BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 1 2007, pages 71–76
doi:10.1093/bioinformatics/btl557

Genetics and population analysis

Odds ratio based multifactor-dimensionality reduction method for detecting gene–gene interactions

Yujin Chung[1], Seung Yeoun Lee[2], Robert C. Elston[3] and Taesung Park[1,*]

[1]Department of Statistics, Seoul National University, San 56-1 Shillim-Dong, Kwanak-Gu, Seoul 151-747, Korea, [2]Department of Applied Mathematics, Sejong University, 98 Gunja-Dong Kwangjin-Gu, Seoul 143-747, Korea and [3]Department of Epidemiology and Biostatistics, Case Western Reserve University, 10900 Euclid Avenue Cleveland, OH 44106-7281, USA

Received on June 27, 2006; revised on September 11, 2006; accepted on October 27, 2006
Advance Access publication November 8, 2006
Associate Editor: Keith A Crandall

ABSTRACT

Motivation: The identification and characterization of genes that increase the susceptibility to common complex multifactorial diseases is a challenging task in genetic association studies. The multifactor dimensionality reduction (MDR) method has been proposed and implemented by Ritchie *et al.* (2001) to identify the combinations of multilocus genotypes and discrete environmental factors that are associated with a particular disease. However, the original MDR method classifies the combination of multilocus genotypes into high-risk and low-risk groups in an *ad hoc* manner based on a simple comparison of the ratios of the number of cases and controls. Hence, the MDR approach is prone to false positive and negative errors when the ratio of the number of cases and controls in a combination of genotypes is similar to that in the entire data, or when both the number of cases and controls is small. Hence, we propose the odds ratio based multifactor dimensionality reduction (OR MDR) method that uses the odds ratio as a new quantitative measure of disease risk.

Results: While the original MDR method provides a simple binary measure of risk, the OR MDR method provides not only the odds ratio as a quantitative measure of risk but also the ordering of the multilocus combinations from the highest risk to lowest risk groups. Furthermore, the OR MDR method provides a confidence interval for the odds ratio for each multilocus combination, which is extremely informative in judging its importance as a risk factor. The proposed OR MDR method is illustrated using the dataset obtained from the CDC Chronic Fatigue Syndrome Research Group.

Availability: The program written in R is available.

Contact: tspark@snu.ac.kr

1 INTRODUCTION

The general strategy for identifying Mendelian disease genes has largely been unsuccessful when applied to identifying susceptibility genes for common complex multifactorial diseases, such as asthma (Altmuller *et al.*, 2001). This is because the Mendelian approach requires each susceptibility factor to have a large independent main effect on disease risk (Moore and William, 2002). The effect of any single genetic variation for a common complex disease may be dependent on other genetic variations (gene–gene interaction) and environmental factors (gene–environment interaction). To address this issue, several methods, such as logistic regression models, multilocus linkage disequilibrium (LD) tests, and Hardy–Weinberg equilibrium tests have been applied. However, most of these methods require a large sample size to model high-order interactions (Moore and William, 2002). Unfortunately, it is not easy to collect large sample size data. Moreover, when logistic regression is used, multicollinearity may occur due to LD.

For moderate sample size data, one method for detecting and characterizing interactions in common complex multifactorial diseases is the multifactor dimensionality reduction (MDR) method (Ritchie *et al.*, 2001). This method detects and characterizes the high-order gene–gene and gene–environment interactions in case-control studies. Using this method, multilocus genotypes are classified into high-risk and low-risk groups, effectively reducing the genotype predictors from n dimensions to one dimension. The new, one-dimensional multilocus genotype variable is evaluated for its ability to classify and predict disease status through cross-validation (CV). The MDR method is model-free, in that it does not assume any particular genetic model. This is important for diseases in which the mode of inheritance is unknown and possibly very complex. Moreover, the MDR method is non-parametric, in that it does not estimate any parameters (Ritchie *et al.*, 2001).

