Anne-Laure Boulesteix, Andreas Bender,
Justo Lorenzo Bermejo, Carolin Strobl

Random forest Gini importance favors SNPs with large minor allele frequency

Technical Report Number 106, 2011
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Title: Random forest Gini importance favors SNPs with large minor allele frequency

Authors: Anne-Laure Boulesteix[1*], Andreas Bender[1], Justo Lorenzo Bermejo[2] and Carolin Strobl[3]
[*] To whom correspondence should be sent

[1] Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, 81377 Munich, Germany, Tel: +49 89 7095 7598, Fax: +49 89 7095 4491

[2] Department of Medical Biometry and Informatics, University of Heidelberg, Im Neuenheimer Feld 305, 69120 Heidelberg, Germany

[3] Department of Statistics, University of Munich, Ludwigstr. 33, 80539 Munich, Germany

Abstract:

The use of random forests is increasingly common in genetic association studies. The variable importance measure (VIM) that is automatically calculated as a by-product of the algorithm is often used to rank polymorphisms with respect to their association with the investigated phenotype. Here we investigate a characteristic of this methodology that may be considered as an important pitfall, namely that common variants are systematically favored by the widely used Gini VIM. As a consequence, researchers may overlook rare variants that contribute to the missing heritability. The goal of the present paper is three-fold: 1) to assess this effect quantitatively using simulation studies for different types of random forests (classical random forests and conditional inference forests, that employ unbiased variable selection criteria) as well as for different importance measures (Gini and permutation-based), 2) to explore the trees and to compare the behaviour of random forests and the standard logistic regression model in order to understand the statistical mechanisms behind the preference for common variants, and 3) to summarize our results and previously investigated properties of random forest VIMs in the context of association studies and to make practical recommendations regarding the methodological choice.

The codes implementing our study are available from the companion website:
http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/020_professuren/boulesteix/ginibias/

Keywords: random forest, variable selection bias, SNPs, genetic association study

Introduction

Random forests, originally suggested by Breiman (2001) ten years ago, have evolved to a standard statistical analysis tool in genetics. They are increasingly used in many genetic studies to rank genetic variants with respect to their association with a disease or trait of interest via the so-called variable importance measures (VIMs), to identify gene-gene interactions, or to investigate the predictive power of genetic data taking into account possible complex non-linear patterns (cf., e.g., Briggs et al 2010, Bureau et al 2005, Cleynen et al 2010, Heidema et al 2006, Liu et al 2011, Lunetta et al 2004, Roshan et al 2011, Weidinger et al 2005). Methodological developments of random forests and new implementations with focus on genetic applications have also been recently addressed in genetic analysis workshops (Schwarz et al 2007, Yang and Gu 2009) and published in genetics or bioinformatics journals including the European Journal of Human Genetics (de Lobel et al, 2010 ; Strobl et al 2008 ; Schwartz et al 2010).

It is now widely known that the commonly used random forest VIMs are biased in favour of the categorical variables with more categories (Strobl et al, 2007). This is because variables with many categories are more likely to yield a good split "by chance", even in the absence of association with the response. They are thus selected as splitting variables more often and/or earlier in the trees, which leads to higher VIMs than those of variables with fewer categories. This bias affects the variable selection in classical random forest algorithms as well as the so-called "Gini VIM" in the case of non-informative predictors. Hothorn et al (2006) have suggested a random forest algorithm based on unbiased variable selection criteria. When these criteria – together with subsampling instead of bootstrap sampling – is used for computing a permutation VIM, this measure is unbiased as discussed in detail by Strobl et al (2007).

In the context of genetic association studies, it has been argued that this kind of bias is irrelevant, since single nucleotide polymorphisms (SNPs) have, by definition, three categories (for example, "AA", "AC" and "CC"). Based on resampling analyses, Calle and Urrea (2010) point out that the Gini VIM shows a better stability than the permutation VIM and consequently recommend its use. In a subsequent study on the stability of VIMs, Nicodemus (2011) suggests that the higher stability of Gini VIM compared to permutation VIM is attributable to a bias in favor of SNPs with large Minor Allele Frequency (MAF). In Nicodemus' resampling analyses, common variants receive high ranks consistently over the subsamples, while rare variants receive consistently low ranks, which induces the apparent stability.

The aim of the present paper is to provide further results and a deeper understanding of the statistical mechanisms responsible for the observations of Calle and Urrea (2010) and Nicodemus (2011). A series of simulations is conducted in order to quantitatively compare the behavior of different types of random forests and different VIMs for SNPs with different MAFs independently of stability issues, and explore the statistical mechanisms behind their behavior. We then summarize our results and previously investigated properties of random forest VIMs in the context of association studies and make recommendations regarding the methodological approach to be used.

Methods

Random forests

Random forests are a classification and regression method based on the aggregation of a large number of decision trees (Breiman 2001). In the most commonly used type of random forests, split selection is performed based on the so-called decrease of Gini impurity (DGI). This version of random forests is implemented in the package 'randomForest' (Breiman, Cutler, Liaw, & Wiener, 2010; Liaw & Wiener, 2002) available in the R system for statistical computing (R Development Core Team 2010), which we use in the simulations with all default parameters. In particular, the number of trees is set to ntree=500 and the number of candidate predictors considered at each split is set to the default value mtry=$p^{1/2}$, where p is the number of predictors. In the rest of this paper, this type of random forests will simply be denoted as "randomForest".

Although this is by far the most widely applied version, the randomForest method has an important pitfall. In the split selection process, predictors may be favoured or disfavoured depending on their scale of measurement or, in the case of categorical predictors, on their number of categories. For example, it has been demonstrated that predictors with many categories are selected more often than predictors with few categories independently of their association with the response. See Strobl et al (2007) for more details. To address this issue, Hothorn et al (2006) developed an alternative class of random forests which are based on conditional hypothesis testing. These random forests use an unbiased splitting criterion and do not share the above pitfall. We also consider this type of random forest in this study taking advantage of the function 'cforest' from the R package 'party' (Hothorn et al, 2010). The number of trees is set to ntree=500 and the number of candidate predictors at each split is set to mtry= $p^{1/2}$ again. Moreover, the p-value threshold acting as a stopping criterion is set to mincriterion=0. All other parameters are set to their default values. In the rest of this paper, this type of random forests will be denoted as “cforest”.

