Genomics 97 (2011) 77–85

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

On the use of multifactor dimensionality reduction (MDR) and classification and regression tree (CART) to identify haplotype–haplotype interactions in genetic studies

Ai-Ru Hsieh [a], Ching-Lin Hsiao [b], Su-Wei Chang [b], Hui-Min Wang [a], Cathy S.J. Fann [a,b,*]

[a] Institute of Public Health, Yang-Ming University, Taipei, Taiwan
[b] Institute of BioMedical Science, Academia Sinica, Nankang, Taipei, Taiwan

ARTICLE INFO

Article history:
Received 18 August 2010
Accepted 14 November 2010
Available online 24 November 2010

Keywords:
Gene–gene interaction
Haplotype
MDR
CART

ABSTRACT

Haplotype-based approaches may have greater power than single-locus analyses when the SNPs are in strong linkage disequilibrium with the risk locus. To overcome potential complexities owing to large numbers of haplotypes in genetic studies, we evaluated two data mining approaches, multifactor dimensionality reduction (MDR) and classification and regression tree (CART), with the concept of haplotypes considering their haplotype uncertainty to detect haplotype–haplotype (HH) interactions. In evaluation of performance for detecting HH interactions, MDR had higher power than CART, but MDR gave a slightly higher type I error. Additionally, we performed an HH interaction analysis with a publicly available dataset of Parkinson's disease and confirmed previous findings that the *RET* proto-oncogene is associated with the disease. In this study, we showed that using HH interaction analysis is possible to assist researchers in gaining more insight into identifying genetic risk factors for complex diseases.

1. Background

Many common diseases, such as coronary heart disease, hypertension, diabetes, and Parkinson's disease, appear to arise from the combined effects of multiple genes, environmental factors, and risk-conferring behaviors [1]. A central goal in human genetics is to understand the relationship between DNA sequence variations and the susceptibility to diseases. Success in this effort will depend critically on the degree of nonlinearity in the relationship between genotypes and phenotypes. Nonlinearities can arise from phenomena such as locus heterogeneity, phenocopy, dependence of genotypic effects on environmental factors (i.e., gene–environment interactions), and genotypes at other loci (i.e., gene–gene interactions or epistasis).

Epistasis is a fundamental component of the genetic architecture of complex traits. Although epistasis is believed to be key to common disorders, traditional statistical methods are generally less practicable for handling high-dimensional data and are not suitable to detect interactions in the absence of independent main effects (i.e., effects that may be detected by a single marker test). This limitation has led to the development and application of different data mining approaches, such as combinatorial-based and tree-based approaches, for handling high-dimensional data.

Multifactor dimensionality reduction (MDR) [3] is one combinatorial-based data mining approach. MDR uses an attribute construction algorithm to create a new discrete attribute by pooling levels from multiple discrete factors [4,5]. This process changes the representation space of the data and therefore makes nonlinear interactions easier to be detected and characterized. The MDR approach has been successfully applied to detect gene–gene interactions in a variety of human diseases including Alzheimer's disease [6], bladder cancer [7], multiple sclerosis [8], and schizophrenia [9]. The Classification and Regression Tree (CART) method [10] is a tree-based data mining approach. Depending on a binary answer to a question, CART divides each node of the tree into two offspring nodes, until the observations at the same node are homogeneous. The goal of CART is to produce an accurate set of data classifiers by uncovering the predictive structure of the problems under consideration.

Both MDR and CART have been used to identify combinations of multi-locus genotypes and discrete environmental factors associated with diseases [11,12]. According to Niu et al. [2], although methods based on single-nucleotide polymorphisms (SNPs) may yield important insights, haplotype-based methods can provide additional statistical power to detect genes involved in complex trait diseases and information on factors that influence the dependency among genetic markers. There are many methods that focus on single-locus or haplotype or SNP–SNP interaction analyses; however, detection of haplotype–haplotype (HH) interactions should be considered as one of the important data mining approaches for genetic association studies.

* Corresponding author. Institute of Biomedical Sciences Academia Sinica, Office 101, 128, Academia Road, Section 2, Nankang, Taipei 115, Taiwan. Fax: +886 2 27823047.
E-mail address: csjfann@ibms.sinica.edu.tw (C.S.J. Fann).

0888-7543/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2010.11.003

To our knowledge, not many papers have discussed HH interactions [32,34–36]. In one such study, Chen et al. [32,34–36] developed a random forests-based approach, called "hapForest," to identify disease-related haplotypes. The authors used "variable importance" as the score to identify significant patterns of haplotypes in association with disease susceptibilities in the forests. In our study, haplotype data were inferred from Haploview [19] and PHASE software [20]. "Index scores" were used to adjust bias arising from haplotype uncertainty. These index scores incorporate all available locus information and give different weights to each individual's possible haplotypes. With haplotype data and index scores, MDR and CART were then used to detect HH interactions.

Although haplotypes typically exist in all human chromosomes, the number and size vary with chromosome and population [13,14]. We are therefore particularly interested in three factors that might influence power and type I error of the two methods: (i) different disease models; (ii) variations in haplotype block size (i.e., the number of markers in each haplotype); and (iii) different haplotype frequencies. By evaluating results from simulations, we address the advantages and disadvantages of both methods as applied to HH interactions. Finally, we perform an HH interaction analysis using the two methods with a publicly available dataset of Parkinson's disease [15], and we compare the results with some current knowledge of the disease.

2. Methods

2.1. Partitioning haplotype blocks

Suppose there are N individuals. Let $G_n=(G_n^1,G_n^2,...,G_n^q)$ denote q diploid genotypes for individual n ($n=1,2,\ldots,N$), where q is the number of designated biallelic SNPs. In general, an individual's haplotype information is not directly observed by using current genotyping techniques. In this study, haplotype blocks are inferred based on the definition of Gabriel et al. [14] for all diploid genotypes, using the Haploview software [19], to compute the 95% confidence bounds of the value of D', which is a standard measurement of linkage disequilibrium (LD), for all pairwise combinations of genotypes. According to Gabriel et al. [14], a pair of genotypes is defined to be in "strong LD" if the upper 95% confidence bound of D' is higher than 0.98 and the lower bound is higher than 0.7.

2.2. Inferring probabilities for all possible haplotype pairs

Suppose there are B blocks determined by Haploview v4.1 software. There are M_i possible haplotype pairs for block i, $i=1, 2, \ldots, B$. Note that we consider biallelic SNPs in this study, so the number of M_i is $C_2^{2^s}+2^s$, where s is the number of SNPs in block i. Let $H_{i(j)}$, $j=1, 2, \ldots, M_i$, denote the jth possible haplotype pair in block i, let $p_n^{H_{i(j)}}$ denote the corresponding haplotype probabilities to $H_{i(j)}$ for individual n ($n=1,2, N$), let $P(H_{i(j)})=(p_1^{H_{i(j)}},p_2^{H_{i(j)}},...,p_N^{H_{i(j)}})$ denote the collection of corresponding haplotype probabilities to $H_{i(j)}$ for N individuals, and let $P_n(H_i\bullet)=(p_n^{H_{i(1)}},p_n^{H_{i(2)}},...,p_n^{H_{i(Mi)}})$ denote the collection of corresponding haplotype probabilities to M_i possible haplotype pairs in block i for individual n ($n=1,2,\ldots,N$). Note that $p_n^{H_{i(j)}}$ is inferred by using PHASE v2.1 software [20]. Furthermore, we consider that heterogeneity may cause the bias of inferring haplotype probabilities. Hence, we carry out the haplotype reconstruction process in PHASE separately for the case and control groups in this study.

