Genetic Epidemiology 29 : 299–312 (2005)

Maximum Likelihood Estimation of Haplotype Effects and Haplotype-Environment Interactions in Association Studies

D.Y. Lin,[1]* D. Zeng,[1] and R. Millikan[2]

[1]*Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina*
[2]*Department of Epidemiology, University of North Carolina, Chapel Hill, North Carolina*

The associations between haplotypes and disease phenotypes offer valuable clues about the genetic determinants of complex diseases. It is highly challenging to make statistical inferences about these associations because of the unknown gametic phase in genotype data. We describe a general likelihood-based approach to inferring haplotype-disease associations in studies of unrelated individuals. We consider all possible phenotypes (including disease indicator, quantitative trait, and potentially censored age at onset of disease) and all commonly used study designs (including cross-sectional, case-control, cohort, nested case-control, and case-cohort). The effects of haplotypes on phenotype are characterized by appropriate regression models, which allow various genetic mechanisms and gene-environment interactions. We present the likelihood functions for all study designs and disease phenotypes under Hardy-Weinberg disequilibrium. The corresponding maximum likelihood estimators are approximately unbiased, normally distributed, and statistically efficient. We provide simple and efficient numerical algorithms to calculate the maximum likelihood estimators and their variances, and implement these algorithms in a freely available computer program. Extensive simulation studies demonstrate that the proposed methods perform well in realistic situations. An application to the Carolina Breast Cancer Study reveals significant haplotype effects and haplotype-smoking interactions in the development of breast cancer. *Genet. Epidemiol.* 29 :299–312, 2005. © 2005 Wiley-Liss, Inc.

Key words: case-control studies; cohort studies; complex diseases; EM algorithm; gene-environment interactions; haplotype analysis; Hardy-Weinberg equilibrium; linkage disequilibrium; profile likelihood; retrospective likelihood; SNPs; unphased genotype

Grant sponsor: National Institutes of Health.
*Correspondence to: Danyu Lin, Ph.D., Department of Biostatistics, University of North Carolina, McGavran-Greenberg Hall, CB #7420, Chapel Hill, NC 27599-7420. E-mail: lin@bios.unc.edu
Received 16 December 2004; Accepted 23 March 2005
Published online 20 October 2005 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20098

INTRODUCTION

Searching for genetic determinants of complex diseases (e.g., hypertension, bipolar disorder, cancer, diabetes, and schizophrenia) is one of the most important challenges in human genetics. Complex diseases are likely affected by an array of genetic and environmental factors, as well as their interactions. It is widely believed that the genetic dissection of these diseases requires association studies, which explore the relationships between genetic variants, such as single-nucleotide polymorphisms (SNPs) and disease phenotypes [Risch, 2000; Botstein and Risch, 2003]. In fact, there is now a proliferation of SNP-based association studies worldwide, thanks to the availability of dense SNP maps across the human genome [International SNP Map Working Group, 2001; International HapMap Consortium, 2003] and the continuing improvement in genotyping efficiency.

There are several options in designing population-based association studies. The simplest is the cross-sectional design, which collects phenotype and SNP data on a random sample of individuals. This design is preferable if the disease of interest is common or if one is interested in some disease-related traits, such as blood pressure. For rare diseases, it is more cost-effective to adopt the case-control design, which collects data retrospectively on a sample of cases (i.e., diseased individuals) and a sample of controls (i.e., disease-free individuals). If one is interested in the age at onset of a disease, then it is desirable to follow a cohort

of at-risk individuals over time and record their times of disease occurrence.

For all the aforementioned study designs, standard statistical methods can be used to assess the association between a particular SNP and a disease phenotype. The information content in any single SNP, however, is very limited. It is desirable to combine information from multiple SNPs through the use of haplotypes. Haplotypes, which are specific combinations of nucleotides at several tightly linked SNPs on the same chromosome of an individual, incorporate the linkage disequilibrium information and correspond directly to protein sequences. The use of SNP-based haplotypes may offer more powerful tests of genetic associations than the use of individual SNPs, especially when the causal SNPs are not typed or when multiple mutations occur in *cis* position [Akey et al., 2001; Fallin et al., 2001; Morris and Kaplan, 2002; Schaid et al., 2002; Zaykin et al., 2002; Botstein and Risch, 2003; Schaid, 2004]. Because the actual number of haplotypes tends to be much smaller than the number of all possible haplotypes, haplotyping also represents a data-reduction strategy.

Routine genotyping procedures do not provide gametic phase information, so that only unphased genotypes rather than haplotypes are directly measured. A number of authors [e.g., Clark, 1990; Excoffier and Slatkin, 1995; Stephens et al., 2001; Zhang et al., 2001; Niu et al., 2002; Qin et al., 2002] developed methods to estimate haplotype frequencies or infer individual haplotypes from unphased genotype data. To make inferences about a haplotype-disease association, one may then relate the probabilistically constructed haplotypes to the disease phenotype through a regression model [e.g., Zaykin et al., 2002]. This approach fails to account for the variation due to haplotype estimation. More important, it produces biased and inefficient estimators of regression parameters, especially when the effect sizes are large or haplotype uncertainty is high [e.g., Kraft et al., 2005].

Several methods have been proposed to make proper inferences about the effects of haplotypes on disease phenotypes. Virtually all these methods are related to likelihood. It is useful to distinguish between the prospective and retrospective likelihoods. For cross-sectional and cohort studies, it is natural to use the prospective likelihood, which pertains to the probability of a phenotype given a genotype. For case-control studies, the sampling is conditional on the case-control status, so it is more appropriate to use the retrospective likelihood, which is the probability of a genotype given a phenotype. For the conventional logistic regression analysis of case-control data, maximizing the prospective or retrospective likelihood yields the same estimator of the odds ratio [Prentice and Pyke, 1979]. This equivalence, however, requires an unrestricted distribution of the exposure, and does not hold when the exposure of interest is the diplotype (i.e., haplotype pair), because its distribution has to be restricted in the statistical analysis due to incomplete exposure data (i.e., phase ambiguity).

Motivated by the equivalence of the prospective and retrospective likelihoods for case-control studies with complete exposure data, Zhao et al. [2003] developed an estimating function to approximate the expectation of the complete-data prospective-likelihood score function given the observed data. This method assumes rare diseases and is not statistically efficient. Epstein and Satten [2003] derived a clever retrospective likelihood for the relative-risk parameter. Currently, this method does not allow for environmental variables. Stram et al. [2003] proposed a conditional likelihood for the odds ratio assuming that cases and controls are chosen with known probabilities from the source population, and did not allow for environmental variables either. As mentioned above, it is desirable to accommodate environmental variables, since complex diseases are likely to be influenced by environmental exposures and gene-environment interactions. For cross-sectional studies, Schaid et al. [2002] and Lake et al. [2003] described likelihood-based methods under generalized linear models. Lin [2004] considered proportional hazards regression for cohort studies. All the aforementioned literature assumes Hardy-Weinberg equilibrium. Simulation results [Lake et al., 2003; Satten and Epstein, 2004] showed that departures from Hardy-Weinberg equilibrium can severely bias the analysis.

