VARIABLE SELECTION METHOD FOR THE IDENTIFICATION OF EPISTATIC MODELS

EMILY ROSE HOLZINGER†

Computational and Statistical Genomics Branch (NHGRI, NIH)
Baltimore, MD 21224, USA
Email: emily.holzinger@nih.gov

SILKE SZYMCZAK

Computational and Statistical Genomics Branch (NHGRI, NIH)
Baltimore, MD 21224, USA
Email: s.szymczak@ikmb.uni-kiel.de

ABHIJIT DASGUPTA

Clinical Trials and Outcomes Branch (NIAMS, NIH)
Bethesda, MD 20892-1468, USA
Email: dasgupab@mail.nih.gov

JAMES MALLEY

Center for Information Technology (NIH)
Bethesda, MD 20892-1468, USA
Email: jmalley@mail.nih.gov

QING LI

Computational and Statistical Genomics Branch (NHGRI, NIH)
Baltimore, MD 21224, USA
Email: liq4@mail.nih.gov

JOAN E. BAILEY-WILSON

Computational and Statistical Genomics Branch (NHGRI, NIH)
Baltimore, MD 21224, USA
Email: jebw@mail.nih.gov

---

† This work was supported in part by the Intramural Research Programs of the National Human Genome Research Institute, the National Institute for Arthritis, Musculoskeletal and Skin Disorders, and the Center for Information Technology, all part of the National Institutes of Health.

Standard analysis methods for genome wide association studies (GWAS) are not robust to complex disease models, such as interactions between variables with small main effects. These types of effects likely contribute to the heritability of complex human traits. Machine learning methods that are capable of identifying interactions, such as Random Forests (RF), are an alternative analysis approach. One caveat to RF is that there is no standardized method of selecting variables so that false positives are reduced while retaining adequate power. To this end, we have developed a novel variable selection method called *relative recurrency variable importance metric* (r2VIM). This method incorporates recurrency and variance estimation to assist in optimal threshold selection. For this study, we specifically address how this method performs in data with almost completely epistatic effects (i.e. no marginal effects). Our results show that with appropriate parameter settings, r2VIM can identify interaction effects when the marginal effects are virtually nonexistent. It also outperforms logistic regression, which has essentially no power under this type of model when the number of potential features (genetic variants) is large. (All Supplementary Data can be found here: http://research.nhgri.nih.gov/manuscripts/Bailey-Wilson/r2VIM_epi/).

1. Introduction

1.1. *Variable selection that allows for interactions*

Thousands of variants have been identified that are associated with complex human traits [1]. However, a large portion of the estimated heritability remains unexplained for many traits [2]. Additionally, these variants often do not improve prediction of complex traits in independent data sets over metrics that are relatively easier to collect (e.g. age, sex, body mass index, family history)[3]. This is likely due, in part, to overly simplistic study designs and modeling methods. The complex nature of biological pathways makes it unlikely that additive main effects explain all of the heritability. Empirical observations in animal model studies show that complex effects are actually pervasive in nature [4]. The identification of these effects would require variant discovery and modeling methods that are robust to interactions, even when main effects are very small or non-existent.

The first step in solving this problem is to separate true signal from noise. Machine learning methods are promising candidates for this task and are currently used in other scientific fields, including drug design [5]. One type of machine learning method is Random Forests (RF) [6]. One limitation of RF is that no standard method exists for selecting a set of "associated" variants with low levels of false positives and adequate power. Parametric analyses produce statistics with generally accepted values for error rates, assuming the parametric model has been exactly and correctly specified. One way to obtain equivalent values is to generate empirical distributions by running thousands of permutation analyses. This is computationally impractical for studies that use high-throughput genomic data, which usually consists of thousands to millions of variables. We propose a more efficient method called r2VIM, which integrates different selection parameters to identify the appropriate threshold between signal and noise [7]. The ultimate goal of this

method is to generate variant sets that include main and interaction effects. These sets can then be assessed using modeling tools for interpretation and prediction purposes to further our understanding of complex human traits.

