J Hum Genet (2008) 53:802–811
DOI 10.1007/s10038-008-0313-z

ORIGINAL ARTICLE

Comparison of multivariate adaptive regression splines and logistic regression in detecting SNP–SNP interactions and their application in prostate cancer

Hui-Yi Lin · Wenquan Wang · Yung-Hsin Liu · Seng-Jaw Soong · Timothy P. York · Leann Myers · Jennifer J. Hu

Received: 1 November 2007 / Accepted: 5 June 2008 / Published online: 8 July 2008

Abstract Single nucleotide polymorphism (SNP) interaction plays a critical role for complex diseases. The primary limitation of logistic regressions (LR) in testing SNP–SNP interactions is that coefficient estimates may not be valid because of numerous terms in a model. Multivariate adaptive regression splines (MARS) have useful features to effectively reduce the number of terms in a model. To study how MARS can address these drawbacks possibly better than LR, the power of MARS and LR with SNPs using the reference-coding and additive-mode scheme was compared using simulated data of ten SNPs for 400 subjects based on 1,000 replications for five interaction models. In overall scenarios, MARS performed better than LR. In the model with a dominant two-way interaction, the power range was 76–96% for MARS and 1–8% for LR in both coding schemes. In the dominant three-way interaction model, the power was 57–85% for MARS and less than 4% for LR. In the prostate cancer example, we evaluated the association between ten SNPs and prostate cancer risk in 649 Caucasians. The best model with one two-way and one three-way interaction was selected using MARS. The findings supported that MARS may provide a useful tool for exploring SNP–SNP interactions.

Keywords Epistasis · Gene–gene interaction · Multivariate adaptive regression splines · MARS · Simulation · Power

Introduction

Identifying genetic factors for complex diseases, such as hypertension, asthma, or cancer, is one of the primary goals of human geneticists. Gene–gene and gene–environment interaction associated with diseases has been discussed recently. In this study, we used the term single nucleotide polymorphism–single nucleotide polymorphism (SNP–SNP) interaction instead of gene–gene interaction because several SNPs with interactions may be in the same gene, and we not only evaluate interactions between genes but also within a gene. Although SNP–SNP interaction detection is conceptually expected to play an important role in defining risk groups for complex diseases (Smith et al. 2002, 2003; Moore 2003; Lin et al. 2006; Hu et al. 2007), the identification of SNP–SNP interaction has been limited, with the majority of studies focusing on identifying the additive effect of SNPs, especially for genome-wide studies with a large number of SNPs (Scuteri et al. 2007; Tomlinson et al. 2007). Identification of such interactions remains difficult because of weak or no marginal effects of

H.-Y. Lin · W. Wang · S.-J. Soong
Medical Statistics Section,
University of Alabama at Birmingham,
Birmingham, AL, USA

Y.-H. Liu
INC Research, Raleigh, NC, USA

T. P. York
Department of Human Genetics,
Virginia Commonwealth University,
Richmond, VA, USA

L. Myers
Department of Biostatistics, Tulane University,
New Orleans, LA, USA

J. J. Hu (✉)
Sylvester Comprehensive Cancer Center
and Department of Epidemiology and Public Health,
University of Miami School of Medicine,
Miami, FL, USA
e-mail: jhu@med.miami.edu

some SNPs, a large number of SNPs to consider, or lack of a priori information about which SNPs interact.

Commonly used case-control methods [i.e., logistic regression (LR)] for gene identification may lack the flexibility to overcome these difficulties. For instance, unconditional LR is typically used to test the association between potential SNPs and a binary outcome, such as “diseased or nondiseased.” As the number of SNPs increases and interactions are taken in to consideration, the number of terms needed in an LR model to statistically describe all possible k-way SNP–SNP interactions of n biallelic loci increases dramatically in a form of $(n!/k!(n-k)!) \times 2^k$ (Wade 2000). Thus, the LR approach usually suffers the data configuration of quasi-complete separation (Albert and Anderson 1984), where not all response levels exist in each covariate combination. The quasi-complete separation may cause invalid estimates of coefficient and an unusually large standard error estimate because the coefficient with quasi-complete separation is theoretically infinite (Webb et al. 2004). In this study, we simply called this effect “empty-cell effect.” When the empty-cell effect exists, the true SNP association may be distorted.

Several statistical methods have been proposed to deal with SNP–SNP interactions, such as multivariate adaptive regression splines (MARS) (Friedman 1991; Cook et al. 2004), multifactor dimensionality reduction (MDR) (Ritchie et al. 2001), combinational partitioning method (CPM) (Nelson et al. 2001), artificial neural networks (ANN) (Veaux et al. 1993), classification and regression trees (CART) (Breiman et al. 1984), and random forests (Bureau et al. 2005). MARS is considered the most flexible compared with CART and traditional LR (Cook et al. 2004), and it has performed better than ANN. In addition, MARS is more powerful than least squares curve fitting using polynomials in testing gene–environmental interactions (York et al. 2006). Several studies have used MARS for detecting SNP–SNP interactions in prostate cancer, breast cancer, ischemic stroke, and hypertension (York and Eaves 2001; Cook et al. 2004; Gu et al. 2006; Lin et al. 2006; Ge et al. 2007; Van Emburgh et al. 2008; Zabaleta et al. 2008).

