Original Paper

Human Heredity

Hum Hered 2011;71:148–160
DOI: 10.1159/000324841

Received: September 27, 2010
Accepted after revision: February 3, 2011
Published online: July 20, 2011

A Bayesian Hierarchical Model for Detecting Haplotype-Haplotype and Haplotype-Environment Interactions in Genetic Association Studies

Jun Li Kui Zhang Nengjun Yi

Department of Biostatistics, Section on Statistical Genetics, University of Alabama at Birmingham, Birmingham, Ala., USA

Key Words

Bayesian methods · Generalized linear models · Genetic associations · Hierarchical models · Haplotype · Haplotype-haplotype interactions · Haplotype-environment interactions

Abstract

Objective: Genetic association studies based on haplotypes are powerful in the discovery and characterization of the genetic basis of complex human diseases. However, statistical methods for detecting haplotype-haplotype and haplotype-environment interactions have not yet been fully developed owing to the difficulties encountered: large numbers of potential haplotypes and unknown haplotype pairs. Furthermore, methods for detecting the association between rare haplotypes and disease have not kept pace with their counterpart of common haplotypes. ***Methods/Results:*** We herein propose an efficient and robust method to tackle these problems based on a Bayesian hierarchical generalized linear model. Our model simultaneously fits environmental effects, main effects of numerous common and rare haplotypes, and haplotype-haplotype and haplotype-environment interactions. The key to the approach is the use of a continuous prior distribution on coefficients that favors sparseness in the fitted model and facilitates computation. We develop a fast expectation-maximization algorithm to fit models by estimating posterior modes of coefficients. We incorporate our algorithm into the iteratively weighted least squares for classical generalized linear models as implemented in the R package glm. We evaluate the proposed method and compare its performance to existing methods on extensive simulated data. ***Conclusion:*** The results show that the proposed method performs well under all situations and is more powerful than existing approaches.

Introduction

Genome-wide and candidate gene association studies based on linkage disequilibrium (LD) are cost effective and statistically efficient tools to unravel the genetic architecture of complex human diseases [1, 2]. Although methods based on individual single nucleotide polymorphisms (SNPs) may lead to significant findings [2, 3], haplotypes comprising multiple SNPs on the same inherited chromosomal region have long been of great interest and have attracted much attention in recent years [4–7]. First, haplotypes are biologically relevant. There is strong evi-

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

Accessible online at:
www.karger.com/hhe

Nengjun Yi
Department of Biostatistics
University of Alabama at Birmingham
Birmingham, AL 35294-0022 (USA)
Tel. +1 205 934 4924, E-Mail nyi@ms.soph.uab.edu

dence that several mutations within a gene may interact (*cis*-interaction) to cause a disease [8]. Haplotype-based methods provide a natural way to capture such *cis*-interactions [9]. Second, the power of single-marker-based methods in association studies depends on LD between the tested marker and the disease susceptibility locus. LD information contained in flanking markers is generally not incorporated, which can result in a reduction in power [10]. Therefore, haplotype-based association methods can be more powerful than those based on single markers since the former fully exploits LD information from multiple markers [9, 11]. Both simulation and empirical studies also support this conclusion [11, 12]. In addition, haplotype-based methods can be more powerful when multiple disease susceptibility alleles occur within a single gene [9].

A number of statistical methods have been proposed to examine the association between haplotypes and human complex diseases [7, 12, 13]. Although many of these approaches have been widely used in association studies, the majority of them only focus on the estimation of marginal effects of haplotypes and the detection of the association between common haplotypes and diseases, while little attention has been paid so far to developing statistical methods for investigating the interacting effects between haplotypes and environmental factors, especially those between haplotypes in different haplotype blocks, and exploring disease associations with rare haplotypes [14, 15].

Increasing evidence suggests that gene-gene and gene-environment interactions play an important role in susceptibility to complex human diseases [16–20]. The investigation of such interactions may provide more insight into disease etiology and ultimately result in new strategies for the treatment and prevention of a disease. As an earlier attempt to explore the interaction between haplotypes and environmental factors, Lake et al. [21] proposed a likelihood-based method in the generalized linear model framework, which has been commonly employed in haplotype-based association studies, because it is publicly available and easy to implement with its R package. This approach, however, is limited by ignoring the interacting effects between haplotype blocks. Subsequently, several methods have been developed to study haplotype-related interactions, but these methods do not consider all potential haplotypes and interactions simultaneously [22–27]. Recently, Guo and Lin [15] proposed a generalized linear model with regularization to detect interacting haplotype effects. However, their method applies an omnibus test and consequently does not provide inference on the effects of specific haplotypes and their interactions. Another concern for haplotype-based methods is that large numbers of haplotypes inferred from genotype data [28–30] often lead to high degrees of freedom for corresponding statistical tests and thus reduce power [14, 15, 31–34]. If interacting effects are considered, such problems become severer. Unfortunately, few statistical methods have been developed to tackle these far-reaching problems in the study of haplotype interactions.

One challenge in haplotype-based association studies is that for haplotypes comprising multiple markers, there might be many rare haplotypes. Because of their low frequencies, the parameter estimates related to rare haplotypes will have large variances, leading to unstable models. Schaid [6] described several approaches to handle rare haplotypes. One approach is to combine all rare haplotypes into one group or rare haplotypes with common ancestral haplotypes, and another is to exclude them from the model, which is equivalent to including them in the baseline group. However, both approaches yield results that may be difficult to interpret. In addition, it has been argued that rare haplotypes may account for a substantial fraction of the multifactorial inheritance of common diseases [35–39], thus the aforementioned approaches may miss the rare haplotypes having true effects. Another approach is to include the effects of each rare haplotype in the model but shrink their effects towards the common mean or towards the effect of a similar haplotype [6, 15, 40]. Recently, Guo and Lin [15] adopted a lasso penalty in their generalized linear model to allow assessment of the effects of rare haplotypes by decreasing the coefficients of unassociated haplotypes to zero so that the associated ones, especially those that are rare, can be estimated.

In this article, we propose a new approach to investigate the association between haplotypes and human diseases based on the hierarchical generalized linear model. The proposed method is built upon a Bayesian framework with weakly informative priors on the coefficients. It can deal with various types of phenotypes in population-based association studies, and can simultaneously fit a large number of effects, including main effects of numerous common and rare haplotypes, environmental effects, haplotype-haplotype interactions, and haplotype-environment interactions. We fit our Bayesian generalized linear models by incorporating an expectation-maximization (EM) algorithm into the usual iteratively weighted least squares as implemented in the R package glm. This strategy leads to stable and flexible computational tools and allows us to apply any generalized linear

model to haplotype-based association studies. We investigated the statistical properties and performance of the proposed method and compared it with three existing methods, the classical generalized linear model, the method of Lake et al. [21], and the method of Guo and Lin [15], through extensive simulation studies. The results show that the proposed method performs well under all situations and is more powerful than the existing approaches.

