Evaluation of Methods for the Detection of Epistasis Using Random Forest

Corinna Lewis Schmalohr

November 16, 2017

A thesis submitted in partial satisfaction of the
requirements for the degree of

Master of Science

in

Biological Sciences

of the

University of Cologne, Germany

Supervisors:
Dr. Mathieu Clément-Ziza
Prof. Andreas Beyer

Abstract

A central objective in biological research is the elucidation of the influence of genetic factors on phenotypes. Non-additive genetic interactions, also called epistasis, are likely to play a central role in this. It has been proposed as one of the causes of the 'missing heritability' that is observed for many complex traits. However, there is a lack of methods that can reliably detect such interactions, especially for quantitative traits.

Random Forest (RF) represents a powerful tool for the modelling of the effect of genetic variants on phenotype variation. This can be primarily attributed to its ability to take interactions into account. However, although it can account for these interactions, it does not specifically detect them. Here, three approaches that exploit RF for the detection of epistasis are proposed: split asymmetry, selection asymmetry, and paired selection frequency. Since they complement each other in different interaction scenarios, combinations of these methods were also created.

The ability of these approaches to detect epistasis was assessed on simulated and real data. They were compared to the commonly used exhaustive pairwise ANOVA approach, as well as to a total of six different two-stage approaches. Overall, the RF-based methods generally performed better at identifying genetic interactions than the other methods. Possible reasons for the performance differences are discussed in this thesis. This work contributes to the long-standing problem of extracting information about the underlying model from a Random Forest.

Acknowledgements

I want to thank all the people who contributed to the work on this thesis. First of all, I am grateful to my advisor, Mathieu Clément-Ziza, for providing the resources and guidance that made this work possible. I thank him for his constant support, for teaching me not only about the field of computational biology, but about good scientific practice in general.

I would also like to thank Andreas Beyer for co-supervising this thesis. I thank him for always taking time to speak about the project and for his valuable suggestions that played an indispensable role in shaping this project.

I want to express my gratitude to my colleagues from the first floor office of the CMMC building, and from the Beyer group, for making my time working on this thesis so enjoyable. I thank Janis for his support in both scientific and administrative matters and for many fruitful discussions. And I want to especially thank Jan for trustingly putting his former project into my hands, for his valuable suggestions and critiques, and for his continuous help in bringing the project forward.

Finally, I thank my family and friends. My parents for their constant generosity, support, and guidance during my entire studies. My sisters for their encouragement and for always being there for me. And I thank Ben, for all the patience, encouragement, and love he has given me, and for always bringing out the best in me.

Contents

1 Introduction

1.1 From Genotype to Phenotype

1.1.1 Genetic Association with Phenotype

Many biological traits are influenced by heritable genetic variation (G), by environmental factors (E), as well as their interaction ($G \times E$). This relationship can be expressed with the formula

$$P = G + E + (G \times E).$$

The genetic influence G can be further partitioned into G_A, the individual additive effects of all genetic variants on the phenotype, the dominance effects G_D, and the effects though gene-gene or marker-marker interactions, G_I. This results in

$$P = G_A + G_D + G_I + E + (G \times E).$$

Understanding the genetic components underlying biological traits has been a central focus of genetic research in humans, as well as in model organisms and agriculturally important species (Buckler et al., 2009; Atwell et al., 2010; Mackay et al., 2012; Aylor et al., 2011). However, in contrast to so-called Mendelian traits, where a single genetic variation causes the manifestation of a specific phenotype, many biological phenotypes follow complex inheritance patterns and are influenced by a multitude of genetic variants (Hill, 2010; Mackay et al., 2009). This makes the identification of the genetic variants that underlie these phenotypes difficult. Examples for such complex phenotypes include susceptibility to Diabetes or Alzheimer's Disease in humans (McCarthy and Zeggini, 2009; Harold et al., 2009), or parasite resistance in plants (Atwell et al., 2010).

In addition to these discrete traits, many quantitative traits (QTs), such as body height in humans or growth rate in yeast also fall into the category of complex traits. Genetic variants that influence QTs are called quantitative trait loci (QTL). The QT can also be a molecular phenotype, such as messenger ribonucleic acid (mRNA) expression or protein abundance. Genetic variants influencing such phenotypes are called expression quantitative trait loci (eQTL) or protein quantitative trait loci (eQTL), respectively. These molecular QTL play an important role in understanding the mode of action of genetic variants which have been associated with a complex trait. In particular, even when genetic variants are successfully associated with a complex trait, it is often challenging to understand the molecular mechanisms behind the influence of the respective variants, especially if they lie in non-coding regions of the genome. The central dogma of molecular biology states that desoxy-ribonucleic acid (DNA) codes for ribonucleic acids (RNAs), which in turn are translated into proteins, whose action leads to a phenotype (Crick, 1970)). Following this dogma, it is expected that the mode of action of the variants can be understood by identifying their effect on intermediate molecular traits (i.e. investigate whether they are eQTL, pQTL or influence another molecular trait).

1.1.2 Missing Heritability

The discrepancy between expected heritability and phenotypic variance that can be explained by known genetic variations was termed 'missing heritability' (Maher, 2008). In humans, heritability estimates from twin studies ranged from 80 to 90% for complex phenotypes such as susceptibility to autism (Sullivan, 2005), susceptibility to schizophrenia (Freitag, 2007), and height (Visscher, 2008). Through Genome-wide association studies (GWASs), a very high number of single nucleotide polymorphisms (SNPs) and other genetic variants have been associated with a wide range of phenotypes. However, cumulatively, these can usually explain only a subset of the heritability of each trait. The unexplained part represents the missing heritability. This implies that there are other, undiscovered genetic factors that influence the phenotypes (Manolio et al., 2009). Missing heritability is also often defined as the difference between the broad-sense heritability $H^2 = G/P$, all the genetic contributions to a population's phenotypic variance, and the narrow-sense heritability $h^2 = G_A/P$, the phenotypic variance due to only additive genetic effects (Bloom et al., 2013).

Several reasons for the observed missing heritability have been proposed. The penetrance, effect sizes, and/or the minor allele frequencies (MAFs) of the causal alleles might be too low, resulting in insufficient statistical power of current studies to detect them. Many researchers try to circumvent this problem by simply increasing the sample size (Bergen and Petryshen, 2012). Another possibility is that part of the heritability is due to epigenetic variants, structural variation, or other genetic factors that were not investigated so far. It is also possible that population structure and environmental factors were not taken into account appropriately, inflating the initial broad-sense heritability estimates and consequently the missing heritability. And finally, it was proposed that genetic interactions, or epistasis, underlie the missing heritability (Maher, 2008).

1.2 Epistasis

1.2.1 Definition and Relevance

Epistasis is defined as interaction between different genetic loci concerning their effect on a certain phenotype. More specifically, it describes the phenomenon when the effect of a genetic variant on a phenotype depends on the allele(s) of one or more 'modifier' variant(s). In the context of QTs, epistasis represents non-additive interactions between markers, i.e. situations where the contribution of two or more genetic variants on a QT differs from the sum of their 'marginal' effects. Marginal effects are the individual, non-interaction effects of markers on a phenotype. There are several models for the different possible manifestations of epistasis (Figure 1). In this thesis, two general types of epistasis will be distinguished, which are defined by their equivalent logical operations. For one, there is AND-epistasis, where an allele at one locus enhances or alleviates the effect of another locus (Figure 1B). Secondly, XOR-epistasis (Figure 1C) describes cases where the effects of alleles at two loci are diminished when they occur together (Phillips, 2008; Purcell and Sham, 2004).

As presented above (Section 1.1.2, p. 2), epistasis has been proposed as one factor underlying missing heritability of complex traits. Therefore, accounting for epistasis might be crucial to unravelling the genetic contribution to phenotypes. In addition, epistasis is often interpreted as a functional relationship between genes, since the interactions can be explained by the variants influencing the same biological process. Identifying interacting genes can thus contribute to elucidate disease mechanisms (Shao et al., 2008; Carlborg and Haley, 2004), to identify common biological functions of genes, and to ultimately reconstruct genetic interaction networks (Phillips, 2008; Tong et al., 2004; Schuldiner et al., 2005; Hannum et al., 2009; Costanzo et al., 2016).

1.2.2 Model Organisms to Study Epistasis

The effort to elucidate the effects of genetic variations on phenotypes can benefit a lot from studying model organisms, because populations with balanced MAFs can be created through targeted crosses, phenotypes can be efficiently and thoroughly measured, and environmental factors controlled (Johnsen et al., 2011; Parks et al., 2013; Flint and Mackay, 2009; Ehrenreich et al., 2009; Costanzo et al., 2010). For example, studies on crosses between two diverse *S. cerevisiae* strains have greatly contributed to our knowledge about the influence of genetic variation on phenotypes (Steinmetz and Davis, 2004; Ehrenreich et al., 2009; Brem and Kruglyak, 2005). Missing heritability has been demonstrated for several traits in yeast, for example for expression

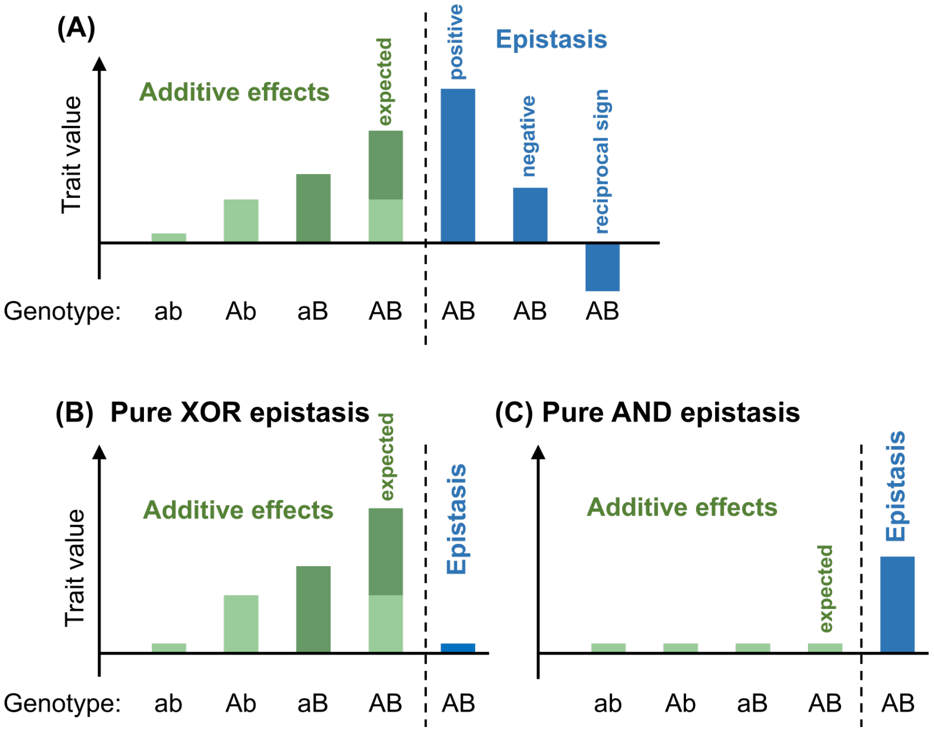

Figure 1: Different types of epistasis for a quantitative trait. On the left side, additive effects are shown, including the phenotype expected under additivity when perturbing both a and b, indicated by the stacked bars. Note that all types of epistasis are also applicable to negative trait values. (A) Examples for positive (enhancing, or synergistic), negative (alleviating, or antagonistic), and reciprocal sign epistasis. (B) Example for pure XOR epistasis, where the marginal effects of two alleles are nullified if they occur together. (C) Example for pure AND epistasis, where two alleles only show an effect if they occur together. XOR and AND-type epistasis are not necessarily exclusive from the epistasis types shown in A.

variation (heritability of 84%, but only 30% explained; Ehrenreich et al. 2010) and diverse growth traits (40% to 96% heritability but 0.21% to 0.84% explained; Bloom et al. 2013). Since the yeast strains are haploid and grown under controlled conditions, dominance effects (G_D) and environmental factors (E) can be neglected, reducing the formula from Section 1.1.1 (p. 1) to

$$P = G_A + G_I.$$

This facilitates the identification of interaction effects significantly.

1.2.3 Detection of Epistasis

The comprehensive detection of epistatic effects has been impeded by three major challenges. One is an experimental challenge: large sample sizes are required in order to sample the landscape of possible combinations of genetic variants, and in order to warrant sufficient statistical power to detect interactions despite their possibly small effect sizes. Another challenge is related to the multiple comparisons problem: the combinatorial number of hypotheses to test results in a strong multiple testing penalty. The high number of comparisons, as well as the needed sample size finally lead to a computational challenge (Niel et al., 2015; Mackay, 2014). The latter can be trivial for problems involving few loci, but can become unmanageable for much larger problems, particularly when multiple traits are investigated, as is the case with eQTL mapping (Picotti et al., 2013).