2.3. Calculating the "index scores" to consider haplotype uncertainty

Haplotype uncertainty is a consequence of inferring haplotype probabilities. Therefore, we calculate index scores to account for the uncertainty [37] of the haplotype reconstruction process in PHASE. The following describes how "index scores" adjust for bias arising from haplotype uncertainty.

A function S of $H_{i(j)}$ is defined as the index score of $H_{i(j)}$ as follows:

$$S\left(H_{i(j)}\right)=\sum_{n=1}^{N} p_n^{H_{i(j)}},\ \text{where}\begin{cases}\sum_{n=1}^{n_1} p_n^{H_{i(j)}}\ \text{for cases}\\ \sum_{n=n_1+1}^{N} p_n^{H_{i(j)}}\ \text{for controls}\end{cases}.$$

This is a summation over all corresponding probabilities of $H_{i(j)}$ for N (that is, $n=1,2,...,n_1$ for cases and $n=n_1+1,n_1+2,.....,N$ for controls) individuals.

The basic idea of "index scores" is that different haplotype pairs of an individual contribute differently to the disease phenotype. A haplotype pair for an individual is weighted by the probability inferred from PHASE software. Therefore, the index score evaluates the contribution that a given haplotype pair makes to an individual's disease phenotype.

2.4. Reducing the amount of data in each haplotype block

Gabriel et al. [14] found that most blocks contained only three to five haplotypes, i.e., most $p_n^{H_{i(j)}}$ are small (e.g., <0.1) or equal to zero. Let $H^*_{i(j)}$, $j=1, 2, \ldots, M_i$ denote the haplotype pair with the j^{th} largest index score $S(H^*_{i(j)})$. We rank all of the index scores of $H_{i(j)}$ for N individuals, i.e., $S(H_{i(1)})$, $S(H_{i(2)})$, …, $S(H_{i(M_i)})$, from the largest to the smallest within each block i, i.e., $S(H^*_{i(1)})>S(H^*_{i(2)})>\cdots>S(H^*_{i(M_i)})$. Hence, a suitable number of groups of haplotype pairs for each block can be determined by combining some haplotype pairs in a group in block i if their corresponding index scores are small (e.g., <0.1).

Suppose the number of groups is g, where $g\le M_i$ (M_i is the number of possible haplotype pairs in block i. Let $K_{i(r)}=(H^*_{i(r)})$, $r=1, 2, \ldots, g-1$ denote the rth group in block i, and there is $H^*_{i(r)}$ contained in this group. Let $K_{i(g)}=(H^*_{i(g)},H^*_{i(g+1)},...,H^*_{i(M_i)})$ denote the gth group in block i, and there are $H^*_{i(g)},H^*_{i(g+1)},\ldots$ and $H^*_{i(M_i)}$ contained in this group. After determining g groups, the M_i possible haplotype probabilities in block i for individual n can be denoted as $P_n(K_i\bullet)=(p_n^{K_{i(1)}},p_n^{K_{i(2)}},...,p_n^{K_{i(g)}})$.

Note that the number of groups (g) in block i is not restricted to a certain number (e.g., 6); however, we assumed $g\le M_i$ in this study because most blocks contained only three to five haplotypes [14]. Therefore, it is not necessary to reduce the amount of data in each haplotype block. However, doing so, can save time when calculating all possible haplotype combinations as described next.

2.5. Constructing haplotype combinations

For all haplotype combinations, we calculate $\prod_{i=1}^{B} p_n^{K_{i(j^*)}}$ for any $j^*=1,2,...,g$. The value of $\prod_{i=1}^{B} p_n^{K_{i(j^*)}}$ is multiplied by 10 and round off to the nearest integer. This integer can be regarded as the number of times that individual n carries its corresponding haplotype combination. A $10N\times B$ matrix is created by combining haplotype combinations for all individuals. For example, suppose $N=2, B=2$ and $g=2$. There are $g^B=2^2$ combinations, i.e., $K_{1(1)}K_{2(1)},K_{1(1)}K_{2(2)},K_{1(2)}K_{2(1)},K_{1(2)}K_{2(2)}$. The value of $\prod_{i=1}^{2} p_n^{K_{i(j^*)}}$ for these four combinations is (0.8,0,0.2,0) for individual *1* and (0,1,0,0) for individual *2*. After multiplying the values by 10 and rounding off to the nearest integer, the number is (8,0,2,0) for individual *1* and (0,10,0,0) for individual *2*. A 20×2 matrix can be created as follows:

$$\begin{array}{c} \\ n=1\\ \\ \\ \\ \\ \\ n=2\\ \\ \\ \\ \\ \end{array}\begin{array}{c} i=1 \quad i=2\\ \left[\begin{array}{cc} K_{1(1)} & K_{2(1)}\\ K_{1(1)} & K_{2(1)}\\ \vdots & \\ K_{1(1)} & K_{2(1)}\\ K_{1(2)} & K_{2(1)}\\ K_{1(2)} & K_{2(1)}\\ K_{1(1)} & K_{2(2)}\\ \vdots & \\ K_{1(1)} & K_{2(2)} \end{array}\right]_{20x2}\end{array}$$

As a result, one unphased dataset, i.e., a genotype dataset with unknown haplotype information, will expand to a large dataset, i.e., a genotype dataset with known haplotype information, reflecting the uncertainties of the haplotypes.

2.6. HH interaction

2.6.1. MDR

MDR is a non-parametric method and it is free of any assumed genetic model. This method reduces data dimensionality by pooling genotypes into either high-risk or low-risk groups for a disease, thereby circumventing the problem of high-order genotype combinations with a low number of observations [3,17–18].

In the past, MDR has successfully identified combinations of multi-locus genotypes and discrete environmental factors that are associated with diseases [11,12]. In this study, we used Haploview and PHASE software to infer haplotype data. We used "index scores" to adjust for bias arising from haplotype uncertainties in haplotype estimation. With above data, MDR was used to detect HH interactions.

Cross-validation (CV) consistency [21] is a measure of the number of times an HH model (an HH interaction) is identified as the best model across the CV subsets. Taking a 10-fold CV for example (i.e., CV consistency ranges from 1 to 10), the phased dataset is divided into a training set (9/10 of the data) and a testing set (1/10 of the data). For each CV subset, the balance testing accuracy is calculated for the testing set and defined as (Sensitivity + Specificity)/2. The average balance testing accuracy is the mean of the 10 CV subsets and is a measure of how well the final HH model predicts the risk status in testing sets. The HH model with the highest CV consistency is selected as the final HH model. Once the final HH model is chosen, permutation testing is used to test the significance of this final HH model by evaluating the magnitude of the average balance testing accuracy. The statistical significance of the MDR results is assessed by comparing the average balance testing accuracy of the observed data to the distribution of average balance testing accuracy under the null hypothesis of no association, which is derived empirically from multiple permuted datasets by randomizing the disease status labels. The null hypothesis is rejected when the Monte Carlo p-value, derived from the permutations, is less than 0.05. The measures as described above and the procedures of the MDR method applied to the case-control haplotype study data is presented schematically in Supplementary Fig. 1. Throughout this study, we used a 10-fold CV and 1000-fold permutation testing, a commonly accepted standard [3,4].