Methods

Generalized Linear Models of Interacting Haplotypes

Suppose that a population-based association study consists of n unrelated individuals, phenotyped for a disease trait, and genotyped for multiple genetic variants (e.g. SNPs) in multiple genomic regions or haplotype blocks. Although our method can deal with various phenotypes, we demonstrate its performance with a binary disease trait as in case-control studies. We use generalized linear models to relate disease status to haplotypes and environmental factors. We simultaneously fit environmental *(E)* effects, main effects of haplotypes *(H)*, and haplotype-haplotype *(H×H)*, and haplotype-environment *(H×E)* interactions. The generalized linear model is expressed as

$$h(\Pr(y_i = 1)) = \beta_0 + \boldsymbol{X}_E\boldsymbol{\beta}_E + \boldsymbol{X}_H\boldsymbol{\beta}_H + \boldsymbol{X}_{HH}\boldsymbol{\beta}_{HH} + \boldsymbol{X}_{HE}\boldsymbol{\beta}_{HE})_i = \boldsymbol{X}_i\boldsymbol{\beta} \quad (1)$$

where h is a link function or transformation which relates the linear predictor $\boldsymbol{X}_i\boldsymbol{\beta}$ to the disease probability $\Pr(y_i = 1)$; β_0 is the intercept, $\boldsymbol{\beta}_E$ and $\boldsymbol{\beta}_H$ are the vectors of environmental effects and all possible haplotype main effects, respectively; $\boldsymbol{\beta}_{HH}$ is the vector of all possible haplotype-haplotype interactions between different regions; $\boldsymbol{\beta}_{HE}$ is the vector of haplotype-environment interactions, and $\boldsymbol{X}_E$, $\boldsymbol{X}_H$, $\boldsymbol{X}_{HH}$, and $\boldsymbol{X}_{HE}$ are the corresponding design matrices of effect predictors. Various link functions are provided in generalized linear models [41], all of which can be adapted in our Bayesian framework. For case-control studies, we can use the logit or probit link function.

Construction of the Design Matrices

Since haplotypes are usually not directly measured, the posterior probabilities of haplotype pairs are first computed from the observed genotype data for each subject to account for this ambiguity using existing methods of haplotype inference [12, 28–30]. These posterior probabilities are then used to compute the estimates of haplotype dosage [42].

The estimate of haplotype dosage is the estimate of the number of copies of a specific haplotype for a subject. For the haplotypes that can be unambiguously resolved based on the observed genotype data, the values of haplotype dosage of a haplotype for a subject can be 0, 1, or 2. But for the haplotypes that cannot be unambiguously resolved, the values of haplotype dosage of a haplotype for a subject would be non-integer, ranging from zero to two, reflecting the possibility that haplotypes are based on the subject's genotypes. For each subject, the sum of haplotype dosage across all haplotypes is equal to two. After obtaining the estimates of haplotype dosage, they can be used to construct the design matrix $\boldsymbol{X}_H$.

Suppose there are W_q possible haplotypes in the q-th haplotype block in the population, $q = 1, 2, ..., Q$, and let d_{iqw}, $w = 1, 2, ..., W_q$, denote the estimate of haplotype dosage of the w-th haplotype in the q-th haplotype block for subject i. Therefore, we can set $(\boldsymbol{X}_H)_i = (d_{i12}, ..., d_{i1W}, ..., d_{iQ2}, ..., d_{iQW_Q})$.

For the environmental factors, the raw values are transformed to have a mean of 0 and a standard deviation of 0.5 [43, 44]. This transformation standardizes all the environmental effects to have a common scale. The matrices of interacting variables, $\boldsymbol{X}_{HE}$ and $\boldsymbol{X}_{HH}$, are set up by simply multiplying two corresponding realizations of $\boldsymbol{X}_E$ and $\boldsymbol{X}_H$.

Prior Distributions

The above model can include a large number of highly correlated predictors, leading to the problems of high dimensionality and collinearity that preclude the use of classical maximum likelihood methods. We handle these problems using a Bayesian approach that places appropriate prior distributions on coefficients to obtain stable estimates. We assume independent Student t priors $t_{\nu_j}(0, s_j^2)$ on coefficients β_j, with ν_j and s_j predetermined. We are motivated to use the t distribution because it allows for flexible modeling, robust inference, and easy and stable computation [43–45]. The distribution $t_{\nu_j}(0, s_j^2)$ can be expressed as a mixture of normal distributions with mean 0 and variance distributed as scaled inverse-χ^2:

$$\beta_j \,|\, \tau_j^2 \sim N(0, \tau_j^2),\ \tau_j^2 \sim \text{Inv-}\chi^2(\nu_j, s_j^2),\ j = 0, 1, ..., J, \quad (2)$$

where J is the total number of effects in the model, and the hyperparameters $\nu_j > 0$ and $s_j > 0$ represent the degrees of freedom and the scale of the distribution, respectively.

The hyperparameters ν_j and s_j control the global amount of shrinkage in the effect estimation; larger ν_j and smaller s_j^2 induce stronger shrinkage and force more effects to be near zero. The method of Yi et al. [45] is used to choose ν_j and s_j. For β_0, $\boldsymbol{\beta}_E$ and $\boldsymbol{\beta}_H$, we employ the weakly informative priors recommended by Gelman et al. [43], i.e. $(\nu_0, s_0) = (1, 10)$ for β_0, and $(\nu_j, s_j) = (1, 2.5)$ for $\boldsymbol{\beta}_E$ and $\boldsymbol{\beta}_H$. For haplotype-environment interactions $\boldsymbol{\beta}_{HE}$, we set $(\nu_j, s_j) = (1, 2.5\, k_H/k_{HE})$, where k_H and k_{HE} are the total numbers of main effects of haplotypes and haplotype-environment interactions, respectively. For haplotype-haplotype interactions $\boldsymbol{\beta}_{HH}$, we set $(\nu_j, s_j) = (1, 2.5\, k_H/k_{\text{HH}})$, where k_{HH} is the total number of haplotype-haplotype interactions. Because there are many more interactions than main effects, these priors apply more stringent restrictions on interactions and allow reliable estimates of main effects and interactions [45].

EM Algorithm for Model Fit

The EM algorithm is used to fit the hierarchical haplotype models with the Student t priors by estimating the marginal posterior modes of the coefficients β_j [44, 45]. We incorporate our algorithm into the iteratively weighted least squares for classical generalized linear models as implemented in the R package glm. The standard iteratively weighted least squares algorithm approximates a generalized linear model by a normal linear model [43, 46]. Specifically, at each iteration, pseudo-data z_i and pseudo-variances σ_i^2 are calculated for each subject i based on the current estimates of parameters, then the generalized linear model likelihood $p(y_i \,|\, \boldsymbol{X}_i\boldsymbol{\beta}, \phi)$ is approximated by a normal likelihood $N(z_i \,|\, \boldsymbol{X}_i\boldsymbol{\beta}, \sigma_i^2)$, and finally the parameters β_j are updated by a weighted linear regression.

Our EM algorithm uses the two-level expression of the *t* prior distribution and treats the unknown variances τ_j^2 as missing data. In each E-step, we replaced the unknown variances τ_j^2 by their posterior expectations,

$$\hat{\tau}_j^2 = \frac{\nu_j s_j^2 + \hat{\beta}_j^2}{1+\nu_j}$$

Given the variances τ_j^2, the priors $\beta_j \mid \tau_j^2 \sim N(0, \tau_j^2)$ are treated as additional 'data points', added to the weighted normal regression $N(z_i \mid \boldsymbol{X}_i\boldsymbol{\beta}, \sigma_i^2)$. In each M-step, the standard iteratively weighted least squares algorithm is applied to the augmented weighted normal regression to update the coefficients β_j. We implement these computations by altering the glm function in R for fitting generalized linear models, inserting the steps for calculating the augmented data and updating the variances into the iterative procedure.

