

# Haplotype-Based Association Analysis in Cohort Studies of Unrelated Individuals

D.Y. Lin\*

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

Exploring the associations between haplotypes and disease phenotypes is an important step toward the discovery of genes that influence complex human diseases. When unrelated subjects are sampled, haplotypes are often ambiguous because of the unknown gametic phase of the measured sites along a chromosome. We consider cohort studies of unrelated subjects which collect data on potentially censored ages of onset of disease along with unphased genotypes and possibly time-varying environmental factors. We formulate the effects of haplotypes and environmental variables on the time to disease occurrence through a semiparametric Cox proportional hazards model, which can accommodate a variety of genetic mechanisms as well as gene-environment interactions. We develop a simple and fast expectation-maximization algorithm to maximize the likelihood for the relative risks and other parameters based on the observable data of unphased genotypes and potentially censored ages of onset. The resultant estimators are consistent, efficient, and asymptotically normal. Simulation studies show that, for practical situations, the parameter estimators are virtually unbiased, the association tests maintain type I errors near nominal levels, the confidence intervals have proper coverage probabilities, and the efficiency loss due to unknown gametic phase is small. © 2004 Wiley-Liss, Inc.

**Key words:** age of onset; association tests; censoring; gene-environment interactions; haplotype effects; maximum likelihood; proportional hazards; SNPs; unphased genotype

Grant sponsor: National Cancer Institute

\*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 8 October 2003; Accepted 26 November 2003

Published online 25 February 2004 in Wiley InterScience (www.interscience.wiley.com)

DOI: 10.1002/gepi.10317

## INTRODUCTION

The recent sequencing of the human genome [International Human Genome Sequencing Consortium, 2001; Venter et al., 2001] revealed 30,000–40,000 functional genes throughout the genome. In addition, millions of single-nucleotide polymorphisms (SNPs) have been identified [International SNP Map Working Group, 2001]. Recently developed molecular techniques [Chee et al., 1996; Wang et al., 1998] enable researchers to genotype biological samples for thousands of SNPs. Furthermore, the continuing increase in genotyping efficiency has made genotyping a large number of subjects economically feasible. These remarkable scientific and technological advances yield unprecedented opportunities to conduct large-scale studies of the associations between genetic variants (e.g., SNPs) and complex human diseases such as cancer, bipolar disorder, diabetes, schizophrenia, and coronary heart dis-

eases [Risch and Merikangas, 1996; Botstein and Risch, 2003].

There are various statistical approaches to evaluating the associations between SNPs and disease phenotypes. One possible approach is to treat each SNP as a separate variable and use a stepwise strategy to process all SNPs systematically [Cordell and Clayton, 2002]. A more appealing approach is to haplotype for multiple SNPs within candidate genes [Hallman et al., 1999; Drysdale et al., 2000; International SNP Map Working Group, 2001; Patil et al., 2001; Stephens et al., 2001]. Since the number of haplotypes within candidate genes is much smaller than the number of all possible haplotypes, haplotyping serves as an effective data-reduction strategy. The use of SNP-based haplotypes, which are specific combinations of allelic variants at a series of tightly linked SNPs, tends to be more powerful for gene mapping than the use of single SNPs, since a haplotype incorporates linkage disequilibrium

information from multiple markers [Fallin et al., 2001; Akey et al., 2001; Zaykin et al., 2002]. Haplotype-based analysis has an additional power advantage over allele-based analysis when multiple disease-susceptibility variants occur within the same gene [Morris and Kaplan, 2002]. Haplotype analysis also provides critical information about the function of a gene.

The determination of haplotype requires the parental origin or gametic phase information, which cannot be easily obtained with current genotyping technologies. Although haplotype ambiguity can potentially be resolved by genotyping the family members of each subject, this strategy is not always an option. Several algorithms [Clark, 1990; Terwilliger and Ott, 1994; Excoffier and Slatkin, 1995; Hawley and Kidd, 1995; Long et al., 1995; Stephens et al., 2001; Zhang et al., 2001; Niu et al., 2002; Qin et al., 2002] were developed to estimate haplotype frequencies from unphased genotype data of unrelated subjects. Statistical inference about the effects of specific haplotype features on disease phenotypes poses additional challenges.

A number of authors [Fallin et al., 2001; Zhao et al., 2000; Schaid et al., 2002; Zaykin et al., 2002; Zhao et al., 2003; Epstein and Satten, 2003; Stram et al., 2003] proposed statistical methods for testing or estimating haplotype-specific relative risks based on unphased genotype data from case-control studies of unrelated subjects. A case-control study identifies a sample of diseased subjects and a separate sample of disease-free subjects, and collects genetic and exposure information on each subject retrospectively. This is a cost-effective design, especially for rare diseases.

The cohort study, which follows a sample of subjects forward in time to ascertain the occurrence of the disease of interest, offers several advantages over the case-control study [Breslow and Day, 1987, p. 11–20]. First, complex human diseases have variable ages of onset. Since the age of onset is likely to be genetically mediated, the subject's age of onset carries more information about the etiology of the disease than the case-control status. Secondly, selection and recall bias inherent in case-control studies can usually be eliminated in cohort studies. Thirdly, the cohort design enables one to collect reliable predisease exposure information and to investigate a full range of diseases in the same study.

There are a number of ongoing cohort studies, including the Cardiovascular Health Study [Fried et al., 1991], the Women's Health Initiative

[Johnson et al., 1999], and the Atherosclerosis Risk in Communities (ARIC) Study [ARIC Investigators, 1989], which aim to explore the environmental and genetic causes of complex diseases. This work was motivated by the ARIC Study, which is conducted by the Collaborative Studies Coordinating Center at the University of North Carolina. ARIC is a prospective cohort study based on 16,000 individuals aged 45–64 years to investigate the etiology of atherosclerosis and other diseases. The investigators are currently studying the associations of various genetic variants with times to the developments of cancer and cardiovascular diseases.