A wide variety of methods to detect epistasis have been developed (Nelson et al., 2001; McKinney et al., 2006; Dong et al., 2008; McKinney et al., 2009). Computationally efficient parametric methods for genome-wide interaction scans on GWAS data for discrete traits (i.e., Wan et al. 2010; Goudey et al. 2013) as well as non-parametric methods such as the multifactor-dimensionality reduction (MDR) method (Ritchie et al., 2001) have been proposed. However, there is a lack of accurate methods for the detection of interactions between genomic loci that affect QTs. These are often still investigated using exhaustive methods, for example by applying pair-wise ANalyses Of VAriance (ANOVAs) to all possible marker pairs to test for non-additive interactions (Purcell et al., 2007). These exhaustive approaches suffer from the computational and statistical pitfalls outlined above. Therefore, some authors use a two-stage approach where markers are pre-selected or weighed prior to interaction testing. The criteria used for the pre-selection can include the presence of marginal effects, knowledge from external databases, or data-mining approaches such as machine learning methods with variable selection (Marchini et al., 2005; Pecanka et al., 2017; Robnik-Šikonja and Kononenko, 2003; Grady et al., 2011). Obviously, the choice of this criterion might introduce a bias and must be chosen carefully to avoid false negatives. Moreover, many existing methods are parametric, which means that they make assumptions about the distribution of the data and the type, scale, and order of interactions, which may not apply to real data (Niel et al., 2015; Ritchie, 2015). The machine learning method Random Forest (RF), however, represents a promising approach for the modelling of phenotypes from genetic association data, and might be used to identify interactions.

1.3 Random Forest

1.3.1 The Random Forest Algorithm

Being a supervised machine learning method, Random Forest (RF) was designed to predict a certain outcome (in this case a phenotype) based on a set of predictors (in this case genetic markers, Breiman 2001). A RF consists of an ensemble of classification and regression trees (CARTs). In each CART, the data is split into two groups based on the alleles of the marker that explains the most variance of the phenotype. The resulting subgroups are sequentially split on other markers, following the same procedure. The trees are fully grown this way, until the number of samples in the final partitions are below a pre-defined threshold, or until there are no more features that can further explain the trait variance (Figure 2).

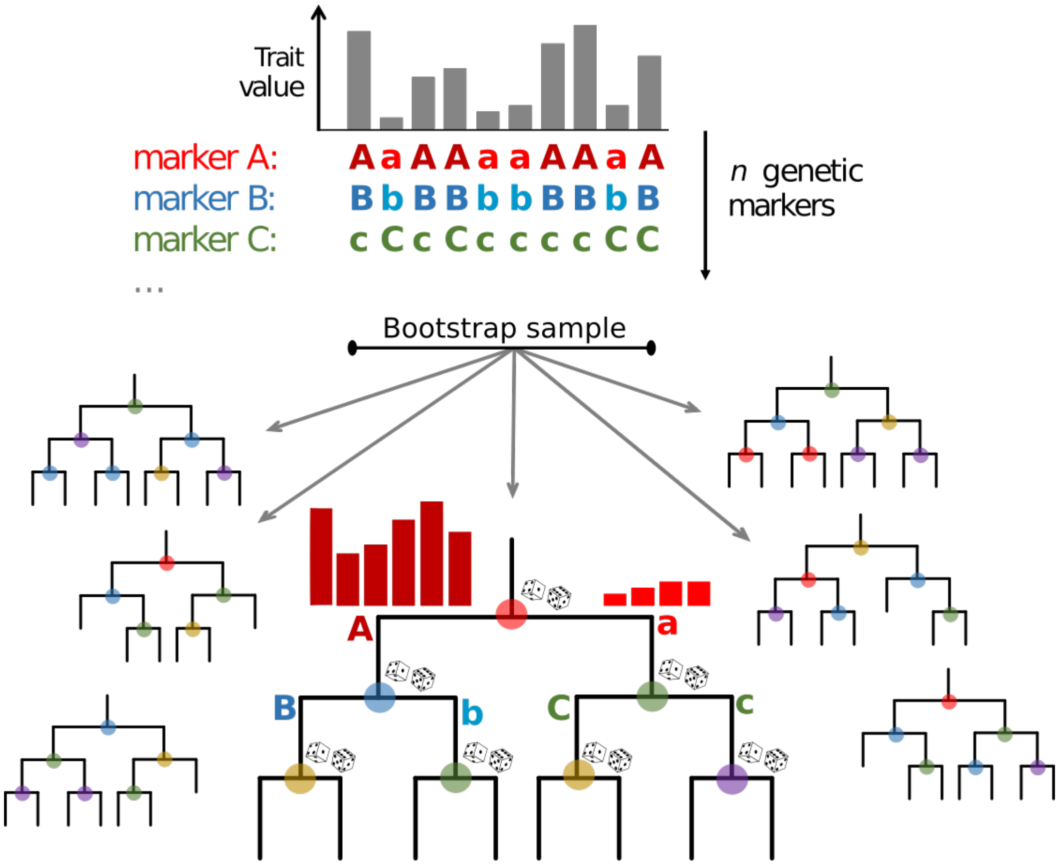

Figure 2: Schematic representation of the Random Forest algorithm. For each tree, a bootstrap sample of the individuals is taken. The marker that explains the most of the variance of the trait value is used to split the data into two groups. Here, the selected marker is represented as a dot at the splitting point in the color corresponding to the marker. The two subgroups are then again split on another marker, following the same procedure. At each splitting point, only a random subset of available markers is evaluated (represented by the dice).

The remarkable predictive power and robustness against overfitting of RFs stems from two separate random sampling steps (Breiman, 2001). For one, each CART is grown on a bootstrap sample of the data. On top of this, only a random subset of markers is considered for the split at each splitting point. When the RF is used for prediction, the collective vote of all trees based on the individuals that were not sampled in the bootstrap, the out-of-bag (OOB) individuals, is used for the prediction. Practically, the RF cannot only be used for prediction, but its structure can be analysed to draw conclusions about the relationship between the predictors (markers) and the outcome (phenotype). Several importance measures exist that indicate which predictors were important for the modelling of the outcome. In a genetic association context, this principle is used to determine which markers influence the phenotype, for example in eQTL mapping (Michaelson et al., 2009).

RF can be used for both quantitative (regression) and discrete (classification) outcome variables, as well as quantitative and discrete predictors. Furthermore, it is a non-parametric method, which means that no assumptions about the distributions of the data are made (Breiman, 2001).

Because of the hierarchical organization of the predictors in each tree, dependencies, or interactions, between features are taken into account. Notably, RF is not limited in the order of interactions it can model. Because of this ability to implicitly account for non-additive effects, RF has been shown to represent a very powerful method for detecting genotype-phenotype relationships, especially in the presence of epistasis, and outperforms other methods in the mapping of QTL (Michaelson et al., 2010; Wright et al., 2016; Stephan et al., 2015; García-Magariños et al., 2009; McKinney et al., 2009; Ackermann et al., 2012).

1.3.2 Detecting Interactions Using Random Forest

A notorious problem of RF remains that, although it is able to model interactions, it does not provide any information about relationships between predictors from the forest. However, that is necessary for detecting epistatic interactions between markers (Wright et al., 2016). There were several attempts at extracting interactions from RF or similar tree-based models. Yoshida and Koike (2011) proposed SNPInterForest, a modification of the RF algorithm combined with an evaluation of the probability of two markers occurring in the same path of a CART, to detect epistasis. However, their method suffers from a high computational burden, and was not sufficiently benchmarked on real data. Aside from that, it is only applicable to discrete phenotypes. The same problems apply to two other tree-based approaches, GWGGI (Wei and Lu, 2014) and the 'maximal conditional chi-square' measure (MCC,Wang et al. 2010). Wright et al. (2016) tested the interaction detection capability of three measures, the 'pairwise permutation importance', the 'joint importance by maximal subtrees', and the 'joint variable importance'. They concluded that, while RF successfully captured the interactions, these measures are not suited to detect epistasis.

One approach for the detection of interactions using RFs that is applicable for QTs, termed the split asymmetry (splitA) approach, was developed and applied by Michaelson (2010) and Picotti et al. (2013), respectively. The approach was subsequently expanded and further benchmarked on simulated and real data (Grossbach, 2015). It is based on the detection of specific signatures of how the trait values will distribute across the splits in the CARTs created by non-additive interactions (details in Section 2.2.2, p. 14). A thorough search of the literature yielded no other approaches that successfully detect epistasis for QTs using RF. Therefore, the splitA represents a unique contribution to QTL research. However, several potential improvements remain to be implemented for the splitA approach. For one, it displayed poor performance at detecting XOR epistasis. This is due to the fact that the interacting markers are unlikely to be selected for a split in the first place, because they individually do not explain any variance. Another problem of the splitA approach is that, for two markers (A and B) in AND-epistasis, marker B will be less likely to be selected on one side of the split on A than on the other side (Figure 3B, red dashed arrow, p. 15). This potentially impedes the detection power of the splitA approach. Lastly, the implementation of the approach could be optimized for improved efficiency and statistical power.

1.4 Aim of the Thesis

Being able to detect epistatic interactions is crucial to elucidating the genetic factors underlying complex traits. However, there is a lack of precise methods for the detection of epistasis, especially for QTs. Because of the high performance of RF for QTL calling, and because of its ability to implicitly account for interactions, a special focus should lie on the development of methods that exploit the structure of RF to detect epistasis.

This thesis builds on previous work from Grossbach (2015) on the detection of interactions from RF. Here, two main questions are adressed: First, can the precision of identifying interactions in RFs be improved? And second, how do RF-based epistasis detection methods perform compared to other approaches?

To address the first issue, the pitfalls of the splitA approach outlined in Section 1.3.2 (p. 7) are circumvented by several measures. First, the splitA approach is enhanced to achieve a higher efficiency. In addition, two novel methods that exploit the structure of RF to detect epistatic interactions are developed in this thesis, namely the selection asymmetry (selA) and the paired selection frequency (pairedSF) approaches. They are expected to complement the splitA method in scenarios where it is not able to identify interactions. Therefore, combinations of these RF-based approaches are also created.

Then, in order to answer the second question, these methods are compared to other common approaches for the detection of epistasis, including an exhaustive ANOVA and several two-step approaches. The methods are benchmarked on simulated data and on real data. Finally, the performances of all approaches are evaluated to identify the most powerful epistasis detection method for QTs.

Being able to not only account for, but to actually identify interactions in RFs would enhance the usability of RF. The methods for detecting feature interactions in RF are applicable beyond the genetic mapping problem.

2 Data and Methods

2.1 Data

2.1.1 Simulated Data

In this thesis, simulated data from Grossbach (2015) was used. In these simulations, traits were simulated based on genotypes from the widely used data from a cross between the two *S. cerevisiae* strains RM11-la (RM) and BY4716 (BY, Brem and Kruglyak 2005). The phenotypes were created by adding marginal and interaction effects with the indicated effect sizes to a baseline of 9, and subsequently adding normally distributed noise with standard deviation $\mathrm{sd} = \mathrm{noiselevel} * \sqrt{\mathrm{mean}}$. In total, 19 different combinations of marginal and epistatic effects with varying effect sizes, different types and orders of epistasis were simulated at eight noise levels (0 to 20%). Each scenario was simulated 32 times. The parameters for the different scenarios are outlined in Table 1 (Grossbach, 2015).

2.1.2 Expression Quantitative Trait Data

Mapping eQTL is a key aspect towards understanding the relationship between genetics and phenotype. Therefore, a dataset from a RM×BY yeast cross that consists of RNA expression information along with genotype information of 3,593 markers for 112 segregants, was selected (unpublished data). The pre-processing of this dataset and the creation of performance measures based on the double knock-out (DKO) data were done together with Jan Grossbach. Since the results of the interaction mapping of the different methods would be evaluated based on the recovery of interactions for growth phenotypes (double knock-out, DKO, dataset, Section 2.1.4, p. 12), 1050 transcripts were selected that code for essential genes according to the Saccharomyces Genome Deletion Project (www-sequence.stanford.edu/group/yeast_deletion_project/deletions3.html). This was done based on the assumption that genetic variants which affect the expression of essential genes are more likely to also effect growth, than variants which affect non-essential genes.

The population structure covariates were created as follows: first, the relatedness, or kinship, between strains was estimated using the R package 'emma' (Kang et al., 2008). The first six eigenvectors, explaining 60% of genotypic variance, were used as representatives for the population structure.

For the benchmark, the markers had to be mapped to nearby genes. For each marker, all markers that were in linkage disequilibrium (LD, absolute correlation coefficient above 0.9) and that were less than 50 kilobases apart, were

Table 1: Outline of the features of the different simulation scenarios. Adapted from Grossbach (2015).