2. Methods

2.1. *r2VIM*

The method r2VIM uses a novel variable selection algorithm based on RF results. RF generates a collection of regression (quantitative outcome) or classification (categorical outcome) trees. In RF, bootstrap samples are drawn to train tree models, and the performance of the trained model is evaluated by testing the tree on the "out-of-bag" (OOB) sample, i.e., the observations not included in the bootstrap sample used for training. This process is repeated over many bootstrap samples and the optimal RF is based on evaluating performance across all the OOB samples. This process reduces the likelihood of overfitting as the model is optimized based on OOB data and not the training data [8]. The VIM is calculated as the difference of an error metric before and after random variable permutation. Variables that result in greater error due to permutation have higher VIMs and are considered more important for prediction purposes. While methods do exist for interpreting the VIM, there is no gold standard method for determining the threshold that best differentiates between noise and functional variables. The random nature of the algorithm can result in variables with high VIMs in one run and low VIMs in another with only a different random seed. To address this, we combine recurrency and a threshold optimization procedure, as described below and illustrated in Supp. Figure 1:

1. *Permutation-based importance score:* Unscaled permutation is used based on previous studies that found this VIM estimation method to be the most reliable [9].

2. *Estimate of null variance*: Assuming that the smallest VIM is negative, the absolute value of the difference between the smallest VIM and zero should be a reliable estimate of VIM variance in null data, as variables with no effect should be symmetrically and randomly distributed around zero [10]. Variables with VIMs greater than the estimated null variance are less likely to be noise variables. In preliminary analyses, we observed that this estimate may be too conservative or liberal for data with different effect types. Therefore, we use the distribution of VIMs to guide threshold selection for this analysis [7].

3. *Recurrency*: Due to the randomness of RF, variables that are deemed important in one run may be declared not important in a second run having only a different random number seed. Variables with relatively high VIMs across many runs are more likely to be true signals. The reasoning here is that stronger predictors will have a higher probability

of being in a top VIM list, and this rate of *recurrency* for a predictor is an estimate of this probability. For this analysis, we run RF five times for each of the 100 simulated datasets to assess false positive and true positive selection at various thresholds. We calculate relative importance score (RIS), which is the VIM divided by the variance estimate. This allows us to compare VIMs across the five runs, and it allows us to select a more appropriate threshold based on the RIS distribution that results from the five runs. We use the median or minimum of the RIS values from the five runs as a "recurrency-corrected" metric. For summarization purposes, we report the median value of this metric for all 100 datasets from each simulation model, unless stated otherwise

2.2. *Data Simulation*

A previous study has assessed r2VIM using simulated data with only main effects [7]. r2VIM performed comparably to linear regression in terms of power and false positive rate. Our study specifically addresses the performance of r2VIM in the presence of interactions with no main effects.

We simulated data sets using genomeSIMLA [11], [12]. We simulated four different types of models, which had either 100 or 1000 total SNPs, with and without correlation between the SNPs (i.e. linkage disequilibrium, or LD). For each model, 100 data set replicates were generated. The genetic effect for the four models consists of two SNPs with an interaction effect and no marginal effects. This genetic effect was used to generate data sets with a penetrance table. This table provides the probability of being a case for each of the nine genotype combinations. The model was generated using the simpen algorithm in genomeSIMLA. The minor allele frequency for both SNPs is 0.4. The target heritability and odds ratio of the effect model are 0.10 and 2.0, respectively. The marginal effects for the genotypes at each locus are all very close to 0. The penetrance values for the genetic models are shown in Supp. Table 1. The outcome is binary (case/control status) with 250 cases/250 controls in the 100 SNP data and 500 cases/500 controls in the 1000 SNP data. Datasets that included LD were generated using forward time population

Table 1. RF parameters for each of the simulated data analyses

Model	RF Parameter Values
100 SNPs (no LD and LD)	mtry = 20, 40 ntree = 200, 600 nodesize = 10
1000 SNPs (no LD and LD)	mtry = 300, 400 ntree = 2000, 6000 nodesize = 100

simulation, as previously described [11]. LD models were selected that had moderate correlation overall, but virtually no correlation between the two functional SNPs. LD plots showing these correlation patterns are shown in Supp. Figure 2.

