Regression-Based Association Analysis with Clustered Haplotypes through Use of Genotypes

Jung-Ying Tzeng,[1] Chih-Hao Wang,[2,3] Jau-Tsuen Kao,[4] and Chuhsing Kate Hsiao[5]

[1]Department of Statistics and Bioinformatics Research Center, North Carolina State University, Raleigh; and [2]Department of Cardiology, Cardinal Tien Hospital, [3]College of Medicine, Department of Medicine, Fu Jen Catholic University, and [4]Department of Clinical Laboratory Sciences and Medical Biotechnology, College of Medicine, and [5]Division of Biostatistics, Institute of Epidemiology, National Taiwan University, Taipei

Haplotype-based association analysis has been recognized as a tool with high resolution and potentially great power for identifying modest etiological effects of genes. However, in practice, its efficacy has not been as successfully reproduced as expected in theory. One primary cause is that such analysis tends to require a large number of parameters to capture the abundant haplotype varieties, and many of those are expended on rare haplotypes for which studies would have insufficient power to detect association even if it existed. To concentrate statistical power on more-relevant inferences, in this study, we developed a regression-based approach using clustered haplotypes to assess haplotype-phenotype association. Specifically, we generalized the probabilistic clustering methods of Tzeng to the generalized linear model (GLM) framework established by Schaid et al. The proposed method uses unphased genotypes and incorporates both phase uncertainty and clustering uncertainty. Its GLM framework allows adjustment of covariates and can model qualitative and quantitative traits. It can also evaluate the overall haplotype association or the individual haplotype effects. We applied the proposed approach to study the association between hypertriglyceridemia and the apolipoprotein A5 gene. Through simulation studies, we assessed the performance of the proposed approach and demonstrate its validity and power in testing for haplotype-trait association.

In the search for genes underlying human complex diseases, one crucial step is to detect the association between the genetic variants and the disease phenotypes. Since a high density of SNPs is being identified and used in genetic studies, jointly analyzing all variants within a gene or chromosomal region for association can be more informative and effective (Stephens et al. 2001). The haplotype, the ordered allele sequences on a chromosome, provides a natural framework for performing joint analysis of multiple markers and is predominantly considered the unit of analysis in association studies. Haplotype analyses are believed to provide high resolution and potentially great power for identifying modest etiological effects of genes (International HapMap Consortium 2003). Following this viewpoint, many statistical methods have been proposed to evaluate haplotype-disease association for case-control samples, including likelihood ratio tests for testing equality of haplotype frequencies between cases and controls (e.g., Sham 1998), tests and inferences for specific haplotype effects under a variety of regression models (e.g., Schaid et al. 2002; Zaykin et al. 2002; Epstein and Satten 2003; Lake et al. 2003; Stram et al. 2003; Zhao et al. 2003; Lin 2004; Zeng and Lin 2005), haplotype-similarity approaches that detect association via excessive haplotype sharing in cases (e.g., Van der Meulen and te Meerman 1997; McPeek and Strahs 1999; Bourgain et al. 2000, 2001, 2002; Tzeng et al. 2003*a*, 2003*b*; Yu et al. 2004), and clustering methods that group homogeneous haplotypes and perform analysis on the unit of haplotype groups (e.g., Seltman et al. 2001, 2003; Molitor et al. 2003*a*, 2003*b*; Durrant et al. 2004; Tzeng 2005).

Whereas the progress in both data availability and data analyses increases the feasibility of haplotype-based association studies, practical implementation indicates that the study findings of such types are not consistently reproducible (Lohmueller et al. 2003; Neale and Sham 2004). Lohmueller et al. (2003) concluded that the inconsistency could be explained largely by a high rate of false-negative results or, equivalently, lack of power. Recently, Chapman and colleagues (Chapman et al. 2003; Clayton et al. 2004) further revealed that analyses-based locus models that regress phenotypes on multiple SNP loci can sometimes be more powerful than haplotype analyses, such as when tag SNPs are used. The main reason is that the locus model uses fewer parameters than does a haplotype model; by modeling only the main effect and low-order interactions of SNPs, the locus model does not spend degrees of freedom on rare haplotypes for which studies would have insufficient power to detect association even if it were present (Clayton et al. 2004).

In contrast to a locus model, haplotype analysis re-

Received June 28, 2005; accepted for publication November 16, 2005; electronically published December 19, 2005.

Address for correspondence and reprints: Dr. Jung-Ying Tzeng, Department of Statistics and Bioinformatics Research Center, North Carolina State University, Campus Box 7566, Raleigh, NC 27695. E-mail: jytzeng@stat.ncsu.edu

Am. J. Hum. Genet. 2006;78:231–242. 0002-9297/2006/7802-0006$15.00

quires a larger number of parameters to capture the abundant haplotype varieties, and the test power is limited by the many degrees of freedom that they use. The power is worsened by the need to adjust for multiple testing when many genes are evaluated. Further difficulties emerge from the fact that complex diseases are derived from intricate genetic and environmental factors (see, e.g., Peltonen and McKusick 2001). Understanding the genetic etiology of complex diseases requires a joint consideration of all potential attributes and sometimes even other auxiliary covariates. The vast quantities of covariates from environmental effects and gene-gene and gene-environment interactions further exacerbate the degrees-of-freedom problem.

Model-based association methods, which incorporate covariate information in association analysis, play an increasingly important role in modern association studies. They facilitate the study of complex gene-disease association. Besides the ability to accommodate polygenic effects, environmental covariates, and interactions among them, model-based analyses can evaluate haplotype effects at either the global level (i.e., evaluating overall haplotype association) or the individual level (i.e., evaluating haplotype-specific association). They also allow modeling of diseases through a variety of clinical phenotypes, from dichotomous to ordinal to quantitative traits. These flexibilities and advantages again reflect the need for efficient usage of haplotype information in a model-based framework for studying association.

Haplotype grouping offers one promising avenue for controlling the issue of degrees of freedom that is encountered in haplotypes-based multiple-marker analysis. It enhances the efficiency of haplotype analysis by using a small number of degrees of freedom to study haplotypes and concentrates statistical power on more-relevant inference. In an earlier study (Tzeng 2005), we introduced an algorithm to cluster related haplotypes to improve the power of association tests. This algorithm adapts the same evolutionary concepts of cladistic analyses and groups rare haplotypes with their closest major haplotypes according to the evolutionary relationships summarized in a haplotype tree. Since many haplotype trees are often virtually likely given the observed data, one key feature of the proposed algorithm is the incorporation of the tree uncertainty in association testing. The algorithm is motivated by and relies on the common disease/common variants assumption (Collins et al. 1997), which conjectures that common modest-risk variants may contribute more to the development of common complex disease than do rare high-risk variants. The algorithm is also built on the recent discovery of the human genome structure that the majority of haplotype diversities are concentrated on a few major categories because of the correlations among proximate SNPs (e.g., Daly et al. 2001; Johnson et al. 2001). Therefore, instead of spending degrees of freedom on rare haplotypes that would result in unstable statistical inference and insufficient testing power, the algorithm reduces the observed haplotype space, in a probabilistic manner, to a core haplotype set that contains fewer polymorphisms but possesses the essential information for studying haplotype-disease association. Such core haplotype diversity presumably mimics the diversity before the occurrence of other events that are not directly related to the evolution of disease mutation—for example, recent marker mutation, gene conversion, genotyping error, and even missing data.

The grouping analysis of Tzeng (2005) is limited to assessing global association between haplotypes and traits. It cannot evaluate the effect of individual haplotypes or accommodate for covariates. Its implementation requires phased haplotypes and empirical evaluation of the significance level. In the present study, we generalized the clustering approach of Tzeng (2005) to a generalized linear model framework and allowed for unphased genotypes. We constructed tests that are based on clustered haplotypes, for assessing association at both global and haplotype-specific levels. The test incorporates two major sources of uncertainties in haplotype analysis—clustering uncertainty and phase uncertainty. Among the many promising regression-based approaches that evaluate individual effects of haplotypes through use of genotypes, we established our work on the score tests developed by Schaid et al. (2002). Their method has been shown to be robust to departure from the Hardy-Weinberg equilibrium and to possess comparable power with retrospective approaches for case-control data that are sampled retrospectively (Satten and Epstein 2004). Through simulation studies, we assessed the performance of the proposed approach and demonstrated its validity and power in testing for haplotype-trait association. We also illustrated the proposed approach through an application to a hypertriglyceridemia study, in which we tested the apolipoprotein A5 gene (*APOA5*), a confirmed risk factor of hypertriglyceridemia.