2.6.1.1. MDR procedure for detecting HH interactions. Below we summarize the steps used to detect HH interactions from genotype data via MDR. We have integrated these steps into a user-friendly software.

Step 1: Perform block partitioning by inputting unphased data (genotype data) using Haploview v4.1 software.
Step 2: Infer each individual's probabilities for all possible haplotype pairs for each block identified in Step 1 using PHASE v2.1 software.
Step 3: Calculate the index score for $H_{i(j)}$, $S(H_{i(j)})$, by using the haplotype probabilities inferred from Step 2.
Step 4: Rank the results from Step 3 to determine g groups in each block.
Step 5: Create a $10N \times B$ matrix by calculating $\prod_{i=1}^{B} p_n^{K_{i(j^*)}}$, i.e., create a phased dataset.
Step 6: Use MDR software to select the final HH model with the highest CV consistency and significant average balance testing accuracy by using the phased dataset created in Step 5.

2.6.2. CART

CART is a non-parametric method and can be used to select predictors and their interactions that are important in determining an outcome variable. In a case-control study, the CART method involves two central steps in construction of classification trees [10]. The first step is a recursive partitioning process that splits the root node (all samples) into two offspring nodes by the value of a predictor variable. The second step is a "pruning" process that removes unnecessary splits from the bottom to the top. The CART method has been proven capable of detecting combinations of multi-locus genotypes and discrete environmental factors associated with some diseases [11,12]. By inferring haplotype data from genotype data as described above for subsequent usage with MDR, these data can also be applied to CART to identify HH interactions.

Since CART is a tree-based approach, some haplotype blocks are prioritized to be used in constructing a classification tree. Haplotype blocks with stronger main effects for the disease phenotype will be closer to the root node [22]. Hence, CART is able to indicate which haplotype block in a significant HH interaction that has a stronger contribution to disease risk. For example, a significant HH interaction $H_{3\cdot} \times H_{2\cdot}$ is selected by CART. The first-appearing haplotype block $H_{3\cdot} = (H_{3(1)}, H_{3(2)}, H_{3(.3)}, H_{3(4)}, H_{3(5)}, H_{3(6)})$ is selected first to split the tree, and the second split is according to the other haplotype block $H_{2\cdot} = (H_{2(1)}, H_{2(2)}, H_{2(.3)}, H_{2(4)}, H_{2(5)}, H_{2(6)})$. Thus, $H_{3\cdot}$ has a stronger contribution to disease risk than $H_{2\cdot}$. Moreover, CART can provide information regarding specific haplotype pairs in a haplotype block that are associated with diseases. For example, an individual with the haplotype pairs $H_{3(2)}$ in $H_{3\cdot}$ and the haplotype pairs $H_{2(1)}$ in $H_{2\cdot}$ is categorized to the case group, indicating that the interactions between these haplotype pairs contribute to disease risk.

In the recursive partitioning process of CART, the root node comprises the total set of haplotype pairs for each individual, i.e., total $H_{i(j)}$. In this study, the R (http://www.r-project.org) [24] package 'tree' [23] was utilized with the following parameters: the Gini index was used as a splitting criterion; a 10-fold CV was used to evaluate overall model fit. The final classification tree was selected as that with the highest CV consistency across each of the 10 CV subsets. Permutations were used to test the significance of the final classification tree by evaluating the magnitude of average balance testing accuracy. The permutation testing was similar to that used for MDR. A 10-fold CV and a 1000-fold permutation testing were used in this study.

2.6.2.1. CART procedure for detecting HH interactions. The CART procedure consists of the following steps. These steps were integrated into a user-friendly software.

Steps 1–5: Follow steps 1–5 for the MDR procedure presented above.
Step 6: Use TREE (R package) to select the final classification tree with the highest CV consistency and significant average balance testing accuracy.

To summarize, for HH interactions we have integrated a series of programs (i.e., R, PHASE, Haploview, MDR and CART) into a convenient software that is available for both Linux and Windows platforms. Software, example datasets, and a user guide are available at the website http://www.csjfann.ibms.sinica.edu.tw/EAG/program/programlist.htm.

2.7. Simulation studies

We divided our simulation studies into two parts. The first one used simulated genotype data without LD, and the second used data with LD. For both simulated data, we assessed the power for identification of SNP–SNP interactions and HH interactions associated with the disease by using MDR and CART. Furthermore, we evaluated the performance of MDR, CART and hapForest [32] for identification of HH interactions associated with the disease.

2.7.1. Generating genotype data without LD

In the first part of the simulation study, we generated 36 SNPs in a case-control association study with 1500 cases and 1501 controls by using the simulation software SNaP [26]. We considered 12 two-locus interaction models, M_i, (where $i = 01–12$) (Table 1) described by Knapp et al. [25]. For the analysis of power, the central loci (3rd and 17th) were assumed to be the disease loci but not included in the power analysis due to the deterministic relationship between these loci and disease status. Instead of using the disease loci data, we assumed high LD between marker 3 and marker 4 ($D' = 0.8$) and between marker 17 and marker 18 ($D' = 0.85$). The 4th and 18th loci were used as quasi-disease markers. Except for these two markers (3rd and 17th loci), the other 34 SNPs used in this simulation study were independent of each other. These 34 loci were used to evaluate type I error.

The following two criteria were used as the definition for "interaction was detected":

1) the HH interaction for the two quasi-disease markers within the two haplotype blocks was identified by MDR or CART as in the final HH model; and
2) the SNP–SNP interaction of the two quasi-disease markers was identified by MDR or CART as in the final SNP–SNP model.

2.7.2. Generating genotype data with LD

A simulation study using 1500 cases and 1501 controls was undertaken by using the simulation software SNaP. We considered the three factors described below when generating genotype data with LD.

1) *Disease model*: In total, 12 disease models, M_i, (where $i = 01–12$) were considered [25], including epistatic models with main effects, epistatic models without main effects, and heterogeneity models (Table 1).
2) *Block size*: Gabriel et al. [14] observed that as few as two or three markers were sufficient to identify regions as blocks. Based on this information, we used three markers as the threshold for determining a block and considered two levels of block size: short (S), having at most three markers in a block, and long (L), having more than three markers in a block.
3) *Haplotype frequency*: Gabriel et al. [14] also found that most blocks contained only three to five haplotypes, and these major haplotypes provided 90% of the information for a given block. Based on this information, we considered two levels of haplotype frequencies: extreme (E), where the frequency of one major haplotype was ≥ 0.6, and average (A), where all possible haplotypes had equal frequency. For the average frequency pattern, we assumed five haplotypes in a block, with equal frequency of 0.2. For the extreme frequency pattern, we assumed three haplotypes—the major haplotype having a frequency of 0.6, and each of the other two haplotypes having a frequency of 0.2.