Although the MDR method provides many useful features, it has several drawbacks. First, its method of determining high-risk or low-risk groups is *ad hoc*—in the sense that it classifies cells, defined by combination of multilocus genotypes, into high-risk or low-risk groups based on a simple comparison of the ratios of the number of cases and controls. Hence, the MDR method is prone to false positive and negative errors when the ratio of the number of cases and controls in a combination of genotypes is similar to that in the entire data, or when both the number of cases and controls in a combination of genotypes is small.

Second, the MDR binary classification does not provide any quantitative measure of disease risk for each combination of genotypes but only provides a binary measure (high or low) of disease risk. Further, the MDR method does not provide any information regarding how well the high-risk group is characterized. Third, the MDR method does not allow comparison of the disease risks between different combinations of genotypes. Thus, it is not possible to identify which combination of genotypes in the high-risk group has the highest risk or which combination in the low-risk group has the lowest risk. In practical applications, however, it is

*To whom correspondence should be addressed.

important to know whether a certain combination of genotypes has a higher risk than other combinations.

In this paper, we propose the odds ratio based multifactor dimensionality reduction (OR MDR) method to overcome the above mentioned limitations of the MDR method. The OR MDR method uses the odds ratio of each combination of genotypes as a quantitative measure of disease risk, so that we can order the combinations of genotypes from the highest to the lowest in terms of the odds ratios. Moreover, a confidence interval of the odds ratio can be obtained, either by using large sample theory or bootstrap samples for each combination of genetoypes.

The MDR method is briefly reviewed in Section 2.1, and the new OR MDR method is proposed in Section 2.2. An example giving a comparison between the results of the MDR and the OR MDR methods is provided in Section 3 using a dataset obtained from the CDC Chronic Fatigue Syndrome Research Group. The discussion and final conclusions are included in Section 4.

2 METHODS

2.1 MDR method

The MDR method has been proposed by Ritchie *et al*. (2001) and Moore and William (2002), and implemented by Hahn *et al*. (2003) and Ritchie *et al*. (2003). It comprises the following two stages. Stage 1 involves choosing the best combination of multifactors. Stage 2 involves classifying the combinations of genotypes into high-risk and low-risk groups.

Figure 1 describes the procedure used to implement the MDR method. First, the data are divided into 10 subsets for CV—nine are classified as training sets and one as an independent test set. Second, the value of n is designated depending upon the number of factors being considered. Then, a set of n genetic and/or environmental factors is selected. The n factors and their possible multifactor classes are represented in n-dimensional space. Next, the ratio of the number of cases to the number of controls within each multifactor class is calculated. Each multifactor class in n-dimensional space is labeled as 'high risk' if the ratio of the number of cases to that of the controls is equal to or exceeds a particular threshold; it is labeled as 'low risk' if that threshold is not exceeded. Thus the n-dimensional space is reduced to one dimension with two levels (low-risk and high-risk). Usually, the threshold is determined as the ratio of the number of cases to the number of controls in the training dataset. The threshold is equal to one in a balanced dataset. Among all the multifactor combinations, the MDR model with the lowest number of misclassified individuals is selected. In order to evaluate the predictive ability of the model, the prediction error for the selected combination of factors is estimated using the independent test data.

After repeating the above procedures for each of the 10 training and test set, a single model that minimizes the average prediction error is selected from the various n-multifactor combinations, and the CV consistency is calculated. CV consistency is a measure of the number of times a particular set of multifactors is identified during the CV (Moore *et al*., 2002b).

Moreover, the whole process is repeated for different values of n, and the best combination is selected from among each possible dimension of combinations by repeating the above procedure. The result is a set of best models; one for each dimension. That is, for each different value of n, we have a list of the best models.

