Published in final edited form as:

Ann Hum Genet. 2012 January ; 76(1): 53–62. doi:10.1111/j.1469-1809.2011.00692.x.

TRM: A Powerful Two-stage Machine Learning Approach for Identifying SNP-SNP Interactions

Hui-Yi Lin^{1,*}, Y. Ann Chen¹, Ya-Yu Tsai¹, Xiaotao Qu¹, Tung-Sung Tseng², and Jong Y. Park¹

¹H. Lee Moffitt Cancer Center & Research Institute, Tampa, Florida, 33612, USA

²School of Public Health, Louisiana State University Health Sciences Center, New Orleans, Louisiana, 70112, USA

SUMMARY

Studies have shown that interactions of single nucleotide polymorphism (SNP) may play an important role for understanding causes of complex disease. Machine learning approaches provide useful features to explore interactions more effectively and efficiently. We have proposed an integrated method that combines two machine learning methods - Random Forests (RF) and Multivariate Adaptive Regression Splines (MARS) - to identify a subset of important SNPs and detect interaction patterns. In this two-stage RF-MARS (TRM) approach, RF is first applied to detect a predictive subset of SNPs, and then MARS is used to identify the interaction patterns among the selected SNPs. We evaluated the TRM performances in four models: three causal models with one two-way interaction and one null model. RF variable selection was based on out-of-bag classification error rate (OOB) and variable important spectrum (IS). First, we compared the selection of important variable of RF and MARS. Our results support that RF_{OOB} had better performance than MARS and RF_{IS} in detecting important variables. We also evaluated the true positive rate and false positive rate of identifying interaction patterns in TRM and MARS. This study demonstrates that TRM_{OOB}, which is RF_{OOB} plus MARS, has combined the strengths of RF and MARS in identifying SNP-SNP interaction patterns in a scenario of 100 candidate SNPs. TRM_{OOB} had greater true positive rate and lower false positive rate compared with MARS, particularly for searching interactions with a strong association with the outcome. Therefore the use of TRM_{OOB} is favored for exploring SNP-SNP interactions in a large-scale genetic variation study.

Keywords

polymorphism; interaction; machine learning

INTRODUCTION

The complexity of diseases such as cancer has led to the general recognition that identifying or targeting a single nucleotide polymorphism (SNP) is not sufficient enough to reveal etiology or develop a treatment strategy (Moore, 2003). A growing number of studies evaluate SNP-SNP interactions to complement the findings from univariate analyses. With the development of modern genotyping technology, large-scale genetic variation studies are

*Corresponding author: Hui-Yi Lin, Ph.D., 12902 Magnolia Drive, Tampa, Florida 33612. Tel: 813-745-6218; Fax: 813-745-6107; Hui-Yi.Lin@moffitt.org.

gaining popularity. Consequently, there is a need to find suitable statistical methods to perform variable selection and identify the interaction patterns.

The common inheritance models for each SNP include the dominant, recessive, additive and genotypic (three-level categorical variable) model. Some approaches search for pairwise interactions based on one pre-specified inheritance model. For example, PLINK package (Purcell *et al.*, 2007, Purcell, 2010) search pairwise interactions for SNPs with an additive model. In Boolean Operation-based Screening and Testing (BOOST) (Wan *et al.*, 2010), each SNP is treated as a genotypic model. If considering multiple inheritance models, the number of interaction combinations (e.g. dominant-dominant and dominant-recessive) increases dramatically as the number of candidate SNPs increases. Therefore, the complexity of genotype combinations makes the all-pairwise search approaches more challenging, especially for higher-order interactions, and aggravates the issue of multiple comparisons. The pre-specified inheritance model is not required for Random Forests (RF) and Multivariate Adaptive Regression Splines (MARS). These two methods will automatically select the suitable model based on model improvement, so this feature makes the interaction search more effectively and efficiently.

In order to select the most powerful genotype combinations for predicting phenotype in studies with a large number of SNPs, one solution is to apply a two-stage approach. In the two-stage approach, a subset of SNPs is selected first and then an intensive search is performed in the second step. Univariate analyses are commonly applied as a screening tool to select SNPs with a significant or marginal main effect. This conventional approach overlooks pure interactions without main effects. Compensating for the drawbacks of the univariate approach, RF provides the variable important measurement for taking interactions into consideration. RF has been applied successfully in several studies (Meng *et al.*, 2007, De Lobel *et al.*, 2010, Briggs *et al.*, 2010) and has been shown to be feasible in a genome-wide association study (Goldstein *et al.*, 2010). However, RF is difficult to directly reveal underlying interaction patterns. In order to identify significant interaction patterns, a two-stage approach is preferred.

For exploring SNP-SNP interaction patterns in the second stage, several statistical methods have been proposed, for example MARS (Friedman, 1991), neural networks (De Veaux *et al.*, 1993), classification and regression trees (CART, (Breiman *et al.*, 1984), multifactor dimensionality reduction (MDR) (Ritchie *et al.*, 2001), logic regression (Kooperberg *et al.*, 2007), combinatorial partitioning method (Nelson *et al.*, 2001), and reconstructability analysis (Shervais *et al.*, 2010). MARS has provided strengths over CART by allowing both main effects and interactions in a model, and is superior to traditional logistic regression by increasing accuracy of identifying interactions without strong main effects (Cook *et al.*, 2004). MARS tends to be more powerful and generates more interpretable results than neural network (De Veaux *et al.*, 1993). MARS has been shown to be a powerful method for identifying SNP-SNP interaction patterns in our previous simulation study (Lin *et al.*, 2008). In this existing work, we compared the true positive rate of MARS and logistic regression using the same three models we used in this study with a penetrance in the risk group of 0.15–0.50 and a penetrance in the reference group of 0.01. Using the model with one dominant-dominant two-way interaction (Model 1 in this study) as an example, the true positive rate of MARS is in a range of 74–97% while the true positive rate of logistic regression using an additive inheritance model is only 5–8%. MARS has been successfully applied for detecting SNP-SNP interaction in several studies (Zabaleta *et al.*, 2008, Zabaleta *et al.*, 2009, Nonyane & Foulkes, 2008).

