Original Paper

Human Heredity

Hum Hered 2011;72:121–132
DOI: 10.1159/000330778

Received: February 18, 2011
Accepted after revision: July 11, 2011
Published online: October 11, 2011

AUC-RF: A New Strategy for Genomic Profiling with Random Forest

M. Luz Calle[a] Victor Urrea[a] Anne-Laure Boulesteix[c] Nuria Malats[b]

[a]Systems Biology Department, University of Vic, Vic, and [b]Centro Nacional de Investigaciones Oncológicas, Madrid, Spain; [c]Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Munich, Germany

Key Words
AUC · Gene identification · Genomic profile · Random forest · ROC curve · Variable selection

Abstract
Objective: Genomic profiling, the use of genetic variants at multiple loci simultaneously for the prediction of disease risk, requires the selection of a set of genetic variants that best predicts disease status. The goal of this work was to provide a new selection algorithm for genomic profiling. ***Methods:*** We propose a new algorithm for genomic profiling based on optimizing the area under the receiver operating characteristic curve (AUC) of the random forest (RF). The proposed strategy implements a backward elimination process based on the initial ranking of variables. ***Results and Conclusions:*** We demonstrate the advantage of using the AUC instead of the classification error as a measure of predictive accuracy of RF. In particular, we show that the use of the classification error is especially inappropriate when dealing with unbalanced data sets. The new procedure for variable selection and prediction, namely AUC-RF, is illustrated with data from a bladder cancer study and also with simulated data. The algorithm is publicly available as an R package, named AUCRF, at http://cran.r-project.org/.

Introduction

In the last few years, a large number of association studies have been carried out with the goal of studying the inherited genetic basis of common diseases. Typically, candidate or tag single nucleotide polymorphisms (SNPs) in candidate genes or randomly distributed SNPs across the genome in genome-wide association studies (GWAS) are genotyped for cases and controls with the wish that those SNPs that are in linkage disequilibrium (LD) with a causal variant will show some signal of association with the phenotype. The success of this strategy will depend on both biological and statistical reasons: the genetic architecture of the disease, controlled by rare high-penetrant mutations, rare disease-causing variants, common susceptibility alleles [1] or a combination of these situations, will affect the power of the specific study design and the statistical approach chosen for analysis. Specially challenging is the study of genetic diseases controlled by common susceptibility alleles, since, as observed in many recent GWAS, common variants are usually low penetrant, with low detectable odds ratios (ORs), typically <1.5, and with little predictive value. Any polymorphism usually explains only 1–8% of the overall disease risk in the population, but the additive effect of several such risk

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0001–5652/11/0722–0121$38.00/0

Accessible online at:
www.karger.com/hhe

M. Luz Calle
Systems Biology Department, University of Vic
Sagrada Familia 7
ES–08500 Vic (Spain)
Tel. +34 93 881 6179, +34 93 881 5519, E-Mail malu.calle@uvic.cat

factors could make up the 20–70% of the overall disease risk that is attributed to genetic factors in most common diseases [2]. In this context, genomic profiling (the use of genetic variants at multiple loci simultaneously for the prediction of disease risk) is of primary interest.

Several statistical and data-mining alternatives have been proposed to explore complex patterns of genetic susceptibility involving multiple loci. Some examples are the multifactor dimensionality reduction method [3], the model-based multifactor dimensionality reduction method [4, 5], or approaches based on random forest (RF) [6] and support vector machine [7]. In this paper we will centre on the RF methodology, a classification algorithm developed by Leo Breiman [8] consisting in the aggregation of multiple classification trees generated from bootstrap samples. The use of RF is increasingly common in genetic epidemiology, and its behaviour in different scenarios including LD has been extensively studied in the last few years.

Genomic profiling consists in the selection of the set of genetic variants that best predicts the disease status and RF can be used for performing this selection process. Diaz-Uriarte and de Andrés [9], in the context of gene expression studies, proposed a backward elimination method for obtaining the optimal subset of variables providing the lowest overall classification error. Their method is implemented in the R package varSelRF and as a web-tool called GeneSrF [10]. Diaz-Uriarte's [10] variable selection method has proven to be useful and reliable in several gene expression studies [11, 12]. However, this method has two main drawbacks or limitations that are especially manifest when analyzing unbalanced data sets: (1) by default, it performs the most voted class RF prediction strategy (described below in the Methods section), and (2) it relies on the use of the out-of-bag (OOB) classification error rate (ER) of RF. On the one hand, as we will illustrate with an example in the Results section, for unbalanced data sets the most voted class strategy tends to classify almost all individuals in the largest class. On the other hand, ER mixes up false-positive (FP) and false-negative (FN) results which can give a false impression of accuracy in unbalanced samples (for instance, in a sample with 80% controls and 20% cases, a method that correctly classifies all controls but classifies cases randomly, ER will be only 10%).

In this paper, we propose an adaptation of the varSelRF approach that overcomes these limitations. The proposed algorithm, namely AUC-RF, uses the receiver operating characteristic (ROC) curve and the area under this curve (AUC) as the predictive accuracy of RF and then selects the set of variables with the highest AUC value. The goal of this work is twofold: to establish AUC as a preferable accuracy measure for RFs and to provide a modified selection algorithm focusing on AUC instead of ER. In the context of a genetic study, the AUC-RF algorithm can be used for genomic profiling, i.e. for identifying the set of variants with the highest combined predictive value of individual risk of disease.

We focus on variable selection for prediction which is different from variable selection for gene finding. In gene finding, the selection is based on statistical significance (joint or marginal) while for prediction the selection should be based on the predictive accuracy of the selected set of variables. The main goal of variable selection for gene finding, using for instance (penalized or not) logistic regression, is to minimize the number of false positives (FPs); the result is a small set of candidates that achieve the established significance threshold. It has been proven that, in general, this set of significant candidates has poor predictive power and that a better strategy for prediction is to be less restrictive and to build a genomic profile with the most promising candidates even if it is clear that this larger set will include many FPs [13]. This approach was used successfully in a GWAS in schizophrenia [14] where the SNPs were selected based on an extremely liberal threshold ($p < 0.5$). Our approach is similar to this in the sense that AUC-RF will select a large number of variables, including both associated and non-associated SNPs. The larger the number of associated SNPs included in the set, the higher the predictive accuracy of the model. The difference (advantage in some settings) is that in AUC-RF the selection is based on the importance of each variable in an RF which allows capturing nonlinear associations.

This work was motivated by the Spanish Bladder Cancer EPICURO Study. The goal was to examine the contribution of the inflammation pathway on bladder carcinogenesis, and a stratified analysis by tobacco smoking risk group was also of interest. While the whole sample was approximately balanced, the stratified samples were strongly unbalanced and, in this case, the use of RF for variable selection based on ER was not satisfactory (see Results section). This was the basis for the new strategy proposed in this paper for variable selection using RF. A simulation analysis has been conducted to evaluate the effectiveness of the AUC-RF method for selecting causative SNPs. Balanced and unbalanced scenarios have been simulated considering different numbers of causal SNPs, relative risks, minor allele frequencies (MAF) and disease prevalence.

Though RF can be used for both discrete and continuous dependent variables, we will explain the approach for a binary dependent variable representing case/control status in the context of a case-control study.

Methods

AUC-RF Algorithm for Variable Selection

The AUC-RF algorithm consists of the following 4 stages (a detailed technical description of stages 1–4 is given later):

Stage 1: Iterative Elimination Process. A first RF is built, using all predictor variables, which provides the ranking of the variables. In the subsequent steps, a fraction of the less important variables according to the initial ranking is eliminated (by default 20%). RF is built with the remaining variables and the OOB-AUC of the reduced model is computed. This is repeated until the number of remaining variables is less or equal than a specified value.