Methods

We begin this section by reviewing the clustering methods of Tzeng (2005). We then integrate the clustering algorithm into a regression framework. Finally, we construct the score test for association that incorporates phase ambiguity and clustering uncertainty on the basis of the work of Schaid et al. (2002) and Tzeng (2005).

The Haplotype-Clustering Method of Tzeng

The fundamental purpose of the clustering algorithm is to group rare haplotypes with their corresponding ancestral haplotypes. Given an evolutionary tree of haplotypes, the algorithm sequentially combines "rare" haplotypes into their one-

step neighboring haplotypes, from the tips of the tree toward the major nodes. Each of the resulting clusters is represented by the most common haplotype, and haplotypes within a cluster are assumed to have the same effect on the disease trait.

Determining "rare" haplotypes requires a trade-off between information and dimensionality, and the algorithm uses an information criterion to find the optimal balance between the two. The information criterion is defined as "the cumulative Shannon information content" (Shannon 1948), with penalty function determined by the number of dimensions and the sample size involved. Denote H_F as the full set of observed haplotypes and H_C as the set of clustered haplotypes. The algorithm obtains H_C by preserving high-frequency haplotypes—that is, to set H_C as the ℓ most frequent haplotypes, where ℓ maximizes the information criterion.

In reality, the evolutionary tree is often unknown and needs to be inferred. Instead of inferring the most-likely tree relationship and performing grouping accordingly, the algorithm assigns each relationship branch a probability. It then clusters haplotypes by considering all relationships according to the probability weights. The branch probability is determined by two factors that were commonly considered in reconstructing a haplotype tree (Crandall and Templeton 1993; Slatkin and Rannala 1997): (1) the relatedness of haplotypes and (2) the age of haplotypes. The algorithm uses haplotype frequencies to indicate the haplotype age. To measure the relatedness of haplotypes, a certain metric of haplotype similarity is used, such as counting the number of matching loci between two haplotypes. When the evolutionary relationships are known, the branch probability is reduced to an indicator function of whether two haplotypes u and v are one-step related. For further detail, see Tzeng (2005).

The general algorithm can be described as follows: first, partition the list H_F into (1) $H^{(0)} = H_C$, the core category, (2) $H^{(1)}$, the one-step neighbors of $H^{(0)}$ that consist of haplotypes different from the core haplotypes by one step of mutation, and (3) $H^{(2)}$, the two-step neighbors of $H^{(0)}$ that consist of haplotypes different from the core haplotypes by two steps of mutation, and continue until the entire space of H_F is exhausted. Let Π_F denote the haplotype frequencies of H_F; correspondingly, Π_F is also decomposed into $\Pi^{(0)},\Pi^{(1)},\ldots,\Pi^{(j)},\ldots,\Pi^{(J)}$. Starting from $j = J$ to $j = 1$, group each element of $H^{(j)}$ to its one-step ancestor in $H^{(j-1)}$ and combine the frequencies. The grouping rule is specified according to the branch probabilities that are stored in the allocation matrix $\boldsymbol{B}^{(j)}$; each row of $\boldsymbol{B}^{(j)}$ describes to whom and how a certain haplotype of $H^{(j)}$ is allocated among $H^{(j-1)}$. As illustrated by Tzeng (2005), this one-step grouping process is equivalent to the matrix operation $\Pi^{(j)\prime}\boldsymbol{B}^{(j)}$, and the overall process can be described as

$$\Pi_C'(=\Pi^{(0)*\prime}) = \Pi^{(0)\prime} + \Pi^{(1)\prime}\boldsymbol{B}^{(1)} + \Pi^{(2)\prime}\boldsymbol{B}^{(2)}\boldsymbol{B}^{(1)} + \cdots + \Pi^{(J)\prime}\boldsymbol{B}^{(J)}\boldsymbol{B}^{(J-1)}\cdots\boldsymbol{B}^{(2)}\boldsymbol{B}^{(1)} .$$

Or, equivalently,

$$\Pi_C' = \Pi_F'\boldsymbol{B} , \qquad (1)$$

where

$$\Pi_F = \begin{bmatrix} \Pi^{(0)} \\ \hline \Pi^{(1)} \\ \hline \Pi^{(2)} \\ \hline \vdots \end{bmatrix} \quad \boldsymbol{B} = \begin{bmatrix} \mathbf{I} \\ \hline \boldsymbol{B}^{(1)} \\ \hline \boldsymbol{B}^{(2)}\boldsymbol{B}^{(1)} \\ \hline \vdots \end{bmatrix} .$$

Suppose there are $(L + 1)$ distinct haplotypes in the population and they are clustered into $(L^* + 1)$ groups. The dimension of $\boldsymbol{B}$ is $(L + 1) \times (L^* + 1)$.

Regression Model with Clustered Haplotypes

Given that the clustering procedure can be implemented via the matrix multiplication in equation (1), it is straightforward to integrate this dimension reduction procedure into a regression framework. Under the regression model, probabilistic clustering of haplotypes can be done by replacing the vector of the haplotype frequencies Π in equation (1) with the data matrix of haplotypes. That is, denote $\boldsymbol{X}_F$ as the haplotype matrix of the full dimension with use of a certain scoring rule; its (h,i) entry, for example, can be the number of copies of haplotype h that individual i possesses. The matrix $\boldsymbol{X}_F$ has dimension $(L + 1) \times n$, where n is the sample size. Then the data matrix of clustered haplotypes, $\boldsymbol{X}_C$, can be obtained by

$$\boldsymbol{X}_C' = \boldsymbol{X}_F'\boldsymbol{B}(\Pi) . \qquad (2)$$

Here, we rewrite the allocation matrix $\boldsymbol{B}$ as $\boldsymbol{B}(\Pi)$ to emphasize the fact that the allocation matrix $\boldsymbol{B}$ is a function of the haplotype frequency Π.

Let Y denote an $n \times 1$ vector of the disease trait values, and let $\boldsymbol{Z}$ denote a $P \times n$ matrix of the P environmental covariates. With the original haplotype data of full dimension, the effects of the genetic and environmental covariates can be modeled by the generalized linear model (GLM):

$$g(\mathrm{EY}) \equiv \eta = \boldsymbol{X}_F'\boldsymbol{\beta}_F + \boldsymbol{Z}'\boldsymbol{\gamma} ,$$

where $\boldsymbol{\beta}_F' = (\beta_{F(0)},\beta_{F(1)},\cdots,\beta_{F(L)})$ is an $(L + 1) \times 1$ vector. The association of haplotypes with the disease traits can be detected by testing $H_0{:}\beta_{F(0)} = \beta_{F(1)} = \cdots = \beta_{F(L)}$.

To reduce the degrees of freedom, we performed an analysis on groups of homogeneous haplotypes, using the following model:

$$g(\mathrm{EY}) \equiv \eta = \boldsymbol{X}_C'\boldsymbol{\beta}_C + \boldsymbol{Z}'\boldsymbol{\gamma} ,$$

where $\boldsymbol{X}_C'$ is obtained by the clustering algorithm of equation (2) and $\boldsymbol{\beta}_C' = (\beta_{C(0)}, \beta_{C(1)}, \cdots, \beta_{C(L^*)})$ with $L^* \leqslant L$. The association test is now performed through the $(L^* + 1)$ parameters of the clustered haplotypes,

$$H_0{:}\beta_{C(0)} = \beta_{C(1)} = \cdots = \beta_{C(L^*)} . \qquad (3)$$

Score Test Incorporating Clustering Uncertainty and Phase Uncertainty

Here, we derive the score test for association in the clustered haplotype space. We first calculate the score function, which is the partial derivative of the log likelihood function, and then use it to construct the score test. To facilitate derivation, we reparameterize $\boldsymbol{\beta}_C$ via a linear transformation

$$\boldsymbol{\beta}_C \equiv \begin{bmatrix} \mu \\ \mu + \alpha_1 \\ \vdots \\ \mu + \alpha_{L^*} \end{bmatrix} = \boldsymbol{A}\begin{bmatrix} \mu \\ \boldsymbol{\alpha} \end{bmatrix}, \text{ with } \boldsymbol{A} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & & & \\ \vdots & & \mathbf{I}_{L^* \times L^*} & \\ 1 & & & \end{bmatrix}.$$

Consequently, the global null hypothesis (3) is equivalent to $H_0{:}\alpha_1 = \alpha_2 = \cdots = \alpha_{L^*} = 0$, and the effect of haplotype h can be examined by $H_0{:}\alpha_h = 0$.