MARS (Friedman 1991) is an automated and flexible data-mining tool that combines the advantages of recursive partitioning (Morgan and Sonquist 1963; Breiman et al. 1984) and spline fitting (De Boor 1978). MARS provides useful features to overcome the limitations of LR in exploring SNP–SNP interactions. MARS can automatically select and transform variables and can identify potential interactions (2001). These features are useful for SNP data analysis. The mode of inheritance (dominant, recessive, and additive) for SNPs and their interactions also can be determined automatically, so the number of parameters in modeling can be dramatically reduced. In addition to detecting which SNP is involved in an interaction, MARS can also detect interaction combinations (or patterns) that can be used to define risk groups. For example, eight parameters are required to present a three-way interaction using LR, but MARS may only need one parameter of high- vs. low-risk subgroup. The outcome variable for MARS can be binary or continuous, and the covariates can be categorical and continuous. Thus, it can be applied for various types of studies, such as gene expression and gene–environment interaction.

Even though several studies have been conducted to compare MARS and other methods using real data sets (Cook et al. 2004; Gu et al. 2006), the power of MARS in assessing SNP–SNP interactions is unknown. Empirical evidence of LR power is also limited despite the well-known disadvantages of LR in detecting SNP–SNP interactions. The objectives of this study were: (1) to compare the power of MARS and LR to detect SNP–SNP interactions for binary outcomes for multiple scenarios; (2) to apply MARS and LR to a real data example of prostate cancer.

Materials and methods

Simulations

To compare the power of MARS and LR, we generated case-control data sets with 400 subjects (200 cases/200 controls) with nonmissing genotypes for ten SNPs. We assume no linkage disequilibrium for the ten SNPs. These SNPs were generated independently based on the Hardy–Weinberg equilibrium with major allele proportions 0.5, 0.75, or 0.9. In the two-way interaction models (Models 1–4), two SNPs (SNP_A and SNP_B) contributed to disease risk. The major allele proportions of SNP_A and SNP_B are $P(A) = 0.5$ and $P(B) = 0.75$. The disease outcomes were generated based on penetrance for the two functional SNPs, which is the conditional probability of disease given the genotype, shown in Table 1. The penetrance in the risk cell (PEN_r) was set to be 0.15, 0.3, or 0.5, and the penetrance in the low-effect cell was equal to 0.01. As shown in Table 1, four different types of two-way interactions for the two functional SNPs associated with the disease outcome were evaluated. Model 1 and Model 2 both had a dominant–dominant interaction but with different disease alleles. The disease alleles in Model 1 were major alleles (A and B) and in Model 2 were minor alleles (a and b). The high-risk genotype combinations in Model 3 were those containing at least one of the aa and bb genotypes. Model 4 was simulated for a dominant–recessive interaction. In Model 1, $P(D|AABB) = P(D|AaBB) = P(D|AABb) = P(D|AaBb) =$

Table 1 Penetrances in the two-way interaction models

	BB	Bb	bb
Model 1			
AA	PEN_r	PEN_r	0.01
Aa	PEN_r	PEN_r	0.01
aa	0.01	0.01	0.01
Model 2			
AA	0.01	0.01	0.01
Aa	0.01	PEN_r	PEN_r
aa	0.01	PEN_r	PEN_r
Model 3			
AA	0.01	0.01	PEN_r
Aa	0.01	0.01	PEN_r
aa	PEN_r	PEN_r	PEN_r
Model 4			
AA	PEN_r	0.01	0.01
Aa	PEN_r	0.01	0.01
aa	0.01	0.01	0.01

$P(\text{A}) = 0.5$, $P(\text{B}) = 0.75$. PEN_r (penetrance in the risk cell) = 0.15, 0.3 or 0.5

PEN_r and $P(D|$ other low-effect cells$) = 0.01$, where D represents the subject is affected; A and B denote major alleles; and a and b denote minor alleles. Let "low-effect cells" represent the genotype combinations with penetrances of 0.01.

The power of Model 5, with a three-way dominant interaction, was also evaluated. In this three-way model, the major allele proportions of SNP_A, SNP_B, and SNP_c are $P(\text{A}) = 0.5$, $P(\text{B}) = 0.75$ and $P(\text{C}) = 0.5$. $P(D|\text{AABBCC}) = P(D|\text{AABBCc}) = P(D|\text{AABbCC}) = P(D|\text{AABbCc}) = P(D|\text{AaBBCC}) = P(D|\text{AaBBCc}) = P(D|\text{AaBbCC}) = P(D|\text{AaBbCc}) = \text{PEN}_r$ and $P(D|$ other low-effect cells$) = 0.01$. PEN_r was 0.15, 0.3 or 0.5. Data simulation and analyses were performed using SAS 9.1 (simulation and LR) and MARS 2.0. Power was calculated based on 1,000 replications for each condition.

Logistic regression variable selection

Two methods were used to parameterize SNP in LR in this study. First, SNP was treated as an additive mode, which is a continuous variable. For example, 0, 1, and 2 were applied to the AA, Aa, and aa genotypes, respectively. Second, each SNP was treated as a categorical variable using the reference-coding scheme with the major homozygous genotype as the reference group. In the reference-coding scheme, two degrees of freedom (DF) are required for each SNP. The advantage of treating an SNP with an additive mode is reducing the number of parameters required for modeling. However, the additive mode assumption may not be applicable to some situations. The stepwise selection of LR with entry and removal criteria at p value 0.05 was used.