Stage 2: Visual Representation of the Elimination Process. The elimination process is visualized with a curve describing the OOB-AUC value of the different RFs (y-axis) as a function of the number of predictor variables (x-axis).

Stage 3: Selection of the Optimal Set of Predictors. The optimal set of predictive variables is considered the one giving rise to the RF with the highest OOB-AUC, denoted by OOB-AUC_{opt}. The number of selected predictors is denoted by k_{opt}.

Stage 4: Predictive Accuracy and Probability of Selection. The obtained OOB-AUC_{opt} value cannot be considered as the genuine predictive accuracy of the selected variables on a new data set. It is inflated by the fact that it is measured on the same training data set that has been used for the selection process. A correction of this overoptimism is required. Also of special concern is the robustness of the rankings and, consequently, of the selected variables [15]. AUC-RF deals with these two important issues by performing a repeated cross-validation analysis. The results of this analysis provide a corrected estimation of the predictive accuracy of the selected variables and an estimate of the probability of selection for each variable [16]. A detailed description is given in the online supplementary file (www.karger.com/doi/10.1159/000330778).

RF Importance Measures

RF, as implemented in the R-package *randomForest* available at http://cran.r-project.org/, provides two different importance measures, mean decrease accuracy (MDA) and mean decrease Gini (MDG). MDA quantifies the importance of a variable by measuring the change in OOB prediction accuracy when the values of the variable are randomly permuted compared to the original observations. The Gini index or Gini impurity of a node in a tree provides a measure of the heterogeneity in cases and controls of the node. The Gini impurity is minimum (zero) when the node is completely homogeneous (the node contains only cases or only controls). For a dichotomous variable (case/control), the Gini index is given by $I = 2p(1 - p)$, where p is the proportion of cases in the node. MDG is the sum of all decreases in Gini impurity due to a given variable (when this variable is used to form a split in RF) normalized by the number of trees.

Strobl et al. [17, 18] studied different mechanisms that can induce bias in the RF importance measures and Calle et al. [19] explored stability of these measures. On the one hand, both the MDG and MDA importance measures may be biased in the case of variables with different scales or in the case of categorical variables with different numbers of categories [15], but in the context of SNP data analysis (almost) all variables are three categorical. On the other hand, in terms of robustness, the ranks based on the MDA provide very unstable results [17]. More research is needed to elucidate the respective advantages and inconveniences of MDG and MDA in general. However, in the context considered, our preliminary study has clearly shown that MDA performs consistently and substantially worse than MDG, probably because of its high instability (data not shown). We will thus use MDG in this paper for both the bladder cancer analysis and the simulation study.

RF Prediction and AUC Computation

The individual class prediction using RF is based on what are called the votes. The principle is that each tree 'votes' for a class and that the predicted class of an individual is finally the class with the most votes. However, the voting procedure differs depending on whether one wants to compute the so-called OOB-ER or rather make predictions for new individuals from a test data set (DT). If the goal is to compute the OOB-ER, those trees for which an individual was OOB (i.e. was not used to build the tree) contribute with a vote to the predictive class for this individual. For dichotomous class prediction (y = 0, y = 1), the votes are two variables (v_0, v_1), where v_0 is the number of votes for class y = 0 and v_1 is the number of votes for class y = 1. The total number $v_0 + v_1$ is the number of trees for which the individual was OOB: approximately a third of the total number of trees when 'replace = TRUE' is used. The OOB-ER is then defined as the proportion of individuals with predicted class different from the true class. If the goal is to predict new individuals from DT, the procedure is similar but in this case all trees in the RF contribute with a vote ($v_0 + v_1 = n_{tree}$), since the individual was never used to build the trees. The default prediction procedure is to predict the most voted class and to provide the OOB-ER, which is used in Diaz-Uriarte's procedure. Alternatively, AUC-RF explores the predictive accuracy of RF through its ROC curve and the corresponding AUC [20]. The AUC-RF procedure computes the AUC based on OOB predictions, similarly to the OOB-ER, hence the notation 'OOB-AUC'. Each individual is characterized by the numbers v_0 or v_1 of trees predicting y = 0 and y = 1, respectively. The ROC curve plots sensitivity against 1 – specificity and can be obtained by varying the cut-off, c, in the prediction procedure based on the votes. The *randomForest* package allows to specify the cut-off as a vector $(1 - c, c)$ and then predicts an individual as $\hat{Y} = 1$ if $\nu_0 \cdot c < \nu_1(1 - c)$. The most voted class strategy corresponds to $c = 0.5$. The OOB-AUC can be calculated directly from the mean rank of the cases, denoted by $\bar{r}_1$, as

$$AUC = \frac{1}{n_0}\left(\bar{r}_1 - \frac{n_1}{2} - \frac{1}{2}\right)$$

where n_1 and n_0 are the number of cases and controls, respectively, and the ranks are based on the proportion of trees yielding y = 1, that is $v_1/(v_0 + v_1)$ [21].

Application

The Spanish Bladder Cancer EPICURO Study is a case-control study conducted in 18 hospitals in 5 areas in Spain (Asturias, Bar-

celona metropolitan area, Vallès/Bages, Alicante, and Tenerife) aiming at evaluating the role of both genetic and environmental factors in bladder carcinogenesis. Eligible cases were aged 21–80 years and had newly diagnosed, histologically confirmed carcinoma of the urinary bladder from 1998 to 2001. Patients who had a previous diagnosis of cancer of the lower urinary tract were not eligible for the study, as were patients with bladder tumors secondary to other malignancies. Controls were selected from patients admitted to participating hospitals with diagnoses thought to be unrelated to the exposures of interest, such as tobacco use, and were individually matched to the cases for age at interview within 5-year categories, sex, ethnic origin and region. In this paper, we centre our attention on the analysis of the joint effect of multiple genes in the inflammation pathway on bladder carcinogenesis for which information on 282 SNPs genotyped in a total of 108 genes in this pathway is available. After excluding patients with >20% missing genotypes, the available sample for analysis consists of 1,150 cases and 1,149 controls. The remaining missing genotypes were imputed using function *rfimpute* provided in the *randomForest* library. Smoking is the most important risk factor for bladder cancer, and gene-smoking interactions have been reported [22, 23]. For this reason, we were interested in performing a stratified analysis by tobacco smoking risk group (current smokers, former smokers and never smokers).

Simulation Study

Linear Effects

We performed a simulation study with the goal of investigating the performance of the proposed AUC-RF method for selecting variables with predictive capacity. In this simulation, we did not consider the method by Diaz-Uriarte since it is not suitable for an unbalanced data set (see the results on the bladder cancer study in the Results section). We generated a set of k causal SNPs and 1,000 – k non-causal SNPs. We followed a strategy similar to Janssens et al. [24] for simulating the causal SNPs that assumed independence and an approximate multiplicative genetic model. All causal SNPs were assumed to be in Hardy-Weinberg equilibrium, to have the same effect size on the response and the same genotype frequencies. The difference from Janssens's approach is that for each causal SNP we fixed the heterozygous relative risk (RR1) instead of the OR. We assumed that the minor homozygous relative risk is $RR2 = RR1^2$. We investigated the role of disease prevalence (p = 0.01, 0.1, 0.2, 0.3), effect size (RR1 = 1.1, 1.3, 1.5), MAF (= 0.1, 0.2, 0.3) on balanced (n_0 = number of controls = n_1 = number of cases = 2,000) and unbalanced (n_0 = 3,000 and n_1 = 1,000) data sets. The number of causal SNPs was k = 10, 50 and, for RR1 = 1.1, also k = 100. This yields a total of 192 scenarios.