3. Results

3.1. *Simulated Data*

We ran r2VIM on the four simulation models specifying five runs of RF per simulated dataset and obtaining the RIS scores for each run and the median and minimum RIS for all five runs for each dataset. We report the median value of the median or minimum RIS for the 100 datasets. We also calculated the detection power (or rate), which is defined as the number of times in the 100 datasets that the median RIS or minimum RIS exceeded a set threshold. We ran RF with different combinations of variable subset sizes (mtry) and number of trees (ntree), as these have the largest effect on performance. The minimum number of samples allowed in a terminal node, called terminal node size (nodesize), is also important, and

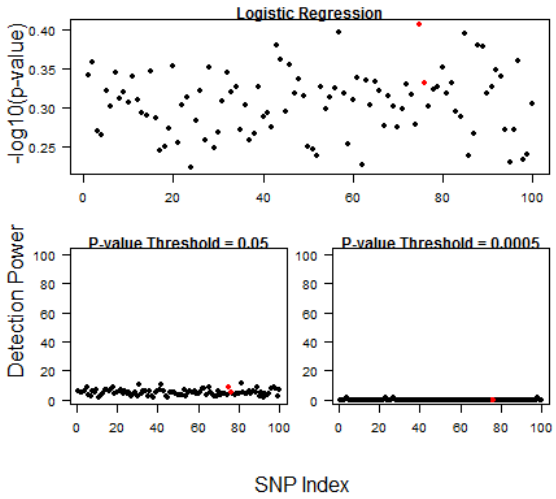

Fig. 1. Results for univariate logistic regression analysis of 100 SNP data with no LD. The top plot shows the median $-\log_{10}$ (p-value) for each SNP across 100 dataset replicates. The bottom two plots show the number of times out of 100 datasets that the p-value was smaller than the specified significance threshold for each SNP. Functional SNPs 75 and 76 are shown in red.

varied across models according to sample size. Table 1 shows the RF parameter values applied. Other parameter settings were consistent across runs. We ran univariate logistic regression for comparison.

Figure 1 shows the results for the logistic regression analysis on the 100 datasets for the 100 SNP data with no LD. We report the median p-value for each SNP across all datasets. As expected, the lack of marginal effects in the simulated model results in virtually no power, even at very liberal selection thresholds. Even if an exhaustive search for all possible two-way interactions was performed, the multiple testing correction would hinder the identification of most true models. Moreover, an ideal analysis would recode the SNPs genotypically, which doubles the number of variables and makes the correction even more stringent.

The results for the 100 SNP data with no LD for the r2VIM analysis are shown in Figure 2. The r2VIM analyses took ~13 seconds per dataset to complete. The median RIS for the analysis with mtry=40 and ntree=600 is shown. Results for all parameter settings can be found in Supp. Figures 3-8. The functional SNPs (75 and 76) are highlighted in red. Note that the positions here are not relevant as all of the SNPs were simulated to be independent. Again, RIS is calculated as raw VIM / variance estimate for that dataset. This allows us to compare importance scores across datasets.

