Estimating Haplotype-Disease Associations With Pooled Genotype Data

D. Zeng and D.Y. Lin*

Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina

The genetic dissection of complex human diseases requires large-scale association studies which explore the population associations between genetic variants and disease phenotypes. DNA pooling can substantially reduce the cost of genotyping assays in these studies, and thus enables one to examine a large number of genetic variants on a large number of subjects. The availability of pooled genotype data instead of individual data poses considerable challenges in the statistical inference, especially in the haplotype-based analysis because of increased phase uncertainty. Here we present a general likelihood-based approach to making inferences about haplotype-disease associations based on possibly pooled DNA data. We consider cohort and case-control studies of unrelated subjects, and allow arbitrary and unequal pool sizes. The phenotype can be discrete or continuous, univariate or multivariate. The effects of haplotypes on disease phenotypes are formulated through flexible regression models, which allow a variety of genetic hypotheses and gene-environment interactions. We construct appropriate likelihood functions for various designs and phenotypes, accommodating Hardy-Weinberg disequilibrium. The corresponding maximum likelihood estimators are approximately unbiased, normally distributed, and statistically efficient. We develop simple and efficient numerical algorithms for calculating the maximum likelihood estimators and their variances, and implement these algorithms in a freely available computer program. We assess the performance of the proposed methods through simulation studies, and provide an application to the Finland-United States Investigation of NIDDM Genetics Study. The results show that DNA pooling is highly efficient in studying haplotype-disease associations. As a by-product, this work provides valid and efficient methods for estimating haplotype-disease associations with unpooled DNA samples. *Genet. Epidemiol.* 28:70–82, 2005. © 2004 Wiley-Liss, Inc.

Key words: association tests; case-control studies; cohort studies; DNA pooling; EM algorithm; gene-environment interactions; haplotype analysis; Hardy-Weinberg equilibrium; linkage disequilibrium; maximum likelihood; pooled DNA samples; SNPs

Grant sponsor: National Institute 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 13 July 2004; Accepted 5 September 2004
Published online 22 November 2004 in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/gepi.20040

INTRODUCTION

Complex human diseases, such as hypertension, diabetes, and schizophrenia, are influenced by multiple genetic variants and environmental exposures, as well as gene-environment interactions. The population associations between genetic variants and disease phenotypes hold great promise for understanding the genetic basis of these diseases. With the availability of dense single-nucleotide polymorphism (SNP) maps across the genome, there is a proliferation of SNP-based association studies [Botstein and Risch, 2003].

Since haplotypes incorporate linkage disequilibrium information from multiple markers, the use of haplotypes tends to produce more powerful tests of associations than the use of individual SNPs [Akey et al., 2001; Fallin et al., 2001; Morris and Kaplan, 2002; Zaykin et al., 2002]. Due to unknown gametic phase, however, haplotypes in general cannot be determined with certainty. It is highly challenging to perform a haplotype analysis based on unphased genotype data. A large number of papers have been published on the estimation of haplotype frequencies and haplotype-phenotype associations; see Zeng and Lin [2004] for a review.

Because of disease heterogeneity, modest genetic effects, and gene-environment interactions, association studies to identify complex disease genes require hundreds or even thousands of subjects so as to achieve sufficient power. Despite the continuing improvement in genotyping efficiency, it is still very costly to genotype a large number of subjects, especially in a comprehensive genome scan. One strategy to reduce genotyping costs is DNA pooling, such that pools of DNA samples rather than individual samples are assayed [Arnheim et al., 1985; Pacek et al., 1993; Barcellos et al., 1997; Daniels et al., 1998; Risch and Teng 1998; Shaw et al., 1998; Amos et al., 2000; Sasaki et al., 2001; Bansal et al., 2002; Barratt et al., 2003; Mohlke et al., 2002; Sham et al., 2002]. An additional advantage of DNA pooling is that the amount of DNA required from each person for each genotype can be dramatically reduced [Mohlke et al., 2002]. Measuring pooled genotypes instead of individual genotypes increases the haplotype uncertainty and further complicates the statistical analysis.

Several authors have studied the estimation of haplotype frequencies based on pooled DNA data. For two SNPs and two or three subjects per pool, Wang et al. [2003] assessed the cost-effectiveness of DNA pooling in the estimation of haplotype frequencies. For multiple SNPs and multiple subjects per pool, Ito et al. [2003] developed a computer program which estimates the haplotype frequencies by the EM algorithm [Dempster et al., 1977] and which estimates the variances of estimated haplotype frequencies by the bootstrap method. Yang et al. [2003] provided analytical variance estimators for the estimated haplotype frequencies, and found that pool sizes of three to four appear optimal. All the existing literature assumes Hardy-Weinberg equilibrium. To our knowledge, there is no published work on the estimation of the effects of haplotypes on disease phenotypes based on pooled DNA data.

Here, we develop simple and efficient statistical methods for making inferences about haplotype-disease associations based on potentially pooled DNA data. We accommodate Hardy-Weinberg disequilibrium and allow arbitrary and unequal pool sizes. We construct appropriate likelihood functions for a variety of study designs, disease phenotypes, and association models. We show that the maximum likelihood estimators have desirable theoretical properties. We develop a computer program which efficiently implements the likelihood-based inference procedures. We evaluate the performance of the proposed methods through Monte Carlo simulation and provide an application to a major genetic study of type 2 diabetes.

METHODS

DEFINITIONS AND ASSUMPTIONS

Suppose that we collect genomic DNA samples from n unrelated subjects. In order to reduce cost, time, and labor, we genotype pools of DNA samples rather than individual samples. Specifically, we construct J DNA pools with potentially different sizes. For $j = 1, \ldots, J$, let n_j denote the number of subjects in the jth pool such that $n = \sum_{j=1}^{J} n_j$. We perform quantitative DNA typing on each DNA pool at M linked loci. We focus on the common situation of biallelic loci, and denote the two alleles at each locus by 0 and 1. Following previous authors [Wang et al., 2003; Ito et al., 2003; Yang et al., 2003], we assume that the number of allele copies for each pool at each locus can be accurately determined by the quantitative DNA typing. For $j = 1, \ldots, J$, let G_j denote the pooled genotype for the jth pool, which indicates the number of allele 1 at each of the M loci. If no DNA samples are pooled, then $n_j = 1$ for all $j = 1, \ldots, J$, and we observe the individual genotypes for all study subjects.

For $j = 1, \ldots, J$ and $i = 1, \ldots, n_j$, let $H_{ji} = (h_{ji1}, h_{ji2})$ denote the haplotype pair for the ith subject of the jth pool, where h_{ji1} and h_{ji2} are two specific sequences of zeros and ones. We do not observe the individual haplotype pairs directly, even when the pool sizes are one. Instead, we observe the pooled genotype G_j $(j = 1, \ldots, J)$. By definition,

$$G_j = \sum_{i=1}^{n_j} (h_{ji1} + h_{ji2}).$$

Because allele frequencies are commonly determined by fluorescence intensities in genotyping assays, occasionally one knows for sure that a given allele is present in the pool but cannot be certain about the exact number of allele copies. In some instances, no information is available at all on some locus or loci. Thus, we allow G_j $(j = 1, \ldots, J)$ to contain missing values. Our methods make maximum use of the available data under the assumption that the missingness occurs at random.