For each scenario, we generated two data sets, a learning data set and DT that was used for validation of the predictive accuracy of the selected set of SNPs. We performed the AUC-RF feature selection algorithm and kept the percentage (Pc) of causal SNPs that AUC-RF picks up and the predictive accuracy of the selected set of SNPs on DT, denoted by test AUC. This predictive accuracy depends on the ability of the algorithm to identify the causal SNPs but also on the predictive capacity of the causal SNPs. Thus, we also computed the predictive ability of the causal SNPs as follows; each individual is assigned a risk score given by

$$\sum_{j=1}^{k}\left(1\left\{\mathbf{G}_j=1\right\}\cdot\log OR1_j+\mathbf{1}\left\{\mathbf{G}_j=2\right\}\cdot\log OR2_j\right)$$

where OR1 is the OR of heterozygous versus major homozygous and OR2 is the OR of minor homozygous versus major homozygous. We computed the AUC of predictions based on the above risk score, denoted by score AUC. The score AUC can be seen as the best empirical predictive accuracy provided by the causal SNPs, if they were known, and will be used as a reference for interpreting the observed predictive accuracy of the AUC-RF method. We repeated this process 100 times for each scenario and averaged the results over the 100 replications.

Nonlinear Effects

The advantage of AUC-RF over logistic regression may be more manifest under a nonlinear model. We have used the simulated data sets at http://discovery.dartmouth.edu/epistatic_data/ for showing the improvement in RF over logistic regression for selecting the causal SNPs in the absence of main effects. These data sets have 1,000 variables, the first 2 being functional through an epistatic effect but without exhibiting a marginal main effect. The remainder (998 variables) were randomly generated. The case-control label is in the last column. Sample sizes include 200, 400, 800 and 1,600, but we have only considered the data sets corresponding to a sample size of 1,600 (800 cases and 800 controls). Different scenarios are considered for different values of MAF (0.2, 0.4) and heritability (0.4, 0.3, 0.2, 0.1, 0.05, 0.025, 0.01). These data sets have been used for evaluating the performance of several methods for exploring gene-gene interactions [25, 26].

Results

Application: The Spanish Bladder Cancer EPICURO Study

We use the never-smoker group, an unbalanced data set consisting of 426 controls and 209 cases, for illustration of the proposed methodology, AUC-RF, and for comparison with varSelRF by Diaz-Uriarte.

The backward elimination process performed by the varSelRF algorithm is depicted in figure 1. As anticipated, the use of the most voted classification strategy and the OOB-ER provides unsatisfactory results in the unbalanced non-smoker data set. The first RF, considering all variables, results in an OOB-ER of 0.32. Though, in some contexts, a predictive error of 32% could be acceptable, in this case, this value only reflects the proportion of cases in the sample, which are almost all incorrectly classified as controls. Indeed, all 426 controls are predicted as controls (ER = 0) but only 12 out of the total 209 cases are classified as cases (ER = 0.94). A similar behaviour is observed for the subsequent RF built in the backward elimination process, providing always an OOB-ER around 0.3 and, consequently, an OOB-ER curve almost flat, which is not useful for identifying the optimal subset of predictors. In this case, the varSelRF feature selection algorithm selects only 3 variables providing OOB-ER = 0.31.

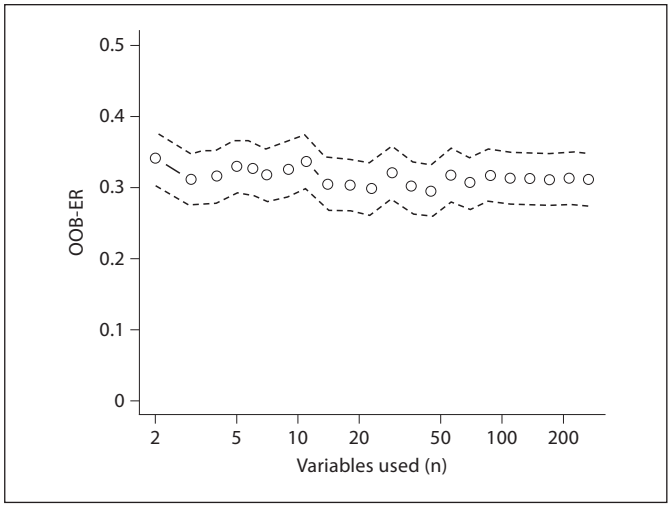

Fig. 1. varSelRF backward elimination procedure.

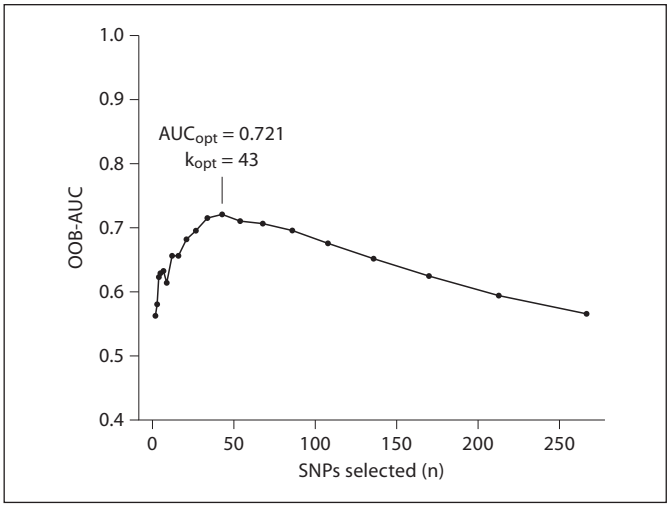

Fig. 2. AUC-RF backward elimination procedure.

The backward elimination process performed by the AUC-RF algorithm using the MDG importance measure can be visualized in figure 2. The points in the curve correspond to the OOB-AUC of consecutive RF obtained with the remaining variables, after the less important variables were removed. They were obtained from the right (all variables) to the left (1 variable). The optimal OOB-AUC is provided by the top 43 more important variables, giving an OOB-AUC_{opt} equal to 0.721. Correction for overfitting was performed with a 5-fold cross-validated (CV) analysis and CV-AUC = 0.56 was obtained.

We also performed the AUC-RF analysis for smokers and former smokers (data not shown) and for the whole sample (without stratifying for smoking). The obtained CV-AUC in each group was 0.54, 0.55 and 0.56, respectively. These results reflect the difficulty in obtaining a useful genomic profile for bladder cancer risk.

An important concern of selection methods, especially when they are based on rankings, is the robustness of the rankings and, consequently, of the selected set of SNPs. It is possible that different sets of variables provide practically the same predictive accuracy. For this reason, it is very important to provide the list of selected variables together with a measure of robustness of this selection. The AUC-RF algorithm implements a repeated CV process that provides the percentage of times that each variable has been selected. In this data set, we repeated 20 times a 5-fold CV process. Table 1 provides the list of the most important SNPs that were selected by AUC-RF at least 70% of the times. We can see that the selection of this set of 18 SNPs is very robust, with the top 2 being selected almost every time. We compared these results with the results of a univariate analysis for every SNP using logistic regression and three genetic models (dominant, recessive and co-dominant; data not shown). Only 3 SNPs achieve significance after adjusting for multiple testing and all 3 belong to the optimal set selected by AUC-RF (they are indicated with an asterisk in table 1), the first 2 follow a dominant genetic model and the 3rd follows a recessive model. This shows that AUC-RF is able to identify those SNPs with a significant marginal effect. We want to emphasize that the goal of this analysis is to show the ability of the AUC-RF to select an SNP signature in the context of an unbalanced case-control study. To this end, we used a subset of SNPs not representative of the whole set genotyped by the bladder cancer study. Thus, the list of the selected SNPs cannot be taken as a definitive result of the inflammatory genetic susceptibility of this neoplasm but as an example of the potential of the bioinformatic tool.

