Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations

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

Submitted: 17th May 2011; Received (in revised form): 2nd August 2011

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 ability to predict 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 favoured 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 article is 3-fold: (i) 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); (ii) 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 (iii) to summarize these results and previously investigated properties of random forest VIMs in the context of genetic association studies and to make practical recommendations regarding the choice of the random forest and variable importance type. All our analyses can be reproduced using R code available from the companion website: http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/020_professuren/boulesteix/ginibias/.

Keywords: random forest; genetic association study; variable importance; variable selection bias; CART; cforest

INTRODUCTION

Random forests, originally suggested by Breiman 10 years ago [1], 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 single nucleotide polymorphisms (SNPs) involved in gene–gene interactions of second order or higher, and to investigate the predictive power of genetic data taking into account possible complex nonlinear patterns [2–9]. Methodological developments of random forests and new implementations with

Corresponding author. Anne-Laure Boulesteix, 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; E-mail: boulesteix@ibe.med.uni-muenchen.de

Anne-Laure Boulesteix is an assistant professor of Computational Molecular Medicine at LMU University of Munich, Germany. She works on statistical methods in bioinformatics. Her research interests include validation in bioinformatics and prediction models with high-dimensional biological data with emphasis on random forests and dimension reduction.

Andreas Bender is a master student of biostatistics at the LMU University of Munich, Germany. He obtained his bachelor’s degree in statistics in 2010. His bachelor thesis was on random forests in genetic association studies.

Justo Lorenzo Bermejo leads the group ‘Statistical Genetics’ at the Institute of Medical Biometry and Informatics, University Hospital Heidelberg, Germany. His research focuses on the identification of genetic risk factors of disease including genetic variants, gene methylation and expression and antigen profiling. In addition to the development of tailored statistical approaches for genetic data, his team supports applied research projects on cancer and heart diseases.

Carolin Strobl is an assistant professor at the Department of Statistics, LMU University of Munich, Germany. Her research interests include the development of unbiased variable selection techniques and variable importance measures as well as applications of machine learning methods in the social- and life sciences.
focus on genetic applications have also been recently addressed in genetic analysis workshops [10, 11] and published in genetics and bioinformatics journals [12–14].

It is now widely known that the commonly used random forest VIMs are biased in favour of the categorical variables with more categories [15]. 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. [16] have suggested a random forest algorithm based on unbiased variable selection criteria. When these criteria—together with subsampling instead of bootstrap sampling—are used for computing a permutation VIM, this measure is unbiased as discussed in detail by Strobl et al. [15].

In the context of genetic association studies, it has been argued that this kind of bias is irrelevant, since SNPs have three categories (homozygous with the common allele, heterozygous, homozygous with the rare allele). Based on resampling analyses, Calle and Urrea [17] 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 [18] suggests that the higher stability of Gini VIM compared to permutation VIM is attributable to a bias in favour of SNPs with large Minor Allele Frequency (MAF). In Nicodemus’ resampling analyses, common variants receive high ranks consistently over the subsamples, while rare variants consistently receive low ranks, which induces the apparent stability.

