doi: 10.1111/j.1469-1809.2006.00262.x

A Combinatorial Searching Method for Detecting A Set of Interacting Loci Associated with Complex Traits

Qiuying Sha[1], Xiaofeng Zhu[2], Yijun Zuo[3], Richard Cooper[2], and Shuanglin Zhang[1,4]*

[1]*Department of Mathematical Sciences Michigan Technological University, Houghton, MI 49931*
[2]*Department of Preventive Medicine and Epidemiology, Loyola University Stritch School of Medicine, Maywood, IL*
[3]*Department of Statistics and Probability Michigan State University, East Lansing, MI 48824*
[4]*Department of Mathematics, Heilongjiang University Harbin 150080, China*

Summary

Complex diseases are presumed to be the results of the interaction of several genes and environmental factors, with each gene only having a small effect on the disease. Mapping complex disease genes therefore becomes one of the greatest challenges facing geneticists. Most current approaches of association studies essentially evaluate one marker or one gene (haplotype approach) at a time. These approaches ignore the possibility that effects of multilocus functional genetic units may play a larger role than a single-locus effect in determining trait variability. In this article, we propose a Combinatorial Searching Method (CSM) to detect a set of interacting loci (may be unlinked) that predicts the complex trait. In the application of the CSM, a simple filter is used to filter all the possible locus-sets and retain the candidate locus-sets, then a new objective function based on the cross-validation and partitions of the multi-locus genotypes is proposed to evaluate the retained locus-sets. The locus-set with the largest value of the objective function is the final locus-set and a permutation procedure is performed to evaluate the overall p-value of the test for association between the final locus-set and the trait. The performance of the method is evaluated by simulation studies as well as by being applied to a real data set. The simulation studies show that the CSM has reasonable power to detect high-order interactions. When the CSM is applied to a real data set to detect the locus-set (among the 13 loci in the ACE gene) that predicts systolic blood pressure (SBP) or diastolic blood pressure (DBP), we found that a four-locus gene-gene interaction model best predicts SBP with an overall p-value = 0.033, and similarly a two-locus gene-gene interaction model best predicts DBP with an overall p-value = 0.045.

Keywords: epistasis, complex disease, gene-gene interaction

Introduction

Searching for a set of susceptibility genes responsible for a complex trait is one of the greatest challenges facing geneticists. There is increasing evidence suggesting that gene-gene and gene-environment interactions play an important role in liability to complex diseases (Risch 2000; Risch *et al.* 1999; Nicolae & Cox 2002; Carrasquillo *et al.* 2002; Olson *et al.* 2002; Hoh & Ott 2003; Trornton *et al.* 2004). Methods to search for a set of marker loci in different genes and to analyze these loci jointly are therefore critical. Most current approaches of association studies in practice essentially evaluate one locus at a time. These methods make the implicit assumption that susceptibility loci can each be identified through their independent, marginal contributions to the trait variability. This simplified approach ignores the possibility that effects of multilocus functional genetic units play a larger role than the single-locus effect in determining trait variability (Nelson *et al.* 2001; Hoh *et al.* 2001; Templeton 2000). Forming haplotypes over multiple neighboring loci in one gene can increase the power of gene mapping studies (Zhao *et al.* 2000; Fallin

*Corresponding author: Shuanglin Zhang, Ph.D., Department of Mathematical Sciences, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931, Phone: (906) 487-2095, Fax: (906) 487-3133. E-mail: shuzhang@mtu.edu *et al.* 2001; Schaid *et al.* 2002; Zhang *et al.* 2003a), but these methods only work locally in a given genomic region. Although various authors have postulated the need for investigating multiple interacting genes jointly (Tiwari & Elston 1998; Cox *et al.* 1999; Templeton 2000; Wilson 2001; Cordell *et al.* 2001; Cordell 2002; Culverhouse *et al.* 2002; Moore & Williams 2002; Moore 2003), only a few viable approaches in this direction exist (Hoh *et al.* 2001).

Two intriguing methods have recently been proposed by Nelson *et al.* (2001) and Ritchie *et al.* (2001, 2003) to allow for the joint analysis of multiple-marker loci for quantitative traits and qualitative traits, respectively. Nelson *et al.*'s (2001) Combinatorial Partitioning Method (CPM) works by evaluating all possible partitions of multi-locus genotypes and retaining only those partitions fulfilling certain optimal criteria. Using the CPM, Nelson *et al.* (2001) detected clinical interactions between loci that individually showed little or no effect on the phenotype. Although 2-way interactions can be analyzed with the CPM, the number of possible partitions with three biallelic loci is over 10^{21}. Clearly, the CPM is not feasible if we analyze the interactions involving more than two loci. The Multifactor Dimensionality Reduction (MDR) method proposed by Ritchie *et al.* (2001, 2003) and recently reviewed by Moore (2004) is designed for detecting and characterizing high-order gene-gene and gene-environment interactions in a balanced case-control design. With the MDR, multilocus genotypes are pooled into high-risk and low-risk groups, reducing the genotype predictor from high dimensions to one dimension. The new one-dimensional multilocus-genotype variable is used to choose the best set of loci from every two- to *L*-locus sets according to classification and prediction errors. However, the MDR method is only applicable to dichotomous traits.

In this paper, we present an alternative method, the Combinatorial Searching Method (CSM). To apply the CSM to detect a set of interacting loci (possibly unlinked) that predict the complex trait, a simple filter is first used to filter all the possible locus-sets and retain the candidate locus-sets, then a new objective function based on the cross-validation and partitions of the multilocus genotypes is proposed to evaluate the retained locus-sets. The locus-set with the largest value of the objective function is the final locus-set and a permutation procedure is used to evaluate the p-value of the test for association between the final locus-set and the trait. The simulation studies show that the CSM has reasonable power to detect high-order interactions. We also apply the method to the ACE data set (Zhu *et al.* 2001) to identify two sets of loci that "best" predict SBP and DBP, respectively.

Methods

The objective of the CSM is to identify a set of loci that predicts the trait variability. Suppose that *K* SNP loci are genotyped in each of the sampled individuals. The application of the CSM to identify a subset of the *K* loci can be divided into three steps as described in Figure 1. Here we describe each of these steps in detail for both quantitative traits and qualitative traits. In the following discussion, a locus-set means a set of loci, and a *l*-locus set means a locus-set with *l* loci.

Step 1: Search for candidate locus-sets

First, we search every single-locus set and retain those that explain a significant amount of trait variability. Next, we search among all the two-locus sets and retain the two-locus sets that explain a significant amount of trait variability, then consider three- to *L*-locus sets (*L* is a pre-specified number). To evaluate the locus-sets, we need a statistical function (or an objective function) that

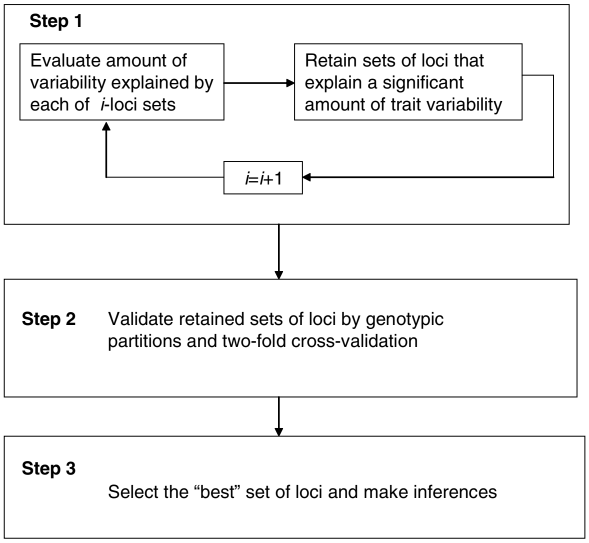
Step 1
Evaluate amount of variability explained by each of *i*-loci sets
Retain sets of loci that explain a significant amount of trait variability
i=*i*+1
Step 2
Validate retained sets of loci by genotypic partitions and two-fold cross-validation
Step 3
Select the "best" set of loci and make inferences

Figure 1 The three steps of the CSM.