The goal of this article is to develop valid and efficient statistical methods for making inferences about the effects of haplotypes on time to onset of the disease of interest in the ARIC Study and other cohort studies of unrelated subjects. Because of loss to follow-up and limited study duration, times to disease developments are potentially censored. We formulate the effects of haplotypes and possibly time-varying environmental factors on the potentially censored time to disease occurrence through a semiparametric proportional hazards model according to Cox [1972]. This formulation allows modeling of dominant, recessive, additive, or other effects of specific haplotype configurations as well as gene-environment interactions. Since unphased genotypes rather than haplotypes are observed, the conventional partial likelihood principle for the Cox model is not applicable. We derive a nonparametric likelihood for the model parameters based on the observable data, and show that this likelihood possesses the familiar asymptotic properties of parametric likelihood. Simulation studies demonstrate that likelihood-based inference procedures perform very well in practical situations.

## METHODS

### DATA AND ASSUMPTIONS

Suppose that we follow a cohort of  $n$  unrelated subjects for the development of certain complex diseases. For  $i = 1, \dots, n$ , let  $T_i$  be the time to occurrence of a particular disease of interest for the  $i$ th subject, and  $C_i$  be the corresponding censoring time; also, let  $G_i$  be the (unphased) multilocus genotype, and  $X_i$  be a set of possibly time-varying covariates representing environmental factors. The (observable) data consist of  $(Y_i, \Delta_i, X_i, G_i)$  ( $i = 1, \dots, n$ ), where  $Y_i = \min$

$(T_i, C_i)$ , and  $\Delta_i = I(T_i \leq C_i)$ . Here and in the sequel,  $I(\mathcal{A})$  is the indicator function for event  $\mathcal{A}$ , which takes the value 1 or 0 dependent on whether event  $\mathcal{A}$  occurs or not. By allowing some genotypes to include missing SNP information, we may assume that the  $G_i$  are known for all  $i$ .

Suppose that each subject is genotyped at a series of  $M$  SNPs. The total number of possible haplotypes is  $K=2^M$ . The actual number of haplotypes consistent with the data is normally much smaller. For  $k = 1, \dots, K$ , let  $h_k$  denote the  $k$ th possible haplotype. If the  $i$ th subject's pair of haplotypes is  $h_k$  and  $h_l$ , then we write  $H_i = (h_k, h_l)$ . The value of  $H_i$  is unknown if the  $i$ th subject is heterozygous at more than one SNP or if any SNP genotype is missing.

We wish to study the effects of haplotypes and environmental factors, and possibly their interactions, on the time to disease occurrence. The method of choice for relating possibly time-varying covariates to the potentially censored failure time is the proportional hazards model of Cox [1972]. We extend this model to the current setting by specifying that the hazard function for  $T_i$  conditional on  $X_i$  and  $H_i = (h_k, h_l)$  takes the form

$$\lambda\{t|X_i, H_i = (h_k, h_l)\} = \lambda_0(t)e^{\beta^T Z_{ikl}(t)}, \quad (1)$$

where  $\lambda_0(t)$  is an unspecified baseline hazard function,  $Z_{ikl}$  is a  $p$ -dimensional function of  $X_i$ ,  $h_k$ , and  $h_l$  that is chosen according to the hypotheses of interest, and  $\beta$  is a  $p$ -vector of regression parameters associated with  $Z_{ikl}$ . The components of  $\beta$  pertain to the (natural) logarithm of the hazard ratio or relative risk. We assume that  $C_i$  is independent of  $T_i$  and  $H_i$  conditional on  $X_i$  and  $G_i$ , and that  $H_i$  is independent of  $X_i$  conditional on  $G_i$ . For notational simplicity, the formulas in the sequel are confined to time-invariant covariates.

To provide some insights into the choice of  $Z_{ikl}$  and the interpretation of the corresponding  $\beta$ , we consider the situation in which one is interested in the effect of a specific haplotype  $h^*$ . Let  $\delta_k = I(h_k = h^*)$ . Then  $\beta^T Z_{ikl}$  takes the form of  $\alpha\delta_k\delta_l + \gamma^T X_i$  for a recessive genetic model,  $\alpha(\delta_k + \delta_l - \delta_k\delta_l) + \gamma^T X_i$  for a dominant genetic model, and  $\alpha(\delta_k + \delta_l) + \gamma^T X_i$  for an additive genetic model (multiplicative on the relative-risk scale), where  $\alpha$  and  $\gamma$  pertain to the log relative risks associated with  $h^*$  and  $X_i$ , respectively. If one is interested in investigating gene-environment interactions, then the products of  $\delta_k$  and  $\delta_l$  with  $X_i$  may be added.

Because of haplotype ambiguity, estimation of model (1) is not feasible without additional assumptions. We suppose that the haplotype pairs

in the underlying population are in Hardy-Weinberg equilibrium, such that  $\Pr\{H_i = (h_k, h_l)\} = \pi_k\pi_l$  ( $i = 1, \dots, n; k, l = 1, \dots, K$ ), where  $\pi_k$  is the population frequency of haplotype  $k$ . This kind of assumption is required of all haplotype analyses, including all previous work on estimating haplotype effects in case-control studies. We do not require any assumptions on linkage disequilibrium, recombination, or other population genetic parameters.