For the ith subject in the jth pool, let Y_{ji} be the disease phenotype of interest, and let X_{ji} be the environmental variables or covariates. In association studies, we are interested in estimating the effects of X_{ji} and H_{ji} on Y_{ji}. Such a relationship can be characterized by the conditional density function $P(Y_{ji}|X_{ji}, H_{ji}; \theta)$, where θ is a set of unknown parameters. If one is interested in the effects of a particular haplotype h^* on a binary trait, then $P(Y_{ji}|X_{ji}, H_{ji}; \theta)$ may correspond to a logistic regression model with the linear predictor

$$\begin{aligned} &\alpha + \beta_1\{I(h_{ji1} = h^*) + I(h_{ji2} = h^*)\} \\ &\quad + \beta_2 I(h_{ji1} = h_{ji2} = h^*) + \beta_3^T X_{ji} \\ &\quad + \beta_4^T X_{ji}\{I(h_{ji1} = h^*) + I(h_{ji2} = h^*)\} \\ &\quad + \beta_5^T X_{ji} I(h_{ji1} = h_{ji2} = h^*), \end{aligned} \tag{1}$$

where $I(A)$ is the indicator function for event A, taking the value 1 or 0 dependent on whether A is true or false, α is the intercept, and $\beta_1, \ldots, \beta_5$ are the log odds ratios; for a quantitative trait, a normal linear regression model with the same linear predictor may be used. We obtain recessive, dominant, and additive models by setting $\{\beta_1 = 0, \beta_4 = 0\}$, $\{\beta_2 = -\beta_1, \beta_5 = -\beta_4\}$, and $\{\beta_2 = 0, \beta_5 = 0\}$, respectively.

The selection of DNA samples for pooling is random, and each sample is selected only once. The selection may depend on the phenotype and covariates, but is assumed to be independent of the haplotype given the phenotype and covariates. In addition, the haplotype and covariates are assumed to be independent. The latter assumption is necessary unless one is willing to model the dependence between the haplotype and covariates.

Let K be the total number of distinct haplotypes. For $k = 1, \ldots, K$, let h_k be the kth possible haplotype, and let π_k be the population frequency of h_k; for $k, l = 1, \ldots, K$, let π_{kl} be the probability that the haplotype pair consists of h_k and h_l. All the existing literature on the estimation of haplotype frequencies based on pooled genotype data assumes Hardy-Weinberg equilibrium, such that $\pi_{kl} = \pi_k \pi_l$ $(k, l = 1, \ldots, K)$. It is possible to relax this assumption. In particular, one may consider the following form of Hardy-Weinberg disequilibrium:

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

where ρ is the inbreeding coefficient or fixation index [Weir, 1996, p. 93]. Excess homozygosity and excess heterozygosity arise when $\rho > 0$ and $\rho < 0$, respectively. In the sequel, we denote the probability density function of H_{ji} by $P(H_{ji}; \gamma)$, where γ represents the parameters in the haplotype distribution, which consists of the π_ks and ρ under condition (2).

We will develop valid and efficient statistical methods for estimating haplotype-disease associations under all commonly used study designs. Rigorous proofs of the theoretical results presented in this article require very advanced mathematical arguments. We will omit such proofs, but refer interested readers to Zeng and Lin [2004] for the kind of arguments that are involved.

COHORT STUDIES

In a cohort study, there is a random sample of N subjects. We observe the phenotypes and covariates for all n subjects together with the pooled genotypes for J pools. Some of the pooled genotypes may be missing, either partially or completely. The phenotype can be discrete or continuous, and $P(Y|X, H; \theta)$ normally takes the form of a generalized linear model [McCullagh and Nelder, 1989]. If the phenotype pertains to repeated measures in a longitudinal study, then the generalized linear mixed model [Breslow and Clayton, 1993] can be used.

The likelihood function for the parameters θ and γ based on the data (Y_{ji}, X_{ji}, G_j) $(j = 1, \ldots, J; i = 1, \ldots, n_i)$ is proportional to

$$\prod_{j=1}^{J} \left\{ \sum_{H_j \in S(G_j)} \prod_{i=1}^{n_j} P(Y_{ji}|X_{ji}, H_{ji}; \theta) P(H_{ji}; \gamma) \right\}, \tag{3}$$

where H_j denotes $(H_{j1}, \ldots, H_{jn_j})$, the n_j haplotype pairs in the jth pool, and $S(G_j)$ consists of all possible combinations of n_j haplotype pairs that are compatible with the pooled genotype G_j such that $\sum_{i=1}^{n_j} (h_{ji1} + h_{ji2}) = G_j$. If G_i is missing, either partially or completely, then the set $S(G_i)$ is enlarged accordingly.

By the arguments of Zeng and Lin [2004], it is possible to estimate or identify γ from the observed data under condition (2), if there is a positive probability for the $2n_j$ haplotypes in some pool j to be identical. Furthermore, if θ is uniquely determined from the distribution $P(Y|X, H; \theta)$ for $H = (h_k, h_k)$ $(k = 1, \ldots, K)$, then θ is also identifiable. In particular, the latter is true of the generalized linear model with linear predictor (1). From now on, we assume that both θ and γ are identifiable.

We can maximize the likelihood function given in (3) directly or by using the expectation-maximization (EM) algorithm [Dempster et al., 1977] described in Appendix A. The resultant maximum likelihood estimators (MLEs) are consistent and asymptotically normal. The MLEs are also asymptotically efficient in that they have the smallest variances among all valid estimators of θ and γ, at least in large samples. The covariance matrix for the MLEs can be estimated by the inverse of the observed Fisher information matrix.

CASE-CONTROL STUDIES WITH KNOWN POPULATION TOTALS

In a case-control study, we select certain numbers of cases and controls from a finite population. Suppose that we know the total numbers of cases and controls in the finite population. Let n be the total number of subjects in the case-control sample, and let N be the size of the finite population. For subjects in the case-control sample, the data are represented in the same way as the cohort studies. For the $(N-n)$ subjects not selected, we let Y_m denote the disease status for the mth subject, $m = n+1, \ldots, N$.

The likelihood function based on such incomplete data takes the form

$$\prod_{j=1}^{J}\left\{\sum_{H_j\in S(G_j)}\prod_{i=1}^{n_j}P(Y_{ji}|X_{ji},H_{ji};\theta)P(H_{ji};\gamma)P(X_{ij})\right\} \times \prod_{m=n+1}^{N}\left\{\sum_{X,H}P(Y_m|X,H;\theta)P(H;\gamma)P(X)\right\}, \qquad (4)$$

where $P(X)$ is the probability density function of X, and the second summation is taken over all possible values of X and H. The distribution of X is an infinite-dimensional nuisance parameter when there are continuous components in X. Thus, it is mathematically and computationally more difficult to deal with (4) than (3). In Appendix B, we present an EM algorithm to maximize (4). By the arguments of Zeng and Lin [2004], the MLEs are consistent, asymptotically normal, and asymptotically efficient. The variances for the MLEs of θ and γ can be estimated by the profile likelihood method described in Zeng and Lin [2004]. It is simpler to make inference about θ and γ based on the likelihood ratio statistics than the Wald statistics.

CASE-CONTROL STUDIES WITH UNKNOWN POPULATION TOTALS

Under the traditional case-control design, the population totals of cases and controls are assumed unknown. Since the sampling is conditional on the case-control status, one should use the retrospective likelihood function

$$\prod_{j=1}^{J}\left\{\frac{\sum_{H_j\in S(G_j)}\prod_{i=1}^{n_j}P(Y_{ji}|X_{ji},H_{ji};\theta)P(H_{ji};\gamma)P(X_{ji})}{\prod_{i=1}^{n_j}\sum_{X,H}P(Y_{ji}|X,H;\theta)P(H;\gamma)P(X)}\right\}. \qquad (5)$$

In general, this function may involve non-identifiable parameters, so that inferences about θ and γ would be intractable.