provides a measure for each of the locus-sets. When we compare among the different locus-sets with the same number of loci, the correlation coefficient between trait values and numerical codes of the multi-locus genotypes is a choice of the objective function. The numerical codes of the multi-locus genotypes can be defined in many ways. In this article, we use the following way to define the numerical code. For a sample of size n, consider a l-locus set ($1 \leq l \leq L$). Let $g_1, \ldots, g_{m+1}$ denote all the distinct multi-locus genotypes observed in the sample, where $m + 1$ is the total number of distinct multi-locus genotypes in the sample. Define a numerical code for the multi-locus genotype of the ith individual as a numerical vector $X_i = (x_{i1}, \ldots, x_{im})$, where

$$x_{ij} = \begin{cases} 1 & \text{if the genotype of } i\text{th individual is } g_j \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

To define the correlation between the trait values and the numerical codes, we first project the multi-dimensional vector X_i into a one-dimensional number by

$$x_i = \sum_{j=1}^{m} \alpha_j x_{ij},$$

where $\alpha_1, \ldots, \alpha_m$ are parameters. We estimate $\alpha = (\alpha_1, \ldots, \alpha_m)^T$ by $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_m)^T$ which maximize the correlation between trait values and one-dimensional genotype scores, that is,

$$\rho^2(\hat{\alpha}) = \max_{\alpha} \rho^2(\alpha),$$

where $\rho(\alpha)$ is the correlation coefficient between the trait values and the one-dimensional codes $x_1, \ldots, x_n$. Using $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_m)^T$, we define a new one-dimensional space or direction that captures the maximum information of correlation between the trait and the genotype code in the initial data. Let $\hat{x}_i = \sum_{j=1}^{m} \hat{\alpha}_j x_{ij}$ denote the genotypic code in the “best” direction and denote $R^2 = \rho^2(\hat{\alpha})$. Then R^2 is the square of the correlation coefficient between the trait values and $\hat{x}_1, \ldots, \hat{x}_n$. Let y_i denote the trait value of the ith individual (for a qualitative trait, denote affected as 1 and unaffected as 0). The above procedure is equivalent to the following linear model setting. Assume that the trait y_i and the numerical genotypic code $X_i = (x_{i1}, \ldots, x_{im})$ follow the linear model

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \cdots + \alpha_m x_{im} + \epsilon_i, \tag{2}$$

then $\hat{\alpha}_1 \ldots, \hat{\alpha}_m$ given above are also the least-squares estimators of $\alpha_1, \ldots, \alpha_m$. Let $\hat{y}_i = \hat{\alpha}_0 + \hat{\alpha}_1 x_{i1} + \cdots + \hat{\alpha}_m x_{im}$ be the predicted trait value of the ith individual under linear model (2). Then $R^2 = \rho^2(\hat{\alpha})$ given above is the square of the correlation coefficient between the trait y_i and predicted trait value $\hat{y}_i$, and R^2 also represents the proportion of the total variance of the trait value explained by the genotype. Although R^2 is a reasonable measure as an objective function to compare the data sets with the same number of independent variables under a linear model, a disadvantage of R^2 is that it will tend to increase as the number of independent variables increases, and thus will favor data sets with more independent variables. Theoretically, two locus-sets with the same number of loci will have the same number of genotypes, therefore the same number of independent variables. However, due to the rare allele frequencies of some markers, some of the genotypes for a specific locus-set may not appear in the sample. Thus, two locus-sets with the same number of loci may have a different number of genotypes and thus a different number of independent variables. Based on these considerations, we propose to use Leave-One-Out Cross-Validation (LOOCV) to calculate R^2. The LOOCV R^2, denoted by R_1^2, is the square of correlation coefficient between the trait y_i and the LOOCV predicted trait value $\hat{y}_{-i}$. To calculate $\hat{y}_{-i}$, remove the ith individual from the sample and use the data of the remaining $n - 1$ individuals to calculate $\hat{\alpha}_j$, the least-squared estimator of α_j ($j = 0, 1, \ldots, m$), then $\hat{y}_{-i} = \hat{\alpha}_0 + \hat{\alpha}_1 x_{i1} + \cdots + \hat{\alpha}_m x_{im}$. Comparing to R^2, R_1^2 is a less biased estimator of the population’s proportion of the trait variability explained by genotypes (Goutte 1997; Stone 1977). Under linear model (2), we are able to give a simple formula to allow quick computation of R_1^2 (Hastie *et al.* 2001, page 216). When R_1^2 is used to compare locus-sets with the same number of loci, many criteria can be used to decide which locus-sets should be retained for further studies, including biological significance (e.g., $R_1^2 \geq 0.05$), the top 1% or the top 10 of the locus-sets for a fixed number of loci, or simply the best. In our simulation studies, for computational consideration, we choose the best one as the retained locus-set (the one with the largest value of R_1^2) for each of the one- to L-locus sets. In the application of the CSM to a real data set, we choose the top 10 locus-sets for each of the one- to L-locus sets.

Step 2: Validate the retained locus-sets

Since the retained locus-sets from step 1 were searched from a large number of locus-sets, model validation is critical in this situation (Coffey *et al.* 2004a, b). In the second step, we will validate and compare the locus-sets retained in the first step. To validate and compare the locus-sets retained in the first step, we need to consider the following problems: (1) When the number of genotypes $m + 1$ is large, we need to use some dimension reduction methods to deal with the sparse data; (2) In the first step, having chosen a locus-set that is "good" or "best" for a particular sample of data, we have no assurance that the locus-set can be reliably applied to other samples, and thus, need to verify the reliability of the retained locus-sets; (3) Although R_1^2 is a good measure to compare locus-sets with the same number of loci, R_1^2 still tends to increase with the number of independent variables, and the locus-sets retained in the first step have different numbers of loci and thus have quite different numbers of independent variables. Based on these considerations, we propose the following method to calculate the value of the objective function used to compare the locus-sets retained in step 1.

To deal with the sparse data due to the large number of multilocus genotypes, we use Nelson *et al.*'s (2001) idea of partitions or groups of multilocus genotypes. Nelson *et al.* (2001) proposed to evaluate every possible partition of the genotypes. It makes the computation infeasible due to the large number of possible partitions if we consider interactions involving more than two biallelic loci. As noted by Culverhouse *et al.* (2004), a large part of the partitions is unnecessary to be evaluated. In fact, a good partition should have the property that genotypes with similar trait values will be in the same group. Based on this consideration, we propose to find the approximate "best" partition of the genotypes by the K-mean clustering method (Richard *et al.* 1998, see Appendix for detail) to group the multilocus genotypes with similar trait values into the same group. For a given locus-set with $m + 1$ multilocus genotypes, we propose to cluster the genotypes into k groups where $k = 2, 3, \ldots, m + 1$. For example, when we consider three biallelic loci, the number of partitions evaluated by Nelson *et al.* (2001)'s method is over 10^{21}, while we only evaluate 26 different cases ($k = 2, \ldots, 27$).

For a given number of groups k, we denote $G_1, \ldots, G_k$ to be the k genotype groups found by the K-mean clustering method. We take $G_1, \ldots, G_k$ as if they were k different genotypes and define a numerical code for the multi-locus genotype of the ith individual as a numerical vector $X_i = (x_{i,1}, \ldots, x_{i,k-1})$, where

$$x_{i,j} = \begin{cases} 1 & \text{if the genotype of } i\text{th individual belongs to } G_j \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

