Computational and Structural Biotechnology Journal 21 (2023) 4354–4360

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Method article

Thresholding Gini variable importance with a single-trained random forest: An empirical Bayes approach

Robert Dunne [a,*], Roc Reguant [b], Priya Ramarao-Milne [b], Piotr Szul [c], Letitia M.F. Sng [b], Mischa Lundberg [b,d], Natalie A. Twine [b,e], Denis C. Bauer [b,e,f,**]

[a] Data61, Commonwealth Scientific and Industrial Research Organisation, Sydney, Australia
[b] Transformational Bioinformatics, Commonwealth Scientific and Industrial Research Organisation, Westmead, Australia
[c] Data61, Commonwealth Scientific and Industrial Research Organisation, Dutton Park, Australia
[d] Diamantina Institute, The University of Queensland, St Lucia, Australia
[e] Macquarie University, Applied BioSciences, Faculty of Science and Engineering, Macquarie Park, Australia
[f] Macquarie University, Department of Biomedical Sciences, Faculty of Medicine and Health Science, Macquarie Park, Australia

ARTICLE INFO

Keywords:
Random forest
Feature selection
Empirical Bayes
Genetic analysis
Machine learning significance
Local FDR

ABSTRACT

Random forests (RFs) are a widely used modelling tool capable of feature selection via a variable importance measure (VIM), however, a threshold is needed to control for false positives. In the absence of a good understanding of the characteristics of VIMs, many current approaches attempt to select features associated to the response by training multiple RFs to generate statistical power via a permutation null, by employing recursive feature elimination, or through a combination of both. However, for high-dimensional datasets these approaches become computationally infeasible. In this paper, we present RFlocalfdr, a statistical approach, built on the empirical Bayes argument of Efron, for thresholding mean decrease in impurity (MDI) importances. It identifies features significantly associated with the response while controlling the false positive rate. Using synthetic data and real-world data in health, we demonstrate that RFlocalfdr has equivalent accuracy to currently published approaches, while being orders of magnitude faster. We show that RFlocalfdr can successfully threshold a dataset of 10^6 datapoints, establishing its usability for large-scale datasets, like genomics. Furthermore, RFlocalfdr is compatible with any RF implementation that returns a VIM and counts, making it a versatile feature selection tool that reduces false discoveries.

1. Introduction

Random forests (RFs) are a non-linear modelling tool that has widespread popularity, from research to industry [1]. This versatility is, in part, due to the ability of RF to process large volumes of data efficiently [2], which is especially useful for genetic data as it is high-dimensional in nature with $p >> n$. Furthermore, RFs require little hyperparameter tuning and are robust against to overfitting due to its bootstrapping method. This minimises the need for data splits, which can be difficult with small sample sizes [3].

Crucially, variable importance measures (VIMs) can be extracted from RFs which enables the selection of features associated with the response via a threshold value. This is especially relevant for disease gene discovery, where highly associated genomic locations are identified as likely having a molecular causation on disease. As such, RF are a step in the direction of interpretable machine learning.

Yet, feature selection using VIMs is vulnerable to false positives and a statistical assessment is needed to confidently identify which features are significantly associated to the outcome. However, here is no theoretically defined VIM in the sense of a parametric quantity that a variable importance estimator should try to estimate [4]. Those that do have a firmer theoretical basis like Shapley values [1,5,6], or tackle the issues of bias in a comprehensive manner like 'conditional variable importance' [7], are too computationally intense to use with high-dimensional data like genomic data.

Instead, there are two main categories of empirical approaches to determine the significance of features associated to the response using RF VIMs: permutation of the response vector and recursive feature

* Corresponding author.
** Corresponding author at: Transformational Bioinformatics, Commonwealth Scientific and Industrial Research Organisation, Westmead, Australia.
E-mail addresses: rob.dunne@data61.csiro.au (R. Dunne), Denis.Bauer@CSIRO.au (D.C. Bauer).

https://doi.org/10.1016/j.csbj.2023.08.033
Received 4 June 2023; Received in revised form 30 August 2023; Accepted 30 August 2023
Available online 1 September 2023

elimination (RFE) [8]. In the permutation approaches, the response vector is permuted k times to calculate p values for significance from the permuted VIM, akin to the standard procedure for estimating false discovery rates. There currently are two state-of-art approaches utilizing permutations: the actual-impurity-reduction (AIR) [9] in combination with the Vita approach [10] and the Permutation IMPortance (PIMP) [11] algorithms. Permutation approaches are problematic for genomic data as their validity relies on the assumption that the statistic for gene j is a function of only the data for gene j, which is violated as VIMs are functionally related [12,13].

The second widely used method is RFE where RF are built recursively, removing a proportion of least important features before a new RF is generated with the remaining variables until a single feature is left. The verSelRF algorithm [14] and its modifications [8] uses the RFE approach. However, the time and compute implication of iteratively training RF can be prohibitive. Even with a smaller dataset of 500 samples × 50,000 gene expression levels (e.g., colon cancer [15]) with, say, $k = 100$ genes that are highly involved with the label, we can define $\rho = k/n$ and $\delta = n/p$ to get $\{\rho, \delta\} = \{0.2, 0.01\}$, putting the problem in a 'difficult' region of $\{\rho, \delta\}$ phase space where many algorithmic methods of feature recovery have a high probability of failing [16,17].