Scenario	Number of purely additive effects	Effect size of the additive loci	Number of epistatic interactions	Type of epistatic interaction	Number of loci involved per epistatic interaction	Effect size of the epistatic interactions	Shared loci between additive and epistatic effects
EA1	0	-	1	AND	2	1	-
EA2	0	-	1	AND	2	0.5	-
EA3	0	-	1	AND	3	1	-
EA4	0	-	1	AND	3	0.5	-
EX1	0	-	1	XOR	2	1	-
EX2	0	-	1	XOR	2	0.5	-
M1	1	1	1	AND	2	1	No
M2	2	1	1	AND	2	1	No
M3	3	1	1	AND	2	1	No
M4	4	1	1	AND	2	1	No
M5	2	1	2	AND	2	0.5	Yes
M6	2	1	2	AND	2	1	Yes
M7	2	1	2	AND	2	0.5	No
M8	2	1	2	AND	2	1	No
M9	2	1	2	XOR	2	1	Both
M10	1	1	1	AND	2	1	Yes
M11	1	1	1	AND	2	1	No
M12	1	3	1	AND	2	1	Yes
M13	1	3	1	AND	2	1	No

grouped. The marker region was then defined as starting directly after the previous marker (position of previous marker + 1) and ending just before the next marker (position of next marker - 1), or at the start or end of each chromosome for its first and last marker, respectively. All genes that overlapped this marker region according to the ENSEMBL annotation (Yates et al., 2015) were then assigned to the marker.

2.1.3 Growth Trait Dataset

For the evaluation of the methods on more complex traits, another yeast cross dataset from Wilkening et al. (2014) was selected. It consisted of genotype and phenotype information for 720 distinct segregants of a cross between the strains BY4742 and SK1. The dataset encompassed measurements for a total of twelve phenotypes: three discrete phenotypes (resistance to Cantharidin, flocculation in YPD, and colony shape, e.g. wrinkled versus smooth) and nine continuous phenotypes, including fitness in rich media (i.e., fitness in YPD, and optical density saturation), growth at high temperatures and high salt concentrations (38°C in YPD, and NaCl in YPD, respectively), growth with non-fermentable carbon sources (medium with 2% alcohol, YPE; medium with 3% glycerol, YPG; and medium with 2% lactic acid, YPL), as well as resistance to 5-fluorouridine. The phenotype values were scaled and centered to mean zero and equal variances.

The dataset also encompassed genotype information for 65,250 markers. Since this number is too high to feasibly perform QTL mapping, and since many markers were highly correlated, markers in high LD were grouped in the following way. For each chromosome, a matrix of absolute Pearson correlation values subtracted from one (1-abs(cor)) was computed and used as a distance measure for a hierarchical clustering of the markers. The clustering tree was then cut at a height of 0.02 to get clusters of highly correlated markers. Most times the grouped markers were located consecutively on the chromosome. This procedure resulted in 2,827 marker groups genome-wide. For each marker group, a 'consensus marker' was created by using the majority vote (rounded mean of alleles which were encoded with 0 or 1) of markers in the group. If the mean was between 0.4 and 0.6 the genotype was assigned as unknown (NA). The resulting MAFs ranged from 0.3 to 0.7, and there were no markers with more than 11% missing genotypes.

The population structure was estimated in the same way as for the eQTL dataset (Section 2.1.2, p. 10). Five principal components, explaining 55 % of genotypic variance, were used as representatives for the population structure.

Markers were mapped to genes in the following way: each (original, non-grouped) marker was defined as starting directly after the previous marker (position of previous marker + 1) and ending just before the next marker (position of next marker - 1), or the start or end of the chromosome for its first and last marker, respectively. If there was a gene overlapping this region according to the ENSEMBL annotation, it was assigned to the marker. Then, all genes that were assigned to the markers in this group were assigned to the consensus marker. This way, a maximum of 10 genes were assigned per marker group.

2.1.4 Double Knock-Out Dataset

In order to evaluate the methods' performance at identifying interactions in the datasets described above, their results were compared to the dataset from Costanzo et al. (2016). This dataset consists of gene-gene interactions that were identified from co-knockouts of yeast genes using the synthetic genetic array (SGA) method. This dataset is referred to as the double knock-out dataset (DKO) in this thesis.

The data from the paper was downloaded, and processed as follows: the datasets for essential (both damp and temperature sensitivity alleles) and non-essential genes were combined. All pairs containing genes that were removed in the latest reference annotation (ENSEMBL) or that were marked as 'dubious' were excluded. Interactions which passed the lenient threshold used in the original paper (p-value ≤ 0.05) were considered significant. A genes × genes matrix was used to represent whether two genes interact (encoded as 1) or not (encoded as 0). Note that gene pairs that were significant in at least one assay were considered to interact, e.g. genes that were found to interact in only one of the two approaches were set to 1. Overall, about 13% of all possible gene pairs were found to interact.

2.1.5 Databases

Three databases were used for this thesis, the Reactome Pathway Database (Fabregat et al., 2016), the STRING database (Szklarczyk et al., 2017), and the Biological General Repository for Interaction Datasets (BioGRID, Chatr-Aryamontri et al. 2017).

The Reactome database pathways were downloaded (version 62, 'lowest level ENSEMBL to pathways', https://reactome.org/download-data) and filtered for *S. cerevisiae* pathways. If any of the genes assigned to a pair of markers occurred together in any pathway, it was considered a true interaction.

STRING interactions for yeast were downloaded (version 10.5, 'protein network data including subscores per channel', https://string-db.org/cgi/download.pl) and a directed network was created from them. The combined scores for the interactions were used as edge weights. Then, all pairwise network distances were computed using the 'distance' function from the 'igraph' R package.

The BioGRID database was downloaded (version 3.4.154, https://downloads.thebiogrid.org/BioGRID) and gene-gene interactions with the label 'genetic' for *S. cerevisiae* extracted. A directed, unweighted graph was created from these interactions using the R package 'igraph' and all pair-wise shortest path distances computed.

2.2 Methods

2.2.1 General Random Forest Methods

All methods were implemented using the R statistical environment (**?**). For the detection of epistasis using RF, a special R implementation 'RandomForestExtended' was used (Grossbach, 2015). It differs from the original implementation from Liaw and Wiener (2002) by its ability to parallelize growing the trees, and to return OOB predictions also for internal tree nodes. For all RF-based epistasis detection approaches, the RF was built the same way. Each time, 30,000 trees were grown, and the minimum final nodesize was set to 5. All other parameters were left at the default.

RF can not deal with missing data in the predictors. Therefore, if there were missing genotypes, they were imputed in 100 iterations. For each iteration, the missing alleles were randomly selected with sampling probabilities according to the respective MAFs at each locus. 300 trees were built based on each iteration, so that the total number of trees was 30,000. The collection of forests was then united into one big forest.

For the growth trait dataset and the eQTL dataset, population structure representatives were included as covariates in the model (i.e., as predictors used to build the RF), as proposed previously (Clement-Ziza et al., 2014).

2.2.2 Random Forest Split Asymmetry

This approach exploits the fact that the splitting behaviours of two interacting markers are influenced by each other if they occur in the same path of a CART. This phenomenon can be observed for both AND-type and XOR-type epistasis. The specific splitting behaviour was used by Michaelson (2010) and Picotti et al. (2013) to detect epistatic interactions and termed split asymmetry (splitA).

Given two hypothetical 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 3, green slopes). The collection of slopes S_{ABl}, S_{ABr} (B was used on the left, or right side of A, respectively), S_{BAl}, and S_{BAr} (A was used on the left, or right side of B, respectively) were collected for all marker pairs from all CARTs in a forest. A Student's t-test was then used to check for slope imbalances for all marker pairs, by comparing S_{ABl} with S_{ABr}, and S_{BAl} with S_{BAr}, respectively. The t-test was only applied if the marker combination occurred in the same path of a CART at least five times throughout the forest.

In the previous implementation of this approach (Grossbach, 2015), only 100 slopes were stored and tested per marker pair due to computational limitations. Here, the memory limitations were circumvented by, instead of storing the actual slope values, sequentially updating the sums of slopes $\sum S_{ABl}$ and $\sum S_{ABr}$, the sums of squares of slopes $\sum(S_{ABl})^2$ and $\sum(S_{ABr})^2$, and the numbers of slopes n_{ABl} and n_{ABr} (here for the case where B was used on the left

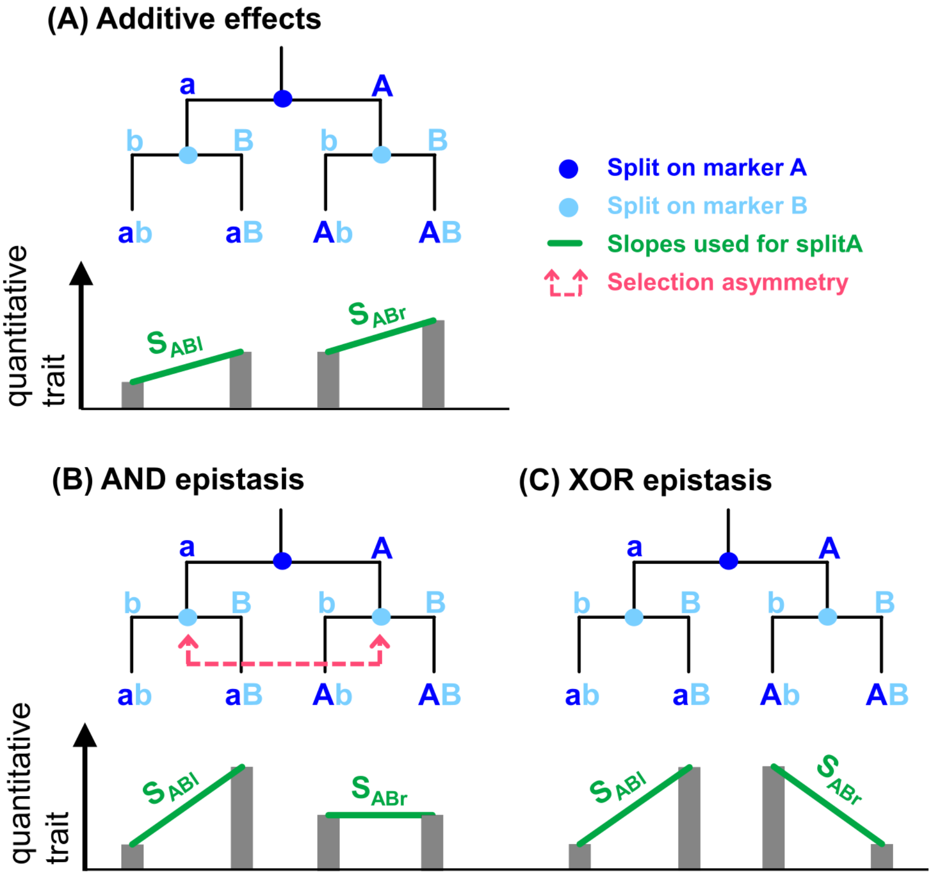

Figure 3: Schematic representation of the detection of epistasis from Random Forest. Shown are hypothetical subtrees that depict the splitting of data on two markers A and B that (A) do not interact, (B) are in AND epistasis, or (C) are in XOR epistasis. The latter two lead to asymmetries in the trait value distribution (indicated by the green slopes), which are exploited in the splitA 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 selA approach. Not only cases where two markers are used directly after each other are considered, but also cases where there are other splits between them.

side of A). The mean and variances of the slopes were calculated as

$$m_{ABl} = \frac{\sum S_{ABl}}{n_{ABl}} \quad \text{and} \quad v_{ABl} = \frac{\sum S_{ABl}^2}{n_{ABl}} - (m_{ABl})^2,$$

respectively. The t-statistic was then calculated as:

$$t_{AB} = \sqrt{\frac{(n_{ABl} * n_{ABr})}{n_{ABl} + n_{ABr}}} * \frac{m_{ABl} - m_{ABr}}{se}$$

with

$$se = \sqrt{\frac{(n_{ABl} - 1) * v_{ABl} + (n_{ABr} - 1) * v_{ABr}}{n_{ABl} + n_{ABr} - 2}}$$

and compared to a t-distribution with $n_{ABl} + n_{ABr} - 2$ degrees of freedom.

The same was done for the cases where marker A was used after marker B, so that each marker pair was tested twice in two independent tests. The two p-values were then combined using the Fisher method (Fisher, 1925). If only one p-value was present (e.g., the marker pair occurred in the same trees in the order AB, but never in the order BA), only this p-value was used. Marker pairs that had an absolute correlation coefficient $R \geqslant 0.9$ were not tested.

These modification not only lead to a better memory efficiency, but also increased statistical power, because more than 100 slopes could be compared. In addition, the implementation was further optimized by exploiting the efficiency of vectorized procedures in R.