To study the reliability of a locus-set and find a suitable objective function to compare the locus-sets with different numbers of loci, we use a bagging version of split-sample analysis (or bagging version of two-fold cross-validation). With this method, we randomly split the sample into two groups with approximately the same sample size. First, we use the first group as a training group and the second group as a test group. Using the data of the training group and new numerical codes in (3), we calculate $\hat{\alpha} = (\hat{\alpha}_0, \hat{\alpha}_1, \ldots, \hat{\alpha}_{k-1})^T$ under linear model (2) and use the prediction equation $\hat{y} = \hat{\alpha}_0 + \sum_{j=1}^{k-1} \hat{\alpha}_j x_j$ to predict trait values of individuals in the training group and in the test group. Then, we use the second group as a training group and the first group as a test group re-do the above procedure. In this way, each individual, individual i for example, has two prediction values: one, denoted by $\hat{y}_i$, is using training group to predict training group (called training to training prediction), the other one, denoted by $\hat{y}_i^*$, is using training group to predict test group (called training to test prediction). We repeat this randomly split procedure many times (100 times in this paper) and calculate the average of the two kinds of prediction values for each of the individuals. Denote average training to training prediction trait values by $\bar{\hat{y}}_1, \ldots, \bar{\hat{y}}_n$ and average training to test prediction values by $\bar{\hat{y}}_1^*, \ldots, \bar{\hat{y}}_n^*$. Denote the square of the correlation coefficient between $y_1, \ldots, y_n$ and $\bar{\hat{y}}_1, \ldots, \bar{\hat{y}}_n$ by $R^2(1)$ and the square of the correlation coefficient between $y_1, \ldots, y_n$ and $\bar{\hat{y}}_1^*, \ldots, \bar{\hat{y}}_n^*$ by $R^2(2)$. The quantity $R^2(2)$ is called the cross-validation correlation, and the quantity $S^2 = (R^2\ (1) - R^2(2))/(R^2(1) + R^2(2))$ is called the shrinkage on cross-validation. Typically, comparing to $R^2(1)$, $R^2(2)$ is a less biased estimator of the population proportion of the trait value variability explained by genotypes. S^2 is a measure that indicates the reliability of the model (how a "good" or "best" locus-set for a particular sample of data can be reliably applied to other samples). The small value of S^2 means that the model or the locus-set is reliable, i.e. the locus-set can predict the trait value as well in any new sample as it does in the sample at hand. So, a "good" locus-set should correspond to a large value of $R^2(2)$ and a small value of S^2 . From this argument, $\frac{R^2(2)}{S}$ can be a reasonable measure to compare different locus-sets. Denote

$$T(k) = \frac{R^2(2)}{S}, \tag{4}$$

where k is the number of genotype groups. For a specific locus-set with $m + 1$ distinct multilocus genotypes observed in the sample, the final value of the objective function of the locus-set is given by

$$T = \max_{2 \leq k \leq m+1} T(k). \tag{5}$$

In summary, for a locus-set with $m + 1$ distinct genotypes in the sample, the algorithm to calculate the objective function proceeds as follows:(starting from $k = 2$)

(1) Divide the genotypes into k groups by using the K-mean clustering method and code the genotypes by equation (3).
(2) Randomly split the sample into two groups with approximately equal sample sizes, and calculate two prediction values $\hat{y}_i$ and $\hat{y}_i^*$ for individual i ($i = 1, \ldots n$).
(3) Repeat step (2) 100 times. Calculate average of the two kinds of prediction values for each individual, and then R^2 (2), S, and $T(k)$.
(4) Repeat steps (1) to (3) for $k = 2, \ldots, m + 1$. Calculate the value of the objective function of this locus-set by equation (5).

Although in each of the randomly split we use only half of the sample to fit the model and to predict the trait values, the final prediction value is the average of prediction values predicted by half of the sample. This is the idea of bagging (Breiman 1996; Friedman & Hall 1999) which is a very popular learning method in machine learning and data mining to improve prediction accuracy. In most of cases, the bagging prediction, average of many prediction values by part of the sample (half sample or 0.632 sample–bootstrap), is more accurate than a single prediction by using the full sample (Breiman 1996; Friedman & Hall 1999, Hastie *et al.* 2001).

In step 1 and step 2, we use two different kinds of cross-validation. We use LOOCV in step 1 due to its computational efficiency. In step 1, we search a large number of locus-sets. Thus, computational efficiency is important. Using LOOCV to calculate R_1^2 allows quick computing, in fact, as quick as computing R^2 based on the full sample. In step 2, we only compare the locus-sets retained in step 1. The number of retained locus-sets is not large. The computation efficiency is not a major concern. However, the dimensions of the retained locus-sets are quite different, and we need to choose the objective function very carefully. In this step, we proposed a measure to combine the predictability and reliability of the model. To measure the reliability of the model, we compare $R^2(1)$ and $R^2(2)$. In order to have an equal number of training to training prediction and training to test prediction for each individuals, we use a 2-fold cross-validation.

Step 3. Choose the best locus-set and make inference

In step 2, each retained locus-set has been assigned a value of the objective function. Larger values of the objective function mean higher predictability and more reliability. We choose the locus-set with the largest value of the objective function as the final locus-set. To assess the statistical significance of the test for association between the final locus-set and the trait, we perform a permutation test. Denote the value of the objective function given by equation (5) from original sample by T_0 . In each permutation, we randomly permute trait values of the individuals and repeat step 1 and step 2 on the randomized data, and calculate the value of the objective function given by equation (5). This permutation process was repeated 1000 times, and denote values of the objective function for the 1000 permuted samples by $T_1, \ldots, T_{1000}$. The estimated p-value of the test for association between the final locus-set and the trait is $\frac{\#\{i: T_i > T_0\}}{1000}$. This p-value is an overall p-value.

The CSM has been implemented using C programing language. The program will be available by Jan. 2006 through http://www.math.mtu.edu/∼shuzhang.

Data simulation

We first evaluate both the type I error and the power of the CSM through simulation studies. For evaluating type I error, we consider the test for association between the final locus-set and the trait. For power evaluation, we consider two kinds of powers: (1) the power to detect the correct set of functional SNPs; (2) the power of the test for association between the final locus-set and the trait. To calculate the power of the test for association (using significance level α), for each replication, we consider there is a significance association only if the p-value $\leq \alpha$ and the final locus-set contains at least one functional SNP. We consider both quantitative traits and qualitative traits. For each simulation scenario, we generate 100 replications of 400 individuals (200 cases and 200 controls for the qualitative trait) and consider ten independent SNPs. Unrelated individuals and their genotypes for the ten unlinked SNPs are generated by assuming Hardy-Weinberg equilibrium and linkage equilibrium.

Data sets for assessing the type I error: To assess the type I error rate of the test for association between the final locus-set and a trait, we generate marker alleles at each of the ten markers independently according to their allele frequencies. The allele frequency of one of the two alleles at each marker is drawn from a beta distribution $B(2, 2\frac{1-q}{q})$ with mean q and variance $\frac{1-q}{2+q}q^2$. We consider three cases: $q = 0.5$, $q = 0.25$ and $q = 0.1$. For a qualitative trait of interest, we randomly assign one individual as a case or a control independent of the genotypes. For a quantitative trait of interest, we generate the trait value of each individual using the model

$$y = \mu + \epsilon$$

where $\mu = 3$ and ϵ is a standard normal random variable.

Table 1 Multilocus penetrance functions and allele frequencies for six two-locus models exhibiting gene-gene interaction in absence of main effects

	Locus B			Minor allele
Model and Locus A	*BB*	*Bb*	*bb*	Frequencies
Model 1				
AA	0	0.1	0	
Aa	0.1	0	0.1	$q_A = q_B = 0.5$
aa	0	0.1	0	
Model 2				
AA	0	0	0.1	
Aa	0	0.05	0	$q_A = q_B = 0.5$
aa	0.1	0	0	
Model 3				
AA	0.08	0.07	0.05	
Aa	0.1	0	0.1	$q_A = q_B = 0.25$
aa	0.03	0.1	0.04	
Model 4				
AA	0	0.01	0.09	
Aa	0.04	0.01	0.08	$q_A = q_B = 0.25$
aa	0.07	0.09	0.03	
Model 5				
AA	0.07	0.05	0.02	
Aa	0.05	0.09	0.01	$q_A = q_B = 0.1$
aa	0.02	0.01	0.03	
Model 6				
AA	0.09	0.001	0.02	
Aa	0.08	0.07	0.005	$q_A = q_B = 0.1$
aa	0.003	0.007	0.02	

Note: Adopted from Ritchie *et al.* (2003).