In this article, we present a general approach to estimating haplotype-disease associations. For case-control studies, we incorporate environmental variables and provide efficient estimators. For cross-sectional and cohort studies, we accommodate more flexible models than the existing literature. In addition, we explore case-cohort and nested case-control designs [Kalbfleisch and Prentice, 2002, p. 339] for cohort studies, under which only a subset of the cohort members needs to be genotyped. For all study designs, we allow Hardy-Weinberg disequilibrium. We describe

appropriate likelihoods for all study designs and disease phenotypes. Except for cross-sectional studies, the likelihoods involve high-dimensional parameters. Thus, there are considerable theoretical and numerical challenges. In two statistical papers [Lin and Zeng, 2006; Zeng et al., 2005], we showed that the maximum likelihood estimators are approximately unbiased, normally distributed, and statistically efficient, and we derived simple and efficient algorithms to compute the maximum likelihood estimators and their variance estimators. In this article, we briefly describe those theoretical results and numerical algorithms; we refer interested readers to the two statistical papers for details. We present an application to case-control data from the Carolina Breast Cancer Study [Newman et al., 1995; Millikan et al., 2003].

METHODS

Suppose that each individual is genotyped at M tightly linked biallelic SNPs. At each SNP site, the two possible alleles are denoted by 0 vs. 1. Thus, each haplotype h is an ordered sequence of M numbers of zeros and ones. The total number of possible haplotypes is $K = 2^M$, although the actual number of haplotypes consistent with the observed data is often much smaller. For $k = 1, \ldots, K$, let h_k denote the kth possible haplotype. Figure 1 shows the four possible haplotypes for two SNPs.

For each individual, the multi-SNP genotype is an ordered sequence of M numbers of zeros, ones, and twos. Let H denote the diplotype (i.e., the pair of haplotypes on the two homologous chromosomes) of an individual, and G the corresponding (unphased) genotype. Note that G codes the number of "1" alleles at each locus. We write $H = (h_k, h_l)$ if the individual's diplotype consists of h_k and h_l, in which case $G = h_k + h_l$. We cannot determine H on the basis of G if the individual is heterozygous at more than one SNP or if any SNP genotype is missing. For the case of two SNPs shown in Figure 1, if $G = (2, 1)$, then $H = (h_3, h_4)$; if $G = (1, 1)$, then $H = (h_1, h_4)$ or $H = (h_2, h_3)$.

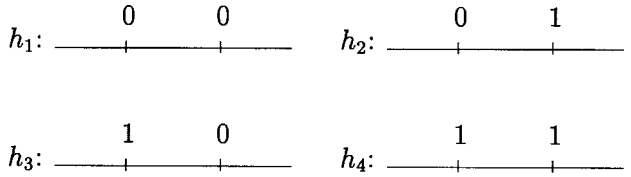

Fig. 1. Possible haplotype configurations with 2 SNPs.

Let Y be the phenotype of interest, and X be a set of environmental variables or covariates. In association studies, we are interested in estimating the effects of X and H on Y. Such a relationship can be characterized by the conditional density function $P(Y \mid X, H; \theta)$ indexed by a set of parameters θ. There are various choices for the association or disease model. Suppose that h^* is the target haplotype of interest, and Y is a binary disease indicator. In the absence of covariates, we may employ a logistic regression model with the linear predictor $\alpha + \beta I(h_k = h_l = h^*)$ under a recessive model, $\alpha + \beta\{I(h_k = h^*) + I(h_l = h^*) - I(h_k = h_l = h^*)\}$ under a dominant model, $\alpha + \beta\{I(h_k = h^*) + I(h_l = h^*)\}$ under an additive model, and $\alpha + \beta_1\{I(h_k = h^*) + I(h_l = h^*)\} + \beta_2 I(h_k = h_l = h^*)$ under a codominant model, where h_k and h_l are the pair of haplotypes in H, and $I(\mathcal{A})$ takes the value 1 or 0, dependent on whether the event $\mathcal{A}$ is true or false. The codominant model includes the other three models as special cases. The codominant model with gene-environment interactions has the linear predictor

$$\alpha + \beta_1\{I(h_k = h^*) + I(h_l = h^*)\} + \beta_2 I(h_k = h_l = h^*) + \beta_3' X + \beta_4'\{I(h_k = h^*) + I(h_l = h^*)\}X + \beta_5' I(h_k = h_l = h^*)X. \quad (1)$$

Here, θ consists of $\beta_1, \ldots, \beta_5$ and α, where α is the intercept and the βs are log odds ratios.

Although we are interested in how H affects Y, we observe G instead of H. With such missing data, it is in general not possible to estimate θ without imposing some restrictions on the distribution of H. Virtually all the published work on haplotype inference requires the diplotype distribution to satisfy Hardy-Weinberg equilibrium, such that

$$\pi_{kl} = \pi_k \pi_l, \quad k, l = 1, \ldots, K \quad (2)$$

where π_{kl} is the probability that H consists of h_k and h_l, and π_k is the probability that a particular haplotype is h_k. It is useful to consider the following forms of departures from Hardy-Weinberg equilibrium:

$$\pi_{kl} = \begin{cases} \pi_k^2 + \rho\pi_k(1 - \pi_k), & k = l, \\ (1 - \rho)\pi_k\pi_l, & k \neq l \end{cases} \quad (3)$$

and

$$\pi_{kl} = \begin{cases} \pi_k^2/(1 - \rho + \rho\sum_{j=1}^K \pi_j^2), & k = l, \\ (1 - \rho)\pi_k\pi_l/(1 - \rho + \rho\sum_{j=1}^K \pi_j^2), & k \neq l \end{cases} \quad (4)$$

where $0 \le \pi_k \le 1$, and $\sum_{k=1}^{K} \pi_k = 1$. In (3), ρ is called the inbreeding coefficient or fixation index [Weir, 1996, p. 93]. Both (3) and (4) reduce to (2) if $\rho = 0$. There are excess homozygosity (i.e., $\pi_{kk} > \pi_k^2, k = 1, \ldots, K$) and excess heterozygosity (i.e., $\pi_{kk} < \pi_k^2, k = 1, \ldots, K$) under $\rho > 0$ and $\rho < 0$, respectively. Satten and Epstein [2004] considered condition (3) for the control population under the case-control design.

We denote the probability distribution of H by $P(H; \gamma)$, where γ includes π_k ($k = 1, \ldots, K$) and ρ. Our methods apply to conditions (2), (3), and (4) unless otherwise indicated. We assume that X is independent of H conditional on G. We use $\mathcal{S}(G)$ to denote the collection of diplotypes that are compatible with genotype G, i.e., the diplotypes (h_k, h_l) such that $h_k + h_l = G$. We allow missing genotypes. If G is missing, then $\mathcal{S}(G)$ is expanded accordingly.

CROSS-SECTIONAL STUDIES

A cross-sectional study selects a random sample of n individuals from the source population and measures the phenotype, genotype, and covariates for each individual. The data consist of (Y_i, X_i, G_i) $(i = 1, \ldots, n)$. The phenotype or trait Y can be discrete or continuous, univariate or multivariate. For a univariate trait, the association model $P(Y \mid X, H; \theta)$ may take the form of a generalized linear model [McCullagh and Nelder, 1989] with the linear predictor given in (1). In particular, the logistic and linear regression models may be chosen for the binary and quantitative traits, respectively. If the trait is measured repeatedly in a longitudinal study, then a generalized linear mixed model [Diggle et al., 2002] may be suitable.