The aim of the present article is to provide further results and a deeper understanding of the statistical mechanisms responsible for the observations of Calle and Urrea [17] and Nicodemus [18]. Simulations based on randomly generated data and real data from HapMap are conducted in order to quantitatively compare the behaviour 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 behaviour. 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 [1]. 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’ [19, 20] available in the R system for statistical computing [21], which we use in the simulations with all default parameters. In particular, the number of trees is set to $n_{tree} = 500$ and the number of candidate predictors considered at each split is set to the default value $m_{try} = p^{1/2}$ in the main analyses, where $p$ is the number of predictors. In additional analyses, we also consider the extreme values $m_{try} = 1$ and $m_{try} = p$ and the smaller and larger values $n_{tree} = 50$ and $n_{tree} = 5000$. In the rest of this article, 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 [15]. To address this issue, Hothorn et al. [16] developed an alternative class of random forests which are based on conditional hypothesis testing. At each split, each candidate predictor is tested for its association with the response and a P-value is output. This P-value is conditional, which means that it reflects the probability to obtain such a high (or a higher) association with the response given the marginal distributions of the response and of the considered predictor. The algorithm of Hothorn et al. [16] is not affected by variable selection bias because it employs P-values as splitting criteria, which are on a comparable scale even for variables of different types, and because it separates the issue of variable selection from that of cut-point selection. Thus it does not share the above mentioned pitfall. We also consider this type of random
forest in this study taking advantage of the function ‘cforest’ from the R package ‘party’ [22]. The number of trees is set to \( ntree = 500 \) and the number of candidate predictors at each split is set to \( mtry = \sqrt{p} \) again. Moreover, the \( P \)-value threshold acting as a stopping criterion is set to min-criterion = 0. All other parameters are set to their default values. In the rest of this article, 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.

### VIMs

For the randomForest method, we consider two types of VIMs: 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 reference [15] for details on VIMs. For the cforest method, we consider the permutation VIM implemented in the function ‘varimp’ from the package ‘party’.

### Simulation design: data generation

The simulated data sets include a binary phenotype \( Y \) and 200 genetically unlinked SNPs in Hardy–Weinberg Equilibrium: 50 SNPs with MAF = 0.05 (SNPs 1–50), 50 SNPs with MAF = 0.1 (SNPs 51–100), 50 SNPs with MAF = 0.25 (SNPs 101–150) and 50 SNPs with MAF = 0.4 (SNPs 151–200). For each simulation setting, 100 data sets are generated and subsequently analyzed. \( Y \) is generated from the additive model

\[
\log(p(Y = 1)/p(Y = 0)) = \beta_0 + \sum_{j=1}^{200} \beta_j \cdot \text{SNP}_j,
\]

where SNPs are coded as 0,1,2 (2 represents the minor homozygous genotype) and \( \beta = (\beta_1, \ldots, \beta_{200}) \) stand for the regression coefficients. In the null-case scenario, we examine non-informative predictors, i.e. \( \beta_1 = \ldots = \beta_{200} = 0 \) and \( \beta_0 \) 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, \( \beta_0 \) is fixed to \(-3\) so that the two groups \( Y = 0 \) and \( Y = 1 \) are of approximately equal size. The remaining coefficients \( \beta_1, \ldots, \beta_{50}, \beta_{53}, \ldots, \beta_{100}, \beta_{103}, \ldots, \beta_{150}, \beta_{153}, \ldots, \beta_{200} \) are equal to zero. The considered total sample sizes are \( n = 50, 200 \) (small studies), \( n = 500, 1000 \) (studies of moderate size) and \( n = 10000 \) (large study).

### Design of the analysis based on HapMap data

In order to investigate the bias in a realistic data setting, we considered a subset of HapMap phase III data [23] retrieved using the HapMart data management system (http://hapmap.ncbi.nlm.nih.gov/biomart/martview). The two populations ‘CEU: Utah residents with Northern and Western European ancestry from the CEPH collection’ and ‘YRI: Yoruba in Ibadan, Nigeria (West Africa)’ were chosen from release 27. Genotypes on chromosome 1 from position 0 to position 213 \( \times 10^6 \) were filtered and monomorphic SNPs were excluded. The remaining SNPs had either <10% or >50% missing values. SNPs with >50% missing values were excluded. The final data set included 9671 SNPs and 350 individuals (176 YRI and 174 CEU). Missing genotypes were imputed using the function ‘rfImpute’ from the package ‘randomForest’ with default values. Ethnicity (YRI or CEU) were considered as response class \( Y \). To assess the bias of the VIMs, 10 non-informative data sets were generated by randomly permuting \( Y \) while preserving the linkage disequilibrium and frequency patterns of the SNPs.