2.2.3 Random Forest 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 3, red dashed arrow). This means that there might not be enough slopes for a marker pair in order for the splitA approach to be able to detect it, although they interact. However, this selection asymmetry (selA) property can also be exploited to detect epistasis. This was done by counting, for each marker pair (A and B), the number of times B was used on the left side, or the right side after a split on A, respectively (n_{ABl} and n_{ABr}). In addition, the number of splits in the tree where marker B was *not* used after a split on marker A were counted ($n_{A\bar{B}l}$ and $n_{A\bar{B}r}$). A binomial test of equal probabilities without Yates' continuity correction was then used to test for this interdependence in the marker selection frequency. The total number of splits following a split on each side of A ($n_{Al} = n_{ABl} + n_{A\bar{B}l}$ and $n_{Ar} = n_{ABr} + n_{A\bar{B}r}$) were used as the numbers of trials , and n_{ABl} and n_{ABr} represented the numbers of successes. Only marker pairs where B was used at least five times on either side of A were tested. Again, each marker pair was tested twice: once for the cases where marker B was used after marker A in the tree, and also for the opposite case. The p-values for the two possible pairs (AB and BA) were combined using the Fisher method. Marker pairs with absolute correlation coefficient above 0.9 were excluded.

2.2.4 Random Forest Paired Selection Frequency

This approach for the detection of epistasis, termed pairedSF, is based on the expectation that interacting markers are more likely to be selected in the same

CART than non-interacting markers. It was inspired by the approach from Yoshida and Koike (2011), but, in contrast to their SNPinterforest, it is not dependent on permutations for the evaluation of the interaction significance.

The number of times two markers were selected in the same tree of a RF (Number of co-occurrences N_{AB}) was compared to the number of times the markers were selected independently of each other ($N_{A\bar{B}}$ and $N_{\bar{A}B}$). For this approach, the order which marker was selected first is irrelevant. These counts were used to build a contingency table (Table 2) and a one-sided Fisher's exact test (Fisher, 1925) was applied to detect co-dependencies between the markers. Note that cases where marker combinations were used several times in the same tree were just counted as one co-occurrence, and cases where two markers occur in the same tree, but not in the same path, were still counted as a co-occurrence.

Table 2: Contingency table used for the paired selection frequency test.

	B	$\bar{\mathbf{B}}$
A	N_{AB}	$N_{A\bar{B}}$
$\bar{\mathbf{A}}$	$N_{\bar{A}B}$	$N_{\bar{A}\bar{B}}$

2.2.5 Random Forest Ensemble Approach

The three RF-based epistasis detection methods described above complement each other for different types of epistasis (i.e., selA theoretically only works for AND-epistasis, and pairedSF is most likely to be able to detect XOR-epistasis). In addition, the methods are applicable in different scenarios due to different statistical requirements: splitA and selA can only be used when a marker was used at least five times on both sides, or on either side of another marker, respectively. Thus, combining the approaches might lead to a better epistasis detection performance.

The p-values generated by the splitA, selA, and pairedSF approaches were combined using the Fisher method to create an ensemble score. In addition, all two-way combinations of methods (splitA + selA, splitA + pairedSF, and selA + pairedSF) were created. Only the p-values of the methods that were available for each marker pair were used.

2.2.6 Exhaustive ANOVA

For all marker pairs (e.g. A and B) with an absolute correlation coefficient $R \leq 0.9$, a linear model that models the phenotype y was computed using the formula

$$y = \sum_{i=1}^{P} (\hat{\beta}_i * p_i) + \hat{\beta}_A * A + \hat{\beta}_B * B + \hat{\beta}_{A\times B}(A \times B),$$

where P represents the collection of population structure covariates p_i. The significance of the interaction term $\hat{\beta}_{AB}$ was then used as the significance of the marker-marker interaction.

2.2.7 Two-Step Random Forest Approach

As mentioned in a previous Section (1.3.1, p. 5), RF importance measures can be used for QTL calling. Here, RF-based QTL calling was performed as described previously (Grossbach, 2015), and used to pre-select markers for interaction testing. To this end, a forest with 20,000 trees was computed and the importance scores (combined score, Grossbach 2015) of every marker extracted. In the case of missing genotypes, they were imputed in 100 iterations as described above (Section 2.2.1, p. 13), and 200 trees were grown on each imputation. The same procedure was repeated 400 times, but with permutations of the phenotype, so that the association between genotype and phenotype was disrupted. The relationship between phenotype and population structure variables was not permuted. This resulted in null distributions of importance scores for every marker. The null distributions were then used to determine the significance of the association of each marker with the phenotype. Markers with a q-value ≤ 0.1 after false discovery rate (FDR, Benjamini and Hochberg 1995) correction were considered significant. An ANOVA was then used to test for interactions within the significant markers (‘both significant’) or between significant and all other markers (‘one significant’), following the procedure described in Section 2.2.6 (p. 17). Only marker pairs with absolute correlation coefficient R leq0.9 were tested.

2.2.8 *t*-Test-Based Two-Step Approach

This approach employs a t-test to pre-select markers for interaction testing. First, an exhaustive t-test was used to identify significant marginal effects. Samples with an unknown genotype for a marker were not used. It should be noted that in this first pre-selection step, population structure could not be taken into account. Variants that passed the significance threshold of q-value ≤ 0.1 after FDR correction were considered to have a marginal effect. An ANOVA was subsequently used to test for interactions for marker pairs where either both markers had a marginal effect, or only one marker had a marginal effect. For this second step the population structure was taken into account. Marker pairs with an absolute correlation coefficient above 0.9 were not tested.

2.2.9 LASSO-Based Two-Step Approach

For this approach, the machine learning method Least Absolute Shrinkage And Selection Operator (LASSO) was used to preselect markers for interaction testing. Since the implementation of LASSO from the ‘glmnet’ R package cannot handle missing values, unknown genotypes had to be imputed. Similar to the procedure used for RF concerning missing genotypes, 500 distinct imputations based on allele frequencies were created. For each imputation, a LASSO model was computed. In each model, the phenotype was regressed against the collection of markers (M) and population structure (P) variables:

$$y = \sum_{i=1}^{P} \hat{\beta}_i * p_i + \sum_{i=1}^{M} \hat{\beta}_i * m_i$$

Like RF, LASSO is deemed to perform well at selecting the most predictive variable out of many correlated predictors, as is necessary for genetic mapping. The parameter selection is done using the regularization condition

$$\hat{\beta}_{lasso} = \underset{\hat{\beta} \in \mathbb{R}}{\operatorname{argmin}} \, ||y - X\hat{\beta}||_2^2 + \lambda \sum |\hat{\beta}_j|.$$

In practice, this means that, depending on the parameter λ, variable coefficients will be sequentially set to zero, leading to a sparse model. The optimal λ was determined using cross-validation.

A marker was assumed to have a marginal effect if its coefficient was nonzero in at least 5 of the 500 models. Finally, using an ANOVA interactions were tested for marker pairs where both markers had marginal effect (‘both’), or one marker had a marginal effect (‘one’). Marker pairs with an absolute correlation coefficient above 0.9 were not tested.

2.2.10 Performance Evaluation on Simulated Data

Each simulation scenario was simulated 32 times. Accordingly, the performance of each method was evaluated by counting the number of simulations where the p-value(s) of the truly interacting marker pair(s) was below the lower 0.5 percentile of p-values (i.e., 99.5% of p-values were higher), at each noise level. For scenarios with more than one epistatic marker pair, all of them had to be detected in order for the simulation to be considered recovered. For three-way interactions, all pair-wise combinations of the interacting markers had to be detected.

Since each simulation scenario encompassed only a very small number of interactions (maximum two), the results for the simulations of the same scenario were merged in order to compute the receiver operating characteristic (ROC) and precision-recall (PR) curves. As the name suggests, PR plots precision against recall (i.e. true positive rate, TPR), and ROC plots TPR against false positive rate (FPR). The respective areas under them are called AUROC and AUPR. The R package ‘precrec’ was used to compute these performance measures (Saito and Rehmsmeier, 2017).

2.2.11 Performance Evaluation Based on DKO Dataset

The DKO dataset (Section 2.1.4, p. 12) was used as a gold standard to evaluate the performance of the methods for the application to real data. A marker pair was considered a true positive if any of the pairwise combinations of the genes assigned to the two markers were found to interact in the DKO. These labels, along with the p-values, were then used to construct ROC, and PR curves, and the respective areas under them, using the ‘precrec’ R package (Saito and Rehmsmeier, 2017). In order to compute empirical p-values for the performance measures, null distributions for AUROC and AUPR were generated from permutations of the DKO reference data. Multiple testing correction was subsequently done using the Bonferroni method.

3 Results and Discussion

3.1 Selection of Epistasis Detection Methods

Three different tests that exploit the structure of RF to detect epistasis were extended or developed in this thesis: the splitA approach, the selA approach, and the pairedSF approach. In addition, combinations of these methods (splitA + selA, splitA + pairedSF, and splitA + pairedSF) as well as an overall ensemble method (splitA + selA + pairedSF) were conceived by combining the p-values of the respective approaches. The implementation of the splitA approach was modified for a better computational efficiency and statistical power (Section 2.2.2, p. 14).

All of these RF-based methods were compared to other common approaches for the detection of interaction between QTL. For one, an exhaustive ANOVA was performed, testing for interaction between all possible marker pairs. Due to the feature selection properties of the RF algorithm, it ultimately pre-selects markers based on their influence on the phenotype. In order to investigate the influence of this pre-selection on the methods' performance, several two-step approaches were implemented and applied as well. To that end, three different pre-selection criteria were chosen: (i) the 't-test-based two-step approach', where markers were selected based on whether they manifested marginal effects according to an exhaustive t-test screen, (ii) the 'LASSO-based two-step approach', where markers were preselected using the feature selection properties of the machine learning algorithm LASSO, and (iii) the RF-based two-step approach, where markers were pre-selected based on RF importance measures (as described in Grossbach (2015) and Section 2.2.7, p. 18). For each criterion, an ANOVA was subsequently used to test for interactions *within* preselected markers ('both' with marginal effect), and between pre-selected markers and all others ('one' with marginal effect), as both procedures are commonly applied (Evans et al., 2006).

It is important to distinguish the 'RF-based' epistasis detection methods (selA, splitA, and pairedSF, as well as their combinations), and the 'two-step RF' approaches ('two-step RF one' and 'two-step RF both'). The RF-based methods differ from the two-step RF approaches by two major criteria: (i) the procedure of marker pair pre-selection (minimum number of co-occurrence in the trees versus importance measure, respectively) and (ii) the method for interaction testing (exploiting the structure of RF versus ANOVA, respectively).

3.2 Benchmark on Simulated Data

3.2.1 Performance on Different Simulation Scenarios

The selected methods (previous Section, 3.1) were benchmarked on the 19 different simulation scenarios (Section 2.1.1, p. 10). The performance of the methods was evaluated by checking whether the *p*-values of the truly interacting markers were among the smallest 5% of *p*-values. It should be noted that the way the interactions were modelled (i.e., using a formula with marginal and interaction terms) conforms to the model used for an ANOVA, which might favour the ANOVA-based approaches.

The performances of the different methods are shown in Figures 4 to 7. In general, all RF-based methods were able to detect both two-way and three-way interactions (Figure 4). Among them, the ensemble method generally performed best for the simulated data benchmark. It also usually outperfomed the other methods at recovering simulated interactions. The exhaustive ANOVA, and the *t*-test-based two-step approach testing marker pairs where one of the markers had a marginal effect ('2stepANOVA one'), were the only methods that performed at a comparable level. The two-step approaches based on LASSO and RF were not able to recover any simulated interactions. These two selection criteria usually identified only a very small number, or even no marginal effects. Apparently, they were not be able to pre-select the interacting markers, even when one of them actually had a marginal effect (Figure 6). However, this could be due to too strict cutoffs for the pre-selection criteria for these two approaches.

The RF-based methods handled interfering marginal effects equally well as, or better than the other methods (Figure 5). The RF-based methods profited a lot from cases where at least one of the interacting markers had a marginal effect (Figure 6). This represents one of the limitations of RF: the modelling relies on the presence of marginal effects. Correspondingly, the two-step approaches applied in this thesis also rely on marginal effects of at least one of the interacting markers. However, AND epistasis will, even in the absence of underlying additive effects, cause the involved markers to appear as having marginal effects, promoting their selection in the RF. This is the reason why the RF-based methods were still able to uncover the interactions in scenarios without any marginal effects. In addition, in a real biological setting it is unlikely that interacting loci lack any marginal effects, which relativises the severity of this limitation.