## ESTIMATION

Model (1) is a semiparametric model with latent covariates in that the distributional form of  $T_i$  is unspecified and  $H_i$  is not directly observable. In the absence of latent covariates, Cox [1972, 1975] provided an ingenious partial likelihood principle for estimating the set of regression parameters  $\beta$ ; the corresponding estimator of the cumulative baseline hazard function  $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$  is commonly referred to as the estimator of Breslow [1972]. In the presence of latent covariates, the partial likelihood function is intractable. We will instead use the so-called nonparametric maximum likelihood method [Bickel et al., 1993], which is a modern approach to the estimation of nonparametric and semiparametric models.

Let  $S(G_i)$  denote the set of haplotype pairs consistent with genotype  $G_i$ . If the genotype is missing at one or more loci, then the set  $S$  is enlarged accordingly. The haplotype pair uniquely determines the genotype, but not vice versa. The likelihood involves the summation of the probability densities under model (1) over all the haplotype pairs that are consistent with the genotype.

Write  $\pi = (\pi_1, \dots, \pi_K)$ . The likelihood function based on the observable data  $(Y_i, \Delta_i, X_i, G_i)$  ( $i = 1, \dots, n$ ) is proportional to

$$L(\beta, \pi, \Lambda) = \prod_{i=1}^n \sum_{k,l} I\{(h_k, h_l) \in S(G_i)\} \left( \Lambda\{Y_i\} e^{\beta^T Z_{ikl}} \right)^{\Delta_i} \times \exp\left\{ -\Lambda(Y_i) e^{\beta^T Z_{ikl}} \right\} \pi_k \pi_l, \quad (2)$$

where  $\Lambda(t)$  is an increasing right-continuous function,  $\Lambda\{Y_i\}$  is the jump size of  $\Lambda(t)$  at  $t = Y_i$ , i.e., the value of  $\Lambda(t)$  at  $t = Y_i$  minus its value right before  $Y_i$ , and  $\sum_{k,l}$  denotes the (double) summation over  $k = 1, \dots, K$  and  $l = 1, \dots, K$ . The maximum likelihood estimators, denoted by  $\hat{\beta}$ ,  $\hat{\pi}$ , and  $\hat{\Lambda}$ , are the values of  $\beta$ ,  $\pi$ , and  $\Lambda$  which maximize  $L(\beta, \pi, \Lambda)$ . It is easy to show that  $\hat{\Lambda}(t)$  is a step function with jumps only at those  $Y_i$  for which

$\Delta_i = 1$ . Thus, the maximization boils down to maximizing  $L(\beta, \pi, \Lambda)$  over the jump sizes of  $\Lambda(t)$  at those time points, along with  $\beta$  and  $\pi$ . We show in Appendix A that this maximization can be carried out efficiently through a simple and elegant expectation-maximization (EM) algorithm [Dempster et al., 1977].

It is important to note that, although the true value of  $\Lambda_0(t)$  in model (1) is usually assumed to be continuous, the likelihood function  $L(\beta, \pi, \Lambda)$  is defined on the extended parameter space in which  $\Lambda(t)$  is not necessarily continuous. As mentioned above, the function  $\Lambda(t)$  that maximizes  $L(\beta, \pi, \Lambda)$  is always a step function which jumps at the observed  $Y_i$  for which  $\Delta_i = 1$ . This is a modern approach to the estimation of infinite-dimensional parameters, and would in fact yield the familiar maximum partial likelihood estimator for  $\beta$  and the Breslow estimator for  $\Lambda_0$  in the absence of latent covariates [van der Vaart, 1998, p. 402–404].

## INFERENCE PROCEDURES

We show in Appendix B that the maximum likelihood estimators  $\hat{\beta}$ ,  $\hat{\pi}$ , and  $\hat{\Lambda}$  are consistent, efficient, and asymptotically normal, with variances and covariances that can be consistently estimated from the observable data. It is convenient to perform association tests by the likelihood ratio statistics. Suppose that one is interested in testing the null hypothesis  $H_0 : \beta^{(1)} = \beta_0^{(1)}$  against the alternative hypothesis  $H_1 : \beta^{(1)} \neq \beta_0^{(1)}$ , where  $\beta^{(1)}$  is a subset of  $\beta$  with  $r$  components. Let  $\hat{\beta}$ ,  $\hat{\pi}$ , and  $\hat{\Lambda}$  be the maximum likelihood estimators of  $\beta$ ,  $\pi$ , and  $\Lambda_0$  under  $H_0$ . Then the likelihood ratio statistic takes the form of  $-2 \log \{L(\hat{\beta}, \hat{\pi}, \hat{\Lambda}) / L(\hat{\beta}, \hat{\pi}, \hat{\Lambda})\}$ , which is approximately chi-squared with  $r$  degrees of freedom. Confidence intervals or regions for  $\beta^{(1)}$  can be obtained by inverting the likelihood ratio tests: the  $(1 - \alpha)100\%$  confidence region for  $\beta^{(1)}$  consists of the values of  $\beta_0^{(1)}$  which are not rejected by the  $\alpha$ -level likelihood ratio tests.

In practice, the true model of disease is unknown, so that one will likely have to test many possible models. Thus, it is useful to develop methods for model selection. Since our approach is likelihood-based, we can apply model selection criteria such as the Akaike information criterion (AIC) [Akaike, 1985] to determine the best model for haplotype effects on disease risk. Specifically, we propose to select the model that

minimizes the AIC, which is given by  $-2 \log L(\hat{\beta}, \hat{\pi}, \hat{\Lambda}) + 2p$ .

