'} \beta + X_E \gamma + X_{hh'} E \nu\}^d} \cdot 2^{1-\delta_{hh'}} p_h p_{h'}}{1+\theta^*(E)}.
\end{aligned} \tag{9}$$

Note that we have replaced α by α^* in this likelihood. By maximizing (9) with respect to parameters α^* , β , γ , and ν , we maximize the retrospective likelihood for a model assuming a saturated distribution for exposure. This allows us to conduct inference on haplotype, environment, and haplotype-environment interaction effects without specifying the (nuisance) distribution of exposure. We note that the likelihood in (9) corresponds to the rare-disease approximation of the profile-likelihood procedure described by Spinka et al. [2005].

RETROSPECTIVE LIKELIHOOD ANALYSES FOR FINELY-STRATIFIED DATA

Case-control studies of genetic diseases sometimes employ finely matched stratifying of the data to remove the effects of potential environmental confounders. In this section, we consider retrospective approaches for the analysis of such matched data that involve relevant stratum-specific parameters. In particular, we propose to use the likelihood of the observed genotype and environment data conditional on the observed disease data and the set of environmental covariates in each stratum. We do not condition on the set of genotypes in each stratum because, as we will show, such conditioning is unnecessary to eliminate stratum-specific intercepts and may thus result in overconditioning. Under this approach, we assume all main and interaction effects are either constant across different strata or appropriately modeled with stratum-specific covariates, i.e. we assume (1) holds but with a stratum-specific intercept.

To develop this retrospective likelihood for finely-stratified data, we let D_{ji} , G_{ji} , and E_{ji} denote the disease, genotype, and environmental data for the i th subject in the j th stratum. Further, we let $\bar{E}_j$ denote the (unordered) set of environmental covariates for the j th stratum, i.e., the values of E without the information about which covariate values were associated with cases and which values were associated with controls. Let n_j denote the number of study participants in the j th stratum, of which n_{0j} are control participants and n_{1j} are case participants. Then, our proposed retrospective likelihood for highly stratified data takes the form

$$\begin{aligned}
L &= \prod_j P \left[\{G_{ji}, E_{ji}, i=1, n_j\} \mid \{D_{ji}, i=1, n_j\}, \bar{E}_j \right] \\
&= \prod_j P \left[\{G_{ji}, i=1, n_j\} \mid \{D_{ji}, E_{ji}, i=1, n_j\}, \bar{E}_j \right] \\
&\quad \times P \left[\{E_{ji}, i=1, n_j\} \mid \{D_{ji}, i=1, n_j\}, \bar{E}_j \right].
\end{aligned} \tag{10}$$

Since $P[G|D, E]$ is already conditional on both D and E , and we further assume genetic effects are constant across strata, we have

$$\begin{aligned}
P \left[\{G_{ji}, i=1, n_j\} \mid \{D_{ji}, E_{ji}, i=1, n_j\}, \bar{E}_j \right] \\
= \prod_{i=1}^{n_j} P \left[G_{ji} \mid D_{ji}, E_{ji} \right]
\end{aligned} \tag{11}$$

where we evaluate $P[G_{ji}|D_{ji}, E_{ji}]$ using results in (3) and (5). From standard theory for conditional logistic regression, it is possible to write

$$P\left[\{E_{ji,i} = 1, n_j\} | \{D_{ji,i=1}, n_j\}, \bar{E}_j\right] = \frac{\prod_{i=1}^{n_j} \theta(E_{ji})^{d_{ji}}}{\sum_{d'_{ji}=1}^{n_j} \prod_{i=1}^{n_j} \theta(E_{ji})^{d'_{ji}}} \quad (12)$$

where the sum in the denominator of (12) is over all vectors d'_{ji} having n_j components, of which n_{0j} are 0s (corresponding to control participants) and n_{1j} are 1s (corresponding to case participants). Equation (12) holds since, within each stratum, permuting over values of D with E fixed is equivalent to permuting over values of E with D fixed. Note that by equation (6), the intercept cancels out of (12) and we have already seen that $P[G|D, E]$ is independent of α . Thus, our proposed likelihood in (10), which is the product of (11) and (12), eliminates stratum-specific intercepts and is thus appropriate for matched or finely-stratified data. We provide details of an efficient ECM algorithm to maximize (10) in the Appendix.

INFERENCE ON HAPLOTYPE-ENVIRONMENT INTERACTIONS

Using the appropriate retrospective likelihood in (8), (9), or (10), we can test hypotheses about or construct estimators of the haplotype-environment interaction parameters v (e.g., $H_0: v = 0$ vs. $H_A: v \neq 0$). For testing hypotheses on v , we can apply either likelihood ratio or score statistics for inference. For common haplotypes, these statistics under H_0 should asymptotically follow a χ^2 with degrees of freedom equal to the number of components of v . For rare haplotypes one may need to apply a permutation procedure for proper inference as in Epstein and Satten [2003].

APPLICATION TO FUSION DATA

We applied our proposed retrospective approaches to data from the FUSION study of type II diabetes [Valle et al., 1998]. Our study sample consists of 796 case subjects with type II diabetes and 415 healthy control subjects. For genotype data, we investigated five tightly-linked SNPs from chromosome 22. Missing genotype rates for the five SNPs ranged between 2.4% and 5.6% with 17.1% of cases and 19.8% of controls missing genotype data at one or more locus. For environmental data, we considered a dichotomous variable denoting whether a subject was overweight (defined by a BMI > 25). Among case participants, 69 had missing BMI values and were removed from our analyses (no control participants were missing BMI). Therefore, our sample for haplotype-environment interaction analysis consists of 727 cases and 415 controls.