Furthermore, the above procedures are performed 10 times using different random number seeds to reduce the probability of observing spurious results due to chance divisions of the data. The average prediction errors and CV consistencies are calculated, and the best combination with a minimum average of prediction errors is selected. From these selected combinations,

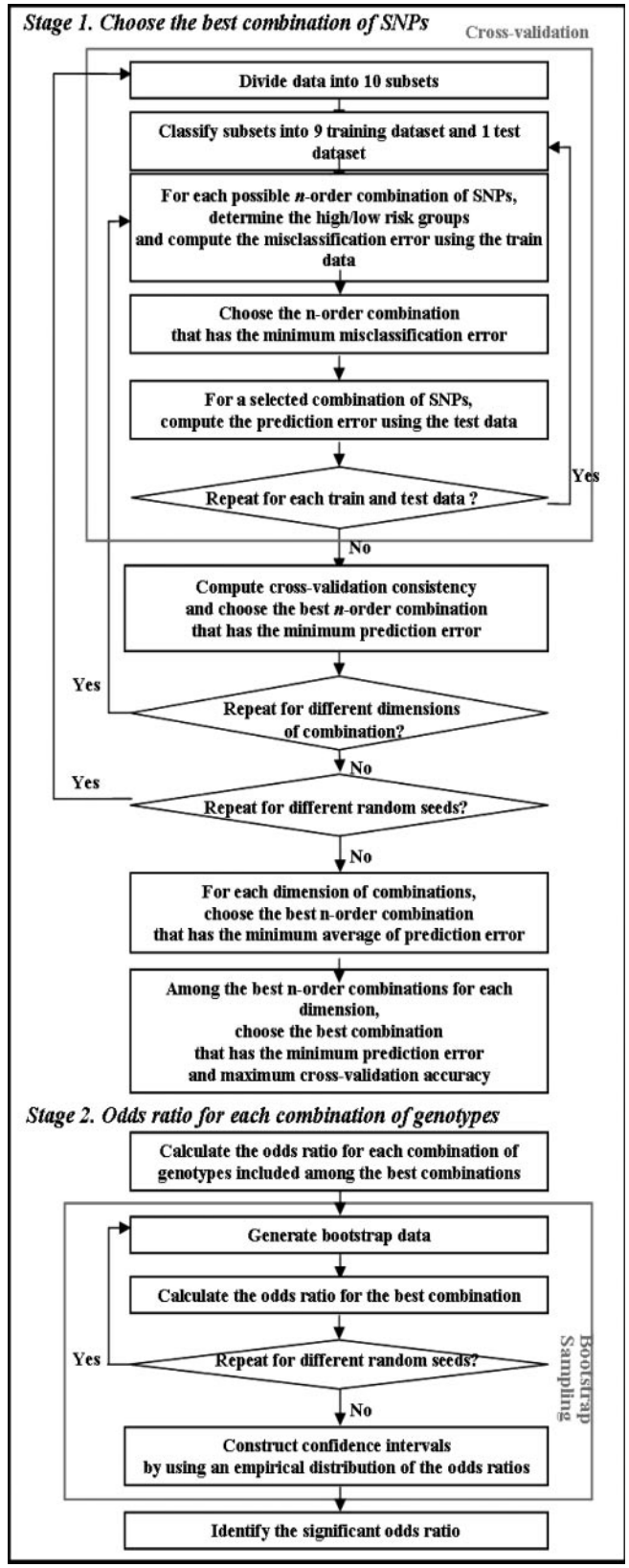

Fig. 1. Summary of OR MDR method.

the model with the combination of genotypes that maximizes the CV consistency and minimizes the prediction error is selected. If the CV consistency is maximal for one model and the prediction error is minimal for another model, then the model with the lowest number of multifactors is selected.

Once the MDR method identifies the best combination of multifactors, Stage 2 of the MDR method is performed. This involves classifying the multilocus genotype levels as high- or low-risk. However, the MDR binary classification of multilocus genotypes into the high-risk and low-risk groups is based on a simple comparison of the ratios of the number of cases and controls to that of each multilocus genotype combination.