For both randomForest and cforest, these settings yield large trees with small terminal nodes. Additionally, all analyses (for both randomForest and cforest) are also performed with one-layer trees – also called “stumps”, i.e. small trees with only two terminal nodes.

Variable importance measures

For the randomForest method, we consider two types of variable importance measures: the mean decrease of Gini impurity (denoted as “Gini VIM”), and the unscaled permutation-based importance measure (“permutation VIM”) both implemented in the function 'importance' from the ‘randomForest’ package. See Strobl et al., 2007, for details on variable importance measures. For the cforest method, we consider the permutation VIM implemented in the function 'varimp' from the package ‘party’.

Simulation design: data generation

The simulated datasets include a binary phenotype Y and 200 genetically unlinked SNPs in Hardy-Weinberg Equilibrium: 50 SNPs with MAF=0.05 (SNPs 1 to 50), 50 SNPs with MAF=0.1 (SNPs 51 to 100), 50 SNPs with MAF=0.25 (SNPs 101 to 150), and 50 SNPs with MAF=0.4 (SNPs 151 to 200). For each simulation setting, 100 data sets are generated and subsequently analyzed. Y is generated from the additive model $\log\left(\frac{p(Y=1)}{p(Y=0)}\right) = \beta_0 + \sum_{j=1}^{200} \beta_j . SNP_j$, where SNPs are coded as 0,1,2 (2 represents the minor homozygous genotype) and $\beta = (\beta_1,...,\beta_{200})$ stand for the regression coefficients. In the null case scenario, we examine non-informative predictors, i.e. $\beta_1 = ... = \beta_{200} = 0$, and β_0 is also set to 0. In the alternative scenario, the coefficients $\beta_1, \beta_{51}, \beta_{101}, \beta_{151}$ (corresponding to four SNPs with MAF = 0.05, 0.1, 0.25, 0.4 respectively) are fixed to log(3), yielding a large genotype odds ratio (OR) of 3, and the coefficients $\beta_2, \beta_{52}, \beta_{102}, \beta_{152}$ (again corresponding to four SNPs with MAF = 0.05, 0.1, 0.25, 0.4 respectively) are set to log(1.5), corresponding to a moderate OR of 1.5. In this setting β_0 is fixed to -3 so that the two groups Y=0 and Y=1 are of approximately equal size. The remaining coefficients $\beta_3,...,\beta_{50}, \beta_{53},...,\beta_{100}, \beta_{103},...,\beta_{150}, \beta_{153},...,\beta_{200}$ are equal to zero. The considered total sample sizes are n=200 (small study), n=500, n=1000 (studies of moderate size), and n=10000 (large study).

Results

Bias in favour of large MAFs in the case of non-informative SNPs

In the null case scenario, i.e. under the null-hypothesis that none of the SNPs in a simulated data set is informative ($\beta_1 = ... = \beta_{200} = 0$) the VIMs should be equally low for all SNPs. Any pattern that deviates from this indicates a systematic bias.

In this null case setting, Part (A) of Table 1 shows the median (with 1st and 3rd quartiles in parentheses) of the VIM of SNPs with MAF=0.05, 0.1, 0.25, 0.4 for the Gini VIM in randomForest (left), the permutation VIM in randomForest (middle), and the permutation VIM in cforest (right) for the sample size n=500. In each setting, the results are aggregated from 100 simulated data sets. It is clear from Table 1 that the Gini VIM is strongly biased in favor of SNPs with large MAF, although all SNPs are non-informative. In contrast, the permutation VIM is unbiased in the case of non-informative predictors (the interquartile ranges displayed in Table 1 cover the value zero) and the obtained VIM pattern does not depend substantially on the random forest type (randomForest or cforest). An interesting feature is that the permutation VIM has no bias but a higher variance for common genetic variants (large MAFs). This will be discussed later. As shown in Figure 1 representing boxplots of the Gini VIM with other sample sizes (n=200, n=1000 and n=10000), the bias does not disappear for large studies (n=10000). Moreover, the variability of the VIM decreases with an increasing sample size so that the differences in VIM between the MAFs appear noticeably more pronounced in the rightmost plot for n=10000 in Figure 1.

These results clearly demonstrate that the Gini VIM is biased towards SNPs with large MAFs. This can have a non-negligible impact on the results of genetic association studies. The bias is substantial and cannot be explained by the previously observed bias (Strobl et al., 2007) in favor of variables with many categories since, at least for n=10000, all SNPs have three categories. So what is the mechanism behind this bias? We will try to answer this question in the next sections.

Is the Gini criterion a biased criterion?

In the intent to identify the source of the bias outlined in the above section, the perhaps most natural idea is that the applied splitting criteria employed by the random forests algorithms might be biased even if all SNPs have the same number of categories but different MAFs. From a theoretical perspective, we know from the literature (Grabmeier and Lambe, 2007) that, in the case of a binary response Y, the Gini criterion yields the same trees as a standard chi-square criterion. This equivalence can be checked by straightforward calculations (see additional file 1). This is an important result with respect to the bias investigated here. Indeed, the chi-square statistic asymptotically follows a chi-square distribution under the null hypothesis of no association between Y and the predictor - independently of the category frequencies. Thus, the Gini criterion is not expected to favor predictors with balanced categories in asymptotical settings. Asymptotic results may, however, not be valid under the investigated scenario, especially for small MAFs and in the deep layers of trees where nodes are typically very small. Thus, we will now investigate the effects in the first splits separately from those in the lower layers of the trees.

What happens in the first split?

Let us first consider the simple test situation that occurs in the first split of the trees by replacing standard trees by one-layer trees (stumps) in our random forests. Part (B) of Table 1 has the same structure as Part (A), the only difference being that standard trees are replaced by stumps. From Table 1 it becomes clear that the absolute size of the VIM and the degree of the bias of the Gini VIM is diminished for the stumps, but the pattern in favor of large MAF is still present for n=500 (the interquartile ranges do not cover zero). Similarly, Figure 2 represents the same boxplots as Figure 1 based on samples of sizes n=200 (left), n=1000 (middle) and n=10000 (right), the only difference being that standard trees are replaced by stumps. It can be seen from Figure 2 that the Gini VIM in stumps is still noticeably biased for n=200 and n=1000. However, in contrast to the large trees considered previously, the Gini VIM is almost unbiased for n=10000 in stumps.

