Exploiting the Structure of Random Forest for the Detection of Epistatic Interactions

Corinna Schmalohr[1,2,*], Jan Grossbach[2,*], Andreas Beyer[1,2], Mathieu Clément-Ziza[1,2]

[1] Center for Molecular Medicine Cologne (CMMC), University of Cologne, Cologne, Germany
[2] Cologne Cluster of Excellence in Cellular Stress Responses in Aging-associated Diseases (CECAD), University of Cologne, Cologne, Germany
[*] These authors contributed equally to this work

Summary. *Epistasis (non-additive genetic interaction) is one possible cause for the 'missing heritability' that is observed for many complex traits. However, there is a lack of sensitive methods for the detection of epistasis. We propose four approaches for the detection of epistasis, which are based on the machine learning algorithm Random Forest: the split asymmetry, the selection asymmetry, the paired selection frequency, and an ensemble method that combines the three approaches. We assess the performance of these methods on simulated and real data, comparing them to the commonly used exhaustive pair-wise ANOVA approach. Our scores perform better than ANOVA on both simulated and real data and we discuss possible reasons for the performance differences. This work contributes to the long-standing problem of extracting information about the underlying model from a Random Forest.*

Introduction

It is a central objective in biology to identify genomic variations that determine complex biological phenotypes and quantitative traits. However, the cumulative contribution of individual genetic variants detected by genome-wide association studies (GWAS) can only partially explain the phenotypic variance due to genetic effects – a phenomenon termed ‘missing heritability’. Epistasis describes non-additive interactions between markers, i.e. situations where the contribution of two or more genetic variants on a quantitative trait differs from the sum of their marginal effects. Epistasis suggests functional relationships between genes, and has been proposed as one possible factor underlying missing heritability (reviewed in [1]). Commonly, two types of epistasis are distinguished: (i) AND-epistasis, where an allele at one locus enhances or alleviates the effect of another locus, and (ii) XOR-epistasis, where the effects of alleles at two loci are diminished when they occur together [2]. However, detecting epistatic interactions in quantitative trait studies remains challenging because of the combinatorial number of hypotheses to test and due to insufficient statistical power [3]. A wide variety of methods to detect epistasis has been developed, yet most methods are not applicable on a genome-wide scale, and/or make assumptions about the distribution of the data and the scale and order of interactions, which may not apply to real data (reviewed in [3]).

We and others have shown that Random Forest (RF) is a very efficient method for detecting genotype-phenotype relationships especially in the presence of epistasis, because it implicitly accounts for non-additive effects [4]–[8]. RF consist of an ensemble of classification and regression trees (CART), which are trained on bootstraps and random samples of the data, leading to a remarkable robustness against overfitting [9]. A notorious problem of RF remains that it is difficult to extract information about relationships between features from the forest. That is, however, necessary for detecting epistatic interactions between markers (i.e. features). Here we present four new approaches that exploit the structure of RF to detect different types of epistatic interactions, the *split asymmetry* test, the *selection asymmetry* test, the *paired selection frequency (paired SF)* test,

and an ensemble method. These methods for detecting non-additive feature interactions are applicable beyond the genetic mapping problem.

Methods

Paired Selection Frequency. This approach is based on the expectation that interacting markers are more likely to be selected in the same tree than non-interacting markers. The number of times two markers were selected in the same tree in the RF are compared to the number of times the markers were selected independently of each other. These counts are used to build a contingency table and a one-sided Fisher's exact test is applied to detect co-dependence between the markers.

Split Asymmetry. Given two markers A and B in the same path of a CART (A before B), the difference in mean phenotype observed after a split on marker B depends on the result of the partitioning on marker A (Figure 1, green lines). We use a Student's *t*-test to check for this imbalance in phenotype differences for all marker pairs. Each marker pair is tested twice in two independent tests. Once for the cases where marker A was used first, and also for the cases where marker B was used first. The two *p*-values are then combined using the Fisher method [10].