The likelihood function for parameters θ and γ is proportional to

$$\prod_{i=1}^{n} \sum_{H_i \in \mathcal{S}(G_i)} P(Y_i | X_i, H_i; \theta) P(H_i; \gamma). \tag{5}$$

We may maximize (5) directly by the standard Newton-Raphson algorithm, or indirectly by the expectation-maximization (EM) algorithm [Dempster et al., 1977]. In the EM algorithm, which is described in Appendix A, the haplotypes are treated as missing data. The maximum likelihood estimators (MLEs) are consistent and asymptotically normal with a covariance matrix that can be consistently estimated by the inverse of the observed Fisher information matrix. In other words, the MLEs of θ and γ are, for large samples, approximately normal with means θ and γ, and with the covariance matrix being the negative inverse of the second derivative matrix of the log-likelihood evaluated at the MLEs. Furthermore, the MLEs are asymptotically efficient in that they have the smallest variances among all possible estimators of θ and γ, at least for large samples.

CASE-CONTROL STUDIES WITH KNOWN POPULATION TOTALS

Under a case-control design, we obtain separate random samples of cases and controls from a source population. Suppose that the total numbers of cases and controls in the source population are known. This information is often available from hospital records, disease registries, and official statistics [Scott and Wild, 1997]. The case-control sample may be drawn from a cohort study, in which case the cohort serves as the source population with known population totals. If the phenotype pertains to a binary disease indicator, then the association model $P(Y \mid X, H; \theta)$ may be a logistic, probit, or complementary log-log regression model. When there are more than two disease categories, the proportional odds model, the multivariate probit, and multivariate logistic regression models may be used.

Let n and N denote the total numbers of individuals in the case-control sample and the source population, respectively. For individuals in the case-control sample, the data take the same form as in the situation of a cross-section study. For the $(N-n)$ individuals not selected, we let Y_j denote the disease status for the jth subject, $j = n+1, \ldots, N$.

The likelihood function can be written as

$$\prod_{i=1}^{n} \left\{ \sum_{H_i \in \mathcal{S}(G_i)} P(Y_i | X_i, H_i; \theta) P(H_i; \gamma) P(X_i | G_i) \right\} \times \prod_{j=n+1}^{N} \left\{ \sum_{X,G} \sum_{H \in \mathcal{S}(G)} P(Y_j | X, H; \theta) P(H; \gamma) P(X | G) \right\} \tag{6}$$

where $P(X \mid G)$ is the conditional density function of X given G, and the summation inside the second product is taken over all possible values of X and G. We refer to $P(X \mid G)$ as a nuisance parameter, in that we are not interested in such parameters, although they cannot be eliminated from the likelihood and thus have to be estimated. If there are continuous covariates, then $P(X \mid G)$ is an infinite-dimensional nuisance parameter in that $P(X \mid G)$ is a continuous function in X for

each G. The presence of infinite-dimensional nuisance parameters poses considerable challenges, both numerically and theoretically.

We can maximize the likelihood by the Newton-Raphson algorithm or by the EM algorithm described in Appendix B. The resultant MLEs are again consistent, asymptotically normal, and asymptotically efficient [Lin and Zeng, 2006]. The variances of MLEs for θ and γ can be estimated by the profile likelihood method [Murphy and van der Vaart, 2000]. We can use the likelihood ratio statistics to make inferences about θ and γ without estimating the variances.

CASE-CONTROL STUDIES WITH UNKNOWN POPULATION TOTALS

Under the traditional case-control design, we measure X and G on n_1 cases ($Y = 1$) and n_0 controls ($Y = 0$) without any knowledge of the population totals. Because the sampling is conditional on the case-control status and the population totals are unknown, it is necessary to use the retrospective likelihood, which takes the form

$$\prod_{i=1}^{n} \frac{\sum_{H_i \in \mathcal{S}(G_i)} P(Y_i|X_i,H_i;\,\theta)P(H_i;\,\gamma)P(X_i|G_i)}{\sum_{X,G}\sum_{H\in\mathcal{S}(G)} P(Y_i|X,H;\,\theta)P(H;\,\gamma)P(X|G)}$$

where $n = n_0 + n_1$. The parameters in this likelihood may not be identifiable, in that different parameter values may yield the same value of the likelihood. When the parameters are identifiable, the MLEs have the desirable theoretical properties [Lin and Zeng, 2006].

Rare disease is the main motivation for the case-control design. For the logistic regression with rare disease, $P(Y \mid X,H;\,\theta)$ is approximately equal to $\exp\{Y(\alpha + \beta'\mathcal{Z}(X,H))\}$, where $\mathcal{Z}(X,H)$ is a specific function of X and H. Then the likelihood becomes

$$\prod_{i=1}^{n} \frac{\sum_{H_i \in \mathcal{S}(G_i)} \exp\{Y_i\beta'\mathcal{Z}(X_i,H_i)\}P(H_i;\,\gamma)P(X_i|G_i)}{\sum_{X,G}\sum_{H\in\mathcal{S}(G)} \exp\{Y_i\beta'\mathcal{Z}(X,H)\}P(H;\,\gamma)P(X|G)}. \quad (7)$$

Like (6), this likelihood involves the infinite-dimensional nuisance parameters $P(X \mid G)$. It is considerably more challenging to deal with (7) than (6), both theoretically and computationally. Nevertheless, the parameters in (7) are identifiable, and the MLEs are consistent, asymptotically normal, and asymptotically efficient [Lin and Zeng, 2006]. An efficient and stable algorithm to obtain the MLEs for β and γ and to estimate their variances is provided in Appendix C.

In the absence of covariates, (7) reduces to the retrospective likelihood of Epstein and Satten [2003]. This is not surprising, since Epstein and Satten [2003] worked with the relative-risk parameter, which is approximately the same as the odds ratio when the disease is rare.

COHORT STUDIES

In a cohort study, we follow a random sample of n at-risk individuals to ascertain their ages at onset of disease. The individuals who are withdrawn prematurely from the study or who are disease-free at the end of the study have censored observations, in that their ages at onset are only known to be beyond their durations of follow-up. Let Y denote the age at onset and C denote the censoring time. We assume that C is independent of Y and H conditional on X and G. The data consist of $(\widetilde{Y}_i, \Delta_i, X_i, G_i)$ $(i = 1,\ldots,n)$, where $\widetilde{Y}_i = \min(Y_i, C_i)$, and $\Delta_i = I(Y_i \le C_i)$. The covariates X are allowed to vary over time.

It is convenient to employ the proportional hazards model [Cox, 1972], which specifies that the hazard function of Y conditional on X and H takes the form

$$\lambda(t|X,H) = \lambda_0(t)\exp\{\beta'\mathcal{Z}(X(t),H)\} \quad (8)$$

where $\mathcal{Z}(X(t),H)$ is a specific function of $X(t)$ and H, β is the corresponding set of log hazard ratio parameters, and $\lambda_0(t)$ is an arbitrary baseline hazard function. Integrating both sides of (8) yields the equivalent representation in terms of the cumulative hazard function: $\Lambda(y|X,H) = \int_0^y \exp\{\beta'\mathcal{Z}(X(t),H)\}d\Lambda_0(t)$, where $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$. Because the proportional hazards assumption may be violated in applications, we consider a broad class of transformation models,

$$\Lambda(y|X,H) = Q\left(\int_0^y \exp\{\beta'\mathcal{Z}(X(t),H)d\Lambda_0(t)\}\right) \quad (9)$$

where Q is an increasing function. If the covariates do not depend on time, then equation (9) can be written in the familiar linear model form: $\Gamma(Y) = -\beta'\mathcal{Z}(X,H) + \varepsilon$, where Γ is an unknown transformation, and ε is an error term with a known distribution function F. The choices of the extreme-value and standard logistic distributions for F yield the proportional hazards model and the proportional odds model [Pettitt, 1984], respectively.

