European Journal of Human Genetics (2008) 16, 735–741
© 2008 Nature Publishing Group All rights reserved 1018-4813/08 $30.00
www.nature.com/ejhg

ARTICLE

Variable selection in logistic regression for detecting SNP–SNP interactions: the rheumatoid arthritis example

Hui-Yi Lin*[,1], Renee Desmond[1], S Louis Bridges Jr[2] and Seng-jaw Soong[1]

[1]*Medical Statistics Section, Department of Medicine, University of Alabama at Birmingham, Birmingham, Alabama, USA;* [2]*Division of Clinical Immunology and Rheumatology, University of Alabama at Birmingham, Birmingham, Alabama, USA*

Many complex disease traits are observed to be associated with single nucleotide polymorphism (SNP) interactions. In testing small-scale SNP–SNP interactions, variable selection procedures in logistic regressions are commonly used. The empirical evidence of variable selection for testing interactions in logistic regressions is limited. This simulation study was designed to compare nine variable selection procedures in logistic regressions for testing SNP–SNP interactions. Data on 10 SNPs were simulated for 400 and 1000 subjects (case/control ratio = 1). The simulated model included one main effect and two 2-way interactions. The variable selection procedures included automatic selection (stepwise, forward and backward), common 2-step selection, AIC- and SC-based selection. The hierarchical rule effect, in which all main effects and lower order terms of the highest-order interaction term are included in the model regardless of their statistical significance, was also examined. We found that the stepwise variable selection without the hierarchical rule, which had reasonably high authentic (true positive) proportion and low noise (false positive) proportion, is a better method compared to other variable selection procedures. For testing interactions, the hierarchical rule effect was obvious. The procedure without the hierarchical rule requires fewer terms in testing interactions, so it can accommodate more SNPs than the procedure with the hierarchical rule. For testing interactions, the procedures without the hierarchical rule had higher authentic proportion and lower noise proportion compared with ones with the hierarchical rule. These variable selection procedures were also applied and compared in a rheumatoid arthritis study.

European Journal of Human Genetics (2008) **16,** 735–741; doi:10.1038/sj.ejhg.5202010; published online 30 January 2008

Keywords: single nucleotide polymorphisms; gene interaction; logistic model; variable selection

Introduction

Many complex diseases, such as breast cancer, ischemic stroke and hypertension have been found to be associated with single nucleotide polymorphism (SNP) interaction combinations.[1–4] Identifying such interactions is considered difficult because of the following reasons: weak or no marginal effects of some SNPs, a large number of SNPs and their interaction to consider, or lack of prior information about which SNPs interact. Several statistical methods have been proposed to deal with SNP–SNP interactions, such as multivariate adaptive regression splines,[4] multifactor dimensionality reduction,[5] combinational partitioning method,[6] neural networks[7]and random forest.[8]

Logistic regression (LR) is commonly used for analyzing SNP–SNP interactions in some situations, such as a study

***Correspondence: Dr H-Y Lin, Department of Medicine, Medical Statistics Section, University of Alabama at Birmingham, WTI 153, 1530 3rd Avenue S, Birmingham, Alabama 35294, USA.**
Tel: +001 205 934 3679; Fax: +001 205 934 4262;
E-mail: hylin@uab.edu

Received 26 July 2007; revised 30 November 2007; accepted 20 December 2007; published online 30 January 2008

with a limited number of SNPs,[9–13] or 2-stage genome-wide studies.[14] The popularity of LR is that it is widely available in most statistical packages and the output in terms of disease risk is easy to understand by non-statistician researchers. In comparison to some newly developed methods, the variable selection of LR achieves similar results within a limited number of SNPs.[15,16] LR also provides reasonable power for genome-wide SNP–SNP interaction studies.[14] In addition, LR can be used as the second-stage variable selection in testing for gene–gene or gene-environment interactions for a two-stage analysis that incorporates a bootstrap procedure as the first stage of selection.[17,18] However, the limitations of variable selection in LRs for testing SNP–SNP interactions have been discussed widely. The primary limitation of LR is that poor model parameter estimates may be generated because some genotype combinations have low frequency or zero responses especially when a large number of SNPs and high order of interactions are considered.[5,15,19]

The variable selection of 'main effects' in an LR has been discussed,[20,21] but variable selection in an LR for testing main effects with interactions has not been well documented. The Hierarchically Well-Formulated Rule (HWFR) is typically applied for interaction analysis in LRs.[22] Under the HWFR, all lower order terms of the highest order interaction term are included in the model, regardless of their statistical significance. Without following the HWFR, the results of interaction terms alone can also be interpretable, and LR can be used to analyze more SNPs because fewer terms are required. Although LR is popular in SNP-interaction studies, assessments of the performance of the LR variable selection procedures on detecting interaction terms are limited.