Within the FUSION dataset, Epstein and Satten [2003] previously found a common 5-SNP haplotype that appeared to increase the risk of type II diabetes. We denote this haplotype as h_R . We used this previous result as a basis to explore potential interactions between h_R and overweight status that may further increase the risk of type II diabetes. We first applied our retrospective cases-only likelihood in (8) to the FUSION data. We obtained estimates of the interaction effect v between h_R and BMI and tested $H_0: v = 0$ using a likelihood-ratio statistic. Next, we applied our retrospective likelihood in (9) to both the FUSION case and control data. For this likelihood, we considered different genetic models for h_R and chose the “best” model as the one with the lowest value of the Akaike Information Criterion (AIC) [Akaike, 1985]. We obtained estimates of the main h_R effect β , the main BMI effect γ , and the interaction effect v . We also tested $H_0: v = 0$ using a likelihood-ratio statistic.

SIMULATIONS

Based on parameter estimates obtained from our retrospective analyses of the FUSION data, we conducted simulation studies to evaluate the bias, type-I error, and power of our

retrospective approaches. We simulated case and control data in a retrospective manner; that is, we simulated haplotype and environmental data conditional on disease status. For control subjects, we assumed independence of haplotype and environment. We simulated control haplotype data under HWE using estimated haplotype frequencies derived from the FUSION analyses. We simulated control environmental data using the non-parametric estimate of the overweight covariate in the FUSION dataset. For case subjects, we simulated data under the model

$$P[H=(h,h'), E=e|D=1] = \frac{e^{X_{hh'}\beta + X_e\gamma + X_{hh'e}v} \cdot 2^{1-\delta_{hh'}} p_h p_{h'} \cdot F_0(e)}{\sum_{h_1, h'_1} e^{X_{h_1 h'_1}\beta + X_e\gamma + X_{h_1 h'_1 e}v} \cdot 2^{1-\delta_{h_1 h'_1}} p_{h_1} p_{h'_1} \cdot F_0(e^*)}$$

where $F_0(e) = P[E=e|D=0]$ is the non-parametric estimate of overweight covariate in the sample of control participants. We assumed $F_0(0) = 0.17$ and $F_0(1) = 0.83$, which corresponds to the observed values in the FUSION sample.

For each data replicate, we assumed the same sample size as the FUSION study: 727 cases and 415 controls. We assumed values for the haplotype main effect β and the environmental main effect γ that are derived from values based on the retrospective analyses of the FUSION data. We then varied the interaction effect v among both null and alternative values to explore the type-I error and power, respectively, of our proposed retrospective approaches. For each level of interaction, we simulated 1,000 datasets and analyzed each dataset using our retrospective case-only approach (under a multiplicative model) and our full retrospective approach (under multiplicative, dominant, and recessive models). For comparison, we also analyzed each dataset using the prospective-likelihood approach of Lake et al. [2003] under multiplicative, dominant, and recessive models.

We also conducted simulations to assess the performance of our retrospective likelihood for finely-stratified data. Assuming the FUSION sample size of 727 cases and 415 controls, we simulated matched strata with each stratum containing one control and one or more cases (the number of cases varied randomly across strata). For each dataset, we assumed values for the haplotype main effect β and the environmental main effect γ that are derived from values based on the retrospective analyses of the FUSION data. We then varied the interaction effect v among both null and alternative values to explore the type-I error and power, respectively, of our method.

RESULTS

ANALYSIS OF FUSION DATA

Table I provides estimates of the risk parameters derived from the application of our proposed retrospective methods to the unmatched case-control data from the FUSION study. For our cases-only approach, we found the estimate of the interaction between the risk haplotype (h_R) and overweight status under the implicit multiplicative model to be close to significant ($\hat{\gamma} = -0.34$, $p = 0.059$ at Bonferroni-corrected significance level $\alpha = 0.013$). As expected, this result is quite similar to our result based on using the full retrospective likelihood (using data from cases and controls) under a multiplicative model ($\hat{\gamma} = -0.33$, $p = 0.061$). The difference in P -values reflects slight differences in the way the variance-covariance matrix was calculated. Additionally, we assessed the interaction effects of h_R and overweight status under both dominant and recessive models, but found no evidence of significance under either model. Table I also reports the AIC [Akaike, 1985] values for each model. Based on these values, the multiplicative model appears to fit the FUSION data the best in this situation.

SIMULATIONS: UNMATCHED DATA

Figure 1 provides the type-I error and power results at $\alpha = 0.05$ for detecting haplotype-environment interactions using the full-retrospective, cases-only and prospective approaches. Under multiplicative and dominant models, the empirical type-I error rates of all approaches considered closely matched the nominal value of 0.05. Under a recessive model, the full-retrospective and prospective approaches were conservative with empirical type-I error rates of 0.036 and 0.030, respectively.