The likelihood takes the form

$$\prod_{i=1}^{n} \sum_{H_i \in \mathcal{S}(G_i)} \lambda(\widetilde{Y}_i|X_i,H_i)^{\Delta_i} \exp\Big\{-\Lambda(\widetilde{Y}_i|X_i,H_i)\Big\} P(H_i;\,\gamma). \tag{10}$$

We maximize this likelihood over β, γ, and $\Lambda_0(t)$, treating $\Lambda_0(t)$ as a step function with jumps only at the observed disease occurrence times. The maximization can be carried out via an optimization algorithm, such as *fminunc* of MATLAB, or the EM algorithm described in Appendix D. The MLEs of β and γ are consistent, asymptotically normal, and asymptotically efficient, and their variances can be estimated by the observed information matrix or by the profile likelihood method [Lin and Zeng, 2006].

CASE-COHORT AND NESTED CASE-CONTROL STUDIES

For large cohorts with rare diseases, it is not cost-effective to genotype all cohort members. In fact, there is little loss of efficiency in the statistical inference by genotyping all the cases and a small fraction of controls as opposed to all controls. The most commonly used sampling procedures include nested case-control sampling, in which a small number (typically in the range of 1–5) of controls are matched to each case by random sampling from the set of individuals who are disease-free and under observation at the time of occurrence of that case, and case-cohort sampling, in which a random subcohort is selected from the entire cohort [Kalbfleisch and Prentice, 2002, p. 339–342].

Suppose that a total of n_c individuals is selected for genotyping out of a cohort of size n. For these individuals, the data consist of $(\widetilde{Y}_i, \Delta_i, X_i, G_i)$ $(i = 1, \ldots, n_c)$. For those not selected for genotyping, the data consist of $(\widetilde{Y}_j, \Delta_j, X_i)$ $(j = n_c{+}1, \ldots, n)$. As in the case of full cohort sampling, we may employ the proportional hazards model given in (8) or the general class of transformation models given in (9).

The likelihood takes the form

$$\prod_{i=1}^{n_c} \sum_{H_i \in \mathcal{S}(G_i)} \lambda(\widetilde{Y}_i|X_i,H_i)^{\Delta_i} \exp\Big\{-\Lambda(\widetilde{Y}_i|X_i,H_i)\Big\} P(H_i;\,\gamma) \times \prod_{j=n_c+1}^{n} \sum_{H} \lambda(\widetilde{Y}_j|X_j,H)^{\Delta_j} \exp\Big\{-\Lambda(\widetilde{Y}_j|X_j,H)\Big\} P(H;\,\gamma) \tag{11}$$

where the summation inside the second product is taken over all possible values of H [Zeng et al., 2005]. We maximize this function over β, γ, and $\Lambda_0(t)$, treating $\Lambda_0(t)$ as a step function with jumps only at the observed times of disease occurrence. The maximization can be carried out through an optimization algorithm or the EM algorithm given in Appendix D. The MLEs of β and γ are consistent, asymptotically normal, and asymptotically efficient, and their variances can be estimated by the observed information matrix or by the profile likelihood method [Zeng et al., 2005].

RESULTS

CAROLINA BREAST CANCER STUDY

Breast cancer is influenced by a variety of genetic and environmental factors. The Carolina Breast Cancer Study (CBCS) is a population-based case-control study designed to identify the causes of breast cancer [Newman et al., 1995; Millikan et al., 2003]. Cases were identified from the North Carolina Central Cancer Registry, and controls were identified from Division of Motor Vehicles and Health Care Financing Administration lists. Cases were enrolled between 1993–2001, with oversampling of African American and younger women. Controls were frequency-matched to cases based on age (± 5 years) and race. Information on established and potential risk factors was obtained from in-person interviews. Blood samples were collected at time of interview.

In total, 2,311 cases (1,417 whites and 894 African Americans) and 2,022 controls (1,234 whites and 788 African Americans) were enrolled. Ages ranged from 26–74. There is a current effort to study the effects of smoking and several candidate genes, as well as their interactions, on the risk of breast cancer. In this article, we focus on three SNPs in the XRCC1 gene: codon 194 C to T, which leads to an amino-acid substitution of Arg to Trp, codon 280 G to A, which leads to an amino-acid substitution of Arg to His, and codon 399 G to A, which leads to an amino-acid substitution of Arg to Gln. For these three SNPs, codons 194, 280, and 399, the genotype data are missing in 11.1%, 11.9%, and 13.2% of cases, and in 10.1%, 10.5%, and 11.1% of controls.

Based on data from the control group, haplotype frequencies are estimated at 0.62, 0.27, 0.07, and 0.04 for haplotypes CGG, CGA, CAG, and TGG, respectively. The other four possible haplotypes have zero estimated frequencies, and the inbreeding coefficient is estimated at 0.04. We fit the logistic regression models with various genetic hypotheses. All models include age, race, and

smoking duration. The last variable has five categories: no active and no environmental tobacco smoking (ETS) after age 18, no active but ETS after 18, $\leq$10 years, 11–20 years, and $>$20 years. The first group serves as the reference in the analysis. A total of 16 subjects had no information on smoking duration, and those subjects were excluded from the analysis. There are no missing data on age or race. We used the retrospective likelihood given in (7), but with a slight modification to account for the unequal sampling probabilities. The modified likelihood involves an offset term, which is the natural logarithm of the ratio of the sampling probability for a case in the specific age-race stratum to the sampling probability for a control in the specific age-race stratum.

Tables I and II display the estimates for the effects of haplotypes and smoking duration, as well as their interactions, for target haplotypes CGG and CGA, respectively. Neither the haplotype effects nor haplotype-smoking interactions are significant for haplotypes CAG and TGG, and thus the results are not shown here. As expected, the effects of age and race are highly significant, though the results are not shown.

It is desirable to select an appropriate model among all potential models. We suggest selecting the model that minimizes the information criterion of Akaike [1985] (AIC), which is $-2\log L+2p$, where L is the likelihood evaluated at the MLEs, and p is the number of parameters in the model. Based on the values of the AIC, we select the dominant model for haplotype CGG and the additive model for haplotype CGA, although the values of the AIC are fairly close between the additive and dominant models for both haplotypes. The results reported in Tables I and II provide evidence for haplotype and smoking effects, as well as haplotype-smoking interactions. Specifically, haplotype CGA has a strong main effect as well as a strong interaction with environmental tobacco smoking in the development of breast cancer. A more comprehensive investigation involving multiple candidate genes and other smoking variables will be reported elsewhere.