2.2 OR MDR method

We propose the OR MDR method to improve the ad hoc classification of the MDR method. Figure 1 illustrates the procedure to implement the OR MDR method. Stage 1 is the same as that of the MDR method explained in section 2.1. In Stage 2, the odds ratio for each combination of genotypes is used as a quantitative measure of disease risk.

To illustrate the OR MDR method, we assume that two single nucleotide polymorphisms (SNPs; SNP1 and SNP2) each with three genotypes are selected as the best model in Stage 1. The two SNPs and a binary variable distinguishing cases and controls yield a $3 \times 3 \times 2$ contingency table for N subjects. The observed cell frequencies are denoted by $\{n_{ijk}\}$, where the subscripts i and j represent the two SNPs and k represents the disease (case) and normal (control) phenotypes. The odds of the disease for the given genotype combination (SNP1 $= i$, SNP2 $= j$) is

$$\frac{\mathrm{P}[\mathrm{Disease} \mid \mathrm{SNP1} = i, \mathrm{SNP2} = j]}{\mathrm{P}[\mathrm{Normal} \mid \mathrm{SNP1} = i, \mathrm{SNP2} = j]} = \frac{\mathrm{P}[\mathrm{SNP1} = i, \mathrm{SNP2} = j \mid \mathrm{Disease}]}{\mathrm{P}[\mathrm{SNP1} = i, \mathrm{SNP2} = j \mid \mathrm{Normal}]} \times \frac{\mathrm{P}[\mathrm{Disease}]}{\mathrm{P}[\mathrm{Normal}]}.$$

Then, the odds for the genotypes (i, j), θ_{ij} is given as follows:

$$\frac{\mathrm{P}[\mathrm{SNP1} = i, \mathrm{SNP2} = j \mid \mathrm{Disease}]}{\mathrm{P}[\mathrm{SNP1} = i, \mathrm{SNP2} = j \mid \mathrm{Normal}]} = \frac{\mathrm{P}[\mathrm{Disease} \mid \mathrm{SNP1} = i, \mathrm{SNP2} = j]}{\mathrm{P}[\mathrm{Normal} \mid \mathrm{SNP1} = i, \mathrm{SNP2} = j]} \Big/ \frac{\mathrm{P}[\mathrm{Disease}]}{\mathrm{P}[\mathrm{Normal}]}.$$

Note that the righthand side is the odds of the disease for the genotypes (i, j) divided by the odds of the disease for all the data disregarding the genotype information. Then, θ_{ij} is estimated as follows

$$\hat{\theta}_{ij} = \frac{n_{ij1}/n_{++1}}{n_{ij2}/n_{++2}},$$

where $n_{++k} = \sum_i \sum_j n_{ijk}$ for $k = 1, 2$. Note that this estimator is different from the ordinary odds ratio estimator, because the marginal sum n_{++k} contains n_{ijk}. We propose using the odds ratio estimator $\hat{\theta}_{ij}$ as a quantitative measure of disease risk for a given genotype combination. If this odds ratio is equal to one, then the odds of the case for the given genotype combination is equal to the odds of the case for the entire data, i.e. the risk associated with the disease for the given genotype combination is the same as the overall risk estimated from all the case and control samples. Thus, this genotype combination is not associated with the disease. On the other hand, the larger the odds ratio (>1), the stronger the positive association between the genotype combination and the disease. Similarly, the lower the odds ratio (<1), the stronger the negative (protective) association between the genotype combination and the disease.

We can use θ_{ij} as a quantitative measure to represent the disease risk for a given genotype combination. As in the MDR method, each multifactor combination in n-dimensional space is labeled as 'high-risk' if the odds ratio exceeds the threshold, or as 'low-risk' if that threshold is not exceeded. If the threshold chosen is one, the binary classification of the OR MDR method is the same as that of the MDR method.