The EM algorithm is initialized by setting each τ_j to a small value (say $\tau_j = 0.1$) and β_j to the starting value provided by the standard iteratively weighted least squares for the classical generalized linear model as implemented in the R function glm. We repeat the E-step and the M-step until convergence. At convergence of the algorithm, we obtain all outputs from the R function glm, including the estimates $\hat{\beta}_j$, standard errors, and p values (for testing $\beta_j = 0$). The standard errors are calculated from the inverse second derivative matrix of the log posterior density evaluated at $\hat{\beta}_j$, [43]. The p values are then determined by the estimates $\hat{\beta}_j$, and their standard errors as in the classical framework.

Simulation Study

We used extensive simulation studies to evaluate the statistical properties and performance of the proposed method. Our simulation utilized real haplotype data, the TGFBR1 haplotype-tagging SNP (htSNP) data reported in a genetic case-control association study of TGFBR1 haplotypes and risk of non-cell lung cancer [47]. The six htSNPs were partitioned into two blocks, one forming 2-SNP haplotypes and the other forming 4-SNP haplotypes, based on the estimation of Lewontin (D') and squared correlation coefficients (r^2). The haplotype frequencies were estimated for the 2-SNP and 4-SNP haplotypes, respectively, and are presented in table 1. Given these haplotype frequencies, we generated case and control subjects, assuming Hardy-Weinberg equilibrium for the haplotype pair of each individual and a logistic regression model for the disease risk. The baseline penetrance of disease (the proportion of affected subjects with a pair of non-disease-associated haplotypes) was set at 10%. A binary variable, smoking status with a proportion of 49% as in Lei et al. [47], was included in the model as a covariate and was considered in haplotype-environment interactions. The results from the proposed method (referred to as BayesGLM) were compared with those from the classical generalized linear model (referred to as GLM), the method of Lake et al. [21] (referred to as ScoreGLM), and the method of Guo and Lin [15] (referred to as rGLM). The method of Lake et al. [21] uses a generalized linear model with a two-step iteration process: the posterior probabilities of haplotype pairs per subject are used as weights to update the regression coefficients, and the regression coefficients are then used to update the posterior probabilities. The method of Lake et al. [21] has been implemented in the freely available software R/haplo.stats (http://mayoresearch.mayo.edu/mayo/research/schaid_lab/software.cfm). Guo and Lin [15] also created an R package to carry out their method and it is available free at the website: http://www.stat.osu.edu/~statgen/SOFTWARE/rGLM/.

Table 1. Haplotype patterns and their frequencies

4-SNP haplotype			2-SNP haplotype		
haplotype	pattern	frequency	haplotype	pattern	frequency
haplo4.1	1111	3.27×10^{-1}	haplo2.1	11	4.28×10^{-1}
haplo4.2	1112	2.71×10^{-2}	haplo2.2	12	3.03×10^{-1}
haplo4.3	1121	7.04×10^{-3}	haplo2.3	21	3.33×10^{-2}
haplo4.4	1211	6.64×10^{-2}	haplo2.4	22	2.36×10^{-1}
haplo4.5	1212	9.50×10^{-9}			
haplo4.6	1221	1.35×10^{-1}			
haplo4.7	1222	2.06×10^{-9}			
haplo4.8	2111	2.78×10^{-3}			
haplo4.9	2121	4.82×10^{-3}			
haplo4.10	2211	1.22×10^{-2}			
haplo4.11	2212	4.14×10^{-1}			
haplo4.12	2222	3.27×10^{-3}			

Simulation Settings

Five scenarios were posed to carry out our evaluation processes. To examine whether the proposed method can be applied to both common and rare haplotypes, we considered a rare haplotype, *haplo4.3*, two moderately rare haplotypes, *haplo4.2* and *haplo4.4*, and a common haplotype, *haplo4.1*, in the 4-SNP haplotype block, and a moderately rare haplotype, *haplo2.3*, in the 2-SNP haplotype block to be associated with the disease in the five scenarios.

In the first two scenarios, we considered only the main effects of haplotypes arising from the 4-SNP haplotype block. Specifically, in the first scenario, we assumed that *haplo4.1* and *haplo4.3* increased the odds of getting disease 2- and 3-fold, respectively, and *haplo4.2* and *haplo4.4* were not associated with the disease. In the second scenario, we assumed that *haplo4.1*, *haplo4.2*, *haplo4.3*, and *haplo4.4* increased the odds of getting disease 2-, 3-, 4-, and 3-fold, respectively, and none of the other 8 haplotypes in the 4-SNP haplotype block were associated with the disease (table 2).

In the third to fifth scenarios, we considered both the main and interacting effects arising between haplotypes in the two haplotype blocks, and between the haplotypes and smoking status. We assumed the effects in a similar way as we did in the first two scenarios (table 2). Of note, in the last scenario, we considered all the main effects of haplotypes and smoking status, and all possible interacting effects between the two haplotype blocks and between the haplotypes and smoking status. In this scenario, there are a total of 81 terms, including 17 marginal and 64 interacting terms (table 2).

Each of these five scenarios had three different sample sizes: 250, 500, and 1,000, with equal numbers of cases and controls. A

Table 2. Marginal and interacting terms and their effects in the five simulation scenarios

Scenario 1		Scenario 2		Scenario 3		Scenario 4		Scenario 5	
term	OR	term	OR	term	OR	term	OR	term	OR
haplo4.3	3	haplo4.3	4	haplo2.3:haplo4.1	4	haplo2.3:haplo4.3 smoke:haplo4.3	5	haplo2.3:haplo4.3 smoke:haplo4.3	5
haplo4.1	2	haplo4.2 haplo4.4	3	haplo4.3 smoke:haplo4.1	3	haplo4.3 haplo2.3:haplo4.1 smoke:haplo2.3	4	haplo4.3 haplo2.3:haplo4.1 smoke:haplo2.3	4
haplo4.2 haplo4.4	1	haplo4.1	2	haplo2.3 haplo4.1 smoke	2	haplo2.3, haplo4.2 haplo4.4 smoke:haplo4.1	3	haplo2.3, haplo4.2, haplo4.4 smoke:haplo4.1	3
		haplo4.5, haplo4.6 haplo4.7, haplo4.8 haplo4.9, haplo4.10 haplo4.11, haplo4.12	1	haplo4.2 haplo4.4	1	haplo4.1 smoke	2	haplo4.1 smoke	2
						haplo2.1, haplo2.2 haplo2.4, haplo4.5 haplo4.6, haplo4.7 haplo4.8, haplo4.9 haplo4.10, haplo4.11 haplo4.12	1	other 70 effects	1

':' Stands for an interaction between two terms (before and after ':').

total of 1,000 replicates were generated under each of these 15 settings. All of the generated data were analyzed using ScoreGLM, GLM, rGLM, and BayesGLM, respectively.