In most case-control studies, the disease of interest is relatively rare. In fact, this is the major reason for the case-control design. For the logistic regression with a rare disease, $P(Y_{ji}|X_{ji},H_{ji};\theta)$ is approximately equal to $\exp\{Y_{ji}\alpha + Y_{ji}\beta^T Z(X_{ji}, H_{ji})\}$, where e^{α} is close to 0, and $Z(X_{ji},H_{ji})$ is a specific function of X_{ji} and H_{ji}. Then (5) becomes

$$\prod_{j=1}^{J}\left[\frac{\sum_{H_j\in S(G_j)}\prod_{i=1}^{n_j}\exp\{Y_{ji}\beta^T Z(X_{ji},H_{ji})\}P(H_{ji};\gamma)P(X_{ji})}{\prod_{i=1}^{n_j}\sum_{X,H}\exp\{Y_{ji}\beta^T Z(X,H)\}P(H;\gamma)P(X)}\right]. \qquad (6)$$

As in the case of (4), this likelihood function involves the distribution of X, which is possibly infinite-dimensional. By extending the arguments of Zeng and Lin [2004], we can show that all the parameters in (6) are identifiable and the MLEs derived from (6) are consistent, asymptotically normal, and asymptotically efficient. In Appendix C, we describe a simple and efficient procedure to obtain the MLEs of β and γ and to estimate their variances. Numerical studies show that the rare disease assumption works well when the disease rate is less than 10% [Zeng and Lin, 2004].

COHORT STUDIES WITH AGE-OF-ONSET PHENOTYPES

If one is interested in the age at onset of a complex disease in a cohort study, then a subject who has not developed the disease during his/her follow-up will have a censored observation on the phenotype, in that the age at onset is beyond the duration of follow-up. For the ith subject in the jth pool, let Y_{ji} denote the potential age-at-onset of disease, H_{ji} the haplotype pair, and X_{ji} the covariates. The proportional hazards model of Cox [1972] specifies that the conditional hazard function of Y_{ji}, given X_{ji} and H_{ji}, takes the form

$$\lambda(y|X_{ji},H_{ji}) = \lambda_0(y)e^{\beta^T Z(X_{ji},H_{ji})} \qquad (7)$$

where λ_0 is a completely arbitrary baseline hazard function, $Z(X_{ji},H_{ji})$ is a specific function of X_{ji} and

H_{ji}, and β is a set of regression parameters pertaining to the log relative-risk.

Denote the potential censoring time on Y_{ji} by C_{ji}, which is assumed to be independent of Y_{ji} conditional on X_{ji}. The data consist of $(\widetilde{Y}_{ji}, \Delta_{ji}, X_{ji}, G_j)$ $(j = 1, \ldots, J; i = 1, \ldots, n_j)$, where $\widetilde{Y}_{ji} = min(Y_{ji}, C_{ji})$ and $\Delta_{ji} = I(Y_{ji} \leq C_{ji})$.

Model (7) is a semiparametric model with latent explanatory variables, in that the baseline hazard function is an infinite-dimensional parameter, and the haplotype pairs are not directly observable. If the individual haplotype pairs were directly measured, then the partial likelihood principle [Cox, 1972, 1975] could be used to estimate the regression parameters β and the cumulative baseline hazard function $\Lambda_0(y) = \int_0^y \lambda_0(t)dt$. The same estimators could also be derived from the so-called nonparametric maximum likelihood method [Bickel et al., 1993], which maximizes the likelihood for the observable data with respect to all the parameters, including the infinite-dimensional ones. With pooled genotype data, the partial likelihood would be intractable, whereas the nonparametric maximum likelihood method still yields simple and efficient estimators. Thus, we maximize the following nonparametric likelihood function over β, γ, and Λ

$$\prod_{j=1}^{J} \left(\sum_{H_j \in S(G_j)} \prod_{i=1}^{n_j} \left[\Lambda\{\widetilde{Y}_{ji}\} e^{\beta^T Z(X_{ji}, H_{ji})} \right]^{\Delta_{ji}} \cdot \right.$$
$$\left. \times \exp\left\{ -\Lambda(\widetilde{Y}_{ji}) e^{\beta^T Z(X_{ji}, H_{ji})} \right\} P(H_{ji}; \gamma) \right), \quad (8)$$

where $\Lambda(t)$ is an increasing step function, and $\Lambda\{\widetilde{Y}_{ji}\}$ is the jump size of $\Lambda(y)$ at $y = \widetilde{Y}_{ji}$, i.e., the value of $\Lambda(y)$ at $y = \widetilde{Y}_{ji}$ minus its value right before $\widetilde{Y}_{ji}$. Denote the resultant estimators by $\widehat{\beta}$, $\widehat{\gamma}$, and $\widehat{\Lambda}$.

It can be shown that $\widehat{\Lambda}$ is an increasing step function with jumps only at the $\widetilde{Y}_{ji}$ associated with $\Delta_{ji} = 1$. Thus, the calculation of $\widehat{\beta}$, $\widehat{\gamma}$, and $\widehat{\Lambda}$ is tantamount to maximizing (8) over β, γ, and the jump sizes of Λ at those time points. We show in Appendix D that this maximization can be carried out efficiently through a simple EM algorithm. The estimators are consistent, asymptotically normal, and asymptotically efficient. The covariance matrix for $\widehat{\beta}$ and $\widehat{\gamma}$ can be obtained by the inverse of the observed Fisher information matrix for β, γ, and the jump sizes of Λ, or by the profile likelihood method.

We have implicitly assumed that the covariates are time-invariant. If the X_{ji} depend on time, then we replace $\Lambda(\widetilde{Y}_{ji}) e^{\beta^T Z(X_{ji}, H_{ji})}$ in (8) and (D1) with $\int_0^{\widetilde{Y}_{ji}} e^{\beta^T Z(X_{ji}(y), H_{ji})} d\Lambda(y)$. The rest of the formulas remain the same.

To accommodate non-proportional hazards relationships, we consider the following class of linear transformation models

$$P(Y_{ji} \leq y | X_{ji}, H_{ji}) = Q(\Lambda(y) e^{\beta^T Z(X_{ji}, H_{ji})}),$$

where Q is a specific increasing function, and Λ is an arbitrary increasing function. The choices of $Q(x) = 1 - e^{-x}$ and $Q(x) = 1 - (1 + x)^{-1}$ correspond to the proportional hazards model and proportional odds model [Pettitt, 1982], respectively. For this class of models, the nonparametric likelihood function takes the form

$$\prod_{j=1}^{J} \left(\sum_{H_j \in S(G_j)} \prod_{i=1}^{n_j} \left[\Lambda\{\widetilde{Y}_{ji}\} e^{\beta^T Z(X_{ji}, H_{ji})} Q'(\Lambda(\widetilde{Y}_{ji}) e^{\beta^T Z(X_{ji}, H_{ji})}) \right]^{\Delta_{ji}} \cdot \right.$$
$$\left. \times \left\{ 1 - Q(\Lambda(\widetilde{Y}_{ji}) e^{\beta^T Z(X_{ji}, H_{ji})}) \right\}^{1 - \Delta_{ji}} P(H_{ji}; \gamma) \right),$$

where $Q'(x) = dQ(x)/dx$. The corresponding MLEs of θ, γ, and Λ can be obtained through an optimization algorithm, such as *fminunc* in the Optimization Toolbox of MATLAB. By the arguments of Zeng and Lin [2004], the MLEs are consistent, asymptotically normal, and asymptotically efficient. Inference on θ and γ can be carried in the same manner as in the case of the proportional hazards model.

RESULTS

APPLICATION TO FUSION STUDY

We consider a case-control sample from the Finland-United States Investigation of NIDDM Genetics Study [Valle et al., 1998]. The sample consists of 796 case subjects with type 2 diabetes and 415 control subjects. The subjects were genotyped at five SNPs in a putative disease susceptibility region on chromosome 22. The distances between adjacent SNPs are less than 300 kb. In the sample, 131 cases and 82 controls have missing genotype information for at least one SNP. The observed haplotype frequencies for the cases and controls are given in Table 1 of Zeng and Lin [2004].