As anticipated, we found our proposed retrospective approaches had increased power to detect interactions relative to the prospective approach of Lake et al. (2003). Also, under a multiplicative model, we found that the full-retrospective and cases-only yielded near-identical power for detecting interactions (the two programs calculated the asymptotic variance-covariance matrix in slightly different ways). This result verifies our theoretical claims for the efficiency of the cases-only analysis because the full-retrospective analysis used a saturated categorical model for the environmental exposure.

We conclude that the cases-only approach for detecting haplotype-environment interactions is valuable and cost-effective since it maintains the power of a full retrospective analysis under a reasonable genetic model, while avoiding the costly genotyping of control subjects.

Table II reports the mean and variance estimates of the haplotype main effect β , the environmental main effect γ , and the interaction effect v under the different genetic models. For the full retrospective and prospective approaches, we estimated the variance-covariance matrix using the empirical observed information matrix [McLachlan and Krishnan, 1997] as suggested by Lake et al. [2003]. For the cases-only approach, we estimated the variance-covariance matrix using the observed information matrix. Under multiplicative and dominant models, we found that the full-retrospective and prospective approaches yielded unbiased parameter estimates and produced variance estimates that closely mirrored the empirical variance. For the multiplicative models, we found that the cases-only approach yielded near-identical results for the v parameter relative to the full-retrospective approach, which again demonstrates the utility of the cases-only approach for inference on haplotype-environment interaction effects. Finally, we note that the estimated/empirical variances for the retrospective/case-only approaches were substantially smaller relative to those for the prospective approach, indicating the retrospective approaches provide more precise parameter estimates.

Under a recessive model, we found that both the full-retrospective and the prospective approaches yielded flawed results using our simulation design. We found the retrospective approach consistently underestimated the haplotype effect β and overestimated the interaction effect v , although the estimated and empirical variance matched up well for all parameters. On the other hand, the prospective approach yielded unbiased parameter estimates, but the estimated variances were approximately five times as large as the empirical variances. These results demonstrate that it will be more difficult to conduct accurate inference on recessive haplotype-interaction effects with realistic sample sizes. We believe the reason the retrospective approach performs poorly in assessing haplotype-environment interaction effects under a recessive model is because there will be few observations that have both two copies of the at-risk haplotype and the environmental covariate for a study having the same characteristics as the FUSION study. To support this idea, we conducted additional simulations under a recessive model using our previous simulation design, but decreased the proportion of overweight controls from 0.83 to 0.50. In this situation, Table III shows that our retrospective approach yielded more consistent results and the type-I error rate for testing $v = 0$ under the null hypothesis was appropriate (0.053 at $\alpha = 0.05$). We also note that we obtained improved performance for the recessive model by keeping the proportion of overweight controls as in the FUSION study but increasing the sample size (data not shown).

We noted occasional convergence problems for both the full retrospective and prospective approaches. Under a recessive model, the full retrospective and prospective approaches failed to converge 0.7–1.2% and 2.3–2.7% of the time, respectively, depending on the true value of v . Under multiplicative and dominant models, both approaches failed to converge less than 0.4% of the time. The cases-only approach, on the other hand, converged for all tested datasets.

SIMULATIONS: FINELY-STRATIFIED DATA

For finely-stratified data, Table IV provides mean and variance estimates of the haplotype main effect β , the environmental main effect γ , and the interaction effect v under multiplicative, dominant, and recessive models. We again approximate the estimated variance-covariance matrix using the empirical observed information matrix as described in McLachlan and Krishnan [1997].

Note that our results in Table IV for finely-matched data are similar to those reported for unmatched data in Table II, even though the data are over-matched (no matching is actually required for these data). Under multiplicative and dominant models, we found our retrospective approach for finely-matched data returned unbiased parameter estimates and yielded variance estimates that were quite similar to the empirical variance. However, under a recessive model, our retrospective approach underestimated the main haplotype effects β and overestimated the haplotype-environment interaction effects v , although the estimated and empirical variances matched up well for all parameters. We again feel this bias arises from the sparsity of key haplotype/exposure combinations in the data. As with the unmatched data, we repeated the simulations with a decreased frequency of overweight subjects in the controls and observed a marked improvement in the estimates (see Table III).

We also examined the type-I error and power of our retrospective approach for finely-matched data under multiplicative, dominant, and recessive models. Assuming nominal $\alpha = 0.05$, we found the empirical type-I error rate was appropriate under all three models (0.049 for multiplicative, 0.046 for dominant, and 0.053 for recessive). Under the alternative model of $v = -0.4$, we found the power of our approach for a given model was quite similar to the analogous power of the retrospective approach for unmatched case-control data (0.416 for multiplicative, 0.243 for dominant, and 0.113 for recessive).

DISCUSSION

We have presented a variety of simple retrospective approaches for evaluating haplotype, environment, and haplotype-environment interaction effects in case-only, unmatched, and matched case-control studies of disease. By making the assumptions of a rare disease and haplotype-environment independence in the target population, we have shown that these approaches yield improved power and efficiency for detecting haplotype-environment interaction effects compared to more traditional prospective approaches. We intend on implementing these novel approaches in the next release of our haplotype-analysis package CHAPLIN, which can be downloaded free of charge from our website (see Electronic-Database Information).