In order to be able to distinguish between potential sources of bias attributed to the VIM and those attributed to the splitting criterion employed in the tree construction process, we also check whether SNPs with large MAF had a greater chance to get selected in the first split of the trees, indicating a selection effect rather than (or complementing) a VIM effect. The frequency of selection over the 500 trees of SNPs with MAF=0.05,0.1,0.25 and 0.4, respectively, are represented as boxplots in Additional Figure 1 (included in the additional file) for n=500 and n=10000, where each box represents the frequencies of selection obtained for 100 simulated data sets. Variable selection is strongly biased in the first split for n=500 in randomForest. In contrast, variable selection in cforest, which is based on p-values of conditional inference tests, is only slightly biased. Thus, conditional inference tests used in cforest seem to automatically correct for the bias in favour of large MAFs, at least partially. Note that the slight remaining bias can be removed by using permutation p-values of conditional inference tests in place of the default criterion based on asymptotic p-values. The

permutation-based procedure, however, is extremely time consuming and cannot be applied to such large sets. For both randomForest and cforest, the bias is neglegible for n=10000.
Up to here our results show that variable selection in the first split is biased in favour of large MAFs when n=500 for randomForest, but almost unbiased when n=10000. This is in agreement with the fact that the Gini VIM calculated from stumps, which is derived directly from the Gini criterion in the first split, is slightly biased when n=500, but almost unbiased when n=10000. We conclude that n=500 is too small for asymptotic results to hold. Otherwise, the bias in variable selection and Gini VIM in stumps would be similar for n=500 and n=10000. For MAF=0.05, the minor homozygous genotype has probability 0.0025, thus yielding an expected frequency of 1.25 in studies of size n=500. It is not surprising that asymptotic properties do not hold in this context. The bias found in the selection frequencies for randomForest employing the Gini split selection criterion are in accordance with the results from the Gini VIM, whereas in the randomForest and to a less extent in the cforest permutation VIM the selection bias yields an increased variance for large MAF (similar effects were found by Strobl et al., 2007 in the case of variables with different numbers of categories).

What happens at the bottom of the tree?

In order to better understand the mechanisms of the bias in subsequent splittings, we further look at the frequency of selection of SNPs with different MAF in the splits of each individual randomForest tree. Figure 3 shows the relative frequency of selection of the SNPs with MAF=0.05 (black), MAF=0.1 (red), MAF=0.25 (green), and MAF=0.4 (blue) against the index of the layer (1 standing for the root node, 2 for its two child nodes, etc) for a simulated data set with n=10000 and non-informative SNPs with randomForest. The frequency of selection for a given MAF is computed as the number of selected SNPs with this MAF in the considered layer divided by the total number of selected SNPs in this layer. We display the results only up to layer 35 because after this there were so few trees left that the results depict merely random fluctuation. Roughly, three distinct regions can be observed in Figure 3. Near the root node (approximatively up to layer 3 at the left side of Figure 3; this area is termed region 1 in the following), the frequency of selection does not seem to depend on the MAF. All four MAFs have frequencies of selection of about 25%. This is in agreement with the fact that variable selection is unbiased in the first split for n=10000, as displayed in Additional Figure 1. For intermediate layers (approximatively between layers 4 and 25 in the middle of Figure 3; this area is termed region 2 in the following), the curves of the four MAFs are approximately parallel, and the frequency of selection substantially increases with the MAF. The difference between MAFs tends to slightly increase with the layer index. For deep layers (approximatively from layer 25 at the right side of Figure 3; this area is termed region 3 in the following), the bias increases noticeably. In the deepest layers, SNPs with MAF=0.05 or 0.1 are almost never selected. In the rest of this section, we suggest explanations for this particular pattern with three distinct regions.

A straightforward explanation for the difference between regions 1 and 2 is that the parent nodes to be split get smaller and smaller as partitioning goes on. Hence, asymptotic unbiasedness of the split selection criteria does not hold anymore for deep layers, even if the sample size available at the root node was large (n=10000). This indicates that asymptotics approximately hold in region 1 but not in region 2.

A potential explanation for the sudden decrease of the frequency of selection of small MAFs in region 3 is that, as splitting goes on, more and more SNPs are not 3-categorical anymore. They may become 2-categorical and ultimately 1-categorical. A 2-categorical predictor has lower chance to be selected than a 3-categorical predictor (Strobl et al 2007) or, perhaps more importantly, a 1-categorical predictor has no chance at all to be selected. Since carriers of rare variants are rare for SNPs with small MAFs, these SNPs become 2-categorical and 1-categorical earlier during the construction of the tree than SNPs with larger MAF, as depicted in Additional Figure 2 that represents the frequency at which variables with MAF=0.05, 0.1, 0.25 and 0.4 are 1-,2-, and 3-categorical against the index of the layer (n=10000, non-informative SNPs). It shows that, indeed, SNPs with small MAF loose categories earlier in the tree building process and will thus be affected by classical variable selection bias as described by (Strobl et al 2007) or, perhaps even more importantly, not be selected at all if they have only one category. This extra source of bias in deeper layers adds to the bias we have previously detected in the stumps, thus certainly explaining the acceleration of the decrease of the frequency of selection for small MAFs in region 3 at the right of Figure 3.

Consequences on Gini VIM and permutation VIM

Overall, SNPs with small MAF have a much lower chance to be selected, independently of their prediction relevance (up to now all SNPs were non-informative). Since the Gini VIM directly depends on the selection criterion, Gini VIM is strongly biased. Our results also show that this effect is aggravated with the depth of the trees. In contrast, permutation VIM is essentially unbiased. The reason for this is most likely that the permutation VIM is based on the decrease of accuracy resulting from permutation for out-of-bag observations, i.e. for independent data that was not used to construct the tree. Therefore, with the permutation importance an increased selection probability is not sufficient for producing a higher VIM. The higher frequency of selection of SNPs with large MAF, however, results in a higher variance of the permutation VIM. The reason for this is that SNPs that are selected in a tree lead to a non-zero decrease of accuracy for this tree. Thus, SNPs that are often selected have non-zero VIM for many trees. Moreover, SNPs that are selected earlier in the trees affect more observations. Both results in a more variable total VIM.

The case of informative SNPs

In the case of informative SNPs (Part (C) of Table 1, for n=500), SNPs with larger MAFs have in average substantially larger importance both with Gini VIM and permutation VIM, independently of the random forest type (randomForest or cforest). All patterns look similar as far as informative SNPs are concerned. Similar pictures may also be obtained with different sample sizes (data not shown).

