OXFORD UNIVERSITY PRESS | Bioinformatics

r2VIM: A new variable selection method for random forests in genome-wide association studies

Journal:	*Bioinformatics*
Manuscript ID:	Draft
Category:	Original Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Szymczak, Silke; National Institutes of Health, National Human Genome Research Institute Holzinger, Emily; National Institutes of Health, National Human Genome Research Institute Dasgupta, Abhijit; National Institutes of Health, National Institute of Arthritis and Musculoskeletal and Skin Diseases Malley, James; National Institutes of Health, Center for Information Technology Molloy, Anne; Trinity College Dublin, School of Medicine Mills, James; National Institutes of Health, Eunice Shriver National Institute of Child Health and Human Development Brody, Lawrence; National Institutes of Health, National Human Genome Research Institute Stambolian, Dwight; University of Pennsylvania, Ophthalmology Bailey-Wilson, Joan; National Institutes of Health, National Human Genome Research Institute
Keywords:	Classification, Feature selection, Genetics, Machine learning, SNPs, Random Forest

BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–6

r2VIM: A new variable selection method for random forests in genome-wide association studies

Silke Szymczak [1,8] Emily R Holzinger [1] Abhijit Dasgupta [2] James D Malley [3] Anne M. Molloy [4] James L Mills [5] Lawrence C Brody [6] Dwight Stambolian [7] and Joan E Bailey-Wilson [1*]

[1]Statistical Genetics Section, Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, MD 21224, USA
[2]Clinical Trials and Outcomes Branch, National Institute of Arthritis and Musculoskeletal and Skin Diseases, National Institutes of Health, Bethesda, MD 20892, USA
[3]Division of Computational Bioscience, Center for Information Technology, National Institutes of Health, Bethesda, MD 20892, USA
[4]Department of Clinical Medicine, School of Medicine, Trinity College Dublin, Dublin 2, Ireland
[5]Division of Intramural Population Health Research, Eunice Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20892, USA
[6]Molecular Pathogenesis Section, Medical Genomics and Metabolic Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA
[7]Department of Ophthalmology, University of Pennsylvania, Philadelphia, PA 19104, USA
[8]Current address: Institute of Medical Informatics and Statistics, Christian-Albrechts University Kiel, Kiel, Germany

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Machine learning methods and in particular random forests (RFs) are a promising alternative to standard single SNP analyses in genome-wide association studies (GWAS). RFs provide variable importance measures (VIMs) to rank SNPs according to their predictive power. However, in contrast to the established genome-wide significance threshold, no clear criteria exist to determine how many SNPs should be selected for downstream analyses.

Results: We propose a new variable selection approach, recurrent relative variable importance measure (r2VIM). Importance values are calculated relative to an observed minimal importance score for several runs of RF and only SNPs with large relative VIMs in all of the runs are selected as important. Evaluations on simulated GWAS data show that the new method controls the number of false-positives under the null hypothesis. Under a simple alternative hypothesis with several independent main effects it is only slightly less powerful than logistic regression. In an experimental GWAS data set, the same strong signal is identified while the approach selects none of the SNPs in an underpowered GWAS.

Availability: The approach is implemented as a R package called `r2VIM` (available at http://research.nhgri.nih.gov/software/r2VIM).

Contact: jebw@mail.nih.gov; szymczak@medinfo.uni-kiel.de

*to whom correspondence should be addressed

1 INTRODUCTION

In the last few years, genome-wide association studies (GWAS) have identified more than one thousand single-nucleotide polymorphisms (SNPs) that are reproducibly associated with more than two hundred phenotypes and quantitative traits (Hindorff *et al.*, 2009). However, these common variants evidently explain only a small proportion of the overall heritability (Manolio *et al.*, 2009). One major problem is that the standard approach in GWAS analyzes each SNP separately and is therefore not designed to identify genetic variants that have a strong joint effect on the phenotype. While the assumption of individual SNPs always acting independently makes little biological sense, deriving and modeling plausible alternatives is still a major challenge.

Nonparametric, model-free statistical learning machines are a family of promising alternatives to classical, model-based statistical methods. Popular learning machines, such as Random Forests (RFs) (Breiman, 2001), are known to be statistically optimal and are computationally efficient when run in parallel on distributed systems. RF is an ensemble method based on a large number of classification and regression trees trained on bootstrap samples and has been successfully applied to identify SNPs influencing susceptibility to disease, e.g. (Jiang *et al.*, 2009; Schwarz *et al.*, 2010; Goldstein *et al.*, 2011).