Combining permutation and RFE approaches, Boruta [18] generates shadow features by permuting the original features and recursively training RFs on this extended set until a stopping criterium is met. A test comparing the VIM of the real features to the maximum of all the shadow features determines which features are significantly associated. Recently rank aggregation techniques on multiple runs of feature selection methods have been proposed and tested on Boruta, Vita, and regularised RF [19] showing that the resulting consolidated features were more accurate and robust. However, for extremely high dimensional data both the use of shadow variables (i.e., Boruta) and/or running feature selection methods multiple times (i.e., rank aggregation) have significant resource impactions.

To address the shortcomings of currently availably feature selection approaches (AIR, Boruta, RFE, and PIMP), we approach the problem from a statistical standpoint. Here, p simultaneous tests can be performed on a genetic dataset $\underset{n \times p}{X}$ with $p >> n$ to get the test statistics $\{t_i\}_{i=1}^{p}$. Under assumptions about the distribution $\{t_i\}_{i=1}^{p}$, statistics z_j can be computed, comparing cases with controls which should give, by the central limit theorem, $z_j \sim N(\Delta_j, 1)$, where Δ_j is the effect size for gene j. Therefore, $|\Delta_j|$ is small for 'null genes' (i.e., genes that show the same activity in cases and controls), while $|\Delta_j|$ is large for genes having much different responses for cases versus controls. Inference for the individual genes gives the p value, $\{p_i\}_{i=1}^{p}$.

This is the position taken by the widely used linear models for microarray data (LIMMA) approach [20]. By using an empirical Bayes argument to make an adjustment to the variance of each gene based on a model for the variance of all genes in the sample, genes with very small variance are prevented from having a greatly inflated t statistic and from appearing significant. Another common practice is to permute the phenotype and derive a matrix of statistics of dimension p genes by M permutations. Approaches to determining the threshold of significance given such a matrix are discussed in [21].

For both approaches, the issue of multiple testing needs to be evaluated and there are three widely used approaches in genetic studies: false discovery rate (FDR) [22], Bonferroni correction [23], and q-value [24]. The FDR approach has demonstrated greater power to detect true positives than the simpler Bonferroni correction, while the q-value builds on the infimum over the p values of the FDR. Further to this, the 'local FDR' approach was proposed to control the FDR using an empirical Bayes estimate of the null distribution [25–28].

Based on Efron's local FDR approach, we developed RFlocalfdr, a method for setting a significance level of the VIM: the mean decrease in impurity (MDI) importances. The RFlocalfdr approach does not involve any refitting of RF and does not use 'shadow variables' [9], making it applicable to extremely high dimensional datasets, including genomics where the number of features may be in the millions.

2. Methods

2.1. Illustration of Efron's empirical Bayes approach

Several points make an ideal situation for an empirical Bayes estimate of the null distribution as discussed in Efron (2010). Firstly, the dataset is composed of two groups: a large group of data that will generate null values of some statistic, and a smaller group that will generate non-null values (Fig. 1 A). This means that there is sufficient data to model the null distribution, which Efron argues one should always do in cases like this, rather than make a distributional assumption. By adopting this approach, the high dimensionality of genetic datasets is now an asset as there are enough data points to estimate the null distribution accurately. Secondly, despite the large 'sample' sizes, the $N(0,1)$ Gaussian distribution has a very poor fit to the z values.

We illustrate Efron's empirical Bayes approach using a dataset from Hedenfalk et al. [28], consisting of a matrix with 3226 rows corresponding to the expression levels of genes and 15 columns corresponding to the 15 samples, divided between tumours with the BRCA1 and BRCA2 mutations. Let t_i be the standard t-statistics arising from the comparisons of cases and controls (i.e., tumours with or without the BRCA1/BRCA2 mutations). Let $z_i = \phi^{-1}(G_0(t_i))$, where ϕ is the standard normal cdf, and G_0 is a putative null cdf for the t-values. G_0 can be a theoretical null or a permutation null. Interestingly, in this case, the permutation density is very similar to the theoretical $N(0,1)$ density (Fig. 1B) so a non-parametric approach does not alleviate the problem.

As per Efron's approach, a histogram of the $\{z_i\}_{i=1}^{p}$ is plotted (Fig. 1B) and it demonstrates that the modelling assumptions are inaccurate as the distribution of z_i is not a $N(0,1)$ as shown in red. Efron discusses some reasons behind this occurrence including failed assumptions, correlations between cases or between features, and unobserved covariates.

The observed distribution is thus modelled as a mixture $f(z) = p_0 f_0(z) + (1 - p_0) f_1(z)$ where $f_0(z)$ is the null distribution of t-statistics and $f_1(z)$ is the distribution of significant t-statistics (Fig. 1b). The modelling process involves using the central mass of the null distribution to fit a Gaussian and then the local FDR is calculated as the ratio of a null density and the observed density of the tails $f_1(z)$. See Efron (2005, 2007, 2008, 2010) for more details.

The possibility of applying this empirical Bayes approach directly on the importances returned by AIR which have a distribution designed to be symmetric about 0 [9] is discussed and dismissed in the supplementary materials.

2.2. Empirical Bayes for MDI Importances

The RFlocalfdr, inspired by Efron's approach as introduced above, is a method for modelling the distribution of MDI importances from RF with a view to setting a significance threshold. This method depends on having a large number of features and a substantial count of trees making it ideally suited for genetic analyses, such as differential gene expression and GWAS.