In summary, the following procedure of data generation, statistical analysis, and comparison of results was applied:

(1) *Genotype Data Generation.* Randomly drew 2 haplotypes (phased haplotype pairs) for each subject from the observed haplotypes (table 1).
(2) *Covariate Data Generation.* Smoking status for each subject was determined from a Bernoulli distribution with the observed proportion of smoking.
(3) *Case/Control Data Generation.* Set up the 'true' values of parameters as described in the simulation settings. Using these 'true' values as well as the generated phased haplotypes and smoking status, an individual was assigned to be a case or control according to the probabilities derived from a classical logistic regression model.
(4) *Model Fit.* The generated phased haplotypes and smoking status were used as predictors to fit four kinds of models based on ScoreGLM, GLM, rGLM, and BayesGLM, respectively.
(5) *Replication.* Steps 1–4 were repeated 1,000 times.
(6) *Statistics Calculation.* (1) Calculated 68 and 95% intervals that covered the 'true' values for each parameter in the model: $|b_j - \hat{b}_j| < z_\alpha se_j$, where b_j is the 'true' value of the j-th parameter, $j = 1, 2, ..., J$, $\hat{b}_j$ is an estimated coefficient of the j-th parameter, z_α is an upper critical value of the standard normal distribution for a desired significance level α, and se is the standard error of the estimated coefficients. (2) Calculated empirical powers for each parameter in the model: $power = 1/R\sum_{r=1}^{R} I_{(p_{rj} \le \alpha)}$, where R is the number of replicates required, p_{rj} is the p value of the j-th parameter in the r-th replicate, and α is the significance level taking three values of 0.05, 0.01, or 0.001.

Results

Nonidentifiability of Parameters in Model Fit

There was one main problem, the nonidentifiability of parameters, which was encountered in the model fit using the classical methods. This problem is first pointed out here because it frequently occurred and resulted in serious problems. Specifically, we found that the standard errors of some predictors in the models were large and hence the coefficients were essentially infinite when using haplo.glm in R/haplo.stats based on ScoreGLM or using glm in R based on GLM, whereas there was no such problem when using the proposed method, BayesGLM (data not shown). We could not evaluate the nonidentifiability of parameters when using rGLM because,

Table 3. Proportions of nonidentifiability of parameters for all of the simulation settings

Sample size	Scenario	ScoreGLM	GLM	BayesGLM
250	1	0.34	0.47	0.00
	2	0.69	0.78	0.00
	3	0.42	0.51	0.00
	4	0.79	0.88	0.00
	5	1.00	1.00	0.00
500	1	0.22	0.33	0.00
	2	0.58	0.67	0.00
	3	0.30	0.38	0.00
	4	0.68	0.73	0.00
	5	1.00	1.00	0.00
1,000	1	0.09	0.16	0.00
	2	0.39	0.54	0.00
	3	0.16	0.21	0.00
	4	0.56	0.61	0.00
	5	1.00	1.00	0.00

as mentioned earlier, rGLM can only perform an overall test based on permutation and consequently does not provide standard errors for each predictor in the model fit.

The further question that might be asked is how often and how serious the problem is. To this end, we summarized the results regarding the nonidentifiability of parameters in the model fit for all of the simulation settings in table 3. We can see that with increasing sample size, the proportions of nonidentifiability of parameters decreased in each of the first four scenarios of ScoreGLM and GLM. Under a fixed sample size, the proportions of nonidentifiability of parameters followed the order: scenario 5 > scenario 4 > scenario 2 > scenario 3 > scenario 1 for both ScoreGLM and GLM. In scenarios 2, 4, and 5 of ScoreGLM and GLM, all of the proportions exceeded 50% except that in scenario 2 of ScoreGLM with a sample size of 1,000 (39%). In contrast, in scenarios 1 and 3, only the proportion in scenario 3 of GLM with a sample size of 250 barely exceeded 50% (51%). For BayesGLM, there was no problem observed with the nonidentifiability of parameters in all of the simulation settings. Obviously, the larger the proportions of nonidentifiability of parameters, the less stable the estimated coefficients [48, 49]. Therefore, our results involving comparisons of the three methods were derived only from the replicates without the nonidentifiability of parameters in scenarios 1 and 3, unless otherwise specified.

Main Effect Model

In scenario 1, only 4 haplotypes in the 4-SNP haplotype block were modeled as main effects for the disease (table 2). The 'true' values prespecified for these 4 haplotypes were first compared to their corresponding estimated coefficients based on the four methods (left column of fig. 1). Under the sample size of 250, wider estimated 68 and 95% intervals that covered the 'true' values calculated based on BayesGLM were observed for each of 4 haplotypes compared to those calculated based on the other three methods, with the only exception that rGLM had little wider estimated intervals than BayesGLM for *haplo4.1* (top left corner of fig. 1). With the increase in sample sizes, however, the superiority of reliability of BayesGLM faded for all of the haplotypes except *haplo4.3* (middle left and bottom left corner of fig. 1), although its two coverage rates maintained a low growth rate. For all of the four methods, *haplo4.3* had lower coverage than the other haplotypes, no matter what sample sizes were considered.

In this and the following subsections, we did not consider rGLM in the evaluation of empirical power as well as type I errors because, as mentioned before, its omnibus test does not produce p values for individual effects. Therefore, the empirical powers were calculated based only on ScoreGLM, GLM, and BayesGLM for *haplo4.1* and *haplo4.3*, from which we tried to evaluate the ability of these methods to detect any disease-predisposing haplotypes. Under the sample size of 250, BayesGLM demonstrated higher probabilities for detecting genetic effects compared to both ScoreGLM and GLM (top right corner of fig. 1). Although the advantage of BayesGLM in the statistical validity was diminishing with the increase in sample size, it still persisted, especially for the rare haplotype, *haplo4.3*, and for the powers under $\alpha = 0.001$ and 0.01 (middle right and bottom right corner of fig. 1). For all of the three methods, a sample size of 500 was sufficient to detect a common haplotype with a power of 90% approximately, and a sample size of 1,000 was sufficient to identify a rare haplotype with a power of 85% approximately.

The empirical type I error rates were also calculated for *haplo4.2* and *haplo4.4* based on ScoreGLM, GLM, and BayesGLM. For the sample sizes of 250 and 500, type I error rates under $\alpha = 0.05$ were a little lower with BayesGLM than with both ScoreGLM and GLM. As the sample size went up to 1,000, all type I error rates decreased to zero.