We initially focused on a retrospective cases-only approach for evaluating haplotype-environment interaction effects. Because the cases-only analysis is based on the retrospective likelihood for case and control data, it enjoys the usual asymptotic optimality of the full likelihood analysis when we assume a saturated model for main environmental effects and a separate multiplicative effect for each haplotype. For situations where different model assumptions are appropriate (e.g., for dominant or recessive effects of haplotypes or to model the multiplicative effects of a restricted number of haplotypes), we require the full retrospective likelihood for case-control data to conduct analysis. Relative to the cases-only procedure, we

note that we can obtain additional efficiency gains on interaction parameters by making additional model assumptions in the full retrospective likelihood (e.g., dose-response model for an environmental covariate or modeling only a restricted number of haplotypes in a multiplicative model).

While our simulations focused on a single haplotype-environment interaction that increases risk of disease, our methods apply to arbitrary design vectors, including design vectors that measure haplotype similarity. For example, haplotype-clustering methods [Durrant et al., 2004; Tzeng et al., 2006] that group haplotypes that likely represent ancestral haplotype categories using cladistic approaches could be used to define groups of haplotypes that have a similar effect on the odds of disease. Design vectors that are based on haplotype clusters could reduce the degrees of freedom for testing and may improve power. Tzeng et al. [2006] developed such a clustering approach that modeled the odds of disease as a function of the haplotype clusters within a prospective-likelihood framework. We plan to investigate whether a retrospective version of this approach improves the power to detect both haplotype and haplotype-environment interaction effects relative to the authors' prospective version.

In the development of our likelihood, we assumed that haplotype frequencies in the target population were in HWE. While the HWE assumption often holds, situations may arise where we would expect departure from HWE, such as when inbreeding occurs or when population stratification exists and is unaccounted for in analysis. In such situations, Satten and Epstein (2004) showed that departure from HWE in the target population can lead to incorrect inference on haplotype effects in retrospective approaches. To resolve this problem, we could model the departure from HWE by estimating a fixation index among the control participants, as described in Satten and Epstein [2004] and Lin and Zeng [2006]. Satten and Epstein [2004] showed that including a single fixation index in the haplotype distribution greatly diminished the bias even in simulated data where individual haplotype pairs had a different fixation index. In addition, we are developing projection methods similar to those described in Allen et al. [2005] to make the retrospective approach robust to misspecification of the haplotype distribution.

Our proposed retrospective approach also requires assumptions of haplotype-environment independence in the target population as well as a rare disease. Regarding the assumption of haplotype-environment independence, we note that haplotypes and environmental covariates could be correlated as a result of unmodeled population stratification (where haplotype and environment are independent within each sub-population). However, even within an unstratified population, correlation between haplotype and environment can still arise in specific situations, such as in haplotype analysis of behavioral exposures (e.g., a haplotype that affects alcohol metabolism is unlikely to be independent of alcohol consumption). To assess the sensitivity of our approach to this assumption, we conducted additional simulations in which we induced correlation between the environmental covariate and the haplotype of interest in the population. Confirming the findings of Spinka et al. [2005], we found that modest violations of the haplotype-environment independence assumption can lead to unreliable inference using asymptotic procedures. If one suspects correlation of haplotype and environmental factors in the target population, then we advocate the use of prospective methods as described in Spinka et al. [2005].

To assess the impact of the rare-disease assumption on our method, we conducted simulations with prospectively generated data under the assumption that the disease prevalence was either 0.05 or 0.10. We chose these values because most researchers employ case-control sampling for diseases with prevalence less than 0.10 in the target population [Kleinbaum et al., 1982, Greenland and Thomas, 1982]. We also note that a prevalence of 0.05 corresponds to the prevalence of type-II diabetes in the Finnish target population utilized by the FUSION case-

control study [Tuomilehto et al., 1991]. Under multiplicative and dominant models, we found that parameter estimates were similar to those in Table II and estimated size/power were similar to those shown in Figure 1. Under recessive models, we found that our approach could lead to slightly anticonservative tests for a disease prevalence of 10%, but yielded proper inference for a disease prevalence of 5%. We feel the anticonservative nature of the test under a recessive model for 10% prevalence is due to the control haplotypes showing modest departure from HWE. To resolve this issue, we can follow the strategy outlined in the previous paragraph and model a fixation index [Satten and Epstein, 2004] within our likelihood framework to obtain valid inference.

We have developed our retrospective approaches for haplotype analysis under the assumptions of rare disease, haplotype-environment independence in the target population, and HWE of haplotypes in the target population. Together, these assumptions imply haplotype-environment independence and HWE of haplotypes in the control population. In fact, all results in this paper could be developed making only the two assumptions that haplotypes and environmental covariates are independent in the control population and that haplotypes in the control population are in HWE. However, we are unaware of any situation other than a rare disease where these assumptions are biologically plausible.

For the implementation of the retrospective approach for matched case-control data, we made an assumption that haplotype frequencies were independent of the strata. While this assumption is reasonable if matching strata are formed based on factors such as age and gender, it may not be reasonable if matching is based on factors such as self-reported ethnicity. In this latter situation, haplotype frequencies may vary across strata, which violates our model assumptions. One potential solution to this problem would be to model and estimate ethnic-specific haplotype frequencies. An alternative would be to conduct separate analyses in each ethnic group, if sample size permitted.

ACKNOWLEDGMENTS

We thank the FUSION study investigators for allowing us to present results from the analysis of FUSION data. We also acknowledge useful conversations with Nilanjan Chatterjee. This work was supported by National Institutes of Health grants T32-GM074909 (to L.C.K.), HG003618 (to M.P.E.), and HL077663 (to A.S.A.).