In this study, we tested the common LR selection procedures including stepwise, forward and backward selection with or without the HWFR. We also evaluate the Akaike information criterion (AIC)-based and Schwartz criterion (SC)-based selection procedures, which are a modified version of the stepwise LR selection procedure to select the final model based on AIC and SC.[23] The AIC[24] and SC[25] criteria are the common model selection criteria in LR. In addition, the two-step variable selection procedure was tested, and the details of this procedure are described in the following section. Among the significant variables in LR, a high percentage of them may be false positive.[20] It is important to validate the variable significance in LR. To know which significant terms in an LR may be truly significant, a bootstrap method[19,26] can be applied. With increasing numbers of SNP interaction studies, a rigorous comparison of the various variable selection procedures and the impact of the HWFR in LR for detecting SNP–SNP interactions is needed. The objective of this study is (1) to evaluate and compare performance of variable selection procedures in LR; and (2) to demonstrate the application of variable selection procedures in LR to a rheumatoid arthritis study.

Methods

Simulation

To evaluate performance of the variable selection procedures in an LR, we generated a case–control dataset with 400 and 1000 subjects (case/control = 200/200 and 500/500, respectively) where each subject had complete data on 10 SNPs. We assume no linkage disequilibrium among the 10 SNPs, so the genotype data on 10 SNPs were generated independently. The 10 SNPs followed Hardy–Weinberg equilibrium with percentages of common alleles from 0.6 to 0.8. We let WW, WV and VV denote homozygous wild type, heterozygous wild/variant type and homozygous variant genotypes, respectively. LRs were used to generate outcome variables, for the following two practical reasons: (1) LRs are the most commonly used methods to display the association between binary outcomes and genetic factors; and (2) this simulation method can allow for the determination of associations between 'multiple' SNP combinations and the outcome variable simultaneously.

The binary outcome variables of disease status were then generated based on the following LR:

$$\ln[p_i/(1-p_i)] = -0.32 + 0.8 * SNP5_i \\ -0.7 * SNP3_i * SNP4_i + 1.5 * SNP9_i * SNP10_i$$

where $i = 1, \ldots, n = 400$ or 1000, and p_i is the probability of disease. Among 10 candidate SNPs (SNP1–10), five of them are functional. Associated with the outcome variable, the inheritance modes of SNPs 3, 4, 5 and 9 were dominant, and the mode of SNP10 was recessive. The coding of the five functional SNPs was as follows:

$$\text{For m} = 3, 4, 5 \text{ and } 9,\ SNPm = \begin{cases} 0, & if\ WW \\ 1, & if\ WV/VV \end{cases}.$$

$$\text{For m} = 10, SNPm = \begin{cases} 0, & if\ WW/WV \\ 1, & if\ VV \end{cases}$$

The model was designed to examine various modes of inheritance (dominant and recessive), interaction type (1- or 2-way interaction) and risk/protective subgroups. In this model, the SNP5 WV/VV genotypes (dominant) and the genotype combination of SNP9 WV/VV and SNP10 VV (dominant–recessive) were positively associated with the disease; the genotype combination of SNP3 WV/VV and SNP4 WV/VV (dominant–dominant) was significantly associated with a lower disease risk.

Variable selection procedures

In practice, the true mode of inheritance for each SNP associated with the outcome is unknown. It is impractical to try different modes for all SNPs, so we treated each SNP as a categorical variable with two dummies using the common homozygous genotype as a reference category. Four primary types of variable selection procedures were examined: automatic selection, AIC-based selection, SC-based selection and two-step selection. The last three methods included an automatic variable selection procedure. We first evaluated six automatic variable selection procedures described in the following section, and then the best one was applied in the AIC-based, SC-based and two-step selection procedures.