This case-control sample is not a DNA pooling study, although DNA pooling was considered for the FUSION Study [Mohlke et al., 2002]. For testing the proposed methods, it is actually preferable to use a data set with individual genotyping, so that DNA pools of various sizes

TABLE I. Estimates of haplotype effects for FUSION study[a]

		Pool size=1			Pool size=2			Pool size=4		
Haplotype	Frequency	Est	SE	*p*-value	Est	SE	*p*-value	Est	SE	*p*-value
01011	0.130	0.056	0.140	0.689	−0.037	0.142	0.794	−0.061	0.153	0.689
01100	0.256	0.352	0.101	<0.001	0.408	0.106	<0.001	0.435	0.119	<0.001
10000	0.014	−0.173	0.402	0.667	0.003	0.418	0.994	0.164	0.458	0.721
10011	0.353	−0.332	0.096	<0.001	−0.323	0.096	0.001	−0.366	0.109	0.001
10100	0.053	0.141	0.223	0.527	−0.007	0.250	0.977	−0.061	0.253	0.810
10110	0.032	0.052	0.247	0.833	0.066	0.247	0.789	0.043	0.249	0.861
11011	0.139	−0.120	0.136	0.377	−0.106	0.143	0.459	−0.101	0.159	0.523
11100	0.011	0.065	0.680	0.924	−1.36	1.43	0.342	−1.65	5.97	0.782

[a] Frequency is estimated population frequencies, assuming 7% disease rate. Est, SE, and *p*-value pertain to estimate of log odds ratio, standard error estimate, and two-sided *p*-value for testing no association, respectively. Additive genetic model is used.

can be considered and the analysis with individual genotype data provides a benchmark for the analysis based on pooled genotype data. We pool the subjects within the cases and controls separately. We choose equal pool sizes except possibly for the last pools. We create the pooled genotype by aggregating the individual genotypes within the same pool, and set the pooled genotype to missing if any individual genotypes in the pool, are missing. We pretend that pooled genotypes rather than individual genotypes were measured, and apply the methods for case-control studies with unknown population totals.

Table I shows the estimates of the effects of haplotypes on the risk of type 2 diabetes with pool sizes of one, two, and four. We report the results for the eight haplotypes with estimated frequencies of at least 1% under the additive genetic model, which was previously shown to be the most appropriate model for this study [Zeng and Lin, 2004]. No matter which pool size was used, one would conclude that haplotype 01100 increases the risk of type 2 diabetes, whereas haplotype 10011 has a protective effect. There are some differences in the point estimate among different pool sizes, especially for the nonsignificant effects. There is only a slight increase in the estimated variance as the pool size increases, except for the last haplotype, which has the lowest frequency.

SIMULATION STUDIES

To evaluate the empirical performance of the proposed methods in practical situations, we conducted Monte Carlo studies mimicking the FUSION Study. We considered 796 cases and 415 controls. We generated haplotypes from the estimated frequencies shown in Table I. We set $\rho = 0.2$ to demonstrate our ability to allow Hardy-Weinberg disequilibrium, although there is little inbreeding in the FUSION data. We considered the additive, dominant, and recessive models for haplotype 01100. We set the log odds ratio β to 0 or 0.35. Also, we set the intercept α to -3.7 so as to yield an approximately 7% disease rate, which is the rate of type 2 diabetes in US and Finnish adults. For each SNP, we set the genotypes to missing according to the proportions of missingness observed in the FUSION data. We considered pooled genotypes with pool sizes of one, two, and four, and assessed the proposed methods for case-control studies with unknown population totals.

The results of these studies are summarized in Table II. The proposed parameter estimators are virtually unbiased. The standard error estimators accurately reflect the true variations of the parameter estimators. The associated confidence intervals have proper coverage probabilities. The Wald tests maintain their type 1 errors near nominal levels. Under the additive model, there is only a small increase of variance and a slight loss of power as the pool size increases. The loss of precision due to pooling is more appreciable under the dominant model and even more so under the recessive model, although the variances increase at a slower pace than the pool size. These results suggest that the proposed inference procedures are adequate for practical use, and that DNA pooling is efficient in association studies, at least under the additive and dominant models.

We also carried out simulation studies for haplotype 10000, which has a frequency of 1.4%. We found that the asymptotic approximations continue to be accurate and pooling to be efficient,

TABLE II. Simulation results under the setup of FUSION study[a]

Model	β	Pool size	Bias	SE	SEE	CP	Power	MSE
Additive	0.0	1	0.0001	0.090	0.089	0.954	0.046	0.0081
		2	0.0003	0.093	0.093	0.951	0.049	0.0086
		4	0.0005	0.096	0.098	0.958	0.042	0.0093
	0.35	1	0.001	0.085	0.085	0.943	0.985	0.0072
		2	0.001	0.088	0.089	0.949	0.978	0.0077
		4	0.004	0.092	0.096	0.957	0.972	0.0085
Dominant	0.0	1	−0.001	0.124	0.122	0.952	0.048	0.0154
		2	−0.003	0.136	0.137	0.951	0.049	0.0186
		4	0.002	0.143	0.147	0.954	0.046	0.0205
	0.35	1	−0.003	0.121	0.121	0.955	0.814	0.0148
		2	0.002	0.135	0.137	0.956	0.735	0.0182
		4	0.003	0.145	0.147	0.955	0.674	0.0211
Recessive	0.0	1	0.0001	0.149	0.150	0.952	0.048	0.0223
		2	−0.0009	0.210	0.211	0.952	0.048	0.0441
		4	−0.0009	0.264	0.267	0.957	0.043	0.0698
	0.35	1	−0.001	0.135	0.140	0.963	0.701	0.0183
		2	−0.004	0.187	0.190	0.952	0.453	0.0348
		4	0.004	0.233	0.239	0.953	0.327	0.0540

[a] Bias and SE are bias and standard error of $\hat{\beta}$, respectively. SEE is mean of standard error estimator. CP is coverage probability of 95% confidence interval. Power is actual type 1 error/power of Wald statistic for testing H_0: $\beta = 0$. MSE is mean squared error. Each entry is based on 1,000 simulated data sets.

although the variances of the parameter estimators tend to be large.

To further investigate the efficiency of DNA pooling, we conducted another set of simulation studies. We generated two SNPs with minor allele frequencies of 0.4 and 0.5, and normalized the linkage disequilibrium coefficient of D such that the frequencies for haplotypes 11, 10, 01, and 00 are $0.2 + 0.2D$, $0.2 - 0.2D$, $0.3 - 0.2D$, and $0.3 + 0.2D$, respectively. We set the inbreeding coefficient to 0.2. We focused on haplotype 11 under the additive, recessive, and dominant models with $\beta = 0.5$. We considered the case-control design with 250 cases and 250 controls or 500 cases and 500 controls, and chose pool sizes of one, two, four, six, and eight. For each configuration, we simulated 1,000 data sets.

The results from the second set of studies are displayed in Figures 1 and 2. Following Yang et al. [2003], we define the relative efficiency for pool size n_j as the mean squared error for pool size 1 multiplied by n_j and divided by the mean squared error for pool size n_j, and consider the pooling efficient if the relative efficiency is greater than 1. We observe that DNA pooling is efficient, especially under the additive and dominant models and under strong linkage disequilibrium. The relative efficiency appears to be higher for the estimation of haplotype-disease associations than for the estimation of haplotype frequencies, especially when the linkage disequilibrium is weak.

DISCUSSION

DNA pooling can offer tremendous savings in genotyping cost and DNA usage. There is a natural concern that the precision of the association analysis may be compromised due to the unavailability of individual genotypes and the increased haplotype uncertainty. Our investigations reveal that pooling is highly efficient in terms of the number of genotyping assays required for the same precision of estimation. The proposed methods enable one to assess the cost-effectiveness of DNA pooling and to conduct the most efficient analysis for the completed DNA pooling studies.