SIMULATION STUDIES

We used Monte Carlo simulation to evaluate the proposed methods in realistic settings. We generated haplotypes from the observed haplotype distribution of the XRCC1 gene in the CBCS study. We set 10% of the genotype data to be missing for each of the three SNPs. We focused on the CGA haplotype. We generated disease incidence from the logistic regression model with additive genetic effects:

$$\begin{aligned}\text{logit}P(Y=1|X,H) &= \alpha+\beta_1\{I(h_k=h^*)+I(h_l=h^*)\}\\ &\quad+\beta_2X+\beta_3\{I(h_k=h^*)+I(h_l=h^*)\}X\end{aligned}$$

where h^* is the CGA haplotype, h_k and h_l are the pair of haplotypes in H, and the environmental variable X is a Bernoulli random variable with 0.3 success probability. The parameters β_1, β_2, and β_3

TABLE I. Estimates of haplotype and smoking-duration effects for haplotype CGG[a]

				Codominant model	
Variable	Recessive model	Dominant model	Additive model	Additive	Recessive
Haplotype	−0.119 (0.115)	−0.348 (0.135)***	−0.169 (0.078)**	−0.331 (0.142)**	0.283 (0.211)
Dur1	0.010 (0.097)	−0.135 (0.158)	−0.114 (0.134)	−0.134 (0.158)	
Dur2	−0.004 (0.121)	−0.091 (0.199)	−0.041 (0.169)	−0.091 (0.199)	
Dur3	0.022 (0.126)	−0.259 (0.220)	−0.127 (0.180)	−0.259 (0.220)	
Dur4	0.260 (0.106)**	0.004 (0.177)	0.118 (0.148)	0.004 (0.177)	
Hap * Dur1	0.171 (0.131)	0.254 (0.164)	0.158 (0.088)*	0.203 (0.176)	−0.091 (0.262)
Hap * Dur2	−0.006 (0.170)	0.105 (0.207)	0.030 (0.113)	0.123 (0.222)	−0.166 (0.334)
Hap * Dur3	0.093 (0.173)	0.379 (0.227)*	0.153 (0.119)	0.384 (0.241)	−0.395 (0.349)
Hap * Dur4	0.099 (0.143)	0.353 (0.182)*	0.149 (0.097)	0.351 (0.194)*	−0.348 (0.285)
Log-likelihood	−9,219.43	−9,216.61	−9,217.61	−9,215.87	

[a]Standard error estimates are shown in parentheses. Dur1, Dur2, Dur3, and Dur4 denote, respectively, no active but ETS after 18, $\leq$10 years, 11–20 years, and $>$20 years of smoking. Under codominant model, genetic effects consist of additive and recessive effects, and main effects of smoking duration are shown in left column.

*$P<0.10$.

**$P<0.05$.

***$P<0.01$.

TABLE II. Estimates of haplotype and smoking-duration effects for haplotype CGA[a]

				Codominant model	
Variable	Recessive model	Dominant model	Additive model	Additive	Recessive
Haplotype	0.337 (0.174)*	0.202 (0.111)*	0.197 (0.085)**	0.171 (0.115)	0.079 (0.239)
Dur1	0.107 (0.087)	0.218 (0.105)**	0.216 (0.101)**	0.218 (0.105)**	
Dur2	0.004 (0.109)	0.003 (0.135)	0.013 (0.130)	0.003 (0.135)	
Dur3	0.065 (0.113)	0.181 (0.136)	0.157 (0.131)	0.181 (0.136)	
Dur4	0.315 (0.095)***	0.373 (0.115)***	0.373 (0.111)***	0.373 (0.115)***	
Hap * Dur1	−0.356 (0.218)*	−0.296 (0.127)**	−0.246 (0.097)**	−0.257 (0.134)*	0.047 (0.302)
Hap * Dur2	−0.097 (0.270)	−0.018 (0.162)	−0.030 (0.123)	0.001 (0.171)	−0.098 (0.377)
Hap * Dur3	−0.096 (0.277)	−0.256 (0.167)	−0.171 (0.125)	−0.267 (0.178)	0.324 (0.395)
Hap * Dur4	−0.190 (0.231)	−0.154 (0.137)	−0.128 (0.104)	−0.131 (0.145)	0.013 (0.323)
Log-likelihood	−9,217.89	−9,216.83	−9,215.99	−9,214.98	

[a]Standard error estimates are shown in parentheses. Dur1, Dur2, Dur3, and Dur4 denote, respectively, no active but ETS after 18, $\leq$10 years, 11–20 years, and $>$20 years of smoking. Under codominant model, genetic effects consist of additive and recessive effects, and main effects of smoking duration are shown in left column.
*$P<0.10$.
**$P<0.05$.
***$P<0.01$.

are the log odds ratios corresponding to the main effect of the CGA haplotype, the main effect of the environmental variable, and the haplotype-environment interaction, respectively. We chose $\alpha = -4.7$ and -3.1 to yield disease rates of approximately 1% and 5%. For making inferences on β_1, we set $\beta_2 = \beta_3 = 0.3$, and varied β_1 from -0.3 to 0.3; for making inferences on β_3, we set $\beta_1 = \beta_2 = 0.3$, and varied β_3 from -0.3 to 0.3. We selected case-control samples with $n_1 = n_0 = 250$, 500, or 1,000. In the analysis, we assumed unknown population totals, and used the algorithm described in Appendix C to calculate MLEs based on (7) and estimate their variances. The results of these studies are summarized in Table III.

The estimators of the haplotype effect and haplotype-environment interaction have little biases. The variance estimators accurately reflect the true variations of the parameter estimators. The confidence intervals have correct coverage probabilities. The association tests have proper type 1 errors and reasonable powers. As expected, the powers for detecting haplotype effects are higher than those for detecting haplotype-environment interactions.

We also conducted simulation studies for the case-cohort design. We generated times to disease occurrence from the proportional hazards model

$$\lambda(t|X,H) = 2t \exp[\beta_1\{I(h_k = h^*) + I(h_l = h^*)\} + \beta_2 X + \beta_3\{I(h_k = h^*) + I(h_l = h^*)\}X]$$

where X and H have the same distributions as in the case-control studies, and β_1, β_2, and β_3 pertain to the log hazard ratios. We chose the values of β_1, β_2, and β_3 in the same manner as in the case-control studies. We considered cohort studies of 5,000 individuals, and chose censoring distributions to produce an average of 250 cases per study. We genotyped all cases and 250, 500, or 1,250 controls in a study. We calculated the MLEs of the log hazard ratios using the algorithm of Appendix D, and estimated their variances by the profile likelihood method. The results reported in Table IV show that the proposed methods perform well. The fact that the variances of the MLEs decrease slowly as the number of controls increases indicates that the case-cohort design is highly cost-effective.

DISCUSSION

All the numerical results presented in this article pertain to the comparison of a target haplotype with all other haplotypes. Another type of analysis is to compare several haplotypes to a reference haplotype within the same model. One may also choose a set of haplotypes as the target, or compare several sets of haplotypes. Furthermore, one may wish to include haplotypes from different genes in the same model. We can modify our methods and algorithms to perform all these types of analyses.