Consider observed data $(Y,\boldsymbol{G},\boldsymbol{Z})$ in which $\boldsymbol{G}$ is the data matrix of unphased genotypes. For each individual i, we treat the observed genotype $\boldsymbol{g}_i$ as an incomplete version of haplotype count $\boldsymbol{x}_{F,i}$, which is the ith column of the design matrix $\boldsymbol{X}_F$. Without losing generality, here we assume that the vector $\boldsymbol{x}_{F,i}$ is normed so that its entries sum to 1. Under the assumption of Hardy-Weinberg equilibrium, $\boldsymbol{x}_{F,i} \sim \frac{1}{2} \times \text{multinomial}(2,\boldsymbol{\Pi}_F)$. The GLM density of trait y_i, given covariates $\boldsymbol{x}_{F,i}$ and $\boldsymbol{z}_i$, is

$$f(y_i|\boldsymbol{x}_{F,i},\boldsymbol{z}_i;\boldsymbol{\alpha},\mu,\phi,\boldsymbol{\gamma},\boldsymbol{\Pi}) = \exp\left[\frac{y_i\eta_i - b(\eta_i)}{a(\phi)} + c(y_i,\phi)\right],$$

where

$$\eta_i = \boldsymbol{x}'_{C,i}\boldsymbol{\beta}_C + \boldsymbol{z}'_i\boldsymbol{\gamma} = \boldsymbol{x}'_{F,i}\,\boldsymbol{B}(\boldsymbol{\Pi})\boldsymbol{A}\begin{bmatrix} \mu \\ \boldsymbol{\alpha} \end{bmatrix} + \boldsymbol{z}'_i\boldsymbol{\gamma},$$

and ϕ is the dispersion parameter (see table 1 of Schaid et al. [2002]). Let $\boldsymbol{\zeta}$ denote the vector of the nuisance parameters $(\mu,\boldsymbol{\gamma},\phi,\boldsymbol{\Pi})$. The likelihood function for $(\boldsymbol{\alpha},\boldsymbol{\zeta})$ on the basis of the data $(Y,\boldsymbol{G},\boldsymbol{Z})$ is

$$\begin{aligned} L(\boldsymbol{\alpha},\boldsymbol{\zeta};Y,\boldsymbol{G},\boldsymbol{Z}) &= \prod_{i=1}^{n}\left\{\sum_{\boldsymbol{x}_{F,i}} f(y_i,\boldsymbol{x}_{F,i},\boldsymbol{g}_i \mid \boldsymbol{z}_i;\ \boldsymbol{\alpha},\boldsymbol{\zeta})\right\} \\ &= \prod_{i=1}^{n}\left\{\sum_{\boldsymbol{x}_{F,i}} f(y_i \mid \boldsymbol{x}_{F,i},\boldsymbol{z}_i;\boldsymbol{\alpha},\boldsymbol{\zeta}) \times \mathrm{P}(\boldsymbol{g}_i \mid \boldsymbol{x}_{F,i}) \times \mathrm{P}(\boldsymbol{x}_{F,i};\boldsymbol{\Pi})\right\}. \end{aligned} \tag{4}$$

Because $\mathrm{P}(\boldsymbol{g}_i|\boldsymbol{x}_{F,i})$ is an indicator function of whether the haplotype count $\boldsymbol{x}_{F,i}$ is compatible with the observed genotype $\boldsymbol{g}_i$, likelihood (4) can be further simplified as

$$L(\boldsymbol{\alpha},\boldsymbol{\zeta};Y,\boldsymbol{G},\boldsymbol{Z}) = \prod_{i=1}^{n}\left\{\sum_{\boldsymbol{x}_{F,i}\in\boldsymbol{g}_i} f(y_i \mid \boldsymbol{x}_{F,i},\boldsymbol{z}_i;\boldsymbol{\alpha},\boldsymbol{\zeta}) \times \mathrm{P}(\boldsymbol{x}_{F,i};\boldsymbol{\Pi})\right\}. \tag{5}$$

The score function for $\boldsymbol{\alpha}$ is the partial derivative of likelihood (5), with respect to $\boldsymbol{\alpha}$. The resulting score statistic, denoted by S_α, is the score function evaluated at the restricted maximum-likelihood estimates under the null hypothesis. S_α is the statistic we use to test haplotype effect; in appendix A, we show the following result:

$$S_\alpha = \sum_{i=1}^{n}\frac{y_i - \bar{y}}{a(\phi)}\,\boldsymbol{B}(\boldsymbol{\Pi})'_{-0}\,\mathrm{E}(X_i \mid \boldsymbol{g}_i)\Big|_{\substack{\alpha=\tilde{\alpha}=0 \\ \zeta=\tilde{\zeta}}},$$

where $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\zeta}}$ are the restricted maximum-likelihood estimates under the null hypothesis, $\boldsymbol{B}(\boldsymbol{\Pi})'_{-0}$ is the matrix $\boldsymbol{B}(\boldsymbol{\Pi})$ with the first column (i.e., the baseline haplotype) removed, and $\mathrm{E}(X_i \mid \boldsymbol{g}_i)$ is the same as that defined by Schaid et al. (2002), the expected haplotype counts given the observed genotypes. We see that the proposed score statistic that accounts for phase and clustering ambiguities is the original score test of Schaid et al. (2002) multiplied by the function of allocation matrix $\boldsymbol{B}(\boldsymbol{\Pi})$.

To construct the test for haplotype-trait association that adjusts for environmental covariates, we need the variance of S_α. We consider the generalized score test, which would ensure the asymptotic null χ^2 distribution even under model misspecification (Boos 1992). Define $\boldsymbol{\Theta} = (\boldsymbol{\alpha},\boldsymbol{\zeta})$ and let $\boldsymbol{V}_\alpha$ denote the variance of S_α. As indicated by Boos (1992),

$$\begin{aligned} \boldsymbol{V}_\alpha = (&D_{\alpha\alpha} - I_{\alpha\zeta}I_{\zeta\zeta}^{-1}D'_{\alpha\zeta} - D_{\alpha\zeta}I_{\zeta\zeta}^{-1}I'_{\alpha\zeta} \\ &+ I_{\alpha\zeta}I_{\zeta\zeta}^{-1}D_{\zeta\zeta}I_{\zeta\zeta}^{-1}I'_{\alpha\zeta})\Big|_{\substack{\alpha=\tilde{\alpha} \\ \zeta=\tilde{\zeta}}}, \end{aligned} \tag{6}$$

where

$$\boldsymbol{D} = \begin{pmatrix} D_{\alpha\alpha} & D_{\alpha\zeta} \\ D'_{\alpha\zeta} & D_{\zeta\zeta} \end{pmatrix} = \sum_{i=1}^{n} s_i(y_i,\boldsymbol{g}_i,\boldsymbol{z}_i,\boldsymbol{\Theta})s'_i(y_i,\boldsymbol{g}_i,\boldsymbol{z}_i,\boldsymbol{\Theta})$$

$$\boldsymbol{I} = \begin{pmatrix} I_{\alpha\alpha} & I_{\alpha\zeta} \\ I'_{\alpha\zeta} & I_{\zeta\zeta} \end{pmatrix} = -\sum_{i=1}^{n}\mathrm{E}\left[\frac{\partial s_i(y_i,\boldsymbol{g}_i,\boldsymbol{z}_i,\boldsymbol{\Theta})}{\partial\boldsymbol{\Theta}'}\right].$$

The individual score function

$$s_i(y_i,\boldsymbol{g}_i,\boldsymbol{z}_i,\boldsymbol{\Theta}) \equiv \frac{\partial}{\partial\boldsymbol{\Theta}}\log L(\boldsymbol{\Theta};y_i,\boldsymbol{g}_i,\boldsymbol{z}_i)$$

is the first derivative of the log likelihood with respect to $\boldsymbol{\Theta}$, and matrix $\boldsymbol{I}$ is the expected Fisher information. We use a hybrid method to estimate $\boldsymbol{I}$—that is, to replace the nonzero elements of $\boldsymbol{I}$ by the observed Fisher information (Kent 1982; Boos 1992).