### Quantitative assessment of the bias

#### 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 (\( \beta_1 = \ldots = \beta_{200} = 0 \)) the VIMs should be equally low for all SNPs. Any pattern that deviates from this indicates a systematic bias.

In this null case setting, panel (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 favour 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 = 50, n = 200, n = 1000, n = 10000). Each box corresponds to 100 (data sets) × 50 (SNPs) = 5000 values.
increasing sample size so that the differences in VIM between the MAFs appear noticeably more pronounced in the rightmost plot for \( n = 10,000 \) 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 in favour of variables with many categories [15] since, at least for \( n = 10,000 \), all SNPs have three categories. So what is the mechanism behind this bias? We will try to answer this question in section ‘Sources of the bias in the null-case scenario’ section.

The case of informative SNPs

In the case of informative SNPs [panel (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. [24]? 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 favour of SNPs with large MAFs, especially in the deep layers of the trees. This bias in variable selection may produce what can be considered as a bias in Gini VIM also in the case of informative predictors. Moreover, we find that SNPs with large MAFs are preferred over SNPs with smaller MAFs even if they have the same OR. This effect is investigated below—and could in a broad sense 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 [13], 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 favouring 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 favours 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 favour 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, \ldots, 50, 53, \ldots, 100, 103, \ldots, 150, 153, \ldots, 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 2 illustrates the ability of the three different variable importance measures to detect informative SNPs (i.e. SNPs with OR \( \neq 0 \)) within the 200 candidate SNPs using ROC methodology for sample
size $n = 500$ (top) and sample size $n = 10,000$ (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 $\text{MAF} = 0.4$ and very low $\text{MAF} = 0.05$ 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 ($n = 10,000$, bottom-right part of the figure), where the Gini VIM ranks all SNPs with $\text{MAF} = 0.4$ better than all SNPs with $\text{MAF} = 0.05$ irrespectively of their OR, hence the rectangular form of the ROC curve. In this case, the two informative SNPs with $\text{OR} = 1.5$ (one with $\text{MAF} = 0.4$, one with $\text{MAF} = 0.05$) 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.

The difference between the left panels and the right panels of Figure 2 illustrates that the distribution of MAFs may have an impact on the effect of the bias in terms of ROC. Quite generally, the impact of the VIM bias in terms of SNP ranking increases with the dispersion of the MAFs. If all SNPs have similar MAFs, no effect is observed, whereas the impact of the bias increases when many SNPs have very small or very large MAF—like in our extreme example with $\text{MAF} = 0.05$ and $\text{MAF} = 0.4$. Moreover, for a fixed subset of informative SNPs, the effect of the bias grows with the proportion of non-informative SNPs that have a large MAF. Further, non-informative SNPs with large MAF are also more likely to mask informative SNPs with low association, especially SNPs involved in complex interaction patterns (epistasis) that are essentially more difficult to identify but commonly believed to be well addressed by random forests.
Example based on real data from HapMap

Figure 3 represents the boxplots of the randomForest Gini VIM (left) and randomForest permutation VIM (right) obtained from the non-informative data sets generated by permutation of HapMap data. Each of the five boxes corresponds to a different MAF interval ([0,0.1[, [0.1,0.2[, [0.2,0.3[, [0.3,0.4[, [0.4,0.5], respectively) and aggregates the results obtained from the 10 permuted data sets. The data structure is here more complex than in the simulated data. MAFs are approximately uniformly distributed on [0,0.5]. Many SNPs are highly associated with each other: genetic linkage is even stronger in the HapMap data than in usual genetic association studies. Moreover, many SNPs depart from Hardy–Weinberg Equilibrium. Even in this realistic and complex scenario, Figure 3 clearly shows that the Gini VIM is again systematically biased in favour of large MAFs. In contrast, the permutation VIM is unbiased. SNPs departing from Hardy–Weinberg Equilibrium are also affected by the bias (data not shown).