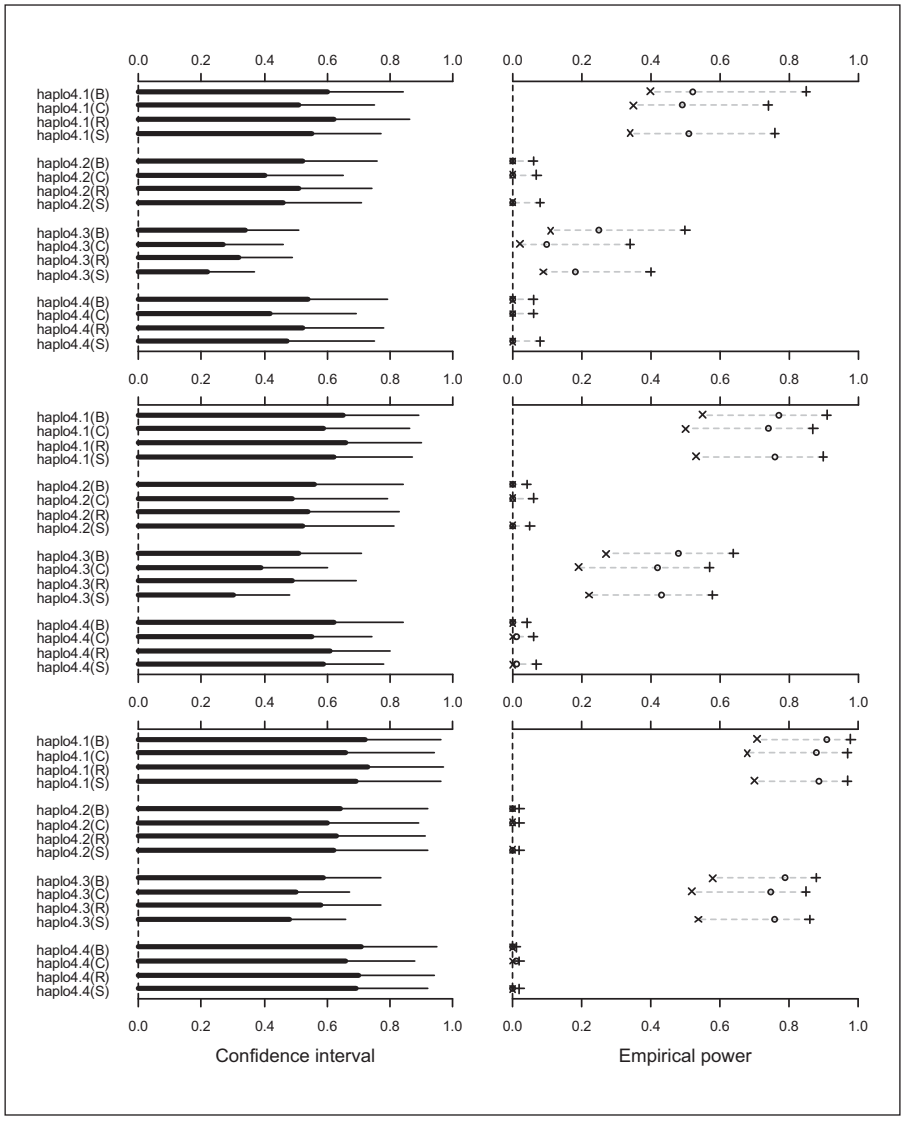

Fig. 1. Main effect model: estimated 68 and 95% coverages of the 'true' values (indicated by bold and thin horizontal lines in the left column, respectively) and empirical powers or type I error rates [empirical powers or type I error rates for α = 0.001 (×), α = 0.01 (o), and α = 0.05 (+)] for each of 4 haplotypes based on the four methods with sample sizes of 250 (top), 500 (middle), and 1,000 (bottom). B = BayesGLM; C = GLM; R = rGLM; S = ScoreGLM.

Main and Interacting Effect Model

In scenario 3, both the main and interacting effects arising between the two haplotype blocks and between the haplotypes and the environmental factor were jointly considered in the model fit for the four methods (table 2). However, since $H \times E$ interactions cannot be fitted using the current version of the rGLM, interaction between *smoke* and *haplo4.1* was set only for ScoreGLM, GLM, and BayesGLM, and since $H \times H$ interactions cannot be fitted using haplo.glm based on ScoreGLM, the interaction between *haplo2.3* and *haplo4.1* was set only for GLM, rGLM, and BayesGLM. So there were a total of 8 terms as predictors included in the model with 6 of them assumed to be disease-associated (fig. 2). Under the sample size of 250, wider estimated 68 and 95% intervals that covered the 'true' values calculated based on BayesGLM were found for each of the 8 predictors compared to those calculated based on the other three methods, with the only exception that rGLM had little wider estimated intervals than BayesGLM regarding smoking status *(smoke)* and *haplo4.1* (top left corner of fig. 2). Although the lead of BayesGLM in the statistical reliability was narrowed with the increase in sample sizes, it continued to exist, especially for the rare haplotype *haplo4.3* and the interacting terms *smoke*:*haplo4.1* and *haplo2.3*:*haplo4.1* (middle left and bottom left corner of fig. 2). For all of the four methods, the rare haplotype and the interacting terms had quite lower coverages

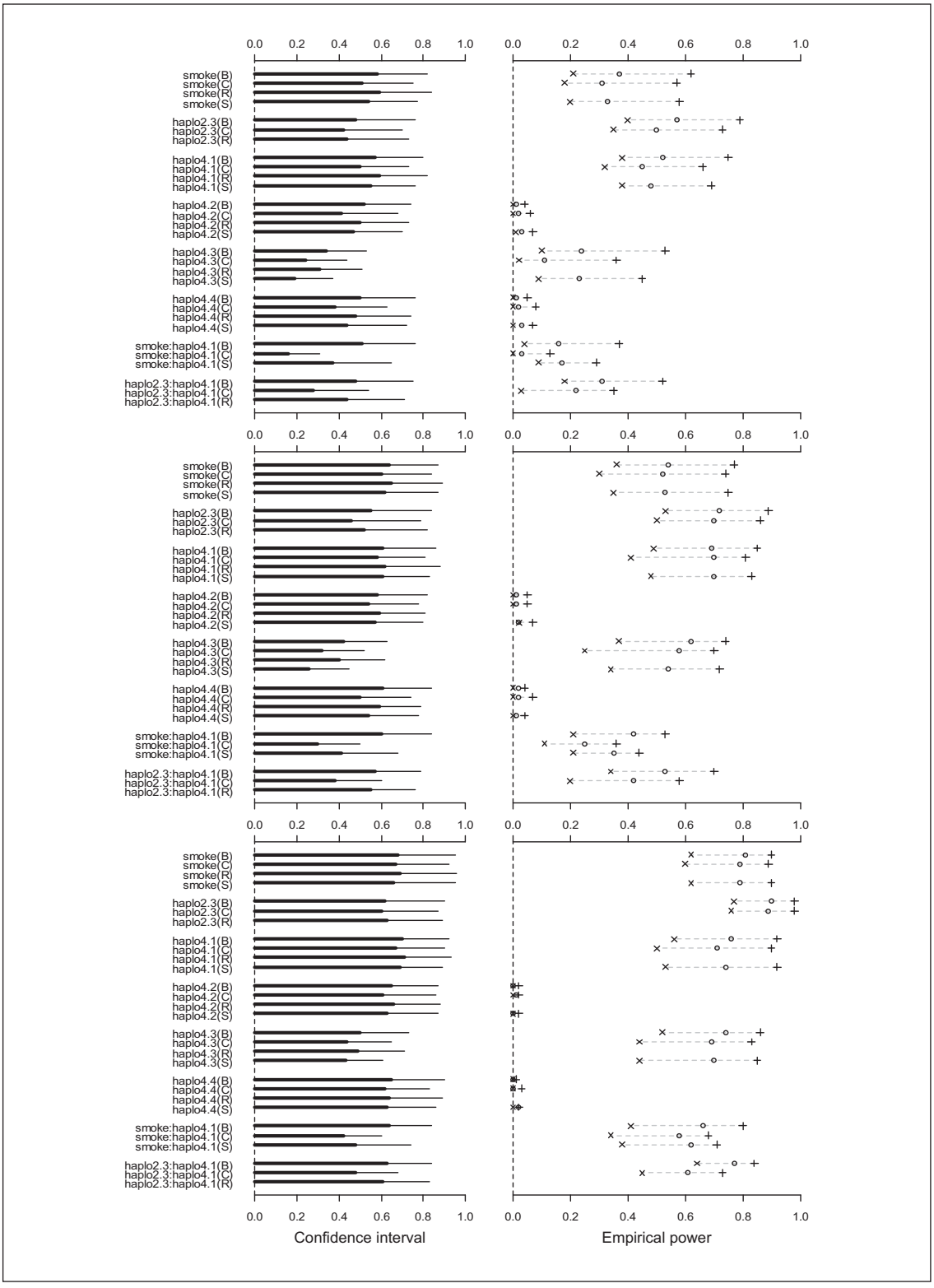

Fig. 2. Main and interacting effect model: estimated 68 and 95% coverages of the 'true' values (indicated by bold and thin horizontal lines in the left column, respectively) and empirical powers or type I error rates [empirical powers or type I error rates for α = 0.001 (×), α = 0.01 (o), and α = 0.05 (+)] for each of 8 predictors based on the four methods with sample sizes of 250 (top), 500 (middle), and 1,000 (bottom). B = BayesGLM; C = GLM; R = rGLM; S = ScoreGLM.

than the other predictors in the model, no matter what sample sizes were considered, which was in agreement with the finding in the foregoing analysis of main effects.