We illustrate the RFlocalfdr method using a vector of MDI importances calculated from RF of the 1000 Genomes Projects [30], described further in the Results section. The density of the log transformed vector of MDI importances is considered (Fig. 2A), and is often the case, multi-model. We can model this as a mixture,

$$f(z) = p_A f_A(z) + p_B f_B(z) \tag{1}$$

and

$$f_B = p_0 f_0(z) + (1 - p_0) f_1(z) \tag{2}$$

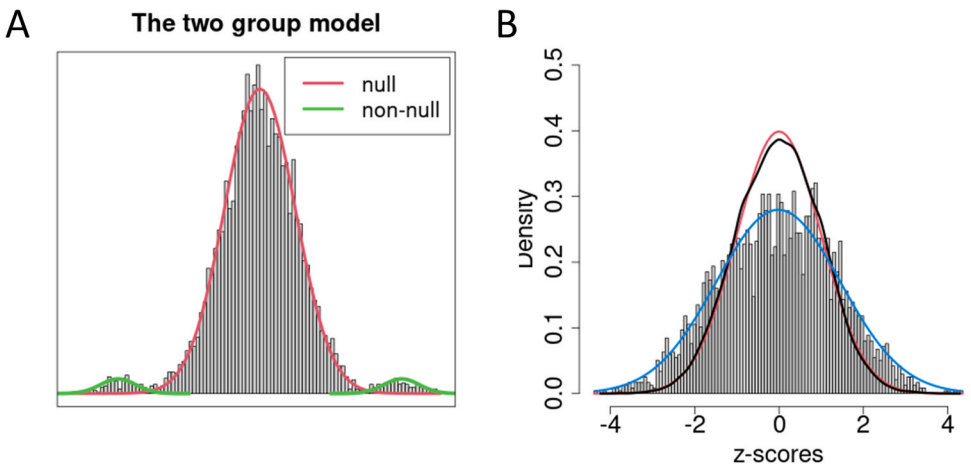

Fig. 1. (A) The two-group model where the data was generated by two processes, one of which produces a set of null statistics (density shown in red) and one which produces non-null statistics (density shown in green). (B) The histogram of z-values for the breast cancer genetic dataset [29]. The red curve shows the N(0, 1) distribution and the black curve shows the permutation null distribution, which is similar to the theoretical N(0,1) curve. The blue curve shows the empirical Bayes Gaussian fit to the data.

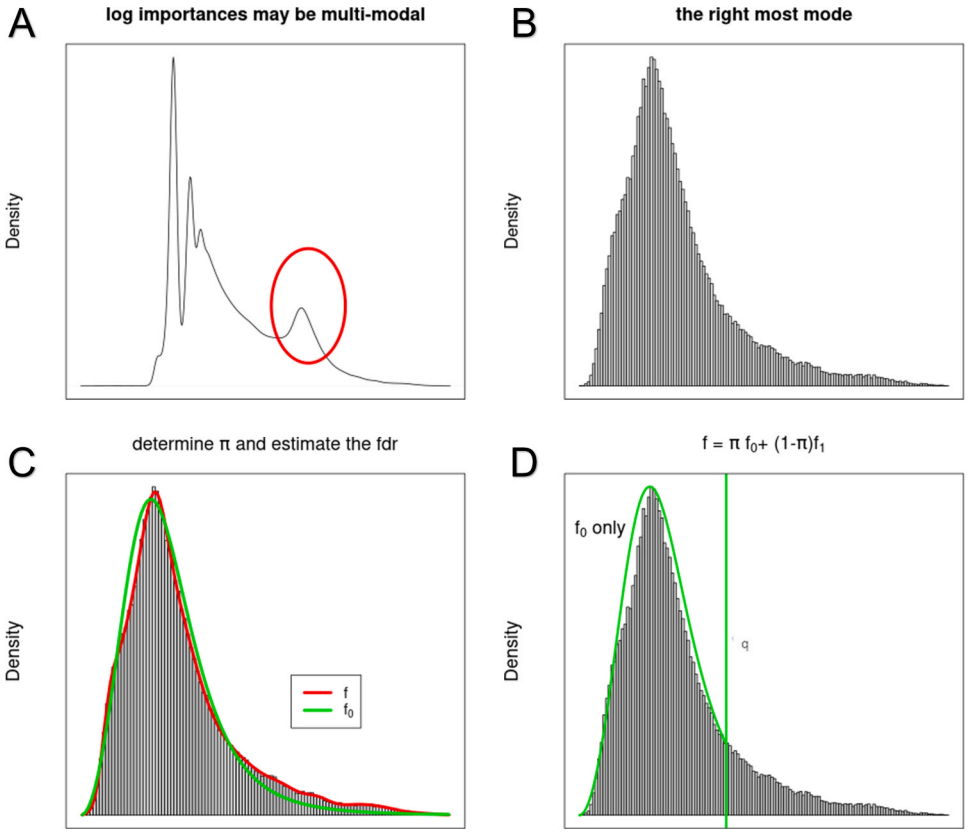

Fig. 2. The steps in estimating the local FDR from distributions of log MDI importances. (A) The density of the log MDI importances shows a multi-modal distribution. The density $p_0 f_0(z) + (1 - p_0) f_1(z)$ is indicated in red. (B) Histogram of log MDI importances of features that were used greater than 30 times in the RF showing the desired distribution. (C) A spline is fit to the observed bin counts from (B) using standard Poisson generalised linear modelling (f, coloured in red). (D) Identify a value q such that to the left of q, f only depends on f_0.