In addition, we can compare combinations of genotypes by using the odds ratios. For example, two genotype combinations (SNP1 $= i$, SNP2 $= j$) and (SNP1 $= i'$,SNP2 $= j'$) can be compared using the two odds ratios θ_{ij} and $\theta_{i'j'}$. That is, if $\theta_{ij} > \theta_{i'j'}$, then (SNP1 $= i$, SNP2 $= j$) has a higher risk than (SNP1 $= i'$, SNP2 $= j'$). Based on the values of θ, we can order the combination of genotypes from the highest risk group to the lowest risk group.

The relative disease risk among the genotype combinations can also be compared by choosing a baseline genotype combination. The most common genotype combination is usually selected as the baseline combination. As an example, suppose the genotype combination (SNP1 $= 1$, SNP2 $= 1$) has the largest frequency. Then, it is selected as a baseline combination. For the genotype combination (SNP1 $= i$, SNP2 $= j$), the ratio θ_{ij}/θ_{11} provides the relative disease risk. In fact, the ratio is the ordinary odds ratio and these can be compared with each other.

The proposed OR MDR method also provides information on the accuracy of the odds ratio estimators. Unlike the original MDR method, the OR MDR method provides a measure of accuracy by deriving a cell-specific confidence interval for θ_{ij}. We consider two types of confidence intervals: one is the usual asymptotic confidence interval for the ratio of two success probabilities derived from the two independent binomial distributions, and the other is a bootstrap confidence interval that can be used when the sample size is not large. For the best combination obtained in Stage 1, the bootstrap samples are generated by randomly selecting cases and controls with replacement for each genotype combination. This resampling procedure is repeated approximately 100 000 times. Then, the empirical distributions of odds ratios for each combination of genotypes are constructed. The empirical confidence interval for each combination of genotypes can be obtained from these bootstrap distributions. Finally, we can use these confidence intervals of the odds ratios to conclude whether a particular genotype combination is significantly more or less associated with the disease than another genotype combination.

3 EXAMPLE

The dataset obtained from the CDC Chronic Fatigue Syndrome Research Group includes gene expression, proteomic, SNP and clinical data. In this paper, we focus only on the SNP data. Information pertaining to the 42 SNPs in the dataset is described in Table 1; more information is available on the website (http://www.camda.duke.edu/camda06/datasets/index.html). Our analysis is based on 55 subjects ever having had chronic fatigue syndrome (CFS) and 54 non-fatigued controls. In this analysis, we used the 35 CFS subjects and 36 non-fatigue subjects who do not have any missing SNP values.

We applied the MDR and the OR MDR methods to all possible combinations of the 42 SNPs up to the fourth order. Table 2 summarizes the CV consistency and the prediction errors obtained from Stage 1 of the OR MDR method, which is identical to the MDR method. One of the two-SNP models has a maximum CV consistency of 6.6 out of 10, and one of the four-SNP models has a minimum prediction error of 0.35. Generally, the combination of SNPs that maximizes the CV consistency and minimizes the prediction error is selected. In our example, however, the CV consistency was maximum for one model and the prediction error was minimum for an other model. Thus, the model with the fewer SNPs was selected, i.e. the two-SNP model comprising rs6196 (NR3C1) and rs140701 (SLC6A4).

In Table 3, the results of Stage 2 of the OR MDR method and the MDR method are compared. The first column represents genotypes of the best combination of two SNPs, rs6196 (NR3C1) and rs140701 (SLC6A4), and the second column shows their frequencies in the cases and controls. The third column shows the binary classification of high-risk and low-risk groups for each combination of genotypes. Three combinations of genotypes are classified as high-risk, five as low-risk, and one empty cell is undetermined.