Data sets for assessing the power: To assess both the power of the CSM to detect the correct locus-set and the power of the test for association between the final locus-set and a trait of interest, we consider two sets of simulations. In the first set of simulations, we consider two-locus epistasis models. Six two-locus epistasis models used in this article are given in Table 1. All six models, also described by Ritchie *et al.* (2003), exhibit interaction effects in the absence of any main effects when Hardy-Weinberg is assumed. To compare with the results of Ritchie *et al.* (2003), we use the same set-up as that in Ritchie *et al.* (2003), i.e. the minor allele frequencies are the same across the ten simulated SNPs and given in Table 1. Of the ten SNPs, there are two functional SNPs and eight nonfunctional SNPs. For a qualitative trait of interest, the genotypes at the nonfunctional SNPs can be generated according to the genotype frequencies under Hardy-Weinberg and linkage equilibrium, and the genotypes across the two functional SNPs can be generated according to the conditional probability of the genotypes given affected or unaffected status. When a trait of interest is quantitative, for each individual, we generate the genotypes across the ten SNPs under Hardy-Weinberg and linkage equilibrium, and generate the trait values using the model

$$y = \mu \cdot Pen(G) + \epsilon,$$

where y denotes the trait value; μ is a constant; $Pen(G)$, given in Table 1, is penetrance (for a case-control design) of this individual's genotype G at the two functional SNPs; ϵ is a standard normal random number. The value of μ can be determined by the value of the heritability. In the first set of simulations, we vary the heritability from 5% to 9%.

In the second set of simulations, we consider one three-locus, one four-locus and one five-locus epistasis model. Thus, of the total 10 simulated SNPs, 3, 4, or 5 are functional epistatic SNPs and there are up to seven nonfunctional SNPs. Allele frequencies for each of the 10 SNPs were selected to match those in the ACE gene samples (Zhu *et al.* 2001). The three-, four-, and five-locus models used in this article, similar to those described by Ritchie *et al.* (2001), are extensions of model 2 in Table 1. Model 2 in Table 1 is a well-characterized model for epistasis, in which the disease risk is dependent on whether two deleterious alleles and two normal alleles are present, from either one locus or both loci (Ritchie *et al.* 2001). Let A and a denote the two alleles at the first disease susceptibility locus, and B and b denote the two alleles at the second disease susceptibility locus. The high risk genotypes are *AAbb*, *AaBb* and *aaBB*. Similar to those of Ritchie *et al.* (2001), we extend this two-locus model to three-locus, four-locus and five-locus epistasis models by adding corresponding homozygous or heterozygous genotypes to the aforementioned high risk genotypes. For example, for the three-locus model, the high risk genotypes are *AAbbcc*, *AAbbcc*, *AAbbcc*, *AAbbcc*, *AabbCc* and *aabbCC*. For a qualitative trait of interest, we assume that all the high risk genotypes have the same penetrance, and all low risk genotypes also have the same penetrance. Let R denote the relative risk of the high risk genotypes to low risk genotypes. We fix the population prevalence to 5%, and vary R from 10 to 20. When a trait of interest is quantitative, for each individual we generate the trait value using the model

$$y = \alpha \cdot I(G) + \epsilon,$$

where y denotes the trait value; α is a constant; $I(G) = 1$ if this individual's genotype G is a high risk genotype, $I(G) = 0$ otherwise; ϵ is a standard normal random number. The constant α can be determined by the value of the heritability. In this set of simulations, we vary the heritability from 10% to 20%.

ACE Data

We apply the CSM to the data set collected by the International Collaborative Study on Hypertension in Blacks (Cooper *et al.* 1997; Zhu *et al.* 2001). The data includes 1,343 individuals in 332 families residing in western Africa. ACE concentration, SBP, DBP, sex, age and body mass index (BMI) are available for this data. 13 SNPs including one insertion/deletion were genotyped. Zhu *et al.* (2001) demonstrated that a two-locus (ACE4 and ACE8) additive model with an additive interaction explained most of the ACE variation. A two-locus (ACE4 and ACE8) epistasis model is also significantly associated with SBP and DBP. For the purpose of this study, we choose 361 unrelated individuals (178 males and 183 females) whose genotype and phenotype data are available.

Results

Simulation Results

We first verify that the CSM has the correct nominal type I error rate for testing for association between the final locus-set and the trait. For 100 replication samples, the standard deviation for the type I error rate are $\sqrt{0.05 \times 0.95/100} \approx 0.022$ and $\sqrt{0.1 \times 0.9/100} \approx 0.03$, for the nominal levels of 0.05 and 0.1, respectively. The 95% Quantile Ranges (QRs) are (0.006, 0.094) and (0.04, 0.16), respectively. The estimated type I errors of the test for different minor allele frequencies and different numbers of L, the pre-specified maximum number of loci which we search for, are summarized in Table 2. It is easy to see that the estimated type I errors of the test are not significantly different from the nominal levels. The pooled p-values for different minor allele frequencies and different values of L are presented in Figure 2. Figure 2 shows that the distribution of p-values of the test is very close to a uniform distribution, also suggesting that the test has the correct type I error rate.

The powers of the CSM to detect the two functional SNPs from the ten SNPs for each of the two-locus epistasis models and to test association between the final

Table 2 Type-I error rate of the test for association between the final locus-set and the trait. The type I error rates are based on 100 replications and 1000 permutations to evaluate the p-value. *L* is the pre-specified maximum number of loci which we search for, i.e. we search all the one- to *L*-locus sets

	Expected allele frequency	Significant level = 0.05					Significant level = 0.1				
		L=1	L = 2	L = 3	L = 4	L = 5	L = 1	L = 2	L = 3	L = 4	L = 5
Qualitative trait	0.1	0.06	0.08	0.05	0.06	0.05	0.15	0.10	0.09	0.11	0.14
	0.25	0.04	0.03	0.06	0.03	0.04	0.07	0.07	0.12	0.07	0.06
	0.5	0.03	0.03	0.03	0.03	0.07	0.12	0.10	0.10	0.07	0.11
Quantitative trait	0.1	0.08	0.04	0.08	0.05	0.05	0.10	0.09	0.13	0.08	0.15
	0.25	0.06	0.07	0.06	0.02	0.02	0.16	0.11	0.12	0.07	0.07
	0.5	0.04	0.04	0.04	0.03	0.07	0.07	0.07	0.09	0.08	0.12

COLOUR FIG.

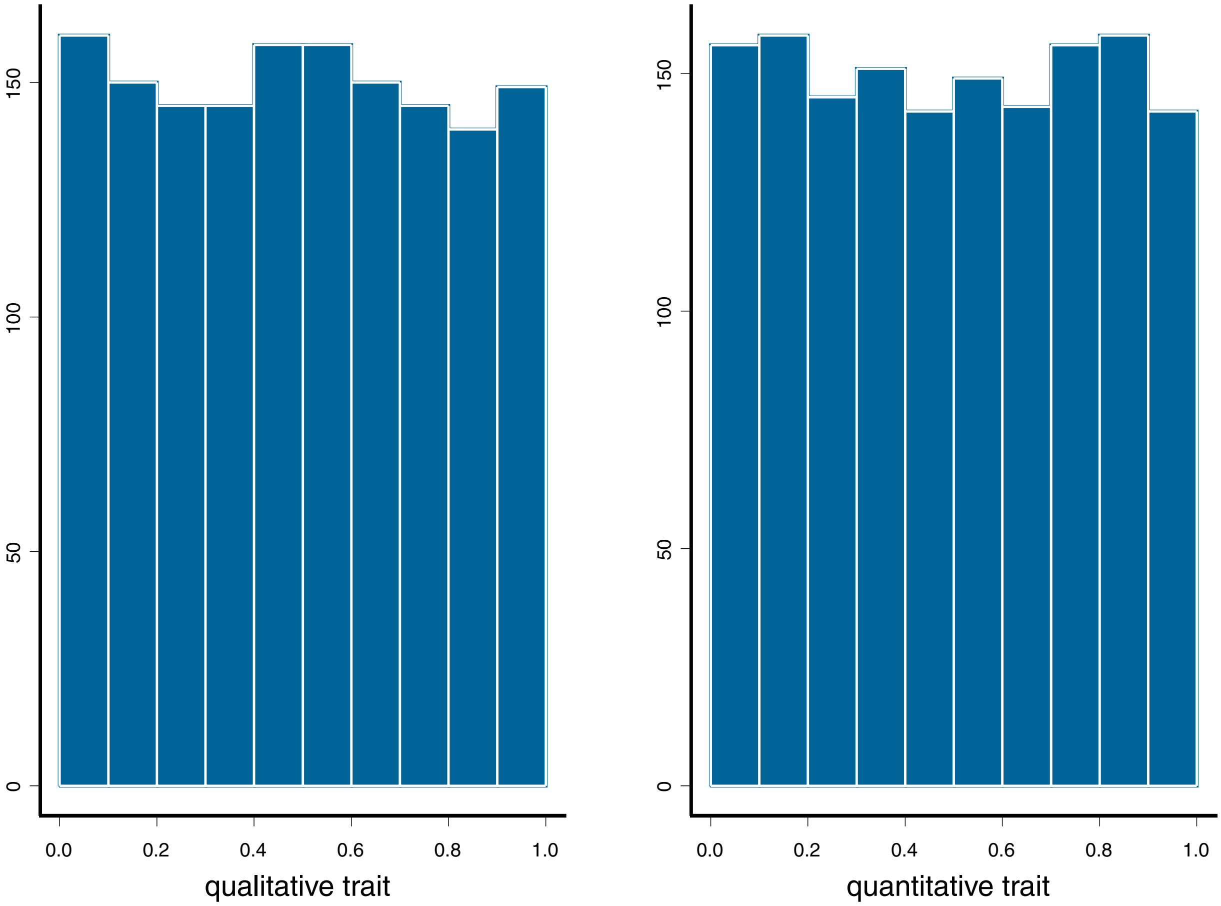

Figure 2 The histograms of the pooled p-values for the three different minor allele frequencies and $L = 1, 2, 3, 4, 5$, where *L* is the pre-specified maximum number of loci which we search for.