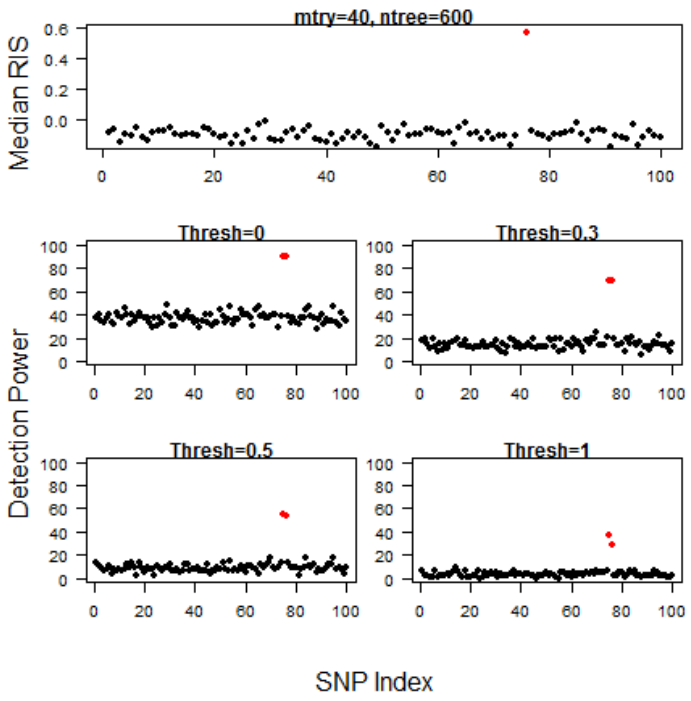

Fig. 2. Results for r2VIM analysis of datasets with 100 SNPs and no LD for mtry=40 and ntree=600. The top plot shows the median RIS for each SNP across 100 dataset replicates. The bottom plots show the number of times out of 100 replicates that the median RIS was smaller than the specified significance threshold for each SNP. Functional SNPs 75 and 76 are shown in red.

We show the number of times each variable was selected at four different selection threshold levels. The thresholds (0, 0.3, 0.5, and 1) were chosen based on the distribution of median RIS values. The results suggest that the median of RIS values produces less false positives than the minimum RIS in this data (Supp. Figures 3, 4, 6, and 7). For this genetic model, a factor threshold of 0.3 optimizes both detection power and false positive rate when applied to the median RIS.

Next, we assessed the effect of LD on r2VIM performance. LD is an important characteristic of genetic data that can have a large effect on power and false positive rate [13]. Figure 3 shows the distribution of median RIS values for all 100 datasets with moderate LD. The functional SNP positions for these simulations (4 and 26) are relevant, as they were selected to be nearly uncorrelated. The detection power thresholds are higher here, as the RIS scores were much higher than those for the data with no LD.

Interestingly, LD increases detection power and the RIS scale for the functional and non-functional SNPs. VIM inflation with variable correlation has been observed before in RF [9]. This feature should be considered when performing SNP filtering based on pairwise LD measures. Of note, LD results in higher detection rates for functional *and* non-functional SNPs due to inflated RIS values and could result in more false positives. However, for this small number of SNPs, many of the nearby non-functional SNPs are in high LD with at least one of the functional SNPs and are not true "false positives" but instead are the result of RF detecting association of a "chromosomal region" with the trait.

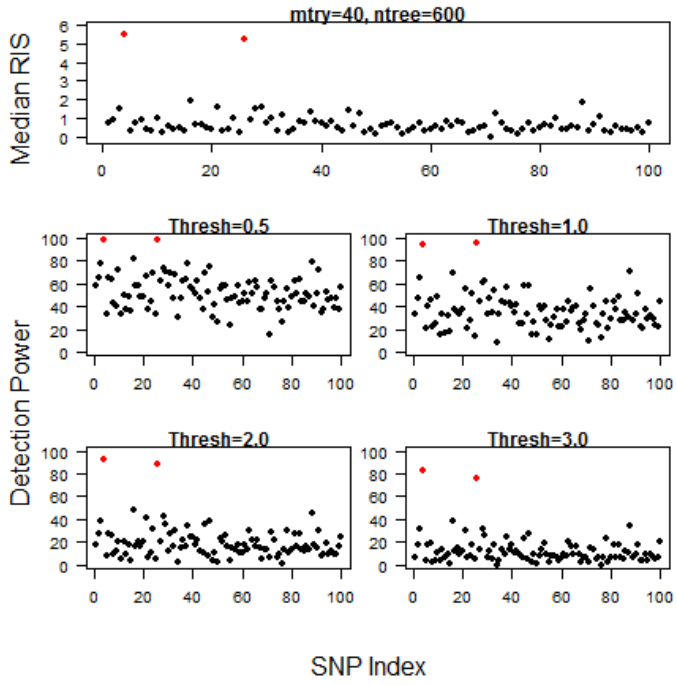

Fig. 3. Results for r2VIM analysis of datasets with 100 SNPs and LD for mtry=40 and ntree=600. The top plot shows the median RIS for each SNP across 100 dataset replicates. The bottom plots show the number of times out of 100 replicates that the median RIS was smaller than the specified significance threshold for each SNP. Functional SNPs 4 and 26 are shown in red.