Simulation Study

Linear Effects

We summarize the results of the simulation study in terms of the percentage of selected causal SNPs, denoted by Pc, in tables 2–4. The predictive accuracy of the selected set of SNPs on DT, denoted test AUC, is reported in tables 5–7. The score AUC is also provided in parentheses as a reference value of the maximum predictive

Table 1. Most important SNPs, MDG and probability of selection (P)

Gene (SNP No.)	MDG	P
abca1 (04)	3.7387	1
masp1 (53)*	2.6158	0.99
ephx2 (04)	2.6030	0.96
il10 (17)	2.1866	0.89
lta (04)	2.0868	0.89
fcgr2a (01)*	2.3368	0.88
ptgs2 (05)	2.0644	0.85
ccr2 (02)*	1.8030	0.80
csf1r (05)	2.0657	0.79
mbl2 (12)	1.9348	0.78
gdf15 (02)	1.8225	0.77
alox5 (10)	1.6049	0.77
tlr2 (04)	1.7858	0.74
il4r (10)	1.6247	0.74
cd86 (02)	1.6072	0.71
alox5 (28)	1.7345	0.70

Table 2. Percentage (Pc) of selected causal SNPs for RR1 = 1.5 in balanced and unbalanced data sets at different prevalences (0.01, 0.1, 0.2, 0.3)

	Pc: balanced				Pc: unbalanced			
	0.01	0.1	0.2	0.3	0.01	0.1	0.2	0.3
k = 10								
MAF = 0.1	99.1	99.6	99.9	100	99.4	99.8	99.8	100
MAF = 0.2	100	100	100	100	100	100	100	100
MAF = 0.3	100	100	100	100	100	100	100	100
k = 50								
MAF = 0.1	96.8	94.3	94.3	94.9	98.1	96.1	96.3	97
MAF = 0.2	100	99.8	99.4	99.3	100	99.7	99.5	99.3
MAF = 0.3	100	100	99.9	99.8	100	99.9	99.7	99.7

k = Number of causal SNPs.

Table 3. Percentage (Pc) of selected causal SNPs for RR1 = 1.3 in balanced and unbalanced data sets at different prevalences (0.01, 0.1, 0.2, 0.3)

	Pc: balanced				Pc: unbalanced			
	0.01	0.1	0.2	0.3	0.01	0.1	0.2	0.3
k = 10								
MAF = 0.1	62.6	72	83.8	93.4	74.9	81.4	90.6	96
MAF = 0.2	96.8	98.5	99.9	99.9	98	98.4	99.6	99.9
MAF = 0.3	99.9	100	100	100	99.5	99.9	100	100
k = 50								
MAF = 0.1	57.3	59.2	64.9	72.1	71.7	72.4	76	83.3
MAF = 0.2	94.2	92.9	92.7	94.3	96.5	94.8	94.8	96.5
MAF = 0.3	99.3	98.6	98.3	98.3	98.9	97.5	97.7	98.2

k = Number of causal SNPs.

Table 4. Percentage (Pc) of selected causal SNPs for RR1 = 1.1 in balanced and unbalanced data sets at different prevalences (0.01, 0.1, 0.2, 0.3)

	Pc: balanced				Pc: unbalanced			
	0.01	0.1	0.2	0.3	0.01	0.1	0.2	0.3
k = 10								
MAF = 0.1	1.9	2.4	4.1	5.8	9.6	11.6	13.8	16.3
MAF = 0.2	14.8	18.8	21.8	27.7	31.5	31.6	40.4	50
MAF = 0.3	36.8	43.8	50.9	61.7	44.7	51.6	58.3	67.2
k = 50								
MAF = 0.1	2.2	2.4	3.6	5	9.8	11.1	13	17
MAF = 0.2	13.4	16.3	20.2	27.3	30	34.1	37.7	45.3
MAF = 0.3	34.5	39.7	45.1	53.4	44	46.8	53.7	60.6
k = 100								
MAF = 0.1	2	2.5	3.2	4.4	9.3	10.1	13.1	15.7
MAF = 0.2	13.6	15.5	18.5	22.4	29.3	31.9	36.5	40.7
MAF = 0.3	33.7	36.2	40.5	47	44.5	46.4	50.4	54.8

k = Number of causal SNPs.

accuracy of the causal SNPs. In order to visualize some of the obtained results, figures 3 and 4 show the results for k = 50 and a balanced data set.

The percentage, Pc, of causal SNPs that AUC-RF is able to pick up is mainly affected by the effect size, followed by MAF and the disease prevalence (fig. 3; tables 2–4). For RR1 = 1.5, the percentage Pc is almost 100% in all cases, i.e. all causal SNPs are identified. When RR1 reduces to 1.3 the efficacy remains for MAF = 0.3 and 0.2 but reduces considerably for MAF = 0.1. For RR1 = 1.1, the percentage of selected causal SNPs reduces drastically to 30–40% for MAF = 0.3, 10–20% for MAF = 0.2 and it is almost 0% for MAF = 0.1. A slight effect of the disease prevalence is observed in some situations: for RR1 = 1.3

Table 5. Test AUC of the selected SNPs and score AUC (in parentheses) of the causal SNPs for RR1 = 1.5 at different prevalences (0.01, 0.1, 0.2, 0.3)

	Predictive accuracy: balanced dataset				Predictive accuracy: unbalanced dataset			
	0.01	0.1	0.2	0.3	0.01	0.1	0.2	0.3
k = 10								
MAF = 0.1	0.63 (0.66)	0.64 (0.67)	0.66 (0.69)	0.69 (0.71)	0.61 (0.66)	0.63 (0.67)	0.65 (0.69)	0.68 (0.71)
MAF = 0.2	0.67 (0.7)	0.69 (0.72)	0.71 (0.74)	0.73 (0.76)	0.66 (0.7)	0.68 (0.72)	0.7 (0.74)	0.73 (0.76)
MAF = 0.3	0.7 (0.73)	0.71 (0.74)	0.73 (0.76)	0.76 (0.78)	0.69 (0.73)	0.7 (0.74)	0.72 (0.76)	0.75 (0.78)
k = 50								
MAF = 0.1	0.78 (0.81)	0.77 (0.82)	0.78 (0.83)	0.79 (0.85)	0.77 (0.81)	0.76 (0.81)	0.77 (0.83)	0.79 (0.85)
MAF = 0.2	0.85 (0.88)	0.84 (0.87)	0.84 (0.88)	0.85 (0.89)	0.84 (0.87)	0.83 (0.87)	0.83 (0.88)	0.84 (0.89)
MAF = 0.3	0.88 (0.9)	0.86 (0.89)	0.86 (0.9)	0.87 (0.91)	0.87 (0.9)	0.85 (0.89)	0.85 (0.89)	0.86 (0.91)

k = Number of causal SNPs.