Table 1. List of 42 SNPs

Gene	SNPs
CRHR1	rs110402, rs242924, rs173365, rs242940, rs7209436, rs1396862
SLC6A4	rs140701, hCV7911132, rs2066713
MAOA	rs979605, rs1801291, rs979606
NR3C1	rs860458, rs258750, rs2918419, rs6188, rs1866388, rs852977 rs6196
COMT	rs6269, rs4633, rs933271, rs4646312, rs5993882, rs740603 rs 165722
TPH2	rs10784941, rs2171363, rs4760816, rs1872824, rs4760750 rs1386486, rs1487280
TH	rs2070762, rs4074905
POMC	rs12473543
MAOB	rs2283729, rs3027452, rs1799836
CRHR2	rs2267710, rs2284217, rs2267714

Table 2. Selection of the best combination of SNPs by Stage 1 of the OR MDR method and the MDR method

The best combination in each dimension	Prediction error	CV consistency
rs6196, rs140701	**0.366505**	**6.6**
rs740603, rs6196, rs140701	0.381994	2
rs1799836, rs2171363, rs140701, rs1396862	0.354444	4

The model with maximum CV consistency and minimum prediction error is indicated in bold type.

However, the MDR method does not provide any information beyond simple binary classification. The three high-risk genotype groups may have different disease risks. Moreover, the MDR method is vulnerable to false positive and negative errors when the ratio of the numbers of cases and controls in a combination of genotypes is similar to that of the entire data, or when the numbers of cases and controls are very small. For example, consider the genotype AA/CC. Its frequency ratio in the cases and controls is equal to one, which is similar to the ratio for the entire data. Although AA/CC is classified as high-risk, a small change in this frequency can change this classification from the high-risk group to the low-risk group. Thus, the classification of AA/CC is vulnerable to false positive error. On the other hand, although AA/TT is classified as high risk, this combination is quite robust to a small change in the frequencies of the cases and controls. Thus, AA/TT appears to show much stronger evidence for its classification as high risk. Unfortunately, the MDR method does not distinguish between these two combinations. Moreover, the number of cases and controls in AG/CC is very small; hence, the MDR method classifies AG/CC as low-risk. However, AG/CC is also vulnerable to false negative error because a small change in its frequencies can cause a change in its classification from the low-risk group to the high-risk group.

In the OR MDR method, however, the odds ratio provides a more rigorous quantitative measure of disease risk. For each combination of genotypes, the fourth column in Table 3 indicates the odds ratios; the fifth column, its rank and the sixth column, its 95% asymptotic confidence interval; the seventh column, the 95% confidence interval from the bootstrap samples. Note that the genotypes AG/CC, AG/CT and AG/TT have 0 frequencies, which made it difficult (impossible in the case of AG/TT) to estimate a confidence interval by either method.

Both asymptotic and bootstrap confidence intervals provided consistent results. There is reasonably good evidence based on the 95% confidence interval that AA/TT is a high risk combination, but the large confidence intervals clearly show that there is little evidence regarding the other combinations—whereas the MDR classifies several of them as high risk. If the upper (lower) limit of the confidence interval for one of the other combinations were less (greater) than 1, that would be good evidence for a truly low (high) risk combination. Only the confidence interval of the cell with the genotype AA/TT does not contain 1. The odds of disease for this genotype are 2.674 times that of the overall odds, showing positive association between the genotype of AA/TT and CFS. Therefore, we conclude that the risk of CFS is positively associated with two genotypes—AA and TT.

The ranks of the odds ratios indicate which is the highest-risk and lowest-risk group. We can also compare two different combinations of genotypes. For example, when the genotype AA/CC is used as a baseline combination of genotypes, the odds of disease for AA/TT is 2.599 (=2.674/1.029) times larger than the odds of disease for AA/CC.