We considered four unlinked haplotype blocks (Table 2): two disease-related haplotype blocks (DR_B1 and DR_B2) and two disease-unrelated haplotype blocks (DU_B1 and DU_B2). The assumptions of block size and haplotype frequency of each haplotype block are shown in Table 2. The central loci (i.e., the 2nd locus of DR_B1 and DR_B2) were the assumed disease loci, but they were not included in the power analysis due to their deterministic relationship with disease status. The two disease-unrelated haplotype blocks (DU_B1 and DU_B2) were used to evaluate type I error. These marker loci within the same block had a high degree of LD ($D' > 0.8$) with each other. Markers in different blocks had low or no LD ($D' < 0.02$).

The 48 different scenarios (i.e., M_iSE, M_iSA, M_iLE, and M_iLA where $i = 01–12$) were simulated using the varied combinations of disease model, block size, and haplotype frequency as described above. For

Table 1
Penetrance table for the two-locus disease models.

M_{01} (Prevalence = 0.1) Epi$_{OUT}$M[a]				M_{02} (Prevalence = 0.1) Epi$_{OUT}$M[a]				M_{03} (Prevalence = 0.1) Epi$_{OUT}$M[a]			
P(A)[b] = 0.21		P(B)[c] = 0.21		P(A)[b] = 0.6		P(B)[c] = 0.199		P(A)[b] = 0.577		P(B)[c] = 0.577	
	BB	Bb	bb		BB	Bb	bb		BB	Bb	bb
AA	0.707[d]	0.707	0	AA	0.778[d]	0.778	0	AA	0.9[d]	0	0
Aa	0.707	0.707	0	Aa	0	0	0	Aa	0	0	0
aa	0	0	0	aa	0	0	0	aa	0	0	0
M_{04} (Prevalence = 0.1) Epi_M[e]				M_{05} (Prevalence = 0.1) Epi$_{OUT}$M[a]				M_{06} (Prevalence = 0.07) Epi_M[e]			
P(A)[b] = 0.372		P(B)[c] = 0.243		P(A)[b] = 0.349		P(B)[c] = 0.349		P(A)[b] = 0.190		P(B)[c] = 0.190	
	BB	Bb	bb		BB	Bb	bb		BB	Bb	bb
AA	0.911[d]	0.911	0	AA	0.799[d]	0.799	0	AA	0[d]	1	1
Aa	0.911	0	0	Aa	0.799	0	0	Aa	1	0	0
aa	0.911	0	0	aa	0	0	0	aa	1	0	0
M_{07} (Prevalence = 0.1) Epi$_{OUT}$M[a]				M_{08} (Prevalence = 0.1) Het[f]				M_{09} (Prevalence = 0.1) Het[f]			
P(A)[b] = 0.194		P(B)[c] = 0.194		P(A)[b] = 0.053		P(B)[c] = 0.053		P(A)[b] = 0.279		P(B)[c] = 0.04	
	BB	Bb	bb		BB	Bb	bb		BB	Bb	bb
AA	1[d]	1	0.512	AA	0.744975[d]	0.74498	0.495	AA	0.8844[d]	0.8844	0.66
Aa	1	0.512	0	Aa	0.74498	0.74498	0.495	Aa	0.66	0.66	0
aa	0.512	0	0	aa	0.495	0.495	0	aa	0.66	0.66	0
M_{10} (Prevalence = 0.074) Het[f]				M_{11} (Prevalence = 0.1) Het[f]				M_{12} (Prevalence = 0.1) Het[f]			
P(A)[b] = 0.194		P(B)[c] = 0.194		P(A)[b] = 0.052		P(B)[c] = 0.052		P(A)[b] = 0.288		P(B)[c] = 0.045	
	BB	Bb	bb		BB	Bb	bb		BB	Bb	bb
AA	1[d]	1	1	AA	0.522[d]	0.522	0.522	AA	1[d]	1	1
Aa	1	0	0	Aa	0.522	0.522	0.522	Aa	0.574	0.574	0
aa	1	0	0	aa	0.522	0.522	0	aa	0.574	0.574	0

[a]Epistatic model without main effects. [b]Frequency of the disease allele at locus 1. [c]Frequency of the disease allele at locus 2. [d]Penetrance of the genotype carrying AA and BB copies of the disease haplotype at locus 1 and 2, respectively. [e]Epistatic model with main effects. [f]Heterogeneity model.

Table 2
Assumptions for four unlinked haplotype blocks for simulations.

Two disease-unrelated haplotype blocks (DU_B1 and DU_B2) and one disease-related haplotype block (DR_B2)	
DU_B1 with 6 markers (L[a]) and average frequency (A[c]) DU_B2 with 3 markers (S[b]) and average frequency (A) DR_B2 with 4 markers (L) and extreme frequency (E[d])	
One disease-related haplotype block (DR_B1)	
Block Size	Haplotype Frequency
DR_B1 with 3 markers (S) DR_B1 with 4 markers (L)	DR_B1 with extreme frequency (E) DR_B1with average frequency (A)

[a]The number of markers is greater than 3. [b]The number of markers is less than or equal to 3. [c]All possible haplotypes with equal haplotype frequency. [d]One major haplotype with haplotype frequency greater than 0.6.

example, M_{03}SA denoted the M_{03} disease model of DR_B1 with short block size (S) and average haplotype frequency (A).

The following three criteria were used as the definition for "interaction was detected":

1) the HH interaction for the two disease-related haplotype blocks was identified by MDR or CART as in the final HH model.
2) the SNP–SNP interaction, i.e., the final SNP–SNP model, of the two loci within the two disease-related haplotype blocks, respectively, was identified by MDR or CART.
3) hapForest identified the two disease-related haplotype blocks having a significant *P*-value.

To summarize, we simulated 12 different scenarios for genotype data without LD and 48 different scenarios for genotype data with LD. For each scenario, 1000 replicates were conducted to assess the power and type I error on the basis of 1000 permutations.

3. Results

3.1. Performance of MDR and CART for detecting SNP–SNP interactions and HH interactions

For the genotype data without LD, the power of MDR and CART for detecting SNP–SNP interactions and HH interactions was the same in all scenarios (86% and 78% for MDR and CART, respectively).

For the genotype data with LD, under the scenarios assuming short block and average frequency, the power of detecting SNP–SNP interactions and HH interactions was about the same for MDR (51% vs 52%) and CART (32% vs 33%). For detecting SNP–SNP interactions in the rest of the scenarios, MDR had much lower power (42%) compared to 82% for detecting HH interactions. Similarly, CART had 28% power for detecting SNP–SNP interactions compared to 53% for detecting HH interactions. The above results (Supplementary Fig. 2) displayed that HH interactions performed better than SNP–SNP interactions for both MDR and CART for the genotype data with LD. We then evaluated the performance of MDR, CART and hapForest [32] for identification of HH interactions associated with disease for the genotype data with LD.