In the LR modeling, we did not apply the hierarchical restriction, which requires including all lower order terms of the highest order interaction term in the model, regardless of their statistical significance. A growing number of studies showed that some SNP–SNP interactions exist without marginal effects (Culverhouse et al. 2002; Moore and Williams 2002; Musani et al. 2007). In addition, the stepwise selection without hierarchical restriction in LR has been shown to have higher true positive and lower false positive findings compared with other commonly used variable selection procedures in LR to detect SNP–SNP interactions (Lin et al. 2008). The two-way interaction models without main effects using SNPs as categorical variables are as follows, and the reference group for the interactions is the combination of AABB, AABb, AAbb, AaBB, and aaBB (Table 1).

$$\begin{aligned}\log\left(\frac{p_i}{1-p_i}\right) &= b_0 + b_1 I_i(\text{SNP}_\text{A} = \text{Aa}) \times I_i(\text{SNP}_\text{B} = \text{Bb}) \\ &+ b_2 I_i(\text{SNP}_\text{A} = \text{aa}) \times I_i(\text{SNP}_\text{B} = \text{Bb}) \\ &+ b_3 I_i(\text{SNP}_\text{A} = \text{Aa}) \times I_i(\text{SNP}_\text{B} = \text{bb}) \\ &+ b_4 I_i(\text{SNP}_\text{A} = \text{aa}) \times I_i(\text{SNP}_\text{B} = \text{bb})\end{aligned}$$

where $i = 1, 2, \ldots, n$ (=400). Let p_i denote the proportion of disease (event) and

$$I_i(\text{SNP}_m = W) = \begin{cases} 1 & \text{if} \quad \text{SNP}_m = W \\ 0 & \text{if} \quad \text{SNP}_m = \text{other genotype(s)} \end{cases}$$

for subject i.

MARS variable selection

The primary unit of the MARS modeling method is the basis function (BF). Unlike conventional modeling, MARS does not select a reference group for each potential predictor in advance. BFs represent the information of one or more variables. The example of the MARS result for Model 1 is as follows.

$$\begin{aligned}&\text{BF2} = (\text{SNP}_\text{A} = \text{AA or SNP}_\text{A} = \text{Aa}) \\ &\text{BF25} = (\text{SNP}_\text{B} = \text{BB}) \times \text{BF2} \\ &\text{BF27} = (\text{SNP}_\text{B} = \text{Bb}) \times \text{BF2} \\ &\quad Y = 0.158236\text{E - }06 + 0.607 \times \text{BF25} + 0.614 \times \text{BF27}.\end{aligned}$$

Y represents the binary phenotype. BF2 represents the dummy variable of SNP_A with AA/Aa = 1 and aa = 0 (dominant), and BF25 represents the dummy variable of the SNP_A and SNP_B combination with "AA/Aa and BB" = 1 and "other combinations with these two SNPs" = 0. BF27 represents the dummy variable with "AA/Aa and Bb" = 1 and “others” $= 0$. Thus, this model successfully selected the designated dominant–dominant interaction.

The strategy of variable selection in MARS is first to overfit a model by performing a forward-stepping search and then to prune it by dropping BFs that contribute the least through a backward deletion process. Each backward step is examined by generalized cross validation (GCV), a criterion for measuring generalized mean square errors. The best MARS model contains the lowest GCV. Researchers can determine the order of interactions for testing, and several parameters can be used to control the selection process. The maximum number of BFs, a control parameter in MARS, is used to control the size of the overfitted model. The guideline for the maximum number of BFs is at least two to four times the size of the “truth,” in accordance with the MARS user’s guide (2001). We allowed for a maximum of 70 BFs, which was large enough for our simulated models.

The final MARS model is determined by the DF penalty applied to BFs. Using a higher DF penalty, a smaller final model is selected. The DF penalty can be manually designated by users or can be automatically estimated by a cross-validation procedure. In this study, a tenfold cross-validation procedure was applied. A model was built using nine tenths of the data (training set), and the remaining one tenth (test set) was used to test this model. This process was repeated ten times to allow each one tenth as the test set and decide the best DF penalty. The set of dummy variables, which represents the combination of levels of the predictors, displayed in the form of BFs in MARS, may not be mutually exclusive. MARS automatically selects them based upon model improvement (2001).

Model evaluation and power calculation

The input data contained a binary outcome and ten SNPs. These SNPs were treated as categorical variables in MARS. In LR, the reference-coding and additive-mode scheme were used. The same coding scheme was applied for all SNPs in the same LR model. All main effects and interactions up to the designated way of interaction were considered, and the final model was selected using the above variable selection procedures of LR and MARS. The power was calculated as the proportion of detecting the designated interaction among 1,000 replicates. In LR, we consider the true interaction was detected if the p value of the Wald test for the designated interaction was less than 0.05. For consistency, any type of the designated interactions selected by MARS was counted, although MARS can detect specific interaction patterns. To evaluate the effects of empty cells on LR with the reference-coding scheme, power was also calculated, stratified by the empty-cell status of the designated interaction.

Real data example: prostate cancer risk

Prostate cancer is the most common cancer in American men. We applied MARS and LR to the data set from a study of prostate cancer risk among a Caucasian population. The details of the study population and the eligibility criteria were described previously (Hu et al. 2004; Hu 2006). We tested ten nonsynonymous SNPs (nsSNPs) in nine deoxyribonucleic acid (DNA)-repair genes of four repair pathways, including: (1) base excision repair (BER): *ADPRT V762A* and *XRCC1 R399Q*; (2) nucleotide excision repair (NER): *ERCC2 D312N/K751Q*, *ERCC5 D1103H*, and *XPC A499V*; (3) mismatch repair (MMR): *MLH1 I219V* and *MSH3 R940Q*; and (4) double-strand-break repair (DSBR): *NBS Q185E* and *XRCC3 T241M*.