We proposed the two-stage RF-MARS (TRM) approach for integrating the advantages of two powerful machine learning methods. RF has been commonly used for selecting

predicting important variables without directly providing specific interaction patterns. MARS is a useful tool for detecting SNP-SNP interaction patterns, but could be inundated by a large amount of non-functional SNPs. In order to understand the mechanism and provide empirical evidence for this approach, we evaluated the performance of MARS and RF in selecting important variables, and compared true positive rate and false positive rate of TRM and MARS in identifying interaction patterns.

MATERIALS AND METHODS

We combined two machine learning method for identifying SNP-SNP interactions. This TRM approach uses RF as a screening tool in the first stage and then applies MARS in the second stage. The methods of RF and MARS are described in greater details below.

MARS

MARS is a nonparametric regression method without an assumption about the underlying functional association between the outcome variable and the covariates. The primary unit of MARS is the basis function, which contains information of one or more variables. When covariates are categorical, the MARS model for an outcome variable y and M covariates can be written as the following equation:

$$\begin{aligned} y = \widehat{f}(x) &= a_0 + \sum_{m=1}^M a_m B_m(x) \\ &= a_0 + \sum_{m=1}^M a_m \prod_{k=1}^{K_m} I(x_{km} \in A_{km}), \end{aligned}$$

where A is a subset of the possible values of x . Here, a_0 and a_m are the coefficients estimated via least square estimation, and K_m is the number of factors (interaction order) in the m^{th} basis function (B_m) (Friedman & Roosen, 1995, Friedman, 1991). An example of the MARS output is as follows:

BF1=(SNP_A in (“GG”));

BF2=(SNP_B in (“CT”, “CC”)) × BF1;

BF3=(SNP_C in (“AA”));

Y=0.5+ 0.3×BF2 + 0.2×BF3

Y represents a binary phenotype. BF1 (basis function 1) represents the dummy variable of SNP_A with GG=1 and AG/AA=0 (recessive model). BF2 represents the dummy variable of the SNP_A and SNP_B combination with “GG and CT/CC”=1 and other combinations of the two SNPs are equal to 0, so BF2 is a recessive-dominant interaction. Thus, this model contains one two-way interaction of SNP_A × SNP_B and one main effect of SNP_C.

A forward-backward stepwise strategy is commonly used to select the best model with the minimum value of the Generalized Cross Validation (GCV) error. The GCV criterion has taken both the residual errors and model complexity into consideration and measures goodness of fit of the model in the following form:

$$GCV(M) = \frac{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{f}_M(x_i))^2}{(1 - C(M)/N)^2},$$

where with $C(M) = M^* (d/2+1)+1$, N is the number of subjects in the data set, M represents the number of nonconstant basis function in the model, and d is the penalty for adding a basis function. The details were described previously (Friedman, 1991).

The highest order of interaction is user defined. In forward selection, the maximum number of basis functions is used for determining how many basis functions are considered. In backward selection, the best MARS model is selected based on the lowest value of GCV, an index for measuring generalized mean square errors. Degree of freedom charged per basis function is used to decide the final model size. If a higher degree of freedom setting is used, a small final MARS model will be selected. The degree of freedom value can be user defined or estimated using a cross-validation procedure.

Random Forests

Random Forests (RF) is a collection of decision trees with binary splits (Breiman, 2001). Each tree is grown based on two steps of randomizations. First, a subset of predictors is randomly selected from all predictors. Second, a bootstrap sample of the data is randomly drawn with replacement from the original sample with the same sample size. RF provides a measurement of variable importance and is a useful feature in genetic studies. For each RF tree, the unused observations, called “out-of-bag (OOB)”, are used for estimating the classification accuracy. The classification accuracy is then calculated after permuting a predictor to destroy the association between the predictor and the outcome. The unscaled permutation accuracy importance, the average of the difference between two accuracies, was applied in this study. Two RF variable selection methods proposed by Diaz-Uriarte (Diaz-Uriarte & Alvarez de Andres, 2006, Diaz-Uriarte, 2010) were applied for selecting a set of important variables. In the RF_{OOB} method, multiple random forests were iteratively fitted. For each forest, the backward variable elimination is applied to select a set of variables based on the smallest OOB error rate. Among all fitted forests, the final set of variables is selected based on the smallest number of variables whose error rate is less than a specified standard error of the minimum OOB error rate of all forests. The RF_{IS} method is based on the importance spectrum of the original data compared with the importance spectrum (IS) from the permuted data, of which the class labels are randomly permuted.

Simulation

We simulated 1,000 subjects (500 cases and 500 controls) and all SNPs were generated following the Hardy-Weinberg equilibrium. For the number of testing SNPs, we set up two scenarios: scenario 1 with 10 SNPs and scenario 2 with 100 SNPs, respectively. Three models with one two-way interaction were generated. The functional SNP_A, SNP_B and SNP_C had common variants, and the major allele frequency were 0.5, 0.75 and 0.5, respectively. The nonfunctional (‘noise’) SNPs were generated independently with a major allele frequency in a range of 0.5 to 0.9.