Does that mean that we should speak of a bias in the case of informative SNPs, too, as argued by Tang et al (2009)? There are two contradicting answers to this question. On one hand, we have seen in the case of non-informative predictors that variable selection is biased in favor of SNPs with large MAFs, especially in the deep layers of the trees. This bias in variable selection produces what can be considered as a bias in Gini VIM also in the case of informative predictors. Moreover, we find that for informative predictors SNPs with large MAFs are preferred over SNPs with smaller MAFs even if they have the same odds ratio. This effect is investigated below – and could also be considered as a bias.

On the other hand, in the case of informative predictors a bias is hard to define. In the null case it was clear that all uninformative SNPs should receive the same VIM – and any deviation from this pattern could clearly be considered as a bias. With informative predictors, on the other hand, it is not clear how the VIM should behave. This is mostly because we are lacking a common definition of what exactly a VIM is supposed to measure, especially in cases where variables are no longer independent, like in Strobl et al. (2008), or here, where one might argue that SNPs with smaller MAFs can carry less information than those with larger MAFs.

Even if we strictly speaking cannot consider the favoring of large MAFs as a “bias” because we are no longer in the null case scenario where the term “bias” is clearly defined, it is important to understand why this strong effect is observed. If a variable with small MAF is permuted, the number of observations that have a different value before and after permutation is limited per se because most observations are – and remain – in the biggest category. Thus, only few observations are susceptible to affect accuracy. The effects (in terms of regression coefficients in a generalized linear model) being equal, permutation of a SNP with large MAF thus leads to a larger average decrease of accuracy than permutation of a SNP with small MAF. This result is not only observed with trees as base learners, but also with simple logistic regression models. To illustrate this, we replace the single trees of the random forests by simple logistic regression models, and compute the permutation VIM exactly in the same way. The result is a pattern similar to the permutation VIM in randomForest (data not shown). Thus, the construction principle of the permutation VIM favors SNPs with large MAF. Whether this effect should be considered as a bias or not depends on the point of view – yet it is a characteristic many users of random forests may not be aware of.

Effect of the bias on SNP ranking

As outlined above, the notion of bias is not well-defined in the case of informative predictors, because there is no natural and universal ordering of the predictors. However, any sensible importance measure is expected to favor informative predictors (SNPs 1,2,51,52,101,102,151,152 with a beta coefficient greater than zero in our simulation design) over non-informative predictors (SNPs 3,…,50,53,…,100,103,…,150,153,…,200 with a beta coefficient of zero in our simulation design). Clearly, the Gini VIM does not fulfil this requirement, since it gives higher importance to non-informative SNPs with MAF=0.4 than to informative SNPs with small MAF=0.05 or 0.1 but OR=1.5 (where a multivariate OR of 1.5 is already quite high for a typical genetic association study) or even OR=3. Figure 4 illustrates the ability of the three different variable importances to detect informative SNPs (i.e. SNPs with OR≠0) within the 200 candidate SNPs using ROC methodology for sample size n=500 (top) and sample size n=10000 (bottom). The plotted curves aggregate the results obtained from the 100 simulated data sets. While in the first column all candidate SNPs are considered, the plots in the second column focus on SNPs with very large MAF =0.4 and very low MAF=0.05 only. From these ROC curves, it can be clearly seen that the permutation VIMs have noticeably better power to detect informative SNPs than the Gini VIM. This is especially striking in very large samples (n=10000, bottom-right part of the figure), where the Gini VIM ranks all SNPs with MAF=0.4 better than all SNPs with MAF=0.05 irrespectively of their OR, hence the rectangular form of the ROC curve. In this case, the two informative SNPs with OR=1.5 (one with MAF=0.4, one with MAF=0.05) are never correctly identified as top-ranking by the Gini VIM, whereas the permutation VIM from randomForest identifies them correctly in most of the 100 simulated data sets, yielding an area under curve near 1. These results clearly show that the Gini VIM is likely to rank many informative SNPs worse than many non-informative SNPs.

Discussion and concluding remarks

In the case of non-informative SNPs, the widely used Gini VIM implemented in the standard randomForest method is biased in favour of SNPs with large MAF. The bias is substantial and can have important consequences in practical studies. Some of the numerous non-informative SNPs with large MAF might mask the effect of interesting SNPs with small MAF. This is a strong argument in favour of the permutation VIM, since in large-scale genetic association studies most of the SNPs are not related to the outcome, i.e. non-informative. The bias in the Gini VIM does not vanish with increasing sample size. That is because the bias originates mainly from the bottom of the trees, where splitting nodes are always small with standard settings, independently of the starting sample size in the root node.

The Gini VIM is computed directly from the splitting criterion itself as the decrease in Gini impurity. We identified two sources of bias. First, the Gini criterion, though asymptotically unbiased, is biased in favour of large MAFs in samples where the least frequent category is very small. This bias, that affects both variable selection and VIM, decreases as sample size increases. In our simulations, the bias was strong in the root node for a sample size of n=500, but almost disappeared for sample size n=10000. The second source of bias is associated to the tree structure and affects the nodes at the bottom of the tree. As splitting goes on, nodes become smaller and smaller. The bias, that is moderate at the top of the tree, becomes dramatic at the bottom of the tree. Note that, as a consequence of splitting, SNPs become 2- or 1-categorical at the bottom of the trees. Since SNPs with small MAF become more rapidly 1- or 2-categorical, they are more affected by this problem.

Even though variable selection is biased in favour of large MAFs in small sample settings, permutation VIMs are unbiased in the case of non-informative predictors since they are based on the accuracy on out-of-sample, i.e. independent data. A non-informative SNP with large MAF might be selected more often than an informative SNP with small MAF. Nevertheless, it will have a permutation VIM of zero in average anyway, since it has no chance to achieve good prediction accuracy for the out-of-sample data.

Thus, our recommendation is to use the permutation VIM and *not* the Gini VIM. To address the stability issue pointed out by Calle and Urrea (2010), it may be worth doing several permutations of the variables instead of only one (default value of the parameter nperm in randomForest and in the varimp function for cforest).