A nice feature of RF are variable importance measures (VIMs) that can be used to order and select the most predictive SNPs. But the actual importance values are difficult to interpret as they depend not only on the signal in the data but also on

the parameters of the algorithm (Genuer *et al.*, 2008). Usually, SNPs are ranked according to decreasing importance values and the top ranked SNPs are declared as important. The number of selected SNPs is often arbitrary and several approaches have been proposed to objectively determine a threshold. A classical statistical test could be used by estimating z-scores and calculating asymptotic p-values (Breiman and Cutler, Random Forests, http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm). However, the power of this test depends on the number of trees, which is a tuning parameter in RF. Therefore, this method is not recommended (Strobl and Zeileis, 2008). As an alternative, the null distribution of the VIMs can be estimated by permuting phenotype status (see e.g. R package rfpermute). Unfortunately, this approach would require at least 1000 runs of RF and is therefore computationally prohibitive for GWAS data sets. Therefore, it is difficult to decide how many SNPs should be selected with the threshold being somewhat arbitrary. As a consequence, no clear criteria exist to decide if RF is able to identify any important SNP or if the study is underpowered. Indeed, simulations have shown that when the effects of the causal SNPs on the trait are low and/or sample size is not extremely large, then most of the SNPs with strongest VIMs are not causally related to the trait (Kim *et al.*, 2009, 2011).

Here we present a novel variable selection procedure called recurrent relative variable importance measure (r2VIM). Several runs of RF are performed each resulting in importance values calculated relative to the observed minimal importance score. Only SNPs with large relative VIMs in all of the runs are declared as important. GWAS data with realistic local linkage disequilibrium patterns were simulated to evaluate false-positives and emprical power compared to logistic regression. Analysis of two experimental GWAS, one that has a strong signal and another one that is underpowered, illustrate the applicability of our new method.

2 METHODS

2.1 Random Forest

RF is a machine learning approach that combines many classification and regression trees into a committee or ensemble (Breiman, 2001). Each tree is built using a bootstrap sample of the data set and at each node the optimal variable is selected from a random subset of all predictor variables. Majority voting over all trees is used to classify a sample using the ensemble. In addition to prediction, RF offers VIMs to assess the predictive power of each variable. The most reliable measurement is the (unscaled) permutation importance (Nicodemus *et al.*, 2010) that measures the difference in prediction accuracy before and after permuting values of the variable, averaged over all trees.

2.2 New variable selection method r2VIM

Our proposed variable selection method r2VIM is based on the permutation importance scheme, a standard component of RF. Our method has three components. First, instead of performing a single run of RF and selecting a few top ranked variables, we propose running RF several times with different random number seeds. Then, trees in each run over several forests will be slightly different leading to random variability in VIMs, where the randomness is partly sample based and partly seed based. Variables more predictive of the outcome will have relatively high importance scores in each of the runs, while other, less predictive variables will have only randomly high importance scores. The second component of the scheme is that variables with little predictive capacity will have importance values close to zero. It is useful to note that the variable importance values in RF will generate importances that may be negative. Therefore, since most SNPs in a GWAS setting are not expected to be associated with the disease, the smallest, usually negative, observed importance score across the variables, SNPs in a GWAS, can be used as an approximate estimate of the variability for variables with no predictive power. The related idea here is that noise variables will have importances randomly and symmetrically above and below zero. For each variable, we define a relative importance score by dividing its value by the the observed minimal importance score. Hence all SNPs with a relative importance score larger than 1 or in general a factor f could then defined to be important (see e.g. (Strobl *et al.*, 2009)), or more accurately, not unimportant. The last part combines the other two components by declaring only those SNPs as important that have relative importance scores $> f$ in each of the runs.

For all analyses presented in this paper we used ten runs and factors f = 1, 3, and 5. As shown in the results, using $f = 1$ identifies too many false-positive SNPs under the null hypothesis. That is, the simple observed minimum, negative importance value is not a good estimate of the variance of importances across noise features, while simple multiples of the observed minimum seem to do quite well.