To determine the effect of added noise, we simulated data sets with 1000 SNPs (998 non-functional and 2 functional). Each analysis took ~40 seconds per dataset to complete. This is still not near the scale of a typical GWAS analysis; however, it is impractical to run r2VIM on 100 GWAS-sized datasets many times. Figures 4 and 5 show the results for the 1000 SNP analyses without and with LD, respectively. When no LD was present, detection power was lower than the 100 SNP data for the RIS thresholds shown; however, the median RIS values still differentiate between the non-functional and functional SNPs. With LD, we observe the same increase in RIS values and detection power for all SNPs. This is even more pronounced in the 1000 SNP data.

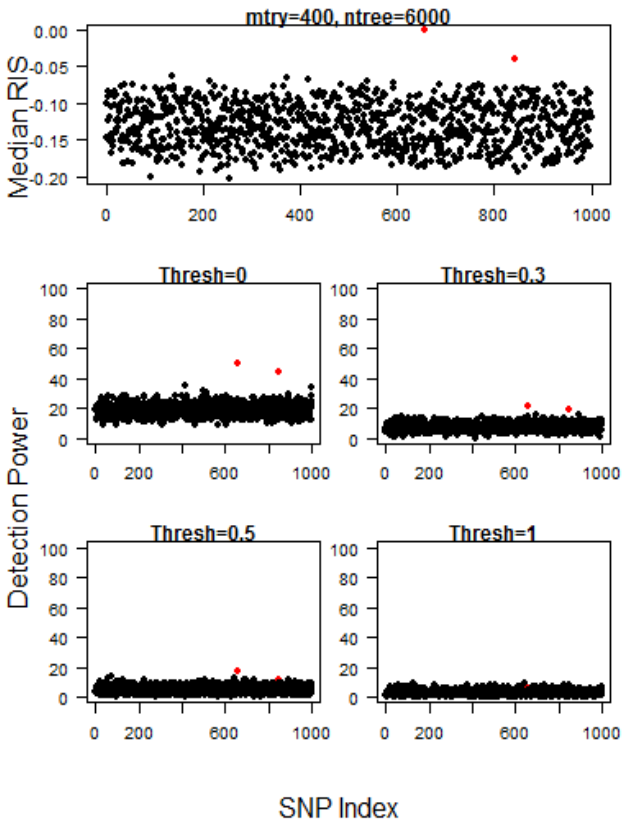

Fig. 4. Results for r2VIM analysis of datasets with 1000 SNPs and no LD for mtry=400 and ntree=6,000. The top plot shows the median RIS for each SNP across 100 dataset replicates. The bottom plots show the number of times out of 100 replicates that the median RIS was smaller than the specified significance threshold for each SNP. Functional SNPs 657 and 844 are shown in red.

Also, the detection power was higher with a larger ntree (Supp. Figures 9-12). This emphasizes the importance of using the correct parameter settings and selection criteria for data with a high noise to signal ratio.