Another sensible recommendation is to limit the depth of the trees. In our analyses, we have shown that stumps are much less affected by the bias than trees with many layers, irrespectively of the chosen algorithm and VIM type. This result is corroborated by the number of SNPs with 2 or 1 categories at the bottom of the trees, that obviously increases much faster for small than for large MAFs. In most applications, it is certainly not a good idea to build stumps instead of large trees, because stumps possibly do not appropriately account for the complexity in the data. It is impossible to make general recommendations here, since the optimal depth of the trees may depend on many parameters including the sample size, the number of SNPs, the proportion of informative SNPs, the supposed presence of interactions and of course the MAF of the considered SNPs. However, we claim that, in general, a compromise between unpruned trees of maximal depth (the default of the randomForest function) and stumps might be preferred with respect to the bias discussed in this paper.

Another related factor that potentially influences the bias in lower layers of the trees is the minimal size of terminal nodes (parameter nodesize in randomForest, minbucket in cforest). Smaller values of the minimal node size tend to produce a higher bias. That is because rare variants might get selected based on the splitting criterion but eventually rejected just because the rare categories (say AC and CC) would form a too small terminal node. This source of bias is expected to affect equally both randomForest and cforest. Strictly speaking, depth and minimal node size should be considered as parameters that should be tuned, e.g., by means of cross- validation, as also indicated by the results of Lin and Jeon (2006).

Our results for informative SNPs also support that even though the permutation importance also favors SNPs with larger MAF, the variable selection properties we have illustrated by means of ROC curves in Figure 4 strongly support the superiority of the permutation VIM over the Gini VIM.

In our analyses representing a particular setting with SNP data, the permutation VIM of randomForest is found to have a similar unbiased behaviour as the permutation VIM of cforest. However, the variance of the permutation VIM increases with the MAF more strongly in randomForest than in cforest. This difference between randomForest and cforest is in agreement with the higher variable selection bias of randomForest depicted in additional Figure 1. It may explain the slightly better performance of cforest in terms of ROC curve in Figure 4. For these reasons, cforest should be preferred to randomForest for the analysis of SNPs, too. Besides the party package, the cforest methodology is now also implemented in the most recent version of the random jungle software (Schwartz et al, 2010) which is particularly designed to handle genome-wide data efficiently.

The effects illustrated by Nicodemus (2011) and in this article – that SNPs with large MAF are systematically preferred over those with small MAF, to some extent even if they are less informative – will to some readers appear as a serious problem, to others only as a natural property of a VIM sensitive to group size. However, especially in the context of large genetic association studies, the interpretation of VIMs may be lead by the expectation that those SNPs ranked highly are actually those with the strongest association to the response – not those whose category frequencies provide the highest value of a VIM with potentially unexpected statistical properties. With this expectation in mind, the permutation VIM has clearly shown its superiority to the Gini VIM.

Acknowledgements: ALB is supported by the LMU-innovativ Project BioMed-S: “Analysis and Modelling of Complex Systems in Biology and Medicine”. AB is supported by grant BO3139/2-1 from the German Research Foundation (Deutsche Forschungsgemeinschaft) headed by ALB: “Resampling methods for the statistical analysis of high-dimensional biological data”. CS is supported by grant STR1142/1-1 (“Methods to Account for Subject-Covariates in IRT-Models”) from the German Research Foundation (Deutsche Forschungsgemeinschaft).

References

Bureau A, Dupuy J, Falls K *et al*: Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology* 2005; 28:171-182.

Breiman L: Random forests. *Machine Learning* 2001; 45:5-32.

Breiman, L., Cutler, A., Liaw, A., & Wiener, M., 2010. randomForest: Breiman and Cutler’s random forests for classification and regression (R package version 4.6-2) URL: http://cran.r-project.org/package=randomForest

Briggs F, Ramsay PP, Madden E et al: Supervised machine learning and logistic regression identifies novel epistatic risk factors with PTPN22 for rheumatoid arthritis. *Genes and Immunity* 2010; 11:199-208.

Cleynen I, Mahachie John JM, Henckaerts L *et al*: Molecular reclassification of Crohn's disease by cluster analysis of genetic variants. *PLoS One* 2010; 5:e12952.

De Lobel L, Geurts P, Baele G *et al*: A screening methodology based on Random Forests to improve the detection of gene–gene interactions. *European Journal of Human Genetics* 2010; 18, 1127–1132.

Heidema AG, Boer JMA, Nagelkerke N *et al:* The challenge for genetic epidemiologists: How to analyze large numbers of SNPs in relation to complex diseases. *BMC Genetics* 2006; 7:23.

Hothorn, T., Hornik, K., & Zeileis, A. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 2006; 15:651– 674.

Hothorn, T., Hornik, K., Strobl, C., & Zeileis, A. (2010). party: A laboratory for recursive part(y)itioning (R package version 0.9-99991) URL: http://cran.r-project.org/package=party

Lin, Y., & Jeon, Y. (2006). Random forests and adaptive nearest neighbors. Journal of the American Statistical Association, 101, 578 –590.

Liu C, Ackermann HH, Carulli JP: A genome-wide screen of gene-gene interactions for rheumatoid arthritis susceptibility. *Human Genetics* 2011 (Epub ahead of print).

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R News, 2(3), 18 –22.

Lunetta KL, Hayward LB,. Segal J *et al:* Screening large-scale association study data: exploiting interactions using random forests. *BMC Genetics* 2004; 5:32.

Nicodemus K: On the stability and ranking of random forest variable importance measures. *Briefings in Bioinformatics* 2011 (in press).

R Development Core Team, 2010. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Roshan U, Chikkagoudar S, Wei Z *et al*: Ranking causal variants and associated regions in genome wide association studies by the support vector machine and random forest. *Nucleid Acids Research* 2011 (Epub ahead of print).

Schwartz DF, Szymczak S, Ziegler A, König IR: Picking single-nucleotide polymorphisms in forests. BMC Proceedings 2007; 1:S59.

Schwartz DF, König I, Ziegler A: On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. *Bioinformatics* 2010; 26:1752-1758.

Strobl C, Boulesteix AL, Zeileis A *et al*: Bias in random forest variable importance measures: Illustrations, sources, and a solution. *BMC Bioinformatics* 2007; 8:25.

Strobl C, Boulesteix AL, Kneib T *et al*: Conditional variable importance for random forests. *BMC Bioinformatics* 2008; 9:307.