These 'quasi-marginal' effects induced by AND-epistasis do not arise for XOR epistasis. Accordingly, the RF-based methods showed very poor performance for the recovery of XOR-epistasis, compared for example to the exhaustive

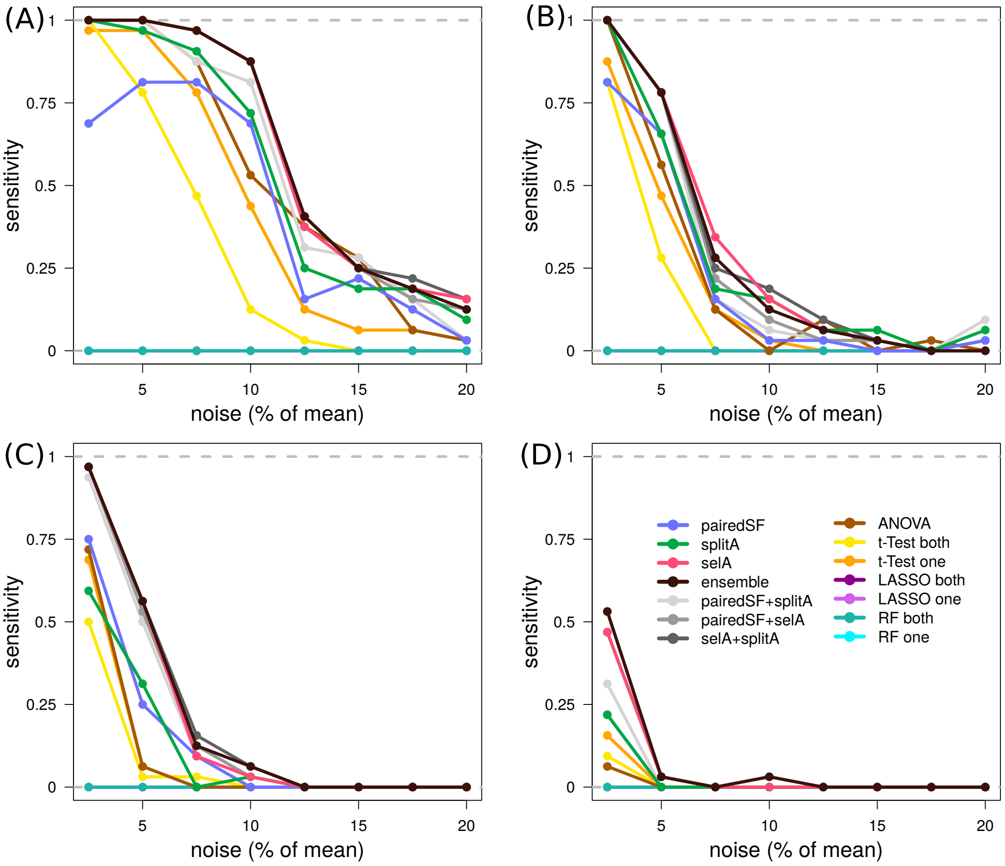

Figure 4: Sensitivity of methods based on simulations of pure two- and three-way AND epistasis. Sensitivity was measured by the proportion of simulations where the p-value of the truly interaction marker pair was below the lower 0.5-percentile of p-values (i.e. 99.5% of p-values were higher). (A) pure (no marginal effects) two-way epistasis, effect size 1, scenario EA1; (B) pure two-way epistasis with effect size 0.5, scenario EA2; (C) pure three-way epistasis, effect size 1, EA3; (D) pure three-way epistasis with effect size 0.5, EA4. For the three-way interactions, all possible two-way combinations of the three interacting markers had to be detected in order for the simulation to be considered recovered.

ANOVA (Figure 7). Surprisingly, adding marginal effects to both interacting markers did not improve the performance of the RF methods (Figure 7C). The pairedSF method, which had generally shown weaker performance for simulation scenarios involving AND-type epistasis, performed comparatively better for XOR epistasis. The biological relevance of XOR epistasis has been a matter of controversy. There are theoretically possible mechanisms that can lead to XOR epistasis for both discrete and quantitative traits (Moore and Williams, 2009). For example, one could consider the hypothetical case of two loci jointly influencing the QT growth rate through XOR epistasis. This could be explained through a threshold model: the alternative allele of

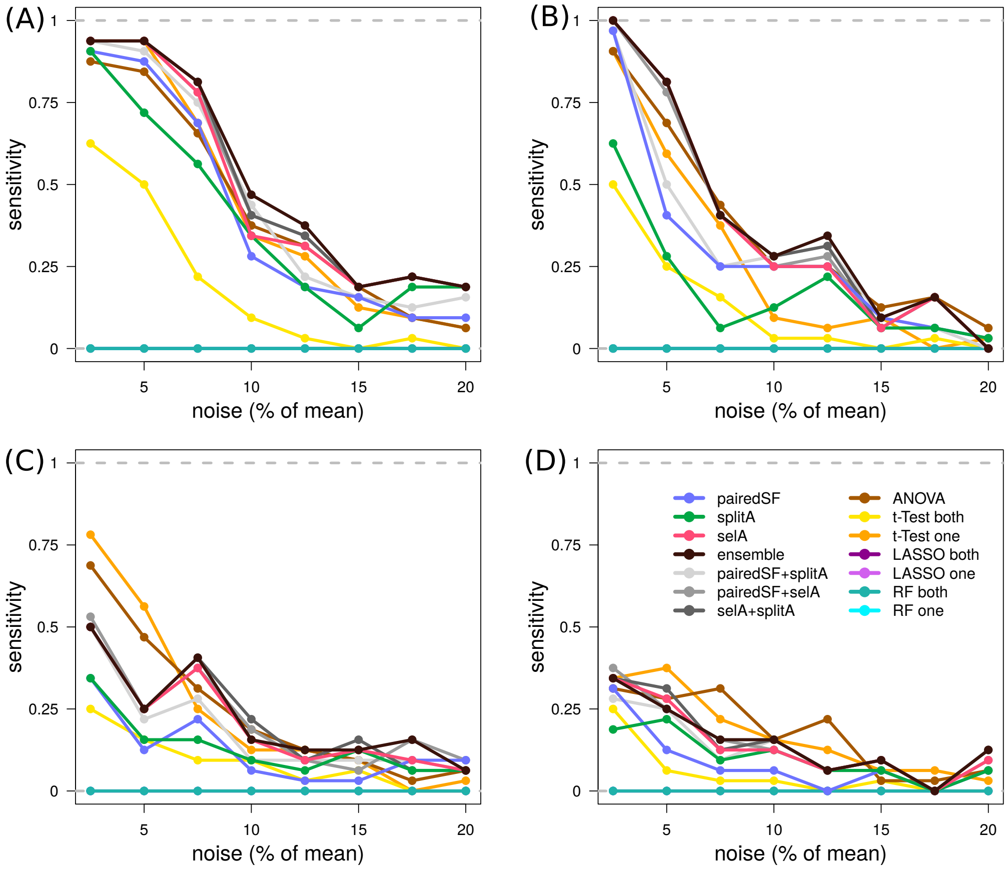

Figure 5: Influence of increasing number of marginal effects on interaction detection sensitivity. Sensitivity was measured by the proportion of simulated interactions for which its p-value was below the lower 0.5-percentile of p-values (i.e. 99.5% of p-values were higher). (A) to (D) represent simulations with one AND-type interaction each, as well as marginal effects for one to four separate markers, respectively (scenarios M1 to M4). The ability to detect interactions decays with increasing number of marginal effects for all methods.

each locus increases the cellular abundance of a certain factor, which in turn promotes growth. However, if both loci are altered, the factor abundance is raised above a critical level, ultimately leading to a detrimental effect on growth rate. Despite the theoretical possibility of its existence, a thorough search of the literature did not uncover any real biological examples for XOR epistasis. This lack of discoveries, however, might also be due to the fact that this type of interaction is difficult to detect (Li and Reich, 2000; Moore et al., 2006). In conclusion, one can assume that pure XOR epistases, without any marginal effects of the involved markers, plays a rather minor role in a real biological setting.

The pairedSF approach tended to perform not as well as the other RF-based

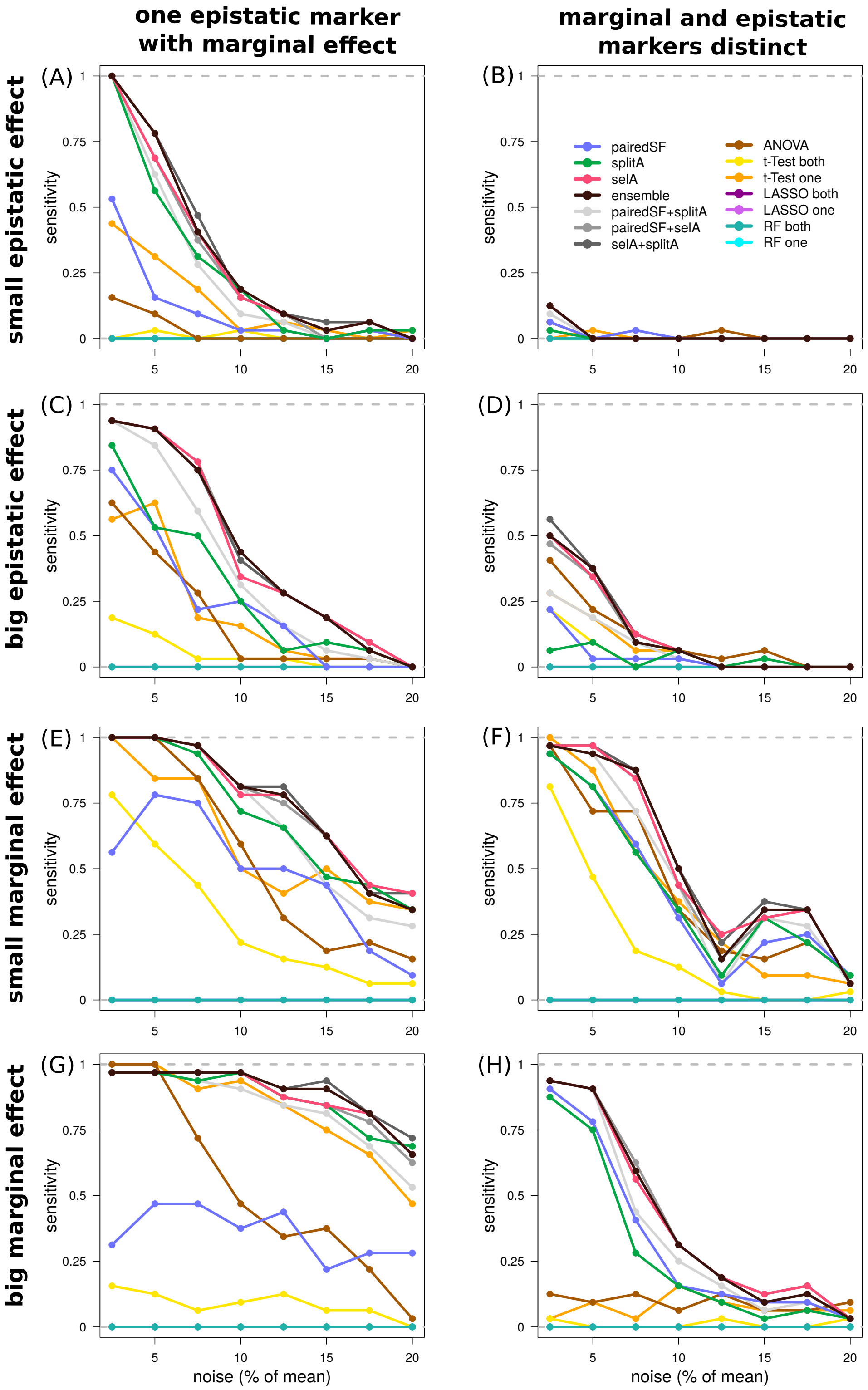

Figure 6: Sensitivity of methods for AND epistasis with and without marginal effects of the interacting markers. Sensitivity was the proportion of simulated interactions where its p-value was among the smallest 5% of p-values. (A) to (H) represent simulation scenarios M5, M7, M6, M8, M10, M11, M12, and M13, respectively. Cases where the interacting markers additionally had a marginal effect (left panel) are compared to cases where the markers with marginal effects were distinct from the interacting markers (right panel). Relevant features of each simulation scenario are indicated left of the plots.