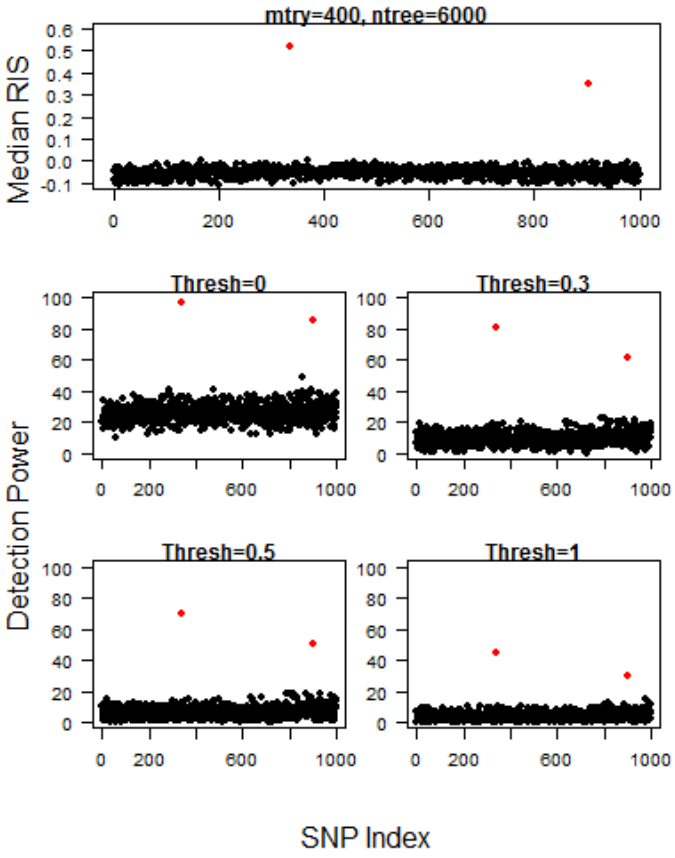

Fig. 5. Results for r2VIM analysis of datasets with 1000 SNPs and LD for mtry=400 and ntree=6,000. The top plot shows the median RIS for each SNP across 100 dataset replicates. The bottom plots show the number of times out of 100 replicates that the median RIS was smaller than the specified significance threshold for each SNP. The functional SNPs 336 and 903 are shown in red.

Recurrency requires more computational resources than a single-run RF analysis, so we assessed the level of performance gain due to recurrency by comparing false positive detection at the RIS threshold of 0.3 for all models (Figure 6). To compare 100 SNP data with 1000 SNP data we divided the number of FPs selected by the total variable count. For the five single runs and the recurrency-corrected runs (median and minimum), both functional SNPs were identified at this threshold. For a single data set, the minimum RIS value reduces false positive selection over all other RIS values. This is in contrast to the summary data analyses where the median RIS value appeared to be optimal (Supp. Figures 2-11). This could be a factor of the summarization itself, as this essentially adds another layer of recurrency to the results. In applied analyses, it will be important to plot both the minimum and median RIS values to assess the distributions of each.

4. Discussion

One of the biggest hurdles in performing a successful analysis of high-throughput data is selecting variables least likely to be noise. The most commonly used methods thus far often

take the results from a univariate analysis to identify predictor variables with some level of marginal effects. This would not be appropriate if interaction effects exist with little or no marginal effects. Our method addresses this by performing relatively fast variable selection that can identify interaction effects when marginal effects are virtually nonexistent.

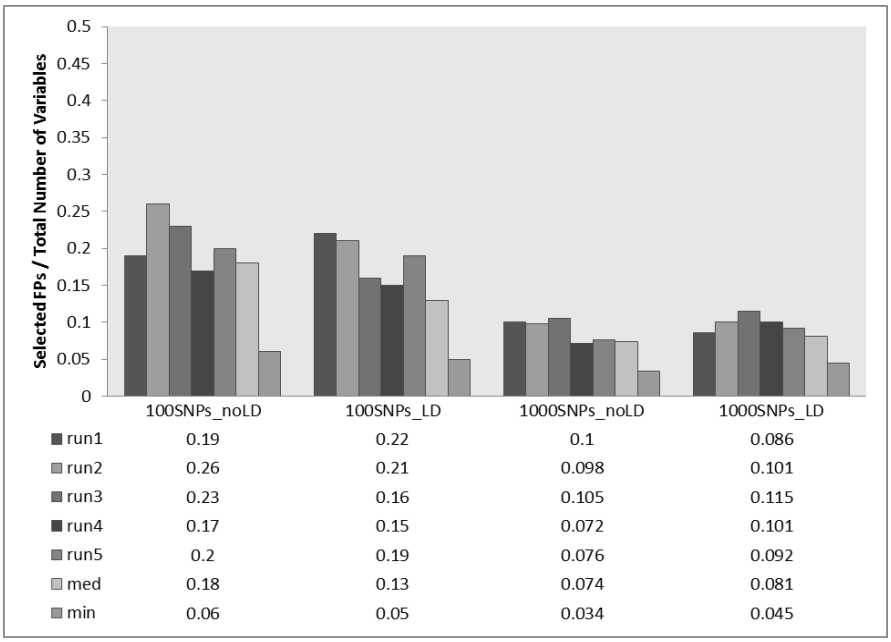

Fig. 6. False positives selected at an RIS threshold of 0.3 for a single dataset for each of the simulation models. The number shown is the number of FPs selected divided by the total number of variables in dataset for comparison purposes.