The empirical powers were calculated for *smoke*, *haplo2.3*, *haplo4.1*, *haplo4.3*, *smoke*:*haplo4.1*, and *haplo2.3*:*haplo4.1* based on ScoreGLM, GLM, and BayesGLM. For *smoke*, the powers were comparable for ScoreGLM, GLM, and BayesGLM, no matter what sample sizes were considered (top three lines in each of three right panels of fig. 2). This is reasonable because, for a common environmental factor with a decent frequency, any statistical test can achieve similar power for detecting it and the possible difference of powers among some tests can be explained by the random variability. For the predictors *haplo2.3*, *haplo4.1*, and *haplo4.3*, the results were almost the same as those in the preceding subsection of main effects. For *smoke*:*haplo4.1*, under the sample size of 250, BayesGLM had higher power only for $\alpha = 0.05$ compared to ScoreGLM (top right corner of fig. 2). With the increase in sample sizes, however, the situation was soon improved and eventually turned around (middle right and bottom right corner of fig. 2). For *haplo2.3*:*haplo4.1*, BayesGLM demonstrated a higher probability for correctly detecting genetically interacting effects under each of three fixed type I error rates and each of three sample sizes compared to both ScoreGLM and GLM (bottom two lines in each of the three right panels of fig. 2).

The empirical type I error rates were also calculated for *haplo4.2* and *haplo4.4* based on ScoreGLM, GLM, and BayesGLM, as in the preceding subsection of main effects, and similar results were observed.

Full Model

In scenario 5, a total of 81 marginal and interacting terms arising between the two haplotype blocks and between the haplotypes and the environmental factor were simultaneously considered (table 2). As can be seen from table 3, however, all the proportions of nonidentifiability of parameters jumped to 1 for both ScoreGLM and GLM in scenario 5. Consequently, the statistical estimations under this situation should be quite unstable and any comparison to them does not make sense. Since the current version of rGLM cannot fit $H \times E$ interactions, rGLM cannot be used to fit the full mode. Therefore, a single model based on BayesGLM was fitted to demonstrate its performance in a case where the number of predictors in a model is huge. As in the analyses of main and interacting effect models in the foregone subsections, the 'true' values prespecified for all predictors in the model were first compared to their corresponding estimated coefficients for each of three sample sizes, respectively (1st, 3rd, and 5th columns of fig. 3). From the graph it can be seen that, along with the increase in sample sizes, the estimated 68 and 95% intervals increased that covered the 'true' values for each of the 81 predictors. We also found that the rare haplotypes *(haplo4.3* and *haplo2.3)* and the interactions *(smoke*: *haplo2.3*, *smoke*:*haplo4.1*, *smoke*:*haplo4.3*, *haplo2.3*:*hap-lo4.1*, and *haplo2.3*:*haplo4.3)* had quite lower coverages than the other predictors in the model, no matter what sample sizes were considered. All these findings were consistent with those observed in the foregoing subsections.

The empirical powers were calculated for a total of 11 disease-associated predictors in the model under each of 3 fixed type I error rates ($\alpha = 0.001$, 0.01, and 0.05; 2nd, 4th, and 6th columns of fig. 3). From the graph it can be seen that although the power increased along with the increase in sample sizes, they started at quite low levels and maintained low growth rates. Under the sample size of 1,000, 8 predictors *(smoke*, *haplo2.3*, *haplo4.1*, *haplo4.3*, *smoke*:*haplo4.1*, *smoke*:*haplo4.3*, *haplo-2.3*:*haplo4.1*, and *haplo2.3*:*haplo4.3)* had an 80% chance or more of being identified under $\alpha = 0.05$, while 3 predictors *(haplo4.2*, *haplo4.4*, and *smoke*:*haplo2.3)* had a 60% chance or more of being identified under $\alpha = 0.05$.

The empirical type I error rates were also calculated for a total of 70 non-disease-associated predictors in the model. As the sample size went up to 500, almost all of the type I error rates shrank to zero.

Discussion

Complex human diseases are believed to be influenced by genetic and environmental factors, and their interactions. However, identifying interacting effects is challenging. In general, the identification and characterization of interactions are limited due to the lack of powerful statistical methods and/or large sample sizes. When numerous interactions are fitted explicitly in a model, the degrees of freedom for the corresponding test statistics would grow rapidly, and, as a result, sufficient power cannot be guaranteed to detect possibly significant effects in the model, especially in a relatively small sample size [50–54]. This issue is also confronted in haplotype-based association studies by classical methods, which usually have insufficient power and are not flexible