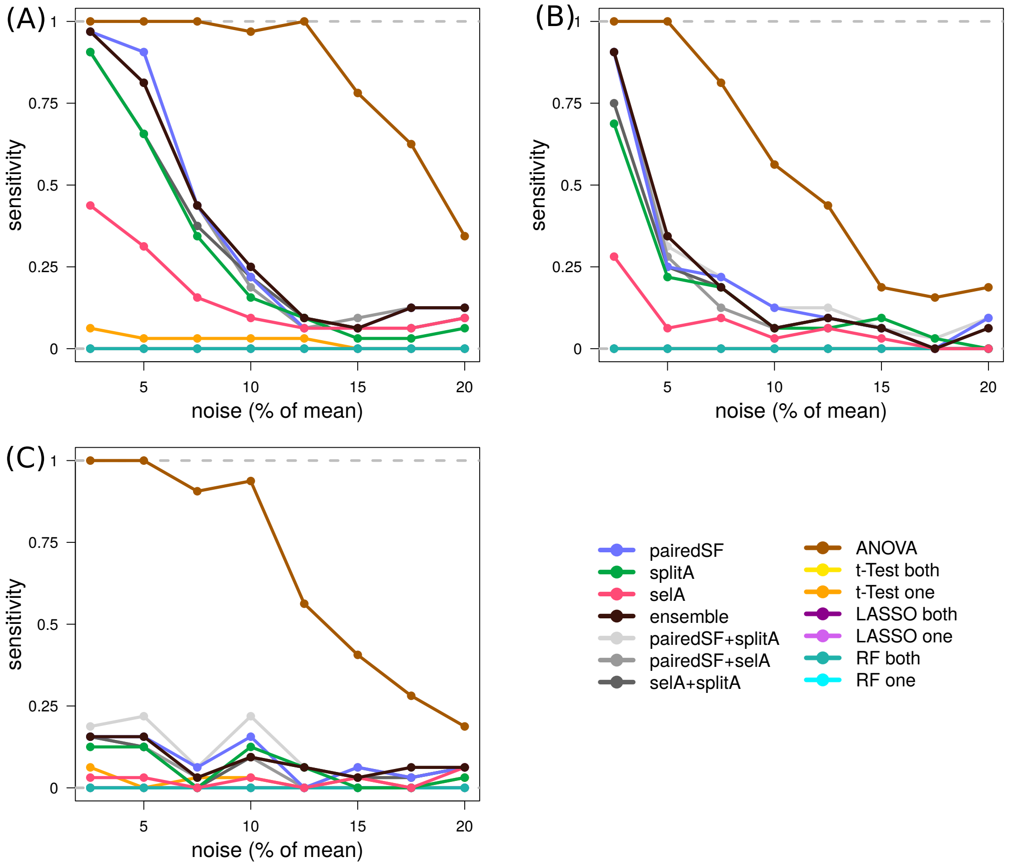

Figure 7: Sensitivity of methods based on simulations of XOR epistasis. Sensitivity was measured by the proportion of simulated interactions for which its p-value was below the lower 0.5-percentile of p-values (i.e. 99.5% of p-values were higher). (A) and (B) show the performances of the methods on simulations of pure XOR epistasis with effect sizes 1 and 0.5, respectively (scenarios EX1 and EX2). The RF-based methods were clearly outperformed by the exhaustive ANOVA. (C) Surprisingly, adding marginal effects to both interacting markers (thereby enabling their selection in the RF, scenario M9), did not improve the ability of the RF-based methods to detect XOR epistasis.

approaches. However, it did surprisingly well considering the relative naivety of this approach. The naivity lies in the fact that counting the co-occurrences of two markers in CARTs did not take into account whether or not these markers actually appear in the same path of the CARTs. Extensive efforts to improve this approach by counting these co-occurrences for CART paths rather than per tree were unsuccessful (results not shown).

3.2.2 Optimization of the RF Parameter mtry

The RF parameter 'mtry' determines the number of predictors that are evaluated at each split. It may be varied based on the data in order to optimize the predictive power of the model, but it has been shown that the default value commonly leads to the best result (Liaw and Wiener, 2002; Díaz-Uriarte and De Andres, 2006). However, in the context of epistasis detection, the predictive power is not necessarily the best criterion to choose the optimal mtry value, but rather the ability to detect interactions. For instance, if a smaller number of markers is evaluated at each split (lower mtry), markers with small or even absent marginal effects have a higher chance of being selected 'by chance'. Consequently, this could improve the ability of RF to capture XOR interactions. Therefore, the performances of the RF-based methods using five different mtry settings were evaluated based on the simulated data, and compared to the exhaustive ANOVA. The mtry values tested were $n/10$, $n/5$, $n/3$ (the default), $n/2$, and n (the total number of predictors). For ease of representation, the results of the 32 iterations of each simulation scenario were combined, which enabled the computation of ROC, PR, as well as the areas under them (AUROC and AUPR). The latter two were used to evaluate the performances at each mtry setting (Figure 8). However, there was no difference between the results for each mtry setting. Therefore, the default setting of mtry $= n/3$ was kept.

3.2.3 Effect of Marker-Marker Correlation

In the previous implementation (Grossbach, 2015) only marker pairs that had an absolute correlation coefficient R ≤ 0.3 were tested, while no such cutoff was applied to the other methods (i.e., the exhaustive ANOVA). This cutoff may have an effect on the performances of the RF-based methods, and was chosen arbitrarily to avoid false positive interactions between markers in LD. Therefore, the RF-based methods as well as the exhaustive ANOVA were applied using four different R cutoff values (0.3, 0.6, 0.9, and 1). Changing the cutoff had next to no effect on epistasis detection performance (results not shown). Since highly correlated markers are unlikely to represent true interactions, marker pairs with an absolute correlation cutoff of $R \geqslant 0.9$ were excluded in the following.

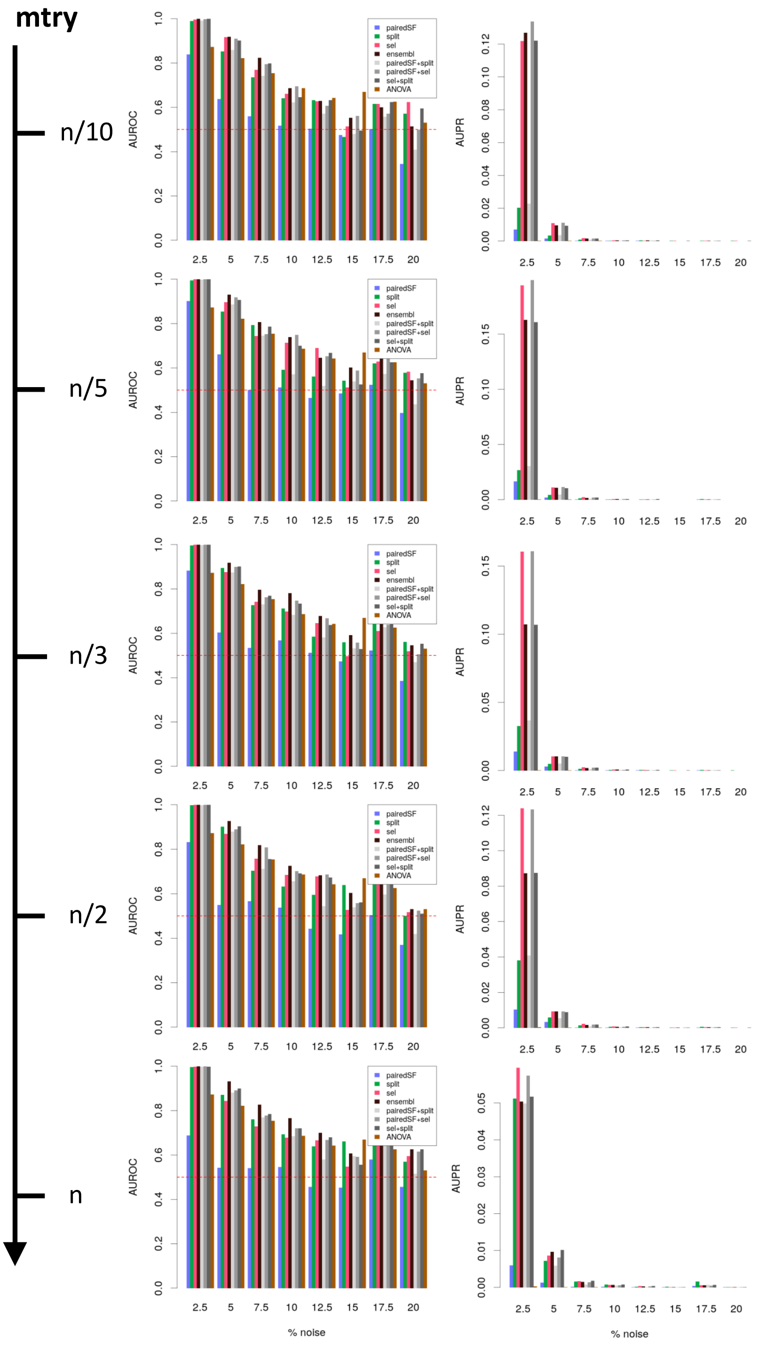

Figure 8: Effect of varying the RF parameter mtry on epistasis detection performance. Evaluation is based on the simulated data. Area under receiver operating characteristic (AUROC,left panel) and area under precision recall curve (AUPR, right panel) were created by merging the results of the 32 simulations of each scenario, for ease of representation. The used mtry setting is indicated for each row. Within the graphs, the bars represent AUROC and AUPR at varying noise levels for each method.

3.3 Application to eQTL Data

3.3.1 Benchmark Based on DKO Dataset

Simulated data is helpful in understanding the strengths and weaknesses of methods, but is never sufficient for the conclusive evaluation of the applicability of methods (discussed in Section 4.2, p. 35). Therefore, the methods were also tested on two different real datasets. Since eQTL mapping represents a common objective for QTL analyses, an eQTL dataset derived from a yeast cross (RM×BY, data unpublished) was selected as the first dataset.

In this benchmark, only the RF-based approaches and the exhaustive ANOVA were applied. These methods were used to find interactions between markers influencing the expression of 1,050 essential genes (i.e., genes that are lethal when knocked-out). The performance was measured by the ability to recover epistatic interactions detected in DKO experiments (Section 2.1.4, p. 12, Costanzo et al. 2016), as previously proposed (Picotti et al., 2013). The ROC, PR, AUROC, and AUPR of the different methods are shown in Figure 9. They were created together with Jan Grossbach.

For all tested methods the AUROC and AUPR were low, although significantly above random (p-value $< 2 \times 10^{-4}$ for all methods). A perfect performance (i.e., AUROC=1) would be impossible in this benchmark, because of two main reasons (also discussed in Section 4.2, p. 35). For one, the reference data (the DKO study) measures survival, while gene expression was analysed here. And second, the DKO data entails a different type of genetic perturbation (whole gene knock-outs versus segregating small 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. All RF-based approaches outperformed the ANOVA based on the AUROC and the AUPR. In contrast to the simulation results, the split asymmetry slightly outperformed the ensemble method.

3.3.2 Comparison to Database Networks

Because of the overall poor performance of the tested methods in the eQTL benchmark based on the DKO data, two different performance evaluation procedures were conceived. For one, a different dataset with a phenotype more similar to the DKO dataset was used (Section 3.4, p. 31). And also, gene interaction databases were used to evaluate the functional similarity between interacting loci. Because the databases integrate several sources of evidence, they were assumed to be more complete and accurate than just one single dataset (e.g., the DKO dataset). In addition, by creating networks from the database interactions, epistatic pairs can be evaluated by the distances of

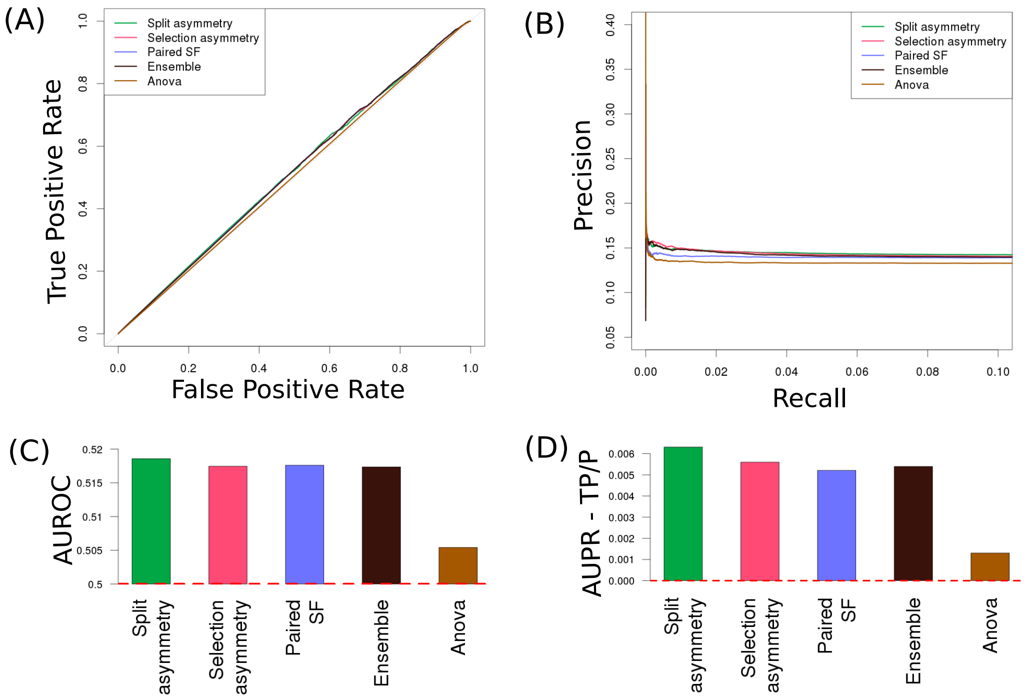

Figure 9: Performance on eQTL data. Performance was evaluated by the ability of the evaluated methods to correctly classify interacting and non-interacting genes, using double knock-out dataset as a gold standard. Shown are the differences between the results of the tested methods and random assignment (red dashed line) for (A) receiver operating characteristic (ROC), (B) precision-recall (PR) curve, (C) area under ROC (AUROC), and (D) area under PR (AUPR). The values expected under random assignment were subtracted from AUROC and AUPR. The RF-based methods outperform the ANOVA, while the split asymmetry (splitA) approach has the best recovery and precision. eQTL: expression quantitative trait locus.

the respectively assigned genes in such networks. Gene-gene distances were determined as the shortest path distances in networks derived from interactions from the STRING and Biogrid databases (Section 2.1.5, p. 13). For each transcript and for each method, the most significant, or the top 10 significant interactions according to q-value (FDR) were identified, respectively. The markers were then mapped to genes (Section 2.1.2, p. 10), so that there was a set of gene-gene interactions for each target transcript and each method. Subsequently, the mean and minimum distance, respectively, were calculated per target and method. The results for the Biogrid and STRING networks are shown in Figures 10 and 11, respectively. Surprisingly, because of the neglectable differences between the methods, this benchmark was not conclusive. The Biogrid network seemed to be densely connected, so that gene-gene distance ranged only from one to three, which hampered the comparison of methods. The results for the STRING database, however, were too similar between the different methods and non-consistent for different approaches, so that no meaningful conclusions could be drawn.