Automatic selection

In the automatic selection procedure, a total of six procedures with combinations of three different selections (forward selection (F), stepwise selection (S) and backward elimination (B)) and the two HWFR statuses (yes/no) were examined. The most commonly used entry and removal criteria with *P*-values of 0.05 were used.

AIC- and SC-based selection

The AIC- and SC-based selection procedures are a modified version of the stepwise LR selection procedure to select the final model based on AIC and SC.[23] The AIC and SC are commonly used criteria for variable selection. The lower the value, the better the model. The purpose of these two methods is to reduce the number of models tested to a manageable size. Shtatland *et al*[23] used stepwise LR with entry and removal criteria of $P=1.0$ to generate a full sequence of models beginning with the null model and ending with the full model. Then, AIC and SC were calculated for each model in the sequence. The best-fit models were selected based on the smallest values of AIC or SC. Due to the large number of SNP-interaction combinations, it is difficult to use $P=1.0$ as the entry and removal criteria, so $P=0.5$ was applied to get the sequence of models.

Two-step selection

For two-step variable selection, bivariable selection to evaluate the association for one SNP with the outcome at a time is not appropriate, because the potential confounders are not included for control. A full model with all candidate variables and a backward elimination regression are recommended.[27,28] In this study, the first step of variable selection was to include all main effects of the 10 SNPs in the model and retained the covariates with a conservative *P*-value less than 0.2. The second step was to use the automatic selection with the best performance to detect possible SNP interactions.

Model evaluation

Results were evaluated based on prediction accuracy. Depending on whether the term was included in the simulated model, two types of terms could be observed: authentic and noise terms. Authentic (AP) and noise proportion (NP) were calculated based on the following equations. AP is the percentage of the authentic term selected in 1000 replications. Although we set up three terms in the simulated model, the three terms may not all be significant during 1000 replications by chance. Thus, the denominator for AP is the observed number of the significant authentic terms in the 1000 simulated models. NP is the percentage of noise terms among total terms in a model. For each condition, the average NP was obtained from 1000 replications.

$$\text{Authentic Proportion (AP)} = \frac{\text{number of times the authentic term was detected}}{\text{number of significant authentic terms}}$$

$$\text{Noise Proportion (NP)} = \frac{\text{number of false positive terms}}{\text{number of total terms in the model}}$$

Variable significance validation: bootstrap analysis

To know which significant terms in an LR may be truly significant, a bootstrap method[21,22] was applied. In this method, an arbitrary number of repeated samples with the original sample size were sampled with replacements from the original data set. Here, we generated 1000 bootstrap data sets from a given simulated data set. For each bootstrap data set, we recomputed LRs by using the best variable selection procedure in this study. The frequencies of significant terms selected in 1000 runs were presented. The lower the frequency that was selected, the lower the chance to be true positive.

Real data example: rheumatoid arthritis

Rheumatoid Arthritis (RA) is an autoimmune disease with complex etiology, which is assumed to be associated with several genetic or environmental factors. For example, the gene–gene and gene-environment interactions of RA involving *HLA-DRB1*, *PTPN22* and smoking has been discussed.[29] However, the inheritance of RA is largely unknown. The presence of anti-cyclic citrullinated peptide antibody can be used to characterize one subset of RA patients. A total of 426 (308 rheumatoid arthritis patients and 118 healthy controls) African-Americans were recruited from the ongoing Consortium for the Longitudinal Evaluation of African-Americans with Early Rheumatoid Arthritis Study.[30] We were interested in the following 10 SNPs in four genes, including: (i) *BTNL2*: rs2076529, rs2076530 and rs9268480; (ii) *MHC* region: rs2239804 and rs2395165; (iii) *FOXO3A*: rs7747393 and rs9285397; and, (iv) *NOTCH4*: rs394657, rs422951 and rs1109771.

These SNPs were treated as three-level categorical variables using the common homozygous genotype as the reference category. Hardy–Weinberg equilibrium was evaluated for all 10 SNPs by using both χ^2 and exact tests. In the RA control group, linkage disequilibrium in the same gene among the 10 SNPs was evaluated using Lewontin's D'. We tested for up to 2-way SNP–SNP interactions with anti-cyclic citrullinated peptide antibody status (positive *vs* negative) among African-American participants. For SNPs with low frequencies (around 5%) in a homozygous genotype, the homozygous genotype was combined with the heterozygous genotype.