The distribution we are interested in Eq. (2), indicated in red in Fig. 2A, is assumed to be unimodal and that it is a mixture of null features and non-null features. Resultingly, we have two tasks: (1) to separate $f_A(z)$ and $f_B(z)$, (2) like Efron's problem, to estimate the empirical null $f_0(z)$ and calculate the local FDR.

For task 1, modelling $p_A f_A(z) + p_B f_B(z)$ by mixtures is quite difficult due to the wide range of distributions between different data sets. However, the modes of $f_A(z)$ are observed to be associated with features that were used a specific (and small) number of times in the RF. For example, the left most peak of $f_A(z)$ in Fig. 2A is composed of features that were used only once in the RF. There will often be another peak (at $-\infty$ as we have taken the log) of features that were not used at any time in the RF. Building on these observations, we denote the number of times each variable is used as C. We then progressively threshold C giving $f_c(z) = f(z) | C > c$. A skew-normal distribution [31] was explored to be a fit for $f_c(z)$, and either a Hartigan's diptest for unimodality or a goodness-of-fit test such as the Cramer-von Mises test is applied. However, as none of these procedures selects a c value giving a satisfactory fit (see supplementary materials), our current procedure is to fit a skew-normal $S_q(z)$ up to the q^{th} quantile of $f_c(z)$, and calculate the L_∞ norm $d_c = \max_z |f_c(z) - S_q(z)|$. The d_c is plot against c, and the minimum value c^* is chosen. The corresponding distribution, $f_{c^*}(z)$, is shifted along the z positive axis so that the smallest value is 0. For this dataset, the selected distribution is shown in Fig. 2B.

From a histogram of the selected distribution resulting from task 1 (Fig. 2B), we start task 2 by fitting a spline to the observed bin counts, denoted as f and shown in red in Fig. 2C. This can be done using standard Poisson generalised linear modelling software, fitting the counts to a

natural cubic spline basis on the midpoints of the bins. By assumption, there is a point q such that to the left of q $f_B \sim f_0(z)$, that is, there is a q such that there are only null features to the left of q. A change point method related to penalised model selection [32] is used to determine q (Fig. 2D) and then p_0 is estimated. Using only the data $< q$, a skew-normal is fit to f_0 using only the truncated range with the formulation of the skew-normal by [33]. The fit is done with non-linear least squares [34]. With f, p_0, and f_0, the local FDR is estimated as $fdr(x) = \frac{p_0 f_0(x)}{f(x)}$ and shown in Fig. 3.

3. Results

We applied the RFlocalfdr approach on five datasets: (1) a synthetic dataset with a strong correlation structure to establish the need for statistical approaches in feature selection, (2) chromosome 22 from the 1000 Genomes Project dataset to demonstrate real-world suitability, (3) a gene expression dataset from Spira et al. [35] to demonstrate suitability with non-genomic high-dimensional datasets, (4) the widely used benchmarking Boston housing dataset from Harrison and Rubinfeld [36] to illustrate the incompatibility of RFlocalfdr with a low-dimensional dataset, (5) a dataset of 10^6 datapoints to show applicability to large-scale genomic datasets like whole genome sequencing. Where possible, the RFlocalfdr approach was compared to four published approaches: AIR, Boruta, REF, and PIMP. Unless otherwise stated, the R package ranger [37] was used to build all RF and the parameters used are given in the supplementary materials.

3.1. Feature selection on a synthetic dataset

The synthetic dataset consists of 'bands' with 'blocks' of {1, 2, 4, 8, 16, 32, 64} of identical features (Fig. 4A). The features are $\in \{0, 1, 2\}$, a common encoding for genomic data where the numbers represent the number of copies of the minor allele. Only band 1 is used to calculate the y vector, and y is 1 if any of X [, (1, 2,4, 8, 16, 32, 64)] is non-zero. The result of this is that y is unbalanced, containing more 1's than 0's. In total, there are 50 bands and 200 observations, so X is 300 × 6350 with 127 non-null features (see supplementary materials for more details). A standard RF was fit to this dataset and the resulting MDI importances were recorded.

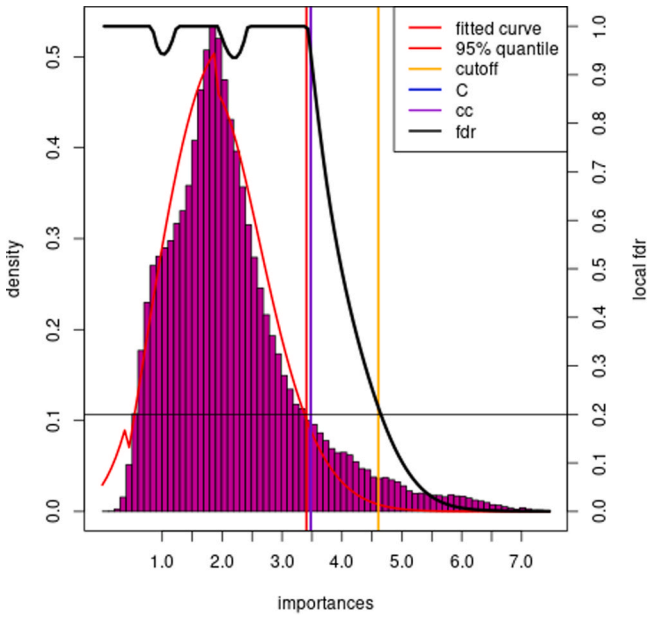

Fig. 3. An example of the plot produced from the RFlocalfdr approach. The FDR curve is shown in black.