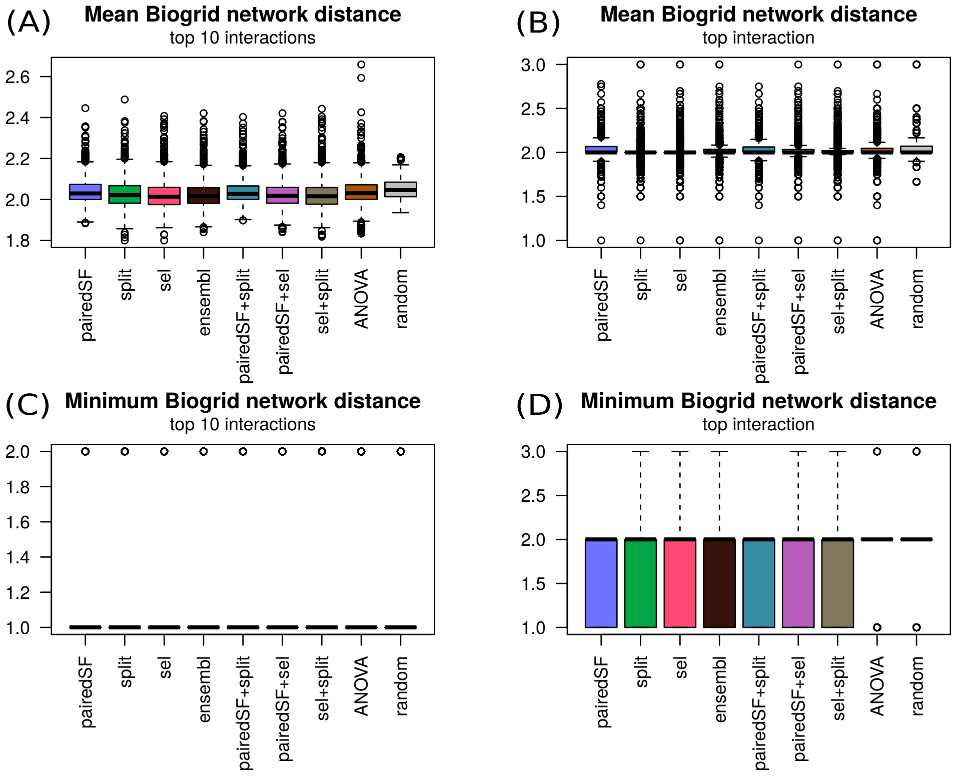

Figure 10: Performance on eQTL data based on Biogrid network. Performance was evaluated by the mean (A and B) or minimum (C and D) Biogrid network distance of genes assigned to the most significant (B, D) or the ten most significant (A, C) marker pairs, respectively. The 'random' group represents results that were produced by randomly selecting 'significant' marker pairs for each target. eQTL: expression quantitative trait locus.

3.4 Benchmark on Growth Trait Data

As suggested above, the phenotypes investigated in the eQTL dataset differed substantially from the phenotype investigated in the DKO dataset (expression of essential genes versus survival), which was suspected to be one reason for the overall small congruence between identified interactions and gold standard. Therefore, another dataset was selected where several growth traits were measured (Section 2.1.3, p. 11, Wilkening et al. 2014), which were expected to show a higher concordance with the DKO dataset. Furthermore, since these growth phenotypes are more complex than gene expression, more interactions are expected for them than for gene expression.

The RF-based methods, the exhaustive ANOVA, as well as the different two-

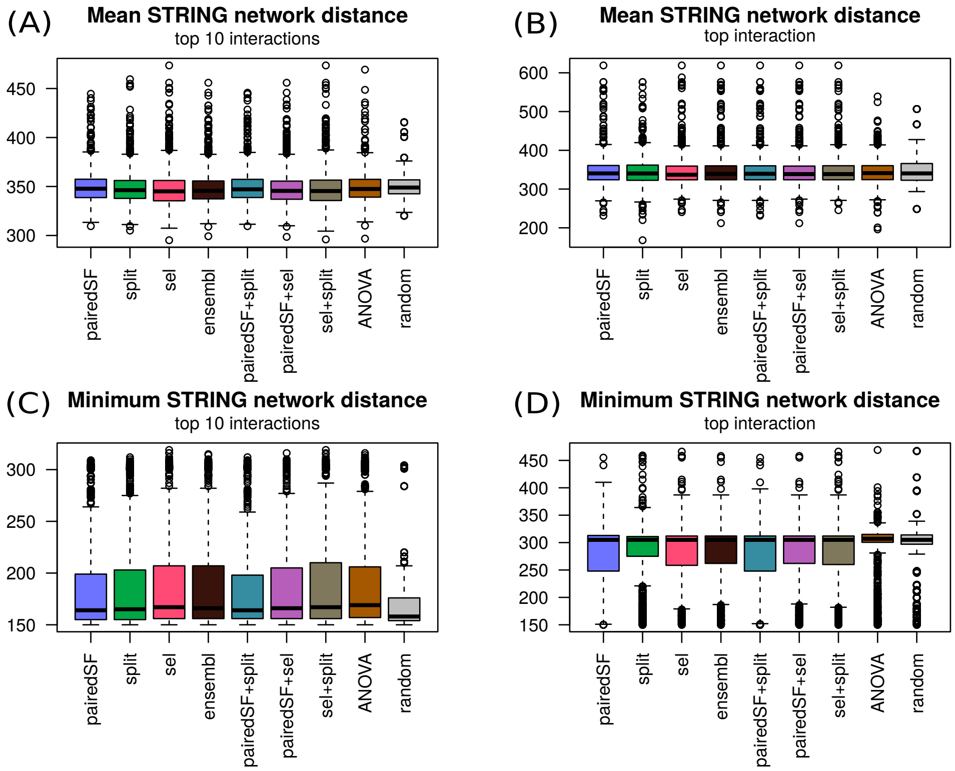

Figure 11: Performance on eQTL data based on Biogrid network. Performance was evaluated by the mean (A and B) or minimum (C and D) STRING network distance of genes assigned to the most significant (B, D) or the ten most significant (A, C) marker pairs, respectively. The 'random' group represents results that were produced by randomly selecting 'significant' marker pairs for each target. eQTL: expression quantitative trait locus.

step approaches were applied to the dataset and their results evaluated based on either (i) the recovery of interactions from the DKO dataset (Section 2.1.4, p. 12, Costanzo et al. 2016), or (ii) based on co-occurrence of respective genes in Reactome pathways (Section 2.1.5, p. 13). The methods were applied to identify interactions for all phenotypes that were measured in this dataset, including discrete and quantitative traits. Here, only the results for the two phenotypes 'Cantharidin resistance' and 'fitness on agar' are shown. Because resistance to Cantharidin is a Mendelian trait and therefore no interactions are possible, it was treated as a negative control. Fitness on agar, on the other hand, was the phenotype that most closely resembled the conditions of the DKO study, so that the results for this phenotype should show a higher overlap with the DKO interactions. The benchmark on the DKO data are shown in Figures 12 and 13 for Cantharidin resistance and fitness on agar, respectively. The results for the benchmark on the Reactome pathways were equivalent and are therefore not shown.

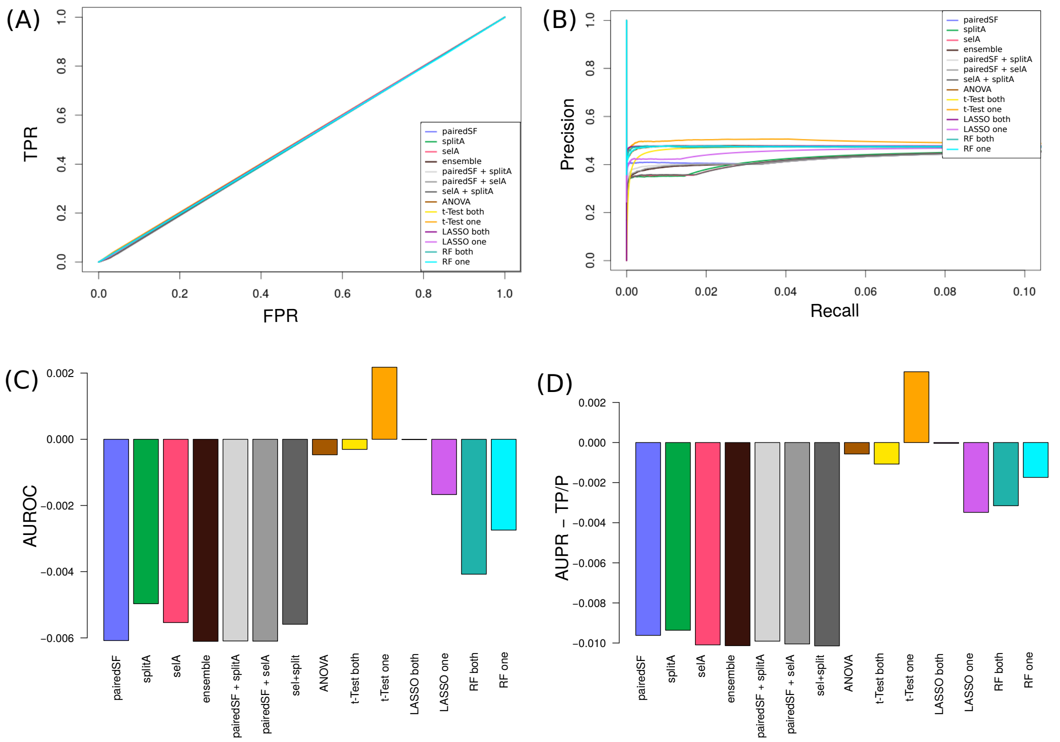

Figure 12: Meaningfulness of interactions for Cantharidin resistance, based on DKO dataset. (A) receiver operating characteristic (ROC), (B) precision recall (PR) curve, (C) area under ROC (AUROC), and (D) area under PR (AUPR) are shown for Cantharidin resistance. The values expected under random assignment were subtracted from AUROC and AUPR. Marker pairs corresponding to gene pairs that were found to interact in the double knock-out (DKO) dataset were treated as true positives. Since Cantharidin resistance is a Mendelian trait in yeast, no meaningful interactions were expected.

The performances of all methods were very similar to what would be expected from random assignment, and the differences between methods were extremely slight. Unlike what was expected, switching to a dataset with growth phenotypes did not increase the recovery of interactions from the DKO dataset. Possible reasons for this are discussed in Section 4.2 (p. 35). However, some trends about the methods' performances can be concluded. The PR and ROC measures indicate that the interactions identified for the negative control 'Cantharidin resistance' were not meaningful. Indeed, AUROC and AUPR were even slightly negative when the areas expected for random assignment were substracted (Figures 12 C and D). In contrast, for the phenotype 'fitness on agar' the interactions of the RF-based methods seemed to be more biologically meaningful than the results from the other methods. As for the benchmark on eQTL data, and opposed to the simulated data, the selA and splitA approaches performed better than the ensemble methods. The t-test-based