## RESULTS

We carried out Monte Carlo simulation studies to assess the performance of the proposed numerical and inferential procedures in practical situations. We generated individuals' haplotypes by randomly drawing pairs of haplotypes from the frequency distribution shown in Table I, which pertains to 5 tightly-linked SNPs on chromosome 22 found in the control sample of the FUSION data reported by Epstein and Satten [2003]. Following those authors, we focused on the additive effects of haplotype 01100 in our simulations. We generated times to disease occurrence according to the following model

$$\lambda \{t | H_i = (h_k, h_l)\} = \lambda \eta (\lambda t)^{\eta-1} e^{\beta(\delta_k + \delta_l)},$$

$$i = 1, \dots, n; k, l = 1, \dots, 32,$$

where, for  $k = 1, \dots, 32$ , the indicator variable  $\delta_k$  equals 1 if  $h_k$  is 01100 and 0 otherwise. We set  $\lambda = 1$  and  $\eta = 2$  to produce an increasing hazard function over time. The censoring times were generated from the uniform  $(0, \tau)$  distribution, where  $\tau$  was chosen to yield approximately 90% censored observations.

We considered relative risks of 1, 1.25, 1.5, and 1.75, i.e.,  $\beta = 0, \log 1.25, \log 1.5, \text{ and } \log 1.75$ . We set  $n = 1,000, 2,000, \text{ and } 5,000$ , which, at the 90% censoring level, yields on average 100, 200, and 500 disease cases, respectively. We assessed the performance of the EM algorithm described in Appendix A. More importantly, we investigated the bias and variance of the maximum likelihood estimator of  $\beta$ , the size/power of the  $\alpha$ -level

TABLE I. Haplotype frequencies for simulation studies

Haplotype	Frequency
00011	0.0042
00110	0.0018
01100	0.2514
01111	0.0019
10011	0.3574
10110	0.0317
11100	0.0110
00100	0.0035
01011	0.1292
01101	0.0012
10000	0.0136
10100	0.0520
11011	0.1391
11111	0.0020

likelihood ratio statistic for testing  $H_0 : \beta = 0$ , and the coverage probability of the  $(1 - \alpha)$  100% confidence interval for  $\beta$ , where  $\alpha$  is set to 0.01 or 0.05. In addition, we assessed the ability of the AIC in selecting the true model among the additive, dominant, and recessive models. Since the haplotypes are known in simulations, we were able to evaluate the variance of the maximum partial likelihood estimator of  $\beta$  and the power of the likelihood ratio test with known haplotypes.

For each combination of simulation parameters, we generated 10,000 simulated data sets. We set the convergence criterion of the EM algorithm to be  $10^{-4}$ . The algorithm converged in less than 20 iterations for all simulated data sets. It took only a few seconds to perform the estimation and testing for a given data set, even with sample size of 5,000, in a SUN BLADE 1000 machine, which is slower than most current personal computers.

Table II reports the performance of the proposed inferential procedures. The maximum likelihood estimator is virtually unbiased, the likelihood ratio tests have proper type I errors, and the confidence intervals achieve desired coverage probabilities. A sample size of 1,000 or 100 cases would be sufficient for detecting a relative risk of

1.75 or larger (power >90% at the 5% level, and >80% at the 1% level), whereas at least 5,000 subjects or 500 cases would be required to have high power (90% at the 5% level) for detecting a relative risk of 1.25 or smaller. When  $\beta = 0$ , the additive, dominant, and recessive models all hold, in which case the AIC selects the additive model about 18% of the time. For nonzero  $\beta$ , only the additive model is true, and the probability of selecting the true model increases with increasing values of  $\beta$  and  $n$ . Comparisons of the variances and powers between the situations of unphased and phased genotypes reveal that unknown gametic phase incurs little loss of efficiency.

Hardy-Weinberg equilibrium is assumed in the theoretical development. This assumption may be violated in practice. Table III shows the performance of the inferential procedures under the same setup as in Table II, but with a common fixation index [Weir, 1996, p. 93]  $f = 0.2$  for each haplotype pair, which corresponds to a severe departure from Hardy-Weinberg equilibrium. The maximum likelihood estimator remains unbiased, and the confidence intervals maintain proper coverage probabilities. The power of the association test and the performance of the AIC turn out to be better under  $f = 0.2$  than under  $f = 0$ ,

TABLE II. Summary statistics for simulation studies under Hardy-Weinberg equilibrium

$\beta$	$n$	Phase unknown						Phase known			
		Bias <sup>a</sup>	SE <sup>b</sup>	Size/power <sup>c</sup>		Coverage <sup>d</sup>		AIC <sup>e</sup>	SE <sup>b</sup>	Size/power <sup>c</sup>	
				$\alpha=0.05$	0.01	$\alpha=0.05$	0.01			$\alpha=0.05$	0.01
0	1,000	-0.008	0.165	0.053	0.011	0.947	0.989	0.18	0.161	0.054	0.012
	2,000	-0.004	0.115	0.050	0.011	0.950	0.989	0.18	0.112	0.050	0.011
	5,000	-0.002	0.073	0.052	0.010	0.948	0.990	0.17	0.071	0.053	0.011
log 1.25	1,000	-0.006	0.161	0.289	0.120	0.949	0.990	0.33	0.158	0.293	0.127
	2,000	-0.003	0.112	0.498	0.272	0.949	0.990	0.44	0.110	0.514	0.288
	5,000	-0.001	0.071	0.873	0.700	0.947	0.989	0.62	0.070	0.888	0.722
log 1.5	1,000	-0.005	0.155	0.722	0.493	0.949	0.990	0.54	0.153	0.738	0.512
	2,000	-0.003	0.108	0.952	0.852	0.949	0.989	0.71	0.107	0.956	0.870
	5,000	-0.001	0.068	1.000	1.000	0.948	0.990	0.89	0.067	1.000	1.000
log 1.75	1,000	-0.004	0.153	0.940	0.832	0.947	0.991	0.70	0.151	0.946	0.848
	2,000	-0.002	0.107	0.998	0.993	0.949	0.990	0.85	0.106	0.999	0.994
	5,000	-0.001	0.067	1.000	1.000	0.949	0.991	0.97	0.066	1.000	1.000

<sup>a</sup>Mean of estimator of  $\beta$  minus true value of  $\beta$ .