Table 1. Information about nine causal SNPs under the alternative hypothesis

SNP	MAF	RR	# SNPs (strong LD)	# SNPs (moderate LD)
11-103959987	0.474	1.3	0	1
22-28469630	0.488	1.3	4	8
17-9807099	0.496	1.3	0	0
1-240799543	0.312	1.5	0	0
7-45984820	0.312	1.5	0	2
5-130104076	0.323	1.5	12	18
14-67463012	0.062	2.0	2	31
18-34645639	0.062	2.0	0	1
3-2770509	0.064	2.0	0	1

Table shows SNP identifier in chromosome and position notation, minor allele frequency (MAF), relative risks (RR) and number of SNPs within a 1 Mb region that are in strong ($r^2 > 0.8$) or moderate LD ($0.3 < r^2 \geq 0.8$).

2.3 Simulation study

To evaluate our new variable selection method we simulated genome-wide data sets with realistic local linkage disequilibrium patterns. Haplotypes from 381 European individuals provided by the 1000 genomes project were used as input data for the software GWAsimulator (Li and Li, 2008) to simulate new haplotypes for a case-control study. 554,813 SNPs from the Illumina Human660W chip were selected and 10 replicates generated. We used total sample sizes of 2000 and 6000 with a balanced number of cases and controls.

To estimate the number of false-positive SNPs, a null hypothesis was simulated where case-control status was not dependent on any SNP but was assigned randomly. For empirical power estimation, a simple alternative hypothesis was generated. Case-control status was determined by nine independent causal SNPs, each with multiplicative (on relative risk level) main effects. No gene-gene interaction effects were simulated. For reasonable power, most of the causal SNPs were common with minor allele frequencies (MAF) of 0.3 or nearly 0.5 and relative risks for one minor allele was set to 1.5 or 1.3. In addition, three less common SNPs with MAF of 0.06 and a relative risk of 2 were included into the model. Detailed information about all nine SNPs can be found in Table 1.

RF analyses were performed with RandomJungle (Schwarz *et al.*, 2010) version 1.2.365. For each of the ten runs per replicate, 1000 classification

trees were generated and 100,000 (about 20%) or 250,000 (about 50%) SNPs randomly selected at each node. The number of samples in terminal nodes was restricted to 100 and 300 for sample sizes of 2000 and 6000, respectively. To make the analyses computationally feasible, depth of trees was limited to 3. Important SNPs were selected using our new variable selection method with factors of 1, 3 and 5, i.e. only SNPs with relative importance scores $>$ 1, 3 and 5 in each of the ten runs were selected. For comparison, logistic regression in PLINK (Purcell *et al.*, 2007) version v1.07 was performed for each SNP separately and SNPs with a p-value smaller than the genome-wide significance level of $5 * 10^{-8}$ were selected. Type I errors and empirical power were estimated for each SNP separately using the proportion of replicates in which a particular SNP was identified. A SNP was declared a false-positive if it was not in LD with any causal SNP.

2.4 Experimental data

We selected two GWAS studies to illustrate application of the new method on real data sets. To compare results from a GWAS with a strong signal we used GWAS data from the Trinity Student Study (TRINITY) which examines traits related to folate and vitamin B12 metabolism in healthy young Irish individuals who were students at Trinity College in Dublin at the time of study enrollment (Desch *et al.*, 2013). The analyzed phenotype is total serum bilirubin (TBIL) measured as a quantitative trait. For illustration purposes, we selected individuals at the extremes of the distribution. 193 individuals with TBIL $>$ 17 and 241 individuals with TBIL $<$ 5 were defined as cases and controls, respectively. Since missing values pose a problem for RF, quality controlled SNPs were imputed with PLINK using CEU individuals from phase 2 of the HapMap project as reference panel resulting in 873,565 common SNPs with complete genotypes.

As a negative control data set we chose a GWAS study with a relatively small sample size so that the power to identify a real effect is very low. Data are from the Age-Related Eye Disease Study (AREDS), a cohort study focusing on risk factors for the development of age-related macular degeneration (AMD) and cataract (Simpson *et al.*, 2013; Stambolian *et al.*, 2013). Mean spherical equivalent (MSE) of both eyes was calculated on study participants without either AMD or cataracts at the first study visit. A binary phenotype hyperopia defined 858 cases as those with MSE $\geq$ +1D and 602 controls with MSE $<$ 0D. Quality-controlled SNPs were imputed using MACH (Li *et al.*, 2010) based on HapMap phase 2 reference panel. To reduce the number of SNPs for analysis, LD pruning was performed using PLINK with pairwise r^2 of 0.99 as threshold. 908,293 common SNPs with complete genotypes remained for analysis.