It is desirable to adjust for the effects of multiple comparisons when considering several haplo-

TABLE III. Simulation results for case-control studies[a]

			Haplotype effect					Haplotype-environment interaction				
$n_0 = n_1$	Disease rate	Effect size	Bias	SE	SEE	CP	Power	Bias	SE	SEE	CP	Power
250	1%	−0.3	−0.004	0.181	0.181	0.952	0.383	−0.007	0.221	0.221	0.953	0.275
		0.0	0.000	0.171	0.172	0.951	0.049	−0.005	0.204	0.205	0.950	0.050
		0.3	0.002	0.167	0.166	0.951	0.446	−0.007	0.195	0.196	0.954	0.322
	5%	−0.3	−0.002	0.179	0.180	0.949	0.381	0.010	0.218	0.222	0.954	0.242
		0.0	0.002	0.169	0.171	0.953	0.047	−0.007	0.205	0.206	0.953	0.047
		0.3	0.010	0.165	0.166	0.951	0.470	−0.040	0.196	0.197	0.946	0.260
500	1%	−0.3	−0.001	0.128	0.127	0.951	0.662	0.000	0.155	0.155	0.949	0.491
		0.0	0.002	0.119	0.121	0.955	0.045	−0.004	0.145	0.144	0.948	0.052
		0.3	0.003	0.116	0.117	0.952	0.740	−0.009	0.136	0.137	0.954	0.557
	5%	−0.3	0.002	0.125	0.127	0.953	0.654	0.015	0.158	0.155	0.946	0.452
		0.0	0.007	0.121	0.120	0.952	0.048	−0.005	0.145	0.145	0.950	0.050
		0.3	0.012	0.117	0.117	0.950	0.759	−0.041	0.138	0.138	0.941	0.465
1,000	1%	−0.3	0.001	0.089	0.090	0.956	0.921	0.002	0.110	0.109	0.952	0.778
		0.0	0.003	0.085	0.085	0.949	0.051	−0.001	0.102	0.101	0.950	0.050
		0.3	0.003	0.082	0.082	0.950	0.956	−0.010	0.097	0.097	0.946	0.850
	5%	−0.3	0.001	0.088	0.089	0.953	0.921	0.015	0.111	0.109	0.943	0.744
		0.0	0.005	0.086	0.085	0.946	0.054	−0.008	0.104	0.102	0.950	0.050
		0.3	0.012	0.083	0.082	0.941	0.967	−0.043	0.098	0.098	0.930	0.744

[a]Bias and SE are bias and standard error of MLE for effect size. SEE, mean of standard error estimator for MLE. CP, coverage probability of 95% confidence interval for effect size. Power pertains to 0.05-level Wald test of null hypothesis of zero effect size. Each entry is based on 5,000 replictates.

TABLE IV. Simulation results for case-cohort studies[a]

		Haplotype effect					Haplotype-environment interaction				
Controls	Effect size	Bias	SE	SEE	CP	Power	Bias	SE	SEE	CP	Power
250	−0.3	−0.008	0.171	0.176	0.957	0.404	−0.009	0.206	0.204	0.951	0.319
	0.0	−0.006	0.149	0.154	0.958	0.042	−0.004	0.187	0.185	0.950	0.050
	0.3	−0.003	0.131	0.134	0.953	0.606	−0.001	0.173	0.172	0.949	0.414
500	−0.3	−0.007	0.158	0.161	0.953	0.474	−0.009	0.205	0.203	0.950	0.320
	0.0	−0.005	0.136	0.139	0.954	0.046	−0.004	0.186	0.185	0.951	0.049
	0.3	−0.002	0.119	0.121	0.954	0.696	−0.001	0.172	0.171	0.950	0.418
1,000	−0.3	−0.007	0.151	0.152	0.954	0.516	−0.009	0.205	0.202	0.948	0.323
	0.0	−0.005	0.130	0.131	0.952	0.048	−0.004	0.185	0.183	0.950	0.050
	0.3	−0.002	0.113	0.113	0.948	0.743	−0.001	0.171	0.169	0.951	0.430

[a]Bias and SE are bias and standard error of MLE for effect size. SEE, mean of standard error estimator for MLE. CP, coverage probability of 95% confidence interval for effect size. Power pertains to 0.05-level Wald test of null hypothesis of zero effect size. Each entry is based on 5,000 replictates.

type configurations in the same study. This is especially important in genome-wide studies, which involve hundreds or thousands of haplotypes. The Bonferroni correction would be overly conservative because of the correlation of haplotypes both within and between regions. Proper multiple testing adjustments can be achieved by permuting the data or by simulating the joint distribution of the test statistics [see Lin, 2005]. These adjustments require the use of appropriate test statistics, such as those presented in this article.

The proposed EM algorithms are relatively fast and have good convergence properties. Naturally, the computing time increases with the observed number of haplotypes. When the number of SNPs is large, the partition-ligation method of Niu et al.

[2002] and Qin et al. [2002] and other modifications can be adapted to improve computation efficiency.

Latent population substructure or stratification may bias the results in association studies. There exist several statistical methods to adjust for the effects of population stratification with the aid of genomic markers that are informative about the population substructure. It is fairly straightforward to incorporate genomic markers into our likelihood framework so as to adjust for population stratification.

This article is concerned with studies of unrelated individuals. Many genetic studies involve multiple family members or relatives. This is particularly the case if association studies are embedded in linkage studies. Haplotype ambiguity can potentially be reduced by using the genotype information from related individuals. Inferences on haplotype effects need to account for the intraclass correlation. We are currently developing methods for inferring haplotype-disease associations in family studies.

We developed a general computer program that implements the proposed methods. This program is posted on the website: http://www.bios.unc.edu/~lin.

REFERENCES

Akaike H. 1985. Prediction and entropy. In: Atkinson AC, Fienberg SE, editors. A celebration of statistics. New York: Springer. p 1–24.

Akey J, Jin L, Xiong M. 2001. Haplotypes vs. single marker linkage disequilibrium tests: what do we gain? Eur J Hum Genet 9: 291–300.

Botstein D, Risch N. 2003. Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. Nat Genet [Suppl] 33: 228–237.

Clark AG. 1990. Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 7:111–122.

Cox DR. 1972. Regression models and life-tables (with discussion). J R Stat Soc B 34:187–220.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). J R Stat Soc B 39:1–38.

Diggle PJ, Heagerty P, Liang K-Y, Zeger SL. 2002. Analysis of longitudinal data, 2nd ed. Oxford: Oxford University Press.

Epstein MP, Satten GA. 2003. Inference on haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet 73:1316–1329.

Excoffier L, Slatkin M. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927.

Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork N. 2001. Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. Genome Res 11:143–151.

International Hapmap Consortium. 2003. The international HapMap project. Nature 426:789–796.

International SNP Map Working Group. 2001. A map of human genome sequence variation containing 14.2 million single nucleotide polymorphisms. Nature 409:928–933.

Kalbfleisch JD, Prentice RL. 2002. The statistical analysis of failure time data. Hoboken, NJ: Wiley.

Kraft P, Cox DG, Paynter RA, Hunter D, De Vivo I. 2005. Accounting for haplotype uncertainty in matched association studies: a comparison of simple and flexible techniques. Genet Epidemiol 28:261–272.

Lake SL, Lyon H, Tantisira K, Silverman EK, Weiss ST, Laird NM, Schaid DJ. 2003. Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. Hum Hered 55:56–65.

Lin DY. 2004. Haplotype-based association analysis in cohort studies of unrelated individuals. Genet Epidemiol 26: 255–264.

Lin DY. 2005. An efficient Monte Carlo approach to assessing statistical significance in genomic studies. Bioinformatics 21: 781–787.

Lin DY, Zeng D. 2006. Likelihood-based inference on haplotype effects in genetic association studies (with discussion). J Am Stat Assoc (in press).

McCullagh P, Nelder JA. 1989. Generalized linear models, 2nd ed. New York: Chapman and Hall.

