

Retrospective analysis of haplotype-based case–control studies under a flexible model for gene–environment association

YI-HAU CHEN

Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, People’s Republic of China

NILANJAN CHATTERJEE*

*Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute,
6120 Executive Boulevard, EPS 8038, Rockville, MD 20852, USA
chattern@mail.nih.gov*

RAYMOND J. CARROLL

Department of Statistics, Texas A&M University, TAMU 3143, College Station, TX 77843-3143, USA

SUMMARY

Genetic epidemiologic studies often involve investigation of the association of a disease with a genomic region in terms of the underlying haplotypes, that is the combination of alleles at multiple loci along homologous chromosomes. In this article, we consider the problem of estimating haplotype–environment interactions from case–control studies when some of the environmental exposures themselves may be influenced by genetic susceptibility. We specify the distribution of the diplotypes (haplotype pair) given environmental exposures for the underlying population based on a novel semiparametric model that allows haplotypes to be potentially related with environmental exposures, while allowing the marginal distribution of the diplotypes to maintain certain population genetics constraints such as Hardy–Weinberg equilibrium. The marginal distribution of the environmental exposures is allowed to remain completely nonparametric. We develop a semiparametric estimating equation methodology and related asymptotic theory for estimation of the disease odds ratios associated with the haplotypes, environmental exposures, and their interactions, parameters that characterize haplotype–environment associations and the marginal haplotype frequencies. The problem of phase ambiguity of genotype data is handled using a suitable expectation–maximization algorithm. We study the finite-sample performance of the proposed methodology using simulated data. An application of the methodology is illustrated using a case–control study of colorectal adenoma, designed to investigate how the smoking-related risk of colorectal adenoma can be modified by “NAT2,” a smoking-metabolism gene that may potentially influence susceptibility to smoking itself.

Keywords: Case–control studies; EM algorithm; Gene–environment interactions; Haplotype; Semiparametric methods.

*To whom correspondence should be addressed.

1. INTRODUCTION

Genetic epidemiologic studies often involve investigation of the association between a disease and a candidate genomic region of biologic interest. Typically, in such studies, genotype information is obtained on multiple loci that are known to harbor genetic variations within the region of interest. An increasingly popular approach for analysis of such multilocus genetic data are haplotype-based regression methods, where the effect of a genomic region on disease risk is modeled through “haplotypes,” the combinations of alleles (gene variants) at multiple loci along individual homologous chromosomes. It is believed that association analysis based on haplotypes, which can efficiently capture inter-loci interactions as well as “indirect association” due to “linkage disequilibrium” of the haplotypes with unobserved causal variant(s), can be more powerful than more traditional locus-by-locus methods (Schaid, 2004).

A technical problem for haplotype-based regression analysis is that in traditional epidemiologic studies, the haplotype information for the study subjects is not directly observable. Instead, locus-specific genotype data are observed, which contain information on the pair of alleles a subject carries on his/her pair of homologous chromosomes at each of the individual loci but does not provide the “phase information,” that is which combinations of alleles appear across multiple loci along the individual chromosomes. In general, the genotype data of a subject will be phase ambiguous whenever the subject is heterozygous at 2 or more loci. Statistically, the lack of phase information can be viewed as a special missing data problem.

Recently, a variety of methods have been developed for haplotype-based analysis of case–control data using the logistic regression model (Zhao *and others*, 2003; Lake *and others*, 2003; Epstein and Satten, 2003; Satten and Epstein, 2004; Spinka *and others*, 2005; Lin and Zeng, 2006; Chatterjee *and others*, 2006). Two classes of methods, namely, “prospective” and “retrospective” have evolved. Prospective methods ignore the retrospective nature of the case–control design. In the classical setting, without any missing data, justification of prospective analysis of case–control data relies on the well-known result about the equivalence of prospective and retrospective likelihoods under a semiparametric model that allows the distribution of the underlying covariates to remain completely nonparametric (Andersen, 1970; Prentice and Pyke, 1979). Even with missing data, the equivalence of the prospective and retrospective likelihood may hold, provided the covariate distribution is allowed to remain unrestricted (Roeder *and others*, 1996). For haplotype-based genetic analysis, however, complete nonparametric treatment of the covariates, including haplotypes, may not be possible due to intrinsic identifiability issues for the phase-ambiguous genotype data (Epstein and Satten, 2003). Thus, in this setting, the proper retrospective analysis of case–control data requires special attention.

An attractive feature of the retrospective likelihood is that it can enhance efficiency of case–control analysis by directly incorporating certain type of covariate distributional constraints that are natural for genetic epidemiologic studies. The assumptions of Hardy–Weinberg equilibrium (HWE) and gene–environment independence are 2 prime examples of such constraints. The HWE model, which specifies simple relationships between “allele” and “genotype” frequencies at a given chromosomal locus or between haplotype and diplotype (pair of haplotypes on homologous chromosomes) frequencies across multiple loci, is a natural law for a random mating large stable population. Often, it is also natural to assume that a subject’s genetic susceptibility, a factor which is determined at birth, is independent of his/her subsequent environmental exposures. However, if these assumptions are violated in some situations, then retrospective methods can produce serious bias in odds ratio estimates (see, e.g. Satten and Epstein, 2004; Chatterjee and Carroll, 2005; Spinka *and others*, 2005). Thus, there is a need for alternative flexible models for specifying the joint distribution of genetic and environmental covariates that could be used to assess the sensitivity of the retrospective methods to underlying assumptions as well as to develop alternative robust methods.

Both Satten and Epstein (2004) and Lin and Zeng (2006) have described retrospective maximum likelihood analysis of case–control data under flexible population genetics models that can relax the HWE assumption. Moreover, Lin and Zeng considered a model that allows the conditional distribution of environmental exposure given unphased genotypes to remain completely nonparametric, but they assumed conditional independence between haplotypes and the environmental factors given the unphased genotypes. If, however, haplotypes are the underlying biologic units through which a mechanism of gene is determined, then it is more natural to allow for direct association between haplotypes and environmental exposures. Moreover, if such association could exist, then quantifying the association between haplotypes and certain type of environmental exposures, such as lifestyle and behavioral factors, would be of scientific interest.

In this article, we propose methods for retrospective analysis of case–control data using a novel model for the gene–environment distribution that can account for direct association between haplotypes and environmental exposures. The model is developed in Section 2. We assume a standard logistic regression model to specify the disease risk conditional on diplotypes and environmental exposures. In addition, we assume a polytomous logistic regression model for specifying the population distribution of the diplotypes conditional on the environmental exposures, with the intercept parameters of the model specified in such a way that the “marginal” distribution of the diplotypes can follow certain population genetic constraints such as HWE. Moreover, by exploiting the equivalence of prospective and retrospective odds ratios under the polytomous regression model, we further incorporate certain constraints on the diplotype–exposure odds ratio parameters that could reflect specific “mode of effects” for the haplotypes. We allow the marginal distribution of the environmental exposure to remain completely nonparametric.

Under the proposed modeling framework, we then describe in Section 3 a “semiparametric” estimating equation method for inference about the finite-dimensional parameters of interest, namely the disease odds ratios, haplotype frequencies, and haplotype–exposure odds ratios. We develop a suitable expectation–maximization (EM) algorithm to account for the phase–ambiguity problem. We study asymptotic theory of the proposed estimator under the underlying semiparametric setting.

In Section 4, we assess the finite-sample performance of the proposed estimator based on case–control data that were simulated utilizing haplotype patterns and frequencies obtained from a real study. In Section 5, we apply the proposed methodology to a case–control study of colorectal adenoma to investigate whether certain haplotypes in the smoking metabolism gene, NAT2, could modify smoking-related risk of colorectal adenoma and whether the same haplotypes could influence an individual’s susceptibility to smoking as well. Section 6 contains concluding remarks. All technical details are in an appendix. A SAS macro is available from the Web site <http://www.stat.sinica.edu.tw/yhchen/download.htm>.

2. NOTATIONS AND PROPOSED MODEL

For haplotype-based studies, the underlying genetic covariate for a subject is defined by “diplotypes,” that is, the 2 haplotypes the individual carries in his/her pair of homologous chromosomes, where each haplotype is the combination of alleles at the loci of interest along an individual chromosome. Following the notation developed in Spinka *and others* (2005), let the diplotype status for a subject be $H^{\text{di}} = (H_1, H_2)$, where H_1 and H_2 denote the constituent haplotypes. We assume that there are J possible haplotypes indexed by h_j for $j = 1, \dots, J$. The diplotypes are then indexed by $h_{j_1, j_2}^{\text{di}} = (h_{j_1}, h_{j_2})$, $j_1 = 1, \dots, J_1$, $j_2 = 1, \dots, J_2$. The diplotype data, however, is not directly observable. Instead, for each subject, the multilocus genotype data G is observed, which contains information on the pair of alleles the individual carries at each individual locus but does not provide the phase information, that is which combination of alleles appears along each of the individual chromosomes. Thus, the same genotype data G could be consistent with multiple diplotypes. We will denote $\mathcal{C}(G)$ to be the set of all possible diplotypes that are consistent with the genotype data G .