Table 6. Test AUC of the selected SNPs and score AUC (in parentheses) of the causal SNPs for RR1 = 1.3 at different prevalences (0.01, 0.1, 0.2, 0.3)

	Predictive accuracy: balanced data set				Predictive accuracy: unbalanced data set			
	0.01	0.1	0.2	0.3	0.01	0.1	0.2	0.3
k = 10								
MAF = 0.1	0.54 (0.6)	0.56 (0.61)	0.58 (0.62)	0.6 (0.64)	0.54 (0.6)	0.55 (0.61)	0.57 (0.62)	0.59 (0.64)
MAF = 0.2	0.59 (0.63)	0.6 (0.64)	0.62 (0.66)	0.64 (0.68)	0.58 (0.63)	0.59 (0.64)	0.61 (0.66)	0.63 (0.68)
MAF = 0.3	0.61 (0.65)	0.62 (0.66)	0.64 (0.68)	0.66 (0.7)	0.6 (0.65)	0.61 (0.66)	0.63 (0.68)	0.65 (0.7)
k = 50								
MAF = 0.1	0.63 (0.71)	0.64 (0.72)	0.66 (0.74)	0.69 (0.76)	0.64 (0.71)	0.64 (0.72)	0.66 (0.73)	0.69 (0.76)
MAF = 0.2	0.73 (0.77)	0.73 (0.78)	0.75 (0.79)	0.76 (0.81)	0.72 (0.77)	0.72 (0.77)	0.74 (0.79)	0.76 (0.81)
MAF = 0.3	0.76 (0.8)	0.77 (0.8)	0.78 (0.82)	0.79 (0.84)	0.76 (0.8)	0.75 (0.8)	0.76 (0.81)	0.78 (0.83)

k = Number of causal SNPs.

and MAF = 0.1 and RR1 = 1.1 and MAF = 0.3, 0.2, and the larger the prevalence, the higher the percentage Pc.

A similar behaviour is observed in terms of predictive accuracy of the set of selected SNPs (fig. 4; tables 5–7). For RR1 = 1.5, the test AUC is very high (around 0.8–0.9), which corresponds to very accurate predictions. Indeed, the obtained test AUC after feature selection is very similar to the score AUC provided by all causal SNPs (given in parentheses). The effect of the genotype frequencies is observed, with MAF = 0.3 giving slightly better results than for MAF = 0.2 and better than for MAF = 0.1. Instead, the disease prevalence effect is not apparent in terms of test AUC. For RR1 = 1.3, the predictive accuracy is around 0.7–0.8 when MAF = 0.2 or 0.3 and around 0.65 when MAF = 0.1. The loss in predictive capacity (comparing the obtained test AUC with the score AUC) is more apparent for low values of MAF (around 7% when MAF = 0.1). In this case, the effect of disease prevalence is not apparent. For RR1 = 1.1, the predictive capacity of the selected set of SNPs is in general very low or non-existent. Note, however, that in this setting the score AUC given by all causal SNPs is also very low. Only for k = 100 and MAF = 0.3 we obtain more acceptable predictive values, around 0.6. This is in accordance with Janssens et al. [27] who state that a genomic profile from a set of causal SNPs with such a weak marginal effect on the phenotype will require a larger number of SNPs to jointly get a useful predictive accuracy. Indeed, looking

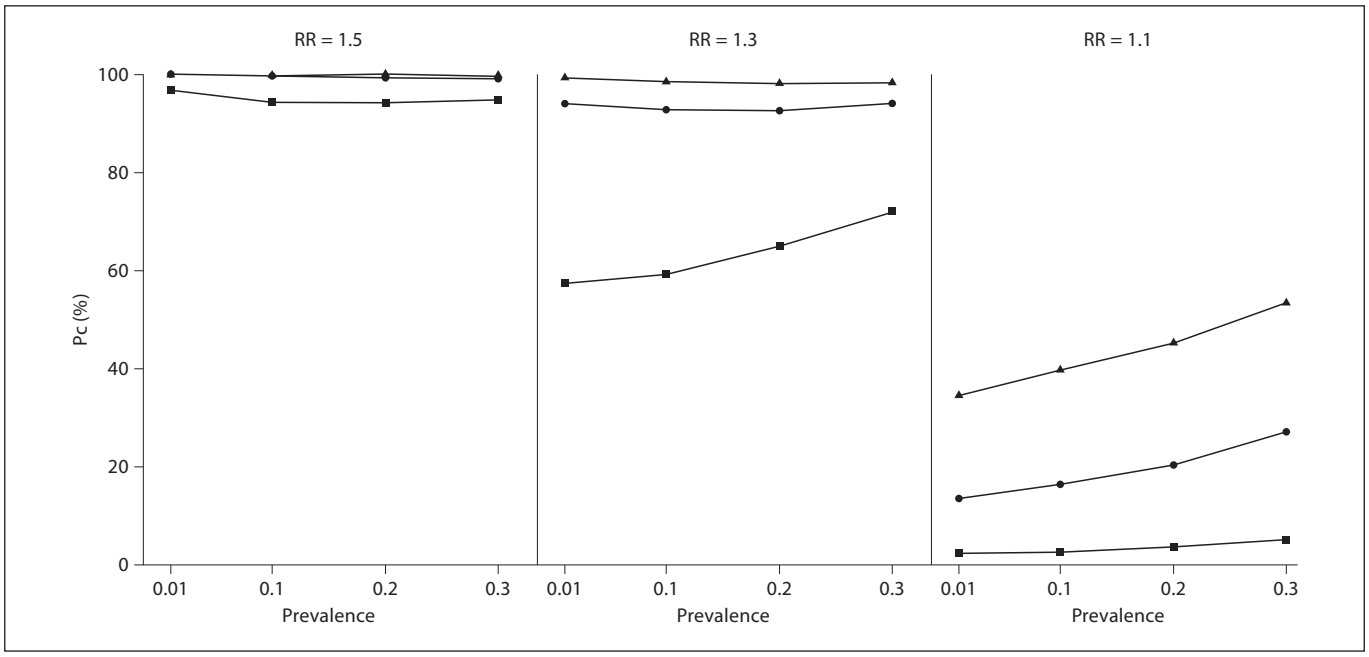

Fig. 3. Percentage of selected causal SNPs (Pc) for k = 50 and balanced data sets. ■: MAF = 0.3; ●: MAF = 0.2; ▲: MAF = 0.1.

Table 7. Test AUC of the selected SNPs and score AUC (in parentheses) of the causal SNPs for RR1 = 1.1 at different prevalences (0.01, 0.1, 0.2, 0.3)

	Predictive accuracy: balanced dataset				Predictive accuracy: unbalanced dataset			
	0.01	0.1	0.2	0.3	0.01	0.1	0.2	0.3
k = 10								
MAF = 0.1	0.5 (0.53)	0.5 (0.53)	0.5 (0.54)	0.5 (0.54)	0.5 (0.52)	0.5 (0.53)	0.5 (0.53)	0.5 (0.54)
MAF = 0.2	0.5 (0.54)	0.51 (0.55)	0.51 (0.55)	0.51 (0.56)	0.5 (0.54)	0.51 (0.54)	0.51 (0.55)	0.52 (0.56)
MAF = 0.3	0.51 (0.55)	0.51 (0.55)	0.52 (0.56)	0.52 (0.57)	0.51 (0.55)	0.51 (0.55)	0.51 (0.56)	0.52 (0.57)
k = 50								
MAF = 0.1	0.5 (0.56)	0.5 (0.57)	0.5 (0.58)	0.51 (0.59)	0.51 (0.55)	0.51 (0.56)	0.51 (0.57)	0.52 (0.58)
MAF = 0.2	0.52 (0.58)	0.52 (0.6)	0.53 (0.61)	0.54 (0.63)	0.52 (0.58)	0.53 (0.59)	0.54 (0.6)	0.55 (0.62)
MAF = 0.3	0.53 (0.6)	0.54 (0.61)	0.55 (0.63)	0.57 (0.65)	0.54 (0.59)	0.54 (0.61)	0.56 (0.62)	0.57 (0.64)
k = 100								
MAF = 0.1	0.51 (0.58)	0.51 (0.59)	0.51 (0.6)	0.51 (0.62)	0.51 (0.57)	0.52 (0.58)	0.52 (0.6)	0.53 (0.61)
MAF = 0.2	0.53 (0.62)	0.53 (0.63)	0.54 (0.65)	0.56 (0.67)	0.54 (0.61)	0.55 (0.62)	0.56 (0.64)	0.57 (0.66)
MAF = 0.3	0.56 (0.64)	0.57 (0.65)	0.58 (0.67)	0.6 (0.69)	0.57 (0.63)	0.57 (0.64)	0.59 (0.66)	0.61 (0.68)

k = Number of causal SNPs.

at tables 5–7, we can observe that the larger the number k of causal SNPs the larger the AUC (both score AUC and test AUC) in all settings. Instead, this effect of the number of causal SNPs is not observed in the efficacy of the AUC-RF method for detecting causal SNPs (tables 2–4). For instance, in table 5 when MAF = 0.1, the percentages Pc of identified causal SNPs are larger for k = 10 than for k = 50.