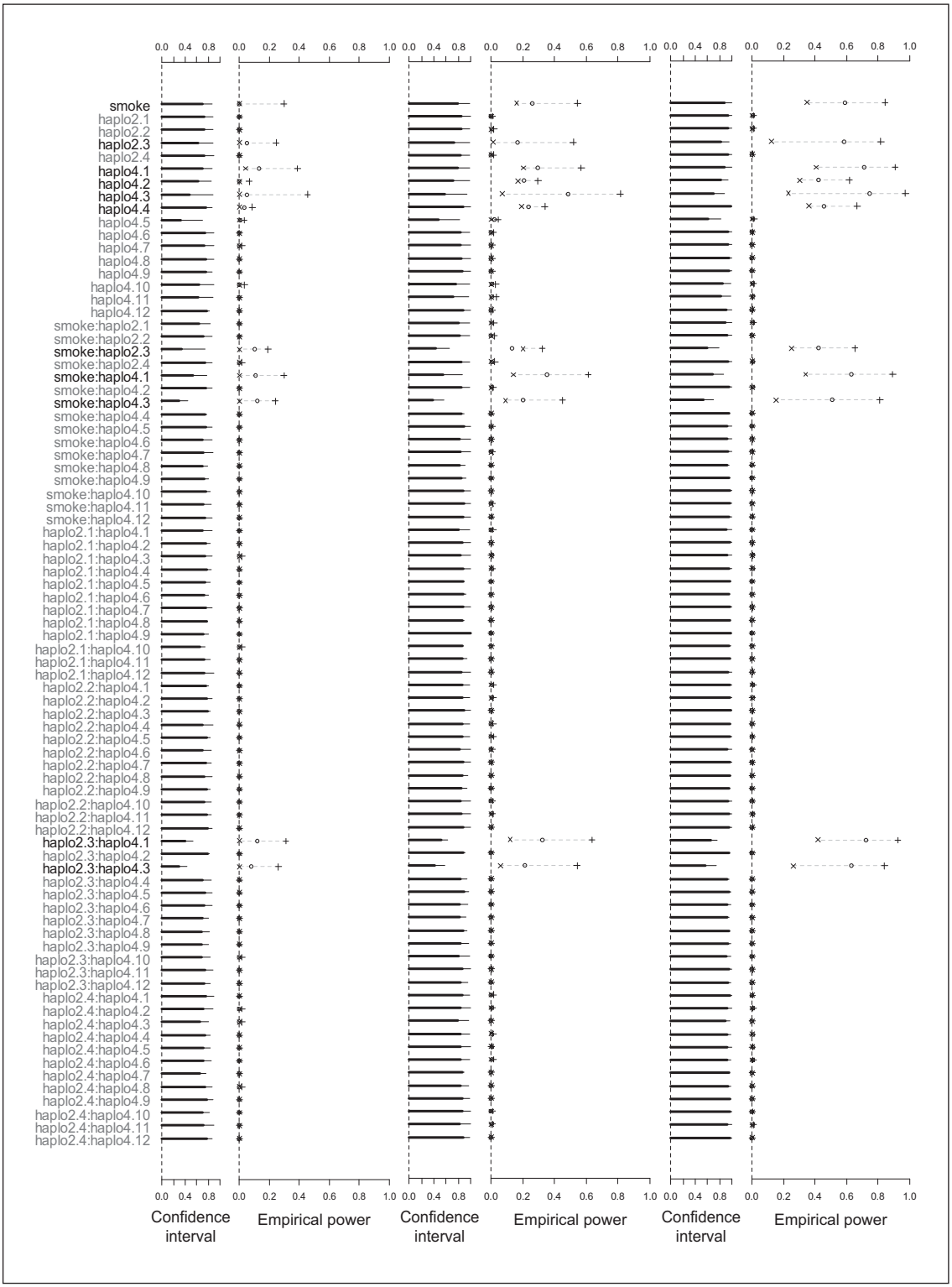

Fig. 3. Full model: estimated 68 and 95% coverages of the 'true' values (indicated by bold and thin horizontal lines in the 1st, 3rd, and 5th columns, respectively) and empirical powers or type I error rates [empirical powers or type I error rates for $\alpha = 0.001$ (×), $\alpha = 0.01$ (o), and $\alpha = 0.05$ (+)] for each of 81 predictors based on BayesGLM with sample sizes of 250 (first 2 columns), 500 (3rd and 4th columns), and 1,000 (last 2 columns). The black labels on the vertical axis stand for the disease-associated predictors, while the gray labels stand for the non-disease-associated predictors.

enough to handle a large number of interactions [14, 21, 26, 55].

The challenges might be further aggravated when rare haplotypes are present. Rare haplotypes can be frequent in genetic association studies and might be produced by common SNPs [15, 56]. As already noted, rare haplotypes, just like other rare genetic variants, could be important disease-predisposing variants and should not be ignored in exploring the genetic susceptibility to common diseases. Regarding statistical modeling, however, rare haplotypes can result in nonidentifiability of parameters, which means the coefficients of predictors cannot be identified or estimated uniquely because of huge, even infinite standard errors [46]. A commonly used approach to this issue in the literature is to pool all rare haplotypes into one single group [7, 13] or pool rare haplotypes with common ancestral haplotypes [33, 57, 58]. These approaches ignore rare haplotypes by lumping them together, and consequently any rare haplotype that might contribute to the risk of disease cannot be identified distinctly.

Statistical methods that can detect the haplotype-related interactions and handle the nonidentifiability of parameters are much needed in research. In the present study, we propose a Bayesian hierarchical generalized linear model with weakly informative priors to simultaneously analyze a large number of effects, including main effects of common and rare haplotypes, environmental effects, and their possible interactions. Our model fitting algorithm takes advantage of the classical generalized linear model procedure, leading to a computationally stable tool. An extensive simulation study was conducted to evaluate the statistical properties and performance of the proposed method, and the results were compared with the classical generalized linear model, the method of Lake et al. [21], and the method of Guo and Lin [15]. The main reason for considering these three methods as reference is that the classical generalized linear model is a flexible and basic approach to analyze case-control data, the method of Lake et al. [21] is the commonly used method for haplotype-based analysis in association studies, and the method of Guo and Lin [15] takes both rare haplotypes and the haplotype interactions between two haplotype blocks into account.

In our simulation study, the identifiability of parameters in model fit was first assessed because it is a common problem in conventional methods. The results show that for ScoreGLM and GLM the estimates of coefficients were substantially nonidentifiable in most of the simulation settings, while for BayesGLM the nonidentifiability of parameters was not observed. This demonstrates the appealing features of the proposed method in terms of robustness of parameter estimation and efficiency of statistical computation over the existing methods, especially if a large number of interactions and some rare haplotypes are included in the model.

With respect to the statistical properties of the proposed method, statistical power is our primary interest in the evaluation processes. The results indicate that the proposed method outperforms ScoreGLM and GLM in terms of statistical power for detecting associations, especially for rare haplotypes and interactions with moderate sample sizes. However, with the increase in the number of predictors fitted in the model, the proposed method had a relative loss of power, but was still acceptable (fig. 3). This is reasonable because, as we already know, the high dimensionality is traded with loss of power in model fit.

The reliability of the proposed method concerning parameter estimation was examined by comparing the 'true' values prespecified for the predictors in the models to their corresponding estimated coefficients. The proposed method can yield better coverage of confidence intervals, especially for the interactions and the rare haplotypes, than ScoreGLM and GLM (fig. 1, 2). However, most of the time, the proposed method unsurprisingly has similar results to rGLM (fig. 1, 2). However, the proposed method provides more features than rGLM in its current implementation. Moreover, although the proposed method is not yet available to practitioners, an R package is currently being developed and will be released soon.

Acknowledgments

We are grateful to two reviewers for helpful suggestions and comments on a draft of the paper. We would like to thank Mohit Limdi for his careful editing, which improved the manuscript, and David B. Allison for his support. This work was supported by the following grants: NIH 2R01 GM069430-06, NIH R01 GM077490, and NIH R01 GM074913.