<sup>b</sup>Standard error of estimator of  $\beta$ .

<sup>c</sup>Size ( $\beta=0$ ) or power ( $\beta \neq 0$ ) of likelihood ratio test for testing  $H_0 : \beta=0$

<sup>d</sup>Coverage probability of  $(1-\alpha)$  100% confidence interval.

<sup>e</sup>Proportion that true model is selected by AIC.

TABLE III. Summary statistics for simulation studies under Hardy-Weinberg disequilibrium

$\beta$	$n$	Phase unknown							Phase known		
		Bias <sup>a</sup>	SE <sup>b</sup>	Size/power <sup>c</sup>		Coverage <sup>d</sup>		AIC <sup>e</sup>	SE <sup>b</sup>	Size/power <sup>c</sup>	
				$\alpha=0.05$	0.01	$\alpha=0.05$	0.01			$\alpha=0.05$	0.01
0	1,000	-0.007	0.149	0.053	0.011	0.947	0.989	0.16	0.147	0.052	0.011
	2,000	-0.003	0.104	0.051	0.011	0.949	0.989	0.16	0.103	0.052	0.010
	5,000	-0.001	0.066	0.050	0.011	0.950	0.989	0.15	0.065	0.052	0.011
log 1.25	1,000	-0.005	0.143	0.339	0.158	0.949	0.990	0.33	0.141	0.348	0.163
	2,000	-0.002	0.100	0.589	0.355	0.949	0.991	0.44	0.099	0.602	0.365
	5,000	-0.001	0.063	0.932	0.812	0.950	0.989	0.65	0.062	0.938	0.825
log 1.5	1,000	-0.004	0.136	0.816	0.625	0.950	0.991	0.56	0.135	0.823	0.638
	2,000	-0.002	0.096	0.981	0.935	0.948	0.991	0.73	0.095	0.984	0.939
	5,000	-0.001	0.060	1.000	1.000	0.950	0.990	0.91	0.060	1.000	1.000
log 1.75	1,000	-0.003	0.133	0.979	0.924	0.950	0.991	0.73	0.132	0.980	0.929
	2,000	-0.001	0.094	1.000	0.999	0.949	0.991	0.88	0.093	1.000	0.999
	5,000	-0.001	0.059	1.000	1.000	0.951	0.990	0.98	0.059	1.000	1.000

<sup>a</sup>Mean of estimator of  $\beta$  minus true value of  $\beta$ .

<sup>b</sup>Standard error of estimator of  $\beta$ .

<sup>c</sup>Size ( $\beta=0$ ) or power ( $\beta \neq 0$ ) of likelihood ratio test for testing  $H_0: \beta=0$

<sup>d</sup>Coverage probability of  $(1-\alpha)$  100% confidence interval.

<sup>e</sup>Proportion that true model is selected by AIC.

because the disequilibrium increases the frequency of the homozygous haplotype pair of 01100.

The results in Tables II and III pertain to the additive model. Additional simulation results (not shown here) revealed that the inferential procedures also perform well under the dominant, and recessive models, except that the power tends to be lower under the dominant model than under the additive model and even lower under the recessive model. Further simulation studies showed that the proposed methods continue to perform well when there are environmental effects.

The disease rate of 10% and haplotype frequency of 0.2514 used in the simulations may be too high for some applications. Additional simulation studies indicated that the bias of the maximum likelihood estimator continues to be negligible, and the coverage of the confidence interval remains adequate for rare diseases and uncommon haplotypes, although the power tends to decrease. The power depends primarily on the number of cases, rather than the number of subjects. Thus, a sample size of 50,000 with disease rate of 1% has similar power to the sample size of 5,000 with disease rate of 10%. The shapes of the failure time and censoring time distributions have little effect on the performance of the proposed methods.

## DISCUSSION

We developed valid and efficient methods for estimating and testing the effects of specific haplotype configurations on the risk of disease in cohort studies of unrelated individuals. The semiparametric proportional hazards formulation yields readily interpretable relative-risk parameters, while allowing the underlying distribution to be completely unspecified. Theoretical and numerical studies show that likelihood-based inference procedures have desirable properties.

For testing the null hypothesis that  $\beta = 0$ , the proposed methods are nonparametric, in that no assumption is imposed on the risk of disease. The only genetic model involved is Hardy-Weinberg equilibrium. The proposed methods are robust to the violation of this assumption, as demonstrated in our simulation studies. We are currently developing methods that do not require Hardy-Weinberg equilibrium. For estimation of haplotype effects and for hypothesis testing involving a subset of  $\beta$ , the proportional hazards assumption is made. It is possible to develop robust inference procedures under misspecified proportional hazards models along the lines of Lin and Wei [1989]. In addition, model-checking techniques analogous to those of Lin et al. [1993] can be derived. We are currently working on such methods.

A naive approach to haplotype analysis would be to infer or impute individuals' haplotypes by an EM algorithm or a Bayesian method, and then use the imputed values in standard Cox regression analysis. Because of the nonlinear nature of the Cox model, such an analysis would be biased even if the imputed values are "unbiased." Furthermore, it would be difficult to properly account for the extra variation due to imputation in the inference on haplotype effects. The proposed approach integrates such a two-stage procedure into a single-stage likelihood-based inference, which is rigorous and efficient.