In appendix B, we list the nonzero entries of matrices $\boldsymbol{D}$ and $\boldsymbol{I}$ that are in $\boldsymbol{V}_\alpha$. As in the work of Schaid et al. (2002), here we also see that the score function for $\boldsymbol{\alpha}$ and the score function for $\boldsymbol{\Pi}$ are independent under the null hypothesis; that is, the covariance between the two score functions is zero. Hence, although the estimate of the haplotype frequency $\boldsymbol{\Pi}$ is required in calculating the score statistic S_α, the variance of the score

statistic is not penalized by the use of estimated haplotype frequencies.

Finally, we assemble S_α and $\mathbf{V}_\alpha$ into the score test for assessing the global and haplotype-specific association. The global score-test statistic for testing $\boldsymbol{\alpha} = 0$ is $T_g \equiv S_\alpha' \mathbf{V}_\alpha^{-1} S_\alpha$. Under the null hypothesis of no haplotype association, T_g follows a χ^2 distribution with L^* df. The haplotype-specific test for haplotype h can be conducted via the test statistic $T_h = S^2_{\alpha(h)}/\mathrm{Diag}(\mathbf{V}_\alpha)_{(h)}$, where the subscript (h) indicates the hth element of a vector. The statistic T_h follows χ^2_1 under the null hypothesis $H_0: \alpha_h = 0$.

Benefiting from the *R* functions developed by Schaid et al. (2002), we implemented the proposed score test in *R* that is based on their codes. The *R* codes are available at the authors' Web site.

Simulation Study

Simulation Scheme

Simulation studies are conducted to evaluate the power and type I error of association tests on the basis of the clustered haplotypes. The haplotype data are generated in a way similar to that of Roeder et al. (2005) and Tzeng (2005). We simulate 100 SNP haplotypes, using a modified Hudson's (Hudson 2002) MS program (Wall and Pritchard 2003). This program generates data under a coalescent model in which the recombination rate varies across the SNP sequence. The scaled recombination rate, $\rho = 4N_e\delta/\mathrm{bp}$, is set to range from 4×10^{-3} to 8×10^{-3} for the recombination cold spots, with 1×10^4 as the effective population size N_e. In the hot spots, ρ is set to be 45 times greater than the rate in the cold spots. The recombination parameters are chosen to mimic the linkage disequilibrium (LD) patterns of the *SELP* gene shown in the SeattleSNP database. The scaled mutation rate for the entire region, $4N_e\mu/\mathrm{bp}$, is set to be 5.6×10^{-4}. The rate is chosen to produce the number of common SNPs (per kb) in the European American sample from the SeattleSNP database. Examining the matrix plots of pairwise correlations (R^2) between SNPs, we see that the original gene data and the simulated haplotype sequences have similar LD patterns and consist of three major blocks. Such a blocky setting allows us to evaluate the grouping method when applying to regions with reduced haplotype diversity.

We discard rare SNPs so that the minor-allele frequencies are >0.05. We then determine the liability locus according to the frequency of the liability allele, q, and the location of the locus. We consider three possible frequencies: $q = 0.1$, 0.3, or 0.5. The positions—that is, whether a liability locus exists in a haplotype block or recombination hot spot—are determined by the entropy-based blocking algorithm of Rinaldo et al. (2005). Once a liability locus is chosen, a haplotype is defined as a segment of six adjacent SNPs in which the third SNP is the liability locus. We sample 400 haplotypes with replacement from the 100 6-SNP haplotypes, randomly pair them to form 200 individuals, and then determine their phenotypes according to the genotypes at the liability locus. This process is repeated 1,000 times to obtain 1,000 data sets.

Now we describe how the phenotypes are determined. Assuming an additive effect of the liability allele, we generate both continuous and binary trait values. We use random sampling for continuous traits and balanced case-control sampling for binary traits, as done by Lake et al. (2003). To mimic a complex disease of a single liability allele with moderate effect, the phenotypes are determined using methods described below.

Continuous traits.—Here, we consider two simple models of quantitative traits. The first model (model I) decomposes the trait value into genetic effect g and environmental effect e: $Y = g + e$. The second model (model II) additionally incorporates a covariate Z: $Y = g + \gamma \times Z + e$. In both models, g has a discrete distribution, where g equals u_2, u_1, and u_0 with probabilities q^2, $2q(1-q)$, and $(1-q)^2$, respectively; e follows a normal distribution with mean ϵ and variance σ_e^2. In the second model, Z is generated from a standard normal distribution. For simplicity, we set $u_j = j - 1$, $\epsilon = 0$, and $\gamma = 1$. The trait values are generated using the normal penetrance function $f(Y \mid j) = \mathrm{N}(u_j, \sigma_e^2)$ for the first model and $f(Y \mid j) = \mathrm{N}(u_j + \gamma \times Z, \sigma_e^2)$ for the second model. We determine σ_e^2 through the heritability of the liability locus h^2, which is defined as

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \ .$$

With simple algebra, we set the environmental variance to

$$\sigma_e^2 = \sigma_g^2 \times \frac{1 - h^2}{h^2} = 2q(1-q) \times \frac{1 - h^2}{h^2} \ .$$

We set $h^2 = 0.1$ so that the average power of the full-dimensional regression analysis is ~0.5 at nominal level 0.01 for a sample with 200 individuals.

Binary Traits.—We generate phenotypes on the basis of a penetrance function f_j, with which an individual was assigned "affected" status with probability f_j if he/she possesses j copies of liability alleles. If $Y = 1$ for affected and $Y = 0$ for unaffected, then $f_j \equiv \mathrm{P}(Y = 1 \mid j)$. Define r to be the relative ratio f_1/f_0 and K the prevalence. Given r, K, and the liability-allele frequency q, we have $f_0 = K/(1 - 2q + 2qr)$, $f_1 = rf_0$, and $f_2 = 2rf_0 - f_0$ under an additive model. Here, we set $r = 2$ and $K = 0.01$. To perform a case-control sampling, two samples are drawn with replacements from the 100 6-SNP haplotypes and are paired to form an individual. Then the individual

with j copies of liability alleles is assigned to be a case if a randomly selected number is less than f_j and otherwise is assigned to be a control. This process is repeated until we obtain 100 cases and 100 controls.

Results

Haplotypes and trait values are generated under nine scenarios, according to the frequency of the disease allele (q =0.1, 0.3, or 0.5) and the diversity (high, moderate, or low) of the haplotype where the disease locus exists. "High diversity" indicates that a disease locus is located in the region of recombination hot spots and that the number of distinct haplotypes is 10–16; "moderate diversity" indicates that a disease locus is located in a haplotype block and that the number of distinct haplotypes is 9–12; "low diversity" indicates that a disease locus is located in a haplotype block and that the number of distinct haplotypes is 5–8. The number of haplotypes for the simulated data set has a range of 5–16; the proposed method tends to retain 4–12 haplotype groups in the analysis.

To study the performance of the proposed method in detecting association, we calculate type I error and power of the clustered score test on the basis of 1,000 simulations. The P values are determined asymptotically by the χ^2 distribution. We also compared the test results with the full dimensional analysis, in which the P values are obtained via permutation test. In evaluating the test performance, we consider the following fits in each simulation; for model I, we fit a regression model without the covariate Z (fit I); for model II, we fit a regression model both with (fit IIa) and without (fit IIb) the covariate; for the binary traits, we fit a model without covariate (fit III).

Table 1 displays the type I error of the global test with use of the proposed method. The values in table 1 are all near the nominal level α for either $\alpha = 0.05$ or 0.01, indicating that the χ^2 distribution adequately approximates the null distribution of the clustered score statistics. The power of the global test is shown in figures 1 ($\alpha = 0.05$) and 2 ($\alpha = 0.01$). In each figure, the four plots correspond to four fits. Fit I and fit IIa have similar patterns, as one would expect. Once the covariate effect is removed from fit IIa, we see similar testing power. Fit IIb does not take into account covariates that should have been modeled, and we see a drop in the power. For each scenario, the power of the clustered score statistics tends to be above the power of the full-dimensional analysis. This suggests that the clustering approach with asymptotic χ^2 P values retains the same or higher power than the full-dimensional test.