We have compared the new algorithm with genomic profiling using logistic regression (selecting those vari-

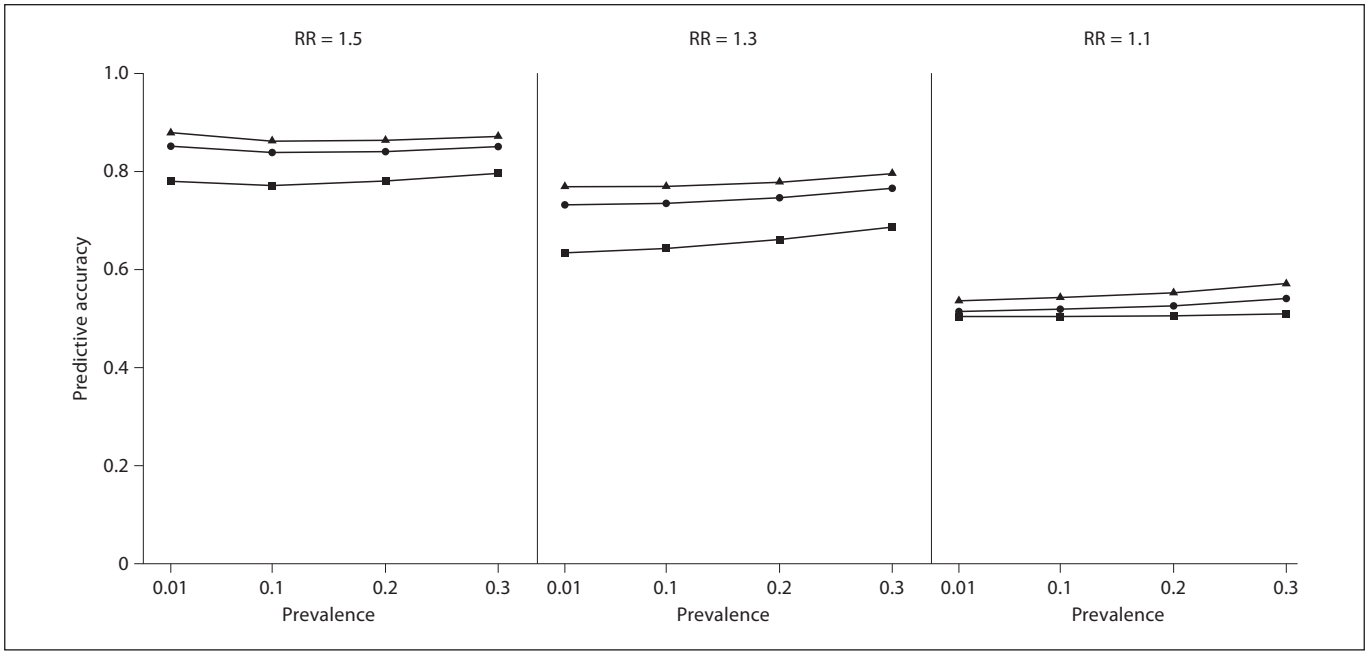

Fig. 4. AUC of the selected SNPs for k = 50 and balanced data sets. ■: MAF = 0.3; ●: MAF = 0.2; ▲: MAF = 0.1.

ables with $p < 0.05$ without adjusting for multiple testing) for the simulated data sets. The following plots, figures 5 and 6, provide the relative improvement in predictive accuracy of AUC-RF versus logistic regression with respect to the predictive accuracy of all causal SNPs in a DT: [AUC-RF (DT) – AUC_logistic (DT)]/ max AUC (DT). Figure 5 corresponds to balanced data sets and figure 6 to unbalanced data sets. The x-axis provides MAF (0.1, 0.2 and 0.3). Different values of prevalence provided very similar results; the plots provide the mean relative improvement for the different prevalence values.

We can see in both plots that when the marginal effect of the causal SNPs is strong (RR = 1.5 and RR = 1.3) the standard logistic regression approach is more effective (approximately a 5% relative improvement of logistic regression to AUC-RF in prediction accuracy). However, when the effect is very small (RR = 1.1) the logistic regression has a poorer performance in identifying causal SNPs than AUC-RF. In this case, the improvement in predictive accuracy of AUC-RF versus logistic regression is more manifest for unbalanced data sets and increases with the number of causal SNPs and MAF. The largest advantage (around 10% of relative improvement) is found in an unbalanced data set scenario, with k = 100 causal SNPs and MAF = 0.3.

Though the results show that there is not a 'universal best method' and that the advantage of one method over the other will depend on the specific situation, the fact that AUC-RF performs much better in the scenario with many SNPs with small effect is very promising. The known genetic variants for most common diseases up to now explain a small proportion of the disease risk. It has been hypothesized that part of the remaining disease risk is given by the joint effect of a large number of variants, with each one having a very low effect.

Nonlinear Effects

We applied AUC-RF and logistic regression for variable selection to the simulated data sets that contain two causal SNPs without main effects. AUC-RF selects the optimal set of variables as exposed in the Methods section. For logistic regression, the usual variable selection procedure is to select those SNPs with a p value smaller than an established significance level α. The results for $\alpha = 0.05$ (not shown) indicated a very poor performance of logistic regression. Since AUC-RF tend to select a larger number of SNPs than logistic regression, one may argue that the observed advantage of AUC-RF over logistic regression is only due to the different numbers of selected SNPs. In order to make the results from both methods comparable, we provide in tables 8 and 9 the