The same parameterization described in the previous section was applied. For the stepwise selection in LR, liberal entry and removal criteria $p = 0.1$ were applied. To thoroughly search for interactions, several MARS model parameters were applied. The maximum BFs of 70 or 100 were used to control the size of the overfitted model. Then, tenfold cross validation or three DF per BF were applied to select the final MARS model. In the control group, the Hardy–Weinberg equilibrium was evaluated for all SNPs using both chi-square and exact tests. Linkage disequilibrium (LD) among the ten SNPs was evaluated using Lewontin’s D'. We tested for up to three-way SNP–SNP interactions for prostate cancer risk (positive vs. negative) among Caucasian participants. The controlling factors included age, family history (yes/no), smoking history (ever smoked at least 100 cigarettes in lifetime), and history of benign prostatic hyperplasia (yes/no). A total of 649 Caucasians with 360 cases and 289 controls had the complete data for the ten SNPs and four controlling factors.

The objective of using this example was to evaluate the associations between SNPs and prostate cancer risk after adjusting for covariates. We used the stepwise LR with forcing the above four covariates to be in the model. MARS does not have a function to force specific covariates in the model. To adjust for the four covariates described above in MARS, an LR with these covariates was conducted and the residuals of this LR were used as an outcome variable in MARS. We applied LR using the terms selected from the final MARS model to calculate odds ratio (OR). To validate variable significance, a bootstrap method with 1,000 runs was applied to LR and MARS for testing up to three-way interactions.

Results

Simulation results

The power of MARS and LR in detecting one two-way or three-way SNP–SNP interaction among ten candidate SNPs is presented in Table 2. For MARS, the power range for Model 1 containing a dominant–dominant interaction with major alleles (A and B) as disease alleles was 74–97%, and the power for Models 2–4 was close to 100%. Using the reference-coding scheme, the power of LR was quite low, especially for Model 1 (<2%). The power of LR for Model 2, which contained a dominant–dominant interaction with minor alleles (a and b) as disease alleles, was 50–78%. The power of LR for Model 3, which contained a two-way interaction with at least one of the aa and bb genotypes, was 61–73%. The power of LR for Model 4, with a dominant–recessive interaction, was 23–44%. The power of LR with the additive-mode scheme was generally higher than that with the reference-coding scheme in this study. The power of LR with the additive-mode scheme was the lowest in Model 1 (<8%) and was the highest in Model 3 (85–99%). In Model 5, with a three-way dominant interaction, the power of MARS was 57–85%; however, the power of LR in both coding schemes was low (<4%).

Among all five interaction models, both MARS and LR had the lowest power in detecting the dominant SNP–SNP interaction in Model 1 and Model 5. LR with the reference-coding scheme had the highest power in Model 2. In general, the power of MARS and LR with the additive-mode scheme increased, as expected, as the PEN_r increased. The power of LR with the reference-coding scheme in Model 3 also increased, whereas the power in other models decreased as PEN_r increased.

Why did the power of some LRs decrease as PEN_r increased? To answer this question, the power of LR with the reference-coding scheme was obtained by stratifying based on empty-cell status of the designated interaction (SNP_A–SNP_B or SNP_A–SNP_B–SNP_c). As shown in Table 3, the empty-cell proportion in the designated interaction, which is the proportion of at least one empty cell in 3 × 3 or 3 × 3 × 3 combination cells, increased as PEN_r increased. Among the two-way interaction models, Model 1 had the highest empty-cell proportions (35–88%) and Model 3 had the lowest ones (2–18%). The range of the empty-cell proportions in Model 5 with a three-way interaction was 67–98%. As we expected, the power of LR to detect SNP_A–SNP_B without an empty cell was much higher than that with at least one empty cell. Therefore, only the power of Model 3 increased as PEN_r increased in LRs with the reference-coding scheme. In Model 5, the power was zero for the designated interaction without an empty cell because of the limited number of simulated runs.

Table 2 Power for multivariate adaptive regression splines (MARS) and logistic regression (LR) to detect the specified single nucleotide polymorphism–single nucleotide polymorphism (SNP-SNP) interactions

Model	PEN_r[a]	Power %		
		MARS	Logistic regression	
			Reference coding[b]	Additive mode[c]
1	0.15	74.2	1.8	5.5
	0.3	85.0	1.7	4.4
	0.5	96.5	1.3	7.7
2	0.15	99.5	78.0	57.0
	0.3	99.9	68.0	73.4
	0.5	100.0	50.0	89.0
3	0.15	97.0	61.4	85.2
	0.3	99.5	70.3	97.3
	0.5	100.0	72.9	98.9
4	0.15	99.3	44.1	49.9
	0.3	100.0	35.2	44.5
	0.5	100.0	23.4	44.7
5 (three-way)	0.15	57.2	0.4	2.5
	0.3	76.2	0.3	1.9
	0.5	85.4	0.1	3.2

[a] Penetrance in the risk cell

[b] SNP was treated as a categorical variable with the major homozygous genotype as the reference group

[c] SNP was treated as a continuous variable with values of 0, 1, and 2

Results of a real-data example: prostate cancer risk

In the control group, all ten SNPs followed the Hardy–Weinberg equilibrium, and no strong pair-wise linkage disequilibrium ($D' > 0.8$) was found. The results are shown in Table 4. In testing the association between each SNP and prostate cancer risk using LR with the reference-coding scheme, Caucasians with *ERCC2 312 DN* and *NN* (heterozygous and variant type) had lower prostate cancer risk compared with ones with *ERCC2 312 DD*. The stepwise selection for up to three-way interactions in LR with the reference-coding scheme also achieved the same result. The same main effect (*ERCC2 312*) was selected in the univariate and stepwise selection in LR with the additive-mode scheme for detecting up to two-way interactions. When testing up to three-way interactions, *ERCC2 312–MSH3 940–ERCC2 751* was selected.