Selection Asymmetry. When two markers A and B interact through AND-epistasis, there is an imbalance in the frequency of splits using B after A (Figure 1, red dashed arrow). We exploit this property to detect epistasis by testing for this interdependence in the marker selection frequency with a binomial test of equal probabilities. Again, we combine the *p*-values for the two possible pairs (AB and BA) using the Fisher method.

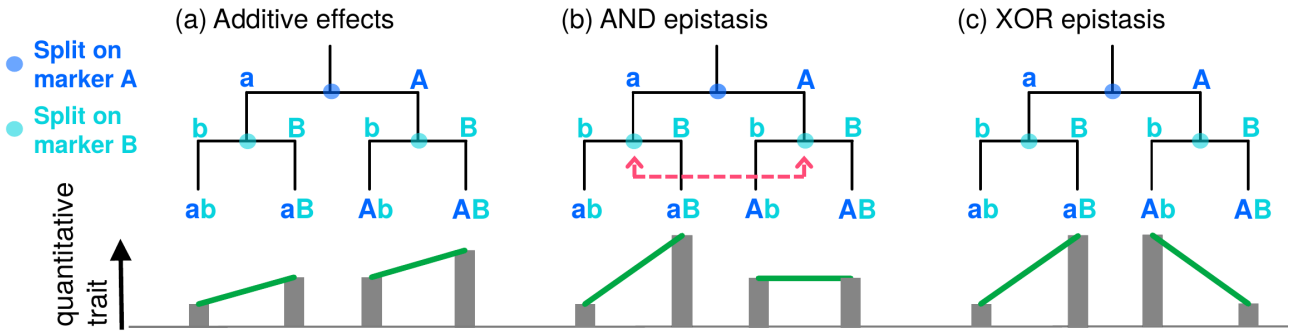

Figure 1: Schematic representation of the detection of epistasis from Random Forest. Shown are example subtrees that depict the splitting of data on two markers A and B that **(a)** do not interact, **(b)** are in AND-type epistasis, or **(c)** are in XOR-type epistasis. The latter two lead to asymmetries in the trait value distribution (indicated by the green lines), which are exploited in the *split asymmetry* approach. In addition, there are unequal probabilities for the selection of marker B for the two partitions created by the split on A (indicated by the red dashed arrow), which is tested for in the *selection asymmetry* approach.

Ensemble Method. The p-values generated by the *Paired SF*, *Split asymmetry* and *Selection asymmetry* approaches were combined using the Fisher method to create an ensemble score.

Simulated Data. Traits were simulated based on genotypes from the widely used *Saccaromyces cerevisiae* BYxRM cross [11]. Different combinations of marginal and epistatic effects with varying effect sizes, different types and orders of epistasis and varying noise levels were simulated 32 times each.

Benchmark on Real Data. In order to evaluate the usability of our method for real data, we applied our methods to an expression QTL (eQTL) dataset of 112 segregant strains from a *Saccaromyces cerevisiae* cross (RMxBY, data unpublished). This dataset encompasses genotype information for

3,593 markers and RNA-seq-based expression data for 1,050 transcripts that correspond to essential genes (i.e. genes that are lethal when knocked-out). Null distributions for the area under receiver operating characteristic (AUROC) and the area under the precision-recall curve (AUPR) were generated from permutations of the double knock-out reference data and were used to compute empirical *p*-values, which were then corrected using the Bonferroni method.

Results and Discussion

Benchmark on Simulated Data. The four RF-based methods were benchmarked on simulated data and compared to an exhaustive pair-wise ANOVA. The ensemble method recovered most simulated interactions (Figure 2). In general, the RF-based methods were more performant when at least one of the interacting markers had a marginal effect (Figure 2b), which is likely to be the case in a real biological setting. However, this represents one of the limitations of RF: the modeling relies on the presence of marginal effects. Accordingly, RF-based methods were not able to recover XOR-epistasis in the absence of marginal effects of the interacting markers (data not shown). Yet, the biological relevance of XOR epistasis without marginal effects remains questionable.