Contract grant sponsor: National Institutes of Health; Contract grant numbers: T32-GM074909, HG003618 and HL077663.

APPENDIX A ECM ALGORITHM FOR CASES-ONLY ANALYSIS

We propose an ECM algorithm to maximize the cases-only likelihood (8). We maximize the likelihood with respect to the haplotype parameters $\tilde{p}$, the haplotype-environment interaction parameter ν , and (if modeled) the haplotype-haplotype interaction parameters ζ while allowing for a Lagrange multiplier to constrain $\sum_h \tilde{p}_h = 1$. Alternatively, we could maximize the cases-only likelihood without applying any constraints by replacing $\tilde{p}_k$ by e^{β_k} . However, such maximization could become unstable given a large number of possible haplotypes.

The ECM algorithm consists of an E-step followed by two conditional-maximization (CM) steps. In the first CM step, we update $\tilde{p}$ and the Lagrange multiplier holding ν and ζ constant. In the second CM step, we update ν and ζ holding $\tilde{p}$ and the Lagrange multiplier fixed. To implement the ECM algorithm, we first assume haplotypes are observed and subsequently construct the complete-data likelihood analog of (8) for a given case participant as

$$L_{Full} = \frac{e^{Z_{hh'} \cdot \zeta + X_{hh' \cdot E} \cdot \nu} \cdot 2^{1 - \delta_{hh'}} \tilde{p}_h \tilde{p}_{h'}}{\sum_{h_1, h'_1} e^{Z_{h_1 h'_1} \cdot \zeta + X_{h_1 h'_1 \cdot E} \cdot \nu} \cdot 2^{1 - \delta_{h_1 h'_1}} \tilde{p}_{h_1} \tilde{p}_{h'_1}}.$$

We derive the score equation for $\tilde{p}_k$ as

$$S_{\tilde{p}_k} = -\lambda + \sum_i \left[\frac{I[h_i=k] + I[h'_i=k]}{\tilde{p}_k} - 2\tilde{p}_k \right]$$

where i indexes the case participant, 1 is a Lagrange multiplier introduced to ensure the constraint $\sum_h \tilde{p}_h = 1$, and

$$\tilde{p}_k = \frac{\sum_h e^{Z_{hk} \cdot \zeta + X_{hk \cdot E} \cdot v} \tilde{p}_h}{\sum_{h_1, h'_1} e^{Z_{h_1 h'_1} \cdot \zeta + X_{h_1 h'_1 \cdot E} \cdot v} \cdot 2^{1-\delta_{h_1 h'_1}} \tilde{p}_{h_1} \tilde{p}_{h'_1}}.$$

At the r th iteration, we assume the values of $\tilde{p}_k$, v and ζ are given by $\tilde{p}_k^{(r)}$, $v^{(r)}$, and $\zeta^{(r)}$, respectively. We first calculate $\tilde{p}_k^{-(r)}$ and subsequently update $\tilde{p}_k$ in the first CM step using

$$\tilde{p}_k^{(r+1)} = \frac{m_k^{(r)}}{\lambda^{(r)} + 2n\tilde{p}_k^{-(r)}}$$

where n denotes the total sample size and

$$m_k^{(r)} = \sum_i \frac{\sum_{h, h' \in S(G_i)} \{I[h=k] + I[h'=k]\} e^{Z_{hh'} \cdot \zeta^{(r)} + X_{hh' \cdot E_i} \cdot v^{(r)}} \cdot 2^{1-\delta_{hh'}} \cdot \tilde{p}_h^{(r)} \tilde{p}_{h'}^{(r)}}{\sum_{h, h' \in S(G_i)} e^{Z_{hh'} \cdot \zeta^{(r)} + X_{hh' \cdot E_i} \cdot v^{(r)}} \cdot 2^{1-\delta_{hh'}} \cdot \tilde{p}_h^{(r)} \tilde{p}_{h'}^{(r)}}$$

is the expected number of haplotypes of type k at the r th iteration conditional on the observed data (the E step). $\lambda^{(r)}$ is chosen so that $\sum_k \tilde{p}_k^{(r+1)} = 1$. Note that $\lambda^{(r)} \rightarrow 0$ as $r \rightarrow \infty$, since $\sum_k \tilde{p}_k^{(r)} = 1$; the Lagrange multiplier is required at early iterations because $\tilde{p}_k^{-(r)}$ is calculated using $\tilde{p}^{(r)}$ instead of $\tilde{p}^{(r+1)}$.

Once we update $\tilde{p}^{(r+1)}$, we can obtain $v^{(r+1)}$ and $\zeta^{(r+1)}$ in the second CM step by maximizing the complete-data log-likelihood with respect to v and ζ holding $\tilde{p}$ fixed at value $\tilde{p}^{(r+1)}$. We use a quasi-Newton algorithm similar to the one proposed in Epstein and Satten (2003) for this purpose, which makes use of the appropriate score vectors derived from the likelihood:

$$\begin{aligned} \frac{\partial \log(L_{full}^{(r+1)})}{\partial \zeta} &= \sum_i Z_{h_i h'_i} - \frac{\sum_{h_1, h'_1} Z_{h_1 h'_1} \cdot e^{Z_{h_1 h'_1} \cdot \zeta + X_{h_1 h'_1 \cdot E_i} \cdot v} \cdot 2^{1-\delta_{h_1 h'_1}} \tilde{p}_{h_1}^{(r+1)} \tilde{p}_{h'_1}^{(r+1)}}{\sum_{h_1, h'_1} e^{Z_{h_1 h'_1} \cdot \zeta + X_{h_1 h'_1 \cdot E_i} \cdot v} \cdot 2^{1-\delta_{h_1 h'_1}} \tilde{p}_{h_1}^{(r+1)} \tilde{p}_{h'_1}^{(r+1)}}, \\ \frac{\partial \log(L_{full}^{(r+1)})}{\partial v} &= \sum_i X_{h_i h'_i \cdot E_i} - \frac{\sum_{h_1, h'_1} X_{h_1 h'_1 \cdot E_i} \cdot e^{Z_{h_1 h'_1} \cdot \zeta + X_{h_1 h'_1 \cdot E_i} \cdot v} \cdot 2^{1-\delta_{h_1 h'_1}} \tilde{p}_{h_1}^{(r+1)} \tilde{p}_{h'_1}^{(r+1)}}{\sum_{h_1, h'_1} e^{Z_{h_1 h'_1} \cdot \zeta + X_{h_1 h'_1 \cdot E_i} \cdot v} \cdot 2^{1-\delta_{h_1 h'_1}} \tilde{p}_{h_1}^{(r+1)} \tilde{p}_{h'_1}^{(r+1)}}. \end{aligned}$$

By using a quasi-Newton algorithm, we are able to maximize the likelihood without inverting the Hessian matrix. Once we have completed the last CM step, we continue iterating the ECM algorithm until the various parameters converge.

APPENDIX B: ECM ALGORITHM FOR FINELY-STRATIFIED DATA

We also use an ECM algorithm to maximize the likelihood for finely-stratified data shown in (10). In the first CM step, we update β , γ , and v , holding the haplotype frequencies p constant. In the second CM step, we update p using a Lagrange multiplier that constrains $\sum_h p_h = 1$, while holding β , γ , and v constant. To implement the ECM algorithm, we assume haplotypes are

observed and subsequently consider the following full-data likelihood analog to (10) for stratum j :

$$L_{Full,j} = \frac{\prod_{i=1}^{n_j} \left(e^{X_{h_{ji}h'_{ji}} \cdot \beta + X_{E_{ji}} \cdot \gamma + X_{h_{ji}h'_{ji} \cdot E_{ji}} \cdot v} \right)^{d_{ji}} \cdot 2^{1-\delta_{h_{ji}h'_{ji}}} p_{h_{ji}} p_{h'_{ji}}}{\sum_{d'_{ji}=1}^{n_j} \prod_{h,h'} \left(\sum_e e^{X_{hh'} \cdot \beta + X_{E_{ji}} \cdot \gamma + X_{hh' \cdot E_{ji}} \cdot v} 2^{1-\delta_{hh'}} p_h p_{h'} \right)^{d'_{ji}}}.$$

Using notation and methods developed by Howard (1972) and Gail et al. [1981], we let

$$U_{ji} = \left(\sum_{h,h'} e^{X_{hh'} \cdot \beta + X_{E_{ji}} \cdot \gamma + X_{hh' \cdot E_{ji}} \cdot v} 2^{1-\delta_{hh'}} p_h p_{h'} \right)^{d'_{ji}}.$$

Then

$$B(n_{1j}, n_j) = \sum_{d'_{ji}} \prod_{i=1}^{n_j} U_{ji}$$

where n_{1j} denotes the number of cases in stratum j . We then use the likelihood components from all strata j to see that the score equation for β is

$$S_{\beta} = \sum_e \sum_{h,h'} f_{hh',e} X_{hh'} - \sum_j \frac{\partial \log B(n_{1j}, n_j)}{\partial \beta}$$

where $f_{hh',e}$ is the expected number of case subjects with haplotypes (h, h') and environmental covariate e . We use the following recursion [Howard, 1972; Gail et al., 1981] to (calculate the necessary pieces of the score equation:

$$B(n_{1j}, n_j) = B(n_{1j} - 1, n_j) + U_{jn_j} \cdot B(n_{1j} - 1, n_j - 1)$$

with initial conditions: $B(0, n) = 1$ for $n \geq 0$ and $B(m; n) = 0$ for $m > n$. Using these relationships, we can also derive analogous score equations for γ and v .

The score equation for p_k is

$$S_{p_k} = \frac{m_k}{p_k} - \sum_j \frac{\partial \log B(n_{1j}, n_j)}{\partial p_k} + 2c$$

where c is the number of controls in the sample, and $2c$ is the Lagrange multiplier which ensures $\sum_K p_k = 1$. At the $(r+1)$ th iteration, we can use the current estimates of β , γ and v to update $p_k^{(r+1)}$ using

$$p_k^{(r+1)} = \frac{m_k}{2c + \sum_j \frac{\partial \log B(n_{1j}, n_j)}{\partial p_k}}.$$

Again, we use the recursion from Howard [1972] and Gail et al. [1981] to calculate $\partial B(n_{1j}, n_j) / \partial p_k$ and $B(n_{1j}, n_j)$ using $\beta^{(r+1)}$, $\gamma^{(r+1)}$, $v^{(r+1)}$, and $p^{(r)}$.