As shown in Fig. 4B, selecting features above a single importance threshold did not achieve a perfect separation of the true positive features from band one (in red) and the false positives from the other bands. This issue was further exacerbated by blocks with fewer features having higher MDI importances as the importances was 'smeared' over the correlated (in this case, identical) features. The effects of correlation on MDI importances are further discussed in the supplementary materials. Therefore, a statistical approach is necessary for true positive selection.

Table 1 shows the performance measures of applying the five feature selection methods to this dataset to identify the features significantly associated with the response (y). Our RFlocalfdr approach resulted in the second highest recall after AIR (0.465 vs 1 respectively), it however had much higher precision (0.621 vs 0.318). Boruta and RFE resulted in the lowest recall (0.016 and 0.173 respectively) but RFE resulted in the highest precision (0.917). See supplementary materials for further details, including a multiple testing correction for the PIMP values.

3.2. Feature selection on real-world datasets

The 1000 Genomes Project dataset was obtained as VCF files from their FTP site with each VCF file containing the genotypes of single nucleotide polymorphisms (SNPs) for every individual. There are 2504 individuals with available genotypes in total and no additional processing was performed. An RF was used to predict the ethnicity of each individual using 1 million SNPs from chromosome 22. The script detailing the analysis in depth can be found in the supplementary materials.

The RFlocalfdr approach selected 6335 SNPs that were significantly associated with ethnicity at an FDR of 0.2. It had comparable performance to the more resource intensive Boruta algorithm, which returned 6773 significant SNPs with an 82 % overlap to RFlocalfdr (Fig. 5). Unlike RFlocalfdr which offers a continuous p value scale for the adjustment of FDR, Boruta offers only two levels and at the most conservative 'confirmed' option, it returns 1443 SNPs, all of which were also discovered by RFlocalfdr. Furthermore, a one-way ANOVA showed a significant association ($P < 2 \times 10^{-16}$) between RFlocalfdr p values and Boruta categories.

As in the simulated example, AIR and PIMP select orders of magnitude (~10x and ~100x respectively) more significant SNPs than RFlocalfdr and Boruta (Table 2), potentially having a higher sensitivity. However, AIR assigns a p value of 0 to all SNPs, making it impossible to tune the specificity which may result in a high false positive rate (Fig. 5). The final feature selection approach, RFE, appeared to have focussed on specificity and returned 59 nested sets, with a set of only 12 SNPs having the smallest prediction error.

The runtimes of each approach were evaluated using the same processing parameters (Table 2). As RFs can be built in parallel, we report runtime in units of RF generation. RFlocalfdr only requires one built of RF and its relative runtime is 1, making it twice as fast as the next fastest method AIR which requires 'shadow variables' to be built. Compared to the RFE approach, RFlocalfdr was 57 times faster, although the runtime of RFE is data dependent and each run of RF in this approach will have a reduced number of features making it faster as it goes. Finally, the RFlocalfdr approach was 100 times faster than the Boruta and PIMP approaches, but we note that the number of evaluations of RF is a user set parameter of Boruta and PIMP. Runtimes with full details are discussed further in the supplementary materials.

We have also tested the RFlocalfdr approach on two non-genomic datasets. Firstly, the gene expression dataset from Spira et al. [35] to demonstrate the generalisability of the approach on non-genomic high-dimensional datasets where RFlocalfdr identified 19 significant genes, see Supplemental Material D. Secondly, on the Boston Housing data from Harrison and Rubinfeld [36]. In this dataset, the RFlocalfdr approach was unsuccessful due to the small number of non-null features ($n = 17$). However, with the addition of 5000 non-informative features, RFlocalfdr identified 15 of the 17 features as significant, see