Parameters for RF analysis were similar to the simulated data. In brief, 1000 trees restricted to a depth of 3 and 5% of the sample size in the terminal nodes were generated in ten runs of RF per data set. Twenty percent of the SNPs were randomly selected at each node and factors 1, 3 and 5 were used for variable selection.

3 RESULTS

3.1 Simulation study

Results under the null hypothesis with case-control status assigned randomly are shown in Figure 1. As expected, no SNP reaches genome-wide significance in all ten replicates for logistic regression. In contrast, the number of false-positive SNPs identified by the new variable selection procedure depends on the factor that is used to define the threshold for declaring SNPs as important (see Figure 1). If a liberal factor of 1 is used, between seven and 13 SNPs are selected across settings. Three and two SNPs on chromosomes 4 and 8 are highly correlated (pairwise $r^2 > 0.8$), resulting in five to 12 independent regions. However, all SNPs have been selected in only one replicate and each SNP is selected either with a sample size of 2000 or 6000. In general, a smaller number of false-positive SNPs is identified for the larger sample size. If the factor is increased

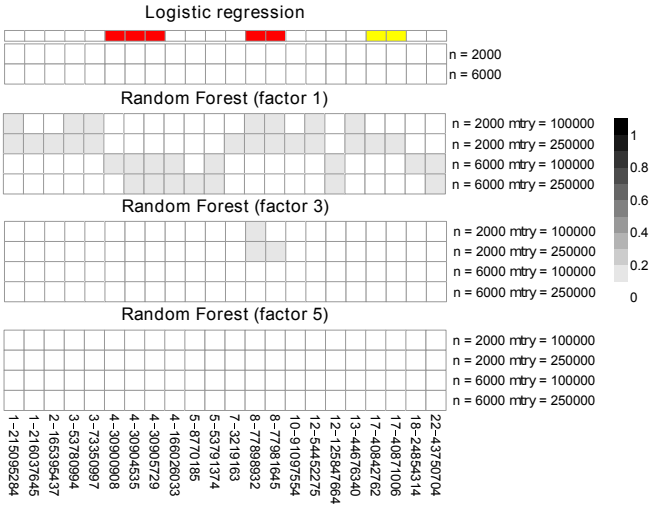

Fig. 1. Heatmaps showing type I error of single SNPs in simulated GWAS for logistic regression and variable selection method with several factors in the different scenarios (different sample sizes and *mtry* parameters in random forest). Columns correspond to SNPs that are selected in at least one approach and are ordered by chromosomal position. Type I error is color-coded in gray with white and black denoting 0 and 1, respectively. In addition, LD information is shown at the top with SNPs in high ($r^2 > 0.8$) and moderate LD ($0.3 < r^2 \leq 0.8$) colored in red and yellow.

to 3, only the region on chromosome 8 is selected for the smaller sample size whereas none of the SNPs is found for the larger one. In addition, if the most stringent factor of 5 is used, type I error is well controlled since the new variable selection procedure declares none of the 500,000 SNPs as important.

Empirical power under the alternative hypothesis is summarized in Figure 2 and Table 2. Detailed information about each SNP that was detected in at least one replicate and with at least one method are given in Supplementary Table 1. With logistic regression eight out of the nine causal SNPs have empirical power $>$ 0 for the smaller sample size. However, only the three common SNPs with relative risks of 1.5 have significant p-values in more than 5 replicates. In the larger data set, all causal SNPs are identified in all ten replicates. All other SNPs with significant p-values are in LD with one of the causal SNPs. The new variable selection method identifies seven and nine causal SNPs with a sample size of 2000 and 6000, respectively. However, power decreases from factor 1 to 3 and 5. The largest reduction in power is observed for the very common SNPs with small effects on chromosomes 17 and 22. Increasing the factor value also reduces the number of selected SNPs that are correlated with one of the causal SNPs. In concordance with results under the null hypothesis, using a factor of 1 results in identification of four to 13 false-positive SNPs (each observed in only a single replicate) that are not correlated with any of the causal SNPs and that are often located on other chromosomes. Interestingly, more false-positives are observed for the larger *mtry* value for both sample sizes. Again, the number is greatly reduced for a factor of 3 with one false-positive SNP identified with *mtry* = 250000. And only causal SNPs or SNPs correlated with causal SNPs are selected with a factor of 5.