To explore the influence of the sample size, we double the sample size and examine the power for $\alpha = 0.05$. For a continuous trait, because of our choice of heritability, the power is almost 1 for both clustering and full-dimensional methods for all the fits. Nevertheless, we can still observe the effect of the sample size through a binary trait. As shown in figure 1, panel fit III (displayed in red), although the power of both methods moves up when sample size is doubled, the relative relationship between the power of the two methods remains roughly similar.

Next, we examine the significant causal haplotypes identified by the clustered approach and by the full-dimensional approach in the global test. We defined the difference count, which equals k if there are k haplotypes identified by the full-dimensional method but not by the clustered method and which equals $-k$ for the reverse situation. The difference counts across the nine scenarios are recorded and the histograms of fit I and fit III are presented in figure 3. Among the simulated data sets, ~75% of the data contain three causal haplotypes, and

Table 1

Type I Error of Global Test through Use of the Clustered Haplotypes Obtained via 1,000 Simulations

Haplotype Diversity	Type I Error							
	Fit I		Fit IIa		Fit IIb		Fit III	
	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$
High:								
$q = .1$	.042	.006	.048	.006	.048	.012	.050	.007
$q = .3$	.032	.002	.028	.002	.040	.002	.039	.010
$q = .5$	.038	.010	.042	.006	.036	.004	.042	.007
Moderate:								
$q = .1$	.044	.012	.048	.006	.046	.004	.052	.012
$q = .3$	.050	.008	.036	.006	.034	.016	.042	.013
$q = .5$	.046	.012	.034	.002	.024	.002	.046	.011
Low:								
$q = .1$	.044	.008	.042	.008	.040	.002	.052	.008
$q = .3$	.044	.006	.046	.008	.036	.008	.032	.002
$q = .5$	.048	.010	.036	.002	.028	.002	.046	.008

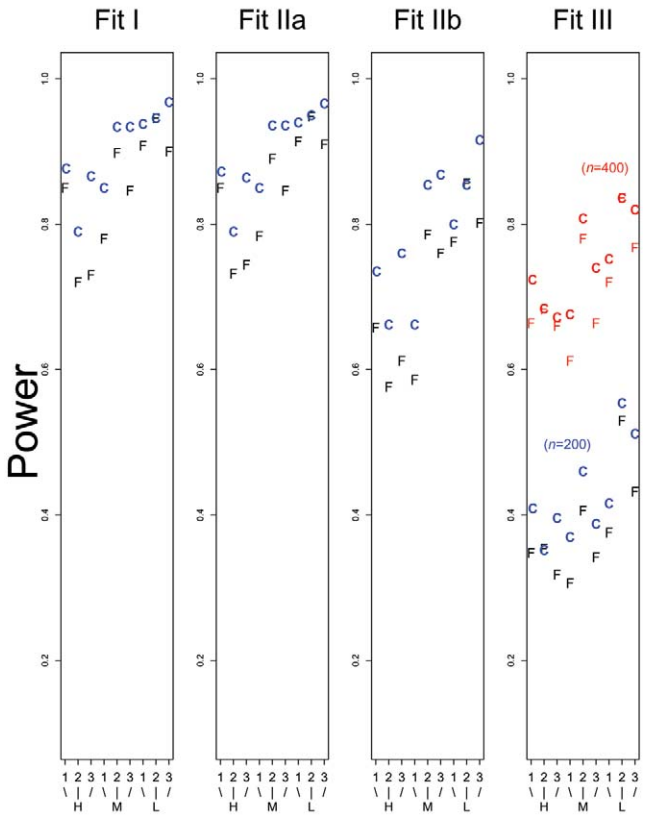

Figure 1 Power of the global score test with the nominal level $\alpha = 0.05$ and sample size $n = 200$ individuals, obtained via 1,000 simulations. The blue C indicates the power of the clustered score test, and the black F indicates the power of the full-dimensional score test. Along the *X*-axis, H = high haplotype diversity, M = moderate haplotype diversity, L = low haplotype diversity, 1 = allele frequency 0.1, 2 = allele frequency 0.3, and 3 = allele frequency 0.5. In the Fit III plot, the results presented in red are the power for doubling the sample size *n* to 400 individuals.

25% of them contain four causal haplotypes. When misdetection occurs, the clustered approach tends to miss one or two causal haplotypes, although it misses three occasionally (the probability is very close to 0, as shown in the histogram plot). Once in a while (~8%), the clustered approach detects extra causal haplotypes that are not identified by the full-dimensional method, but most of the time (~80%) the clustered test detects the same haplotypes as does the full-dimensional test.

Finally, we studied the power and type I error of the haplotype-specific test. The results are displayed in table 2. The trait values are generated in the same way as described above, except that here we predetermine a causal haplotype instead of a causal SNP. We set the frequency of the causal haplotype to 0.1 and consider the haplotype diversity to be low or high. We also consider the scenario of common haplotype frequency (0.4) in a low-diversity setting, which allows us to assess the performance of the clustering method in the least-favorable setting. From table 2, we see that the type I error rates of the haplotype-specific test are around the nominal level. When the causal haplotype is rare in the population (i.e., 0.1), we see power improvement by the clustered score test compared with the full-dimensional test. The power values of the clustered tests are similar to the full-dimensional analysis for a common causal haplotype with a limited haplotype diversity.

Data Application to the Hypertriglyceridemia Study

We applied the proposed method to the study of hypertriglyceridemia conducted at the National Taiwan University Hospital. Hypertriglyceridemia, the elevation of plasma triglyceride concentrations, is a common metabolic disorder in the general population, and its correlation with the risk of cardiovascular diseases remains a subject of enormous attention (Assmann et al. 1996; Gaziano et al. 1997; Jeppesen et al. 1998; Cullen 2000).

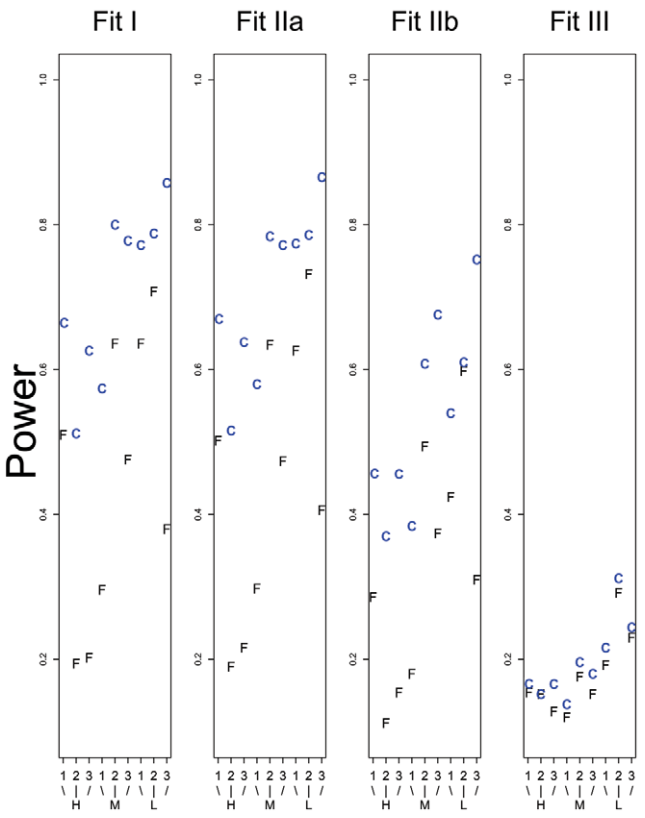

Figure 2 Power of the global score test with the nominal level $\alpha = 0.01$ and sample size $n = 200$ individuals, obtained via 1,000 simulations. The blue C indicates the power of the clustered score test, and the black F indicates the power of the full-dimensional score test. Along the *X*-axis, H = high haplotype diversity, M = moderate haplotype diversity, L = low haplotype diversity, 1 = allele frequency 0.1, 2 = allele frequency 0.3, and 3 = allele frequency 0.5.