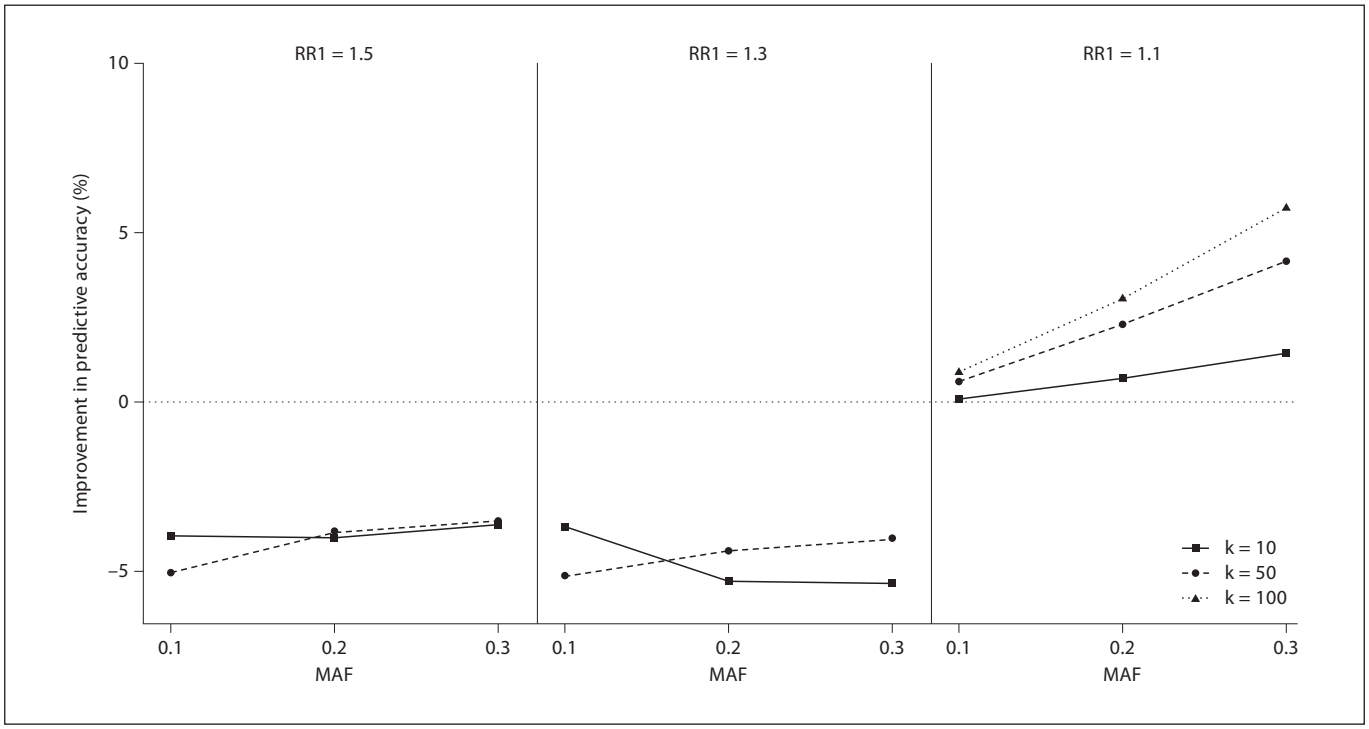

Fig. 5. Relative improvement in predictive accuracy of AUC-RF versus logistic regression with respect to the predictive accuracy of all causal SNPs in a balanced DT.

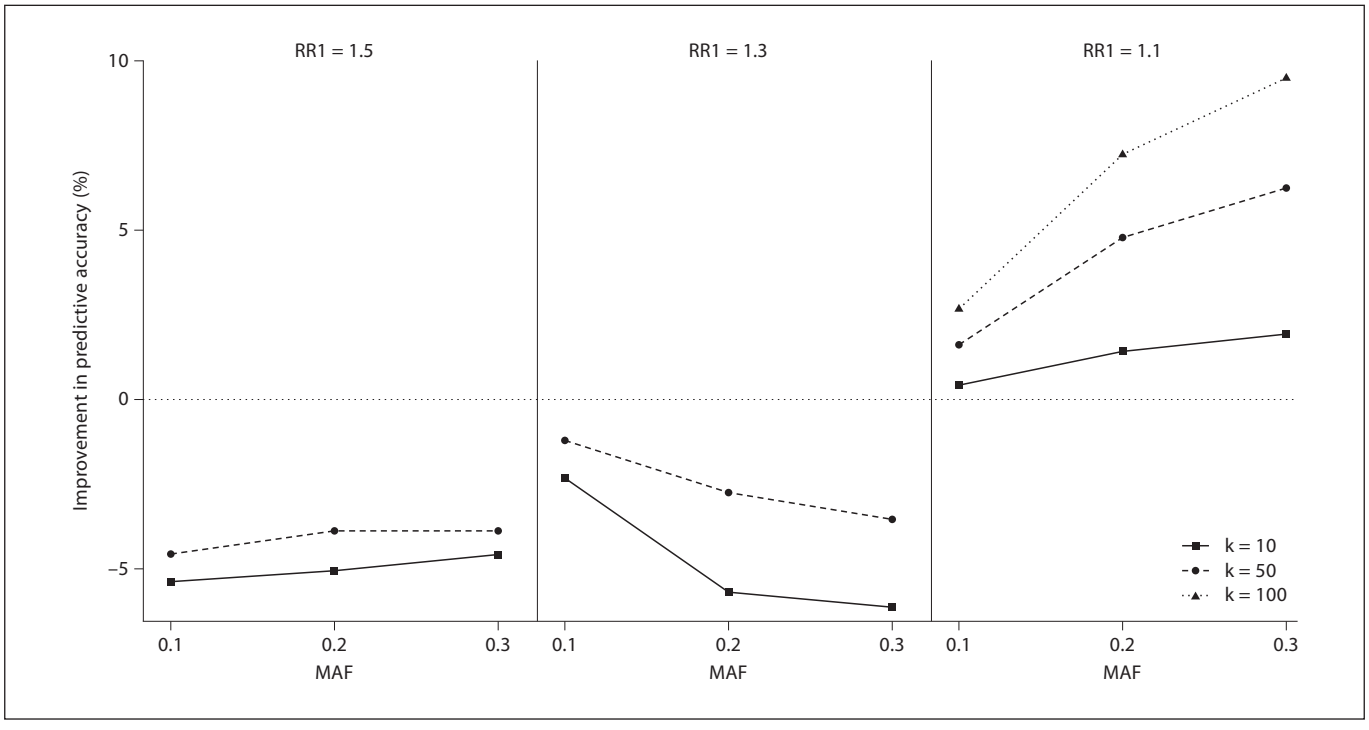

Fig. 6. Relative improvement in predictive accuracy of AUC-RF versus logistic regression with respect to the predictive accuracy of all causal SNPs in an unbalanced DT.

Table 8. Percentage of times that 2 causal SNPs are selected (p2) and percentage of times that only 1 causal SNP is selected (p1) using AUC-RF and logistic regression for variable selection, with the same number of selected SNPs in both methods and MAF = 0.2

Heritability	AUC-RF		Logistic regression	
	p2	p1	p2	p1
0.4	92	6	0	5
0.3	85	7	0	9
0.2	60	19	2	24
0.1	28	20	2	25
0.05	14	21	2	27
0.025	12	23	2	26
0.01	5	25	2	25

Table 9. Percentage of times that 2 causal SNPs are selected (p2) and percentage of times that only 1 causal SNP is selected (p1) using AUC-RF and logistic regression for variable selection, with the same number of selected SNPs in both methods and MAF = 0.4

Heritability	AUC-RF		Logistic regression	
	p2	p1	p2	p1
0.4	81	9	0	7
0.3	61	15	0	10
0.2	42	20	0	20
0.1	25	17	0	25
0.05	23	17	4	22
0.025	17	18	3	25
0.01	6	23	3	22

results of AUC-RF and logistic regression variable selection with the same number of selected SNPs (the number of selected SNPs, k_{opt}, is determined by AUC-RF and then we select the top k_{opt} SNPs with the smallest p values according to logistic regression). The performance of both methods is described through p2, the percentage of times that the 2 causal SNPs are selected, and p1, the percentage of times that only 1 causal SNP is selected, using AUC-RF and logistic regression. The results in tables 8 and 9 are conclusive: AUC-RF outperforms logistic regression in all scenarios. The power of logistic regression for selecting SNPs in the absence of main effects is null (p1 only reflects probability of random selection) while AUC-RF is able to detect them with high probability when the heritability parameter is not very small.