Given the diplotype data H^{di} and a set of environmental covariate X , we assume that the risk of the disease is given by the logistic regression model

$$\text{logit}\{\text{pr}(D = 1|H^{\text{di}}, X)\} = \beta_0 + m(H^{\text{di}}, X, \beta_1), \quad (2.1)$$

for some known function $m(\cdot, \beta_1)$. Often one further imposes structural assumptions on the odds ratio parameters β_1 by modeling the effect of the diplotypes through constituent haplotypes according to a “dominant,” “additive,” or “recessive” mode of effect (Wallenstein *and others*, 1998). For example, a logistic regression model which assumes an additive effect for each copy of a haplotype corresponds to

$$m\{H^{\text{di}} = (h_{j_1}, h_{j_2}), X, \beta_1\} = \beta_X X + \beta_{h_{j_1}} + \beta_{h_{j_2}} + \beta_{h_{j_1}:X} X + \beta_{h_{j_2}:X} X, \quad (2.2)$$

where β_X is the main effect of X , $\beta_{h_{j_k}}$ is the main effect of haplotype h_{j_k} , $k = 1, 2$, and $\beta_{h_{j_k}:X}$ is the interaction effect of X with haplotype h_{j_k} , $k = 1, 2$. Such modeling may be necessary due to identifiability considerations (Epstein and Satten, 2003) and is desirable when the effects of the haplotypes themselves are of direct scientific interest.

Unlike Spinka *and others* (2005), who assumed independence of H^{di} and X , we assume a general polytomous logistic regression for the conditional distribution of H^{di} given X :

$$\text{log} \left[\frac{\text{pr}\{H^{\text{di}} = (h_{j_1}, h_{j_2}) | X\}}{\text{pr}\{H^{\text{di}} = (h_{j_1}^*, h_{j_2}^*) | X\}} \right] = \gamma_{0j_1j_2} + \gamma_{1j_1j_2} X, \quad (2.3)$$

where $h_0^{\text{di}} = (h_{j_1}^*, h_{j_2}^*)$ is a chosen reference diplotype. Observe that model (2.3) allows association between H^{di} and X through the regression parameters $\gamma_{1j_1j_2}$. Let γ_0 and γ_1 denote the vectorized forms for the parameters $\gamma_{0j_1j_2}$ and $\gamma_{1j_1j_2}$. Let $q_{\text{hap}}(h^{\text{di}}|x, \gamma_0, \gamma_1)$ denote $\text{pr}(H^{\text{di}} = h^{\text{di}}|X = x)$ as defined by model (2.3). We allow the marginal distribution of X , denoted by $F(x)$, to remain completely unspecified. If H^{di} were directly observable, then, in principle, no further assumptions are necessary, and one can estimate γ_0 and γ_1 together with the odds ratio parameters of the disease risk using the profile likelihood approach developed by Chatterjee and Carroll (2005). In the presence of phase ambiguity, however, the diplotypes being not directly observable, further constraints on the parameters γ_0 and γ_1 are needed for the purpose of identifiability. In the following, we show how certain natural genetic models can be used to impose these constraints.

Given that genetic susceptibility may influence environmental exposures and not vice versa, for causal interpretation of parameters it is more natural to consider a model for the environmental exposures given the diplotypes. However, the odds ratios associated with the distributions $[X|H]$ and $[H|X]$ being the same, the parameters in γ_1 can be interpreted as measures of “diplotype effects” on the distribution of exposure. Thus, it is natural to specify the γ_1 parameters according to certain mode of effects of the underlying haplotypes. For example, assuming an additive effect for the haplotypes, one can write $\gamma_{1j_1j_2} = \gamma_{1,j_1} + \gamma_{1,j_2}$, which allows the diplotype effects to be determined by a reduced set of “haplotype effect” parameters $\gamma_{1,j}$; in this case, γ_1 would denote the vectorized form for the parameters $\gamma_{1,j}$. Similarly, other commonly used models, such as dominant or recessive models, could be used to impose natural constraints on the γ_1 parameters in model (2.3). We also observe that the parametric model (2.3), combined with the non-parametric distribution $F(x)$, imposes a semiparametric model on the distribution of $[X|H]$ with a density

$$\text{pr}(x|h) = \frac{q_{\text{hap}}(h^{\text{di}}, x, \gamma_0, \gamma_1) dF(x)}{\int q_{\text{hap}}(h^{\text{di}}, v, \gamma_0, \gamma_1) dF(v)}.$$

This class of semiparametric models includes the parametric submodel where $X|H^{\text{di}} = h^{\text{di}}$ follows a multivariate normal distribution with mean $\mu_{h^{\text{di}}}$ and common variance–covariance matrix Σ . In this case, it is

easy to see that $\gamma_{1h^{\text{di}}} = (\mu_{h^{\text{di}}} - \mu_{h_0^{\text{di}}})^T \Sigma^{-1}$, which is a measure of the shift in the mean of the distribution of X due to differences in the diplotypes.

The parameter γ_0 in model (2.3) defines the population diplotype frequencies for a baseline value of the exposure X . It is common to use population genetics models, such as HWE, to specify a relationship between diplotype and haplotype frequencies. However, observe that if the diplotypes can influence certain environmental exposures, then the frequencies of the diplotypes within exposure categories may not follow the HWE constraints although the underlying population, as a whole, may be in HWE. Thus, the population-level marginal haplotype-pair distribution is assumed to follow HWE and is characterized by the parameters $\theta = (\theta_2, \dots, \theta_J)$ so that

$$\log \left[\frac{\text{pr}\{H^{\text{di}} = h^{\text{di}} = (h_{j_1}, h_{j_2})\}}{\text{pr}\{H^{\text{di}} = (h_1, h_1)\}} \right] = \theta_{j_1} + \theta_{j_2} \equiv \theta_{h^{\text{di}}}, \quad (2.4)$$

where h_1 denotes the chosen reference haplotype and $\theta_1 = 0$. Let

$$q_{\text{HWE}}(h^{\text{di}}, \theta) = \text{pr}(H^{\text{di}} = h^{\text{di}}, \theta) = \frac{\exp(\theta_{h^{\text{di}}})}{1 + \sum_{h_*^{\text{di}}} \exp(\theta_{h_*^{\text{di}}})}$$

be the marginal frequency for the diplotype h^{di} . Recall that in the proposed model, γ_0 is defined as an implicit function of γ_1 , θ , and $F(x)$ through the relationship

$$q_{\text{HWE}}(h^{\text{di}}, \theta) = \int q_{\text{hap}}(H^{\text{di}} = h^{\text{di}} | X, \gamma_0, \gamma_1) dF(X). \quad (2.5)$$

Note that F is left unspecified, and hence the model proposed is semiparametric.

3. SEMIPARAMETRIC ESTIMATING EQUATION INFERENCE

3.1 Estimation with known haplotypes

In what follows, where there can be no confusion, we will write h for h^{di} .

Let $\mathcal{H}(x) = \exp(x) / \{1 + \exp(x)\}$ be the logistic distribution function. Write the risk model probability as

$$p_{\text{risk}}(H^{\text{di}}, D, X, \beta_0, \beta_1) = [\mathcal{H}\{\beta_0 + m(H^{\text{di}}, X, \beta_1)\}]^D [1 - \mathcal{H}\{\beta_0 + m(H^{\text{di}}, X, \beta_1)\}]^{1-D}.$$

Recall that $q_{\text{hap}}(h | X, \gamma_0, \gamma_1) = \text{pr}(H^{\text{di}} = h | X; \gamma_0, \gamma_1)$ is the conditional model of H^{di} given X that is specified as in (2.3).

To start with, consider the ideal case that the phase information is known so that H^{di} is observed. Since F is treated nonparametrically, assume that F is discrete and has mass δ_k at x_k , $k = 1, \dots, K$, where $\{x_1, \dots, x_K\}$ are the distinct values of X that are observed in the case–control sample. Let n_{dkh} be the number of subjects in the sample with $(D = d, X = x_k, H^{\text{di}} = h)$. Ignoring the dependence of γ_0 on F tentatively, the log-likelihood of the case–control data can then be written as

$$l = \sum_{d=0}^1 \sum_{k=1}^K \sum_h n_{dkh} \log \{ p_{\text{risk}}(h, d, x_k, \beta_0, \beta_1) q_{\text{hap}}(h | x_k, \gamma_0, \gamma_1) \delta_k \} \\ - \sum_{d=0}^1 n_d \log \left\{ \sum_{k=1}^K \sum_h p_{\text{risk}}(h, d, x_k, \beta_0, \beta_1) q_{\text{hap}}(h | x_k, \gamma_0, \gamma_1) \delta_k \right\}.$$

Maximizing l with respect to δ for fixed values of $\omega = (\beta, \gamma_0, \gamma_1)$ then leads to

$$\delta_k = \frac{\sum_{i=1}^n I(X_i = x_k)}{\sum_d \sum_h (n_d/\pi_d) p_{\text{risk}}(h, d, x_k, \beta_0, \beta_1) q_{\text{hap}}(h|x_k, \gamma_1, \gamma_0)} \quad (3.1)$$

and the profile log-likelihood

$$l(\omega, \hat{\delta}(\omega)) = \sum_{i=1}^n \mathcal{L}(D_i, H_i^{\text{di}}, X_i, \mathcal{B}, \gamma_1, \gamma_0) = \sum_{i=1}^n \log \left\{ \frac{S(D_i, H_i^{\text{di}}, X_i, \mathcal{B}, \gamma_1, \gamma_0)}{\sum_{d_*=0}^1 \sum_{h_*} S(d_*, h_*, X_i, \mathcal{B}, \gamma_1, \gamma_0)} \right\}, \quad (3.2)$$

where

$$\pi_d = \text{pr}(D = d) = \sum_{h,k} p_{\text{risk}}(h, d, x_k, \beta_0, \beta_1) q_{\text{hap}}(h|x_k, \gamma_1, \gamma_0) \delta_k$$

and

$$S(d, h, x, \mathcal{B}, \gamma_1, \gamma_0) = \exp\{d(\kappa - \beta_0)\} p_{\text{risk}}(h, d, x, \beta_0, \beta_1) q_{\text{hap}}(h|x, \gamma_0, \gamma_1),$$

with $\mathcal{B} = (\beta_0, \beta_1, \kappa)^{\text{T}}$ and $\kappa = \beta_0 + \log(n_1/n_0) - \log\{\text{pr}(D = 1)/\text{pr}(D = 0)\}$. The calculation is similar to that in Chatterjee and Carroll (2005).