Although we focused on SNP-based haplotypes in this article, the proposed methods are applicable to microsatellite loci and other genotype data. Furthermore, the proposed methods can accommodate missing genotype data, as indicated previously.

As demonstrated in the simulation studies, the proposed EM algorithm is fast and has desirable convergence properties. For the kind of haplotype distribution used in the simulations, the computing is trivial, even for large cohorts. Naturally, the computing time increases with an increasing number of haplotypes. For a large number of SNPs, the partition-ligation method of Niu et al. [2002] and Qin et al. [2002] can be adapted to reduce the computing burden. We plan to implement the proposed methods in a software package that will be made available to the general public.

The proposed methods assume that, except for random missingness, genotype data are available on all cohort members. Because of continuing reduction in genotyping cost, it is realistic to genotype all subjects in a large cohort study. In fact, this is being done in the motivating ARIC Study. For very large cohorts with rare diseases, it would be cost-effective to adopt the case-cohort design [Prentice, 1986] or nested case-control design [Thomas, 1977], so that only a subset of the cohort members needs to be genotyped. We are currently extending the approach taken in this article to such designs.

As mentioned above, the power of the association test tends to be low for rare haplotypes. Roughly speaking, the variance of the test statistic is inversely proportional to  $r(1-r)$ , where  $r$  is the proportion of the haplotype of interest in the sample. If  $r$  is very small, the power will be unacceptably low, in which case alternative analysis strategies may be preferred. If several

haplotypes have similar relative risks, then one may collapse those haplotypes or combine the relative-risk estimates so as to increase power and precision.

It is widely recognized that the presence of latent population stratification/substructure can cause bias in association studies based on unrelated individuals. A number of statistical methods have been developed to adjust for the effects of population substructure, given the existence of a series of genomic markers that is informative about the population substructure. None of these methods deal with censored phenotypes or haplotype analysis. It would be possible to extend the proposed methods so as to accommodate potential population substructure.

This article is concerned with cohort studies of unrelated individuals. There are several ongoing family-based cohort studies, including the Family Heart Study [Higgins et al., 1996], which also require haplotype analysis. We plan to develop appropriate haplotype-based association methods for family studies with unphased genotypes and potentially censored age-of-onset phenotypes.

## ACKNOWLEDGMENTS

The author is grateful to the Editor-in-Chief and two referees for their readings of the paper, and to Dr. Gerardo Heiss for his helpful discussion regarding the ARIC Study and the Family Heart Study.

## REFERENCES

- Akaike H. 1985. Prediction and entropy. In: Atkinson AC, and 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.
- ARIC Investigators. 1989. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. *Am J Epidemiol* 129:687-702.
- 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 NE. 1972. Discussion of the paper by D.R. Cox. *J R Stat Soc B* 34:216-217.
- Breslow NE, Day NE. 1987. Statistical methods in cancer research: the design and analysis of cohort studies. Lyon: International Agency for Research on Cancer.

- Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, Lockhart DJ, Morris MS, Fodor SPA. 1996. Accessing genetic information with high density DNA arrays. *Science* 274:610–614.
- Clark AG. 1990. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol Biol Evol* 7:111–122.
- Cordell HJ, Clayton DG. 2002. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *Am J Hum Genet* 70:124–141.
- 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.
- 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.
- Drysdale CM, McGraw DW, Stack CB, Stephens JC, Judson RS, Nandabalan K, Arnold K, Ruano G, Liggett SB. 2000. Complex promoter and coding region  $\beta_2$ -adrenergic receptor haplotypes alter receptor expression and predict *in vivo* responsiveness. *Proc Natl Acad Sci USA* 97:10483–10488.
- 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.
- Fried LP, Borhani NO, Enright P, Furberg CD, Gardin JM, Kronmal RA, Kuller LH, Manolio TA, Mittelmark MB, Newman A, O'Leary D, Psaty B, Rautaharju P, Tracy R. 1991. The Cardiovascular Health Study: design and rationale. *Ann Epidemiol* 1:263–276.
- Hallman DM, Groenemeijer BE, Jukema JW, Boerwinkle E. 1999. Analysis of lipoprotein lipase haplotypes reveals associations not apparent from analysis of the constitute loci. *Ann Hum Genet* 63:499–510.
- Hawley ME, Kidd KK. 1995. HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J Hered* 86:409–411.
- Higgins M, Province M, Heiss G, Eckfeldt J, Ellison RC, Folsom AR, Rao DC, Sprafka M, Williams R. 1996. NHLBI Family Heart Study: objectives and design. *Am J Epidemiol* 143:1219–1228.
- International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921.
- International SNP Map Working Group. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409:928–933.
- Johnson SR, Anderson GL, Barad DH, Stefanick ML. 1999. The Women's Health Initiative: rationale, design, and progress report. *J Br Menopause Soc* 5:155–159.
- Lin DY, Wei LJ. 1989. The robust inference for the Cox proportional hazards model. *J Am Stat Assoc* 84:1074–1078.
- Lin DY, Wei LJ, Ying Z. 1993. Checking the Cox model with cumulative sums of Martingale-based residuals. *Biometrika* 80:557–572.
- Long JC, Williams RC, Urbanek M. 1995. An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am J Hum Genet* 56:799–810.
- Morris RW, Kaplan NL. 2002. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet Epidemiol* 23:221–233.
- 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.
- Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294:1719–1723.
- Prentice RL. 1986. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* 73:1–11.
- 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, Merikangas K. 1996. The future of genetic studies of complex human diseases. *Science* 273:1516–1517.
- 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.
- 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 L, 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.
- Terwilliger JD, Ott J. 1994. Handbook of human genetic linkage. Baltimore: Johns Hopkins University Press.
- Thomas DC. 1977. Addendum to: methods of cohort analysis: appraisal by application to asbestos mining. By Liddell FDK, McDonald JC, Thomas DC. *J R Stat Soc A* 140:469–491.
- van der Vaart AW. 1995. Efficiency of infinitely dimensional estimators. *Stat. Neerl.* 49:9–30.
- van der Vaart AW. 1998. Asymptotic statistics. Cambridge: Cambridge University Press.
- van der Vaart AW, Wellner JA. 1996. Weak convergence and empirical processes. New York: Springer.
- Venter J, Adams M, Myers E, Li P, Mural R, Sutton G, Smith H, et al. 2001. The sequence of the human genome. *Science* 291:1304–1351.
- Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, et al. 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280:1077–1082.
- 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.
- Zhang S, Pakstis A, Kidd K, 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 JH, Curtis D, Sham PC. 2000. Model-free analysis and permutation tests for allelic associations. *Hum Hered* 50:133–139.
- 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 MAXIMUM LIKELIHOOD ESTIMATION