For each SNP, an allele with a capital letter denotes a major allele and one with a lower case letter denotes a minor allele. All testing models were based on our previous study (Lin *et al.*, 2008). As shown in Table 1, Model 1 had one dominant-dominant interaction with the major allele (A and B) as the risk allele. In Model 2, the high-risk group was comprised of subjects with at least one of the homozygous minor-allele genotypes (aa and bb). Model 3 also had

one dominant-dominant interaction but the risk allele was a minor allele (a and b). The disease status was generated based on the penetrance, a conditional probability of phenotype given the genotype. A wide range of penetrance contrast between the risk and reference genotype group was chosen to comprehensively evaluate TRM, MARS and RF performance in various situations. The penetrances in the risk genotype group (PEN_r) were 0.02, 0.03, 0.04, and 0.05, and the penetrance in the reference genotype group was 0.01.

For MARS alone method, the maximum number of basis functions was set at 15 for scenario 1 with 10 SNPs, and was set at 70 for scenario 2 with 100 SNPs, respectively. In the TRM approach, 15 maximum number of basis functions was applied for all conditions. Ten-fold cross-validation was applied to select the degree of freedom charged per basis function. For the RF analyses, R package randomForest was used in this study (Liaw & Wiener, 2002), and 500 individual classification trees were grown in the forest. The number of randomly selected predictors was the square root of the number of predictors, which is the default in the randomForest R package manual (Breiman et al., 2009). For scenarios 1 and 2, the number of predictors was set at 3 and 10, respectively. The unscaled permutation accuracy importance index was applied to measure the relative variable importance using out-of-bag observations. Studies (Breiman, 2001, Goldstein et al., 2010) have shown 250 RF trees are enough for achieving small out-of-bag error rate. In the RF_{OOB} calculation, we used 500 trees to build the first forest and 200 trees for all additional forests. Using this tree number setting, stable results were observed (results not shown). Among all fitted forests, the final set of variables was selected based on the smallest number of variables whose error rate is less than one standard error of the minimum OOB error rate of all forests. For the RF_{IS} approach, 100 random permutations of the class labels were performed.

We were interested in testing two endpoints: (1) important variables and (2) SNP-SNP interaction patterns. For identifying important variables, true positive rate was defined as the detected proportion of functional SNPs listed in the MARS model or RF top candidate list among 500 simulation runs. For detecting SNP-SNP interaction patterns, we compared the performance of the MARS and TRM approach. True positive rate of detecting interaction patterns was measured as the frequency of the designated two-way interaction in a MARS model among 500 simulations. Showing two functional SNPs simultaneously was considered as a correct selection regardless of the inheritance model. We also evaluated the false positive rate of MARS, TRM and RF in the null model with 100 candidate SNPs. The false positive rate was defined as the frequency of the selected terms under the null model among 500 simulation runs. In this null model, disease penetrances were generated regardless of genotypes' distribution. Data simulation and analyses were performed by using SAS 9.1. MARS was performed using the MARS 2.0 (Salford Systems), and Random Forest was performed using randomForest and varSelRF R package.

Real data application: prostate cancer aggressiveness

We were interested in exploring SNPs in the estrogen receptor (ER) pathway associated with prostate cancer aggressiveness. We used 1,151 prostate cancer cases (492 non-aggressive and 659 aggressive) obtained from the Cancer Genetic Markers of Susceptibility (CGEMS) prostate cancer genome-wide data set (<http://cgems.cancer.gov/data>) initiated by the National Cancer Institute. Prostate cancer aggressiveness was defined as patients with Gleason scores ≥ 7 or stage $\geq$ III. The study samples, selected from the Prostate, Lung, Colon and Ovarian Cancer Screening Trial enrollment between 1993 to 2001, are non-Hispanic whites without a prior history of prostate cancer before randomization. All men diagnosed with prostate cancer between time of enrollment and 2001 are considered for inclusion in the CGEMS. In order to increase the number of aggressive prostate cancer patients, the aggressive prostate cancer patients in the CGEMS were enrolled until the end of 2003 (Yeager *et al.*, 2007).

Polymorphisms in the ER-related genes have been proposed to be associated with prostate cancer risk (Tanaka *et al.*, 2003, Naylor, 2007). We examined 149 SNPs in the six ER-related genes (*CYP1B1*, *SRD5A2*, *ESR1*, *ESR2*, *CYP19A1* and *CYP11A1*). The Hardy-Weinberg equilibrium was examined in the control group (n=1,101), which was not used for the association test in this study. Studies show that variable importance ranking for SNPs with strong linkage disequilibrium is biased (Goldstein *et al.*, 2010, Nicodemus & Malley, 2009). Thus, linkage disequilibrium among all 149 SNPs was examined using r^2 in the Haploview Tagger (de Bakker *et al.*, 2005). In order to improve accuracy of variable importance ranking in the TRM approach, a total of 106 SNPs were included for further analyses after selecting one SNP in each pair with strong linkage disequilibrium of $r^2 > 0.8$. RF does not permit missing data, therefore the missing genotypes were imputed using IMPUTE version 2.0 (Howie *et al.*, 2009) with the HapMap3 CEU+TSI data as the reference population. IMPUTE has the advantage of being able to impute the missing genotypes while keeping the original genotyped data, and has been shown to be a powerful and reliable imputation method (Nothnagel *et al.*, 2009). The missing values among these 106 SNPs were low. The median missing proportion is 0.3% and the maximum is 4.5%. In order to preserve all SNPs in the analyses, our analyses were based on the combined dataset composed of the original genotyped data and the “best-guess” imputed data.