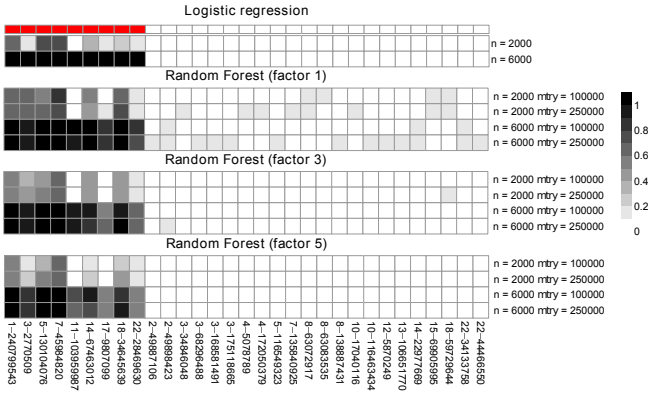

Fig. 2. Heatmaps showing empirical power of single SNPs in simulated GWAS for logistic regression and variable selection method with several factors in the different scenarios (different sample sizes and *mtry* parameters in random forest). Only the nine causal SNPs (marked in red on top) and false-positive SNPs that are uncorrelated to each causal SNP are shown in columns and ordered by chromosomal position. Empirical power is color-coded in gray with white and black denoting 0 and 1, respectively.

Table 2. Number of SNPs in simulated GWAS with empirical power > 0 for logistic regression (LR) and random forest (RF).

method	n	*mtry*	factor	total	causal	high LD	mod LD	low LD	FP
LR	2000			40	8	15	11	4	0
LR	6000			98	9	16	15	15	0
RF	2000	100000	1	38	7	13	10	2	4
RF	2000	100000	3	28	7	12	6	2	0
RF	2000	100000	5	24	7	10	5	2	0
RF	2000	250000	1	40	8	12	9	2	8
RF	2000	250000	3	25	7	9	5	2	1
RF	2000	250000	5	23	7	9	5	2	0
RF	6000	100000	1	51	9	16	10	5	3
RF	6000	100000	3	41	9	16	6	4	0
RF	6000	100000	5	37	9	16	6	4	0
RF	6000	250000	1	63	9	16	12	6	13
RF	6000	250000	3	42	9	16	7	4	1
RF	6000	250000	5	37	9	16	5	4	0

Table shows method, sample size (n), *mtry* parameter and factor for RF, total number of SNPs, number of SNPs in high ($r^2 > 0.8$), moderate ($0.5 < r^2 \geq 0.8$) and low LD ($0.3 < r^2 \leq 0.5$) LD with any causal SNP as well as number of false-positive SNPs (FP).

3.2 Experimental data

The two experimental GWAS data sets have different results. Figure 3 shows a very strong signal on chromosome 2 for the TRINITY study. Ninety-eight SNPs in this region are genome-wide significant with a minimal p-value of $2.342 * 10^{-29}$ (see Figure 3a). Fourty-two, 35 and 34 SNPs in the same region are selected by the new RF variable selection method using factors of 1, 3 and 5, respectively. They have large minimal relative importance scores with a maximum of 240.13 (see Figure 3b). Supplementary Figure 1 compares P-values and minimal relative importance scores for SNPs that were selected by either method on chromosome 2. P-values are

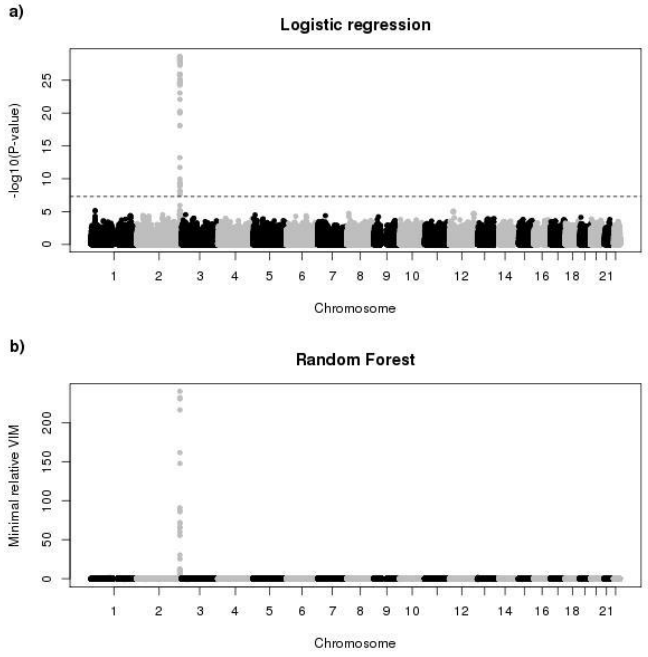

Fig. 3. Manhattan plots for TRINITY data set. a) P-values of logistic regression for each SNP. Dotted line denotes genome-wide significance level of $5 * 10^{-8}$. b) Minimal relative variable importance (VIM) based on RF analysis for each SNP.