locus-set and the trait are given in Table 3. For the qualitative trait, the powers of the MDR (Ritchie *et al.* 2003) from Table II of Ritchie *et al.* (2003) are also given for the purpose of comparison. Table 3 shows that, for the qualitative trait, the powers of the CSM and MDR (Ritchie *et al.* 2001, 2003) are very similar. However, the CSM is slightly more powerful than the MDR under model 5 and 6. For both the qualitative trait and quantitative trait, the power of the test for association at 5% significance level is very similar to

Table 3 The power of the CSM to detect correct functional interacting loci and power of the test for association between the final locus-set and the trait under two-locus interaction models. The numbers in parenthesis are from Table II of Ritchie *et al.* (2003)

			Power to detect the	Power of the test (%)
		Model	Correct locus-set (%)	Significance level = 0.05
qualitative trait		1	100 (100)	100
		2	100 (100)	100
		3	99 (99)	99
		4	98 (99)	96
		5	87 (82)	88
		6	89 (84)	92
	heritability			
quantitative trait	0.05	1	76	75
		2	76	73
		3	70	68
		4	69	68
		5	68	65
		6	73	70
	0.07	1	98	93
		2	94	90
		3	89	86
		4	90	89
		5	89	85
		6	90	90
	0.09	1	99	100
		2	99	98
		3	97	95
		4	96	94
		5	90	92
		6	92	94

the power of detecting correct functional SNPs. For all the cases, more than 95% of the p-values for the correctly identified locus-set (functional SNPs) are less than 0.05 (results not shown). For the quantitative trait, we evaluate the power of the CSM for different value of heritability. Both the power of correctly detecting the functional SNPs and the power of the test for association increase as the heritability increases. However, the power of testing for association increases more rapidly than that of correctly detecting the functional SNPs. When the heritability is 5%, the power of testing for association at the 5% significance level is slightly less than that of correctly detecting the functional SNPs. When the heritability increases to 9%, the power of testing for association is almost the same as that of correctly detecting the functional SNPs.

The powers of the CSM under three-locus to five-locus epistasis models are summarized in Table 4. For the qualitative trait, we consider different values of relative risk. The power of correctly detecting the functional SNPs increases as the relative risk increases. However, the power of testing for association is 100% for all the cases considered here. The reason is that even if the final locus-set is not the set of functional SNPs, the final locus-set will include some or all of the functional SNPs and the association between the final locus-set and the trait is still strong enough. For the quantitative trait, we consider different heritabilities. The power of correctly detecting the functional SNPs and the power of the association test are very similar and both increase as the heritability increases, which is consistent with the results for the two-locus epistasis models. Comparing the performance of the CSM for the qualitative trait and for the quantitative trait, it seems that the CSM performs better for the quantitative trait than for the qualitative trait. Consider the five-locus model as an example. Table 4 shows that, for the qualitative trait, the power of the association test is 100% but the power of correctly detecting the functional SNPs is only 60% (relative risk = 10). This means that the CSM can only correctly detect the functional SNPs that are strongly associated with the qualitative trait. However, for the quantitative trait (heritability = 0.15), the power of correctly detecting the functional SNPs is 59% when the power of the association test is 61%. It seems that the CSM can correctly detect the functional SNPs that have a moderate

Qualitative trait Relative risk	Number of functional loci	Power of correct detection(%)	Power of the test (%) $\alpha = 0.05$
	3	88	100
10	4	80	100
	5	60	100
	3	94	100
15	4	90	100
	5	75	100
	3	99	100
20	4	97	100
	5	80	100
Quantitative trait heritability	3	79	74
0.1	4	53	50
	5	33	30
	3	98	95
0.15	4	76	78
	5	59	61
	3	100	100
0.2	4	89	92
	5	69	72

Table 4 The power of the CSM to detect correct functional interacting loci and power of the test for association between the final locus-set and the trait under three-, four- and five-locus interaction models

effect on the quantitative trait. To compare the performance of the objective function T with $R^2(2)$, the cross-validation correlation, Figure 3 summarizes the values of the objective function T and $R^2(2)$ for a quantitative trait. It is clear that the objective function performs better than $R^2(2)$ to distinguish the set of functional SNPs from the other locus-sets.

For both the qualitative trait and the quantitative trait, when the relative risk or heritability is fixed, the power to identify the correct functional SNPs tends to decrease when the order of interaction is higher. This phenomenon is consistent in many cases including different relative risks (qualitative trait) and different values of heritability (quantitative trait). We believe that this is a real phenomenon. It is surprising that, using a similar model but with an infinite relative risk (penetrance of high risk genotype is 20% while penetrances of other genotypes are 0), Ritchie *et al.* (2001)'s results showed that the power to identify the correct functional SNPs tends to increase as higher-order interactions are modeled. We have applied the CSM to the case of infinite relative risk and got the same results as that of Ritchie *et al.* (2001), i.e. the power to identify the correct functional SNPs tends to increase as the order of interactions increases (results not shown). We believe that this is a special case. Why the case of infinite relative risk is special will require further investigation.

Analysis of ACE data

To apply the CSM to the ACE data, we considered three traits: ACE, SBP and DBP. We first adjusted ACE level, SBP or DBP by age, sex and BMI as covariates. We used the residuals as the adjusted trait values. We applied the CSM to the data to search for the locus-set that best predicts the adjusted trait values for each of the three traits. For ACE, using all 361 individuals, one one-locus model (polymorphism ACE8) had the largest value of the objective function and the p-value of the test for association between polymorphism ACE8 and ACE level was 0 based on 1000 permutations (Table 5). Figure 4(a) summarizes the mean trait values of the three genotypes AA, AG, and GG at the ACE8 locus. The three genotypes were divided into two groups: AA in one group and AG and GG in another group, i.e. the final model was a recessive disease model.

For SBP and DBP, there was no statistically significant evidence of an independent main effect of any of the 13 polymorphisms whenever the entire sample, males only or females only was used. When the CSM was applied to the entire sample and female only, the