Importantly, there are limitations to this method. For example, RF does not allow for any missing data. Missingness is very common in high-throughput data; therefore procedures such as imputation or complete removal of variables with missing data are required to run RF. Additionally, this paper only tests data sets with a binary (case/control) outcome using classification trees. A quantitative outcome would require regression trees, and more testing needs to be done to determine how this affects performance. Although this method supplies a more regimented procedure for RF threshold selection, it is still highly dependent on many factors. For example, the minimum RIS was optimal for certain analyses, while the median was preferable for others. Currently, RF variable selection requires exploring various aspects of the results to determine the best selection method. Finally, sequence data is becoming increasingly available for research. To this end, it will be necessary to optimize r2VIM for rare variant detection, as well as quantitative high-throughput data, such as RNA sequence levels.

For a fair comparison to other methods, tools designed for interaction identification should be assessed on the same data. However, this is not a trivial task. An exhaustive interaction analysis

using a regression model is also underpowered to identify highly epistatic models if the SNPs are coded in an allelic, or additive, manner as they were for the RF analysis. To quickly illustrate this, we tested the correct SNP pair and nine random SNP pairs using logistic regression (y=snp1 + snp2 + snp1*snp2) in a 100 SNP / no LD simulated dataset. The true model interaction term had a p-value of 0.24, which is not borderline significant even before multiple testing correction. Additionally, three of the nine random SNP pair analyses had interaction p-values lower than the correct model. The logistic regression model would have more power to find the interactions if they are coded in a genotypic manner due to the simulation design. However, this encoding would require doubling the number of variables that need to be exhaustively tested. RF, on the other hand, often identifies the model without re-coding the SNPs. Further tests will involve recoding the SNPs genotypically and testing in RF, logistic regression, and other methods.

The impact of adding variables with main effects to the interaction model must also be assessed, as biological data is likely to include many different types of effects. Future work will involve simulating different effect types to assess the impact on RIS distribution and detection power.

After variable selection, the next step is to model the subset for interpretation and prediction purposes. This requires a method robust to interaction and marginal effects. Machine learning methods are also an attractive candidate for this step [14], [15]. It will be important to recode the data so that genotypic effects can be seen, especially for possible interactions. This could also be done at the selection step; however, as it doubles the number of variables in the dataset, it is not an ideal procedure in data with already high levels of noise. Fortunately, r2VIM appears to be able to identify non-additive interaction effects even with the standard additive encoding. After selection, however, it is more computationally feasible to recode the subset of candidate SNPs to generate more informative prediction models. It is also useful to note that the two methods proposed here (noise detection by removal of SNPs with negative variable importance measures and recurrency), could be used with many other machine learning schemes, such as neural nets, boosting and support vector machines.

The ultimate goal of r2VIM is to provide a tool that can perform powerful selection while taking into account main and interaction effects. Non-linear interactions are especially difficult to identify unless specific tools robust to these effects are used. Our results suggest that using proper threshold selection procedures, RF can identify these types of effects even in the extreme situation of virtually no marginal effects.

References

[1] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, and H. Parkinson, "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D1001–D1006, Dec. 2013.

[2] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis,

C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, Oct. 2009.