In evaluating the likelihood contribution from the jth pool, one needs to consider all possible combinations of n_j haplotype pairs that are compatible with the pooled genotype. This kind

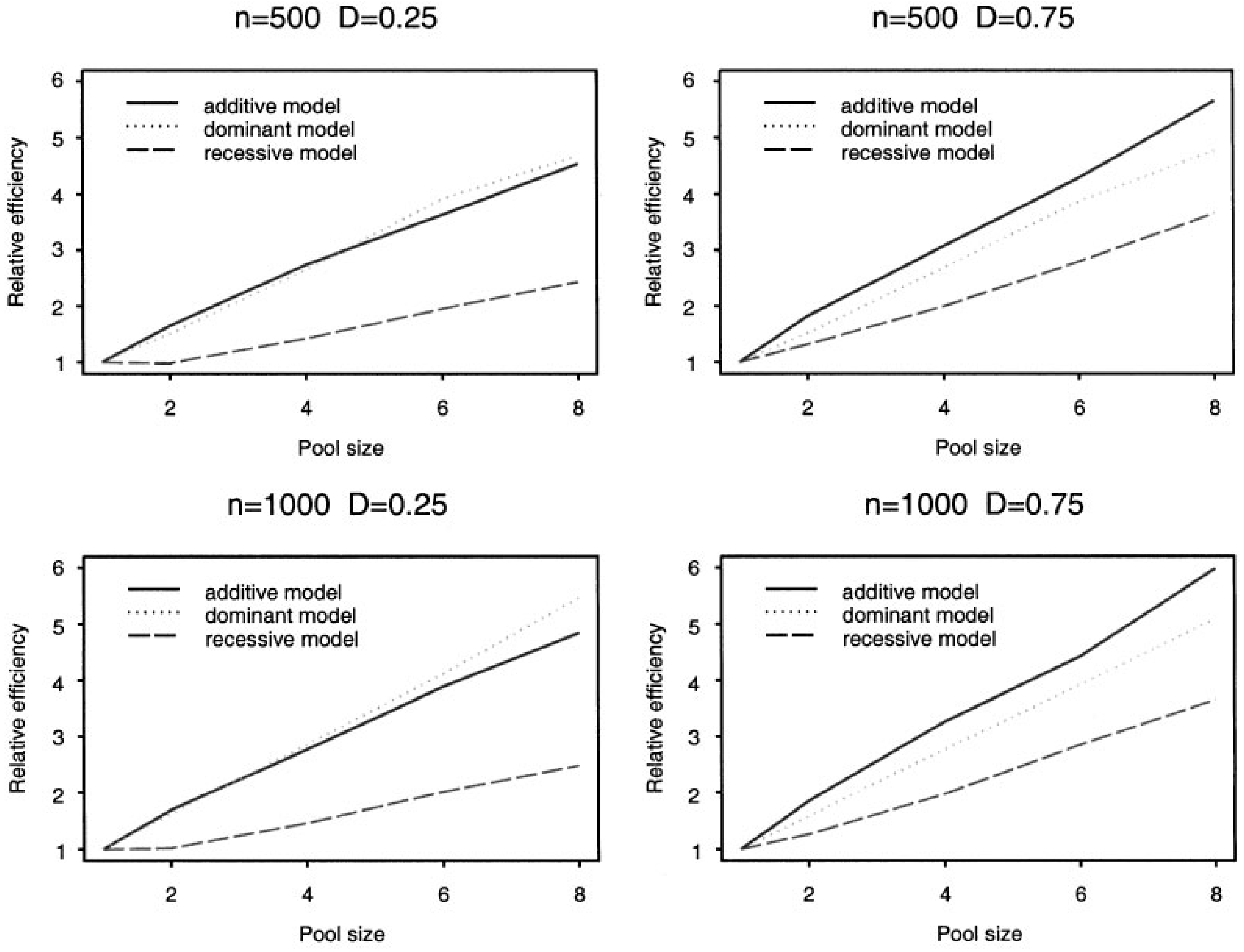

Fig. 1. Relative efficiencies in estimating haplotype effects.

of calculation is demanding. For this reason, the existing literature on the estimation of haplotype frequencies based on pooled genotype data is confined to studies with small numbers of SNPs, small pool sizes, and small sample sizes. The efficient algorithm described in the second paragraph of Appendix A makes it possible to consider larger numbers of SNPs and larger pool sizes, and to allow arbitrarily large sample sizes. It takes 9 min to perform one analysis of the FUSION data with pool size of 4 on an IBM BladeCenter HS20 machine.

Despite our efficient algorithm, we can only handle fairly small pool sizes (<10) with the current computer memory. Some researchers have suggested pools of 30–40 individuals. It is difficult, however, to extract useful haplotype information from such large pools. In order to compensate for the substantial loss of precision due to the use of large pools, one would need to construct a large number of pools, which in turn would require a very large number of subjects. Thus, it may be a good comprise to have relatively small pool sizes, especially as the cost of genotyping continues to decline.

If the pool sizes are all equal to one, then we observe individual genotypes but we still do not directly observe individual haplotypes. Several methods exist for estimating haplotype-disease associations based on individual genotype data; see Zeng and Lin [2004] for a review. For this special case, the methods presented here are more general and more versatile than the existing ones.

We have focused on biallelic SNPs, the most common form of genetic markers. Our methods can be easily extended to other genetic variants, such as microsatellite loci. The computing time will depend on the number of haplotype

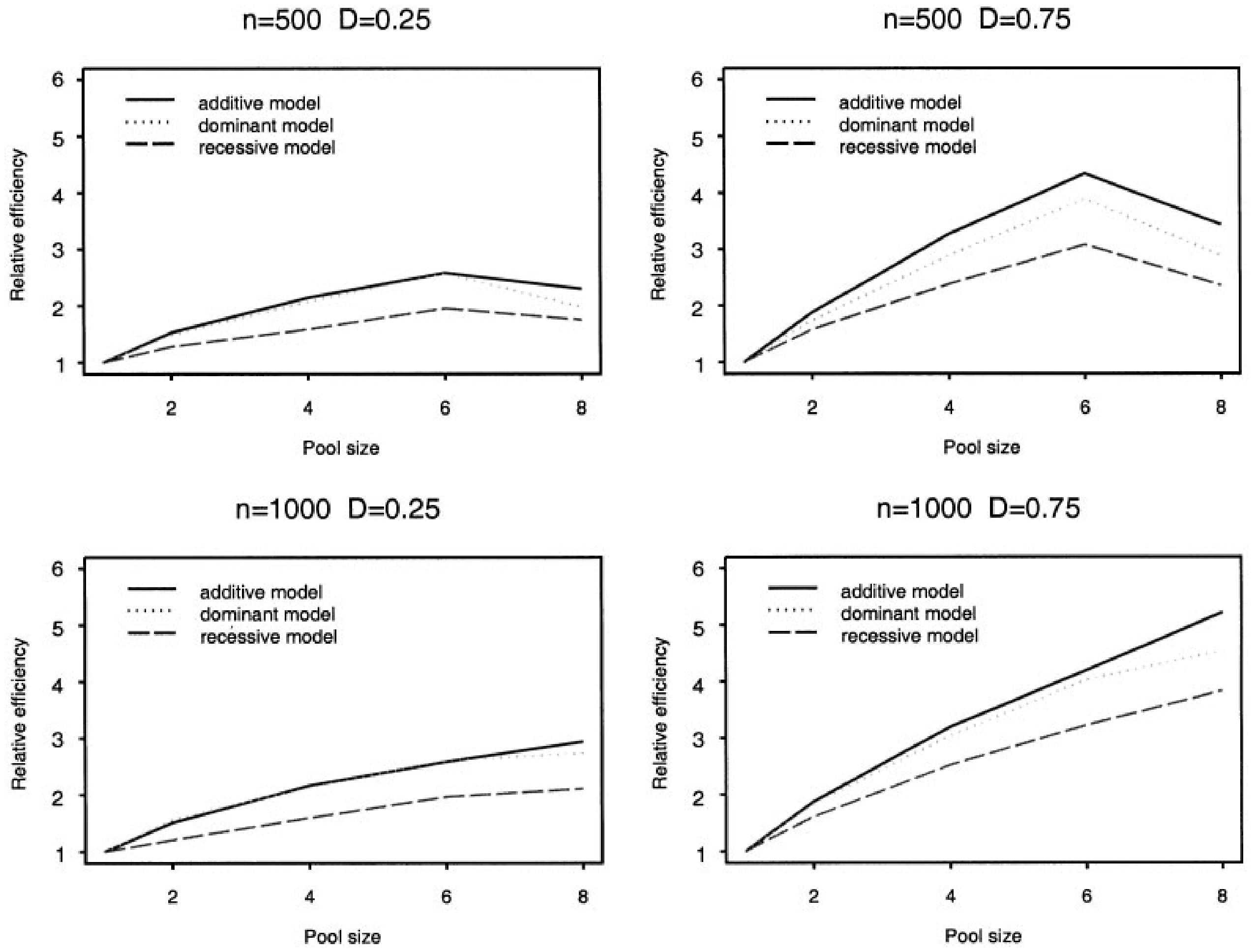

Fig. 2. Relative efficiencies in estimating haplotype frequencies.

combinations that are compatible with pooled genotypes.