Sources of the Bias in the Null Scenario

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 forest 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 [24] 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 outlined in Supplementary File S1. 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 favour 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 favour of large MAFs is still present for n = 500. Similarly, Figure 4 represents the same boxplots as Figure 1 based on samples of sizes n = 50, n = 200, n = 1000 and n = 10,000, the only difference being that standard trees are replaced by stumps. It can be seen from Figure 4 that the Gini
VIM in stumps is still noticeably biased for \( n = 50 \), \( n = 200 \) 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 Supplementary Figure S1 (included in Supplementary Data) for \( n = 500 \) and 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.

Since the Gini criterion is equivalent to the chi-square statistic in binary splits, Supplementary Figure S1 also tells us that the commonly used chi-square test is biased in these settings for \( n = 500 \), but not for \( n = 10000 \). This is an expected result considering the expected frequencies for the minor homozygous genotype. Indeed, for MAF \( = 0.05 \), the expected count of minor variant homozygotes is only \( 0.05^2 \times 500 = 1.25 \) for \( n = 500 \) and \( 0.05^2 \times 1000 = 2.5 \) for \( n = 1000 \). Even for MAF \( = 0.1 \) the total expected count of minor variant homozygotes does not exceed 5 for \( n = 500 \) and 10 for \( n = 1000 \). However, the importance of the bias is striking and probably surprising to many users of the chi-square statistic. Note that this bias of the chi-square statistic is not specific to SNP predictors. In additional analyses (data not shown), we also find a bias with binary predictors or with 3-categorical predictors that do not fulfil the Hardy–Weinberg Equilibrium.

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 negligible for \( n = 10\,000 \).

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 \( n = 10\,000 \). 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 biased when \( n = 500 \), but almost unbiased when \( n = 10\,000 \).

**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 5 shows the relative frequency of selection of the SNPs with MAF = 0.05, MAF = 0.1, MAF = 0.25 and MAF = 0.4 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 \( n = 10\,000 \) 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 5. Near the root node (approximately up to layer 3 at the left side of Figure 5; 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 \( n = 10\,000 \), as displayed in Supplementary Figure S1. For intermediate layers (approximately between layers 4 and 25 in the middle of Figure 5; 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 (approximately from layer 25 at the right side of Figure 5; 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 (\( n = 10\,000 \)). 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 [15] 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 Supplementary Figure S2 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 (\( n = 10\,000 \), non-informative SNPs). It shows that, indeed, SNPs with small MAF lose categories earlier in the tree building process and will thus be affected by classical variable selection bias as described in [15] 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 5.

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 (we are considering the null-case scenario only). This 'selection bias', however, does not lead to a VIM bias in the case of the permutation VIM. 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 were not used to construct the tree. Even if non-informative SNPs with high MAF are selected more often due to the selection bias, they have no chance to improve the average out-of-bag accuracy, and thus do not receive higher VIMs. 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 the Gini VIM is different. A SNP that is selected in a tree contributes to the VIM for this tree, while a SNP that is not selected does not contribute. In contrast to the permutation VIM, these contributions do not average to zero because the Gini criterion is by definition always positive. Moreover, even if there were no selection bias (i.e. if all SNPs were selected equally frequently, for instance because mtry = 1), the Gini VIM would be biased since it is directly computed from the Gini criterion itself, which is biased. This aspect is further discussed in ‘Discussion and concluding remarks’ section and illustrated in Supplementary Figure S3. In a few words, the bias of the Gini criterion translates into a VIM bias in two ways: indirectly through the selection bias, and directly because the Gini VIM is computed from the Gini criterion.

DISCUSSION AND CONCLUDING REMARKS
Recommendations
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, especially if these SNPs are involved in complex interaction patterns that are typically much more difficult to identify than main effects. 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. We identified two sources of bias for the Gini VIM. First, the Gini criterion, though asymptotically unbiased, is biased in favour of large MAFs for sample sizes as large as n = 1000. 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 2- or 1-categorical, they are more affected by this problem.