Tang R, Sinnwell JP, Li J, Rider DN, de Andrade M, Biernacka JM: Identification of genes and haplotypes that predict rheumatoid arthritis using random forests. *BMC Proceedings* 2009; 3:S68.

Yang W, Gu C : Selection of important variables by statistical learning in genome-wide association analysis. *BMC Proceedings* 2009; 3:S70.

Weidinger S, Baurecht H, Wagenpfeil S *et al*: Analysis of the individual and aggregate genetic contributions of previously identified SPINK5 , KLK7 and FLG polymorphisms to eczema risk. *Journal of Allergy and Clinical Immunology* 2008; 122:560-568.

		Gini randomForest	Perm randomForest x 10^5	Perm cforest x 10^5
(A) Large trees – non-informative SNPs				
MAF=0.05		0.58 [0.52 – 0.67]	-1.9 [-13 – 9.4]	-4.3 [-13 – 5.4]
MAF=0.1		0.87 [0.78 – 0.97]	-1.8 [-17 – 14]	-4.3 [-17 – 9.8]
MAF=0.25		1.5 [1.4 – 1.7]	-3.4 [-26 – 21]	-4.3 [-22 – 16]
MAF=0.4		1.9 [1.7 – 2.0]	-3.2 [-30 – 26]	-6.5 [-25 – 15]
(B) Stumps – non-informative SNPs				
MAF=0.05		0.01 [0 – 0.02]	0 [-2.1 – 0]	-1.1 [-7.6 – 1.1]
MAF=0.1		0.01 [0 – 0.03]	0 [-4.3 – 0]	-2.2 [-9.8 – 2.2]
MAF=0.25		0.02 [0.01 – 0.04]	0 [-9 – 1.2]	-3.3 [-13 – 3.3]
MAF=0.4		0.02 [0.01 – 0.05]	0 [-12 – 2.1]	-4.3 [-15 – 3.3]
(C) Large trees – informative SNPs				
MAF=0.05	OR=3	1.5 [1.1 – 2.0]	88 [40 – 188]	128 [56 – 258]
MAF=0.1	OR=3	3.0 [2.1 – 3.9]	246 [116 – 433]	414 [210 – 578]
MAF=0.25	OR=3	5.9 [5.1 – 6.8]	670 [509 – 819]	926 [689 – 1217]
MAF=0.4	OR=3	7.7 [6.5 – 9.2]	909 [708 - 1200]	1303 [1011 – 1627]
MAF=0.05	OR=1.5	0.63 [0.49 – 0.82]	2.2 [-9.5 – 16]	0.54 [-7.9 – 18]
MAF=0.1	OR=1.5	0.95 [0.77 – 1.3]	9.3 [-9.9 – 40]	4.9 [-7.6 – 47]
MAF=0.25	OR=1.5	1.8 [1.6 – 2.2]	36 [4.2 – 87]	42 [0.5 – 84]
MAF=0.4	OR=1.5	2.2 [1.9 – 2.7]	47 [0.27 – 112]	42 [0.8 – 121]
MAF=0.05	OR=0	0.51 [0.45 – 0.61]	-4.7 [-11 – 5.6]	-3.3 [-9.8 – 2.2]
MAF=0.1	OR=0	0.76 [0.67 – 0.88]	0.71 [-16 – 12]	0 [-9.7 – 10]
MAF=0.25	OR=0	1.3 [1.2 – 1.4]	-3.0 [-26 – 16]	-5 [-20 – 13]
MAF=0.4	OR=0	1.6 [1.5 – 1.7]	-2.0 [-33 – 19]	-10 [-21 – 10]

Table 1. (A): Median variable importance [1st quartile – 3rd quartile] in the null scenario with n=500 and large trees. (B) Median variable importance [1st quartile – 3rd quartile] in the null scenario with two-node trees (stumps). (C) Median variable importance [1st quartile – 3rd quartile] in the informative scenario with large trees built with the standard settings of the functions randomForest and cforest. Note that the permuation importances (3rd and 4th columns) are multiplied by 10^5 for the sake of clarity.

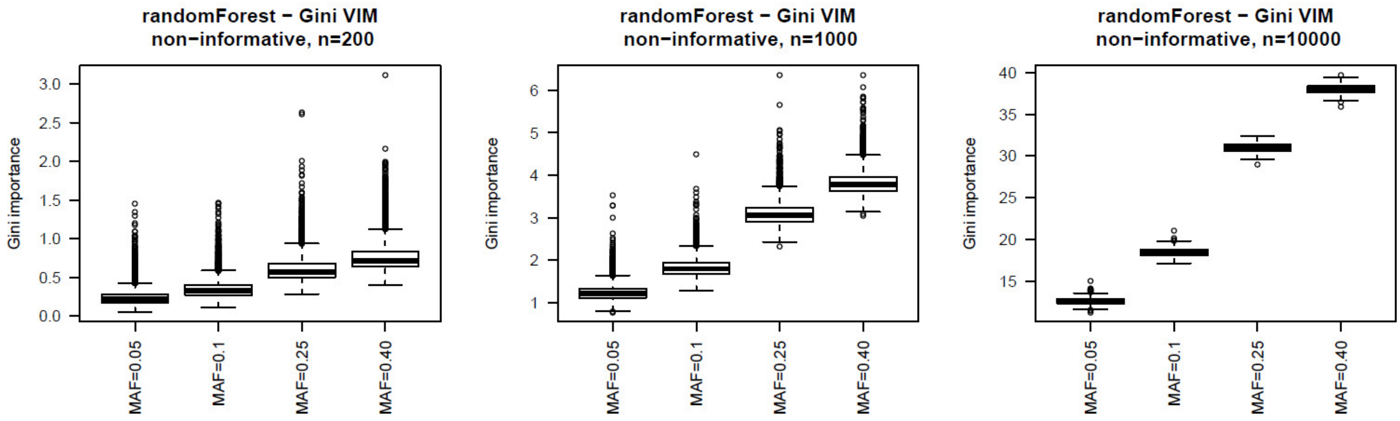

Figure 1. Boxplot of VIMs in the null scenario for different sample sizes.

Variable importance of SNPs with MAF=0.05,0.1,0.25,0.4 in the null scenario (non-informative SNPs). **Left:** Gini VIM based on randomForest. **Middle:** Permutation VIM based on randomForest. **Right:** Permutation VIM based on cforest. Each box corresponds to 100 (data sets) x 50 (SNPs) = 5000 values.