As noted by Chatterjee and Carroll (2005), the parameter β_0 is separable from κ and hence is theoretically identifiable. In practice, however, there is usually little information about β_0 available in the observed data, and hence the information matrix is nearly singular. One way to bypass this problem is to use external information on the disease prevalence $\text{pr}(D = 1)$, while another way is to use the rare-disease approximation when the disease is rare. The estimation method described below can be applied to both the 2 cases of $\text{pr}(D = 1)$ being known and the rare-disease approximation being made, with suitable definitions on \mathcal{B} and $S(d, h, x, \mathcal{B}, \gamma_0, \gamma_1)$. When $\text{pr}(D = 1)$ is known, κ depends on β_0 only, hence here we define $\mathcal{B} = (\beta_0, \beta_1)^{\text{T}}$. When the disease is rare so that

$$p_{\text{risk}}(h, d, x, \beta_0, \beta_1) \approx \exp[d\{\beta_0 + m(h, x, \beta_1)\}], \quad (3.3)$$

we have

$$S(d, h, x, \mathcal{B}, \gamma_1, \gamma_0) = \exp[d\{\kappa + m(h, x, \beta_1)\}] q_{\text{hap}}(h|x, \gamma_0, \gamma_1).$$

Note that β_0 does not appear in this expression, and hence we define $\mathcal{B} = (\kappa, \beta_1)^{\text{T}}$.

Our goal is to estimate the parameters $(\mathcal{B}, \theta, \gamma_1)$ based on the profile log-likelihood (3.2), where γ_0 is defined as an implicit function of (θ, γ_1, F) through (2.5), and we write $\gamma_0 = \mathcal{G}(\theta, \gamma_1, F)$. Let $\Omega = (\mathcal{B}, \gamma_1, \gamma_0)$, $\Omega^* = \{\mathcal{B}, \gamma_1, \mathcal{G}(\theta, \gamma_1, F)\}$, and $\Phi = (\mathcal{B}, \gamma_1, \theta)$. Let $\mathcal{L}_{\Omega}(\cdot)$ and $\mathcal{L}_{\Phi}(\cdot)$ be, respectively, the derivatives of $\mathcal{L}(\cdot)$ with respect to Ω and Φ , and \mathcal{G}_{θ} and \mathcal{G}_{γ_1} the derivatives of $\mathcal{G}(\cdot)$ with respect to θ and γ_1 . We then have

$$\mathcal{L}_{\Phi}(\cdot) = \frac{\partial \Omega^*}{\partial \Phi} \mathcal{L}_{\Omega}(\cdot) = \{\mathcal{L}_{\mathcal{B}}^{\text{T}}(\cdot), (\mathcal{L}_{\gamma_1} + \mathcal{G}_{\gamma_1} \mathcal{L}_{\gamma_0})^{\text{T}}(\cdot), (\mathcal{G}_{\theta} \mathcal{L}_{\gamma_0})^{\text{T}}(\cdot)\}^{\text{T}}.$$

Explicit expressions for \mathcal{G}_{γ_1} and \mathcal{G}_{θ} are given in Appendix C. Also, the information matrix is given by

$$\mathcal{I} = \frac{\partial \Omega^*}{\partial \Phi} \mathcal{I}_{\Omega\Omega} \left(\frac{\partial \Omega^*}{\partial \Phi} \right)^{\text{T}},$$

where $\mathcal{I}_{\Omega\Omega} = E(-\mathcal{L}_{\Omega\Omega})$, with $\mathcal{L}_{\Omega\Omega}$ the second derivative of \mathcal{L} with respect to Ω ; note that the terms involving second derivatives of Ω^* do not appear in the information matrix because $E(\mathcal{L}_{\Omega}) = 0$, which is

a direct consequence of the Lemmas A.1 and A.2 in Appendix A. We propose to obtain the estimate of Φ by solving the estimating equation

$$\begin{aligned} 0 &= n^{-1/2} \sum_{i=1}^n \mathcal{L}_\Phi\{D_i, H_i^{\text{di}}, X_i, \mathcal{B}, \gamma_1, \mathcal{G}(\theta, \gamma_1, \widehat{F})\} \\ &\equiv n^{-1/2} \sum_{i=1}^n \mathcal{L}_\Phi(D_i, H_i^{\text{di}}, X_i, \Phi), \end{aligned} \quad (3.4)$$

where we have substituted an estimate \widehat{F} for F in $\mathcal{G}(\cdot)$; that is, for each fixed value of (θ, γ_1) , we solve γ_0 from

$$q_{\text{HWE}}(h, \theta) = \int q_{\text{hap}}(h|x, \gamma_0, \gamma_1) d\widehat{F}(x). \quad (3.5)$$

One convenient choice of \widehat{F} is the empirical estimate \widehat{F}_{emp} , which is given by

$$\widehat{F}_{\text{emp}}(x) = \widehat{F}_{\text{emp},1}(x)\text{pr}(D = 1) + \widehat{F}_{\text{emp},0}(x)\text{pr}(D = 0)$$

for the case where $\text{pr}(D = 1)$ is known, where $\widehat{F}_{\text{emp},1}(x)$ and $\widehat{F}_{\text{emp},0}(x)$ are the empirical distributions of X in the case and control samples, and is given by $\widehat{F}_{\text{emp}}(x) = \widehat{F}_{\text{emp},0}(x)$ for the case where the rare-disease assumption can be made. An alternative choice of $\widehat{F}(x)$ would be the profile likelihood estimate (3.1). Numerical calculations not given here show that the latter choice requires more computational efforts while yielding results very similar to those given by the empirical estimate \widehat{F}_{emp} .

3.2 Estimation with ambiguous haplotype data

Now, we turn to the more practical case where the haplotype data cannot be observed directly and must be inferred from the unphased genotype data, that is, the haplotype information may be subject to ambiguity. In this case, we apply an EM-like algorithm to the “complete data” estimating equation (3.4). Let G_i denote the observed unphased genotype of subject i and $\mathcal{C}(G_i)$ the set of diplotypes that are consistent with G_i . When only G_i instead of H_i^{di} is observed for each subject, we propose to obtain the estimate $\widehat{\Phi}$ for $\Phi = (\mathcal{B}, \gamma_1, \theta)$ as the solution of the weighted version of (3.4):

$$\begin{aligned} 0 &= n^{-1/2} \sum_{i=1}^n \sum_h \widehat{w}_i\{h, \mathcal{B}, \gamma_1, \mathcal{G}(\theta, \gamma_1, \widehat{F}_{\text{emp}})\} \mathcal{L}_\Phi\{D_i, h, X_i, \mathcal{B}, \gamma_1, \mathcal{G}(\theta, \gamma_1, \widehat{F}_{\text{emp}})\} \\ &\equiv n^{-1/2} \sum_{i=1}^n \bar{\mathcal{L}}_\Phi\{D_i, G_i, X_i, \Phi, \mathcal{G}(\theta, \gamma_1, \widehat{F}_{\text{emp}})\}, \end{aligned} \quad (3.6)$$

where using the short-hand notation that $\widehat{\gamma}_0 = \mathcal{G}(\theta, \gamma_1, \widehat{F}_{\text{emp}})$, the weights are given by

$$\begin{aligned} \widehat{w}_i\{h, \mathcal{B}, \gamma_1, \mathcal{G}(\theta, \gamma_1, \widehat{F}_{\text{emp}})\} &= \text{pr}(H^{\text{di}} = h | D_i, X_i, G_i, \mathcal{B}, \gamma_1, \widehat{\gamma}_0) \\ &= I\{h \in \mathcal{C}(G_i)\} \frac{p_{\text{risk}}(h, D_i, X_i, \mathcal{B}) q_{\text{hap}}(h | X_i, \gamma_1, \widehat{\gamma}_0)}{\sum_{h_* \in \mathcal{C}(G_i)} p_{\text{risk}}(h_*, D_i, X_i, \mathcal{B}) q_{\text{hap}}(h_* | X_i, \gamma_1, \widehat{\gamma}_0)}. \end{aligned} \quad (3.7)$$

The limiting version of the weights is given as

$$w(h, \Omega) = \frac{I\{h \in \mathcal{C}(G)\} S(D, h, X, \mathcal{B}, \gamma_1, \gamma_0)}{\sum_{h_* \in \mathcal{C}(G)} S(D, h_*, X, \mathcal{B}, \gamma_1, \gamma_0)} = \text{pr}(H^{\text{di}} = h | D, G, X, \Omega). \quad (3.8)$$

Solving the estimating equation (3.6) can be implemented simply by an EM-like algorithm as follows: starting with an initial value for Φ and hence an initial value for γ_0 , we

- i) calculate the weights $\{\widehat{w}_i\}$ from (3.7) and
- ii) solve (3.6) to obtain an updated estimate of Φ using the weights $\{\widehat{w}_i\}$ given in (i); note that within this step we also need to solve (3.5) to obtain updated value of γ_0 .