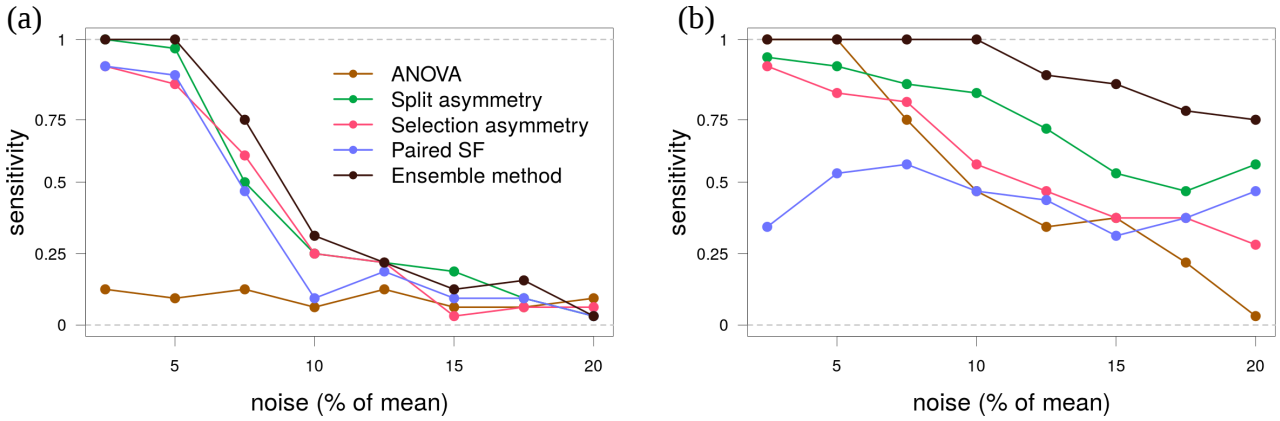

Figure 2: Sensitivity of methods based on representative simulation scenarios. Here, sensitivity is the proportion of simulated interactions that were recovered. An interaction was regarded as recovered if its *p*-value was below the lower 0.5-percentile of *p*-values (i.e. 99.5% p-values were higher). **(a)** Simulation scenario with AND-epistasis between two markers and marginal effects of two unrelated markers. **(b)** Simulation scenario with AND-epistasis between two markers, one of which has an additional marginal effect.

Benchmark on Real Data. We assessed the biological relevance of the proposed methods by applying them to an eQTL dataset in comparison to an exhaustive pair-wise ANOVA approach. The performance was measured by the ability to recover epistatic interactions detected in double knock-out experiments [12], as previously proposed [6]. All RF-based approaches outperformed the ANOVA (Figure 3) based on the AUROC and the AUPR. In contrast to the simulation results, the *split asymmetry* outperformed the ensemble method. AUROC and AUPR were low for all tested methods, although significantly above random (p-value$<2*10^{-4}$ for all methods). A perfect performance (i.e. AUROC=1.0) is impossible in this benchmark, because the reference data (the double knock-out study) measures a different phenotype (growth *versus* expression), and because it entails a different type of genetic perturbation (gene knock-outs *versus* segregating genetic variants). Thus, the reference data can in this case only be used for a relative comparison of different methods, but not for an evaluation of their absolute performance.

Conclusions

Here we propose methods that exploit RF for the detection of epistasis. These methods outperform the exhaustive pair-wise tests in simulated and real data. Since Random Forest makes no assumptions about the model complexity it is – in principle – possible to extend this approach from two-way interactions to higher order interactions. Our scores for detecting non-additive feature interactions in RF are applicable beyond genetic mapping where the model structure of a RF needs to be analysed.