very similar for a long region of 100 kb because of strong linkage disequilibrium, whereas only four SNPs at about 234.33 have very large relative importance scores. Two additional SNPs, one on chromosome 1 and the other one on chromosome 13, are selected with a factor of 1. However, if a more stringent factor is used they are not declared as important and p-values of logistic regression are larger than 0.1 for both SNPs. Detailed information about all selected SNPs on chromosome 2 can be found in Supplementary Table 2.

Results for the underpowered AREDS study are summarized in Figure 4. Using logistic regression, no SNP is genome-wide significant and the smallest p-value of $3.011 * 10^{-7}$ is observed for a SNP on chromosome 7 (see Figure 4a). Similarly, the new variable selection method selects none of the SNPs even with the most liberal factor 1 and minimal relative importance scores are much smaller than 1 (see Figure 4b). Again, SNPs on chromosome 7 have the largest minimal importance scores.

4 DISCUSSION

In this work, we presented a new approach for RF to select important variables, i.e. SNPs in GWAS. Evaluations on simulated GWAS data showed that this new method controls the number false-positives and has only slightly less power than the standard approach logistic regression.

Further research is needed to evaluate this promising method in more realistic situations. Since this work was designed as a proof-of-concept study we simulated common SNPs with effects

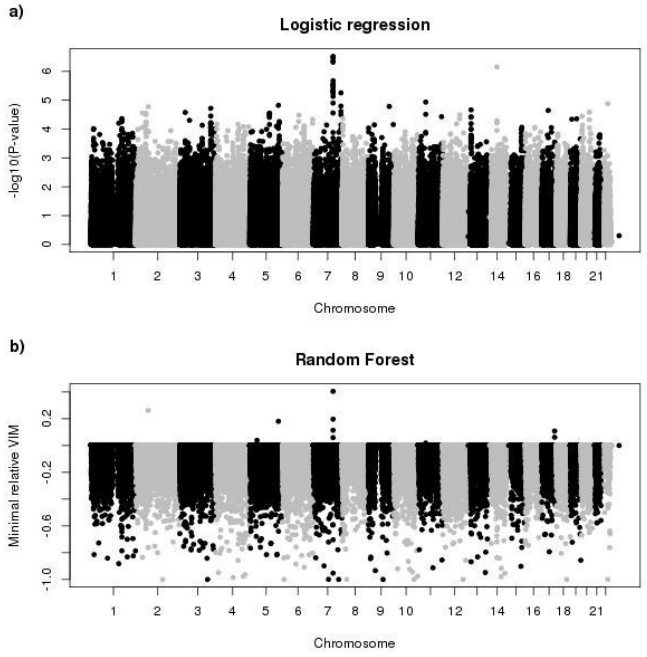

Fig. 4. Manhattan plots for AREDS data set. a) P-values of logistic regression for each SNP. b) Minimal relative variable importance (VIM) based on RF analysis for each SNP.

that are larger than the ones observed in real studies. A power comparison using more realistic effect sizes, however, would require larger sample sizes so that there is a chance to detect the signal. Another limitation of the current study is the very simple alternative hypothesis with case-control status determined by a small number of SNPs interacting independently. We expect RF in combination with the new variable selection procedure to be more powerful than single SNP analyses in more complex scenarios including gene-gene or gene-environment interactions.

Our new variable selection method introduces an additional parameter that determines the threshold in each run. Our simulations show that a fairly stringent parameter is needed to fully control the number of false-positives SNPs that are identified. However, this approach leads to reduced power. Depending on the costs of follow-up analyses and experiments, more liberal thresholds might be preferred in situations where sensitivity is more important.