In some applications, one may be interested in treating each SNP as a separate unit. Then phase uncertainty is not a concern. With pooled DNA data, however, individual genotypes in general cannot be determined with certainty. It is possible to modify the likelihood-based approach taken in this article to estimate the effects of multiple SNPs on the disease phenotype with pooled SNP-genotype data.

The choice between haplotype analysis and single-marker analysis depends on a number of factors, including the nature of the SNP-disease association, the number and positions of disease-causing SNPs, the extent and strength of linkage disequilibrium, and the selection of markers. Haplotype analysis is likely to be more powerful than single-marker analysis if the causal SNPs are not typed or if multiple SNPs act in *cis* rather than in *trans*. Haplotype analysis also serves as an effective data-reduction strategy, since the observed number of haplotypes tends to be much smaller than the theoretical number.

Yang et al. [2003] suggested the likelihood ratio statistic for testing the equality of haplotype distributions between cases and controls in case-control studies. This test has $(K-1)$ degrees of freedom and may not be very powerful if the majority of K haplotypes are unrelated to the disease. More important, this approach does not provide a quantification of the effects of haplotypes on the disease phenotype; neither does it allow assessment of gene-environment interactions. By contrast, the proposed methods enable one to efficiently estimate the effects of haplotypes and environment variables on the risk of disease.

Since our approach is built on likelihood, we can use likelihood-based model selection criteria, such as the information criterion (AIC) of Akaike [1985]. Specifically, we may select the model that minimizes the AIC, which is given by $-2logL + 2p$, where L is the likelihood evaluated at the MLEs, and p is the number of parameters in the model.

All the existing work on haplotype analysis, as well as our work, assumes that the number of allele copies in each DNA pool can be accurately determined. There are several sources of errors in quantitative genotyping assays, including unequal amounts of DNA from individual samples in the pool, preferential amplification of one of the alleles, and experimental errors. The errors due to unequal contributions from individual samples are often negligible. Differential allelic amplification can be corrected by a factor that is obtained from reference samples of known allele frequencies [Sham et al., 2002]. The proposed methods should be adequate when the measurement errors are small relative to the sampling errors. We can formally adjust for measurement errors in the inference on haplotype-disease associations if the error rates can be determined, perhaps by comparing the pooled and individual genotypes for a group of subjects and by genotyping a set of DNA samples multiple times.

We have developed a general computer program which implements the proposed methods. This program is available from the authors upon request.

ACKNOWLEDGMENTS

The authors are grateful to the FUSION Study Group for sharing their data, and to Michael Boehnke and Laura Scott for transmitting the data.

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.

Amos C, Frazier M, Wang W. 2001. DNA pooling in mutation detection with reference to sequence analysis. Am J Hum Genet 66:1689–1692.

Arnheim N, Strange C, Erlich H. 1985. Use of pooled DNA samples to detect linkage disequilibrium of polymorphic restriction fragments and human disease: studies of the HLA class II loci. Proc Natl Acad Sci USA 82:6970–6974.

Bansal A, van den Boom D, Kammerer S, Honisch C, Adam G, Cantor C, Kleyn P, Braun A. 2002. Association testing by DNA pooling: an effective initial screen. Proc Natl Acad Sci USA 99:16871–16874.

Barcellos L, Klitz W, Field L, Tobias R, Bowcock A, Wilson R, Nelson M, Nagatomi J, Thomson G. 1997. Association mapping of disease loci, by use of a pooled DNA genomic screen. Am J Hum Genet 61:734–747.

Barratt BJ, Payne F, Rance HE, Nutland S, Todd JA, Clayton DG. 2003. Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. Ann Hum Genet 66:393–405.

Bickel PJ, Klassen CAJ, Ritov Y, Wellner JA. 1993. Efficient and adaptive estimation in semiparametric models. Baltimore: Johns Hopkins University Press.

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.

Breslow N, Clayton DJ. 1993. Approximate inference in generalized linear mixed models. J Am Statist Assoc 88:9–25.

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

Cox DR. 1975. Partial likelihood. Biometrika 62:269–276.

Daniels J, Holmans P, Williams N, Turic D, McGuffin P, Plomin R, Owen M. 1998. A simple method for analyzing microsatellite allele image patterns generated from DNA pools and its application to allelic association studies. Am J Hum Genet 62:1189–1197.

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.

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.

Ito T, Chiku S, Inoue E, Tomita M, Morisaki T, Morisaki H, Kamatani N. 2003. Estimation of haplotype frequencies, linkage-disequilibrium measures, and combination of haplotype copies in each pool by use of pooled DNA data. Am J Hum Genet 72:384–398.

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

Mohlke K, Erdos M, Scott L, Fingerlin T, Jackson A, Silander K, Hollstein P, Boehnke M, Collins F. 2002. High-throughput screening for evidence of association by using mass spectrometry genotyping on DNA pools. Proc Natl Acad Sci USA 99:16928–16933.

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

Pacek P, Sajantila A, Syvanen AC. 1993. Determination of allele frequencies at loci with length polymorphism by quantitative analysis of DNA amplified from pooled samples. PCR Methods Appl 2:313–317.

Pettitt AN. 1982. Inference for the linear model using a likelihood based on ranks. J R Statist Soc B 44:234–243.

Risch N, Teng J. 1998. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. Genome Res 8:1273–1288.

Sasaki T, Tahira T, Suzuki A, Koichiro H, Kukita Y, Baba S, Hayahi K. 2001. Precise estimation of allele frequencies of single-nucleotide polymorphisms by a quantitative SSCP analysis of pooled DNA. Am J Hum Genet 68:214–218.

Sham P, Bader J, Craig I, O'Donovan M, Owen M. 2002. DNA pooling: A tool for large-scale association studies. Nat Rev Genet 3:862–871.

Shaw S, Carrasquillo M, Kashuk C, Puffenberger E, Chakravarti A. 1998. Allele frequency distributions in pooled DNA samples:

applications to mapping complex disease genes. Genome Res 8:111–123.

Valle T, Tuomilehto J, Bergman RN, Ghosh S, Hauser ER, Eriksson J, Nylund S, Kohtamaki K, Tuomilehto-Wolf E, Toivanen L, Vidgren G, Ehnholm C, Blaskchak J, Langefeld CD, Watanabe RM, Magnuson V, Ally DS, Hagopian W, Ross E, Buchanan TA, Collins F, Boehnke M. 1998. Mapping genes for non-insulin dependent diabetes mellitus: Design of the Finland-United States investigation of NIDDM genetics (FUSION. study. Diabetes Care) 21:949–958.

Wang S, Kidd K, Zhao H. 2003. On the use of DNA pooling to estimate haplotype frequencies. Genet Epidemiol 24:74–82.

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

Yang Y, Zhang J, Hoh J, Matsuda F, Xu P, Lathrop M, Ott J. 2003. Efficiency of single-nucleotide polymorphism haplotype estimation from pooled DNA. Proc Natl Acad Sci USA 100:7225–7230.

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. 2004. Likelihood-based inference on haplotype effects in genetic association studies. Technical Report. University of North Carolina, Department of Biostatistics.