Both effects of individual SNPs and SNP-SNP interactions were examined. In MARS and RF, SNPs are treated as categorical variables so the additive model of genotypes was not taken into consideration. For consistency, the additive model was not evaluated in the univariate analyses. For each SNP, different inheritance models (genotypic, dominant and recessive model) were assessed using the Wald test in logistic regression models, and the best model was selected based on the minimum p-value. For adjusting multiple comparisons, false discovery rate (FDR) adjusted p-value (Benjamini & Hochberg, 1995) was applied.

For exploring SNP-SNP interaction patterns, we applied the MARS and TRM_{OOB} (RF_{OOB} plus MARS), which were shown to be more powerful than TRM_{IS} (RF_{IS} plus MARS) in detecting interaction patterns among a large number of candidate SNPs in our simulation study. The maximum number of basis functions was set at 30 for the MARS alone method. The final MARS model was selected based on 10-fold cross-validation used to select the degree of freedom charged per basis function. For RF, 500 classification trees were generated. In the first stage of the TRM_{OOB} approach, the top candidate SNPs were selected based on minimizing the OOB errors using the unscaled permutation accuracy importance index in RF. In the RF_{OOB}, we used the suggested default setting of 5,000 trees to build the first forest and 2,000 trees for all additional forests (Diaz-Urriarte, 2010). In the second stage, MARS with the maximum basis functions of 15 was applied to exploring SNP-SNP interaction patterns among the top candidate SNPs. The other parameter settings in the second stage of TRM_{OOB} were the same as the MARS alone method.

The heterozygous advantage model is not easy to interpret in general, so the MARS basis function with a heterozygous genotype alone was replaced by either a dominant or recessive model. This depended upon which model was more significantly associated with the outcome. The significance level of each covariate was not provided in MARS, so logistic regression was applied for selecting the final model. All MARS selected covariates were included in a logistic regression model and the insignificant (p-value > 0.05) covariates were excluded.

RESULTS

Simulation

For selecting important variables, RF_{OOB} tended to be the most powerful method compared with MARS and RF_{IS} (Figures 1 and 2). Under all scenarios, RF_{OOB} was more powerful than RF_{IS} . For 10 candidate SNPs (Figure 1), the true positive rate of RF_{OOB} was slightly better than or similar to MARS, and both of them were more powerful than RF_{IS} . In Model 1 with 10 candidate SNPs, RF_{OOB} was more powerful than MARS. In Models 2 and 3, RF_{OOB} and MARS had similar performance. For the dataset with 100 candidate SNPs (Figure 2), RF_{OOB} was more powerful in selecting important variables than RF_{IS} , and both performed better than MARS except for $PENr$ of 0.02. For low penetrance and 100 candidate SNPs, the true positive rate of MARS is similar to RF_{OOB} .

For selecting important variables, the impact on true positive rate of MARS and RF due to a large number of noise SNPs was shown in Figure 3. The number of noise SNPs had bigger impact on MARS but smaller impact on RF methods. As the number of noise SNPs increased from 8 to 98, the true positive rate decreased in both MARS and RF_{OOB} methods, but the degree of penalty in true positive rate varied. In dominant-dominant Model 1 with a penetrance of 0.03, the true positive rate in MARS declined 39% (from 51% to 12%) but the true positive rate in RF_{OOB} only decreased 16% (from 58% to 42%) when number of noise SNPs increased from 8 to 98. For an interaction with large penetrance contrast, the penalty in RF due to noise variables was minor. In Model 1 with a penetrance of 0.05, the true positive rate in MARS decreased 38% (from 79% to 41%) but the true positive rate in RF_{OOB} only dropped 4% (from 79% to 75%). The impact of the number of noise SNPs' on RF_{IS} was different from the other two methods. For detecting an interaction with a low penetrance contrast, a large number of noise SNPs slightly decreased true positive rate of RF_{IS} . For identifying interactions with a large penetrance contrast, the true positive rate of RF_{IS} tended to increase as the number of noise SNPs increased based on Figure 3.

For identifying specific SNP-SNP interaction patterns, the true positive rate of MARS and TRM (TRM_{OOB} and TRM_{IS}) was compared. As the penetrance contrast increased, the detection true positive rate increased for both the MARS and TRM approach. As expected, the TRM_{OOB} approach performed better than TRM_{IS} approach in detecting interaction patterns because RF_{OOB} was more powerful in selecting important variables than RF_{IS} in the first stage of TRM. Therefore we only compared TRM_{OOB} and MARS for selecting the most powerful method. For 10 candidate SNPs (Figure 4), MARS had similar or higher true positive rate than TRM_{OOB} . In the scenario 2 of testing 100 SNPs, TRM_{OOB} had greater true positive rate than MARS except for identifying an interaction with a low penetrance contrast ($PENr=0.02$). For example, Model 3 used 100 candidate SNPs with a penetrance of 0.03 (Figure 5), and the true positive rate in the TRM_{OOB} and MARS approach was 88% and 64%, respectively. When a penetrance contrast was low ($PENr=0.02$), the true positive rate was higher in MARS than in the TRM methods. For Model 3 with 100 candidate SNPs and a penetrance of 0.02 (Figure 5), the true positive rate of MARS and TRM_{OOB} was 39% and 27%, respectively.