For each hypothesis and each sample size we only simulated ten replicates to reduce computation time. Simulating one replicate and converting the data into appropriate input formats for PLINK and RandomJungle took approximately 4 hours on the high performance Biowulf Linux cluster at the National Institutes of Health, Bethesda, Md. (http://biowulf.nih.gov). We restricted the size of the trees in each forest, so that a single run of RF was performed in about 8 hours. We checked the effect of the depth parameter by generating trees that were only restricted by node size for some of the replicates with similar or slightly worse results (data not shown).

Similarly, we made several decisions regarding the analysis of the two experimental GWAS data sets for illustration purposes. The first was to dichotomize the provided quantitative traits because our simulation study was focused on case-control studies. Although we were still able to identify the strong signal in the TRINITY data , this approach is usually less powerful and therefore not recommended (Yang *et al.*, 2010). In the AREDS data set, we reduced the number of SNPs by LD pruning. In a real study we would not recommend to remove SNPs, but rather use RF to select the important variables. In some smaller simulation studies (Nicodemus and Malley, 2009; Walters *et al.*, 2012), LD seemed to be a problem in identifying the true causal SNP in regions with moderate and high LD, but in our simulations the causal SNP usually had the highest power. However, additional simulation studies are needed to fully explore the effect of LD in a genome-wide setting because our causal SNPs were not located in regions with very high LD and especially not in very long LD blocks.

RF identified a much smaller region in the TRINITY data compared to the large number of SNPs with similar p-values based on logistic regression. Linkage and association studies have shown that genetic variants in this region influence TBIL levels (Kronenberg *et al.*, 2002; Johnson *et al.*, 2009) and the most promising candidate gene is Uridine diphosphate glucuronosyl transferase 1 family, polypeptideA1 (*UGT1A1*) which is involved in bilirubin metabolism. Interestingly, three out of the four SNPs with the largest median importance scores are located in introns of the gene.

In conclusion, our new variable selection approach is a promising tool for joint analysis of GWAS data that helps to identify interesting regions for follow up studies while limiting the number of false-positives.

ACKNOWLEDGEMENT

This work was supported by the Intramural Research Programs of the National Human Genome Research Institute (NIH), National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIH) and Center for Information Technology (NIH) and utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health (http://biowulf.nih.gov). The authors acknowledge the contributions made by the study participants in the Trinity Student Study (TSS). The TSS GWAS work was supported in part by the Intramural Research Programs of the National Human Genome Research Institute, the Eunice Shriver National Institute of Child Health and Development of the National Institutes of Health (NIH) and the Health Research Board, Dublin, Ireland. Y Kim is now working for Macrogen Inc. CDC is the recipient of an NHGRI Health Disparities Research Fellowship.

REFERENCES

Breiman, L. (2001). Random forests. *Mach Learn*, **45**, 5–32.

Desch, K. C., Ozel, A. B., Siemieniak, D., Kalish, Y., Shavit, J. A., Thornburg, C. D., Sharathkumar, A. A., McHugh, C. P., Laurie, C. C., Crenshaw, A., Mirel, D. B., Kim, Y., Cropp, C. D., Molloy, A. M., Kirke, P. N., Bailey-Wilson, J. E., Wilson, A. F., Mills, J. L., Scott, J. M., Brody, L. C., Li, J. Z., and Ginsburg, D. (2013). Linkage analysis identifies a locus for plasma von willebrand factor undetected by genome-wide association. *Proc Natl Acad Sci U S A*, **110**, 588–593.

Genuer, R., Poggi, J.-M., and Tuleau, C. (2008). Random forests: Some methodological insights. Research Report RR-6729, INRIA.

Goldstein, B. A., Polley, E. C., and Briggs, F. B. S. (2011). Random forests for genetic association studies. *Stat Appl Genet Mol*, **10**, 32.

Hindorff, L., Sethupathy, P., Junkins, H., Ramos, E., Mehta, J., Collins, F., and Manolio, T. (2009). Potential etiologic and functional implications of genome-wide

association loci for human diseases and traits. *Proc Natl Acad Sci U S A*, **106**, 9362–9367.

Jiang, R., Tang, W., Wu, X., and Fu, W. (2009). A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinformatics*, **10**, S65.

Johnson, A. D., Kavousi, M., Smith, A. V., Chen, M.-H., Dehghan, A., Aspelund, T., Lin, J.-P., van Duijn, C. M., Harris, T. B., Cupples, L. A., Uitterlinden, A. G., Launer, L., Hofman, A., Rivadeneira, F., Stricker, B., Yang, Q., O'Donnell, C. J., Gudnason, V., and Witteman, J. C. (2009). Genome-wide association meta-analysis for total serum bilirubin levels. *Hum Mol Genet*, **18**, 2700–2710.