Results

Comparison of automatic variable selection procedures

In Figure 1, we compared the six combinations of automatic variable selection procedures (S, B and F with or without following the HWFR) in LR. For impact of the HWFR, procedures following the HWFR had higher percentages to detect dominant SNP5 main effect in the model than procedures not following the HWFR (Figure 1a). However, the procedures without the HWFR had dramatically higher percentages to detect the 2-way interactions without main effects (Figures 1b and c). For 400 subjects, the percentages in detecting the main effect in stepwise selection were 98 *vs* 83% for with and without following the HWFR, respectively. Under the same condition, the percentages in detecting the dominant–dominant SNP3*SNP4 interaction were 3 and 62% for the selection with and without following the HWFR, respectively. The percentage of detecting the dominant-recessive SNP9*SNP10 interaction in stepwise selection with the HWFR and without the HWFR were 1 and 60%, respectively. In general, the procedures without the HWFR had lower NP compared with ones with the HWFR. For example, the stepwise selection without the HWFR could reduce NP to less than half (32%) compared with those with the HWFR (68%) in a sample size of 1000.

For testing the authentic terms, F and S selection without the HWFR similarly had better performance compared with B elimination without the HWFR. For the variable selection procedures with the HWFR, B elimination had the lowest percentage to detect the main effect but had the highest percentage to detect the 2-way interactions. As the percentage of noise terms, S without the HWFR had the lowest NP, and B with the HWFR had the highest NP.

The sample size impact was different for the procedures with or without the HWFR. In the procedures with the HWFR, both AP and NP increased when sample size increased. However, AP increased and NP decreased in the procedures without the HWFR when sample size increased. Based on the above results, S without the HWFR was the preferable variable selection procedure because it had a reasonably high AP and the lowest NP.

Comparison of stepwise, AIC-based, SC-based and 2-step variable selection

S selection without the HWFR, the best among all six combinations of automatic selection and the HWFR, was applied in the AIC-based, SC-based and 2-step variable