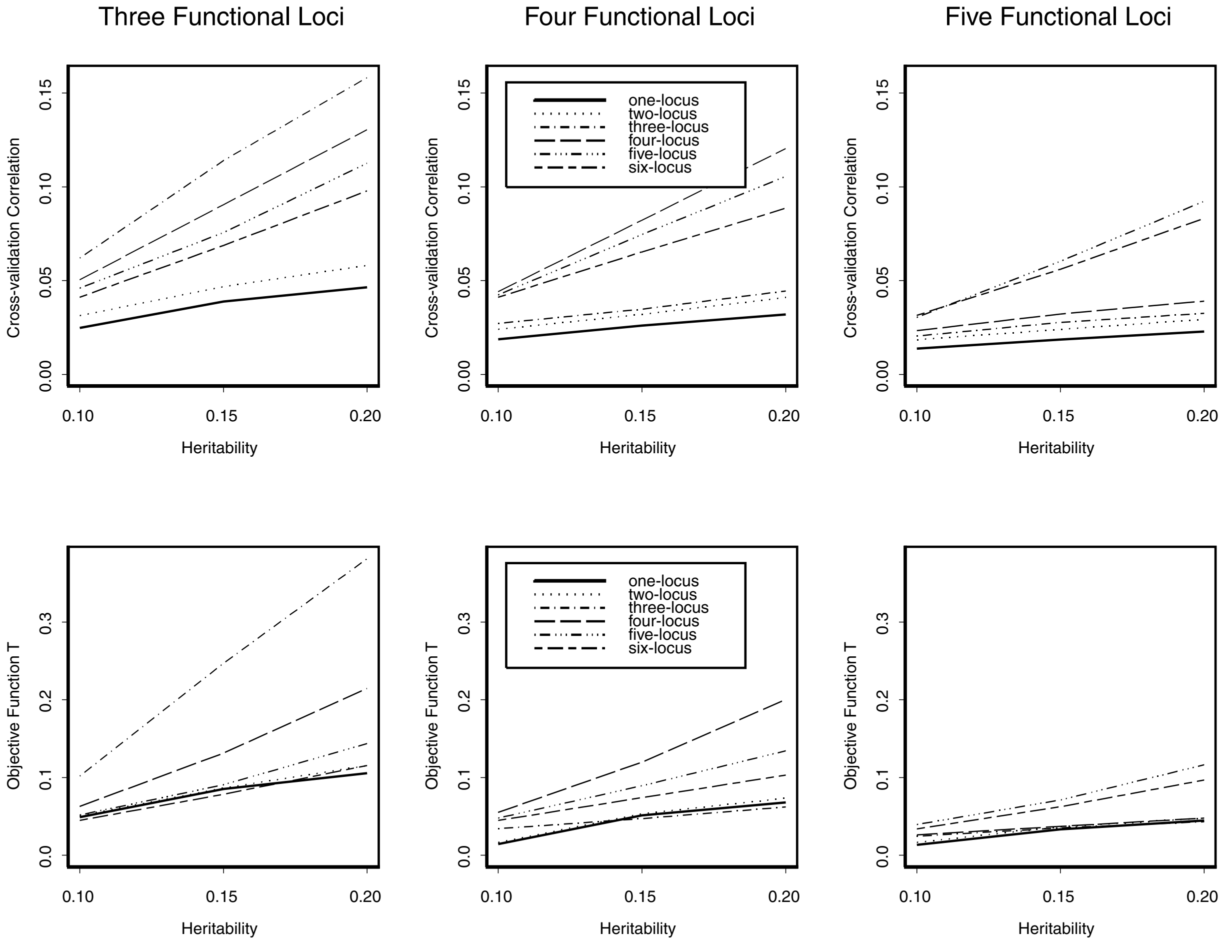

Figure 3 Comparisons of the cross-validation coefficient $R^2(2)$ and objective function T.

Table 5 The values of the objective function of the "best" locus-set for different number of loci and the p-value of the test for association between the "best" locus-set and the trait. The multilocus model with maximum value of the objective function is indicated in boldface type for each of the three traits

	Value of T of the "best" locus-set for each number of loci		
Number of loci	ACE	SBP (male)	DBP (male)
1	**0.45**	0.050	0.040
2	0.40	0.048	**0.069**
3	0.25	0.091	0.057
4	0.23	**0.110**	0.018
5	0.12	0.071	0.012
6	0.12	0.040	0.010
p-value of the "best" locus-set	<0.001	0.033	0.045

p-value was larger than 15% both for SBP and DBP. When the CSM was applied to the sample of males only, one two-locus model with a p-value = 0.045 has the largest value of the objective function for DBP, and one four-locus model with a p-value = 0.033 has the largest value of the objective function for SBP. The two-locus model for DBP included the polymorphisms ACE1 and ACE8. The mean trait values of each of the two-locus genotypes and the groups of the genotypes are summarized in Figure 4(b). Figure 4(b) shows the epistasis or gene-gene interaction between ACE1 and ACE8; that is, the influence of the genotypes on the trait at one locus is dependent on the genotypes at another locus. The four locus model for SBP included the polymorphisms ACE1, ACE4, ACEs2.1 and ACE8. The mean trait values for each of the four-locus genotypes also show the

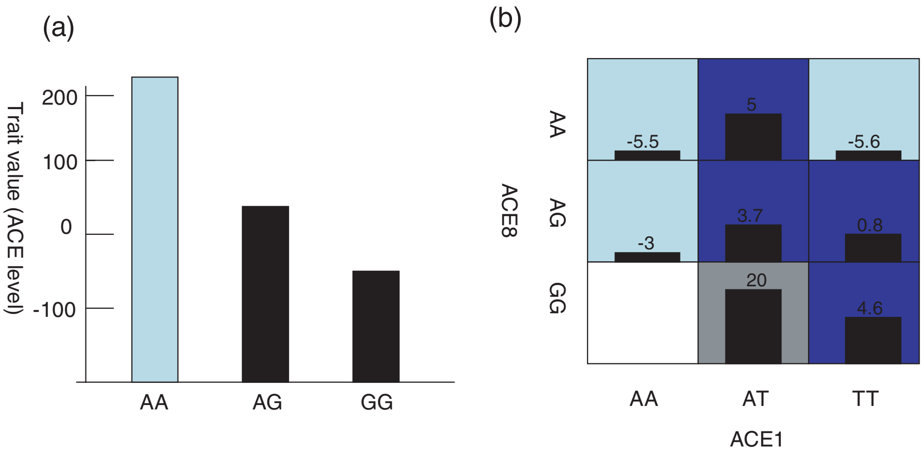

Figure 4 Summary of the trait values (trait values after adjusted for age, sex, BMI, and weight) of the genotypes. (a) Average trait (ACE) values corresponding to the three genotypes at ACE1 locus. The two bars with the same color are clustered in one group. (b) Average trait (SBP) values corresponding to the nine genotypes of the two-locus genotypes (ACE1 locus and ACE8 locus). The boxes with the same color are in one group. The box without trait-value bar indicates no individual with this genotype.

gene-gene interaction among the four loci (results not shown).

Discussion

The CSM is proposed to identify a group of interaction loci that has both high predictability and high reliability. The method evaluates sets of SNP markers at various positions in the genome (in one candidate gene or in different candidate genes) while keeping the overall type I error in control. The development of the method was motivated by the success of data-reduction methods by genotype partitioning for quantitative traits (Nelson *et al.* 2001) and dimension-reduction methods by dividing the genotypes into high-risk and low-risk groups for qualitative traits (Ritchie *et al.* 2001, 2003). The primary advantage of the CSM is that it can detect and characterize multiple interacting loci affecting either a qualitative trait or a quantitative trait. Using simulation studies, we have shown that, under some particular epistasis models, the CSM has reasonable power to identify the locus-set with high-order gene-gene interaction for both quantitative traits and qualitative traits. We applied the CSM to the ACE data set to identify the locus-set associated with ACE concentration level, the locus-set associated with SBP, and the locus-set associated with DBP.

It has been suggested that there are several polymorphisms with the ACE gene additively contributing to the variation of ACE level (Zhu *et al.* 2001; Bouzekri *et al.* 2003), and ACE8 has a much stronger effect than other markers. When we applied the CSM to the ACE data, we only detected a one-locus model involving ACE8 associated with ACE level. The reason is that the CSM is more powerful to detect the interaction of multiple loci. When several polymorphisms additively affect the trait and one has a much stronger effect than other markers, the CSM may only pick the polymorphism most strongly affecting the trait, which is consistent with the idea that ACE8 is the strongest polymorphism affecting ACE level (Zhu *et al.* 2001; Bouzekri *et al.* 2003). For SBP and DBP, the CSM detects a two-locus and a four-locus models in males, respectively. Both models include ACE8, which is also present in the epistasis model identified by Zhu *et al.* (2001). The slightly different models identified between this study and Zhu *et al.* (2001) could be potentially due to the strong linkage disequilibrium between the SNPs in the 3'end of the ACE genes (Bouzekri *et al.* 2003) and the different sample sizes used. In fact, Zhu *et al.* (2001) used the entire sample and this study only used a subset of the sample. Besides, Zhu *et al.* (2001) only searched the interaction model for SBP and DBP based on the model obtained from ACE levels, which is rather limited. The CSM is more ambitious because it exhaustively searches all the possible interaction models. Surprisingly, the CSM can still identify interaction models for both SBP and DBP when the sample size is relatively small.