APPENDIX A

EM ALGORITHM FOR MAXIMIZING (3)

We derive an EM algorithm for the maximization of (3) by treating the H_{ji} as missing data. Suppose that condition (2) holds with $\rho \geq 0$. Let B_{ji} be a Bernoulli variable with success probability ρ, and let Q_{1ji} and Q_{2ji} be discrete random variables with probability density functions $P(Q_{1ji} = (h_k, h_k)) = \pi_k$ and $P(Q_{2ji} = (h_k, h_l)) = \pi_k \pi_l$. Then $B_{ji}Q_{1ji} + (1 - B_{ji})Q_{2ji}$ has the same distribution as H_{ji}, and we can treat B_{ji}, Q_{1ji}, and Q_{2ji} instead of H_{ji} as missing. With this data augmentation, the complete-data likelihood is proportional to

$$\prod_{j=1}^{J}\prod_{i=1}^{n_j}\left\{P(Y_{ji}|X_{ji}, H_{ji}; \theta)\rho^{B_{ji}}(1-\rho)^{1-B_{ji}}\right.$$
$$\left.\times\prod_{k=1}^{K}\pi_k^{B_{ji}I(Q_{1ji}=(h_k,h_k))}\prod_{k,l=1}^{K}(\pi_k\pi_l)^{(1-B_{ji})I(Q_{2ji}=(h_k,h_l))}\right\}.$$

Let Y_j denote $(Y_{j1}, \ldots, Y_{jn_j})$ and X_j denote $(X_{j1}, \ldots, X_{jn_j})$. In the M-step of the EM algorithm, we solve the following score equation for θ:

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j} E\{\partial log P(Y_{ji}|X_{ji}, H_{ji}; \theta)/\partial\theta | Y_j, X_j, G_j\} = 0;$$

we estimate ρ and π_k by

$$\widehat{\rho} = n^{-1}\sum_{j=1}^{J}\sum_{i=1}^{n_j} E\{B_{ji}|Y_j, X_j, G_j\}$$

and

$$\widehat{\pi}_k = c^{-1}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left[E\{B_{ji}I(Q_{1ji} = (h_k, h_k))|Y_j, X_j, G_j\}.\right.$$
$$\left. +2\sum_{l=1}^{K} E\{(1 - B_{ji})I(Q_{2ji} = (h_k, h_l))|Y_j, X_j, G_j\}\right],$$

where c is a normalizing constant such that $\sum_k \widehat{\pi}_k = 1$. The conditional expectations in the above expressions are calculated in the E-step as follows: for any function $\omega(B_{ji}, Q_{1ji}, Q_{2ji})$,

$$E\{\omega(B_{ji}, Q_{1ji}, Q_{2ji})|Y_j, X_j, G_j\}$$
$$= \frac{\sum\limits_{H_j \in S(G_j)} \omega(B_{ji}, Q_{1ji}, Q_{2ji}) \prod_{i=1}^{n_j} P(Y_{ji}|X_{ji}, H_{ji}; \theta)P(H_{ji}; \gamma)}{\sum\limits_{H_j \in S(G_j)} \prod_{i=1}^{n_j} P(Y_{ji}|X_{ji}, H_{ji}; \theta)P(H_{ji}; \gamma)}, \tag{A.1}$$

where θ and γ are evaluated at their current estimates. If condition (2) with $\rho \geq 0$ does not hold, then the estimator of γ in the M-step does not have an explicit form, and we solve instead the following score equation:

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j} E\{\partial log P(H_{ji}; \gamma)/\partial\gamma | Y_j, X_j, G_j\} = 0.$$

The main computational burden lies in the evaluation of (A1). The following representation greatly facilitates this evaluation. We can express (A1) as

$$\frac{\sum\limits_{G_{ji} \leq G_j}\left\{\sum\limits_{H_{ji} \in S(G_{ji})} \omega(B_{ji}, Q_{1ji}, Q_{2ji})P(Y_{ji}|X_{ji}, H_{ji}; \theta)P(H_{ji}; \gamma)\right\}\Gamma_{ji}(G_j - G_{ji})}{\sum\limits_{G_{ji} \leq G_j}\left\{\sum\limits_{H_{ji} \in S(G_{ji})} P(Y_{ji}|X_{ji}, H_{ji}; \theta)P(H_{ji}; \gamma)\right\}\Gamma_{ji}(G_j - G_{ji})},$$

where

$$\Gamma_{ji}((g_1, \ldots, g_M)) = \sum_{G_{j1}+\cdots G_{j,i-1}+G_{j,i+1}+\cdots G_{j,n_j}=(g_1,\ldots,g_M)}$$
$$\times\prod_{s \neq i}\left\{\sum_{H_{js} \in S(G_{js})} P(Y_{js}|X_{js}, H_{js}; \theta)P(H_{js}; \gamma)\right\},$$

$G_{ji} \leq G_j$ means that all the M SNP-genotype values of the ith subject in the jth pool are less than or equal to those of the jth pool, and $G_j - G_{ji}$

pertains to the difference of the two M-vectors. To calculate $\Gamma_{ji}((g_1, \ldots, g_M))$, we introduce $(n_j - 1)$ M-variate polynomials

$$\Omega_{js}(x_1, \ldots, x_M) = \sum_{(g_1,\ldots,g_M)\in\mathcal{G}} C^s_{g_1,\ldots,g_M} x_1^{g_1} \cdots x_M^{g_M},$$

$$s = 1, \ldots, n_j, s \neq i,$$

where $\mathcal{G}$ denotes the set of all possible genotypes in M SNPs and

$$C^s_{g_1,\ldots,g_M} = \sum_{H_{js}\in S((g_1,\ldots,g_M))} P(Y_{js}|X_{js}, H_{js}; \theta)P(H_{js}; \gamma).$$

Then $\Gamma_{ji}(g_1, \ldots, g_M)$ corresponds to the coefficient of $x_1^{g_1} \cdots x_M^{g_M}$ in the product $\prod_{s\neq i} \Omega_{js}(x_1, \ldots, x_M)$. These coefficients can be obtained, for example, from the Symbolic Math Toolbox of MATLAB.

APPENDIX B

EM ALGORITHM FOR MAXIMIZING (4)

Assume that condition (2) holds with $\rho \geq 0$. Let B_{ji}, Q_{1ji}, Q_{2ji}, Y_j and X_j be as defined in Appendix A. For the subjects not selected, we attach X_m, H_m, B_m, Q_{1m}, and Q_{2m} to Y_m $(m = n+1, \ldots, N)$. In the M-step, we solve the following score equation for θ,

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j} E\{\partial log P(Y_{ji}|X_{ji}, H_{ji}; \theta)/\partial\theta|Y_j, X_j, G_j\}$$
$$+ \sum_{m=n+1}^{N} E\{\partial log P(Y_m|X_m, H_m; \theta)/\partial\theta|Y_m\} = 0;$$

we estimate ρ and π_k by

$$N^{-1}\left\{\sum_{j=1}^{J}\sum_{i=1}^{n_j} E(B_{ji}|Y_j, X_j, G_j) + \sum_{m=n+1}^{N} E(B_m|Y_m)\right\}$$

and

$$c^{-1}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left[E\{B_{ji}I(Q_{1ji} = (h_k, h_k))|Y_j, X_j, G_j\}\right.$$
$$\left. + 2\sum_{l=1}^{K} E\{(1 - B_{ji})I(Q_{2ji} = (h_k, h_l))|Y_j, X_j, G_j\}\right]$$
$$+ c^{-1}\sum_{m=n+1}^{N}\left[E\{B_m I(Q_{1m} = (h_k, h_k))|Y_m\}\right.$$
$$\left. + 2\sum_{l=1}^{K} E\{(1 - B_m)I(Q_{2m} = (h_k, h_l))|Y_m\}\right],$$

where c is the normalizing constant. In addition, we estimate $P(X)$ as a mass function at the distinct values of the X_{ji}: the mass at x is equal to

$$N^{-1}\left[\sum_{j=1}^{J}\sum_{i=1}^{n_j} I(X_{ji} = x) + \sum_{m=n+1}^{N} E\{I(X_m = x)|Y_m\}\right].$$

The conditional expectations in the above expressions are calculated in the E-step: for $j = 1, \ldots, J$, $i = 1, \ldots, n_j$, and for any function $\omega(B_{ji}, Q_{1ji}, Q_{2ji})$, the conditional expectation $E\{\omega(B_{ji}, Q_{1ji}, Q_{2ji})|Y_j, X_j, G_j\}$ is given in (A1); for $m = n + 1, \ldots, N$ and for any function $\omega(B_m, Q_{1m}, Q_{2m}, X_m)$,

$$E\{\omega(B_m, Q_{1m}, Q_{2m}, X_m)|Y_m\}$$
$$= \frac{\sum_{B,Q_1,Q_2,X} \omega(B, Q_1, Q_2, X)P(Y_m|X, H; \theta)P(H; \gamma)P(X)}{\sum_{B,Q_1,Q_2,X} P(Y_m|X, H; \theta)P(H; \gamma)P(X)},$$

where θ, γ, and $P(X)$ are evaluated at their current estimates.