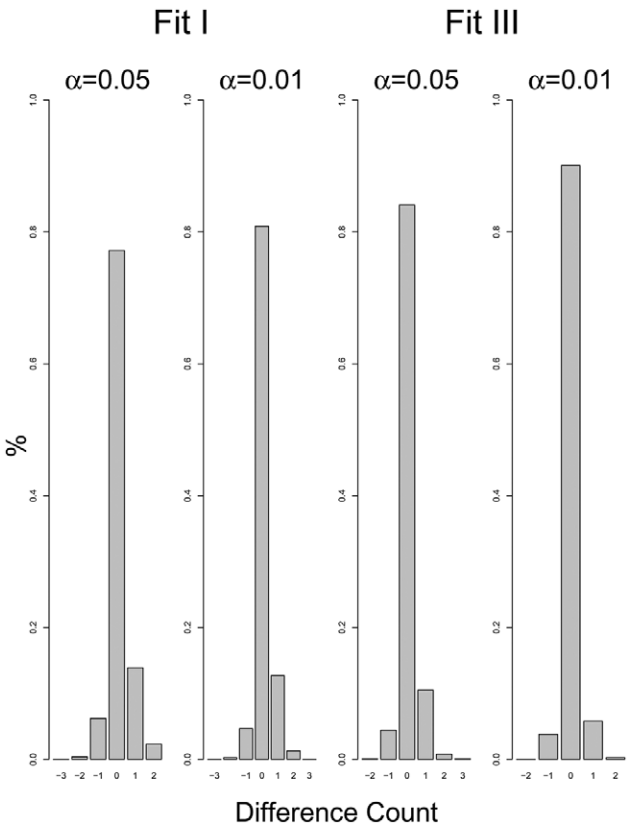

Figure 3 Histogram of difference count: the difference of number of significant haplotypes identified by the full-dimensional analysis and identified by the clustered analysis in the global score test. The *Y*-axis indicates the relative frequency.

Recent research has suggested the association of the variations in the apolipoprotein C-III gene with the differences in triglyceride levels (Ordovas et al. 1991; Peacock et al. 1994; Waterworth et al. 2000, 2001). One main objective of the present study was to investigate the role of genetic polymorphisms in the apolipoprotein C-III gene in hypertriglyceridemia susceptibility.

The present study recruited 290 affected individuals whose serum triglyceride levels were >400 mg/dl and 303 healthy individuals as controls. The controls were recruited through health examinations conducted in the National Taiwan University Hospital. The exclusion criteria were secondary hyperlipoproteinemia, hypertension, diabetes mellitus, medications of lipid-lowering agents, and endocrine or metabolic disorders. All subjects were residents of Taiwan and provided signed informed consent before participating in the study. The study protocol was approved by the Medical Ethics Committee of National Taiwan University Hospital. DNA samples from both the case and control subjects were extracted and were amplified by the PCR technique in a GeneAmpR PCR System (Applied Biosystems Division of Perkin-Elmer).

In particular, Kao et al. (2003) studied *APOA5* on chromosome 11q23 and identified novel variants in this gene region. As a proof check of the proposed method, we applied the proposed score test on the same *APOA5* data set. The polymorphic sites considered in *APOA5* include IVS3+476, c.457, c.553, c.1177, and c.1259; these five SNPs compose the haplotypes in the analysis. We incorporated three environmental covariates—age, sex, and BMI—in the regression model and used the continuous triglyceride level as the dependent variable. We also performed the analysis, using the dichotomized trait values with $Y = 1$ for serum triglyceride >400 mg/dl and $Y = 0$ otherwise. The results of both trait types were similar.

The expectation-maximization algorithm was used to reconstruct 14 distinct haplotypes from the 5-locus genotypes. In the haplotype-clustering regression, we obtained four haplotype groups, represented by GGGCT, GGTCT, AGGCC, and GAGTT, in which these four most-frequent haplotypes explained 95.8% of the total haplotype variation. The global score-test statistic has 3 df and is highly significant (66.78 for the continuous trait and 95.28 for the binary trait; both *P* values are $< 1 \times 10^{-6}$). The first three haplotypes were found to be significant. The haplotype-specific score statistics (*P* values) are 64.16 ($< 1 \times 10^{-6}$), 38.94 ($< 1 \times 10^{-6}$), and 7.62 (.0058) for the continuous trait and 86.86 ($< 1 \times 10^{-6}$), 58.06 ($< 1 \times 10^{-6}$), and 20.43 (6×10^{-6}) for the binary trait. Among these haplotypes, haplotypes GGTCT and AGGCC were shown elsewhere to be associated with increased plasma triglyceride concentration (Pennacchio et al. 2001; Kao et al. 2003). The full-dimensional score test of Schaid et al. (2002) also identified the same three haplotypes to be significantly associated with triglyceride level. The score statistics (and *P* values) of the haplotype-specific test are 64.64 ($< 1 \times 10^{-6}$), 90.25 ($< 1 \times 10^{-6}$), and 8.35 (.0038) for GGGCT, GGTCT, and AGGCC, respectively.

Discussion

Haplotype analyses will likely continue to play an important role in studying common complex diseases (Schaid 2004). Haplotypes effectively capture the joint marker correlation and the evolutionary history; the progressive knowledge of the haplotype structure holds great promise for the use of haplotype information to understand genetic risk factors (International HapMap Consortium 2003). However, naive haplotype analyses can lead to limited performance, since they require many degrees of freedom (Schaid 2004), many of which are expended on rare haplotypes (Chapman et al. 2003). To fully realize the potential of haplotype analyses, new haplotype-based methods are needed—methods that do not dilute

Table 2

Power and Type I Error of Haplotype-Specific Test Obtained via 1,000 Simulations

Haplotype Diversity and Frequency of Causal Haplotype	Power and Type I Error					
	$\alpha = .05$			$\alpha = .01$		
	Full-Dimensional Power	Clustered Power	Clustered Type I Error	Full-Dimensional Power	Clustered Power	Clustered Type I Error
Fit I:						
High diversity:						
$q = .1$	.778	.984	.046	.758	.952	.006
Low diversity:						
$q = .1$	.910	.990	.046	.862	.958	.012
$q = .4$	.998	.998	.054	.984	.986	.010
Fit III:						
High diversity:						
$q = .1$	.438	.666	.045	.296	.424	.010
Low diversity:						
$q = .1$	.500	.642	.042	.270	.386	.007
$q = .4$	.768	.786	.052	.524	.544	.007

their power to detect interesting gene-trait relationships among many distinct haplotypes.

In the present study, we introduce one such test for assessing haplotype-phenotype association. To overcome the degrees-of-freedom problem, our strategy is to analyze groups of homogeneous haplotypes within which haplotypes share similar effects on phenotypes. The proposed method uses unphased genotypes and provides several advantages in performing haplotype analyses. It offers an integrated procedure for haplotype analysis, including phase reconstruction, haplotype clustering, and inference of haplotype effects. By combining the merits of the GLM score test of Schaid et al. (2002) and the probabilistic clustering technique of Tzeng (2005), the proposed approach incorporates uncertainties that arise from missing haplotype phase and unknown haplotype tree in the assessment of haplotype-phenotype association. The method is constructed under a model-based framework; hence, it can accommodate a wide range of trait values. It allows simultaneous consideration of the multiple environmental and genetic factors that underlie complex traits. It can also be used to evaluate either the overall haplotype association or the individual haplotype effects. Simulation results show that the clustered score test has correct type I error rates and can improve power to detect association at either the global or haplotype-specific level when compared with the full-dimensional method.

The proposed method also has its limitations. Motivated by the common disease/common variants hypothesis, the clustered test is designed to identify common polymorphisms with small or large effect. It is incapable of detecting rare variants with large effect because rare haplotypes are not retained in the clustered haplotype space. Next, established on the Tzeng (2005) algorithm, the proposed method also inherits its major assumption that the haplotype diversity is due mainly to mutation; other diversifying forces, such as recombination, are negligible in evolution. As a result, the proposed method would be more appropriate to apply to tightly linked DNA regions. Furthermore, unlike several proposed clustering methods for fine mapping (e.g., that of Molitor et al. 2003*a,* 2003*b*), our method does not take into account the location and/or order of the markers. For example, if one permutes the SNPs and applies our clustering algorithm, we would expect to get the same answers. Thus, our method is more suitably applied for studying haplotype association—such as the goal of candidate-gene studies—than for mapping purposes.