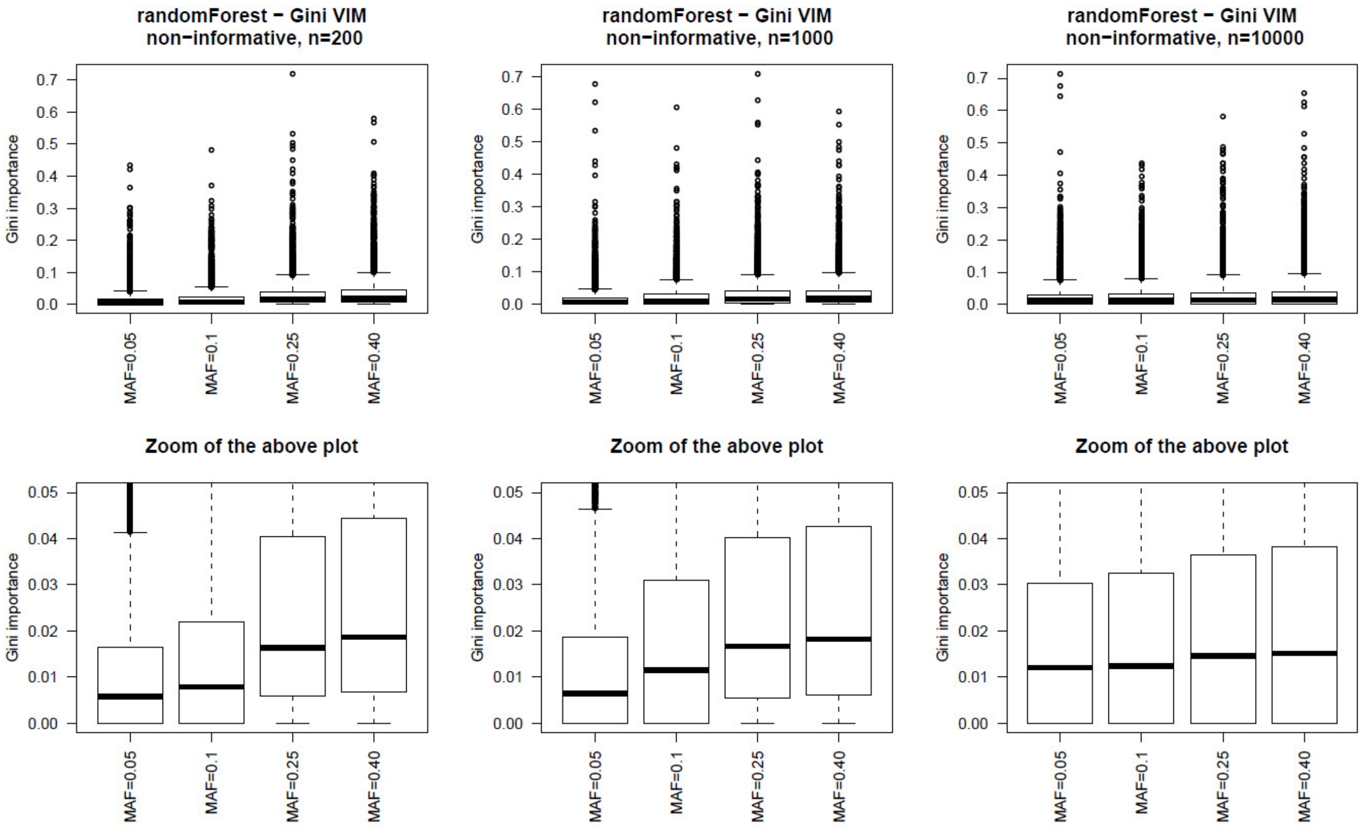

Figure 2. Boxplot of VIMs based on stumps in the null scenario for different sample sizes. Variable importance of SNPs with MAF=0.05,0.1,0.25,0.4 based on trees with two-nodes ("stumps") in the null scenario (non-informative SNPs). **Left:** Gini VIM based on randomForest. **Middle:** Permutation VIM based on randomForest. **Right:** Permutation VIM based on cforest. Each box corresponds to 100 (data sets) x 50 (SNPs) = 5000 values. Zooms of the boxplots are displayed in the bottom row.