The MARS one-way model was the same as the one selected from LRs with the reference-coding scheme. For testing up to two-way interactions in MARS, we observed that individuals with the genotype combination of *ERCC2 312 DN/NN* and *MSH3 940 RR* had lower prostate cancer

Table 3 Power of logistic regression (LR) with the reference-coding scheme by empty-cell status of the designated interaction

Model	PEN_r[a]	Empty-cell % in the designated interaction[b]	Power %		
			Overall	Empty cell in the designated interaction[b]	
				Yes	No
1	0.15	34.9	1.8	0.3	2.6
	0.3	66.5	1.7	0.9	3.3
	0.5	88.3	1.3	0.7	6.0
2	0.15	3.9	78.0	2.6	81.1
	0.3	22.5	68.0	8.0	85.4
	0.5	48.4	50.0	8.7	88.8
3	0.15	2.3	61.4	4.3	62.7
	0.3	7.7	70.3	27.3	73.9
	0.5	17.8	72.9	37.6	80.5
4	0.15	21.1	44.1	30.3	47.8
	0.3	53.8	35.2	24.9	47.2
	0.5	78.9	23.4	17.7	44.5
5 (three-way)	0.15	66.5	0.4	0.3	0.5
	0.3	93.4	0.3	0.3	0
	0.5	98.3	0.1	0.1	0

[a] Penetrance in the risk cell

[b] $SNP_A \times SNP_B$ in Models 1–4, $SNP_A \times SNP_B \times SNP_C$ in Model 5

Table 4 Model comparison of prostate cancer risk in Caucasians

Method	Variable	Adjusted OR (95% CI)[a]	*P* value	AIC[b]	BIC[c]	Bootstrap (selected/total)
Univariate LR (R)						
Stepwise LR (R), up to three-way	*ERCC2 312* (*DN* vs. *DD*)	0.69 (0.49–0.97)	0.035	873.7	905.0	(LR: R) 223/1,000
	(*NN* vs. *DD*)	0.44 (0.27–0.72)	0.001			
MARS, one-way						(MARS) 249/1,000
Univariate LR (A)	*ERCC2 312*	0.67 (0.53–0.84)	<0.001	871.8	898.6	275/1,000
Stepwise LR (A), up to two-way						
Stepwise LR (A), up to three-way	*ERCC2 312*	0.46 (0.32–0.66)	<0.001	866.8	898.1	275/1,000
	ERCC2 312 × *MSH3 940* × *ERCC2 751*	1.09 (1.02–1.17)	0.010			92/1,000
MARS, up to two-way	(*ERCC2 312 DN/NN* + *MSH3 940 RR*) vs. others	0.56 (0.41–0.78)	<0.001	871.6	898.4	161/1,000
MARS, up to three-way	(*ERCC2 312 DN/NN* + *MSH3 940 RR*) vs. others	0.60 (0.43–0.84)	0.003	863.3	894.6	161/1,000
	(*ERCC2 312 DD* + *XPC 499 AA* + *XRCC1 399 QQ*) vs. others	6.99 (1.59–30.85)	0.010			220/1,000

LR logistic regression, *R* reference-coding scheme, *A* additive-mode scheme, *MARS* multivariate adaptive regression splines

[a] Odds ratio and 95% confidence interval for adjusting for age, family history, smoking history, and history of benign prostatic hyperplasia

[b] Akaike information criterion

[c] Bayesian information criterion

risk [OR = 0.56, 95% confidence interval (CI) = 0.41–0.78]. A model containing a two-way and a three-way interaction was detected using up to 70 BFs and three DF per BF. The interaction selected in the two-way MARS model was also included, and one three-way interaction was detected. The genotype combination of *ERCC2 312 DD*, *XPC 499 AA* and *XRCC1 399 QQ* is associated with a significantly higher prostate cancer risk (OR = 6.99, 95% CI = 1.59–30.85).

Akaike information criterion (AIC) (Akaike 1974) and Bayesian information criterion (BIC) (Schwarz 1978), which are the common model selection criteria in LR, were used to select the final model. The lower the criterion value, the better the model. Based on the lower value of both criteria, the MARS three-way model is better than other models. Among 1,000 bootstrap data sets, *ERCC2 312* was the most commonly selected term (223–275 out of 1,000). The three-way interaction *ERCC2 312 DD–XPC 499 AA–XRCC1 399 QQ* selected from MARS also had a relatively high frequency (220 out of 1,000) to be associated with prostate cancer risk. However, the three-way interaction detected by using LR with the additive-mode scheme was rarely selected in the bootstrap data sets.

To present both disease distribution and ORs in the specific genotype combinations, the final MARS model can be displayed in a tree plot, as shown in Fig. 1. This example demonstrated that MARS can effectively reduce data dimensionality. Only two parameters were needed in the model that contained one two-way and one three-way interaction without specific inherent mode assumption. This example shows that MARS is more powerful than LR in detecting SNP–SNP interactions. It should be noted that SNP–SNP interactions we found here were data driven. More studies with larger sample size are needed to confirm our novel findings.