The first SNP rs6196 is located in nuclear receptor subfamily 3; group C, member 1 glucocorticoid receptor (NR3C1), which regulates glucocorticoid levels in the blood. NR3C1 was shown to have a significant association with CFS; this supports the hypothesis that medically unexplained chronic fatigue is heterogeneous and presents preliminary evidence of the genetic mechanisms underlying a few of the putative conditions (Smith *et al*., 2006). In particular, different classes of subjects with unexplained fatigue were distinguished by gene polymorphsims that were involved in either hypothalanmic-pituitary-adrenal (HPA) axis function or neurotransmitter systems, including proopiomelanocortin (POMC), NR3C1, monoamine oxidase A (MAOA), monoamine oxidase B (MAOB) and tryptophan hydroxylase 2 (TPH2). Recently, Geortzel *et al*. (2006) showed that 28 well-selected SNPs could predict with 76% accuracy whether a person has CFS, and that the top three important genes are TPH2, catechol-O-methyltransferase (COMT) and NR3C1. rs6196, in particular, is a special case of missense mutations in which a change in one nucleotide results in the substitution of one amino acid that results in a non-functional protein.

The other SNP rs140701 is located in solutes carrier family 6, neurotransmitter transporter, serotonin, member 4 (SLC6A4). Neuroendocrine axis assessment is one of the best and safest approaches for the assessment of specific neurotransmitter function. Based on the neuroendocrine responses in fatiguing disorders, Chaudhuri and Behan (2004) derived a biological model of central fatigue.

We expect rs140701, which is the sixth intronic SNP in SLC6A4, to play an important role as a transcription regulator. By examining the evolutionary origin and mechanisms of the differential transcriptional regulation of SLC6A4, Soeby *et al*. (2005) addressed the possible impact of the second intronic variable number of tandem repeats (VNTR) on behavior and disease, and found new putative binding sites for several transcription factors in the VNTRs of the mammalian SLC6A4 gene. Further, Soeby *et al*. showed that the

Table 3. Comparison between the results of Stage 2 of the MDR and OR MDR methods SNPs[a]

SNPs[a]	Cell frequency[b]	High/Low-risk[c]	Odds ratio	Rank[d]	95% Asymptotic[e] CI	95% Bootstrap[f] CI
AA/CC	10:10	High	1.029	2	(0.489, 2.162)	(0.441, 2.400)
AG/CC	0:1	Low	0	7	.	.
GG/CC	5:8	Low	0.643	5	(0.233, 1.775)	(0.187, 1.646)
AA/CT	3:3	High	1.029	2	(0.223, 4.756)	(0.205, 5.143)
AG/CT	0:1	Low	0	7	.	.
GG/CT	1:4	Low	0.257	6	(0.030, 2.189)	(0.000, 1.543)
AA/TT	13:5	High	2.674	1	(1.065, 6.714)	(1.018, 8.229)
AG/TT	0:0			.		
GG/TT	3:4	Low	0.771	4	(0.186, 3.201)	(0.171, 6.171)

[a]The genotype before the solidus is rs6196 (NR3C1 gene) and that after the the solidus is rs140701 (SLC6A4 gene).
[b]The first number indicates the number of cases and the second the number of controls in each cell.
[c]The threshold is the ratio of the total number of cases and controls, 0.9722.
[d]Order of odds ratio. The high risk genotype has the largest odds ratio value.
[e]Asymptotic confidence interval.
[f]Confidence interval based on the 100 000 bootstrap samples.

intronic VNTR has been selectively targeted through mammalian evolution to fine tune the transcriptional regulation of SLC6A4 expression.

In summary, it has been shown that rs6196 of NR3C1 regulates the HPA axis and rs140701 of SLC6A is a neurotransmitter transporter. Thus, we hypothesize that rs6196 and rs140701 are important CFS-related polymorphisms. In CFS-susceptible individuals, environmental stressors induce changes in the neuroendocrine axis mainly through the HPA axis and the norepinephrine system. Our analysis reveals a possible interaction between rs6196 in NR3C1 and rs140701 in SLC6A that is expected to play an important role in the biological mechanism of CFS.