3.2. Simulation results for detecting HH interactions using genotype data with LD

3.2.1. Power analysis of MDR and CART for detecting HH interactions

The impact of disease model (Fig. 1A), block size (Fig. 1B), and haplotype frequency (Fig. 1C) on power was explored separately. For example, the average power of each method under M_{01} was defined as the sum of the power for the scenarios M_{01}SE, M_{01}SA, M_{01}LE, and M_{01}LA divided by 4. The average power for the 12 disease models ranged between 74% and 82% for MDR and between 32% and 60% for CART (Fig. 1A). For CART, the average power for epistatic models with main effects, i.e., M_{04} and M_{06}, and a heterogeneity model with higher penetrance, i.e., M_{10}, was higher than the other disease models. MDR appeared to outperform CART for detecting HH interactions when the disease model was allowed to vary. For block size, the average power for S and L blocks was 70–84% for MDR and 39–51% for CART (Fig. 1B). Both methods gave a higher average power for long blocks. For haplotype frequency, the average power range for A and E frequencies was 70–83% for MDR and 37–53% for CART (Fig. 1C). Both methods gave a higher average power for haplotypes with extreme frequencies.

In addition, we assessed the power of MDR and CART to detect HH interactions under each given scenario (Fig. 1D). For most of the scenarios, MDR was robust to variation of the three factors based on its power performance. However, for some of the scenarios, e.g., for block size S and haplotype frequency A, MDR showed a 30% decrease in power, down to 58%. This was likely due to a high misclassification rate in these scenarios where many cells of a contingency table had a case-control ratio of nearly 1.0 (see Step 7 of Supplementary Fig. 1). That is, many individuals could be grouped as either high or low risk, so the average balance testing accuracy was low. The power of CART was greatly influenced by the choice of disease model and haplotype frequencies. Because of the binary splits, CART was more likely to detect HH interactions in the presence of a strong main effect. With an extreme haplotype frequency, epistatic models with main effects (M_{04} and M_{06}) and the heterogeneity model with higher penetrance (M_{10}) had higher power than all the other scenarios, regardless of the block size. Because the power of CART was susceptible to the first split, if the first haplotype block selected had no significant main effects then the resulting tree could be interpreted as being less reliable. Thus, CART was deemed better for detecting HH interactions in the presence of a strong main effect.

3.2.2. Type I error analysis of MDR and CART for detecting HH interactions

The change of type I error for disease model (Fig. 1A), block size (Fig. 1B), and haplotype frequency (Fig. 1C) was explored. The average type I error for the 12 disease models was 5.4–6.9% and 2.6–4.9% for MDR and CART, respectively (Fig. 1A). For block size, the average type I error for S and L blocks was 5.9–6.6% for MDR and 3.5–4.4% for CART (Fig. 1B). The average type I error range for A and E haplotype frequencies was 6.2–6.3% and 3.9–4.0% for MDR and CART, respectively (Fig. 1C). These three factors did not substantially affect type I error for either method.

In addition, the effect of varying all three factors on type I error for MDR and CART for detecting HH interactions was also examined (Fig. 1D). The range of type I error for the 48 scenarios for MDR was 4.9–7.5%, and most of the type I error for CART was less than 5% (2.4% on average). The type I error for MDR and CART was robust to different scenarios, but CART appeared to control it slightly better than MDR.

3.2.3. Evaluating performance of MDR, CART and hapForest for detecting HH interactions

For the simulation study of power, hapForest was greatly influenced by block size. HapForest tended to have higher power (95%) for long blocks regardless of disease model and haplotype frequency. When assuming a short block and extreme frequency, epistatic models with main effects and heterogeneity models with higher penetrance also attained satisfactory power (88%). Under the remaining scenarios, the power of hapForest was not satisfactory (16%). The type I error values of hapForest under the 48 scenarios ranged from 3.1% to 7.6%, indicating that the three factors did not substantially affect its type I error.

In summary, only MDR was robust to variation of the three factors in terms of power, whereas the other two approaches were

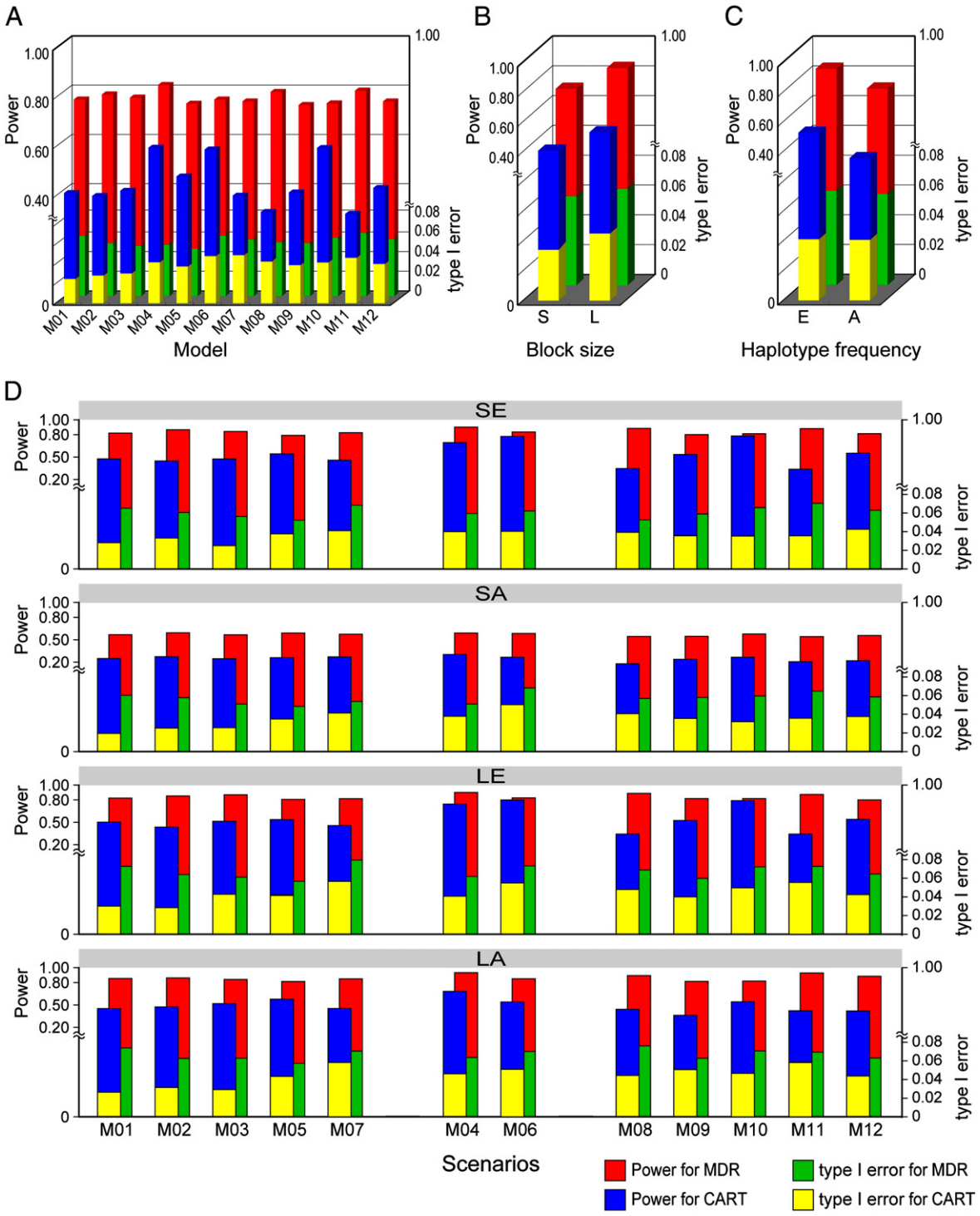

Fig. 1. Results of MDR and CART for detecting HH interactions for simulation studies. The *x*-axis demonstrates 12 disease models in (A), 2 block sizes in (B), 2 haplotype frequencies in (C) and 48 scenarios in (D). The left *y*-axis shows the power given by MDR (red) and CART (blue), respectively. The right *y*-axis indicates the type I error given by MDR (green) and CART (yellow), respectively. The *z*-axis represents the two methods.

considerably affected by these factors. The type I error for each of the approaches was robust to the variations in the factors.