The results of simulated null model showed that the false positive rate in RF_{OOB} was large than in RF_{IS} for selecting important variables. The mean of RF_{IS} false positive rate was 0.4% with a range of 0–1%, and the mean of the RF_{OOB} false positive rate was 5.8% with a range of 1.6–10.4%. For testing the null model, MARS tended to select more false positive covariates than TRM_{OOB} . The average number of covariates (including main effects and interactions) per model was 5 and 32 for TRM_{OOB} and MARS, respectively. The false positive rate per covariate was also lower in TRM_{OOB} than in MARS (mean $\pm$ standard

deviation= $0.4\% \pm 0.5\%$ vs. $0.7\% \pm 0.6\%$, respectively). Among the 500 simulation runs, 94% and 78% of covariates were selected less than 5 runs in TRM_{OOB} and MARS, respectively.

Real data results

We are interested in evaluating ER-related SNPs associated with prostate cancer aggressiveness. All 106 SNPs followed the Hardy-Weinberg equilibrium in the control group. The main effects of individual SNPs were examined using logistic regressions. Using the minimum p-value among different inheritance models (genotypic, dominant and recessive), there were 15 SNPs with raw p values less than 0.05 (Table 2). However, none were significant after adjusting multiple comparisons using the FDR adjusted p-value.

For exploring ER related SNP-SNP interactions, the MARS alone and TRM_{OOB} method were applied to test up to two-way interactions. Using the TRM_{OOB} approach, four SNPs were selected from the 106 candidate SNPs. These SNPs were rs3751591, rs3751592, rs9322336, and rs9972359. These four SNPs were then included in MARS for exploring SNP-SNP interactions. The final models selected using these two methods are shown in Table 3. One main effect (rs3751591) and two two-way interaction (rs3751591*rs9322336 and rs3751591*rs9972359) were selected in both the MARS and TRM_{OOB}. For evaluating SNP interactions among 106 candidate SNPs, TRM_{OOB} selected fewer covariates than MARS. Using MARS, 15 basis functions and 15 SNPs were selected. Using the TRM_{OOB} approach, 5 basis functions and 4 SNPs were selected.

DISCUSSION

Our study provides empirical evidence to demonstrate TRM's ability for combining strengths of two machine learning methods. RF provides variable important scores by taking interactions into consideration, but is challenged to directly reveal the underlying interactions. MARS is a powerful tool for identifying patterns but is overwhelmed by a large amount of non-functional SNPs. In our simulation study for testing 100 SNPs, TRM_{OOB} has been shown to have greater true positive rate and fewer false-positive findings to identify SNP-SNP interaction patterns compared with MARS and TRM_{IS}. TRM_{OOB} had the capability for searching interactions with a strong association with the outcome, which is the crucial for most researchers. For detecting interactions with a small number of candidate SNPs and with a low penetrance contrast, MARS tended to be more powerful.

In the first stage of TRM for selecting important variables, RF_{OOB} is the most powerful method, and the high-dimensional data effect on true positive rate is minor in RF_{OOB} than in MARS. Our results demonstrate how the number of non-functional SNPs weakened the true positive rate of MARS in identifying patterns. This supports the decision of using RF_{OOB} as the screening tool. For comparing the performance of RF_{OOB} and RF_{IS}, RF_{OOB} was more powerful than RF_{IS} in all scenarios. However, the false positive rates in RF_{OOB} were much higher than in RF_{IS}. The false positive issue is of less concern for a screening tool because it can be modified by the method used in the second stage. We also compared true positive rate of RF_{OOB} and MARS in terms of selecting important variables. The accuracy of variable selection in RF_{OOB} tends to be higher than in MARS and, consequently, validates why TRM_{OOB} tends to be more powerful than MARS in identifying SNP-SNP interactions. RF is not only powerful in identifying interactions with none or minor main effects but also good at detecting factors with main effects only. The performance of RF and the conventional univariate test in detecting SNPs with main effects have shown to be similar (Lunetta et al., 2004). In our prostate cancer data analyses, the four SNPs (rs3751591, rs3751592, rs9322336, and rs9972359) on the top six of the main effect analyses (Table 2) were also selected using RF_{OOB}.

In the second stage of TRM, MARS provides custom variable selection in building a prediction model. The ultimate goal of genetic association studies is to identify biomarkers that can gain significant prediction value additional to the existing prediction model. Thus, it is preferable to apply a variable selection strategy that can consider multiple predictors (both main effects and interactions) simultaneously and can control for known confounding factors. MARS contained both useful features in conducting a prediction model. Some statistical methods used in genetic association studies (such as MDR (Ritchie *et al.*, 2001) and CART (Breiman *et al.*, 1984)) are not tailor-made for variable selection in building a prediction model because they cannot evaluate multiple interactions and main effects at the same time. As for variable selection in MARS, the significance level (p-value) of each covariate is not provided. With regard to the selection of maximum number of basic functions in MARS, the suggested maximum number of basis functions is at least two to four times the size of the truth based on the MARS user's guide. In practice, the value of maximum number of basis functions is selected based on empirical experience to choose a big enough value for including sufficient candidate SNPs but not too many to overwhelm the modeling. In addition to variable selection strategies in logistic regression, our results show that the majority of false positive rate per covariate is low (<1%) in the null model. This suggests the bootstrap analysis may be applied to validate the variable significance of MARS selected covariates in the future studies.