The data with known phase information would be  $(Y_i, \Delta_i, X_i, H_i)$  ( $i = 1, \dots, n$ ), which are referred to as the full data. Write  $\theta = (\beta, \pi, \Lambda)$ . The full-data log-likelihood for  $\theta$  is, up to a constant,

$$l_F(\theta) = \sum_{i=1}^n \sum_{k,l} I\{H_i = (h_k, h_l)\} [\Delta_i \log \Lambda\{Y_i\} + \Delta_i \beta^T Z_{ikl} - e^{\beta^T Z_{ikl}} \Lambda(Y_i) + \log \pi_k + \log \pi_l].$$

The conditional expectation of  $l_F(\theta)$ , given the observable data  $(Y_i, \Delta_i, X_i, G_i)$  ( $i = 1, \dots, n$ ) is

$$\tilde{l}(\theta) = \sum_{i=1}^n \sum_{k,l} \rho_{ikl}(\theta) [\Delta_i \log \Lambda\{Y_i\} + \Delta_i \beta^T Z_{ikl} - e^{\beta^T Z_{ikl}} \Lambda(Y_i) + \log \pi_k + \log \pi_l], \quad (\text{A1})$$

where  $\rho_{ikl}(\theta) = \Pr\{H_i = (h_k, h_l) | Y_i, \Delta_i, X_i, G_i\}$  ( $i = 1, \dots, n; k, l = 1, \dots, K$ ), which pertains to the posterior distribution of the haplotype. It follows from Bayes' rule that, under the assumptions stated in Data and Assumptions,

$$\rho_{ikl}(\theta) = \frac{I\{(h_k, h_l) \in \mathcal{S}(G_i)\} \exp\{\Delta_i \beta^T Z_{ikl} - e^{\beta^T Z_{ikl}} \Lambda_0(Y_i)\} \pi_k \pi_l}{\sum_{k',l'} I\{(h_{k'}, h_{l'}) \in \mathcal{S}(G_i)\} \exp\{\Delta_i \beta^T Z_{ik'l'} - e^{\beta^T Z_{ik'l'}} \Lambda_0(Y_i)\} \pi_{k'} \pi_{l'}}.$$

We apply the EM algorithm to (A1): the E-step evaluates  $\rho_{ikl}$ , while the M-step maximizes  $\tilde{l}(\theta)$ , given  $\rho_{ikl}$ . Specifically, the  $(m+1)$ th iteration consists of replacing  $\rho_{ikl}(\theta)$  in (A1) with  $\rho_{ikl}(\hat{\theta}^{(m)})$  and maximizing the resultant  $\tilde{l}(\theta)$  over  $\beta, \pi$ , and the jump sizes of  $\Lambda(t)$  at the  $Y_i$  associated with  $\Delta_i = 1$ . The estimator of  $\pi$  at the  $(m+1)$ th iteration takes the form

$$\hat{\pi}_k^{(m+1)} = \frac{\sum_{i=1}^n \sum_{l=1}^K \rho_{ikl}(\hat{\theta}^{(m)})}{n}, k = 1, \dots, K. \quad (\text{A2})$$

The estimator  $\hat{\beta}^{(m+1)}$  for  $\beta$  at the  $(m+1)$ th iteration is the root of the estimating function

$$U^{(m+1)}(\beta) = \sum_{i=1}^n \Delta_i \left\{ \sum_{k,l} \rho_{ikl}(\hat{\theta}^{(m)}) Z_{ikl} - \frac{\sum_{j=1}^n I(Y_j \geq Y_i) \sum_{k,l} \rho_{jkl}(\hat{\theta}^{(m)}) e^{\beta^T Z_{jkl}} Z_{jkl}}{\sum_{j=1}^n I(Y_j \geq Y_i) \sum_{k,l} \rho_{jkl}(\hat{\theta}^{(m)}) e^{\beta^T Z_{jkl}}} \right\}, \quad (\text{A3})$$

and the corresponding estimator of  $\Lambda_0(t)$  is

$$\hat{\Lambda}^{(m+1)}(t) = \sum_{i=1}^n \frac{I(Y_i \leq t) \Delta_i}{\sum_{j=1}^n I(Y_j \geq Y_i) \sum_{k,l} \rho_{jkl}(\hat{\theta}^{(m)}) e^{\hat{\beta}^{(m+1)T} Z_{jkl}}} \quad (\text{A4})$$

Equation (A2) is similar to the M-step in the conventional EM estimation of haplotype frequencies [Excoffier and Slatkin, 1995], except that the posterior probabilities  $\rho_{ikl}(\theta)$  are conditional on the disease phenotype in addition to the genotype. Equations (A3) and (A4) are reminiscent of the familiar partial likelihood score function for  $\beta$  and the Breslow estimator for  $\Lambda_0(t)$ , and would in fact reduce to the latter if the phase information were known. As in the case of the partial likelihood score function, the negative derivative matrix of  $U^{(m+1)}(\beta)$  with respect to  $\beta$  is always positive semidefinite, so that the standard Newton-Raphson algorithm can be used to solve the equation  $U^{(m+1)}(\beta) = 0$ .