3.3. Application of MDR and CART to Parkinson's disease data

We applied our proposed MDR and CART procedures to test for HH interactions using publicly available case-control data on Parkinson's disease [15]. The genotype data on 408 K SNPs included a combination of Illumina Infinium I and HumanHap 300 assays for 270 individuals with idiopathic Parkinson's disease and 271 neurologically normal controls.

With a large number of markers such as SNPs and haplotypes, it is not easy to identify which haplotypes should be considered for HH interaction testing. Marchini et al. [27] suggested that a two-stage strategy be implemented, in which all loci that met some lenient threshold in a single-locus search should be subsequently assessed for possible multi-locus

genotype associations. We followed the three procedures described bellow and applied Marchini's strategy in our single-locus analysis to determine important HH interactions in the Parkinson's disease dataset. These procedures were:

1) *Single-locus analysis*: Single-locus tests were performed using Armitage's trend test [28] with 1000 permutations using the PROC CASECONTROL statistical procedure in the SAS/Genetics™ package (SAS Institute, Inc. Cary, NC, USA). The SNPs included for further study were selected based on a loose threshold (p-value$<10^{-3}$) for a dataset with a large number of markers. The final determined SNPs, including 10 kb of flanking sequence of the significant SNPs, were used in subsequent analyses.
2) *Haplotype analysis*: Haplotype association tests were performed with 1000 permutations using the PROC HAPLOTYPE statistical procedure in the SAS/Genetics™ package. The haplotype with main effect was defined based on a threshold with p-value$<10^{-4}$.
3) *HH interaction analysis*: We used our provided software to select important HH interactions in the Parkinson's disease dataset and compared our results with some knowledge of Parkinson's disease.

3.3.1. Results for Parkinson's disease data

A total of 11,655 SNPs were excluded according to the following quality control criteria: minor allele frequency<0.01, genotype call rates<0.95, and departure from Hardy-Weinberg equilibrium (i.e., p-value$<10^{-4}$). The dataset used for the final analysis contained 384,936 SNPs (i.e., 97% of the total) and 541 individuals (270 cases and 271 controls).

In total, 422 significant SNPs were identified with a p-value$<10^{-3}$. Expanding to cover 10 kb of flanking sequence of the 422 significant SNPs, a total of 1429 SNPs were included in the subsequent analyses. Using Haploview and these 1429 SNPs, 311 haplotype blocks were identified, and each individual's haplotypes for each block were reconstructed using PHASE software. Of the 311 haplotype blocks, 11 had main haplotype effects (i.e., p-value$<10^{-4}$).

All possible HH interactions (more than 48,000) were exhaustively examined. Numerical results and gene information for the identified top 20 HH interactions by MDR and CART and the corresponding haplotypes are summarized in Supplementary Table 1 and Supplementary Table 2, respectively. For MDR, 19 of the top 20 HH interactions had no main haplotype effects (Supplementary Table 1). The most significant HH interaction found by MDR was *ZFAT1*RAC2*. These genes have not been reported to be related to Parkinson's disease, although they are involved in brain neoplasms and nervous system dysfunction [30,31].

For CART, 4 of the top 20 HH interactions had main haplotype effects, and all involved *RET* (Supplementary Table 2). *RET* plays a crucial role in neural crest development and reportedly is linked to Parkinson's disease [29]. The most significant HH interaction identified by CART contained two haplotype blocks: (1) rs727048 and rs5756587 located in *RAC2* and (2) rs7202238 and rs11648686 located in *TOX3*. *RAC2* was identified to have more contributions to Parkinson's disease than *TOX3*, as *RAC2* was the first haplotype block chosen and had main haplotype effects. Moreover, the most significant HH interaction that made the greatest contribution to Parkinson's disease risk in this analysis consisted of the haplotype pair of (CT,CT), (TC,CT), (CC,CT), (CC,CC), (CC,TC), or (TC,TC) in the first haplotype block and the haplotype pair of (TA,TA), (TA,CA) (TA,CG), (CA,CA), (CA,CG), or (CG,CG) in the second haplotype block. Like *RAC2*, *TOX3* is associated with brain neoplasms and nervous system dysfunction [31], and *TOX3* has also been reported to regulate calcium-dependent transcription in neurons [33].

To summarize the Parkinson's disease results, MDR appears to be capable of detecting HH interactions associated with the disease even without a main haplotype effect. CART is more useful for capturing HH interactions with main haplotype effects than for those interactions lacking a main haplotype effect.

4. Discussion

The goal of this study was to use MDR and CART with the "index score" to identify disease-related HH interactions and to investigate the influence of block size, haplotype frequency, and disease model on the performance of these two methods. Both methods used CV to avoid over-fitting while permutation testing was used to assess statistical significance. The combination of CV and permutation testing could help to avoid inflated type I error due to multiple testing [3,16,38–40].

In general, the results of both MDR and CART showed that HH interactions had higher power than SNP–SNP interactions for all the 48 scenarios in simulated genotype data with LD. The results suggest that haplotype-based analyses have greater power over single-locus analyses when SNPs are in strong LD with the risk locus. We also compared the detection of SNP–SNP interactions and the top 20 HH interactions from MDR and CART using a Parkinson's disease dataset. Thirteen of the top 20 HH interactions were detected by SNP–SNP interactions. Eight of the top 20 SNP–SNP interactions were not located in the haplotype blocks. The remaining 12 SNP–SNP interactions located in the haplotype blocks were also detected by HH interactions. These results showed that HH interactions can provide more information than SNP–SNP interactions if disease-related SNPs are located in haplotype blocks. In other words, SNP–SNP interactions can better provide information if disease-associated SNPs are outside haplotype blocks. SNP–SNP interactions or HH interactions cannot completely replace each other. Therefore, we suggest that SNP–SNP interactions and HH interactions should be considered as complementary to each other and used in parallel to thoroughly analyze datasets.

One conventional and still widely used approach for detecting HH interactions is logistic regression. In our simulation studies, we also assessed the results of MDR, CART, logistic regression, and hapForest in the detection of HH interactions. The drawback of logistic regression is that the usual maximum likelihood estimates of the log-odds ratio parameters are biased for small samples. In such situations, it is very likely that the maximum likelihood estimates will not converge. This drawback is the major reason for the lower power of logistic regression compared to other methods in our simulation results. We also have used logistic regression to obtain the odds ratio by recoding interactions based on the risk level in MDR. But the maximum likelihood estimate still cannot converge in some simulated data. Using different codings cannot circumvent empty or sparse samples in some combinations of HH interactions because this is the inherent limitation of logistic regression. The drawback of hapForest is that even though the approach can identify haplotypes with main effects and/or interactions of disease-related haplotypes, it is difficult to distinguish whether the disease-related haplotypes result from single haplotype effects or from HH interactions. Because HH interaction tests are not provided in hapForest, statistical tests are needed to detect possible HH interactions. By contrast, MDR and CART can detect HH interactions directly. Moreover, MDR and CART can handle the sparseness of data in high dimensions, and can account for nonlinear HH interactions, so that HH interactions missed by logistic regression are more likely to be detected.