The algorithm is iterated between the 2 steps until convergence. Note that the weights $\{\widehat{w}_i\}$ are only used in solving Φ from (3.6) and are not required in solving γ_0 from (3.5).

3.3 Asymptotic theory

Make the following series of definitions. Expectations denoted as $E_{\text{cc}}(\cdot)$ are taken under the case-control sampling design, that is, for any random vector Y , $E_{\text{cc}}(Y) = \sum_{d=0}^1 \mu_d E(Y|D = d)$, where $\mu_d = \text{plim } n_d/n$, $d = 0, 1$. Then, define

$$\begin{aligned} \bar{\mathcal{L}} &= E_{\text{cc}} \left\{ -\frac{\partial}{\partial \Phi^T} \bar{\mathcal{L}}_{\Phi}(D, G, X, \Phi) \right\} = \frac{\partial \Omega^*}{\partial \Phi} \bar{\mathcal{L}}_{\Omega\Omega} \left(\frac{\partial \Omega^*}{\partial \Phi} \right)^T, \\ \bar{\mathcal{L}}_{\Omega}(D, G, X, \Omega) &= \sum_h w(h, \Omega) \mathcal{L}_{\Omega}(D, h, X, \Omega), \\ \bar{\mathcal{L}}_{\Omega\Omega} &= E_{\text{cc}}\{-\partial \bar{\mathcal{L}}_{\Omega}(\cdot)/\partial \Omega^T\} = E_{\text{cc}}\{\bar{\mathcal{L}}_{\Omega}(D, G, X, \Omega) \bar{\mathcal{L}}_{\Omega}^T(D, G, X, \Omega)\}. \end{aligned} \quad (3.9)$$

Note that the second derivative of Ω^* does not appear in $\bar{\mathcal{L}}$ since $E(\bar{\mathcal{L}}_{\Phi}) = E(\mathcal{L}_{\Phi}) = 0$, and the last identity in (3.9) is given by Lemma A.3 in Appendix A.

Define $\widehat{p}_{\text{emp}}(D_i)$ to be the mass of $\widehat{F}_{\text{emp}}(X_i)$, which is equal to $\sum_{d=0}^1 \pi_d I(D_i = d)/n_d$ if $\pi_d = \text{pr}(D = d)$ is known and is equal to $I(D_i = 0)/n_0$ when the rare-disease approximation is used. Let $\mathbf{q}_{\text{hap}}(X, \gamma_1, \gamma_0) = \{q_{\text{hap}}(h|X, \gamma_1, \gamma_0)\}$ be the vector collection over h of $q_{\text{hap}}(h|X, \gamma_1, \gamma_0)$ for all diplotypes except the reference diplotype, and let $\mathbf{q}_{\text{HWE}}(\theta)$ be defined similarly. Define

$$\begin{aligned} Q_n &= \sum_{i=1}^n \frac{\partial}{\partial \gamma_0} \mathbf{q}_{\text{hap}}(X_i, \gamma_0, \gamma_1) \widehat{p}_{\text{emp}}(D_i), \\ Q &= \int \frac{\partial}{\partial \gamma_0} \mathbf{q}_{\text{hap}}(x, \gamma_0, \gamma_1) dF(x) = E \left\{ \frac{\partial}{\partial \gamma_0} \mathbf{q}_{\text{hap}}(X, \gamma_0, \gamma_1) \right\}, \\ k(X, D) &= nQ^{-1} \{ \mathbf{q}_{\text{hap}}(X, \gamma_1, \gamma_0) - \mathbf{q}_{\text{HWE}}(\theta) \} \widehat{p}_{\text{emp}}(D_i), \\ \bar{\mathcal{L}}_{\Phi\gamma_0} &= (\partial \Omega^* / \partial \Phi) \bar{\mathcal{L}}_{\Omega\gamma_0}, \\ \bar{\mathcal{L}}_{\Omega\gamma_0} &= E(-\partial \bar{\mathcal{L}}_{\Omega} / \partial \gamma_0^T), \\ \mathcal{K}(X, D) &= \bar{\mathcal{L}}_{\Phi\gamma_0} k(X, D), \end{aligned}$$

where $\bar{\mathcal{L}}_{\Omega\gamma_0}$ is the obvious submatrix of $\bar{\mathcal{L}}_{\Omega\Omega}$.

THEOREM 3.1 Let

$$\bar{\mathcal{E}}(D, G, X, \Phi) = \bar{\mathcal{L}}_{\Phi}(D, G, X, \Phi) - E\{\bar{\mathcal{L}}_{\Phi}(D, G, X, \Phi)|D\} - \{\mathcal{K}(X, D) - E(\mathcal{K}(X, D)|D)\}.$$

Suppose that $E_{cc}\{\bar{\mathcal{E}}(\cdot)\bar{\mathcal{E}}^T(\cdot)\}$ exists and the matrix $\bar{\mathcal{I}}$ is invertible. Then, $n^{1/2}(\hat{\Phi} - \Phi)$ is asymptotically normal with mean zero and covariance matrix

$$\bar{\Gamma} = \bar{\mathcal{I}}^{-1} E_{cc}(\bar{\mathcal{E}} \bar{\mathcal{E}}^T) \bar{\mathcal{I}}^{-1}. \quad (3.10)$$

REMARK 3.2 The asymptotic variance $\bar{\Gamma}$ can be readily estimated by replacing each component matrix with its empirical counterpart. Lemma A.3 gives useful expressions to facilitate this computation.

REMARK 3.3 In our numerical experiments, the estimated covariance based on formula (3.10) is very close to that based on the “naive” covariance estimate obtained by naively treating the estimating equation (3.6) as a genuine score equation; namely, treating the \hat{F}_{emp} plugged into $\mathcal{G}(\cdot)$ as the true covariate distribution F . In this case, by applying Proposition 1(ii) in Chatterjee and Carroll (2005), the naive covariance estimate can be obtained simply as the empirical counterpart of the matrix $\bar{\mathcal{I}}^{-1} - \bar{\mathcal{I}}^{-1} \Psi \bar{\mathcal{I}}^{-1}$, where $\Psi = \sum_{d=0}^1 (n_d/n) [E\{\bar{\mathcal{L}}_{\Phi}(D, G, X, \Phi) | D = d\}]^{\otimes 2}$. Whether this naive estimate performs well in general is unknown, and we suggest using the estimate based on (3.10).

4. SIMULATIONS

4.1 Finite-sample performance under correct model

In this section, we study the finite-sample performance of the proposed estimator using simulated data generated under the proposed modeling framework. We simulated haplotypes following published data (Epstein and Satten, 2003) on haplotype patterns and frequencies for 5 single-nucleotide polymorphisms (SNPs) in a putative susceptibility gene for diabetes (see Table 1). The simulations involved a single environmental covariate X , assumed to follow a standard normal distribution in the population. Given X , the diplotypes (haplotype pair) for an individual were generated from a polytomous logistic regression of the form (2.3), where the diplotype-specific odds ratios were further specified according to an additive model of the form $\gamma_{1j_1j_2} = \gamma_{1,j_1} + \gamma_{1,j_2}$, where j_1 and j_2 denote the index for 14 haplotypes shown in Table 1. We assume $\gamma_{1,4} = \gamma_{1,5} = -0.4$ and $\gamma_{1,12} = 0.4$, and all the other $\gamma_{1,j} = 0$. The parameters

Table 1. Marginal haplotype frequencies and grouped haplotypes in the simulation. Here, j is index for the original haplotypes while j' is index for the grouped haplotypes

j	Haplotype	Frequency	j'
1	10011	0.3534	1
2	00100	0.0055	1
3	00110	0.0048	1
4	01011	0.1232	2
5	01100	0.2504	3
6	01101	0.0062	1
7	01111	0.0059	1
8	10000	0.0106	4
9	00011	0.0052	1
10	10100	0.0510	5
11	10110	0.0317	6
12	11011	0.1351	7
13	11100	0.0110	8
14	11111	0.0060	1

$\gamma_{0j_1j_2}$'s in model (2.3) are then specified in such a way that the marginal diplotype distribution follow HWE with haplotype frequencies given in Table 1.

For generating disease outcome, we chose the haplotype "01100" ($j = 5$) to be causal and used the logistic model

$$\text{logit}\{\text{pr}(D = 1|H^{\text{di}}, X)\} = \beta_0 + \beta_H Z(5) + \beta_X X + \beta_{HX} Z(5)X,$$

where $Z(5)$ denotes the number of the copies of the causal haplotype contained in H^{di} . The true value of the parameter vector $(\beta_0, \beta_H, \beta_X, \beta_{HX})$ was set to $(-3.0, 0.2, 0.1, 0.3)$. A case-control sample with 600 controls and 600 cases was then sampled. The results were based upon 1000 simulated data sets.