REFERENCES

- Akaike, H. Prediction and entropy.. In: Atkinson, AC.; Feinberg, SE., editors. A celebration of statistics. Springer; New York: 1985. p. 1-24.
- 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, Tsiatis AA. Locally-efficient robust estimation of haplotype-disease association in family-based studies. *Biometrika* 2005;92:559–571.
- Carroll RJ, Wang S, Wang CY. Prospective analysis of logistic case-control studies. *J Am Stat Assoc* 1995;90:157–169.
- Chapman JM, Cooper JD, Todd JA, Clayton DG. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered* 2003;56:18–31. [PubMed: 14614235]
- 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]
- Epstein MP, Satten GA. Inference on haplotype effects in case-control studies using unphased genotype data. *Am J Hum Genet* 2003;73:1316–1329. [PubMed: 14631556]
- Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 1995;12:921–927. [PubMed: 7476138]
- Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork N. Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE variation and Alzheimer's disease. *Genome Res* 2001;11:143–151. [PubMed: 11156623]
- Gail MH, Lubin JH, Rubinstein LV. Likelihood calculations for matched case-control studies and survival studies with tied death times. *Biometrika* 1981;68:703–707.
- Greenland S, Thomas DC. On the need for the rare-disease assumption in case-control studies. *Am J Epidemiol* 1982;116:547–553. [PubMed: 7124721]
- Howard S. Discussion of “Regression Models and Life Tables.”. *J R Stat Soc B* 1972;34:210–211.
- Kleinbaum, DG.; Kupper, LL.; Morgenstern, H. *Epidemiologic research: principles and Quantitative Methods*. Lifetime Learning Publications; Belmont, CA: 1982.
- Kraft P, Cox DG, Paynter RA, Hunter D, De Vivo I. Accounting for haplotype uncertainty in matched association studies: a comparison of simple and flexible techniques. *Genet Epidemiol* 2005;28:261–272. [PubMed: 15637718]
- Lake SL, Lyon H, Tantisira K, Silverman EK, Weiss ST, Laird NM, Schaid DJ. Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. *Hum Hered* 2003;55:56–65. [PubMed: 12890927]
- Lin DY, Zeng D. Likelihood-based inference on haplotype effects in genetic association studies. *J Am Stat Assoc* 2006;101:89–104.
- Lin DY, Zeng D, Millikan R. Maximum-likelihood estimation of haplotype effects and haplotype-environment interactions in association studies. *Genet Epidemiol* 2005;29:299–312. [PubMed: 16240443]
- McLachlan, GJ.; Krishnan, T. *The EM algorithm and extensions*. Wiley; New York: 1997.
- Morris RW, Kaplan NL. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet Epidemiol* 2002;23:221–233. [PubMed: 12384975]
- Olson JM, Wijsman EM. Design and sample-size considerations in the detection of linkage disequilibrium with a disease locus. *Am J Hum Genet* 1994;55:574–580. [PubMed: 8079996]
- Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med* 1994;13:153–162. [PubMed: 8122051]
- Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika* 1979;66:403–412.
- Rosenberg PS, Che A, Chen BE. Multiple hypothesis testing strategies for genetic case-control association studies. *Stat Med* 2006;25:3134–3149. [PubMed: 16252274]
- Satten GA, Epstein MP. Comparison of prospective and retrospective methods for haplotype inference in case-control studies. *Genet Epidemiol* 2004;27:192–201. [PubMed: 15372619]
- Spinka C, Carroll RJ, Chatterjee N. Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. *Genet Epidemiol* 2005;29:108–127. [PubMed: 16080203]

- Tavtigian S, Simard J, Teng D, Abtin V, Baumgard M, Beck A, Camp J, et al. A candidate prostate cancer susceptibility gene at chromosome 17p. *Nat Genet* 2001;27:172–180. [PubMed: 11175785]
- Tuomilehto J, Korhonen HJ, Kartovaara L, Salomaa V, Stengard JH, Pitkanen M, Aro A, Javela K, Uusitupa M, Pitkaniemi J. Prevalence of diabetes mellitus and impaired glucose tolerance in the middle-aged population of three areas in Finland. *Intl J Epidemiol* 1991;20:1010–1017.
- Tzeng JY, Roeder K. Comment on 'Likelihood-based inference on haplotype effects in genetic association studies 2006. *J Am Stat Assoc* 101:111–114.
- Tzeng JY, Wang CH, Kao JT, Hsiao CK. Regression-based association analysis with clustered haplotypes through use of genotypes. *Am J Hum Genet* 2006;78:231–242. [PubMed: 16365833]
- Valle T, Tuomilehto J, Bergman RN, Ghosh S, Hauser ER, Eriksson J, Nylund SJ, Kohtamaki K, Toivanen L, Vidgren G, Tuomilehto-Wolf E, Ehnholm C, Blaschak J, Langefeld CD, Watanabe RM, Magnuson V, Ally DS, Hagopian WA, Ross E, Buchanan TA, Collins F, Boehnke M. Mapping genes for NIDDM: design of the Finland-United States Investigation of NIDDM Genetics (FUSION) study. *Diab Care* 1998;21:949–958.
- Yang Q, Khoury MJ, Flanders WD. Sample size requirements in case-only designs to detect gene-environment interaction. *Am J Epidemiol* 1997;146:713–720. [PubMed: 9366618]

Fig. 1.