APPENDIX C

CALCULATION OF MLES OF θ AND γ BASED ON (6)

Because the distribution of X is a potentially infinite-dimensional nuisance parameter, we wish to calculate the MLEs of β and γ by profiling (6) over the distribution function of X. By algebraic manipulations similar to those given in Appendix 4.4 of Zeng and Lin [2004], we can show that profiling the logarithm of (6) over the distribution function of X is equivalent to profiling the following function over the scalar parameter μ:

$$\sum_{j=1}^{J} \log\left[\sum_{H_j\in S(G_j)} \cdot \prod_{i=1}^{n_j} \frac{\exp\{Y_{ji}\beta^T Z(X_{ji}, H_{ji})\}P(H_{ji}; \gamma)p^{Y_{ji}}\{(1-p)\mu\}^{1-Y_{ji}}}{\sum_{y=0}^{1}\sum_{H} exp\{y\beta^T Z(X_{ji}, H)\}P(H; \gamma)p^{y}\{(1-p)\mu\}^{1-y}}\right],$$

where p is the proportion of cases in the case-control sample. The above expression is the log-likelihood function for a cohort study in which the conditional distribution of Y_{ji} and H_{ji} given X_{ji} has the probability density function

$$\widetilde{P}(Y_{ji}, X_{ji}, H_{ji}; \theta, \gamma) = \frac{\exp\{Y_{ji}\beta^T Z(X_{ji}, H_{ji})\}P(H_{ji}; \gamma)p^{Y_{ji}}\{(1-p)\mu\}^{1-Y_{ji}}}{\sum_{y=0}^{1}\sum_{H}\exp\{y\beta^T Z(X_{ji}, H)\}P(H; \gamma)p^{y}\{(1-p)\mu\}^{1-y}}$$

and in which the pooled genotypes G_j instead of the individual haplotypes H_{ji} are observed. Thus, we can use the EM algorithm described in Appendix A upon replacing $P(Y_{ji}|X_{ji}, H_{ji}; \theta) \times P(H_{ji}; \gamma)$ with $\widetilde{P}(Y_{ji}, X_{ji}, H_{ji}; \theta, \gamma)$. Furthermore, the inverse of the observed Fisher information matrix can be used to estimate the covariance matrix of the MLEs of θ and γ. In our variance estimation, we approximate the observed Fisher information matrix by the empirical covariance matrix of the observed score function. Under condition (2) with $\rho \geq 0$, we can use the data augmentation $H_{ji} = B_{ji}Q_{1ji} + (1 - B_{ji})Q_{2ji}$ introduced in Appendix A. The M-step can then be simplified if we express $\widetilde{P}(Y, X, H = BQ_1 + (1-B)Q_2; \theta, \gamma)$ as

$$\frac{\exp\left\{\xi_0 Y + Y\beta^T Z(X, H) + \sum_{k=1}^{K} \xi_k Z_k\right\}}{\sum_{Y,B,Q_1,Q_2} \exp\left\{\xi_0 Y + Y\beta^T Z(X, H) + \sum_{k=1}^{K} \xi_k Z_k\right\}},$$

where $Z_k = BI(Q_1 = (h_k, h_k)) + 2(1 - B)\sum_l I(Q_2 = (h_k, h_l))$ $(k = 1, \ldots, K)$, 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 algorithm.

APPENDIX D

EM ALGORITHM FOR MAXIMIZING (8)

Assume that condition (2) holds with $\rho \geq 0$. Let B_{ji}, Q_{1ji} and Q_{2ji} be as defined in Appendix A. Also, let $\widetilde{Y}_j$ denote $(\widetilde{Y}_{j1}, \ldots, \widetilde{Y}_{jn_j})$ and Δ_j denote $(\Delta_{j1}, \ldots, \Delta_{jn_j})$. In the M-step of the EM algorithm, we estimate β by solving the following equation:

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}\Delta_{ji}\left[E\{Z(X_{ji}, H_{ji})|\widetilde{Y}_j, \Delta_j, X_j, G_j\} - \frac{\sum_{v=1}^{J}\sum_{u=1}^{n_v} I(\widetilde{Y}_{vu} \geq \widetilde{Y}_{ji})E\{Z(X_{vu}, H_{vu})e^{\beta^T Z(X_{vu}, H_{vu})}|\widetilde{Y}_v, \Delta_v, X_v, G_v\}}{\sum_{v=1}^{J}\sum_{u=1}^{n_v} I(\widetilde{Y}_{vu} \geq \widetilde{Y}_{ji})E\{e^{\beta^T Z(X_{vu}, H_{vu})}|\widetilde{Y}_v, \Delta_v, X_v, G_v\}}\right] = 0,$$

and estimate $\Lambda_0(t)$ by

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}\frac{I(\widetilde{Y}_{ji} \leq t)\Delta_{ji}}{\sum_{v=1}^{J}\sum_{u=1}^{n_v} I(\widetilde{Y}_{vu} \geq \widetilde{Y}_{ji})E\{e^{\beta^T Z(X_{vu}, H_{vu})}|\widetilde{Y}_v, \Delta_v, X_v, G_v\}}.$$

In addition, we estimate ρ and π_k by

$$n^{-1}\sum_{j=1}^{J}\sum_{i=1}^{n_j} E\{B_{ji}|\widetilde{Y}_j, \Delta_j, X_j, G_j\}$$

and

$$c^{-1}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\Bigg[E\{B_{ji}I(Q_{1ji} = (h_k, h_k))|\widetilde{Y}_j, \Delta_j, X_j, G_j\} + 2\sum_{l=1}^{K} E\{(1 - B_{ji})I(Q_{2ji} = (h_k, h_l))|\widetilde{Y}_j, \Delta_j, X_j, G_j\}\Bigg],$$

where c is the normalizing constant. In the above expressions, the conditional expectation $E[\omega(B_{ji}, Q_{1ji}, Q_{2ji}, X_{ji}|\widetilde{Y}_j, X_j, \Delta_j, G_j]$ is evaluated in the E-step according to the formula

$$\frac{\sum_{H_j \in S(G_j)} \omega(B_{ji}, Q_{1ji}, Q_{2ji}, X_{ji})\prod_{i=1}^{n_j}\exp\left\{\Delta_{ji}\beta^T Z(X_{ji}, H_{ji}) - \Lambda(\widetilde{Y}_{ji})e^{\beta^T Z(X_{ji}, H_{ji})}\right\}P(H_{ji}; \gamma)}{\sum_{H_j \in S(G_j)}\prod_{i=1}^{n_j}\exp\left\{\Delta_{ji}\beta^T Z(X_{ji}, H_{ji}) - \Lambda(\widetilde{Y}_{ji})e^{\beta^T Z(X_{ji}, H_{ji})}\right\}P(H_{ji}; \gamma)}, \qquad (D1)$$

where β, γ, and Λ are evaluated at their current estimates.