Even though split selection in randomForest 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. Similar patterns were obtained with both simulated data and permuted real data from the HapMap study. Our results for informative SNPs illustrated in Figure 2 also support the superiority of the permutation VIM over the Gini VIM.

Our recommendation is thus to use the permutation VIM and not the Gini VIM. To address the stability issue pointed out by Calle and Urrea [17], 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). More research is needed to address this topic.
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 Supplementary Figure S1. It may explain the slightly better performance of cforest in terms of ROC curve in Figure 2. 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 [25] which is particularly designed to handle genome-wide data efficiently.

Influence of forest parameters
At the light of our analyses of stumps, another sensible recommendation would be 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 article. Another related factor that potentially influences the bias in lower layers of the trees is the minimal size of terminal nodes (parameter nsize in randomForest, minbucket in cforest). Larger 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 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 [26].

A further important parameter is the number mtry of candidate predictors considered at each split. The amplitude of the bias is expected to decrease when mtry decreases. In the extreme case of mtry = 1, there is no variable selection bias, since at each split the selected SNP is simply the only (randomly chosen) candidate SNP. The Gini VIM, however, is still biased—although to a lesser extent. This is because the VIM bias is not only due to the variable selection bias, as outlined in ‘Sources of the bias in the null-case scenario’ section and discussed in reference [15]. Supplementary Figure S3 shows the Gini VIM obtained with n = 500 for different extreme values of mtry: mtry = 1, mtry = \sqrt{200} = 14 and mtry = 200. The observed bias is substantial for all three values of mtry, although the specific bias pattern seems to depend on mtry. In conclusion, we cannot generally recommend a particular mtry value to address the investigated bias. The choice of mtry is a delicate issue and should rather be seen as tuning problem.

The choice of the number of trees does not influence the bias mechanism in each individual tree. It may, however, enhance the ratio between the bias and the within-MAF variance of the VIM. This is because the Gini VIM is built by aggregation over all trees. If the forest includes many trees, the average is computed over many numbers, and thus less variable. In this sense, a large ntree tends to make the bias/variance ratio larger, as illustrated in Supplementary Figure S4 showing the Gini VIM in the null-case scenario for n = 500 with ntree = 50 (left), ntree = 500 (middle) and ntree = 5000 (right). Note, however, that a larger variance is not a desirable feature and that we do not recommend to choose a small ntree. In most applications the prediction accuracy of the forest becomes substantially better with increasing ntree because the stability and smoothness of the prediction increases.

Concluding remarks
The effects illustrated by Nicodemus [18] 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.

SUPPLEMENTARY DATA
Supplementary data are available online at http://bib.oxfordjournals.org/.

Key points
- The Gini VIM is biased in favour of SNPs with large MAF. These SNPs get systematically larger importance values even if all investigated SNPs are non-informative (i.e. not related with the response).
- With the Gini importance informative SNPs with small MAF are often ranked worse than non-informative SNPs with large MAF, yielding a misleading ranking of the SNPs. This effect does not decrease but increases with the sample size.
- The permutation importance measure is unbiased in the case of non-informative SNPs. The regression coefficients being all equal, the permutation importance measure tends to increase with the MAF in the case of informative SNPs. Whether this effect should be denoted as a bias or not, is a matter of interpretation.
- The closest variant of random forests is based on an unbiased splitting criterion, so that split selection with closest does not systematically favour SNPs with large MAF. The usage of closest is recommended to avoid the investigated bias.

FUNDING
LMU-innovativ Project BioMed-S: ‘Analysis and Modelling of Complex Systems in Biology and Medicine’ (to A.L.B.); German Research Foundation (Deutsche Forschungsgemeinschaft): ‘Resampling methods for the statistical analysis of high-dimensional biological data’ (grant 3139/2-1 to A.L.B.); German Research Foundation (Deutsche Forschungsgemeinschaft) ‘Methods to Account for Subject-Covariates in IRT-Models’ (grant STR1142/1-1 to C.S.).

References