Finally, our method is derived on the basis of perspective likelihood, which can be less efficient than retrospective approaches for case-control samples when haplotypes have a nonmultiplicative effect on the disease odds (Satten and Epstein 2004). In our simulation, we also observed a less-significant power gain in fit III (case-control data) when $\alpha = 0.01$. However, our proposed method is just one of many possible ways to integrate regression methods with dimension-reduction techniques. One of our key findings is the simple presentation of the clustering algorithm through a linear transformation: $\boldsymbol{X}_C^{\prime} = \boldsymbol{X}_F^{\prime}\boldsymbol{B}(\Pi)$. This deduction offers a convenient path for extending our results to a wide range of regression-based methods, including the retrospective method of Epstein and Satten (2003). We plan our further research along this path.

Acknowledgments

The authors thank the reviewers for their constructive and detailed comments, which improved the manuscript. J.Y.T. was supported by National Institutes of Health grant GM45344 and National Science Foundation grant DMS-0504726.

Appendix A

Let $S_\alpha(Y,\boldsymbol{G},\boldsymbol{Z},\boldsymbol{\alpha},\boldsymbol{\zeta})$ denote the score function of the observed data $(Y,\boldsymbol{G},\boldsymbol{Z})$ for $\boldsymbol{\alpha}$. As set forth by Louis (1982), $S_\alpha(Y,\boldsymbol{G},\boldsymbol{Z},\boldsymbol{\alpha},\boldsymbol{\zeta})$ is the expectation of the complete-data score function given the observed data—that is, $S_\alpha(Y,\boldsymbol{G},\boldsymbol{Z},\boldsymbol{\alpha},\boldsymbol{\zeta}) = \mathrm{E}\,[S_\alpha(Y,\boldsymbol{X}_F,\boldsymbol{Z},\boldsymbol{\alpha},\boldsymbol{\zeta}) \mid \boldsymbol{G}]$. Hence,

$$
\begin{aligned}
S_\alpha(Y,\boldsymbol{G},\boldsymbol{Z},\boldsymbol{\alpha},\boldsymbol{\zeta}) &= \sum_{i=1}^{n} \mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\alpha}} \log L(\boldsymbol{\alpha},\boldsymbol{\zeta}; y_i,\boldsymbol{x}_{F,i},\boldsymbol{z}_i) \,\middle|\, \boldsymbol{g}_i\right] \\
&= \sum_{i=1}^{n} \mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\alpha}} [\log f(y_i|\boldsymbol{x}_{F,i},\boldsymbol{z}_i;\boldsymbol{\alpha},\boldsymbol{\zeta}) + \log \mathrm{P}(\boldsymbol{x}_{F,i};\Pi)] \,\middle|\, \boldsymbol{g}_i\right] \\
&= \sum_{i=1}^{n} \mathrm{E}\left[\frac{y_i - b'(\eta)}{a(\phi)}\ \boldsymbol{B}(\Pi)'_{-0} X_i \mid \boldsymbol{g}_i\right] \\
&= \sum_{i=1}^{n} \frac{y_i - \mathrm{E}(y_i)}{a(\phi)}\ \boldsymbol{B}(\Pi)'_{-0} \mathrm{E}(X_i \mid \boldsymbol{g}_i)\ .
\end{aligned}
$$

Appendix B

Let $\Gamma = (\mu,\gamma)$. The expected Fisher information function of the observed data $(Y,\boldsymbol{G},\boldsymbol{Z})$, $\boldsymbol{I}$, is

$$
\boldsymbol{I} = \begin{pmatrix} I_{\alpha\alpha} & I_{\alpha\Gamma} & I_{\alpha\phi} & I_{\alpha\Pi} \\ I'_{\alpha\Gamma} & I_{\Gamma\Gamma} & I_{\Gamma\phi} & I_{\Gamma\Pi} \\ I'_{\alpha\phi} & I'_{\Gamma\phi} & I_{\phi\phi} & I_{\phi\Pi} \\ I'_{\alpha\Pi} & I'_{\Gamma\Pi} & I'_{\phi\Pi} & I_{\Pi\Pi} \end{pmatrix}\ ,
$$

where

$$
\begin{aligned}
I_{\alpha\phi} &= 0_{L^* \times 1}\ , \\
I_{\alpha\Pi} &= 0_{L^* \times (L+1)}\ , \\
I_{\Gamma\phi} &= 0_{(1+P) \times 1}\ , \\
I_{\Gamma\Pi} &= 0_{(1+P) \times (L+1)}\ .
\end{aligned}
$$

and

$$
I_{\phi\Pi} = 0_{1 \times (L+1)}\ .
$$

The hybrid estimate of $\boldsymbol{I}$ is obtained by replacing the nonzero entries of $\boldsymbol{I}$ with the observed Fisher information (denoted by $\boldsymbol{i}$):

$$
\boldsymbol{I} = \begin{pmatrix} i_{\alpha\alpha} & i_{\alpha\Gamma} & 0 & 0 \\ i'_{\alpha\Gamma} & i_{\Gamma\Gamma} & 0 & 0 \\ 0 & 0 & i_{\phi\phi} & 0 \\ 0 & 0 & 0 & i_{\Pi\Pi} \end{pmatrix}\ .
$$

Hence, equation (6) can be simplified as

$$
\boldsymbol{V}_\alpha = D_{\alpha\alpha} - i_{\alpha\Gamma} i_{\Gamma\Gamma}^{-1} D'_{\alpha\Gamma} - D_{\alpha\Gamma} i_{\Gamma\Gamma}^{-1} i'_{\alpha\Gamma} + i_{\alpha\Gamma} i_{\Gamma\Gamma}^{-1} D_{\Gamma\Gamma} i_{\Gamma\Gamma}^{-1} i'_{\alpha\Gamma}\ .
$$

Recall that

$$
\boldsymbol{D} = \sum_{i=1}^{n} s_i(y_i,\boldsymbol{g}_i,\boldsymbol{z}_i,\boldsymbol{\Theta}) s'_i(y_i,\boldsymbol{g}_i,\boldsymbol{z}_i,\boldsymbol{\Theta})
$$

and that Louis (1982) proposed

$$
s_i(y_i,\boldsymbol{g}_i,\boldsymbol{z}_i,\boldsymbol{\Theta}) = \mathrm{E}\,[s_i(y_i,\boldsymbol{x}_{F,i},\boldsymbol{z}_i,\boldsymbol{\Theta}) \mid \boldsymbol{g}_i]
$$

and

$$
\begin{aligned}
\boldsymbol{i} = \sum_{i=1}^{n} \Bigg\{ & \mathrm{E}\left[-\frac{\partial s_i(y_i,\boldsymbol{x}_{F,i},\boldsymbol{z}_i,\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} \ \middle|\ \boldsymbol{g}_i \right] \\
& - \mathrm{E}\left[s_i(y_i,\boldsymbol{x}_{F,i},\boldsymbol{z}_i,\boldsymbol{\Theta}) s'_i(y_i,\boldsymbol{x}_{F,i},\boldsymbol{z}_i,\boldsymbol{\Theta}) \ \middle|\ \boldsymbol{g}_i \right] \\
& + \mathrm{E}\,[s_i(y_i,\boldsymbol{x}_{F,i},\boldsymbol{z}_i,\boldsymbol{\Theta}) \mid \boldsymbol{g}_i]\, \mathrm{E}\,[s'_i(y_i,\boldsymbol{x}_{F,i},\boldsymbol{z}_i,\boldsymbol{\Theta}) \mid \boldsymbol{g}_i] \Bigg\}\ .
\end{aligned}
$$