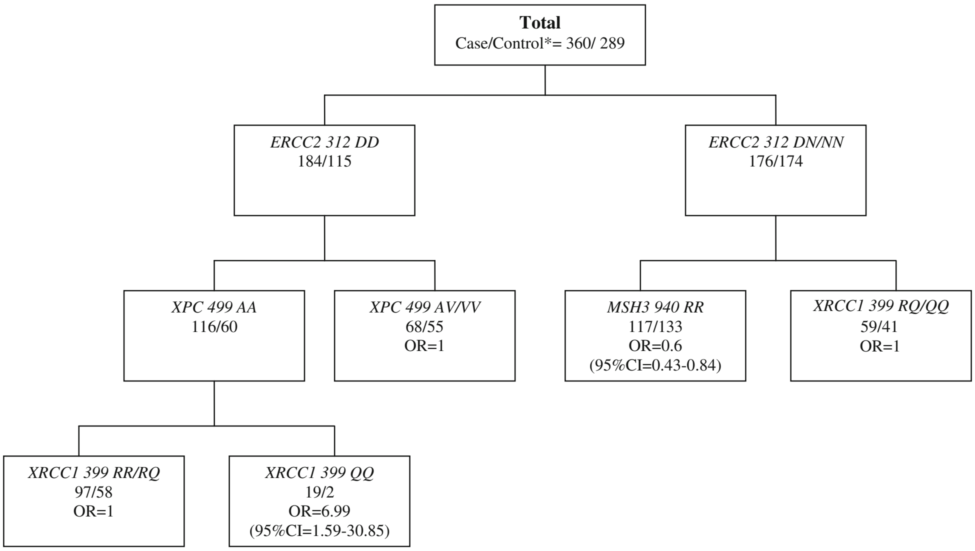
Discussion

MARS may overcome some limitations of LR and was demonstrated to be more powerful in detecting SNP–SNP interactions. The comparison of MARS and LR for detecting SNP–SNP interactions is summarized in Table 5, and the strengths of MARS are listed as follows. MARS can be applied in various types of studies because of the flexibility of outcome and covariates. The useful features of MARS in detecting SNP–SNP interactions include flexible reference group selection, automatic genotype combination, and automatic interaction pattern detection. As with other traditional modeling, MARS can include multiple terms (main effects and interactions) in a model simultaneously, and genetic interactions can be evaluated after adjusting for potential confounding factors. This study shows that empty-cell effect has a minor impact on MARS compared with LR with the reference-coding scheme. In addition, MARS is not restricted by the hierarchical rule. As for result interpretation, the terms selected from MARS can be easily converted into a logistic model, so the final model can be displayed by a logistic model. Thus, the straightforward result interpretation and comprehensive model diagnosis method in LR can be applied to MARS.

The power of modeling to detect SNP–SNP interactions depends on experiment design (Gauderman 2002), sample size, genotype frequencies determined by allele proportions, penetrance contrast between the comparative and reference groups, interaction patterns, and variable parameterization in modeling. In general, the larger sample size and penetrance contrast between the risk and low-effect cells, the higher the chance that the interaction can be detected. In addition, some interaction patterns tend to have lower power to be detected, such as a model containing cells with low genotype frequencies and low penetrances.

We can gain insight by comparing the power between MARS and LR though the five interaction models. Both MARS and LR had the lowest power to detect the two-way (Model 1) or three-way (Model 5) SNP–SNP interaction, which had a dominant–dominant interaction with major alleles as the disease alleles, compared with the other three models. The possible reason is that all the low-penetrance

Fig. 1 Multivariate adaptive regression splines (MARS) model of prostate cancer risk for Caucasians



*: Frequency of Case/ Control for complete data in the model
OR adjusted for age, family history, smoking history and BPH

Table 5 Comparison of logistic regression and multivariate adaptive regression splines (MARS) for detecting single nucleotide polymorphism–single nucleotide polymorphism (SNP–SNP) interactions

	Logistic regression	MARS
Outcome	Binary	Binary or continuous
Covariate	Categorical and continuous	Categorical and continuous
Automatically detect interaction patterns to define risk/protective subgroup	No	Yes
Automatically categorized an SNP into appropriate mode of inheritance	No	Yes
SNP parameterization		
Predefined reference group (reference coding)	Yes	No
Additive-mode assumption (additive mode)	Yes	No
Interference from the empty-cell effect	Severe (reference coding)	Minor
	Minor (additive mode)	
Number of parameters for an SNP, using a three-way interaction as an example	Up to eight (reference coding); one (additive mode)	May only need one parameter for high-risk group vs. others

cells had low genotype frequencies, so the number of cases in these cells was low or closes to zero. This enlarged the standard errors of the model parameters that related to these cells, so the power of Model 1 and Model 5 was the lowest among the testing models.

Unlike traditional modeling, MARS does not need to preselect a reference group for categorical covariates. The reference group selection for each SNP in MARS is automatic based on model improvement. Inappropriate reference group selection due to modeling may diminish the true magnitude of penetrance contrast between the risk and low-effect groups. Because of the flexibility of MARS in selecting the reference group, the penetrance contrast between the reference and comparison group is close to the true contrast between the risk and low-effect groups. In LR with the reference-coding scheme, the reference group was preselected and fixed. Among these four two-way interaction models, LR with the reference-coding scheme had the highest power in detecting the dominant–dominant interaction with minor alleles as the disease alleles in Model 2. This is because the reference group selection in LR and in Model 2 was consistent. In this way, true penetrance contrast is not reduced by the cross-distribution of risk cells in both the comparative and reference group.