Kim, Y., Wojciechowski, R., Sung, H., Mathias, R. A., Wang, L., Klein, A. P., Lenroot, R. K., Malley, J., and Bailey-Wilson, J. E. (2009). Evaluation of random forests performance for genome-wide association studies in the presence of interaction effects. *BMC Proc*, **3 Suppl 7**, S64.

Kim, Y., Li, Q., Cropp, C. D., Sung, H., Cai, J., Simpson, C. L., Perry, B., Dasgupta, A., Malley, J. D., Wilson, A. F., and Bailey-Wilson, J. E. (2011). Performance of random forests and logic regression methods using mini-exome sequence data. *BMC Proc*, **5 Suppl 9**, S104.

Kronenberg, F., Coon, H., Gutin, A., Abkevich, V., Samuels, M. E., Ballinger, D. G., Hopkins, P. N., and Hunt, S. C. (2002). A genome scan for loci influencing anti-atherogenic serum bilirubin levels. *Eur J Hum Genet*, **10**, 539–546.

Li, C. and Li, M. (2008). GWAsimulator: a rapid whole-genome simulation program. *Bioinformatics*, **24**, 140–142.

Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol*, **34**, 816–834.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.

Nicodemus, K. K. and Malley, J. D. (2009). Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics*, **25**, 1884–1890.

Nicodemus, K. K., Malley, J. D., Strobl, C., and Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, **11**, 110.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, **81**, 559–575.

Schwarz, D. F., König, I. R., and Ziegler, A. (2010). On safari to random jungle: a fast implementation of random forests for high-dimensional data. *Bioinformatics*, **26**, 1752–1758.

Simpson, C. L., Wojciechowski, R., Yee, S. S., Soni, P., Bailey-Wilson, J. E., and Stambolian, D. (2013). Regional replication of association with refractive error on 15q14 and 15q25 in the age-related eye disease study cohort. *Mol Vis*, **19**, 2173–2186.

Stambolian, D., Wojciechowski, R., Oexle, K., Pirastu, M., Li, X., Raffel, L. J., Cotch, M. F., Chew, E. Y., Klein, B., Klein, R., Wong, T. Y., Simpson, C. L., Klaver, C. C. W., van Duijn, C. M., Verhoeven, V. J. M., Baird, P. N., Vitart, V., Paterson, A. D., Mitchell, P., Saw, S. M., Fossarello, M., Kazmierkiewicz, K., Murgia, F., Portas, L., Schache, M., Richardson, A., Xie, J., Wang, J. J., Rochtchina, E., Group, D. C. C. T. D. I. C. R., Viswanathan, A. C., Hayward, C., Wright, A. F., Polasek, O., Campbell, H., Rudan, I., Oostra, B. A., Uitterlinden, A. G., Hofman, A., Rivadeneira, F., Amin, N., Karssen, L. C., Vingerling, J. R., Hosseini, S. M., Dring, A., Bettecken, T., Vatavuk, Z., Gieger, C., Wichmann, H.-E., Wilson, J. F., Fleck, B., Foster, P. J., Topouzis, F., McGuffin, P., Sim, X., Inouye, M., Holliday, E. G., Attia, J., Scott, R. J., Rotter, J. I., Meitinger, T., and Bailey-Wilson, J. E. (2013). Meta-analysis of genome-wide association studies in five cohorts reveals common variants in RBFOX1, a regulator of tissue-specific splicing, associated with refractive error. *Hum Mol Genet*, **22**, 2754–2764.

Strobl, C. and Zeileis, A. (2008). Danger: High power! - exploring the statistical properties of a test for random forest variable importance. Technical Report 017, Department of Statistics, University of Munich.

Strobl, C., Malley, J., and Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests. *Psychol Methods*, **14**, 323–348.

Walters, R., Laurin, C., and Lubke, G. H. (2012). An integrated approach to reduce the impact of minor allele frequency and linkage disequilibrium on variablen importance measures for genome-wide data. *Bioinformatics*, **28**, 2615–2623.

Yang, J., Wray, N. R., and Visscher, P. M. (2010). Comparing apples and oranges: equating the power of case-control and quantitative trait association studies. *Genet Epidemiol*, **34**, 254–257.