In the prostate cancer analyses, all main effects are not significant after adjusting for multiple comparisons. However, some of these SNPs were selected in the TRM model in different forms (main effects or interactions). This means that it may not be suitable for concluding important variables merely based on the multiple comparison results of the univariate analyses. All SNPs selected in these two TRM_{OOB} and MARS models are either in the *CYP19A1* (rs3751591 and rs9972359) or *ESR1* (rs9322336, rs543650, and rs9322343) genes. Our findings suggest these two genes may interact and play a role in prostate cancer aggressiveness. Both genes are related to the balance of androgen and estrogen in the prostate gland, which play a significant role for maintaining prostate health (Carruba, 2007, Ellem & Risbridger, 2010). Previous studies have been shown that *CYP19A1* (Beuten *et al.*, 2009) and *ESR1* (Nicolaiew *et al.*, 2009, McIntyre *et al.*, 2007) may be associated with prostate cancer risk. Furthermore, the SNP-SNP interactions between *CYP19A1* and *ESR1* genes are observed to be associated with premature ovarian failure (Kim *et al.*, 2011). These studies support our findings of the interactions between *CYP19A1* and *ESR1* associated with prostate cancer aggressiveness.

In summary, we have proposed TRM, a two-stage approach by integrating RF and MARS, as a powerful strategy to select important variables and identify interaction patterns. As shown in our simulation study with 100 SNPs, using the TRM_{OOB} has a greater true positive rate and a low false positive rate when compared with MARS and TRM_{IS}. This advantage may be even more pronounced for studies with a large number of candidate SNPs. As we discussed previously, MARS tended to have higher true positive rate in identifying interactions in a study with a small number of candidate SNPs, and with a low penetrance contrast. Therefore, performing both TRM_{OOB} and MARS is recommended in order to comprehensively search for possible interactions in studies with limited number SNPs. In large-scale genetic variation studies, the TRM_{OOB} may provide a useful means to explore SNP-SNP interactions.

Acknowledgments

We thank William J. Fulp and Bhupendra Rawal at the Moffitt Cancer Center & Research Institute and Yung-Hsin Liu at INC Research for programming assistance. We also thank anonymous reviewers for their very helpful suggestions and comments. This study was supported by the National Cancer Institute (R01CA128813, PI: JYP) and American Cancer Society (IRG-93-092-14, PI: WJP/HYL).

References

- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Stat Methodol.* 1995; 57:289–300.
- Beuten J, Gelfond JA, Franke JL, Weldon KS, Crandall AC, Johnson-Pais TL, Thompson IM, Leach RJ. Single and multigenic analysis of the association between variants in 12 steroid hormone metabolism genes and risk of prostate cancer. *Cancer Epidemiol Biomarkers Prev.* 2009; 18:1869–80. [PubMed: 19505920]
- Breiman L. Random Forests. *Machine Learning.* 2001; 45:5–32.
- Breiman L, Cutler A, Liaw A, Wiener M. Breiman and Cutler's random forests for classification and regression. R package version 4.5–30. 2009
- Breiman, L.; Friedman, JH.; Olshen, RA.; Stone, CJ. Classification and regression trees. Belmont, CA: Wadsworth; 1984.
- Briggs FB, Goldstein BA, Mccauley JL, Zuvich RL, De Jager PL, Rioux JD, Ivinson AJ, Compston A, Hafler DA, Hauser SL, Oksenberg JR, Sawcer SJ, Pericak-Vance MA, Haines JL, Barcellos LF. Variation within DNA repair pathway genes and risk of multiple sclerosis. *Am J Epidemiol.* 2010; 172:217–24. [PubMed: 20522537]
- Carruba G. Estrogen and prostate cancer: an eclipsed truth in an androgen-dominated scenario. *J Cell Biochem.* 2007; 102:899–911. [PubMed: 17786930]
- Cook NR, Zee RY, Ridker PM. Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. *Stat Med.* 2004; 23:1439–53. [PubMed: 15116352]
- De Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. Efficiency and power in genetic association studies. *Nat Genet.* 2005; 37:1217–23. [PubMed: 16244653]
- De Lobel L, Geurts P, Baele G, Castro-Giner F, Kogevinas M, Van Steen K. A screening methodology based on Random Forests to improve the detection of gene-gene interactions. *Eur J Hum Genet.* 2010; 18:1127–32. [PubMed: 20461113]
- De Veaux RD, Psychogios DC, Ungar LH. A comparison of two nonparametric estimation schemes: MARS and Neural Networks. *Comput Chem Eng.* 1993; 17:819–837.
- Diaz-Uriarte, R. Package 'varSelRF'. 2010. <http://ligarto.org/rdiaz/Software/Software.html>
- Diaz-Uriarte R, Alvarez De Andres S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics.* 2006; 7:3. [PubMed: 16398926]
- Ellem SJ, Risbridger GP. Aromatase and regulating the estrogen:androgen ratio in the prostate gland. *J Steroid Biochem Mol Biol.* 2010; 118:246–51. [PubMed: 19896534]
- Friedman J. Multivariate adaptive regression splines. *Annals of Statistics.* 1991; 19:1–66.
- Friedman JH, Roosen CB. An introduction to multivariate adaptive regression splines. *Stat Methods Med Res.* 1995; 4:197–217. [PubMed: 8548103]
- Goldstein BA, Hubbard AE, Cutler A, Barcellos LF. An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. *BMC Genet.* 2010; 11:49. [PubMed: 20546594]
- Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 2009; 5:e1000529. [PubMed: 19543373]
- Kim S, Pyun JA, Kang H, Kim J, Cha DH, Kwack K. Epistasis between CYP19A1 and ESR1 polymorphisms is associated with premature ovarian failure. *Fertil Steril.* 2011; 95:353–6. [PubMed: 20797716]
- Kooperberg C, Bis JC, Marcianti KD, Heckbert SR, Lumley T, Psaty BM. Logic regression for analysis of the association between genetic variation in the renin-angiotensin system and myocardial infarction or stroke. *Am J Epidemiol.* 2007; 165:334–43. [PubMed: 17082497]
- Liaw, A.; Wiener, M. R News. 2002. Classification and regression by randomForest *R News*.
- Lin HY, Wang W, Liu YH, Soong SJ, York TP, Myers L, Hu JJ. Comparison of multivariate adaptive regression splines and logistic regression in detecting SNP-SNP interactions and their application in prostate cancer. *J Hum Genet.* 2008; 53:802–11. [PubMed: 18607530]

- Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P. Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet.* 2004; 5:32. [PubMed: 15588316]
- Mcintyre MH, Kantoff PW, Stampfer MJ, Mucci LA, Parslow D, Li H, Gaziano JM, Abe M, Ma J. Prostate cancer risk and ESR1 TA, ESR2 CA repeat polymorphisms. *Cancer Epidemiol Biomarkers Prev.* 2007; 16:2233–6. [PubMed: 18006911]
- Meng Y, Yang Q, Cuenco KT, Cupples LA, Destefano AL, Lunetta KL. Two-stage approach for identifying single-nucleotide polymorphisms associated with rheumatoid arthritis using random forests and Bayesian networks. *BMC Proc.* 2007; 1(Suppl 1):S56. [PubMed: 18466556]
- Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered.* 2003; 56:73–82. [PubMed: 14614241]
- Naylor SL. SNPs associated with prostate cancer risk and prognosis. *Front Biosci.* 2007; 12:4111–31. [PubMed: 17485362]
- 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; 11:458–70. [PubMed: 11230170]
- Nicodemus KK, Malley JD. Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics.* 2009; 25:1884–90. [PubMed: 19460890]
- Nicolaiew N, Cancel-Tassin G, Azzouzi AR, Grand BL, Mangin P, Cormier L, Fournier G, Giordanella JP, Pouchard M, Escary JL, Valeri A, Cussenot O. Association between estrogen and androgen receptor genes and prostate cancer risk. *Eur J Endocrinol.* 2009; 160:101–6. [PubMed: 18952763]
- Nonyane BA, Foulkes AS. Application of two machine learning algorithms to genetic association studies in the presence of covariates. *BMC Genet.* 2008; 9:71. [PubMed: 19014573]
- Nothnagel M, Ellinghaus D, Schreiber S, Krawczak M, Franke A. A comprehensive evaluation of SNP genotype imputation. *Hum Genet.* 2009; 125:163–71. [PubMed: 19089453]
- Purcell, S. PLINK(version 1.07). 2010. <http://pngu.mgh.harvard.edu/purcell/plink/>
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007; 81:559–75. [PubMed: 17701901]
- Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet.* 2001; 69:138–47. [PubMed: 11404819]
- Shervais S, Kramer PL, Westaway SK, Cox NJ, Zwick M. Reconstructability analysis as a tool for identifying gene-gene interactions in studies of human diseases. *Stat Appl Genet Mol Biol.* 2010; 9:Article18. [PubMed: 20361857]
- Tanaka Y, Sasaki M, Kaneuchi M, Shiina H, Igawa M, Dahiya R. Polymorphisms of estrogen receptor alpha in prostate cancer. *Mol Carcinog.* 2003; 37:202–8. [PubMed: 12891629]
- Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NL, Yu W. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am J Hum Genet.* 2010; 87:325–40. [PubMed: 20817139]
- Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, Fearnhead P, Yu K, Chatterjee N, Wang Z, Welch R, Staats BJ, Calle EE, Feigelson HS, Thun MJ, Rodriguez C, Albanes D, Virtamo J, Weinstein S, Schumacher FR, Giovannucci E, Willett WC, Cancel-Tassin G, Cussenot O, Valeri A, Andriole GL, Gelmann EP, Tucker M, Gerhard DS, Fraumeni JF Jr, Hoover R, Hunter DJ, Chanock SJ, Thomas G. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet.* 2007; 39:645–9. [PubMed: 17401363]
- Zabaleta J, Lin HY, Sierra RA, Hall MC, Clark PE, Sartor OA, Hu JJ, Ochoa AC. Interactions of cytokine gene polymorphisms in prostate cancer risk. *Carcinogenesis.* 2008; 29:573–8. [PubMed: 18174250]
- Zabaleta J, Su LJ, Lin HY, Sierra RA, Hall MC, Sartor AO, Clark PE, Hu JJ, Ochoa AC. Cytokine genetic polymorphisms and prostate cancer aggressiveness. *Carcinogenesis.* 2009; 30:1358–62. [PubMed: 19474090]

Figure 1. True positive rate of selecting important variables with 10 candidate SNPs. RF(OOB): the backward variable elimination based on the smallest OOB error rate in Random Forest; RF(IS): selection based on the importance spectrum in Random Forest. Penetrance of the risk group varied from 0.02 to 0.05, and penetrance of the reference group was 0.01.

Figure 2.

True positive rate of selecting important variables with 100 candidate SNPs. RF(OOB): the backward variable elimination based on the smallest OOB error rate in Random Forest; RF(IS): selection based on the importance spectrum in Random Forest. Penetrance of the risk group varied from 0.02 to 0.05, and penetrance of the reference group was 0.01. MARS

Figure 3.