The empty-cell effect had minor impact on MARS compared with the impact on LR with the reference-coding scheme. For testing SNP–SNP interaction, the empty-cell phenomenon is common. For just a two-way interaction, the empty-cell proportion in LR may be as high as 90%. As the penetrance contrast between the risk and low-effect subgroups increased, the power of MARS for all testing models increased, as expected. In LR with the reference-coding scheme, however, power decreased as the penetrance contrast increased in some interaction models (Models 1, 2, 4, and 5). The primary reason is that some low-effect cells (with penetrance 0.01) had no cases. The designated interaction was severely distorted by the empty-cell effects in LR. As the $\mathrm{PEN_r}$ increased, the higher chance of the empty-cell effect occurred and therefore the lower power the LR had to detect the true SNP–SNP interactions. Model 3, whose risk groups contained at least one variant genotype, had the fewest number of cells with low penetrances. These cells with low penetrances also had a higher frequency of subjects, so the empty-cell effect had minor impact on Model 3 compared with other models.

Although the empty-cell effect makes a minor impact on LR with the additive-mode scheme, its additive mode assumption is only reasonable in some situations. For example, the power of LR with the additive-mode scheme was higher in Models 2 and 3 than the other two models. That is because the additive mode using the major homozygous genotype as a baseline is consistent with the designated penetrance distribution. In contrast, MARS can automatically combine empty cells into others, so the power of MARS still increased with minor interference by the empty-cell effect. This study result demonstrates how severely the empty-cell effect impacts LR with SNPs using the reference-coding scheme and MARS in correctly detecting SNP–SNP interactions.

LR with SNPs using the reference-coding scheme had low power for two primary reasons: the empty-cell effect and the preselected reference group. This study shows MARS performed better than LR with SNPs using both the reference-coding and additive-mode schemes. Besides the two coding schemes we examined in this study, Cockerham's (1954) coding scheme also has been used in LR to detect SNP–SNP interactions. In this scheme, two dummy variables (say x and z) are applied for an SNP, with $x = 1$ and $z = -0.5$ for one homozygote genotype, $x = 0$ and $z = 0.5$ for the heterozygote genotype, and $x = -1$ and

$z = -0.5$ for the other homozygote genotype. It has been shown that LR with the additive-mode scheme is sufficient to detect SNP–SNP interactions comparing with LR using the Cockerhams' coding scheme (North et al. 2005; Barhdadi and Dube 2007). We can expect that the empty-cell effect has impact on LR using Cockerhams's coding scheme, which uses two parameters for each SNP. The empty-cell effect also interferes with the performance of MDR (Ritchie et al. 2001), which is a popular method for testing SNP–SNP interactions. The dichotomizing process in MDR interferes with the empty-cell effect, especially in detecting high-order interactions for a small sample size (Park and Hastie 2008).

Although MARS had useful traits for detecting SNP–SNP interactions, it had the following weaknesses. There are several ways to present the same interaction combination. The interaction combinations selected from MARS may not be the best in terms of the number of degrees of freedom. For example, in Model 1, the best term was (SNP_A = AA/Aa) × (SNP_B = BB/Bb) with one DF. MARS may display the same two-way SNP–SNP interaction by two terms: (SNP_A = AA/Aa) × (SNP_B = BB) and (SNP_A = AA/Aa) × (SNP_B = Bb) with two DF. However, this drawback may be conquered by manually reselecting the best combinations among the terms selected by MARS. In addition, the original MARS design is for continuous outcomes. Although MARS can also be applied for binary outcomes, the model prediction is not restricted within 0 and 1 as probabilities (2001). This weakness can be solved by using MARS as a screening tool to select the significant terms and then by using LR to display the final model. In addition, MARS is not designed for identifying genetic heterogeneity. Before applying MARS to detect genetic interactions, the cluster analysis was recommended to detect genetic heterogeneity (Schork et al. 2001).

A revised logistic regression called penalized logistic regression (PLR) has been proposed by Park and Hastie (2008). PLR using quadratic penalization can improve the unstable model coefficient estimates, and the empty-cell effects when the number of parameters grows large. However, with PLR, it is difficult to avoid the effect of a preselected reference group, which has been shown in this study to be an important issue in detecting SNP–SNP interactions. In summary, this study shows MARS is a powerful method to exploring SNP–SNP interactions. In addition to comparing it with LR, it is important in future studies to compare the performance of MARS with other statistical methods that assess SNP–SNP interactions.

Acknowledgments We thank Laura Gallitz for her help on the manuscript editing, and we also thank anonymous reviewers for very helpful suggestions and comments. This research was supported by National Cancer Institute grants, CA73629 and CA090898 and American Cancer Society (CNE-101119) to J.J. Hu.

References

MARS user guide (2001) Salford Systems, San Diego

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automat Contr 19:716–723

Albert A, Anderson A (1984) On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71:1–10

Barhdadi A, Dube MP (2007) Two-stage strategies to detect gene × gene interactions in case-control data. In: BMC proceedings. Genetic analysis workshop 15, p S135. St. Pete Beach, FL, USA

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth, Belmont

Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, Van Eerdewegh P (2005) Identifying SNPs predictive of phenotype using random forests. Genet Epidemiol 28:171–182

Cockerham CC (1954) An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. Genetics 39:859–882

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

Culverhouse R, Suarez BK, Lin J, Reich T (2002) A perspective on epistasis: limits of models displaying no main effect. Am J Hum Genet 70:461–471

De Boor C (1978) A practical guide to splines. Springer, New York

Friedman JH (1991) Multivariate adaptive regression splines. Ann Stat 19:1–66

Gauderman WJ (2002) Sample size requirements for association studies of gene–gene interaction. Am J Epidemiol 155:478–484