In the implementation of the CSM, one question is that the multilocus genotype g of a specific individual in the test group may not appear in the training group. In this case, we use $\overline{y}$, the sample mean of the training group, as the predicted trait value of this individual. As an alternative method, we may choose a genotype g^* in the training group that is most similar to g, and use the trait value of g^* as the predicted trait value of this individual. In this article, we use a linear model to model qualitative traits. An alternative method for qualitative traits is using a logistic model. Under a logistic model, if we use the estimated conditional probability of an individual being affected given his/her genotype as the predicted trait value, the method described in this article can be applied to the logistic model directly. However, the performance of the method based on a logistic model needs further investigation.

Culverhouse *et al.* (2004) also proposed a promising Restrict Partition Method (RPM) to limit the "good" partitions. The RPM uses testing methods to merge genotypes (if trait means are not significantly different) which greatly reduced the computation effort comparing to the CPM by exhaustive search. The "good" partitions found by the RPM and by K-mean clustering are very similar. However, in applying the RPM, each locus-set needs several tests and each test needs a permutation procedure to evaluate the p-value which makes the RPM still computational demanding if we search a large number of locus-sets. Furthermore, in applying the RPM, the overall p-value of the final locus-set is the Bonferroni correction of the individual p-value. The Bonferroni correction may be very conservative due to the correlation between locus-sets, and a large number of permutations is required in order to get a small (0.05 for example) overall p-value.

Considering the complex and unclear nature of gene-gene interaction, like most of other methods to detect a set of loci with possible interaction (Nelson *et al.* 2001; Ritchie *et al.* 2001, 2003; Culverhouse *et al.* 2004), the CSM searches all possible combinations of the multilocus genotypes without considering their partial order. For example, one marker with the genotypes AA, Aa, and aa, we may combine AA and aa as one group and Aa as the other group. Considering all possible genotype combinations allows us to find all possible interactions. In other hand, if the genotypes really have the partial order, we may lose power by searching a larger number of recombinations. Furthermore, without considering the partial order of the genotypes may make the CSM models difficult to interpret. This is illustrated clearly in the two-locus model in Figure 4. There are no obvious trends or patterns in the distribution of high-risk and low-risk groupings across the two-dimensional genotype space. Interpretation of multi-locus models with interactions is always a difficult task due to the complex nature of gene-gene interactions plus the different meanings of interaction in statistics and genetics literature(Cordell *et al.* 2001; Moore & Williams 2005). Another shortcoming of the CSM is that we are not readily clear whether the final model contains interaction effects or only main effects. Also, using the partitioning method, the CSM is hard to model the additive relation and thus there would be a decrease in the CSM power if several loci have additive effects. It is clearly an important topic by using genotype partitions and also incorporating additive effect in future studies.

Some other unsolved questions need to be addressed. One is that genotyping errors have deleterious effects on association analysis (Akey *et al.* 2001) and thus will affect the CSM method, especially for the case that the genotyping error depends on the trait value. The easiest solution to the error problem is increasing quality control in the laboratory. Another avenue to be explored is incorporating error frequencies in the analysis model as it did for a special disequilibrium test (Gordon *et al.* 2001). Finally, population stratification is a problem in any population-based association study, our method being no exception. If the sampled individuals have different ethnic backgrounds and different ethnic backgrounds are associated with different trait values and different allele frequencies, this will affect the CSM. If there is evidence that there may be different ethnic backgrounds in the sample, our recommendation is to use methods such as the Genomic Control Method, which requires genotyping a set of unlinked SNPs across the genome (Devlin & Roeder 1999; Pritchard *et al.* 2000; Bacanu *et al.* 2000; Zhang & Zhao 2001; Zhang *et al.* 2003b; Chen *et al.* 2003; Zhu *et al.* 2002).

Acknowledgments

We thank the reviewer for their constructive comments. This work was supported by National Institute of Health (NIH) grants R01 GM069940, R03 HG 003613, R01 HG003054, R03 AG024491, and NSF grant 0421756.

Appendix: The K-mean Clustering Method

For a given locus-set, let $g_1, \ldots, g_{m+1}$ denote all the distinct multilocus genotypes observed in the sample. Let n_i and $\overline{y}_i$ denote the number of individuals and the average trait value, respectively, of the individuals who have genotype g_i.

To use the K-mean clustering method to divide the $m + 1$ genotypes into k groups, we first give the initial values of the centers for the k groups and denote the initial value of the center of group l by $C_l (l = 1, \ldots, k)$, then the process involves the following steps:

1. For each genotype $g_i (i = 1, \ldots, m + 1)$, calculate the distance between $\overline{y}_i$ and the center of each group. Assign g_i to group j if C_j is the nearest center among the k centers, that is, $|\overline{y}_i - C_j| = \min_{1 \le l \le k} |\overline{y}_i - C_l|$. In this way, we partition the $m + 1$ genotypes into k groups. Denote the k groups by $G_1, \ldots, G_k$.
2. Update the center for each of the groups. We use the average trait values of all individuals whose genotype belongs to G_l as the center of group $G_l (l = 1 \ldots .k)$, that is, $C_l = \frac{1}{m_l} \sum_{\{i: g_i \in G_l\}} n_i \overline{y}_i$, where m_l is the number of individuals whose genotype belongs to G_l.
3. Repeat step 1 and step 2 until no more reassignments take place.

Let $a = \min_{1 \le i \le m+1} \overline{y}_i$ and $b = \max_{1 \le i \le m+1} \overline{y}_i$. When we implement this method, we choose the initial values of the k centers as

$$C_l = \frac{l - 0.5}{k}(b - a), \; l = 1, 2, \ldots, k.$$

References

Akey, J. M., Zhang, K., Xiong, M., Doris, P. & Jin, L. (2001) The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. *Am J Hum Genet* **68**, 1447–1456.

Bacanu, S. A., Devlin, B. & Roeder, K. (2000) The power of genomic control. *Am J Hum Genet* **66**, 1933–1944.

Bouzekri, N., Zhu, X., Jiang, Y., McKenzie, C. A., Luke, A., Forrester, R., Adeyemo, A., Kan, D., Farrall, M., Anderson, S., Cooper, R. S. & Ward, R. (2004) Angiotensin I-converting enzyme polymorphisms, ACE level and blood pressure among Nigerians, Jamaicans and African-Americans, *Eur J Hum Genet* **12**, 460–468.

Breiman, L. (1996) Bagging predictor. *Machine Learning* **26**, 123–140.

Carrasquillo, M. M., McCallion, A. S., Puffenberger, E. G., Kashuk, C. S., Nouri, N. & Chakravarti, A. (2002) Genome-wide association study and mouse model identify interaction between RETand EDNRB pathways in Hirschsprung disease. *Nat Genet* **32**, 237–44.

Chen, H. S., Zhu, X., Zhao, H. & Zhang, S. L. (2003) Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. *Ann. Hum Genet* **67**, 250–264.

Coffey, C. S., Hebert, P. R., Krumholz, H. M., Morgan, T. M., Williams, S. M. & Moore, J. H. (2004a) Reporting of model validation procedures in human studies of genetic interactions. *Nutrition* **20**(1), 69–73.

Coffey, C. S., Hebert, P. R., Ritchie, M. D., Krumholz, H. M., Gaziano, J. M., Ridker, P. M., Brown, N. J., Vaughan, D. E. & Moore, J. H. (2004b) An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene-gene interactions on risk of myocardial infarction: the importance of model validation. *BMC Bioinformatics* **5**, 49.

Cooper, R. S., Rotimi, C., Ataman, S., McGee, D., Osotimehin, B., Kadiri, S., Muna, W., Kingue, S., Fraser, H., Forrester, T., Bennett, F. & Wilks, R. (1997) Hypertension prevalence in seven populations of African origin. *Am J Public Health* **87**, 160–168.

Cordell, H. J. (2002) Epistatsis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet* **11**, 2463–2468.

Cordell, H. J., Todd, J. A., Hill, N. J., Lord, C. J., Lyons, P. A., Peterson, L. B., Wicker, L. S. & Clayton, D. G. (2001) Statistical modeling of interlocus interactions in a complex disease: rejection of the multiplicative model of epistasis in type 1 diabetes. *Genetics* **158**, 357–367.