When analyzing the data, we only used the unphased genotype information. We did not assume the causal haplotype to be known. Thus, in both the disease-risk model (2.1) and the diplotype-frequency model (2.3), we choose the most common haplotype "10011" as a reference and estimated a separate regression parameter for each of the non-referent haplotypes. Since rare haplotypes may lead to unreliable estimates of the associated regression parameters, when estimating β and γ_1 , rare haplotypes with frequency $<1\%$ are grouped into the reference haplotype. The resulting 8-grouped haplotypes are labeled as $h_{j'}$, $j' = 2, \dots, 8$; see Table 1 for details about how the haplotypes are grouped.

In each simulation, we obtain 2 sets of estimates from the proposed method, one using the rare-disease approximation (3.3) and the other using the known value of the population disease prevalence. Results shown in Table 2 show that both sets of estimates are essentially unbiased. Also, the standard error estimates are in close agreement with the true values, and the coverage probabilities are close to the nominal value (95%). As expected, the estimates for θ and γ_1 are generally more efficient using external information on the disease prevalence than when using the rare-disease approximation, but no such efficiency gain is observed for the parameters β in the disease-risk model. Similar conclusion can be drawn from the simulations with a Bernoulli covariate (success probability = 0.5), showing the applicability of the proposed method to the categorical covariate. Detailed results for this latter set of simulations are included in the supplementary material available at *Biostatistics* online.

4.2 Model robustness

Here, we consider a simulation study where we generate the data in such a way that the polytomous model for diplotype frequencies may not exactly hold. The main goal is to give an indication of the robustness of the estimate of the association parameters (β) from the proposed method when the model for $[H^{\text{di}}|X]$ is misspecified.

Following the argument of causality in Section 2, if $[X|H^{\text{di}}]$ follows a normal distribution with constant variance, then the polytomous model is exact. So, to show a modest violation of the polytomous model, for given diplotype we generate the data on X from

$$\left[X|H^{\text{di}} = (h_{j_1}, h_{j_2}) \right] \sim \lambda_{j_1} + \lambda_{j_2} + \epsilon,$$

where the diplotype data are again generated from the distribution in Table 1, ϵ is a t -distribution (d.f. = 3) truncated at ± 5 , $\lambda_4 = \lambda_5 = -1.2$, $\lambda_{12} = 1.2$, and all the other λ_j are 0. The disease status data are generated from the same logistic model as in the previous simulation. The simulated data on 600 cases and 600 controls are then analyzed with the proposed method, where the analysis models for the disease risk, $[H^{\text{di}}|X]$, and the marginal diplotype distribution are specified the same as those in the previous simulation. As a comparison, we also fit to the simulated data a model with the haplotype-environment (H-X) independence assumption, i.e. $\text{pr}(H^{\text{di}}|X) = \text{pr}(H^{\text{di}}) = q_{\text{HWE}}(H^{\text{di}})$, using the method proposed by Spinka *and others* (2005). The rare-disease approximation is made when applying both the 2 methods.

Table 2. Simulation results for the case with an additive genetic law. Here, mean is the mean over 1000 simulated data sets, SE is the standard deviation of the estimates, \widehat{SE} is the mean of the estimated standard deviation of the parameter estimates, CP is the coverage probability of the 95% confidence interval, β 's denote risk parameters, θ 's characterize marginal haplotype frequencies, and γ 's denote the haplotype–environment association parameters

Parameter	pr($D = 1$) known				Rare-disease approximation			
	Mean	SE	\widehat{SE}	CP	Mean	SE	\widehat{SE}	CP
$\beta_{h_4} = 0$	-0.011	0.150	0.151	0.946	0.012	0.152	0.150	0.941
$\beta_{h_5} = 0.2$	0.199	0.105	0.110	0.960	0.197	0.108	0.110	0.949
$\beta_{h_8} = 0$	0.000	0.450	0.461	0.977	-0.025	0.436	0.461	0.977
$\beta_{h_{10}} = 0$	0.003	0.224	0.232	0.965	0.006	0.225	0.230	0.963
$\beta_{h_{11}} = 0$	-0.002	0.248	0.258	0.969	0.000	0.232	0.256	0.982
$\beta_{h_{12}} = 0$	0.000	0.155	0.150	0.938	0.009	0.150	0.149	0.953
$\beta_{h_{13}} = 0$	-0.024	0.736	0.737	0.977	-0.013	0.619	0.665	0.993
$\beta_X = 0.1$	0.098	0.126	0.129	0.965	0.105	0.129	0.128	0.957
$\beta_{h_4:X} = 0$	0.000	0.149	0.151	0.958	0.002	0.156	0.148	0.947
$\beta_{h_5:X} = 0.3$	0.310	0.108	0.111	0.954	0.296	0.104	0.109	0.965
$\beta_{h_8:X} = 0$	0.012	0.474	0.518	0.981	0.020	0.443	0.502	0.953
$\beta_{h_{10}:X} = 0$	-0.015	0.244	0.241	0.952	0.004	0.245	0.237	0.937
$\beta_{h_{11}:X} = 0$	0.020	0.255	0.268	0.962	-0.034	0.246	0.263	0.953
$\beta_{h_{12}:X} = 0$	-0.003	0.142	0.139	0.958	0.000	0.138	0.137	0.953
$\beta_{h_{13}:X} = 0$	0.030	0.918	0.853	0.975	-0.016	0.577	0.709	0.955
$\theta_1 = -4.16$	-4.226	0.362	0.361	0.975	-4.219	0.386	0.365	0.951
$\theta_2 = -4.30$	-4.349	0.347	0.348	0.962	-4.376	0.377	0.356	0.957
$\theta_3 = -1.05$	-1.057	0.093	0.095	0.956	-1.060	0.105	0.100	0.939
$\theta_4 = -0.34$	-0.353	0.066	0.071	0.960	-0.357	0.075	0.075	0.949
$\theta_5 = -4.04$	-4.083	0.297	0.277	0.948	-4.080	0.286	0.277	0.967
$\theta_6 = -4.09$	-4.140	0.307	0.300	0.965	-4.136	0.317	0.302	0.961
$\theta_7 = -3.51$	-3.538	0.291	0.281	0.958	-3.535	0.293	0.296	0.959
$\theta_8 = -4.22$	-4.294	0.397	0.381	0.969	-4.296	0.406	0.382	0.973
$\theta_9 = -1.94$	-1.950	0.147	0.150	0.952	-1.943	0.159	0.157	0.955
$\theta_{10} = -2.41$	-2.421	0.158	0.166	0.960	-2.417	0.160	0.174	0.971
$\theta_{11} = -0.96$	-0.961	0.091	0.090	0.946	-0.966	0.101	0.095	0.941
$\theta_{12} = -3.47$	-3.520	0.380	0.360	0.960	-3.528	0.389	0.385	0.967
$\theta_{13} = -4.08$	-4.109	0.288	0.292	0.967	-4.111	0.290	0.293	0.961
$\gamma_{1,4} = -0.4$	-0.402	0.100	0.102	0.954	-0.407	0.113	0.106	0.923
$\gamma_{1,5} = -0.4$	-0.404	0.072	0.075	0.946	-0.421	0.079	0.080	0.945
$\gamma_{1,8} = 0$	0.010	0.301	0.325	0.952	-0.002	0.317	0.341	0.955
$\gamma_{1,10} = 0$	0.008	0.160	0.159	0.931	-0.008	0.172	0.168	0.957
$\gamma_{1,11} = 0$	-0.011	0.186	0.177	0.938	0.006	0.180	0.186	0.951
$\gamma_{1,12} = 0.4$	0.407	0.097	0.097	0.940	0.408	0.103	0.101	0.951
$\gamma_{1,13} = 0$	-0.004	0.389	0.418	0.929	0.021	0.435	0.461	0.947

Table 3. Simulation results for the case of a misspecified conditional diplotype distribution given covariates. Here, mean is the mean over 1000 simulated data sets, SE is the standard deviation of the estimates, \widehat{SE} is the mean of the estimated standard deviation of the parameter estimates, CP is the coverage probability of the 95% confidence interval, β 's denote risk parameters, θ 's characterize marginal haplotype frequencies, and γ 's denote the haplotype–environment association parameters