Millikan R, Eaton A, Worley K, Biscocho L, Hodgson E, Huang WY, Geradts J, Iacocca M, Cowan D, Conway K, Dressler L. 2003. HER2 codon 655 polymorphism and risk of breast cancer in African Americans and whites. Breast Cancer Res Treat 79:355–364.

Morris RW, Kaplan NL. 2002. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genet Epidemiol 23:221–233.

Murphy SA, van der Vaart AW. 2000. On profile likelihood (with discussion). J Am Stat Assoc 95:449–465.

Newman B, Moorman P, Millikan R, Qaqish BF, Geradts J, Aldrich TE, Liu ET. 1995. The Carolina Breast Cancer Study: integrating population-based epidemiology and molecular biology. Breast Cancer Res. Treat. 34:51–60.

Niu T, Qin ZS, Xu X, Liu JS. 2002. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 70:157–169.

Pettitt AN. 1984. Proportional odds models for survival data and estimates using ranks. Appl Stat 26:183–214.

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

Qin ZS, Niu T, Liu JS. 2002. Partition-ligation-expectation maximization algorithm for haplotype inference with single-nucleotide polymorphisms. Am J Hum Genet 71: 1242–1247.

Risch N. 2000. Searching for genetic determinants in the new millennium. Nature 405:847–856.

Satten GA, Epstein MP. 2004. Comparison of prospective and retrospective methods for haplotype inference in case-control studies. Genet Epidemiol 27:192–201.

Schaid DJ. 2004. Evaluating associations of haplotypes with traits. Genet Epidemiol 27:348–364.

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. 2002. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70:425–434.

Scott AJ, Wild CJ. 1997. Fitting regression models to case-control data by maximum likelihood. Biometrika 84:57–71.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989.

Stram DO, Pearce CL, Bretsky P, Freedman M, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Thomas DC. 2003. Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. Hum Hered 55:179–190.

Weir BS. 1996. Genetic data analysis II. Sunderland, MA: Sinauer Associates, Inc.

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG. 2002. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 53:79–91.

Zeng D, Lin DY, Avery CL, North KE, Bray MS. 2005. Efficient semiparametric estimation of haplotype-disease associations in two-stage cohort studies. Technical report. Department of Biostatistics, University of North Carolina at Chapel Hill.

Zhang S, Pakstis AJ, Kidd KK, Zhao H. 2001. Comparisons of two methods for haplotype reconstruction and haplotype frequency estimation from population data. Am J Hum Genet 69:906–912.

Zhao LP, Li SS, Khalid N. 2003. A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. Am J Hum Genet 72:1231–1250.

APPENDIX A

EM ALGORITHM FOR MAXIMIZING (5)

We provide an EM algorithm for the maximization of (5) by regarding H_i as missing data. The complete-data likelihood is $\prod_{i=1}^{n} P(Y_i|X_i, H_i; \theta)P(H_i; \gamma)$. In the M-step of the EM algorithm, we solve the following equations for θ and γ via the Newton-Raphson algorithm:

$$\begin{aligned} \sum_{i=1}^{n} E\{\partial \log P(Y_i|X_i, H_i; \theta)/\partial\theta | Y_i, X_i, G_i\} = 0, \\ \sum_{i=1}^{n} E\{\partial \log P(H_i; \gamma)/\partial\gamma | Y_i, X_i, G_i\} = 0. \end{aligned} \tag{A1}$$

The conditional expectations in the above expressions are calculated in the E-step as follows: for any random variable V_i,

$$E(V_i|Y_i, X_i, G_i) = \frac{\sum_{H_i \in \mathcal{S}(G_i)} V_i P(Y_i|X_i, H_i; \theta)P(H_i; \gamma)}{\sum_{H_i \in \mathcal{S}(G_i)} P(Y_i|X_i, H_i; \theta)P(H_i; \gamma)} \tag{A2}$$

where θ and γ are evaluated at their current estimates.

Under condition (3) with $\rho \geq 0$, the estimate of γ can be obtained in a closed form. Specifically, let B_i be a Bernoulli variable with success probability ρ, and let Q_{1i} and Q_{2i} be discrete random variables with probability functions $P(Q_{1i} = (h_k, h_k)) = \pi_k$ and $P(Q_{2i} = (h_k, h_l)) = \pi_k\pi_l$. Then $B_iQ_{1i} + (1-B_i)Q_{2i}$ has the same distribution as H_i, and we can regard B_i, Q_{1i}, and Q_{2i} instead of H_i as missing. With this data augmentation, the complete-data likelihood is proportional to

$$\prod_{i=1}^{n} \left\{ P(Y_i|X_i, H_i; \theta)\rho^{B_i}(1-\rho)^{1-B_i} \prod_{k=1}^{K} \pi_k^{B_i I(Q_{1i}=(h_k,h_k))} \prod_{k,l=1}^{K} (\pi_k\pi_l)^{(1-B_i)I(Q_{2i}=(h_k,h_l))} \right\}.$$

Then (A1) has an explicit solution for γ, which consists of ρ and π_k,

$$\hat{\rho} = n^{-1} \sum_{i=1}^{n} E(B_i|Y_i, X_i, G_i)$$

and

$$\widehat{\pi}_k = c^{-1} \sum_{i=1}^{n} \left[E\{B_i I(Q_{1i} = (h_k, h_k)) | Y_i, X_i, G_i\} + 2 \sum_{l=1}^{K} E\{(1 - B_i) I(Q_{2i} = (h_k, h_l)) | Y_i, X_i, G_i\} \right]$$

where c is a normalizing constant such that $\sum_k \widehat{\pi}_k = 1$.

APPENDIX B

EM ALGORITHM FOR MAXIMIZING (6)

We provide an EM algorithm by treating H_i as missing data for all individuals, and X_i as missing data for those not selected. For the latter individuals, we attach X_j and H_j to Y_j ($j = n+1, \ldots, N$). In the M-step, we solve the following equations for θ and γ,

$$\sum_{i=1}^{n} E\{\partial \log P(Y_i | X_i, H_i; \theta)/\partial\theta | Y_i, X_i, G_i\} + \sum_{j=n+1}^{N} E\{\partial \log P(Y_j | X_j, H_j; \theta)/\partial\theta | Y_j\} = 0,$$

$$\sum_{i=1}^{n} E\{\partial \log P(H_i; \gamma)/\partial\gamma | Y_i, X_i, G_i\} + \sum_{j=n+1}^{N} E\{\partial \log P(H_j; \gamma)/\partial\gamma | Y_j\} = 0.$$

In addition, we estimate $P(X \mid G)$ by

$$\frac{\sum_{i=1}^{n} I(X_i = X, G_i = G) + \sum_{j=n+1}^{N} E\{I(X_j = X, G_j = G) | Y_j\}}{\sum_{i=1}^{n} I(G_i = G) + \sum_{j=n+1}^{N} E\{I(G_j = G) | Y_j\}}.$$

The conditional expectations are calculated in the E-step: for $i = 1, \ldots, n$, the conditional expectations are given in (A2); for $j = n+1, \ldots, N$,

$$E(V_j | Y_j) = \frac{\sum_{G,X} \sum_{H \in \mathcal{S}(G)} V_j P(Y_j | X, H; \theta) P(H; \gamma) P(X | G)}{\sum_{G,X} \sum_{H \in \mathcal{S}(G)} P(Y_j | X, H; \theta) P(H; \gamma) P(X | G)}$$

where θ, γ, and $P(X \mid G)$ are evaluated at their current estimates. Under condition (3) with $\rho \geq 0$, the data augmentation $H_i = B_i Q_{1i} + (1 - B_i) Q_{2i}$ introduced in Appendix A can be used to yield an explicit estimate of γ.