References

1 Botstein D, Risch N: Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet 2003;33(suppl):228–237.
2 Risch N, Merikangas K: The future of genetic studies of complex human diseases. Science 1996;273:1616–1617.
3 Chapman NH, Wijsman EM: Genome screens using linkage disequilibrium tests: optimal marker characteristics and feasibility. Am J Hum Genet 1998;63:1872–1885.
4 Clark AG: The role of haplotypes in candidate gene studies. Genet Epidemiol 2004;27:321–333.
5 Davidson S: Research suggests importance of haplotypes over SNPs. Nat Biotechnol 2000;18:1134–1135.
6 Schaid DJ: Evaluating associations of haplotypes with traits. Genet Epidemiol 2004;27:348–364.
7 Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA: Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 2002;70:425–434.
8 Fitze G, Cramer J, Ziegler A, Schierz M, Schreiber M, Kuhlisch E, Roesner D, Schackert HK: Association between c135G/A genotype and RET proto-oncogene germline mutations and phenotype of Hirschsprung's disease. Lancet 2002;359:1169–1170.
9 Morris RW, Kaplan NL: On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genet Epidemiol 2002;23:221–233.
10 Kaplan NL, Morris RW: Issues concerning association studies for fine mapping a susceptibility gene for a complex disease. Genet Epidemiol 2001;20:432–457.
11 Akey J, Jin L, Xiong M: Haplotypes vs. single marker linkage disequilibrium tests: what do we gain? Eur J Hum Genet 2001;9:291–300.
12 Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG: Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 2002;53:79–91.
13 Zhao LP, Li SS, Khalid N: A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. Am J Hum Genet 2003;72:1231–1250.
14 Becker T, Schumacher J, Cichon S, Baur MP, Knapp M: Haplotype interaction analysis of unlinked regions. Genet Epidemiol 2005;29:313–322.
15 Guo W, Lin S: Generalized linear modeling with regularization for detecting common disease rare haplotype association. Genet Epidemiol 2009;33:308–316.
16 Cheverud JM, Routman EJ: Epistasis and its contribution to genetic variance components. Genetics 1995;139:1455–1461.
17 Wolf JB, Brodie ED III, Wade MJ: Epistasis and the Evolutionary Process. New York, Oxford University Press, 2000.
18 Moore JH: The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered 2003;56:73–82.
19 Carlborg Ö, Haley CS: Epistasis: too often neglected in complex trait studies? Nat Rev Genet 2004;5:618–625.
20 Moore JH: A global view of epistasis. Nat Genet 2005;37:13–14.
21 Lake SL, Lyon H, Tantisira K, Silverman EK, Weiss ST, Laird NM, Schaid DJ: Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. Hum Hered 2003;55:56–65.
22 Lin DY, Zeng D, Millikan R: Maximum likelihood estimation of haplotype effects and haplotype-environment interactions in association studies. Genet Epidemiol 2005;29:299–312.
23 Spinka C, Carroll RJ, Chatterjee N: Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. Genet Epidemiol 2005;29:108–127.
24 Kraft P, Cox DG, Paynter RA, Hunter D, De Vivo I: Accounting for haplotype uncertainty in matched association studies: a comparison of simple and flexible techniques. Genet Epidemiol 2005;28:261–272.
25 Lin DY, Zeng D: Likelihood-based inference on haplotype effects in genetic association studies. J Am Stat Assoc 2006;101:89–118.
26 Kwee LC, Epstein MP, Manatunga AK, Duncan R, Allen AS, Satten GA: Simple methods for assessing haplotype-environment interactions in case-only and case-control studies. Genet Epidemiol 2007;31:75–90.
27 Chen YH, Chatterjee N, Carroll RJ: Retrospective analysis of haplotype-based case control studies under a flexible model for gene environment association. Biostatistics 2008;9:81–99.
28 Excoffier L, Slatkin M: Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 1995;12:921–927.
29 Niu T, Qin Z, Xu X, Liu JS: Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 2002;70:157–159.
30 Stephens M, Smith NJ, Donnelly P: A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 2001;68:978–989.
31 Seltman H, Roeder K, Devlin B: Transmission/disequilibrium test meets measured haplotype analysis: family-based association analysis guided by evolution of haplotypes. Am J Hum Genet 2001;68:1250–1263.
32 Sha Q, Dong J, Jiang R, Zhang S: Tests of association between quantitative traits and haplotypes in a reduced-dimensional space. Ann Hum Genet 2005;69:715–732.
33 Tzeng JY: Evolutionary-based grouping of haplotypes in association analysis. Genet Epidemiol 2005;28:220–231.
34 Liu J, Papasian C, Deng HW: Incorporating single-locus tests into haplotype cladistic analysis in case-control studies. PLoS Genet 2007;3:e46.
35 Liu PY, Zhang YY, Lu Y, Long JR, Shen H, Zhao LJ, Xu FH, Xiao P, Xiong DH, Liu YJ, Recker RR, Deng HW: A survey of haplotype variants at several disease candidate genes: the importance of rare variants for complex diseases. J Med Genet 2005;42:221–227.
36 Zhu X, Fejerman L, Luke A, Adeyemo A, Cooper RS: Haplotypes produced from rare variants in the promoter and coding regions of angiotensinogen contribute to variation in angiotensinogen levels. Hum Mol Genet 2005;14:639–643.
37 Yende S, Angus DC, Ding J, Newman AB, Kellum JA, Li R, Ferrell RE, Zmuda J, Kritchevsky SB, Harris TB, Garcia M, Yaffe K, Wunderink RG, for the Health ABC Study: 4G/5G plasminogen activator inhibitor-1 polymorphisms and haplotypes are associated with pneumonia. Am J Respir Crit Care Med 2007;176:1129–1137.
38 Semsei AF, Erdélyi DJ, Ungvári I, Kámory E, Csókay B, Andrikovics H, Tordai A, Cságoly E, Falus A, Kovács GT, Szalai C: Association of some rare haplotypes and genotype combinations in the MDR1 gene with childhood acute lymphoblastic leukaemia. Leuk Res 2008;32:1214–1220.
39 Kitsios GD, Zintzaras E: An *NOS3* haplotype is protective against hypertension in a Caucasian population. Int J Hypertens 2010;2010:865031.
40 Molitor J, Marjoram P, Thomas D: Application of Bayesian spatial statistical methods to analysis of haplotypes effects and gene mapping. Genet Epidemiol 2003;25:95–105.
41 McCullagh P, Nelder JA: Generalized Linear Models, ed 2. London, Chapman & Hall, 1989.
42 Stram DO, Pearce CL, Bretsky P, Freedman M, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Thomas DC: Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. Hum Hered 2003;55:179–190.
43 Gelman A, Jakulin A, Pittau MG, Su YS: A weakly informative default prior distribution for logistic and other regression models. Ann Appl Stat 2008;2:1360–1383.
44 Yi N, Banerjee S: Hierarchical generalized linear models for multiple quantitative trait locus mapping. Genetics 2009;181:1101–1113.

45 Yi N, Kaklamani VG, Pasche B: Bayesian analysis of genetic interactions in case-control studies, with application to adiponectin genes and colorectal cancer risk. Ann Hum Genet 2011;75:90–104.
46 Gelman A, Carlin JB, Stern HS, Rubin DB: Bayesian Data Analysis, ed 2. London, Chapman & Hall, 2003.
47 Lei Z, Liu RY, Zhao J, Liu Z, Jiang X, You W, Chen X, Liu X, Zhang K, Pasche B, Zhang H: TGFBR1 haplotypes and risk of non-small-cell lung cancer. Cancer Res 2009;69:7046–7052.
48 Albert A, Anderson JA: On the existence of maximum likelihood estimates in logistic regression models. Biometrika 1984;71:1–10.
49 Lesaffre E, Albert A: Partial separation in logistic discrimination. J R Stat Soc Ser B 1989; 51:109–116.
50 Luan JA, Wong MY, Day NE, Wareham NJ: Sample size determination for studies of gene-environment interaction. Int J Epidemiol 2001;30:1035–1040.
51 Boks MP, Schipper M, Schubart CD, Sommer IE, Kahn RS, Ophoff RA: Investigating gene environment interaction in complex diseases: increasing power by selective sampling for environmental exposure. Int J Epidemiol 2007;36:1363–1369.
52 Mukherjee B, Ahn J, Gruber SB, Rennert G, Moreno V, Chatterjee N: Tests for gene-environment interaction from case-control data: a novel study of type I error, power and designs. Genet Epidemiol 2008;32:615–626.
53 Cordell HJ: Genome-wide association studies: detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 2009; 10:392–404.
54 Thomas D: Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. Annu Rev Public Health 2010;31:21–36.
55 Hein R, Beckmann L, Chang-Claude J: Comparison of different haplotype-based association methods for gene-environment (G×E) interactions in case-control studies when haplotype-phase is ambiguous. Hum Hered 2009;68:252–267.
56 Souverein OW, Zwinderman AH, Jukema JW, Tanck MW: Estimating effects of rare haplotypes on failure time using a penalized Cox proportional hazards regression model. BMC Genet 2008;9:9.
57 Seltman H, Roeder K, Devlin B: Evolutionary-based association analysis using haplotype data. Genet Epidemiol 2003;25: 48–58.
58 Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P, Morris AP: Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. Am J Hum Genet 2004;75:35–43.