To start the EM algorithm, we first estimate  $\pi$  under  $\beta = 0$ . This can be accomplished by using the conventional EM algorithm for estimating haplotype frequencies [Excoffier and Slatkin, 1995]. Given the initial estimate of  $\pi$ , we iterate

the proposed EM algorithm until the average change in the parameter estimates between two consecutive iterations is less than certain convergence criterion, such as  $10^{-4}$  or  $10^{-5}$ .

## APPENDIX B

### ASYMPTOTIC PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATORS

Here, we describe the asymptotic properties of the maximum likelihood estimators  $\hat{\beta}$ ,  $\hat{\pi}$ , and  $\hat{\Lambda}$ . We also indicate how these theoretical results are derived. The proofs involve advanced techniques from modern empirical process theory [van der Vaart and Wellner, 1996] and semiparametric estimation theory [Bickel et al., 1993]. The technical details are beyond the scope of this article, but are available from the author.

Write  $l(\theta) = \log L(\theta)$ , where  $L(\theta)$  is given in (2). Let  $\theta_0$  be the true value of  $\theta$ , and  $\hat{\theta}$  be the

maximum likelihood estimator of  $\theta_0$ . We claim that  $\widehat{\theta}$  is consistent for  $\theta_0$ . To prove this claim, we first show that  $\widehat{\Lambda}(t)$  is not allowed to diverge to infinity. Since  $\widehat{\theta}$  maximizes  $l(\theta)$ , it is clear that  $l(\widehat{\theta}) - l(\theta) \geq 0$  for any  $\theta$ . We construct a function  $\Lambda^*(t)$  by replacing  $\widehat{\beta}$  and  $\widehat{\pi}$  in  $\widehat{\Lambda}(t)$  with  $\beta_0$  and  $\pi_0$ . We then show that, if  $\widehat{\Lambda}(t)$  diverges, then  $l(\widehat{\beta}, \widehat{\pi}, \widehat{\Lambda}) - l(\beta_0, \pi_0, \Lambda^*)$  must be negative eventually, which results in a contradiction. Because  $\widehat{\Lambda}(t)$  is not allowed to diverge, Helly's selection theorem implies that there exists a convergent subsequence of  $\widehat{\theta}$  which converges to a well-defined limit, say,  $\bar{\theta}$ . The fact that  $l(\bar{\theta}) - l(\theta_0) \geq 0$  for any  $n$ , whereas due to the positiveness of the Kullback-Leibler information, the limit of  $l(\theta)$ , say  $\bar{l}(\theta)$ , is maximized at  $\theta_0$  entails that  $\bar{l}(\bar{\theta}) = \bar{l}(\theta_0)$ . The problem now reduces to the identifiability of the parameters: the parameters are identifiable if and only if  $\bar{l}(\theta) = \bar{l}(\theta_0)$  implies that  $\theta = \theta_0$ . It can be shown that the parameters are indeed identifiable, given the assumptions stated in Data and Assumptions, together with some mild regularity conditions.

To establish the asymptotic distribution of  $\widehat{\theta}$ , we consider parametric submodels of the form:  $\theta_\epsilon = \theta + \epsilon(d_1, d_2, \int_0^t d_3(s)d\Lambda(s))$ , where  $d_1$  and  $d_2$

are constant vectors with the same dimensions as  $\beta$  and  $\pi$ , respectively, and  $d_3(t)$  is a function with bounded variation. The score operator is given by  $S(\theta) = n^{-1}dl(\theta_\epsilon)/d\epsilon|_{\epsilon=0}$ . Let  $\bar{S}(\theta)$  be the limit of  $S(\theta)$ . By the Donsker theorem [van der Vaart and Wellner, 1996, p. 130],  $n^{1/2}\{S(\theta_0) - \bar{S}(\theta_0)\}$  converges to a zero-mean Gaussian process. We can verify that  $\bar{S}(\theta)$  is Fréchet-differentiable, and its derivative at  $\theta_0$  is continuously invertible. It then follows from Theorem 3.3.1 of van der Vaart and Wellner [1996, p. 310] that  $n^{1/2}(\widehat{\theta} - \theta_0)$  converges to a zero-mean Gaussian process whose covariance matrix can be consistently estimated by the inverse of the observed Fisher information matrix. Specifically, the general random variable  $n^{1/2}[d_1^T \times (\widehat{\beta} - \beta_0) + d_2^T(\widehat{\pi} - \pi_0) + \int_0^\infty d_3(t)d\{\widehat{\Lambda}(t) - \Lambda_0(t)\}]$  is asymptotically zero-mean normal with a variance that is consistently estimated by  $nd^T \mathcal{I}^{-1}d$ , where  $d$  is the vector comprising  $d_1, d_2$ , and the values of  $d_3(Y_i)$  associated with  $\Delta_i = 1$ , and  $\mathcal{I}$  is the negative second derivative matrix of  $\mathcal{J}(\theta)$  with respect to  $\beta, \pi$ , and the jump sizes of  $\Lambda(t)$  at the  $Y_i$  for which  $\Delta_i = 1$  evaluated at  $\theta = \bar{\theta}$ . Finally, it follows from Proposition 1 of van der Vaart [1995] that  $\widehat{\theta}$  is efficient in that it has the smallest asymptotic variance among all regular estimators.