APPENDIX C

CALCULATIONS OF MLES OF β AND γ BASED ON (7)

It is desirable to calculate the MLEs of β and γ by profiling the nuisance conditional density functions $P(X \mid G)$ out of the likelihood given in (7), i.e., by maximizing (7) over $P(X \mid G)$ first. Although $P(X \mid G)$ are potentially infinite-dimensional, profiling (7) over $P(X \mid G)$ is tantamount to profiling the following

function over the small set of parameters $\{\mu_G\}$

$$\sum_{i=1}^{n}\left\{Y_i\log\sum_{H_i\in\mathcal{S}(G_i)}e^{\beta'\mathcal{Z}(X_i,H_i)}P(H_i;\,\gamma)+(1-Y_i)\log\sum_{H_i\in\mathcal{S}(G_i)}P(H_i;\,\gamma)\right\}$$
$$+\sum_{i=1}^{n}(1-Y_i)\log\sum_{G}\mu_G-\sum_{i=1}^{n}\sum_{G}I(G_i=G)\log\left\{\sum_{H_i\in S(G_i)}e^{\beta'\mathcal{Z}(X_i,H_i)}P(H_i;\,\gamma)-\mu_G+n_1^{-1}m_G\sum_{G}\mu_G\right\}$$

where m_G is the number of times $G_i = G$ in the sample [Lin and Zeng, 2006]. The covariance matrix for the MLEs of β and γ can be estimated by the sandwich estimator or the profile likelihood method.

If X is independent of G, then profiling (7) over $P(X \mid G)$ is equivalent to profiling the following function over the scalar parameter μ:

$$\sum_{i=1}^{n}\log\left[\sum_{H_i\in S(G_i)}\frac{\exp\{Y_i\beta'\mathcal{Z}(X_i,H_i)\}P(H_i;\,\gamma)p^{Y_i}\{(1-p)\mu\}^{1-Y_i}}{\sum_{Y=0}^{1}\sum_{H}\exp\{Y\beta'\mathcal{Z}(X_i,H)\}P(H;\,\gamma)p^{Y}\{(1-p)\mu\}^{1-Y}}\right]$$

where $p = n_1/n$ [Lin and Zeng, 2006]. This expression is the log-likelihood function for a cross-sectional study in which the conditional distribution of Y_i and H_i given X_i has the probability density function

$$\tilde{P}(Y_i,X_i,H_i;\,\theta,\gamma)=\frac{\exp\{Y_i\beta'\mathcal{Z}(X_i,H_i)\}P(H_i;\,\gamma)p^{Y_i}\{(1-p)\mu\}^{1-Y_i}}{\sum_{Y=0}^{1}\sum_{H}\exp\{Y\beta'\mathcal{Z}(X_i,H)\}P(H;\,\gamma)p^{Y}\{(1-p)\mu\}^{1-Y}}$$

and in which G_i instead of H_i is observed. Thus, we can use the EM algorithm of Appendix A upon replacing $P(Y_i \mid X_i,H_i;\,\theta)\,P(H_i;\,\gamma)$ with $\tilde{P}(Y_i,X_i,H_i;\,\theta,\gamma)$. In addition, the covariance matrix of the MLEs of θ and γ can be estimated by the inverse of the observed Fisher information matrix. Under condition (3) with $\rho \geq 0$, we can use the data augmentation $H_i = B_iQ_{1i}+(1-B_i)Q_{2i}$ described in Appendix A. The M-step can then be simplified if we express $\tilde{P}(Y,\,X,\,H = BQ_1 + (1-B)Q_2;\,\theta,\,\gamma)$ as

$$\frac{\exp\{\xi_0Y+Y\beta'\mathcal{Z}(X,H)+\sum_{k=1}^{K}\xi_kW_k\}}{\sum_{Y,B,Q_1,Q_2}\exp\{\xi_0Y+Y\beta'\mathcal{Z}(X,H)+\sum_{k=1}^{K}\xi_kW_k\}}$$

where $W_k = BI(Q_1 = (h_k,h_k))+(1-B)\,\Sigma_l I(Q_2 = (h_k,\,h_l))\,+\,I(Q_2 = (h_l,\,h_k))$ $(k = 1,\ldots,\,K)$, and work with the new parameters $(\beta,\,\xi_0,\,\xi_1,\ldots,\xi_K)$. Because the above density function yields a concave log-likelihood, the corresponding MLEs are unique and can be easily obtained by the Newton-Raphson or EM algorithm.

APPENDIX D

EM ALGORITHMS FOR MAXIMIZING (10) AND (11)

We provide an EM algorithm for the maximization of (11) under model (8) by regarding H_i as missing for those individuals selected for genotyping, and G_i as missing for those not selected. Note that (10) is a special case of (11) with $n_c = n$. In the M-step of the EM algorithm, we estimate β and γ by solving the

following equations

$$\sum_{i=1}^{n} \Delta_i \left[\widehat{E}\{\mathcal{Z}(X_i, H_i)\} - \frac{\sum_{j=1}^{n} I(\tilde{Y}_j \geq \tilde{Y}_i)\widehat{E}\{\mathcal{Z}(X_j, H_j)e^{\beta' \mathcal{Z}(X_j, H_j)}\}}{\sum_{j=1}^{n} I(\tilde{Y}_j \geq \tilde{Y}_i)\widehat{E}\{e^{\beta' \mathcal{Z}(X_j, H_j)}\}} \right] = 0,$$

$$\sum_{i=1}^{n} \widehat{E}\{\partial \log P(H_i;\, \gamma)/\partial \gamma\} = 0$$

and estimate $\Lambda_0(t)$ by

$$\sum_{i=1}^{n} \frac{I(\tilde{Y}_i \leq t)\Delta_i}{\sum_{j=1}^{n} I(\tilde{Y}_j \geq \tilde{Y}_i)\widehat{E}\{e^{\beta' \mathcal{Z}(X_j, H_j)}\}}.$$

The conditional expectations $\widehat{E}(V_i)$ are calculated in the E-step: for $i = 1, \ldots, n_c$,

$$\widehat{E}(V_i) = \frac{\sum_{H_i \in S(G_i)} V_i \exp\{\Delta_i \beta' \mathcal{Z}(X_i, H_i) - e^{\beta' \mathcal{Z}(X_i, H_i)} \Lambda_0(\tilde{Y}_i)\} P(H_i;\, \gamma)}{\sum_{H_i \in S(G_i)} \exp\{\Delta_i \beta' \mathcal{Z}(X_i, H_i) - e^{\beta' \mathcal{Z}(X_i, H_i)} \Lambda_0(\tilde{Y}_i)\} P(H_i;\, \gamma)}$$

where β, γ, and Λ_0 are evaluated at their current estimates; for $j = n_c+1, \ldots, n$, the summation over $H_i \in S(G_i)$ is replaced by the summation over all possible values of H_i.

The data augmentation of Appendix A can again be used to obtain an explicit estimate of γ under condition (3) with $\rho \geq 0$. EM algorithms can also be developed for the class of models given in (9) [see Zeng et al., 2005].