Parameter	Proposed				H-X independence			
	Mean	SE	\widehat{SE}	CP	Mean	SE	\widehat{SE}	CP
$\beta_{h_4} = 0$	0.006	0.165	0.178	0.967	-0.339	0.143	0.163	0.421
$\beta_{h_5} = 0.2$	0.196	0.118	0.126	0.962	-0.217	0.103	0.115	0.033
$\beta_{h_8} = 0$	-0.037	0.452	0.455	0.962	-0.065	0.436	0.434	0.957
$\beta_{h_{10}} = 0$	0.027	0.220	0.224	0.960	0.018	0.219	0.222	0.960
$\beta_{h_{11}} = 0$	-0.014	0.248	0.250	0.955	-0.013	0.237	0.247	0.955
$\beta_{h_{12}} = 0$	0.026	0.133	0.151	0.977	-0.003	0.122	0.149	0.977
$\beta_{h_{13}} = 0$	-0.029	0.629	0.626	0.970	-0.103	0.547	0.613	0.985
$\beta_X = 0.1$	0.124	0.076	0.074	0.938	0.260	0.050	0.061	0.226
$\beta_{h_4:X} = 0$	-0.012	0.082	0.082	0.943	-0.345	0.053	0.064	0.000
$\beta_{h_5:X} = 0.3$	0.328	0.069	0.064	0.913	-0.107	0.042	0.047	0.000
$\beta_{h_8:X} = 0$	-0.031	0.244	0.280	0.965	-0.017	0.144	0.212	0.977
$\beta_{h_{10}:X} = 0$	-0.006	0.132	0.130	0.940	-0.004	0.070	0.098	0.987
$\beta_{h_{11}:X} = 0$	-0.001	0.136	0.142	0.960	-0.001	0.083	0.109	0.977
$\beta_{h_{12}:X} = 0$	-0.038	0.064	0.070	0.930	0.317	0.041	0.054	0.000
$\beta_{h_{13}:X} = 0$	0.006	0.331	0.351	0.923	0.044	0.168	0.294	0.977
$\theta_1 = -4.16$	-4.185	0.406	0.364	0.955	-4.218	0.384	0.360	0.967
$\theta_2 = -4.30$	-4.313	0.367	0.343	0.957	-4.364	0.383	0.348	0.957
$\theta_3 = -1.05$	-1.053	0.0960	0.097	0.957	-1.041	0.098	0.103	0.965
$\theta_4 = -0.34$	-0.340	0.075	0.071	0.938	-0.331	0.072	0.079	0.965
$\theta_5 = -4.04$	-4.066	0.279	0.271	0.950	-4.079	0.286	0.273	0.945
$\theta_6 = -4.09$	-4.144	0.296	0.294	0.957	-4.133	0.287	0.296	0.965
$\theta_7 = -3.51$	-3.563	0.312	0.304	0.957	-3.546	0.310	0.298	0.970
$\theta_8 = -4.22$	-4.228	0.390	0.360	0.948	-4.290	0.403	0.376	0.970
$\theta_9 = -1.94$	-1.947	0.153	0.157	0.967	-1.947	0.158	0.158	0.967
$\theta_{10} = -2.41$	-2.411	0.172	0.174	0.955	-2.419	0.173	0.175	0.960
$\theta_{11} = -0.96$	-0.984	0.091	0.091	0.945	-0.975	0.091	0.101	0.985
$\theta_{12} = -3.47$	-3.561	0.396	0.385	0.960	-3.513	0.404	0.377	0.960
$\theta_{13} = -4.08$	-4.119	0.280	0.286	0.955	-4.115	0.282	0.289	0.967

The results shown in Table 3 reveal that, for the estimation of the association parameters β , the proposed method may be quite robust to modest misspecification of the model for $[H^{\text{di}}|X]$. On the other hand, the estimates from the H-X independence method does result in substantial bias, especially for parameters corresponding to haplotypes for which $[X|H^{\text{di}}]$ have nonzero mean; for example, the estimate for the interaction parameter between h_5 and X is severely biased with the H-X independence method. The estimates for the marginal haplotype-frequency parameters θ seem to be robust to misspecification of $[H^{\text{di}}|X]$ for both the 2 methods.

5. CASE–CONTROL STUDY OF COLORECTAL ADENOMA STUDY, NAT2 HAPLOTYPE, AND SMOKING

We illustrate the proposed modeling and estimating methodologies with an application to a case–control study of colorectal adenoma, a precursor of colorectal cancer. The study involved 628 prevalent advanced

adenoma cases and 635 gender-matched controls, selected from the screening arm of the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial at the National Cancer Institute, USA (Gohagan *and others*, 2000; Moslehi *and others*, 2006). One of the main objectives of this study is to assess whether smoking-related risk of colorectal adenoma may be modified by certain haplotypes in NAT2, a gene known to be important in metabolism of smoking-related carcinogens. In addition, since NAT2 is involved in the smoking metabolism pathway, potentially it can influence an individual’s addiction to smoking. Thus, it was also of interest to identify potential haplotypes that could influence an individual’s susceptibility to smoking.

Genotype data were available on 6 SNPs. We initially applied the EM algorithm proposed by Li *and others* (2003) for haplotype-frequency estimation to derive 7 common haplotypes with estimated frequency greater than 0.5%, which are then included in our association analysis with the most frequent haplotype served as the reference haplotype. Subjects were categorized as “never,” “former,” or “current” smokers. We fit a logistic regression model (2.1) assuming an additive effect for each haplotype other than the reference one; see (2.2). The haplotype–environment interaction terms include only those for the haplotype “101010” with “Smk1” and “Smk2,” 2 dummy variables for former and never smokers, because they are the only promising interactions according to preliminary analysis. The disease-risk model was further adjusted for “age,” recorded in years, and “gender.” A polytomous logistic regression (2.3) is specified for the conditional distribution of diplotypes given the environmental covariates Smk1 and Smk2 with the marginal diplotype distribution being specified by the HWE constraints. The main parameters of interest include the disease-haplotype odds ratio parameters β_1 , the haplotype–environment odds ratio parameters γ_1 , and the marginal haplotype frequencies in the whole population. The marginal distribution for the environmental covariates is left unspecified. For estimation of regression parameters β and γ_1 , we grouped haplotypes with frequency less than 2% into the reference haplotype. The rare-disease approximation was made in deriving the estimating equation, and the EM algorithm proposed in Section 3 is utilized to accommodate the unphased genotype data.

Results from this application are displayed in Table 4. It is clear that current smokers can have significantly elevated risk for colorectal adenoma relative to nonsmokers, adjusting for gender and age. Relative to the reference haplotype “001100,” all the other haplotypes are associated with reduced risk for colorectal adenoma, but the statistical evidence is not significant. However, the significance of the interaction $101010 \times \text{Smk2}$ suggests that smoking-related risk of adenoma was much reduced for carriers of the haplotype 101010 than non-carriers. The finding is consistent with previous laboratory and epidemiologic studies that have identified the haplotype 101010, known as “NAT2*4,” as a rapid metabolizer for smoking-related carcinogens. The estimates for the parameter γ_1 for the conditional diplotype distribution reveal that the susceptibility to smoking seems not to be influenced by any haplotypes we considered. Finally, the estimates for the marginal haplotype frequencies derived from the estimates of θ are quite close to those obtained by the EM algorithm of Li *and others* (2003) applied to the genotype data of the controls.

To check if the analysis is sensitive to model specification for the conditional distribution of diplotypes given the environmental covariates, we further fit the model (2.3) with various choices of the environmental covariates. The results (not shown) for the association parameters β and the marginal haplotype frequencies are fairly consistent across the analyses.

6. CONCLUDING REMARKS

The model we have proposed for gene–environment association is suitable when the underlying haplotypes of a genomic region may causally influence the environmental exposure(s) under study. The model, however, requires special treatment for environmental factors, such as ethnicity or geographic region(s), which may be associated with the genomic region under study, not because of any causal relationship but merely due to population stratification. Suppose, in addition to the main environmental exposure X ,

Table 4. *Results of the colorectal adenoma study. Haplotypes 001010 and 001110 are grouped into the reference haplotype in the disease-risk and conditional diplotype distribution models*

Variable	Coefficient estimate	SE	P-value
Disease-risk model			
Smk1	0.057	0.164	0.730
Smk2	1.112	0.209	<0.0001
001100 (reference)	—	—	—
100011	−0.116	0.094	0.215
101100	−0.146	0.270	0.589
110010	−0.095	0.255	0.708
101010	−0.091	0.154	0.554
101010 × Smk1	0.101	0.212	0.633
101010 × Smk2	−0.564	0.274	0.039
Female	0.002	0.127	0.985
Age	0.052	0.011	<0.0001
Conditional diplotype distribution			
Smk1			
100011	−0.084	0.104	0.421
101100	−0.020	0.298	0.945
110010	0.208	0.287	0.469
101010	−0.050	0.154	0.747
Smk2			
100011	−0.020	0.130	0.877
101100	0.321	0.349	0.358
110010	0.460	0.337	0.172
101010	0.294	0.214	0.168
Marginal haplotype frequency			
001100	0.378	0.013	—
100011	0.303	0.012	—
101100	0.027	0.0047	—
110010	0.028	0.0048	—
101010	0.239	0.012	—
001010	0.005	0.0014	—
001110	0.019	0.0028	—

there is a set of environmental factors S which could be used to divide the underlying population into K strata that are likely to be genetically heterogenous. In such a situation, a natural model for describing the association between diplotypes H^{di} and environmental factors $W = (X, S)$ is given by

$$\log \left[\frac{\text{Pr} \{ H^{\text{di}} = (h_{j_1}, h_{j_2}) \mid X, S \}}{\text{Pr} \{ H^{\text{di}} = (h_{j_1^*}, h_{j_2^*}) \mid X, S \}} \right] = \gamma_{0j_1j_2}(S) + \gamma_{1j_1j_2}X, \quad (6.1)$$

where the stratum-specific intercept parameters $\gamma_{0j_1j_2}(S)$ should be specified in such a way that the diplotype frequencies, after marginalized over X , follow population genetics constraints, such as HWE, within each stratum defined by S . The disease-risk model could be also extended to include S as a risk factor. The proposed estimating equation methodology can be easily modified to estimate the gene–environment interaction and association parameters of interest under these extended models.

ACKNOWLEDGMENTS

Chen’s research was supported by the National Science Council of the People’s Republic of China (NSC 95-2118-M-001-022-MY3). Chatterjee’s research was supported by the Intramural Research Program of the National Cancer Institute. Carroll’s research was supported by a grant from the National Cancer Institute (CA57030) and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106). A SAS macro is available from the Web site <http://www.stat.sinica.edu.tw/yhchen/download.html>. *Conflict of Interest*: None declared.