CART can differentiate the contributions that each haplotype block makes to disease risk for detecting HH interactions. That is, the first haplotype block chosen contributes more to the disease than the second one in an HH interaction. Additionally, CART can explicitly reveal disease-related haplotype pairs within each haplotype block in an HH interaction, whereas MDR only provides information on finding disease-associated HH interactions. A notable feature of CART is the influence of the first split on the tree structure. If the first haplotype block selected has no significant main effects, the resulting tree could be less reliable. Because MDR does not rely on binary splits as it performs a systematic search through all possible HH interactions, it

can identify more interactions than CART. The computational time for MDR, however, grows exponentially as the number of haplotype blocks increases. This is a limitation for most combinatorial-based data mining approaches. Because the MDR and CART approaches are fundamentally different, the two methods should be considered complementary to each other when studying HH interactions in various disease models.

Fung et al. (2006) reported 26 significant SNPs using five tests of single-locus association. Our single-locus analyses confirmed 21 of the 26 SNPs reported by Fung et al. when a threshold p-value$<10^{-3}$ was used. Of the five SNPs not detected in our study, three were filtered out by quality control criteria and two did not pass the Armitage trend test criterion (i.e., p-value$>10^{-3}$). The results of HH interaction detected from the Parkinson's disease dataset were consistent with our simulation findings. For example, for CART, *RET*, which contained two SNPs, rs3004212 and rs1480597, had main haplotype effects that interacted with different partners. Four out of the top 20 HH interactions had these two SNPs that composed of the haplotype located in *RET* (Supplementary Table 2). For MDR, *RET* contained three SNPs, rs2075914, rs3004214 and rs2505513, with no main haplotype effects (Supplementary Table 1). MDR is useful to detect HH interactions regardless of main haplotype effects, whereas CART prefers those HH interactions with main haplotype effects. Thus, combining the information of MDR and CART in detecting HH interactions will help researchers analyze datasets thoroughly.

We also compared the difference between SNP data with and without filtering for detecting HH interactions by using MDR and CART. We selected SNPs on chromosome 22 in the Parkinson's disease dataset with a p-value$<10^{-3}$ for filtered data (62 SNPs) and used all 7071 SNPs for unfiltered data. Most (87.5%) of the associated HH interactions found by using the unfiltered SNP data were also detected with the filtered SNP data. Only one significant HH interaction (p-value$=0.0001$) detected using the unfiltered SNP data was not found using the filtered data. Our results demonstrate that use of a lenient threshold (e.g. p-value$<10^{-3}$) in a dataset with a large number of markers to filter SNP data for subsequent HH analyses seems to be satisfactory and computationally efficient compared with using all SNP data. These results are consistent with what has been suggested by Marchini et al. (2005). Therefore, we only reported the computational burden of searching for all possible HH interactions in chromosome 22 for filtered data. It took 3 min to analyze 541 individuals (271 controls and 270 cases) and 62 SNPs (13 haplotype blocks). For filtered Genome-wide analysis on 1429 SNPs consisting of 311 haplotype blocks, it took about 1 h to detect HH interactions using our windows-based workstation with 2.41 GHz CPU. Our reanalysis of the Parkinson's disease dataset not only confirmed a landmark finding in genetic association studies but also discovered some potentially new candidate genes related to the disease. We caution, however, that the sample size in the Parkinson's disease dataset is relatively small, and hence these candidate genes require further investigation. Our research illustrates the important role of HH interactions in Parkinson's disease and shows that the two methods, MDR and CART, are useful for analyzing real data.

5. Conclusions

When disease-related SNPs are outside haplotype blocks, SNP–SNP interactions can better provide information. HH interactions can provide more information than SNP–SNP interactions when disease-associated SNP markers are located in haplotype blocks. Such information is of considerable interest in genetic association studies. Currently, there are many studies that use state-of-the-art genotyping techniques, the human genome will eventually be thoroughly studied. In this HH interaction study, we used two data mining tools and to adjust for haplotype uncertainties arising from inference. According to our results, these two data mining tools, i.e., MDR and CART are very useful in overcoming complexities due to large numbers of haplotypes. Our findings have shown that SNP–SNP and HH interaction analyses should complement each other in dissecting possible risk factors in genetic studies.

Supplementary materials related to this article can be found online at doi:10.1016/j.ygeno.2010.11.003.

Acknowledgments

We would like to thank the reviewers, for their variable and constructive comments and suggestions which have helped to improve the manuscript. We were additionally grateful to Dr. Hsin-Chou Yang (Institute of Statistical Science Academia Sinica, Nankang, Taipei, Taiwan) and Dr. Hsun-Chih Kuo (Department of Statistics, National Chengchi University, Taipei, Taiwan) for their suggestions on data analysis. This project was partially funded by a grant from the Taiwan National Science Council (NSC 96-2628-B-001-015-MY2, 98-2628-B-001-008).