We have

$$
\begin{aligned}
D_{\alpha\alpha} &= \sum_{i=1}^{n} \left(\frac{y_i - b'(\eta_i)}{a(\phi)}\right)^2 \boldsymbol{B}(\Pi)'_{-0} \mathrm{E}(\boldsymbol{x}_{F,i}|g) \mathrm{E}(\boldsymbol{x}'_{F,i}|g) \boldsymbol{B}(\Pi)_{-0}\ , \\
D_{\alpha\Gamma} &= \sum_{i=1}^{n} \left(\frac{y_i - b'(\eta_i)}{a(\phi)}\right)^2 \boldsymbol{B}(\Pi)'_{-0} \mathrm{E}(\boldsymbol{x}_{F,i}|g)[1\ z'_i]\ , \\
D_{\Gamma\Gamma} &= \sum_{i=1}^{n} \left(\frac{y_i - b'(\eta_i)}{a(\phi)}\right)^2 \begin{bmatrix} 1 \\ z_i \end{bmatrix} [1\ z'_i]\ , \\
i_{\alpha\Gamma} &= \sum_{i=1}^{n} \frac{b''(\eta)}{a(\phi)} \boldsymbol{B}(\Pi)'_{-0} \mathrm{E}(\boldsymbol{x}_{F,i}|g)[1\ z'_i]\ ,
\end{aligned}
$$

and

$$
i_{\Gamma\Gamma} = \sum_{i=1}^{n} \frac{b''(\eta)}{a(\phi)} \begin{bmatrix} 1 \\ z_i \end{bmatrix} [1\ z'_i]\ .
$$

Web Resources

The URL for data presented herein is as follows:

Authors' Web site, http://www4.stat.ncsu.edu/~tzeng/Softwares/Hap-Clustering/R/ (for *R* codes for implementing the proposed test)

References

Assmann G, Schulte H, von Eckardstein A (1996) Hypertriglyceridemia and elevated levels of lipoprotein(a) are risk factors for coronary events in middle-aged men. Am J Cardiol 77:1179–1184

Boos DD (1992) On generalized score tests. Am Stat 46:327–333

Bourgain C, Génin E, Holopainen P, Mustalahti K, Mäki M, Partanen J, Clerget-Darpoux F (2001) Use of closely related affected individuals for the genetic study of complex diseases in founder populations. Am J Hum Genet 68:154–159

Bourgain C, Genin E, Ober C, Clerget-Darpoux F (2002) Missing data in haplotype analysis: a study on the MILC method. Ann Hum Genet 66:99–108

Bourgain C, Genin E, Quesneville H, Clerget-Darpoux F (2000) Search for multifactorial disease susceptibility genes in founder populations. Ann Hum Genet 64:255–265

Chapman JM, Cooper JD, Todd JA, Clayton DG (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Hum Hered 56:18–31

Clayton D, Chapman J, Cooper J (2004) Use of unphased multilocus genotype data in indirect association studies. Genet Epidemiol 27: 415–428

Collins FS, Guyer MS, Charkravarti A (1997) Variations on a theme: cataloging human DNA sequence variation. Science 278:1580–1581

Crandall KA, Templeton AR (1993) Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. Genetics 134:959–969

Cullen P (2000) Evidence that triglicerides are an independent coronary heart disease risk factor. Am J Cardiol 86:943–949

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29: 229–232

Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P, Morris AP (2004) Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. Am J Hum Genet 75: 35–43

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

Gaziano JM, Hennekens CH, O'Donnell CJ, Breslow JL, Buring JE (1997) Fasting triglycerides, high-density lipoprotein, and risk of myocardial infarction. Circulation 96:2520–2525

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338

International HapMap Consortium (2003) The International HapMap Project. Nature 426:789–796

Jeppesen J, Hein HO, Suadicani P, Gyntelberg F (1998) Triglyceride concentration and ischemic heart disease: an eight-year follow-up in the Copenhagen Male Study. Circulation 97:1029–1036

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA (2001) Haplotype tagging for the identification of common disease genes. Nat Genet 29:233–237

Kao JT, Wen HC, Chien KL, Hsu HC, Lin SW (2003) A novel genetic variant in the apolipoprotein A5 gene is associated with hypertriglyceridemia. Hum Mol Genet 12:2533–2539

Kent JT (1982) Robust properties of likelihood ratio tests. Biometrika 69:19–27

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

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

Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 33:177–182

Louis TA (1982) Finding the observed information matrix when using the EM algorithm. J R Statist Soc B 44:226–233

McPeek MS, Strahs A (1999) Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. Am J Hum Genet 65:858–875

Molitor J, Marjoram P, Thomas D (2003*a*) Application of Bayesian spatial statistical methods to analysis of haplotypes effects and gene mapping. Genet Epidemiol 25:95–105

——— (2003*b*) Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. Am J Hum Genet 73:1368–1384

Neale BM, Sham PC (2004) The future of association studies: gene-based analysis and replication. Am J Hum Genet 75:353–362

Ordovas JM, Civeira F, Genest J Jr, Craig S, Robbins AH, Meade T, Pocovi M, Frossard PM, Masharani U, Wilson PWF, Salem DN, Ward RH, Schaefer EJ (1991) Restriction fragment length polymorphisms of the apolipoprotein A-I, C-III, A-IV gene locus: relationships with lipids, apolipoproteins, and premature coronary artery disease. Atherosclerosis 87:75–86

Peacock RE, Hamsten A, Johansson J, Nilsson-Ehle P, Humphries SE (1994) Associations of genotypes at the apolipoprotein AI-CIII-AIV, apolipoprotein B and lipoprotein lipase gene loci with coronary atherosclerosis and high density lipoprotein subclasses. Clin Genet 46: 273–282

Peltonen L, McKusick VA (2001) Genomics and medicine: dissecting human disease in the postgenomic era. Science 291:1224–1229

Pennacchio LA, Olivier M, Hubacek JA, Cohen JC, Cox DR, Fruchart JC, Krauss RM, Rubin EM (2001) An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. Science 294:169–173

Rinaldo A, Bacanu SA, Devlin B, Sonpar V, Wasserman L, Roeder K (2005) Characterization of multilocus linkage disequilibrium. Genet Epidemiol 28:193–206

Roeder K, Bacanu SA, Sonpar V, Zhang X, Devlin B (2005) Analysis of single-locus tests to detect gene/disease associations. Genet Epidemiol 28:207–219

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

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

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

Seltman H, Roeder K, Devlin B (2001) Transmission/disequilibrium test meets measured haplotype analysis: family-based association analysis guided by evolution of haplotypes. Am J Hum Genet 68:1250–1263

——— (2003) Evolutionary-based association analysis using haplotype data. Genet Epidemiol 25:48–58

Sham P (1998) Statistics in human genetics. Arnold, New York

Shannon CE (1948) A mathematical theory of communication. Bell System Tech J 27:379–423, 623–656

Slatkin M, Rannala B (1997) Estimating the age of alleles by use of intraallelic variability. Am J Hum Genet 60:447–458

Stephens J, Schneider J, Tanguay D, Choi J, Acharya T, Stanley S, Jiang R, et al (2001) Haplotype variation and linkage disequilibrium in 313 human genes. Science 293:489–493

Stram DO, Pearce CL, Bretsky P, Freedman M, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Thomas DC (2003) Modeling

and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. Hum Hered 55:179–190

Tzeng JY (2005) Evolutionary-based grouping of haplotypes in association analysis. Genet Epidemiol 28:220–231

Tzeng JY, Byerley W, Devlin B, Roeder K, Wasserman L (2003*a*) Outlier detection and false discovery rates for whole-genome DNA matching. J Am Stat Assoc 98:236–246

Tzeng J-Y, Devlin B, Wasserman L, Roeder K (2003*b*) On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. Am J Hum Genet 72:891–902

Van der Meulen MA, te Meerman GJ (1997) Haplotype sharing analysis in affected individuals from nuclear families with at least one affected offspring. Genet Epidemiol 14:915–919

Wall JD, Pritchard JK (2003) Assessing the performance of the haplotype block model of linkage disequilibrium. Am J Hum Genet 73: 502–515

Waterworth DM, Talmud PJ, Bujac SR, Fisher RM, Miller GJ, Humphries SE (2000) Contribution of apolipoprotein C-III gene variants to determination of triglyceride levels and interaction with smoking in middle-aged men. Arterioscler Thromb Vasc Biol 20:2663–2669

Waterworth DM, Talmud PJ, Humphries SE, Wicks PD, Sagnella GA, Strazullo P, Alberti KG, Cook DG, Cappuchio FP (2001) Variable effects of the APOC3-482C→T variant on insulin, glucose and triglyceride concentrations in different ethnic groups. Diabetalogia 44: 245–248

Yu K, Gu CC, Province M, Xiong CJ, Rao DC (2004) Genetic association mapping under founder heterogeneity via weighted haplotype similarity analysis in candidate genes. Genet Epidemiol 27:182–191

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

Zeng D, Lin DY (2005) Estimating haplotype-disease associations with pooled genotype data. Genet Epidemiol 28:70–82

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