[3] B. Godman, A. E. Finlayson, P. K. Cheema, E. Zebedin-Brandl, I. Gutiérrez-Ibarluzea, J. Jones, R. E. Malmström, E. Asola, C. Baumgärtel, M. Bennie, I. Bishop, A. Bucsics, S. Campbell, E. Diogene, A. Ferrario, J. Fürst, K. Garuoliene, M. Gomes, K. Harris, A. Haycox, H. Herholz, K. Hviding, S. Jan, M. Kalaba, C. Kvalheim, O. Laius, S.-A. Lööv, K. Malinowska, A. Martin, L. McCullagh, F. Nilsson, K. Paterson, U. Schwabe, G. Selke, C. Sermet, S. Simoens, D. Tomek, V. Vlahovic-Palcevski, L. Voncina, M. Wladysiuk, M. van Woerkom, D. Wong-Rieger, C. Zara, R. Ali, and L. L. Gustafsson, "Personalizing health care: feasibility and future implications," *BMC Med.*, vol. 11, p. 179, 2013.

[4] W. Huang, S. Richards, M. A. Carbone, D. Zhu, R. R. H. Anholt, J. F. Ayroles, L. Duncan, K. W. Jordan, F. Lawrence, M. M. Magwire, C. B. Warner, K. Blankenburg, Y. Han, M. Javaid, J. Jayaseelan, S. N. Jhangiani, D. Muzny, F. Ongeri, L. Perales, Y.-Q. Wu, Y. Zhang, X. Zou, E. A. Stone, R. A. Gibbs, and T. F. C. Mackay, "Epistasis dominates the genetic architecture of Drosophila quantitative traits," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 39, pp. 15553–15559, Sep. 2012.

[5] J. C. Gertrudes, V. G. Maltarollo, R. A. Silva, P. R. Oliveira, K. M. Honório, and A. B. F. da Silva, "Machine learning techniques and drug design," *Curr. Med. Chem.*, vol. 19, no. 25, pp. 4289–4297, 2012.

[6] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random Forests," in *Ensemble Machine Learning*, C. Zhang and Y. Ma, Eds. Boston, MA: Springer US, 2012, pp. 157–175.

[7] S. Szymczak, E. Holzinger, A. Dasgupta, J. Malley, and J. Bailey-Wilson, "A new variable selection method for random forests in genome-wide association studes," *Bioinformatics*, In Preparation.

[8] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[9] K. K. Nicodemus, J. D. Malley, C. Strobl, and A. Ziegler, "The behaviour of random forest permutation-based variable importance measures under predictor correlation," *BMC Bioinformatics*, vol. 11, no. 1, p. 110, 2010.

[10] C. Strobl, J. Malley, and G. Tutz, "An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests.," *Psychol. Methods*, vol. 14, no. 4, pp. 323–348, 2009.

[11] T. L. Edwards, W. S. Bush, S. D. Turner, S. M. Dudek, E. S. Torstenson, M. Schmidt, E. Martin, and M. D. Ritchie, "Generating Linkage Disequilibrium Patterns in Data Simulations Using genomeSIMLA," *Lect. Notes Comput. Sci.*, vol. 4793, pp. 24–35, 2008.

[12] S. M. Dudek, A. A. Motsinger, D. R. Velez, S. M. Williams, and M. D. Ritchie, "Data simulation software for whole-genome association and other studies in human genetics," *Pac.Symp.Biocomput.*, vol. 11, pp. 499–510, 2006.

[13] S. A. Tishkoff and B. C. Verrelli, "Role of evolutionary history on haplotype block structure in the human genome: implications for disease mapping," *Curr.Opin.Genet.Dev.*, vol. 13, no. 6, pp. 569–575, Dec. 2003.

[14] E. Holzinger, S. M. Dudek, A. T. Frase, R. M. Krauss, M. W. Medina, and M. D. Ritchie, "ATHENA: A Tool for Meta-Dimensional Analysis Applied to Genotypes and Gene Expression Data to Predict HDL Cholesterol Levels," presented at the Pacific Symposium on Biocomputing, 2013, vol. 18, pp. 385–396.

[15] A. Dasgupta, S. Szymczak, J. H. Moore, J. E. Bailey-Wilson, and J. D. Malley, "Risk estimation using probability machines," *BioData Min.*, vol. 7, no. 1, p. 2, 2014.