APPENDIX A: BASIC LEMMAS

The following lemmas are required to derive the asymptotic distribution of the proposed estimator. Lemma A.1 below is in fact Lemma 3 of Chatterjee and Carroll (2005).

LEMMA A.1 Under the case–control sampling design where the total sample size $n = n_1 + n_0$ tends to infinity but the sampling proportions for the cases and controls, that is, n_1/n and n_0/n , remain fixed at μ_1 and μ_0 , we have for any measurable function $M(D, H^{\text{di}}, X)$ of data (D, H^{di}, X) ,

$$\begin{aligned} E_{\text{cc}}\{M(D, H^{\text{di}}, X)\} &= \sum_{d=0}^1 \mu_d E\{M(D, H^{\text{di}}, X)|D = d\} \\ &= \int E^*\{M(D, H^{\text{di}}, X)|X = x\} \lambda(x) dF(x), \end{aligned}$$

where $E^*(\cdot|X)$ denotes the expectation with respect to the joint distribution of (D, H^{di}) given X defined by

$$p^*(H^{\text{di}}, D|X, \mathcal{B}, \gamma_1, \gamma_0) = \frac{S(D, H^{\text{di}}, X, \mathcal{B}, \gamma_1, \gamma_0)}{\sum_d \sum_h S(d, h, X, \mathcal{B}, \gamma_1, \gamma_0)} \quad (\text{A.1})$$

and $\lambda(x) = \sum_d \sum_h S(d, h, x, \mathcal{B}, \gamma_1, \gamma_0)$.

Lemma A.2 below provides an explicit expression for the estimating function $\bar{\mathcal{L}}_{\Phi}(\cdot)$.

LEMMA A.2 Write

$$\begin{aligned} S(D, H^{\text{di}}, X, \Phi) &= S\{D, H^{\text{di}}, X, \mathcal{B}, \gamma_1, \mathcal{G}(\theta, \gamma_1, F)\}, \\ p^*(D, H^{\text{di}}|X, \Phi) &= \frac{S(D, H^{\text{di}}, X, \Phi)}{\sum_{d_*} \sum_{h_*} S(d_*, h_*, X, \Phi)}, \\ \mathcal{R}(D, H^{\text{di}}, X, \Phi) &= \frac{\partial}{\partial \Phi} \log S(D, H^{\text{di}}, X, \Phi). \end{aligned}$$

Then

$$\mathcal{L}_{\Phi}\{D, H^{\text{di}}, X, \mathcal{B}, \gamma_1, \mathcal{G}(\theta, \gamma_1, F)\} = \mathcal{R}(D, H^{\text{di}}, X, \Phi) - E^*\{\mathcal{R}(\cdot)|X\}. \quad (\text{A.2})$$

Proof: By definition,

$$\mathcal{L}_{\Phi}\{D, H^{\text{di}}, X, \mathcal{B}, \gamma_1, \mathcal{G}(\theta, \gamma_1, F)\} = \frac{\partial}{\partial \Phi} \log p^*(D, H^{\text{di}}|X, \Phi),$$

and direct calculation yields

$$\frac{\partial}{\partial \Phi} \log p^*(D, H^{\text{di}}|X, \Phi) = \mathcal{R}(D, H^{\text{di}}, X, \Phi) - \sum_d \sum_h \frac{\mathcal{R}(d, h, X, \Phi) S(d, h, X, \Phi)}{\sum_{d_*} \sum_{h_*} S(d_*, h_*, X, \Phi)},$$

which proves the result.

Lemma A.3 provides explicit forms for the information matrices.

LEMMA A.3 Let $\mathcal{I}_{\Omega\Omega} = E_{\text{cc}}\{-\mathcal{L}_{\Omega\Omega}(D, H^{\text{di}}, X, \Omega)\}$, where $\mathcal{L}_{\Omega\Omega}$ is the second derivative of $\mathcal{L}(\cdot)$ with respect to Ω and $\bar{\mathcal{I}}_{\Omega\Omega} = E_{\text{cc}}\{-\partial \bar{\mathcal{L}}_{\Omega}(\cdot)/\partial \Omega^{\text{T}}\}$. Then

$$\begin{aligned} \mathcal{I}_{\Omega\Omega} &= E_{\text{cc}}\{\mathcal{L}_{\Omega}(D, H^{\text{di}}, X, \Omega) \mathcal{L}_{\Omega}^{\text{T}}(D, H^{\text{di}}, X, \Omega)\}, \\ \bar{\mathcal{I}}_{\Omega\Omega} &= E_{\text{cc}}\{\bar{\mathcal{L}}_{\Omega}(D, G, X, \Omega) \bar{\mathcal{L}}_{\Omega}^{\text{T}}(D, G, X, \Omega)\}. \end{aligned}$$

Proof: The first identity has been given in Lemma 4 of Chatterjee and Carroll (2005). To show the second identity, applying the chain rule we have

$$\begin{aligned} E_{\text{cc}}\{-\bar{\mathcal{L}}_{\Omega\Omega}(D, G, X, \Omega)\} &= E_{\text{cc}}[E\{-\mathcal{L}_{\Omega\Omega}(D, H^{\text{di}}, X, \Omega)|D, G, X\}] \\ &\quad - E_{\text{cc}} \left\{ \sum_{h \in \mathcal{C}(G)} \mathcal{L}_{\Omega}(D, h, X, \Omega) \frac{\partial w(h, \Omega)}{\partial \Omega^{\text{T}}} \right\}. \end{aligned} \quad (\text{A.3})$$

The first term of (A.3) equals

$$E_{\text{cc}}\{-\mathcal{L}_{\Omega\Omega}(D, H^{\text{di}}, X, \Omega)\} = E_{\text{cc}}\{\mathcal{L}_{\Omega}(D, H^{\text{di}}, X) \mathcal{L}_{\Omega}^{\text{T}}(D, H^{\text{di}}, X)\}.$$

By the definition of $w(h, \Omega)$ given in (3.8), it easy to see that

$$w(h, \Omega) = p^*(H^{\text{di}} = h|D, X, \Omega),$$

where the joint density $p^*(D, H^{\text{di}}|X, \Omega)$ is defined in (A.1). Recalling further that $\mathcal{L}_{\Omega}(D, h, X, \Omega) = \partial \log p^*(D, H^{\text{di}} = h|X, \Omega)/\partial \Omega$ and $\bar{\mathcal{L}}_{\Omega}(D, G, X, \Omega) = \sum_h w(h, \Omega) \mathcal{L}_{\Omega}(D, h, X, \Omega)$, we have the identity

$$\frac{\partial w(h, \Omega)}{\partial \Omega} = \mathcal{L}_{\Omega}(D, h, X, \Omega) w(h, \Omega) - \bar{\mathcal{L}}_{\Omega}(D, G, X, \Omega) w(h, \Omega).$$

Hence, the second term of (A.3) leads to

$$E_{\text{cc}}\{\mathcal{L}_{\Omega}(D, H^{\text{di}}, X, \Omega) \mathcal{L}_{\Omega}^{\text{T}}(D, H^{\text{di}}, X, \Omega)\} - E_{\text{cc}}\{\bar{\mathcal{L}}_{\Omega}(D, G, X, \Omega) \bar{\mathcal{L}}_{\Omega}^{\text{T}}(D, G, X, \Omega)\}.$$

The desired result thus follows.

APPENDIX B: PROOF OF THEOREM

We will first obtain the asymptotic expansion of the proposed estimating equation (3.6), by which the large sample distribution theory for $\hat{\Phi}$ can be derived immediately from the central limit theorem.

For our estimator $\widehat{\Phi} = (\widehat{\mathcal{B}}, \widehat{\gamma}_1, \widehat{\theta})$, a standard Taylor series expansion of (3.6) yields

$$\begin{aligned}
0 &= n^{-1/2} \sum_{i=1}^n \bar{\mathcal{L}}_{\Phi} \{D_i, G_i, X_i, \mathcal{B}, \gamma_1, \mathcal{G}(\theta, \gamma_1, F)\} \\
&\quad + n^{-1/2} \sum_{i=1}^n [\bar{\mathcal{L}}_{\Phi} \{D_i, G_i, X_i, \mathcal{B}, \gamma_1, \mathcal{G}(\theta, \gamma_1, \widehat{F}_{\text{emp}})\} \\
&\quad - \bar{\mathcal{L}}_{\Phi} \{D_i, G_i, X_i, \mathcal{B}, \gamma_1, \mathcal{G}(\theta, \gamma_1, F)\}] \\
&\quad - \bar{\mathcal{I}} n^{1/2} (\widehat{\Phi} - \Phi) + o_p(1).
\end{aligned} \tag{B.1}$$

In view of (2.5) and (3.5), let $\gamma_0 = \mathcal{G}(\theta, \gamma_1, F)$ and $\widehat{\gamma}_0 = \mathcal{G}(\theta, \gamma_1, \widehat{F}_{\text{emp}})$. Recall that $\widehat{\gamma}_0$ is solved from (3.5):

$$0 = \sum_{i=1}^n \mathbf{q}_{\text{hap}}(X_i, \gamma_1, \widehat{\gamma}_0) \widehat{p}_{\text{emp}}(D_i) - \mathbf{q}_{\text{HWE}}(\mathbf{h}, \theta).$$