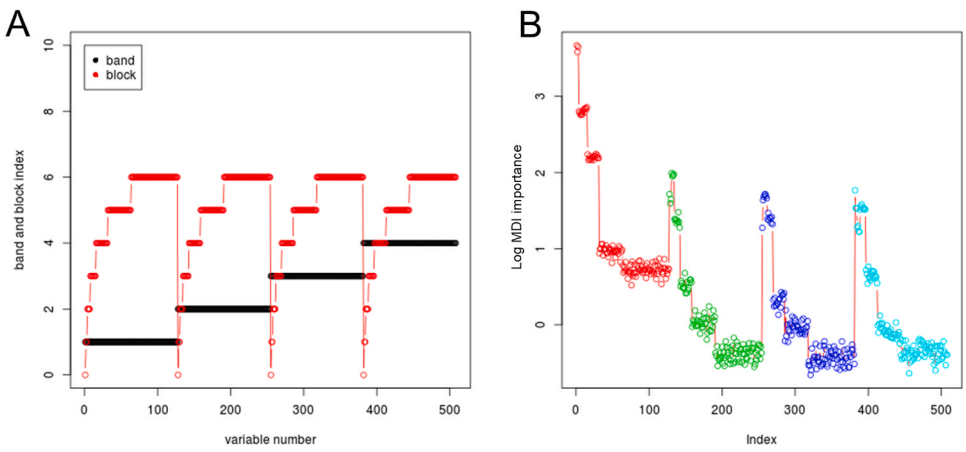

Fig. 4. (A) The synthetic data is structured into bands and blocks. The colour and the *y*-axis show which band/block each feature/variable belongs to, not the feature value. Each 'band' contains 'blocks' of sizes 1, 2, 4, 8, l6, 32, and 64. Each block consists of correlated (identical variables), where each variable is $\in \{0, 1, 2\}$. The dependent variable *y* is 1 if any of *X*[, *c*(1, 2, 4, 8, 16, 32, 64)] is non-zero, so only band 1 has a relationship to the dependent variable. (B) The log MDI importances from the RF on the synthetic dataset, arranged by feature number and coloured by band. It is impossible to threshold the MDI importances to recover the only non-null features (coloured in red).

Table 1
Performance measures of feature selection for the simulated dataset by five methods of feature selection. The best outcomes for each category are in bold face.

Method	True Positives	False Positives	Recall	Precision
AIR	127	273	**1**	0.318
Boruta	2	2	0.016	0.500
RFE	22	2	0.173	**0.917**
RFlocalfdr	59	36	0.465	0.621
PIMP	39	556	0.307	0.067

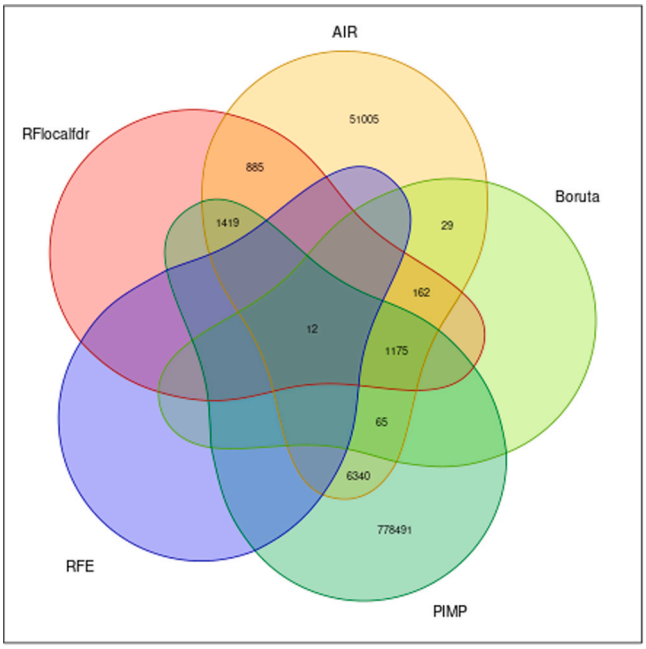

Fig. 5. A Venn diagram of the overlaps in features (i.e., SNPs) classified as significant by AIR, Boruta, RFE, RFlocalfdr, and PIMP. AIR and PIMP are the outliers with more than 10^4 unique SNPs (i.e., not found by the other approaches) for each approach.

Supplemental Material E.

3.3. Feature selection on large-scale dataset (10^6 datapoints)

This dataset was presented in Bayat et al. [2] with 10,000 samples and 6 million SNPs as features. Although based on the 1000 Genomes Project dataset, it has a synthetic *y* value allowing us to calculate accuracy while operating on a dataset with true biological structure.

However, due to the size of the dataset, ranger could not be applied, and we had to evaluate two other Apache Spark-based RF implementations: VariantSpark [2] and ReForest [38]. ReForest fitted the model in about 15 h, compared to VariantSpark which took 5 h. We hence used VariantSpark going forward.

Using the MDI and counts reported by VariantSpark, RFlocalfdr selected 38 SNPs as significant, including the 5 SNPs used to generate the simulated *y* value. This resulted in a 100 % success rate with an 87 % false discovery rate. Given the runtimes reported in Table 2, fitting this large-scale dataset would require up to 62 days for the other feature selection algorithms to complete. Furthermore, none of the other methods accepts MDI importances and counts of variables as is, which is returned by VariantSpark and ReForest. Hence, RFlocalfdr is the only approach capable of selecting variants associated with a trait using whole-genome size datasets in a feasible timeframe and out-of-the-box.

Table 2
The number of significant SNPs returned by each feature selection approach and their runtimes. Runtimes are expressed in multiples of the runtime of a single Ranger fit for the given hardware configuration, hence, the runtimes describe the number of 'refits' that each method requires. The processing time outside of the RF fit is negligible, by comparison, in all cases.

Method	SNPs Returned as Significant	Runtime (as multiples of a single RF)
AIR	61,092	1.5–2
Boruta	6773	100
RFE	12	50–60
RFlocalfdr	6335	1
PIMP	787,502	100

4. Discussion

In this paper, we present RFlocalfdr, a method for calculating the significance threshold of RF MDI importances for detecting label-associated features using an empirical Bayes approach. The accuracy of RFlocalfdr was shown to be comparable to the currently published more resource intensive techniques in terms of performance metrics and also demonstrated advantages. This includes: (1) computational efficiency as it requires only a single fit of RF particularly in comparison to RFE or permutation methods such as PIMP, (2) broad applicability to any RF implementation that returns MDI importances and counts of variables use, and (3) it provides continuous *p* values, which allows for tailored sensitivity and specificity selections.