References

[1] G. Davey Smith, S. Ebrahim, S. Lewis, A.L. Hansell, L.J. Palmer, P.R. Burton, Genetic epidemiology and public health: hope, hype, and future prospects, Lancet 366 (9495) (2005) 1484–1498.
[2] T. Niu, Z.S. Qin, X. Xu, J.S. Liu, Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms, Am. J. Hum. Genet. 70 (2002) 157–169.
[3] M.D. Ritchie, L.W. Hahn, N. Roodi, L.R. Bailey, W.D. Dupont, F.F. Parl, J.H. Moore, Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer, Am. J. Hum. Genet. 69 (2001) 138–147.
[4] M.D. Ritchie, A.A. Motsinger, Multifactor dimensionality reduction for detecting gene–gene and gene–environment interactions in pharmacogenomics studies, Pharmacogenomics 6 (2005) 823–834.
[5] J.H. Moore, The ubiquitous nature of epistasis in determining susceptibility to common human diseases, Hum. Hered. 56 (2003) 73–82.
[6] E.R. Martin, M.D. Ritchie, L. Hahn, S. Kang, J.H. Moore, A novel method to identify gene–gene effects in nuclear families: the MDR-PDT, Genet. Epidemiol. 30 (2006) 111–123.
[7] A.S. Andrew, H.H. Nelson, K.T. Kelsey, J.H. Moore, A.C. Meng, D.P. Casella, T.D. Tosteson, A.R. Schned, M. Karagas, Concordance of multiple analytical approaches demonstrates a complex relationship between DNA repair gene SNPs, smoking and bladder cancer susceptibility, Carcinogenesis 27 (2006) 1030–1037.
[8] D. Brassat, A.A. Motsinger, S.J. Caillier, H.A. Erlich, K. Walker, L.L. Steiner, B.A. Cree, L.F. Barcellos, M.A. Pericak-Vance, S. Schmidt, S. Gregory, S.L. Hauser, J.L. Haines, J.R. Oksenberg, M.D. Ritchie, Multifactor dimensionality reduction reveals gene–gene interactions associated with multiple sclerosis susceptibility in African Americans, Genes Immun. 7 (2006) 310–315.
[9] S. Qin, X. Zhao, Y. Pan, J. Liu, G. Feng, J. Fu, J. Bao, L. He, An association study of the N-methyl-D-aspartate receptor subunit gene (GRIN1) and NR2B subunit gene (GRIN2B) in schizophrenia with universal DNA microarray, Eur. J. Hum. Gene 13 (2005) 807–814.
[10] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth International Group, Belmont, CA, 1984.
[11] L. Briollais, Y. Wang, I. Rajendram, V. Onay, E. Shi, J. Knight, H. Ozcelik, Methodological issues in detecting gene–gene interactions in breast cancer susceptibility: a population-based study in Ontario, BMC Med. 5 (2007) 22.
[12] A.S. Andrew, M.R. Karagas, H.H. Nelson, S. Guarrera, S. Polidoro, S. Gamberini, C. Sacerdote, J.H. Moore, K.T. Kelsey, E. Demidenko, P. Vineis, G. Matullo, DNA repair polymorphisms modify bladder cancer risk: a multi-factor analytic strategy, Hum. Hered. 65 (2008) 105–118.
[13] M.J. Daly, J.D. Rioux, S.F. Schaffner, T.J. Hudson, E.S. Lander, High-resolution haplotype structure in the human genome, Nat. Genet. 29 (2001) 229–232.
[14] S.B. Gabriel, S.F. Schaffner, H. Nguyen, J.M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S.N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E.S. Lander, M.J. Daly, D. Altshuler, The structure of haplotype blocks in the human genome, Science 296 (2002) 2225–2229.
[15] H.C. Fung, S. Scholz, M. Matarin, J. Simon-Sanchez, D. Hernandez, A. Britton, J.R. Gibbs, C. Langefeld, M.L. Stiegert, J. Schymick, M. Okun, R.J. Mandel, H.H. Fernandez, K.D. Foote, R.L. Rodriguez, E. Peckham, F. Wavrant De Vrieze, K. Gwinn-Hardy, J.A. Hardy, A.B. Singleton, Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data, Lancet Neurol. 5 (2006) 911–916.
[16] J. Hoh, A. Wille, J. Ott, Trimming, weighting, and grouping SNPs in human case-control association studies, Genome Res. 11 (12) (2001) 2115–2119.
[17] M.D. Ritchie, L.W. Hahn, J.H. Moore, 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 (2003) 150–157.

[18] L.W. Hahn, M.D. Ritchie, J.H. Moore, Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions, Bioinformatics 19 (2003) 376–382.
[19] J.C. Barrett, B. Fry, J. Maller, M.J. Daly, HaploView: Analysis and visualization of LD and haplotype maps, Bioinformatics 21 (2005) 263–265.
[20] M. Stephens, N.J. Smith, P. Donnelly, A new statistical method for haplotype reconstruction from population data, Am. J. Hum. Genet. 68 (2001) 978–989.
[21] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, IJCAI (1995) 1137–1145.
[22] A. Bureau, J. Dupuis, K. Falls, K.L. Lunetta, B. Hayward, T.P. Keith, P. Van Eerdewegh, Identifying SNPs predictive of phenotype using random forests, Genet. Epidemiol. 28 (2005) 171–182.
[23] RipleyB.D. , Package 'tree'[R package Version 1.0-27], http://cran.r-project.org/web/packages/tree/index.html.
[24] R Development Core Team, R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2008, http://www.r-project.org.
[25] M. Knapp, S.A. Seuchter, M.P. Baur, Two-locus disease models with two marker loci: the power of affected-sib-pair tests, Am. J. Hum. Genet. 55 (1994) 1030–1041.
[26] M. Nothnagel, Simulation of LD block-structured SNP haplotype data and its use for the analysis of case-control data by supervised learning methods, Am. J. Hum. Genet. 71 (suppl 4) (2002) A2363.
[27] J. Marchini, P. Donnelly, L.R. Cardon, Genome-wide strategies for detecting multiple loci influencing complex diseases, Nat. Genet. 37 (2005) 413–417.
[28] P. Armitage, Tests for linear trends in proportions and frequencies, Biometrics 11 (1955) 375–386.
[29] C.M. Backman, L. Shan, Y.J. Zhang, B.J. Hoffer, S. Leonard, J.C. Troncoso, P. Vonsatel, A.C. Tomac, Gene expression patterns for GDNF and its receptors in the human putamen affected by Parkinson's disease: a real-time PCR study, Mol. Cell. Endocrinol. 252 (1-2) (2006) 160–166.
[30] O. Shmueli, S. Horn-Saban, V. Chalifa-Caspi, M. Shmoish, R. Ophir, H. Benjamin- Rodrig, M. Safran, E. Domany, D. Lancet, GeneNote: whole genome expression profiles in normal human tissues, C. R. Biol. 326 (10–11) (2003) 1067–1072.
[31] J. Ng, T. Nardine, M. Harms, J. Tzu, A. Goldstein, Y. Sun, G. Dietzl, B.J. Dickson, L. Luo, Rac GTPases control axon growth, guidance and branching, Nature 416 (2002) 442–447.
[32] X. Chen, C. Liu, M. Zhang, H. Zhang, A forest-based approach to identifying gene and gene gene interactions, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 19199–19203.
[33] S.H. Yuan, Z. Qiu, A. Ghosh, TOX3 regulates calcium-dependent transcription in neurons, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 2909–2914.
[34] T. Becker, J. Schumacher, S. Cichon, M.P. Baur, M. Knapp, Haplotype interaction analysis of unlinked regions, Genet. Epidemiol. 29 (2005) 313–322.
[35] J. Zhang, J.J. Regieli, M. Schipper, M.M. Entius, F. Liang, J. Koerselman, H.J. Ruven, Y. van der Graaf, D.E. Grobbee, P.A. Doevendans, Inflammatory gene haplotype-interaction networks involved in coronary collateral formation, Hum. Hered. 66 (2008) 252–264.
[36] W. Makarasara, N. Kumasaka, A. Assawamakin, A. Takahashi, A. Intarapanich, C. Ngamphiw, S. Kulawonganunchai, U. Ruangrit, S. Fucharoen, N. Kamatani, S. Tongsima, pHCR: a parallel haplotype configuration reduction algorithm for haplotype interaction analysis, J. Hum. Genet. 54 (11) (2009) 634–641.
[37] B. Kulle, A. Frigessi, H. Edvardsen, V. Kristensen, L. Wojnowski, Accounting for haplotype phase uncertainty in linkage disequilibrium estimation, Genet. Epidemiol. 32 (2) (2008) 168–178.
[38] E. Arehart, S. Gleim, B.C. White, J. Hwa, J.H. Moore, Multifactor dimensionality reduction analysis identifies specific nucleotide patterns promoting genetic polymorphisms, BioData Min. 2 (2) (2009).
[39] C.S. Greene, D.S. Himmelstein, H.H. Nelson, K.T. Kelsey, S.M. Williams, A.S. Andrew, M.R. Karagas, J.H. Moore, Enabling personal genomics with an explicit test of epistasis, Pac. Symp. Biocomput. 15 (2010) 327–336.
[40] J. Li, L. Ji, Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix, Heredity 95 (2005) 221–227.