Type-I error and power at $\alpha = 0.05$ of full retrospective, cases-only, and prospective approaches for detecting haplotype-environment interactions. The true value of the main haplotype effect β is 0.65 and the true value of the main environmental effect γ is 1.39. The proportion of overweight controls is 0.83. We varied the haplotype-environment interaction parameter ν among values between -0.8 and 0 (null model). Results are based on 1,000 replicates.

TABLE I
Application of the cases-only and full retrospective methods to the FUSION data

Method	Model	$\hat{\beta}$	$\hat{\gamma}$	$\hat{v}$	P^*	AIC
Cases-only Full retrospective	Multiplicative	NA	NA	-0.34	0.059	NA
	Multiplicative	0.65	1.39	-0.33	0.061	7732.9
	Dominant	0.68	1.39	-0.41	0.095	7738.8
	Recessive	0.70	1.21	-0.48	0.177	7744.0

β denotes the main effect of the risk haplotype (h_R). γ denotes the main effect of overweight status, and v denotes the interaction effect between h_R and overweight status. P^* : P -value for LR test of $H_0: v=0$

TABLE II

Simulation results for haplotype-environment interactions

Model	True v	Full retrospective			Cases-only		Prospective	
		$\hat{\beta}$	$\hat{\gamma}$	$\hat{v}$	$\hat{v}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{v}$
Mult.	0.0	Mean	1.414	-0.004	-0.004	0.685	1.431	-0.037
		Variance	0.071	0.078	0.070	0.121	0.100	0.133
	-0.4	Est. variance	0.074	0.090	0.066	0.112	0.104	0.123
		Mean	1.392	-0.388	-0.388	0.666	1.397	-0.411
Dom.	0.0	Variance	0.068	0.052	0.052	0.107	0.095	0.117
		Est. variance	0.072	0.054	0.052	0.096	0.086	0.108
	-0.4	Mean	1.423	-0.009	NA	0.670	1.420	-0.010
		Variance	0.106	0.149	NA	0.210	0.120	0.229
Rec.	0.0	Est. variance	0.100	0.140	NA	0.177	0.103	0.195
		Mean	1.415	-0.406	NA	0.652	1.402	-0.399
	-0.4	Variance	0.076	0.104	NA	0.159	0.092	0.177
		Est. variance	0.083	0.114	NA	0.169	0.095	0.188
	0.0	Mean	1.407	0.099	NA	0.668	1.407	-0.003
		Variance	0.054	0.317	NA	0.583	0.056	0.638
	-0.4	Est. variance	0.051	0.374	NA	0.119	0.048	0.133
		Mean	1.402	-0.313	NA	0.636	1.393	-0.354
	-0.4	Variance	0.049	0.349	NA	0.611	0.067	0.683
		Est. variance	0.050	0.365	NA	0.129	0.047	0.145

β denotes the main effect of the risk haplotype, γ denotes the main effect of the environmental covariate and v denotes the interaction effect between risk haplotype and environment. The true value of β is 0.65 and the true value of γ is 1.39. The proportion of overweight controls is 0.83. Results are based on 1,000 replicates.

TABLE III
Simulation results using retrospective approach assuming recessive model and decreased proportion of overweight control subjects

True ν	Unmatched				Finely-matched			
	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\nu}$	$\hat{\gamma}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\nu}$	$\hat{\gamma}$
0.0	Mean	1.391			0.627	1.401		0.022
	Variance	0.084	0.020	0.104	0.080	0.023	0.090	0.090
-0.4	Estimated variance	0.083	0.020	0.092	0.086	0.025	0.095	0.095
	Mean	0.630	1.396	-0.386	0.628	1.402	-0.389	-0.389
	Variance	0.085	0.019	0.104	0.079	0.024	0.099	0.099
	Estimated variance	0.079	0.019	0.105	0.082	0.024	0.110	0.110

β denotes the main effect of the risk haplotype, γ denotes the main effect of the environmental covariate, and ν denotes the interaction effect between risk haplotype and environment. The true value of β is 0.65 and the true value of γ is 1.39. Datasets consists of 727 cases and 415 controls. The proportion of overweight controls is 0.50. For the finely-matched data, each stratum contains one control and one or more cases. Results are based on 1,000 replicates.

Simulation results for finely-matched data

TABLE IV

True ν	Multiplicative			Dominant			Recessive		
	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\nu}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\nu}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\nu}$
0.0	Mean	1.416	0.015	0.661	1.439	-0.016	0.563	1.415	0.087
	Variance	0.073	0.091	0.145	0.110	0.143	0.351	0.060	0.351
-0.4	Estimated variance	0.078	0.100	0.147	0.110	0.147	0.383	0.059	0.389
	Mean	0.647	1.426	0.656	1.412	-0.405	0.544	1.416	-0.300
	Variance	0.057	0.075	0.117	0.088	0.112	0.339	0.057	0.352
	Estimated variance	0.061	0.081	0.118	0.090	0.117	0.375	0.058	0.387

β denotes the main effect of the risk haplotype, γ denotes the main effect of the environmental covariate, and ν denotes the interaction effect between risk haplotype and environment. The true value of β is 0.65 and the true value of γ is 1.39. Datasets consist of 727 cases and 415 controls. Each stratum contains one control and one or more cases. The proportion of overweight controls is 0.83. Results are based on 1,000 replicates.