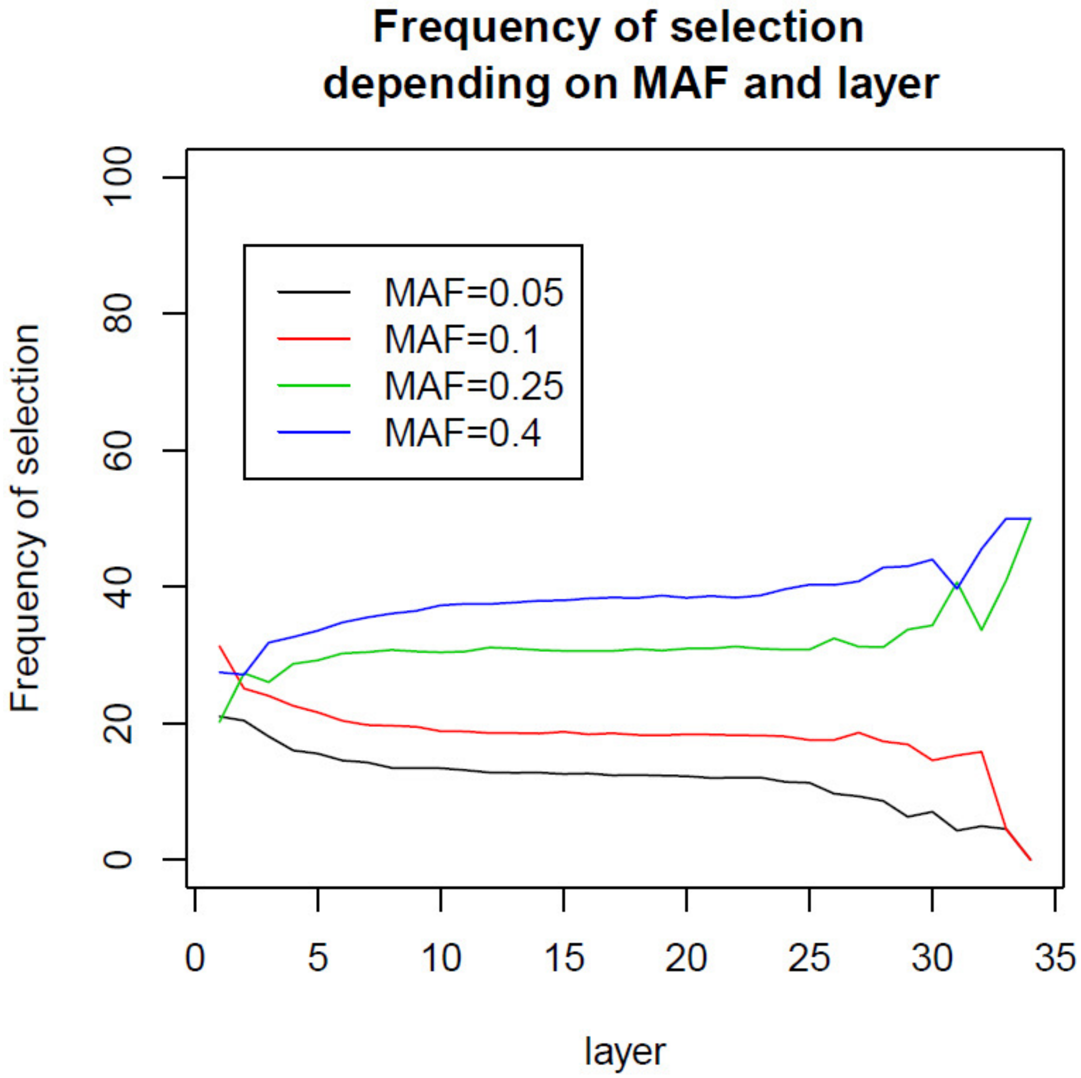

Figure 3. Frequency of selection by MAF in the different layers of the tree.

Frequency of selection (with randomForest) of SNPs with MAF=0.05 (black), 0.1 (red), 0.25 (green), 0.4 (blue) against the index of the layer for a simulated data set (n=10000, non-informative SNPs).

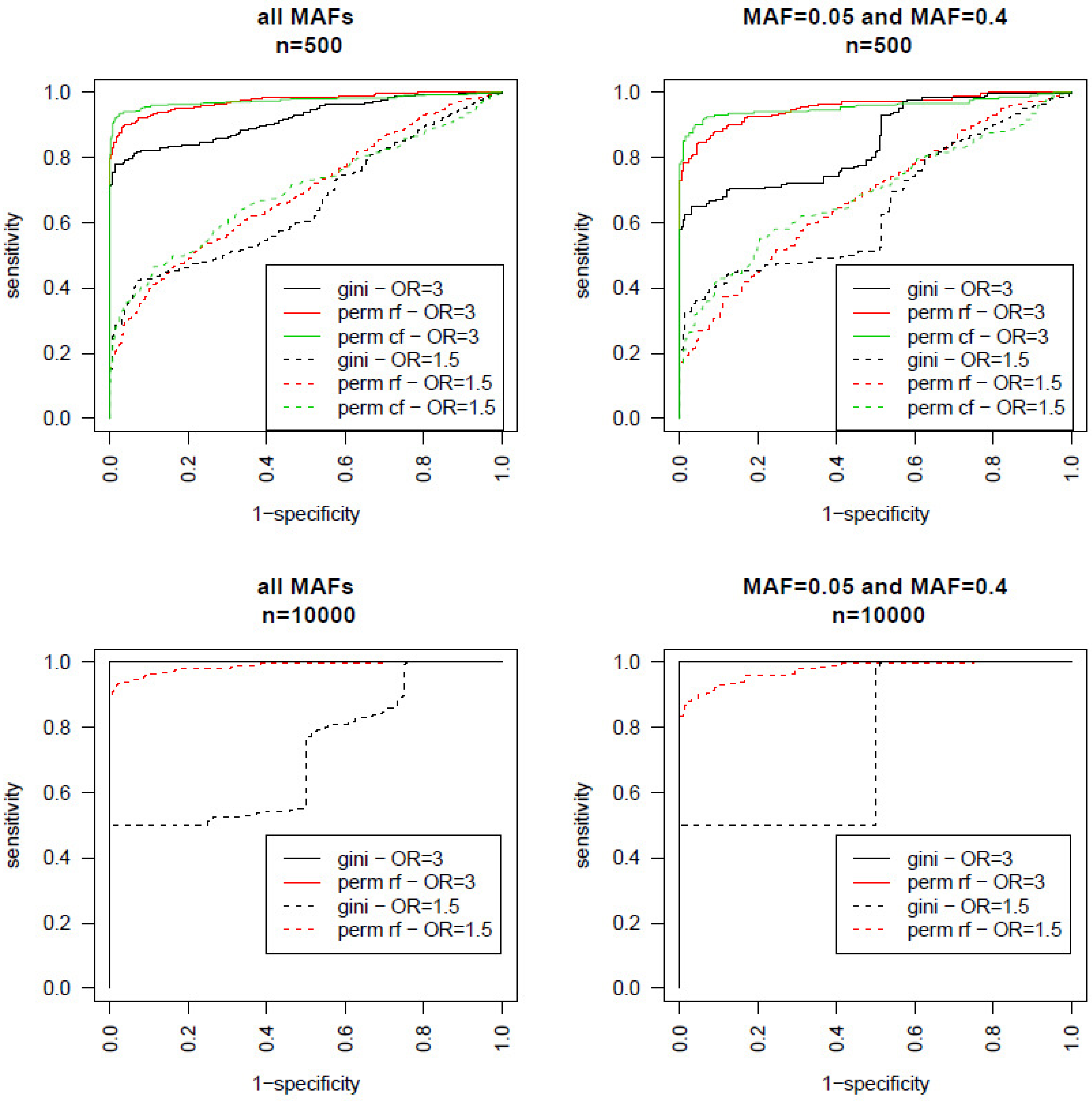

Figure 4. ROC curves in the informative scenario.

The x-axis "sensitivity" is the proportion of informative SNPs (with OR≠0) that are detected i.e. have a VIM above the considered threshold. The y-axis "1-specificity" is the proportion of non-informatives SNP (with OR=0) that are detected. **Top:** n=500. **Bottom:** n=10000. **Left:** all 200 SNPs. **Right:** only SNPs with very large (0.4) or very low (0.05) MAF. For n=10000 cforest was not used for computational reasons.