4 DISCUSSION AND CONCLUSION

In this paper, we proposed the OR MDR method that uses the odds ratio as a quantitative measure of disease risk. Similar to the original MDR method, the OR MDR method is a non-parametric approach and assumes no particular genetic model. In addition, as in the case of the MDR method, the OR MDR method uses CV to select optimal models.

However, the OR MDR method has several advantages over the original MDR method that uses a binary measure of disease risk. First, the OR MDR method is based on the odds ratio for each combination of genotypes and reveals more information regarding the effect of a certain genotype combination on the disease risk, since the quantitative value of the odds ratio represents the strength of the association between the genotypes and disease. Second, the OR MDR method provides a more solid statistical validation by providing a confidence interval for each combination of genotypes. In particular, when the number of cases is similar to the number of controls, or when both the number of cases and controls is too small, the validity of the MDR approach in determining the high-risk and low-risk groups is questionable. On the other hand, the confidence interval from the OR MDR method provides much more information for the high-risk and low-risk classification. If the upper (lower) limit of the confidence interval for one or other of the combinations is less (greater) than 1, that is an evidence for a truly low (high) risk combination. We expect the OR MDR method to play a more important role than the MDR method in the identification of gene–gene interactions in real data.

However, similar to the MDR method, the OR MDR method has the limitation that comes with having empty cells because it cannot classify an empty cell as high risk or low risk. Further, confidence intervals cannot then be estimated by either the asymptotic method or bootstrap method. To solve this problem, a method based on the continuity correction needs to be developed. Such a method will be presented in a separate paper in the near future.

Finally, note that the odds ratio we have used for the OR MDR method is different from the ordinary odds ratio. One of the main reasons why we use θ for the OR MDR method is that we want to include MDR as a special case of OR MDR. That is, if OR MDR uses a binary classification with a threshold value of 1, then it is equivalent to MDR.

ACKNOWLEDGEMENTS

The authors would like to thank the associate editor and two anonymous referees whose comments were extremely helpful. This work was supported by the National Research Laboratory Program of Korea Science and Engineering Foundation (M10500000126), the Brain Korea 21 Project of the Ministry of Education, a grant of the Korea Health 21 R&D Project, Ministry of Health & Welfare, R. O. K(03-PJ10-PG13-GD01-0002), and grants from the U.S. Public Health Service: resource grant RR03655 from the National Center for Research Resources; research grant GM-28356 from the National Institute of General Medical Sciences; and Cancer Center Support Grant P30CAD43703 from the National Cancer Institute.

Conflict of Interest: none declared.

REFERENCES

Altmuller,J. *et al*. (2001) Genomwide scans of complex human disease: true linkage is hard to find. *Am. J. Hum. Genet*., **69**, 936–950.

Chaudhuri,A. and Behan,P.O. (2004) Fatigue in neurological disorders. *The Lancet*, **363**, 978–988.

Goertzel,B.N. *et al*. (2006) Combinations of single nucleotide polymorphisms in neuroendocrine effector and receptor genes predict chronic fatigue syndrome. *Pharmacogenomics*, **7**, 475–483.

Hahn,L.W. *et al*. (2003) Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. *Bioinformatics*, **19**, 376–382.

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

Moore,J.H. *et al*. (2002b) Symolic discriminant analysis of microarray data in autoimmune disease. *Genet. Epidemiol*., **23**, 57–69.

Ritchie,M.D. *et al*. (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet*., **69**, 138–147.

Ritchie,M.D. *et al*. (2003) Power of multifactor-dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genetic Epidemiolog*., **24**, 150–157.

Smith,A.K. *et al*. (2006) Polymorphisms in genes regulating the HPA axis associated with empirically delineated classes of unexplained chronic fatigue. *Pharmacogenomics*, **7**, 387–394.

Soeby,K. *et al*. (2005) Serotonin transporter: evolution and impact of polymorphic transcriptional regulation. *Am. J. Med. Genet. B. Neuropsychiatr Genet*., **5**, 53–57.