Ge D, Zhu H, Huang Y, Treiber FA, Harshfield GA, Snieder H, Dong Y (2007) Multilocus analyses of renin–angiotensin–aldosterone system gene variants on blood pressure at rest and during behavioral stress in young normotensive subjects. Hypertension 49:107–112

Gu D, Su S, Ge D, Chen S, Huang J, Li B, Chen R, Qiang B (2006) Association study with 33 single-nucleotide polymorphisms in 11 candidate genes for hypertension in Chinese. Hypertension 47:1147–1154

Hu JJ (2006) DNA repair pathways: genetic determinants of disparities in prostate and colon cancer. In: The 97th annual meeting of American association for cancer research. Washington, DC

Hu JJ, Keku TO, Galanko J, Velasco-Gonzalez C, Daniel B, Sandler RS (2007) DNA-repair genetic polymorphisms and racial difference of colon cancer risk. American Association Cancer Research, Los Angeles

Hu JJ, Hall MC, Grossman L, Hedayati M, McCullough DL, Lohman K, Case LD (2004) Deficient nucleotide excision repair capacity enhances human prostate cancer risk. Cancer Res 64:1197–1201

Lin HY, Desmond R, Louis Bridges S Jr, Soong SJ (2008) Variable selection in logistic regression for detecting SNP–SNP interactions: the rheumatoid arthritis example. Eur J Hum Genet 16(6):735–741

Lin HY, Hall MC, Clark PE, Phillips JJ, Hu JJ (2006) Gene–gene interactions of DNA-repair nsSNPs in prostate cancer recurrence. In: The 97th annual meeting of American association for cancer research, Washington, DC

Moore JH (2003) The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered 56:73–82

Moore JH, Williams SM (2002) New strategies for identifying gene–gene interactions in hypertension. Ann Med 34:88–95

Morgan JN, Sonquist JA (1963) Problems in the analysis of survey data, and a proposal. J Am Stat Assoc 58:415–434

Musani SK, Shriner D, Liu N, Feng R, Coffey CS, Yi N, Tiwari HK, Allison DB (2007) Detection of gene × gene interactions in genome-wide association studies of human population data. Hum Hered 63:67–84

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

North BV, Curtis D, Sham PC (2005) Application of logistic regression to case-control association studies involving two causative loci. Hum Hered 59:79–87

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

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69:138–147

Schork NJ, Fallin D, Thiel B, Xu X, Broeckel U, Jacob HJ, Cohen D (2001) The future of genetic case-control studies. Adv Genet 42:191–212

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Scuteri A, Sanna S, Chen WM, Uda M, Albai G, Strait J, Najjar S, Nagaraja R, Orru M, Usala G, Dei M, Lai S, Maschio A, Busonero F, Mulas A, Ehret GB, Fink AA, Weder AB, Cooper RS, Galan P, Chakravarti A, Schlessinger D, Cao A, Lakatta E, Abecasis GR (2007) Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. PLoS Genet 3:e115

Smith TR, Miller MS, Lohman K, Lange EM, Case LD, Mohrenweiser HW, Hu JJ (2002) Polymorphisms of XRCC1 and XRCC3 genes and susceptibility to breast cancer. Cancer Lett 190:183–190

Smith TR, Levine EA, Perrier ND, Miller MS, Freimanis RI, Lohman K, Case LD, Xu J, Mohrenweiser HW, Hu JJ (2003) DNA-repair genetic polymorphisms and breast cancer risk. Cancer Epidemiol Biomarkers Prev 12:1200–1204

Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, Spain S, Penegar S, Chandler I, Gorman M, Wood W, Barclay E, Lubbe S, Martin L, Sellick G, Jaeger E, Hubner R, Wild R, Rowan A, Fielding S, Howarth K, Silver A, Atkin W, Muir K, Logan R, Kerr D, Johnstone E, Sieber O, Gray R, Thomas H, Peto J, Cazier JB, Houlston R (2007) A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat Genet 39:984–988

Van Emburgh BO, Hu JJ, Levine EA, Mosley LJ, Case LD, Lin HY, Knight SN, Perrier ND, Rubin P, Sherrill GB, Shaw CS, Carey LA, Sawyer LR, Allen GO, Milikowski C, Willingham MC, Miller MS (2008) Polymorphisms in drug metabolism genes, smoking, and p53 mutations in breast cancer. Mol Carcinog 47:88–99

Veaux RDD, Psichogios DC, Ungar LH (1993) A comparison of two nonparametric estimation schemes: MARS and neural networks. Comput Chem Eng 17:819–837

Wade MJ (2000) Epistasis and evolutionary process. Oxford University Press, New York

Webb MC, Wilson JR, Chong J (2004) An analysis of quasi-complete binary data with logistic model: application to alcohol abuse data. J Data Sci 2:273–285

York TP, Eaves LJ (2001) Common disease analysis using multivariate adaptive regression splines (MARS): genetic analysis workshop 12 simulated sequence data. Genet Epidemiol 21(Suppl 1):S649–S654

York TP, Eaves LJ, van den Oord EJ (2006) Multivariate adaptive regression splines: a powerful method for detecting disease–risk relationship differences among subgroups. Stat Med 25:1355–1367

Zabaleta J, Lin HY, Sierra RA, Hall MC, Clark PE, Sartor OA, Hu JJ, Ochoa AC (2008) Interactions of cytokine gene polymorphisms in prostate cancer risk. Carcinogenesis 29:573–578