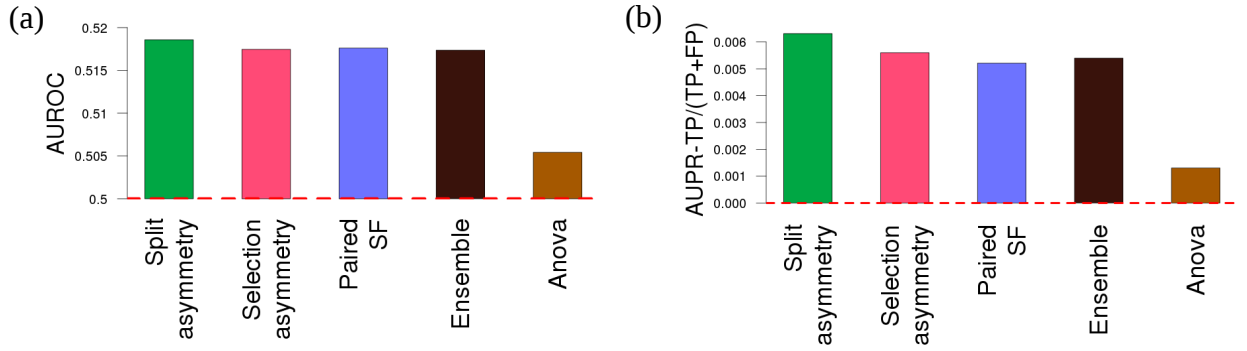

Figure 3: Performance on real data. Performance was evaluated by the ability of the evaluated methods to correctly classify interacting and non-interacting genes, using double knock-out growth data as a gold standard. Shown are the differences between the results of the tested methods and random assignment (red dashed line) for **(a)** AUROC and **(b)** AUPR. The RF-based methods outperform the ANOVA, while the *split asymmetry* approach has the best recovery and precision.

References

[1] E. E. Eichler *et al.*, "Missing heritability and strategies for finding the underlying causes of complex disease," *Nat. Rev. Genet.*, vol. 11, no. 6, pp. 446–450, 2010.

[2] Ö. Carlborg and C. S. Haley, "Epistasis: too often neglected in complex trait studies?," *Nat. Rev. Genet.*, vol. 5, no. 8, pp. 618–625, 2004.

[3] C. Niel, C. Sinoquet, C. Dina, and G. Rocheleau, "A survey about methods dedicated to epistasis detection," *Front. Genet.*, vol. 6, 2015.

[4] J. J. Michaelson, R. Alberts, K. Schughart, and A. Beyer, "Data-driven assessment of eQTL mapping methods," *BMC Genomics*, vol. 11, no. 1, p. 502, 2010.

[5] M. Clement-Ziza *et al.*, "Natural genetic variation impacts expression levels of coding, non-coding, and antisense transcripts in fission yeast," *Mol. Syst. Biol.*, vol. 10, no. 11, pp. 764–764, 2014.

[6] P. Picotti *et al.*, "A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis," *Nature*, vol. 494, no. 7436, pp. 266–270, 2013.

[7] M. N. Wright, A. Ziegler, and I. R. König, "Do little interactions get lost in dark random forests?," *BMC Bioinformatics*, vol. 17, p. 145, 2016.

[8] J. Stephan, O. Stegle, and A. Beyer, "A random forest approach to capture genetic effects in the presence of population structure," *Nat. Commun.*, vol. 6, p. 7432, 2015.

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

[10] R. A. Fisher, *Statistical methods for research workers*. Genesis Publishing Pvt Ltd, 1925.

[11] R. B. Brem and L. Kruglyak, "The landscape of genetic complexity across 5,700 gene expression traits in yeast," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 5, pp. 1572–1577, 2005.

[12] M. Costanzo *et al.*, "A global genetic interaction network maps a wiring diagram of cellular function," *Science*, vol. 353, no. 6306, 2016.