Making a further Taylor series expansion, we have

$$0 = n^{1/2} \left\{ \sum_{i=1}^n \mathbf{q}_{\text{hap}}(X_i, \gamma_1, \widehat{\gamma}_0) \widehat{p}_{\text{emp}}(D_i) - \mathbf{q}_{\text{HWE}}(\theta) \right\} + Q_n n^{1/2} (\widehat{\gamma}_0 - \gamma_0) + o_p(1).$$

Note that

$$\begin{aligned}
Q_n &= \sum_{i=1}^n \frac{\partial}{\partial \gamma_0} \mathbf{q}_{\text{hap}}(X_i, \gamma_0, \gamma_1) \widehat{p}_{\text{emp}}(D_i) = \int \frac{\partial}{\partial \gamma_0} \mathbf{q}_{\text{hap}}(X, \gamma_0, \gamma_1) d\widehat{F}_{\text{emp}}(X) \\
&= Q + o_p(1).
\end{aligned}$$

Hence,

$$\begin{aligned}
n^{1/2} (\widehat{\gamma}_0 - \gamma_0) &= n^{1/2} Q^{-1} \sum_{i=1}^n \{ \mathbf{q}_{\text{hap}}(X_i, \gamma_1, \gamma_0) - \mathbf{q}_{\text{HWE}}(\theta) \} \widehat{p}_{\text{emp}}(D_i) + o_p(1) \\
&\equiv n^{-1/2} \sum_{i=1}^n k(X_i, D_i) + o_p(1).
\end{aligned}$$

An explicit expression for Q is given in Appendix C. Consequently, the middle 2 terms in the expansion (B.1) reduce to

$$\begin{aligned}
&n^{-1/2} \sum_{i=1}^n [\bar{\mathcal{L}}_{\Phi} \{D_i, G_i, X_i, \mathcal{B}, \gamma_1, \mathcal{G}(\theta, \gamma_1, \widehat{F}_{\text{emp}})\} - \bar{\mathcal{L}}_{\Phi} \{D_i, G_i, X_i, \mathcal{B}, \gamma_1, \mathcal{G}(\theta, \gamma_1, F)\}] \\
&= -\bar{\mathcal{I}}_{\Phi \gamma_0} n^{1/2} (\widehat{\gamma}_0 - \gamma_0) + o_p(1) \\
&= -n^{-1/2} \sum_{i=1}^n \mathcal{K}(X_i, D_i) + o_p(1).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
n^{1/2}(\widehat{\Phi} - \Phi) &= n^{-1/2}\bar{\mathcal{I}}^{-1} \sum_{i=1}^n [\bar{\mathcal{L}}_{\Phi}\{D_i, G_i, X_i, \mathcal{B}, \gamma_1, \mathcal{G}(\theta, \gamma_1, F)\} - \mathcal{K}(X_i, D_i)] + o_p(1) \\
&= n^{-1/2}\bar{\mathcal{I}}^{-1} \sum_{i=1}^n \bar{\mathcal{L}}_{\Phi}\{D_i, G_i, X_i, \mathcal{B}, \gamma_1, \mathcal{G}(\theta, \gamma_1, F)\} \\
&\quad - E[\bar{\mathcal{L}}_{\Phi}\{D_i, G, X, \mathcal{B}, \gamma_1, \mathcal{G}(\theta, \gamma_1, F)\}|D_i] \\
&\quad - [\mathcal{K}(X_i, D_i) - E\{\mathcal{K}(X, D_i)|D_i\}] + o_p(1).
\end{aligned}$$

Note that in the last equality above, we have used the fact that

$$\sum_{i=1}^n E[\bar{\mathcal{L}}_{\Phi}\{D_i, G, X, \mathcal{B}, \gamma_1, \mathcal{G}(\theta, \gamma_1, F)\}|D_i] = 0$$

and $\sum_{i=1}^n E\{\mathcal{K}(X, D_i)|D_i\} = 0$, which follow directly from Lemmas A.1 and A.2. This completes the proof.

APPENDIX C: EXPRESSIONS FOR THE DERIVATIVES OF $\mathcal{G}(\theta, \gamma_1, F)$ AND $\mathcal{G}(\theta, \gamma_1, \widehat{F}_{\text{emp}})$

Recall that $\mathcal{G}(\theta, \gamma_1, F)$ is the solution of γ_0 to (2.5). Here, we define γ_1 to be the vectorized form for the diplotype-effect parameters subject to certain mode of effects (e.g. additive effect) of haplotypes, and define η to be the vectorized form for the full set of diplotype effects $\{\gamma_{1j_1j_2}\}$. Differentiating both sides of (2.5) with respect to γ_1 , we have

$$0 = \left\{ \int \frac{\partial \mathbf{q}_{\text{hap}}(x, \gamma_0, \gamma_1)}{\partial \gamma_0} dF(x) \right\} \frac{\partial \gamma_0}{\partial \gamma_1} + \int \frac{\partial \mathbf{q}_{\text{hap}}(x, \gamma_0, \gamma_1)}{\partial \eta} \left(\frac{\partial \eta}{\partial \gamma_1^{\text{T}}} \otimes x^{\text{T}} \right) dF(x).$$

Let

$$Q(x) = \text{diag}\{\mathbf{q}_{\text{hap}}(x, \gamma_0, \gamma_1)\} - \mathbf{q}_{\text{hap}}(x, \gamma_0, \gamma_1)\mathbf{q}_{\text{hap}}(x, \gamma_0, \gamma_1)^{\text{T}}$$

and $Q = \int Q(x)dF(x)$. Then

$$\mathcal{G}_{\gamma_1} = \frac{\partial \gamma_0}{\partial \gamma_1} = -Q^{-1} \int Q(x) \left(\frac{\partial \eta}{\partial \gamma_1^{\text{T}}} \otimes x^{\text{T}} \right) dF(x).$$

Similarly, letting

$$R = \text{diag}\{\mathbf{q}_{\text{HWE}}(\theta)\} - \mathbf{q}_{\text{HWE}}(\theta)\mathbf{q}_{\text{HWE}}(\theta)^{\text{T}},$$

we have

$$\mathcal{G}_{\theta} = \frac{\partial \gamma_0}{\partial \theta} = Q^{-1} R \frac{\partial \theta_{h^{\text{di}}}}{\partial \theta^{\text{T}}}.$$

The derivatives of $\mathcal{G}(\theta, \gamma_1, \widehat{F}_{\text{emp}})$ can be obtained by replacing dF in the above quantities with $d\widehat{F}_{\text{emp}}$.

REFERENCES

- ANDERSEN, J. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, Series B* **32**, 283–301.
- CHATTERJEE, N. AND CARROLL, R. J. (2005). Semiparametric maximum likelihood estimation in case-control studies of gene-environmental interactions. *Biometrika* **92**, 399–418.
- CHATTERJEE, N., CHEN, J., SPINKA, C. AND CARROLL, R. J. (2006). Comment on the paper likelihood based inference on haplotype effects in genetic association studies by D. J. Lin and D. Zeng. *Journal of the American Statistical Association* **101**, 108–110.
- EPSTEIN, M. AND SATTEN, G. (2003). Inference of haplotype effects in case-control studies using unphased genotype data. *American Journal of Human Genetics* **73**, 1316–1329.
- GOHAGAN, J. K., PROROK, P. C., HAYES, R. B., KRAMER, B. S. (2000). The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: History, organization, and status. *Controlled Clinical Trials* **21**, 251S–272S.
- LAKE, S. L., LYON, H., TANTISIRA, K., SILVERMAN, E. K., WEISS, S. T., LAIRD, N. M. AND SCHAID, D. J. (2003). Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. *Human Heredity* **55**, 56–65.
- LI, S. S., KHALID, N., CARLSON, C. AND ZHAO, L. P. (2003). Estimating haplotype frequencies and standard errors for multiple single nucleotide polymorphisms. *Biostatistics* **4**, 513–522.
- LIN, D. Y. AND ZENG, D. (2006). Likelihood-based inference on haplotype effects in genetic association studies (with discussion). *Journal of the American Statistical Association* **101**, 89–118.
- MOSLEHI, R., CHATTERJEE, N., CHURCH, T. R., CHEN, J., YEAGER, M., WEISSFIELD, J., HEIN, D. W. AND HAYES, R. B. (2006). Cigarette smoking, n-acetyltransferase genes and the risk of advanced colorectal adenoma. *Pharmacogenomics* **7**, 819–829.
- PRENTICE, R. L. AND PYKE, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–411.
- ROEDER, K., CARROLL, R. J. AND LINDSAY, B. G. (1996). A nonparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association* **91**, 722–732.
- SATTEN, G. A. AND EPSTEIN, M. P. (2004). Comparison of prospective and retrospective methods for haplotype inference in case-control studies. *Genetic Epidemiology* **27**, 192–201.
- SCHAID, D. J. (2004). Evaluating associations of haplotypes with traits. *Genetic Epidemiology* **27**, 348–364.
- SPINKA, C., CARROLL, R. J. AND 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.
- WALLENSTEIN, S., HODGE, S. E. AND WESTON, A. (1998). Logistic regression model for analyzing extended haplotype data. *Genetic Epidemiology* **15**, 173–181.
- ZHAO, L. P., LI, S. S. AND KHALID, N. A. (2003). Method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. *American Journal of Human Genetics* **72**, 1231–1250.

[Received October 13, 2006; revised March 2, 2007; accepted for publication March 14, 2007]