Of the other methods tested in this paper, only RFE offers a similar capability to adjust the sensitivity/specificity trade-off through the selection of other nested sets with a larger or smaller number of associated

features. In contrast, Boruta and AIR do not offer this capability as there is no criteria for sub-setting the set of associated features, and in practice, AIR tends to report a vastly larger number of significant features. Whether this large number of reportedly significantly features is desirable would depend on the context of the dataset, subsequent analysis, and the false discovery tolerance.

As shown in Bayat et al. [2], the ranger implementation of RF was unable to process datasets of $>$ 3.2 M features and 10,000 samples with 488 GB of memory. For ranger to fit RF with a large-scale dataset of 10^6 datapoints, several hundred cores and several terabytes of RAM will be required. Therefore, feature selection approaches that require multiple RF refits (e.g., RFE, PIMP) in conjunction with RF implementations where computational requirements exponentially increase with data size (e.g., ranger, ReForest) will eventually become infeasible. This is particularly true for genetic analyses as availability of large-scale genomic datasets become commonplace. The RFlocalfdr approach circumvents these challenges as it only requires a single RF fit and can be applied to any RF implementation that returns MDI importances and counts of features used, such as the highly scalable VariantSpark implementation.

RFlocaldfr is highly applicable in problems where there are sufficient null variables to allow the accurate estimation of the null density. In addition, where the variables are all of the same type and on the same scale, the variable importances will be less susceptible to bias.

issues [7]. However, these caveats still leave large areas of applicability for RFlocaldfr, for example, high-throughput biomolecular data generally meets these criteria.

In conclusion, it is our expectation that RFlocalfdr with its direct and real-time capability of detecting trait-associated SNPs will greatly assist the analysis of high-dimensional data, such as genomic data.

Software and data availability statement

The RFlocalfdr approach is available as an R package through github (https://github.com/parsifal9/RFlocalfdr) and a python version is included in the VariantSpark github repository (https://github.com/aehrc/VariantSpark/blob/master/python/varspark/stats/lfdr.py). The python script has been tested using python version 3.8.12 and requires the following libraries: numpy v1.21.2 for numeric transformations, pandas v1.4.1 to create and manage data frames, patsy v0.5.2 to create the cubic regression splines, scipy v1.7.3 to fit the data using the least square method, to compte the cdf and percentile point functions, and statsmodels v0.13.2 to fit a generalised linear model. The user of the python script is only required to run the 'fit' function which requires a data frame as input with a single column containing the log-transformed RF MDI importances for each variant. Currently the function will return a touple with the estimated false discovery rate and a data frame with the p values for the statistically significant features.

The code used to generate the synthetic dataset is provided in the github repository (https://github.com/parsifal9/RFlocalfdr/blob/main/vignettes/simulated.Rmd). The 1000 Genomes Project phase 3 dataset can be obtained and downloaded from their FTP site as VCF files (https://www.internationalgenome.org/data-portal/data-collection/phase-3). Access and availability of the large-scale dataset with 10^6 datapoints of 10,000 samples and 6 million features is described in Bayat et al. [2].

Authors' contributions

RD conceived, designed, and implemented the RFlocalfdr method. RD, RR, PRM, PS, LMFS, ML acquired, analyse, and interpreted the data. RD, RR, LMFS, NAT, DCB drafted and edited the manuscript. All authors read and approved the final manuscript.

CRediT authorship contribution statement

Robert Dunne: Conceptualisation, Methodology, Software, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualisation. **Roc Reguant**: Software, Writing – review & editing. **Priya Ramarao-Milne**: Writing – review & editing. **Piotr Szul**: Investigation, Software, Formal analysis. **Letitia M.F. Sng**: Investigation, Resources, Writing – review & editing. **Mischa Lundberg**: Investigation, Writing – review & editing. **Natalie A. Twine**: Writing – review & editing, Supervision. **Denis C. Bauer**: Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare no conflict of interest.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.08.033.