Discussion

In this work, we propose a new feature selection strategy using RF which is based on optimization of the AUC in a backward elimination process which provides the set of variables that best predicts the outcome. We have illustrated with data from a real bladder cancer study that the default RF most voted class prediction strategy together with the use of ER provides unsatisfactory results in unbalanced data sets. However, even for balanced data sets, the use of the AUC is preferable to ER because ER is dependent on the case/control rates in the sample which not necessarily represent the case/control rates in the population. In comparison to single-cut-off approaches, our AUC-based approach has the major advantage that it does not depend on a specific arbitrary cut-off value but implicitly incorporates all possible cut-off values into a single measure of accuracy. The use of the AUC is especially appealing after the recent increasing interest in this measure in the molecular and genetic epidemiology field [24, 27–33]. Wray et al. [21] related the maximum value of the AUC of a genetic risk predictor model with the heritability and prevalence of the disease. They proved that the maximum AUC is particularly constrained for more common or low heritability diseases. Moreover, the use of the AUC instead of ER as an accuracy measure does not induce any additional computational effort compared to standard RFs. Thus, our procedure could probably be easily integrated into a software implementing RFs for genome-wide data, e.g. the RandomJungle tool [34].

In real applications, it is very usual to have correlated SNPs due to LD. AUC-RF handles SNPs in LD exactly as RFs, hence it suffers from the same limitations of RF in this context, that is diminished variable importance for the true causal SNPs. In this context, the AUC-RF method can be combined with existing strategies for RF when SNPs are in LD [35].

In the proposed approach, the same initial ranking is used for all iterations in the backward elimination process. Jiang et al. [36] proposed a backward elimination strategy for variable selection using RF similar to Diaz Uriarte's method, the main difference being the recomputation of the importance of the remaining variables at each step of the elimination process. In our opinion, a potential drawback of this strategy is that it might accen-

tuate the over-fitting problem already existing in any elimination process. In future research, an improvement might also be obtained by using a permutation variable importance based on AUC decrease rather than accuracy decrease when ranking the variables at the beginning of the procedure.

Acknowledgments

This work was partially supported by grant MTM2008-06747-C02-02 from the Ministerio de Educación y Ciencia (Spain), grant 050831 from the Marató de TV3 Foundation, grant 2009SGR-581 from the AGAUR-Generalitat de Catalunya and the LMU-innovativ Project BioMed-S. Víctor Urrea is the recipient of a pre-doctoral FPU fellowship award from the Spanish Ministry of Education.

References

1 Fletcher O, Houlston RS: Architecture of inherited susceptibility to common cancer. Nat Rev Cancer 2010;10:353–361.
2 Ioannidis JPA: Genetic associations: false or true? Trends Mol Med 2003;9:135–138.
3 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.
4 Calle ML, Urrea V, Vellalta G, et al: Improving strategies for detecting genetic patterns of disease susceptibility in association studies. Stat Med 2008;27:6532–6546.
5 Calle ML, Urrea V, Malats N, van Steen K: mbmdr: an R package for exploring gene-gene interactions associated with binary or quantitative traits. Bioinformatics 2010;26:2198–2199.
6 Bureau A, Dupuis J, Falls K, et al: Identifying SNPs predictive of phenotype using random forests. Genet Epidemiol 2005;28:171–182.
7 Wei Z, Wang K, Qu H, et al: From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genet 2009;5:e1000678.
8 Breiman L: Random forests. Mach Learn 2001;45:5–32.
9 Diaz-Uriarte R, de Andrés SA: Gene selection and classification of microarray data using random forest. BMC Bioinformatics 2006;7:3.
10 Diaz-Uriarte R: GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. BMC Bioinformatics 2007;8:328.
11 Habermann JK, Doering J, Hautaniemi S, et al: The gene expression signature of genomic instability in breast cancer is an independent predictor of clinical outcome. Int J Cancer 2009;124:1552–1564.
12 Torri A, Beretta O, Ranghetti A, et al: Gene expression profiles identify inflammatory signatures in dendritic cells. PLoS One 2010;5:e9404.
13 Yip W, Lange C: Quantitative trait prediction based on genetic marker-array data, a simulation study. Bioinformatics 2011;27:745–748.
14 International Schizophrenia Consortium, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, et al: Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 2009;460:748–752.
15 Boulesteix AL, Slawski M: Stability and aggregation of ranked gene lists. Brief Bioinform 2009;10:556–568.
16 Pepe MS, Longton G, Anderson GL, et al: Selecting differentially expressed genes from microarray experiments. Biometrics 2003;59:133–142.
17 Strobl C, Boulesteix A-L, Zeileis A, et al: Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 2007;8:25.
18 Strobl C, Boulesteix A-L, Kneib T, et al: Conditional variable importance for random forests. BMC Bioinformatics 2008;9:307.
19 Calle ML, Urrea V: Letter to the editor: stability of random forest importance measures. Brief Bioinform 2011;12:86–89.
20 Pepe MS: The Statistical Evaluation of Medical Tests for Classification and Prediction. New York, Oxford, 2003.
21 Wray NR, Yang J, Goddard ME, et al: The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet 2010;6:e1000864.
22 García-Closas M, Malats N, Silverman D, et al: NAT2 slow acetylation, GSTM1 null genotype, and risk of bladder cancer: results from the Spanish bladder cancer study and meta-analyses. Lancet 2005;366:649–659.
23 Samanic C, Kogevinas M, Dosemeci M, et al: Smoking and bladder cancer in Spain: effects of tobacco type, timing, environmental tobacco smoke, and gender. Cancer Epidemiol Biomarkers Prev 2006;15:1348–1354.
24 Janssens AC, Aulchenko YS, Elefante S, et al: Predictive testing for complex diseases using multiple genes: fact or fiction? Genet Med 2006;8:395–400.
25 Moore JH, Gilbert JC, Tsai C-T, Chiang FT, Holden W, Barney N, White BC: A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. J Theor Biol 2006;241:252–261.
26 Wan X, Yang C, Yang Q, Xue H, Fan X, Tang N, Yu W: BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. Am J Hum Genet 2010;87:325–340.
27 Janssens AC, Moonesinghe R, Yang Q, et al: The impact of genotype frequencies on the clinical validity of genomic profiling for predicting common chronic diseases. Genet Med 2007;9:528–535.
28 Pepe MS, Janes H, Longton G, et al: Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol 2004;159:882–890.
29 Lu Q, Elston RC: Using the optimal receiver operating characteristic curve to design a predictive genetic test, exemplified with type 2 diabetes. Am J Hum Genet 2008;82:641–651.
30 Jakobsdottir J, Gorin MB, Conley YP, et al: Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. PLoS Genet 2009;5:e1000337.
31 Kraft P, Wacholder S, Cornelis MC, et al: Beyond odds – ratios communicating disease risk based on genetic profiles. Nat Rev Genet 2009;10:264–269.
32 Lu Q, Obuchowski N, Won S, et al: Using the optimal robust receiver operating characteristic (ROC) curve for predictive genetic tests. Biometrics 2010;66:586–593.
33 Moonesinghe R, Liu T, Khoury MJ: Evaluation of the discriminative accuracy of genomic profiling in the prediction of common complex diseases. Eur J Hum Genet 2010;18:485–489.
34 Schwartz D, König I, Ziegler A: On safari to Random Jungle: a fast implementation of random forests for high-dimensional data. Bioinformatics 2010;26:1752–1758.
35 Meng Y, Yu Y, Cupples LA, Farrer LA, Lunetta LK: Performance of random forest when SNPs are in linkage disequilibrium. BMC Bioinformatics 2009;10:78.
36 Jiang H, Deng Y, Chen H, et al: Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. BMC Bioinformatics 2004;5:81.