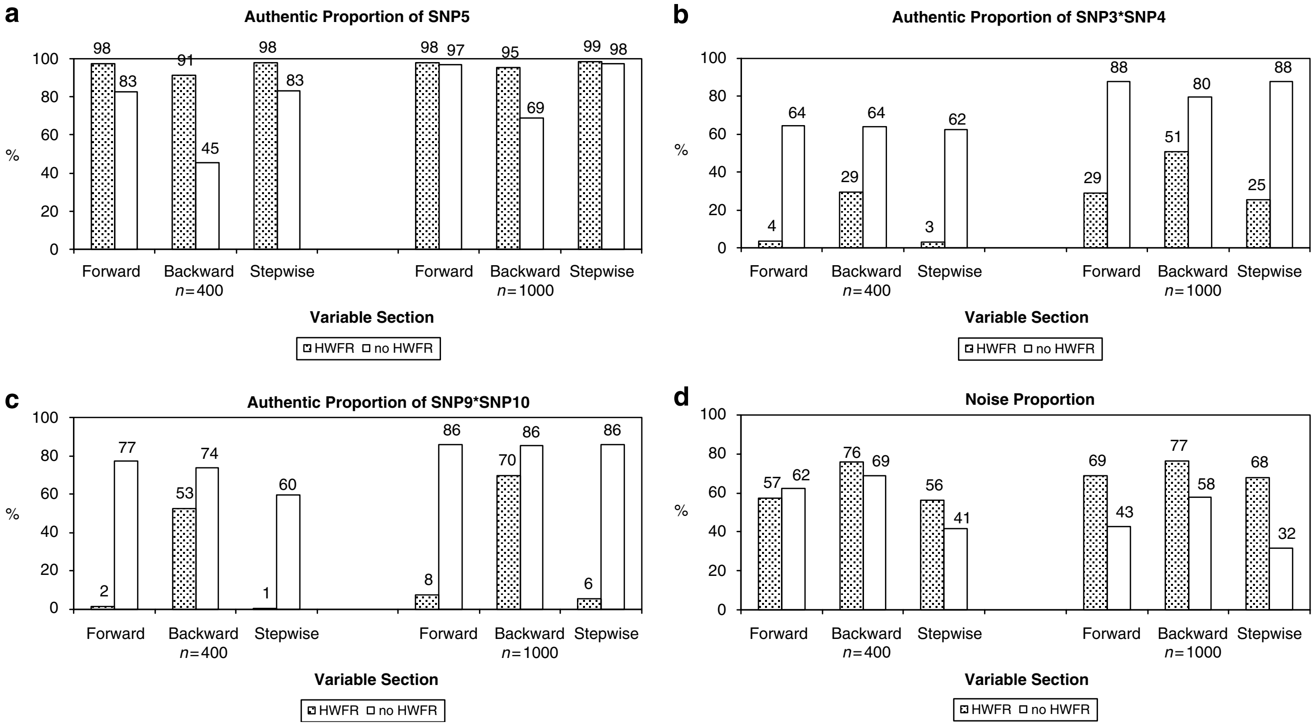

Figure 1 Comparison of stepwise, backward and forward selection and the Hierarchically Well-Formulated Rule (HWFR) in logistic regressions.

selection procedures. As shown in Figure 2, the AIC-based selection had not only the highest AP but also the highest NP among all three authentic terms. For 400 subjects, the AP of the AIC-based selection was even greater than 100%. The SC-based selection had the lowest AP and the lowest NP. Thus, the AIC-based selection was the most liberal procedure and the SC-based selection was the most conservative procedure. The S without the HWFR had the second highest AP and the second lowest NP. For S without the HWFR, the range of AP was 60–83% and the NP was 41% for a sample size of 400. For a sample size of 1000, the range of AP for S without the HWFR increased to 86–98% and the NP decreased to 32%.

Bootstrap analysis

To validate variable significance, a bootstrap method was applied in a given simulated data set. For this simulated data set, six significant terms were selected in the LR using stepwise selection without the HWFR. This bootstrap method successfully ranked the variables by the order of variable significance. As shown in Table 1, three authentic terms (SNP5, SNP3*SNP4, SNP9*SNP10) had the top three highest frequencies in 1000 bootstrap runs. This shows that bootstrap method is a useful method for validating variable significance.

Result of real data example: rheumatoid arthritis

In the RA control group, three strong linkages ($D' > 0.8$) were found in rs2076529 and rs2076530; rs7747393; and rs9285397; and rs394657 and rs422951. Thus, rs2076530, rs9285397 and rs422951 were excluded from the analyses. The remaining seven SNPs were in Hardy–Weinberg equilibrium. As shown in Table 2, two main effects (rs2395165 and rs7747393) and one 2-way interaction (rs9268480*rs2239804) were chosen for the stepwise selection and AIC-based selection. The SC-based selection procedure only selected the same 2-way interaction as the stepwise and AIC-based selections. The two-step procedure selected another 2-way interaction (rs2076529*rs2239804). The model selected from the stepwise without the HWFR, which was the same as the one based on the AIC-based selection, had the lowest value of AIC. The 2-way interaction rs9268480*rs2239804 was selected 412 times among 1000 LRs based on the bootstrap data. Therefore, the rs9268480*rs2239804 interaction had a higher chance to be truly associated with anti-cyclic citrullinated peptide antibody. This also demonstrates that the term with the highest chance to be true positive was included in the stepwise selection without the HWFR, which is this simulation study recommended method.

Table 1 Frequency of significant terms in 1000 bootstrap data

Term	*Frequency (selected/total)*
SNP3*SNP4	785/1000
SNP5	658/1000
SNP9*SNP10	596/1000
SNP4*SNP5	456/1000
SNP1*SNP3	358/1000
SNP7*SNP8	245/1000