Impacts of non-functional SNPs on MARS and RF. RF(OOB): the backward variable elimination based on the smallest OOB error rate in Random Forest; RF(IS): selection based on the importance spectrum in Random Forest. Penetrance of the risk group varied from 0.02 to 0.05, and penetrance of the reference group was 0.01. All models contained two functional SNPs, so the number of non-functional SNPs was 8 and 98 in the two scenarios.

Figure 4.

True positive rate of identifying SNP-SNP interaction patterns with 10 candidate SNPs. RF(OOB): the backward variable elimination based on the smallest OOB error rate in Random Forest; RF(IS): selection based on the importance spectrum in Random Forest. Penetrance of the risk group varied from 0.02 to 0.05, and penetrance of the reference group was 0.01. In the TRM (two-stage RF-MARS) approach, the SNPs selected in RF(OOB) or RF(IS) were included in MARS for further analyses.

Figure 5.

True positive rate of identifying SNP-SNP interaction patterns with 100 candidate SNPs. RF(OOB): the backward variable elimination based on the smallest OOB error rate in Random Forest; RF(IS): selection based on the importance spectrum in Random Forest. Penetrance of the risk group varied from 0.02 to 0.05, and penetrance of the reference group was 0.01. In the TRM (two-stage RF-MARS) approach, the SNPs selected in RF(OOB) or RF(IS) were included in MARS for further analyses.

Table 1

Simulated two-way interaction models.

Model 1:			
SNP _A \SNP _B	BB	Bb	bb
AA	PEN _r	PEN _r	0.01
Aa	PEN _r	PEN _r	0.01
aa	0.01	0.01	0.01
Model 2:			
SNP _A \SNP _B	BB	Bb	bb
AA	0.01	0.01	PEN _r
Aa	0.01	0.01	PEN _r
aa	PEN _r	PEN _r	PEN _r
Model 3:			
SNP _A \SNP _B	BB	Bb	bb
AA	0.01	0.01	0.01
Aa	0.01	PEN _r	PEN _r
aa	0.01	PEN _r	PEN _r

Major allele: A, B; Probability(A)=0.5, Probability(B)=0.75
PEN_r (Penetrance in the risk genotype group)= 0.02, 0.03, 0.04 or 0.05

Table 2

Individual estrogen receptor SNPs associated with prostate cancer aggressiveness using the CGEMS data.

SNP	Gene	Model ¹	Major/Minor allele	OR (95% CI) ²	Raw p-value
rs3751591	CYP19A1	Rec	A/G	4.16 (1.59–10.88)	0.0037
rs3751592	CYP19A1	Rec	T/C	1.71 (1.11–2.62)	0.0149
rs3936674	ESR1	Geno (AA vs. GG) (GA vs. GG)	C/T	0.79 (0.54–1.16)	0.0153
rs9972359	CYP19A1	Dom	T/C	1.36 (1.06–1.74)	0.0164
rs3798577	ESR1	Rec	T/C	0.71 (0.54–0.95)	0.0188
rs9322336	ESR1	Dom	T/C	0.75 (0.59–0.96)	0.0220
rs1952586	ESR2	Rec	A/G	0.22 (0.06–0.81)	0.0222
rs532010	ESR1	Rec	T/C	0.69 (0.50–0.95)	0.0240
rs3784308	CYP19A1	Dom	A/G	1.58 (1.05–2.38)	0.0277
rs9371236	ESR1	Dom	A/G	2.16 (1.07–4.34)	0.0307
rs9322343	ESR1	Dom	A/G	1.97 (1.05–3.70)	0.0343
rs2978381	ESR2	Rec	T/C	0.72 (0.52–0.98)	0.0379
rs2881766	ESR1	Rec	T/G	0.50 (0.26–0.97)	0.0403
rs3020368	ESR1	Dom	C/T	1.36 (1.01–1.83)	0.0447
rs11632903	CYP19A1	Dom	C/T	0.78 (0.60–1.00)	0.0483

¹ Genetic model (Rec: Recessive, Dom: Dominant, Geno: Genotypic)

² Odds ratio (95% confidence interval)

Table 3

Model comparison of prostate cancer aggressiveness.

Method	Model	Comparative genotype group Aggressive/All	Reference genotype group Aggressive/All	OR (95% CI) ^I	p-value
TRM _{OOB}	rs3751591 GG vs. AA/AG	27/32=0.84	632/1119=0.56	6.13 (2.28–16.48)	0.0003
	(rs3751591 AA/AG + rs9322336 TT) vs. others	425/719=0.59	234/432=0.54	1.35 (1.06–1.73)	0.0171
	(rs3751591 AA/AG + rs9972359 TC/CC) vs. others	444/755=0.59	215/396=0.54	1.34 (1.04–1.72)	0.0234
MARS	rs3751591 GG vs. AA/AG	27/32=0.84	632/1119=0.56	6.47 (2.40–17.44)	0.0002
	(rs3751591 AA/AG + rs9322336 TT) vs. others	425/719=0.59	234/432=0.54	1.44 (1.12–1.86)	0.0045
	(rs3751591 AA/AG + rs9972359 TC/CC) vs. others	444/755=0.59	215/396=0.54	1.35 (1.05–1.74)	0.0197
	(rs543650 GG + rs9322343 AA) vs. others	214/403=0.53	445/748=0.59	0.71 (0.55–0.92)	0.0082

^I Odds ratio (95% confidence interval)