References

[1] Lundberg S.M., Erion G.G., Lee S.-I. Consistent Individualized Feature Attribution for Tree Ensembles. ArXiv180203888 Cs Stat 2019.
[2] Bayat A, Szul P, O'Brien AR, Dunne R, Hosking B, Jain Y, et al. VariantSpark: cloud-based machine learning for association study of complex phenotype and large-scale genomic data. GigaScience 2020:9. https://doi.org/10.1093/gigascience/giaa077.
[3] Janitza S, Hornung R. On the overestimation of random forest's out-of-bag error. PLOS ONE 2018;13:e0201904. https://doi.org/10.1371/journal.pone.0201904.
[4] Grömping U. Variable importance assessment in regression: linear regression versus random forest. Am Stat 2009;63:308–19. https://doi.org/10.1198/tast.2009.08199.
[5] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30:4765–74.
[6] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2:56–67. https://doi.org/10.1038/s42256-019-0138-9.
[7] Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinforma 2008;9:307. https://doi.org/10.1186/1471-2105-9-307.
[8] Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics data sets. Brief Bioinform 2019;20:492–503. https://doi.org/10.1093/bib/bbx124.
[9] Nembrini S, König IR, Wright MN, Valencia A. The revival of the Gini importance? Bioinformatics 2018;34:3711–8. https://doi.org/10.1093/bioinformatics/bty373.
[10] Janitza S, Celik E, Boulesteix A-L. A computationally fast variable importance test for random forests for high-dimensional data. Adv Data Anal Cl 2016:1–31. https://doi.org/10.1007/s11634-016-0270-x.
[11] Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics 2010;26:1340. https://doi.org/10.1093/bioinformatics/btq134.
[12] Witten DM, Tibshirani R. Testing significance of features by lassoed principal components. Ann Appl Stat 2008;2:986–1012. https://doi.org/10.1214/08-AOAS182.
[13] Huynh-Thu VA, Saeys Y, Wehenkel L, Geurts P. Statistical interpretation of machine learning-based feature importance scores for biomarker discovery. Bioinformatics 2012:28. https://doi.org/10.1093/bioinformatics/bts238.
[14] Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. BMC Bioinforma 2006;7:3. https://doi.org/10.1186/1471-2105-7-3.
[15] LaPointe LC, Pedersen SK, Dunne R, Brown GS, Pimlott L, Gaur S, et al. Discovery and validation of molecular biomarkers for colorectal adenomas and cancer with application to blood testing. PLoS ONE 2012:7. https://doi.org/10.1371/journal.pone.0029059.
[16] Donoho D., Stodden V. Breakdown point of model selection when the number of variables exceeds the number of observations. 2006 IEEE Int. Jt. Conf. Neural Netw. Proc., IEEE; 2006, p. 1916–1921.
[17] Donoho D, Tanner J. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. Philos Trans R Soc Lond Math Phys Eng Sci 2009;367:4273–93. https://doi.org/10.1098/rsta.2009.0152.
[18] Kursa MB, Rudnicki WR. Feature selection with the boruta package. J Stat Softw 2010;36:1–13.
[19] Pfeifer B, Holzinger A, Schimek MG. Robust random forest-based all-relevant feature ranks for trustworthy AI. Chall. Trust. AI added-value health. IOS Press 2022:137–8. https://doi.org/10.3233/SHTI220418.

[20] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47. https://doi.org/10.1093/nar/gkv007.
[21] Churchill GA, Doerge RW. Empirical threshold values for quantitative trait mapping. Genetics 1994;138:963–71.
[22] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 1995;57:289–300.
[23] Korthauer K, Kimes PK, Duvallet C, Reyes A, Subramanian A, Teng M, et al. A practical guide to methods controlling false discoveries in computational biology. Genome Biol 2019;20:118. https://doi.org/10.1186/s13059-019-1716-1.
[24] Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci 2003;100:9440–5. https://doi.org/10.1073/pnas.1530509100.
[25] Efron B. Local False Discovery Rates 2005.
[26] Efron B. Correlation and large-scale simultaneous significance testing. J Am Stat Assoc 2007;102:93–103. https://doi.org/10.1198/016214506000001211.
[27] Efron B. Microarrays, empirical bayes and the two-groups model. Stat Sci 2008;23:1–22. https://doi.org/10.1214/07-STS236.
[28] Efron B. Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Cambridge: Cambridge University Press,; 2010. https://doi.org/10.1017/CBO9780511761362.
[29] Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, et al. Gene-expression profiles in hereditary breast cancer. N Engl J Med 2001;344:539–48.
[30] Fairley S, Lowy-Gallego E, Perry E, Flicek P. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. Nucleic Acids Res 2020;48:D941–7. https://doi.org/10.1093/nar/gkz836.
[31] Azzalini A. The R package \textttsn: The skew-normal and related distributions such as the skew-t and the SUN (version 2.0.2). 2022.
[32] Gauran IIM, Park J, Lim J, Park D, Zylstra J, Peterson T, et al. Empirical null estimation using zero-inflated discrete mixture distributions and its application to protein domain data. Biometrics 2018;74:458–71. https://doi.org/10.1111/biom.12779.
[33] Ashour SK, Abdel-hameed MA. Approximate skew normal distribution. J Adv Res 2010;1:341–50. https://doi.org/10.1016/j.jare.2010.06.004.
[34] Elzhov T.V., Mullen K.M., Spiess A.-N., Bolker B. minpack.lm: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK, Plus Support for Bounds. 2022.
[35] Spira AE, Beane J, Pinto-Plata V, Kadar A, Liu G, Shah V, et al. Gene expression profiling of human lung tissue from smokers with severe emphysema. Am J Respir Cell Mol Biol 2004.
[36] Harrison D, Rubinfeld DL. Hedonic housing prices and the demand for clean air. J Environ Econ Manag 1978;5:81–102. https://doi.org/10.1016/0095-0696(78)90006-2.
[37] Wright MN, Ziegler A. Ranger: a fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw 2017;77:1–17. https://doi.org/10.18637/jss.v077.i01.
[38] Lulli A, Oneto L, Anguita D. ReForeSt: random forests in apache spark. In: Lintas A, Rovetta S, Verschure PFMJ, Villa AEP, editors. Artif. Neural Netw. Mach. Learn. – ICANN 2017. Cham: Springer International Publishing; 2017. p. 331–9. https://doi.org/10.1007/978-3-319-68612-7_38.