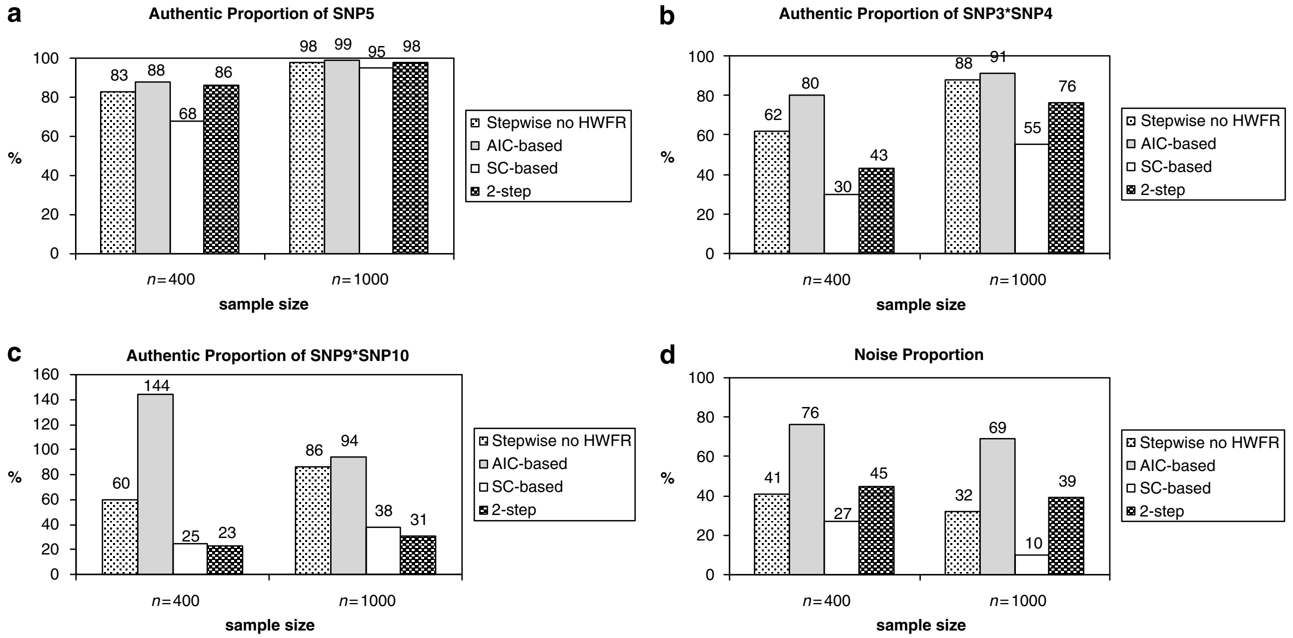

Figure 2 Comparison of stepwise without the Hierarchically Well-Formulated Rule (HWFR), AIC-based, SC-based and 2-step variable selection in logistic regressions.

Table 2 Result of variable selection in the Consortium for the Longitudinal Evaluation of African-Americans with Early Rheumatoid Arthritis study

		P-value			*Bootstrap %*
SNP (genotype grouping)[a]	*Stepwise, no HWFR*	*AIC-based*	*SC-based*	*Two-step*	*(1000 runs)*
rs2395165 (**TT**, CT, CC)	0.0392	0.0392	—	—	11.8
rs7747393 (**CC**, CT+TT)	0.0294	0.0294	—	—	8.8
rs9268480*rs2239804 (**CC**, CT+TT)*(**AA**, AG, GG)	<0.0001	<0.0001	0.0001	—	41.2
rs2076529*rs2239804 (**AA**, AG+GG)* (**AA**, AG, GG)	—	—	—	0.0003	15.6
AIC[b]	305.1	305.1	311.5	313.9	—
SC[c]	325.9	325.9	321.9	324.3	—

[a] Reference genotype is in bold.
[b] Akaike information criterion. The lower value the better the model-fit.
[c] Schwarz criterion. The lower value the better the model-fit.

Discussion

For testing interactions, the HWFR effect was obvious. The variable selection procedure without the HWFR can be more powerful in testing interactions with no or weak main effects. The procedure without the HWFR requires fewer terms in testing high-order interactions so it can accommodate more SNPs than the procedure with the HWFR. One more advantage of the selection without the HWFR is that the AP increases and NP decreases as the sample size increases.

In exploratory SNP association studies, we would prefer to have high AP and relatively low NP. Among forward, backward and stepwise automatic selection without the HWFR, all three methods had similar AP in detecting interactions, but the stepwise selection had the lowest NP and the backward selection had the highest NP. Thus, stepwise selection without the HWFR is the best among the six automatic selection procedures. Comparing stepwise, AIC-based, SC-based and 2-step selection, the AIC-based selection was the most liberal procedure and the SC-based selection was the most conservative procedure. The AIC-based selection was suggested for prediction, and the SC-based selection was suggested for description and interpretation.[23,31] The stepwise variable selection without the HWFR, which had the second highest AP and the second lowest NP, is a better method compared to other variable selection procedures.