Cox, N. J., Frigge, M., Nicolae, D. L., Concannon, P., Hanis, C. L., Bell, G. I. & Kong, A. (1999) Loci on chromosome 2 (NIDDM1) and 15 interact to disease susceptibility to diabetes in Mexican American. *Nat Genet* **21**, 213–215.

Culverhouse, R., Klein, T. & Shannon, W. (2004) Detecting epistatic interactions contributing to quantitative traits. *Genet Epidemiol* **27**, 141–152

Culverhouse, R., Suarez, B. K., Lin, J. & Reich, T. (2002) A perspective on epistasis: limits of models displaying no main effect. *Am J Hum Genet* **70**, 461–71.

Devlin, B. & Roeder, K. (1999) Genomic control for association studies. *Biometrics* **55**, 997–1004.

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

Friedman, J. H. & Hall, P. (1999) On bagging and nonlinear estimation. Stanford University, Department of Statistics. Technical Report.

Gordon, D., Heath, S. C., Liu, X. & Ott, J. (2001) A transmission disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. *Am J Hum Genet* **69**, 371–380.

Goutte, C. (1997) Note on free lunches and cross-validation. *Neural Computation* **9**(6), 1245–1249.

Hastie, T., Tibshirani, R. & Friedman, J. (2001) The elements of statistical learning; data mining, inference, and prediction. Springer Verlag, New York.

Hoh, J. & Ott, J. (2003) Mathematical multi-locus approachs to loculizing complex human trait genes. *Nature Reviews Genetics* **4**, 701–709.

Hoh, J., Wille, A. & Ott, J. (2001) Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Res* **11**, 2115–2119.

Johnson, R. A. & Wichern, D. W. (1998) Applied multivariate statistical analysis, Prentice Hall, New Jersey.

Moore, J. H. (2003) The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered* **56**, 73–82.

Moore, J. H. (2004) Computational analysis of gene-gene interactions using multifactor dimensionality reduction. *Expert Review of Molecular Diagnostics* **4**(6), 795–803.

Moore, J. H. & Williams, S. M. (2002) New strategies for identifying genegene interactions in hypertension. *Ann Med* **34**, 88–95.

Moore, J. H. & Williams, S. M. (2005) Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *BioEssays* **27**(6), 637–646.

Nelson, M. R., Kardia, S. L., Ferrell, R. E. & Sing, C. F. (2001) A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res* **11**, 458–470.

Nicolae, D. L. & Cox, N. J. (2002) MERLIN...and the geneticist's stone?, *Nat Genet* **30**, 3–4.

Olson, J. M., Goddard, K. A. & Dudek, D. M. (2002) A second locus for very-late-onset Alzheimer disease: a genome scan reveals linkage to 20p and epistasis between 20p and the amyloid precursor protein region. *Am J Hum Genet* **71,** 154–61.

Pritchard, J. K., Stephens, M., Rosenberg, N. A. & Donnelly, P. (2000) Association mapping in structured population, *Am J Hum Genet* **67,** 170–181.

Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M. & Poland, G. A. (2002) Score test for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* **70**, 425–434.

Risch, N. J. (2000) Search for genetic determinations in the new millenium. *Nature* **405**, 847–856.

Risch, N., Spiker, D., Lotspeich, L., Nouri, N., Hinds, D., Hallmayer, J., Kalaydjieva, L., McCague, P., Dimiceli, S., Pitts, T., Nguyen, L., Yang, J., Harper, C., Thorpe, D., Vermeer, S., Young, H., Hebert, J., Lin, A., Ferguson, J., Chiotti, C., Wiese-Slater, S., Rogers, T., Salmon, B., Nicholas, P. & Myers, R. M. (1999) A genomic screen of autism: evidence for a multilocus etiology. *Am J Hum Genet* **65**, 493–507.

Ritchie, M. D., Hahn, L. W. & Moore, J. H. (2003) Power of Multifactor-Dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet Epidemiol* **24**, 150–157.

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F. & Moore, J. H. (2001) Multifactor-Dimensionality reduction reveals high-order interactions among Estrogen-Metabolism genes in sporadic brest cancer. *Am J Hum Genet* **69**, 138–147.

Stone, M. (1977) An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. J R Stat Soc B **38**, 44–47.

Templeton, A. R. (2000) Epistasis and complex trait. In Epistasis and the evolutionary process (eds.Wade, M., Brodie, B. III & Wolf, J.). Oxford University Press, Oxford.

Thornton-Wells, T. A., Moore, J. H. & Haines, J. L. (2004) Genetics, statistics and human disease: analytical retooling for complexity?, *Trends in Genetics* **20**(12), 640–647

Tiwari, H. & Elston, R. C. (1998) Restrictions on components of variance for epistatic models. *Theor Popul Biol* **54**, 161–174.

Wilson, S. R. (2001) Epistasis and its possible effects on transmission disequilibrium tests. *Ann Hum Genet* **62**, 565–575.

Zhang, S., Sha, Q., Chen, H. S., Dong, J. & Jiang, R. (2003a) Transmission/disequilibrium test based on haplotype sharing for tightly linked markers. *Am J Hum Genet* **73**, 566–579.

Zhang, S. L. & Zhao, H. (2001) Quantitative similarity-based association tests using population samples. *Am J Hum Genet* **69**, 601–614.

Zhang, S. L., Zhu, X. & Zhao, H. (2003b) On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. *Genetic Epidemiol* **24**, 45–56.

Zhao, H., Zhang, S. & Merikangas, K. R. *et al.* 2000) Transmission/disequilibrium tests using multiple tightly linked markers. *Am J Hum Genet* **67**(4), 936–946

Zhu, X., Zhang, S. L., Zhao, H. & Cooper, R. S. (2002) Association mapping, using a mixture model for complex traits. *Genetic Epidemiol* **23**, 181–196.

Zhu, X., Bouzekri, N., Southam, L., Cooper, R. S., Adeyemo, A., McKenzie, C. A., Luke, A., Chen, G., Elston, R. C. & Ward, R. (2001) Linkage and association analysis of angitensin I-converting enzyme (ACE)-gene polymorphisms with ACE concentration and bllod pressure, *Am J Hum Genet* **68**, 1139–1148.

Received: 19 January 2006
Accepted: 19 January 2006

MARKED PROOF

Please correct and return this set

Please use the proof correction marks shown below for all alterations and corrections. If you wish to return your proof by fax you should ensure that all amendments are written clearly in dark ink and are made well within the page margins.

Instruction to printer	*Textual mark*	*Marginal mark*
Leave unchanged	• • • under matter to remain	Stet
Insert in text the matter indicated in the margin	⋏	New matter followed by ⋏
Delete	├─┤ through matter to be deleted	₰
Delete and close up	⊖ through matter to be deleted	₰
Substitute character or substitute part of one or more word(s)	/ through letter or ├─┤ through word	New letter or new word
Change to italics	— under matter to be changed	⊔⊔
Change to capitals	≡ under matter to be changed	≡
Change to small capitals	= under matter to be changed	=
Change to bold type	〰 under matter to be changed	〰
Change to bold italic	≂ under matter to be changed	⊔⊔ 〰
Change to lower case	Encircle matter to be changed	≢
Change italic to upright type	(As above)	⊔̸
Insert 'superior' character	/ through character or ⋏ where required	⋎ under character e.g. $\overset{3}{\curlyvee}$
Insert 'inferior' character	(As above)	⋏ over character e.g. $\underset{2}{\curlywedge}$
Insert full stop	(As above)	⊙
Insert comma	(As above)	,
Insert single quotation marks	(As above)	$\overset{‘}{\curlyvee}$ and/or $\overset{’}{\curlyvee}$
Insert double quotation marks	(As above)	$\overset{“}{\curlyvee}$ and/or $\overset{”}{\curlyvee}$
Insert hyphen	(As above)	(hy)
Start new paragraph	_⌐	_⌐
No new paragraph	⌒	⌒
Transpose	⊔⊓	⊔⊓
Close up	linking ⁐ letters	⁐
Insert space between letters	⋏ between letters affected	#
Insert space between words	⋏ between words affected	#
Reduce space between letters	⊤ between letters affected	⊤
Reduce space between words	⊤ between words affected	⊤