Variable selection of ‘main effects’ in an LR has been discussed. The stepwise, forward and backward automatic selections in LR provide similar results in detecting main-effect covariates with the same entry and removal criterion. It was recommended that variables included in an LR should be nearly uncorrelated.[26] The variable selections in LR are very unstable and the percentage of noise variables is high.[27] Unlike the variable selection in main effects only, we observed that the automatic selection procedures performed differently in detecting interactions.

The stepwise selection without the HWFR, the best variable selection procedure among the testing procedures, had 30–40% noise terms in the model. This result is consistent with other studies for detecting main effect in LR.[20,21] Although AP was the primary focus for exploratory SNP studies, the high NP can interfere with researchers' judgment. The bootstrap method provides us a useful tool to rank and validate the variable significance so we may have an idea which terms may be true positive.

One must use caution when applying LR in testing SNP–SNP interactions. The limitations of variable selection in LRs for testing SNP–SNP interactions have been widely discussed.[10,22,28,29] The primary limitation is the empty cell-effect that some genotype combinations have low frequency or zero responses. In an LR, two dummy parameters are usually needed for each SNP and more parameters are required for their interactions. As the number of SNPs and/or the order of testing interaction increases, the number of genotype combinations increases and this empty-cell effect becomes severe. The maximum likelihood estimate of LR coefficients may be invalid because of the empty-cell effect. Secondly, the correlation between SNPs can weaken power of variable selection in an LR. Thirdly, the issue of genetic heterogeneity, in which an identical or similar phenotype is caused by different mutations, is not handled well by LR. The cluster analysis was recommended to detect genetic heterogeneity before detecting for SNP–SNP interactions.[30]

In this study, we can gain insight to the accuracy of the commonly used 2-step variable selection procedure. The APs in the commonly used 2-step selection were similar in testing main effects but were much lower in testing interactions compared with the stepwise selection without the HWFR. The NPs in the two-step selection were greater than the stepwise selection without the HWFR and the SC-based selection. As the number of SNPs increases, an appropriate two-step selection to narrow down the testing terms before using LR to test gene–gene or gene-environment interactions is necessary.[32] The common two-step selection we tested in this study has limited power for the interaction without main effects because the first stage only tests for main effects. Several two-stage methods have been proposed.[17,33]

In conclusion, this study provided empirical evidence for comparing variable selections in an LR. With limited candidate SNPs, stepwise selection without the HWFR is a better procedure for testing SNP–SNP interactions in an LR. The bootstrap method is useful for validating variable significance. Future research is needed for the issue of sample size and number of SNPs in variable selection of an LR and comparison of an LR with other statistical methods in testing SNP–SNP interactions.

Acknowledgements

We appreciate use of genotyping data from Dr David Reich of the Broad Institute and anti-cyclic citrullinated peptide antibody data from Dr Michael Holers of the University of Colorado. We thank Yung-Hsin Liu (INC Research) for programming support and Laura Gallitz (University of Alabama at Birmingham) for manuscript editing.

References

1 Van Emburgh BO, Hu JJ, Levine EA *et al*: Polymorphisms in drug metabolism genes, smoking, and p53 mutations in breast cancer. *Mol Carcinog* 2007; **47**: 88–99.

2 Ge D, Zhu H, Huang Y *et al*: Multilocus analyses of renin–angiotensin-aldosterone system gene variants on blood pressure at rest and during behavioral stress in young normotensive subjects. *Hypertension* 2007; **49**: 107–112.

3 Gu D, Su S, Ge D *et al*: Association study with 33 single-nucleotide polymorphisms in 11 candidate genes for hypertension in Chinese. *Hypertension* 2006; **47**: 1147–1154.

4 Cook NR, Zee RY, Ridker PM: Tree and spline-based association analysis of gene–gene interaction models for ischemic stroke. *Stat Med* 2004; **23**: 1439–1453.

5 Ritchie MD, Hahn LW, Roodi N *et al*: Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2001; **69**: 138–147.

6 Nelson MR, Kardia SL, Ferrell RE, Sing CF: A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res* 2001; **11**: 458–470.

7 Ripley B: *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press, 1996.

8 Bureau A, Dupuis J, Falls K *et al*: Identifying SNPs predictive of phenotype using random forests. *Genet Epidemiol* 2005; **28**: 171–182.

9 Hoffjan S, Ostrovnaja I, Nicolae D *et al*: Genetic variation in immunoregulatory pathways and atopic phenotypes in infancy. *J Allergy Clin Immunol* 2004; **113**: 511–518.

10 Bae Y, Park C, Han J *et al*: Interaction between GNB3 C825T and ACE I/D polymorphisms in essential hypertension in Koreans. *J Hum Hypertens* 2007; **21**: 159–166.

11 Miao X, Zhang X, Zhang L *et al*: Adenosine diphosphate ribosyl transferase and x-ray repair cross-complementing 1 polymorphisms in gastric cardia cancer. *Gastroenterology* 2006; **131**: 420–427.

12 Zhang X, Miao X, Guo Y *et al*: Genetic polymorphisms in cell cycle regulatory genes MDM2 and TP53 are associated with susceptibility to lung cancer. *Hum Mutat* 2006; **27**: 110–117.

13 Huber A, Keck CC, Hefler LA *et al*: Ten estrogen-related polymorphisms and endometriosis: a study of multiple gene–gene interactions. *Obstet Gynecol* 2005; **106**: 1025–1031.

14 Marchini J, Donnelly P, Cardon LR: Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* 2005; **37**: 413–417.

15 Coffey CS, Hebert PR, Ritchie MD *et al*: 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* 2004; **5**: 49.

16 Oestergaard MZ, Tyrer J, Cebrian A *et al*: Interactions between genes involved in the antioxidant defence system and breast cancer risk. *Br J Cancer* 2006; **95**: 525–531.

17 Hoh J, Wille A, Zee R *et al*: Selecting SNPs in two-stage analysis of disease association data: a model-free approach. *Ann Hum Genet* 2000; **64**: 413–417.

18 Zee RY, Hoh J, Cheng S *et al*: Multi-locus interactions predict risk for post-PTCA restenosis: an approach to the genetic analysis of common complex disease. *Pharmacogenomics J* 2002; **2**: 197–201.

19 Park MY, Hastie T: Penalized logistic regression for detecting gene interactions. *Biostatistics* 2008; **9**: 30–50.

20 Derksen S, Keselman HJ: Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *Br J Math Stat Psychol* 1992; **45**: 265–282.

21 Zellner D, Keller F, Zellner GE: Variable selection in logistic regression models. *Communications in Statistics: Simulation and Computation* 2004; **33**: 787–805.

22 Kleinbaum DG: *Logistic Regression: A Self-Learning Text*. New York: Springer, 1994.

23 Shtatland ES, Cain E, Barton MB: The perils of stepwise logistic regression and how to escape them using information criteria and the output delivery system. *SAS Users Group International* 2001; **26**: 222.

24 Akaike H: A new look at the statistical model identification. *IEEE Transactions on Automatic Control AC* 1974; **19**: 716–723.

25 Schwarz G: Estimating the dimension of a model. *Ann Stat* 1978; **6**: 461–464.

26 Efron B, Tibshirani R: *An Introduction to the Bootstrap*. Boca Raton: CHAPMAN & HALL/CRC, 1993.

27 Sun GW, Shook TL, Kay GL: Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *J Clin Epidemiol* 1996; **49**: 907–916.

28 Mantel N: Why step-down procedures in variable selection. *Technometrics* 1970; **12**: 621–625.

29 Kallberg H, Padyukov L, Plenge RM *et al*: Gene–gene and gene-environment interactions involving HLA-DRB1, PTPN22, and smoking in two subsets of rheumatoid arthritis. *Am J Hum Genet* 2007; **80**: 867–875.

30 Bridges Jr SL, Hughes LB, Mikuls TR *et al*: Early rheumatoid arthritis in African-Americans: the CLEAR Registry. *Clin Exp Rheumatol* 2003; **21**: S138–S145.

31 Burnham KP, Anderson DR: *Model Selection and Multimodel Inference: A Practical Information-Theorectic Approach*, 2nd edn. New York: Springer, 2002.

32 Heidema AG, Boer JM, Nagelkerke N, Mariman EC, van der AD, Feskens EJ: The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases. *BMC Genet* 2006; **7**: 23.

33 Skol AD, Scott LJ, Abecasis GR, Boehnke M: Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 2006; **38**: 209–213.