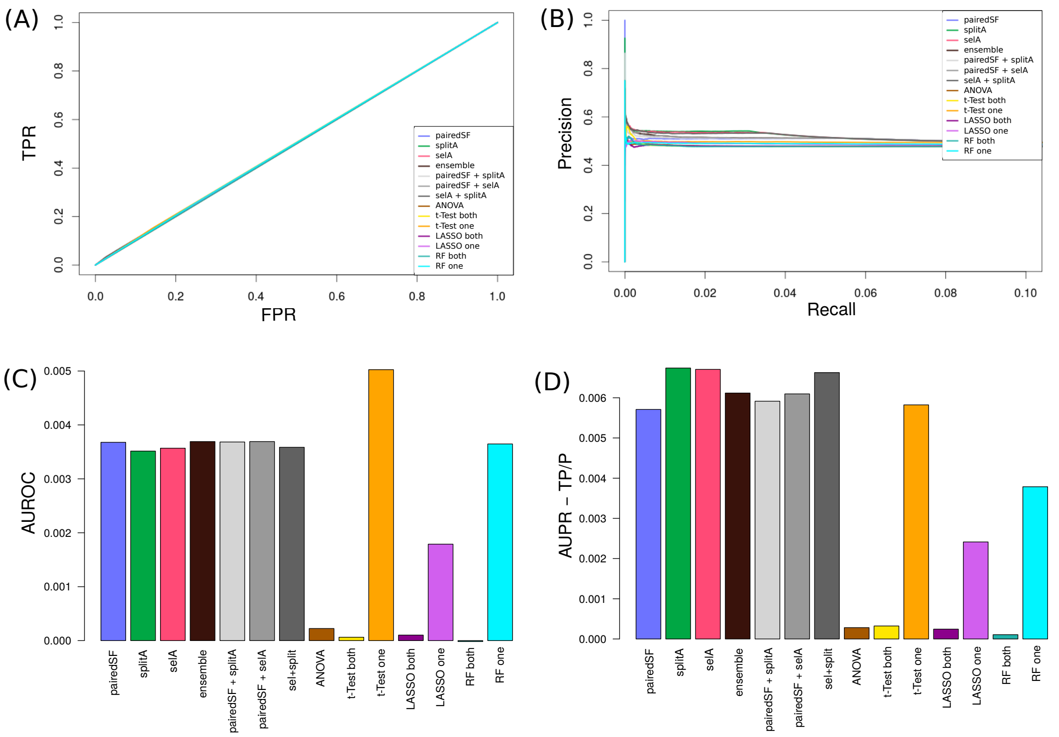

Figure 13: Meaningfulness of interactions for fitness on agar, based on DKO dataset.(A) receiver operating characteristic (ROC), (B) precision recall (PR) curve, (C) area under ROC (AUROC), and (D) area under PR (AUPR) are shown for fitness in agar. The values expected under random assignment were subtracted from AUROC and AUPR. Marker pairs corresponding to gene pairs that were found to interact in the double knock-out (DKO) dataset were treated as true positives. This phenotype was very similar to the phenotype investigated in the DKO dataset.

two-step approach also performed considerably well. However, the AUPR for the RF-based methods was larger. In this benchmark, the PR was more informative than the ROC. Thus, this benchmark represents supportive evidence for the applicability of the RF-based methods for real data. It furthermore implicates that the RF-based approaches outperform the other methods for epistasis detection. It should be noted that the exhaustive ANOVA and all two-step approaches investigating ‘both’ interactions performed particularly poorly in this benchmark. However, in contrast to the simulated data (Section 3.2.1, p. 22), the LASSO- and RF-based two-step approaches testing for marker pairs with ‘one’ marginal effect, were able to detect meaningful interactions in this benchmark. This supports the assumption that the pre-selection criteria were not appropriate for the simulated data.

4 Conclusion and Outlook

4.1 Results Summary

The three RF-based approaches, namely splitA, selA, pairedSF, and their combinations were applied to simulated data and to two real datasets. They were compared to an exhaustive ANOVA, as well as to several two-step approaches (Section 3.1, p. 21).

In the application to simulated data, the RF-based approaches, in particular the ensemble approach combining all three of them, outperformed the other methods in almost all simulation scenarios, with the exception of simulation scenarios with XOR epistasis. For these, the RF-based approaches were unable to identify the truly interacting markers, as opposed to some of the other methods. Varying the RF parameter mtry, or changing the correlation cutoff that was used to exclude marker pairs, had no effect on the performance of the methods.

The methods were also evaluated on real data. For the identification of eQTL interactions, the RF-based approaches outperformed the exhaustive ANOVA with respect to the ability to recover interactions identified in the reference DKO dataset. However, the overall low overlap between reference dataset and the investigated dataset allowed no definitive comparison of the method's performance. A benchmark based on database interactions was also not conclusive.

Finally, all methods were applied to a second real dataset with growth traits. The overall performance at recovering interactions from the DKO dataset was low and the differences between the methods minor. However, the RF-based approaches tended to give results that were biologically more relevant than the other methods.

4.2 Limitations of the Study

Simulated data always has to be treated with caution. They only represent a model of the real world, and by definition models are incomplete. Even if all phenomena of interest are included in the simulation, the complexity of a real biological setting cannot be represented (Wooley et al., 2005). In addition, the way the data is modelled is always biased by the expectations of the investigator, which might not apply to real data. The main advantage of simulations is the fact that the underlying 'truth' is known, so that it is straightforward to evaluate the performance of different methods. Yet, simulated data should not be overrated as a means to make absolute conclusions about the applicability of the methods, but rather as a way to understand in which cases, and why the methods will, or will not work.

The reverse applies to real data: obviously it will automatically represent actual biology, but because the underlying ‘truth’ is unknown, it is hard to determine whether a method was successful at uncovering real interactions or not. As a consequence, when comparing several methods, authors often refrain from using real data and restrict themselves to simulated data (Shang et al., 2011; García-Magariños et al., 2009; Yoshida and Koike, 2011; Wan et al., 2010). However, the evaluation of methods based on real data is critical to assess their performance. In this respect, this thesis, like its predecessors (Picotti et al., 2013; Grossbach, 2015) stands out against other studies comparing several interaction methods. In order to use real data, one has to rely on prior knowledge in the form of previous publications or database information as a gold standard. Doing so encompasses several problems. For one, it depends on the completeness and correctness of the reference dataset, which is unlikely to ever be absolute. For instance, existing databases are limited to interactions that were identified using the respective conventional methods. Therefore, this might favour methods that simply reproduce prior knowledge, and place methods which are able to uncover novel associations at a disadvantage.

In this thesis, the differences in characteristics between the reference datasets were a substantial problem. For the DKO benchmark, deviations between environmental factors such as lab conditions, variations between the used yeast strains, and the diverging investigated phenotypes were bound to induce non-reproducibility between datasets. In addition, small genetic variations were investigated in this thesis, whereas the reference dataset from Costanzo et al. (2016) encompassed whole gene knock-outs. Consequently, the markers had to be mapped to nearby genes. Intergenic variants may not have been treated appropriately, considering the fact that eQTL can act in ‘trans’ in yeast, i.e. the loci might be remote from the genes whose expression they affect (Clement-Ziza et al., 2014; Brem et al., 2002; Ackermann et al., 2013). Furthermore, the interacting loci might not necessarily act through the expression of two genes, but rather through epigenetic factors such as DNA methylation, or histone modifications.

Considering all these factors, it is not surprising that the interactions identified by the various methods in this study only partially recovered known gene-gene interactions. In fact, this study was not aimed at completely reconstructing the findings of previous knowledge, but simply to use these to evaluate and compare the different methods.

4.3 Conclusive Evaluation of the Methods

In the real data benchmarks in this thesis, the recovery of known interactions was very low for all methods, so that these results cannot offer an exhaustive conclusion about the performance of the methods. Further tests, especially on real data, are needed to conclusively determine which methods perform best at the detection of genetic interactions (Section 4.4, p. 37). However, taking together the results on the simulated data and both real data benchmarks, several conclusions can be drawn. It is important keep in mind that this thesis did not evaluate all existing epistasis detection methods. Therefore, all conclusions about the methods' performances can only be considered in comparison to the approaches evaluated in this method.

The three RF-based approaches presented in this thesis, and their combinations, are able to detect epistasis from genetic association data. Particularly the ensemble approach, which combines all three RF-based approaches seemed to perform relatively well at this task. It generally achieved the highest precision for the simulated data. When applied to real data, it seemed to produce the biologically most meaningful results, compared to the other tested methods.

As stated previously, the three RF-based approaches complement each other for different interaction scenarios in theory (Section 2.2.5, p. 17) and in practice (Section 3.2.1, p. 22). The pairedSF approach, for example, although it did not perform as well as the other two approaches for AND-epistasis, was more likely to detect XOR-epistasis, although not as well as the exhaustive ANOVA. Additionally taking into to account the tendency of the three-way ensemble to outperform the two-way combinations of RF-based approaches, exploiting all three approaches seems to be useful.

The RF algorithm offers many advantages over other modelling techniques considering genetic association studies, as discussed in Section 1.3 (p. 5). Its ability to account for interactions likely represents a considerable factor leading to the superior performance at detecting QTL compared to other methods. However, so far these interactions could only be *captured*, but not *detected* (Wright et al., 2016). Therefore, the approaches for the identification of interactions within the forest presented in this thesis represent a valuable enhancement of the applicability of RF for genetic association analyses.

4.4 Outlook

As outlined in the previous Section, further investigations are needed to elucidate which approach might be best for the identification of genetic interactions. The benchmarks applied in this thesis mostly investigated the relationship between the interacting loci. Considering the overall poor performance of all methods in this thesis, this approach is not absolutely adequate to evaluate the methods' performance. An alternative would be to focus on the plausibility of the association of the interacting markers with the investigated phenotype. For example, in the context of eQTL mapping, one could investigate the connection between the target transcript and the genes assigned to the two interacting markers. In essence, all analyses of real data performed in this study could be repeated with the focus on the plausibility of the regulator-target association instead of the regulator-regulator association. This is in line with the main objective of QTL mapping, namely to discover meaningful relationships between genetic factors and phenotypes.

The underlying objective when studying epistasis is the expectation that this will help to explain some of the missing heritability observed for many biological traits. Bloom and colleagues (2013; 2015) set about this task by studying very large yeast crosses. They estimated the contribution of interaction effects from the difference between narrow-sense and broad-sense heritability and claimed that the contribution of genetic interactions to QT variation is less than that of additive effects (9% versus 43% on average, respectively). The pairwise interactions identified by them, however, could explain only a minority of this missing heritability. This could mean that their approach was unable to exhaustively identify all pairwise interactions, or that higher-order interactions play a substantial role. Since RF is not limited in the order of interactions it can model, the latter issue could be tested by investigating the discrepancy between heritability that can be explained by RF and by conventional approaches that only account for two-way interactions. In addition, the RF-based epistasis detection approaches presented in this thesis might be able to identify additional two-way interactions, ultimately elucidating a larger proportion of the missing heritability.

As stated above, one of the advantages of RF is its ability to account for higher-order interactions. Therefore, the question arises whether these can also be extracted from the forest. Indeed, the approaches presented in this thesis can be extended from two-way epistasis to higher-order interactions. However, the search space for interaction grows exponentially with the order of interaction that is investigated, making an exhaustive search for three-way interactions infeasible. The feature selection properties of RF removes part of this multiple testing burden. In addition one could also exploit the fact that a three-way AND-type interaction will also lead to detectable two-way interactions between all pairwise combinations of the involved loci. One could restrict tests for three-way interaction to such 'interaction triangles'.

While studies on yeast represent a valuable resource for the investigation of the influence of genetic factors on complex traits, it remains a relatively simple model organism. Efficient implementations of RF that are applicable on a genome-wide scale were developed (Wright and Ziegler, 2015) and RF was previously used for GWAS in humans (Botta et al., 2014; Chen and Ishwaran, 2012). In order for the approaches presented here to be applicable to more complex species including humans, they have to be adapted to work on diploid genotypes. This requires taking dominance effects into account and leads to a higher complexity of possible interaction models. The former can be solved by treating the genetic markers as discrete factors with three levels. The latter makes the actual testing for interactions much more complicated. The pairedSF approach is not affected by the change to a diploid model and therefore would be straightforward to extend to other species. The splitA and selA approaches could also be applicable for some diploid interaction types, but this would require some fundamental modifications to their implementations.

Finally, the approaches presented in this thesis using RF for the detection of epistasis represent a valuable resource for various scientific applications. It would be interesting to investigate which novel genetic associations can be discovered using RF. Furthermore, any study that uses RF for modelling might profit from being able to extract information about the relationships between the predictors.

References

Ackermann, M., Clément-Ziza, M., Michaelson, J. J., and Beyer, A. (2012). Teamwork: Improved eQTL Mapping Using Combinations of Machine Learning Methods. *PLoS ONE*, 7(7):e40916.

Ackermann, M., Sikora-Wohlfeld, W., and Beyer, A. (2013). Impact of natural genetic variation on gene expression dynamics. *PLoS Genetics*, 9(6):e1003514.

Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., et al. (2010). Genome-wide association study of 107 phenotypes in a common set of Arabidopsis thaliana inbred lines. *Nature*, 465(7298):627–631.

Aylor, D. L., Valdar, W., Foulds-Mathes, W., et al. (2011). Genetic analysis of complex traits in the emerging Collaborative Cross. *Genome Research*, 21(8):1213–1222.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

Bergen, S. E. and Petryshen, T. L. (2012). Genome-wide association studies (GWAS) of schizophrenia: does bigger lead to better results? *Current opinion in psychiatry*, 25(2):76–82.

Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T.-L. V., and Kruglyak, L. (2013). Finding the sources of missing heritability in a yeast cross. *Nature*, 494(7436):234–237.

Bloom, J. S., Kotenko, I., Sadhu, M. J., et al. (2015). Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. *Nature Communications*, 6:8712.

Botta, V., Louppe, G., Geurts, P., and Wehenkel, L. (2014). Exploiting SNP Correlations within Random Forest for Genome-Wide Association Studies. *PLoS ONE*, 9(4).

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Brem, R. B. and Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 102(5):1572–1577.

Brem, R. B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568):752–755.

Buckler, E. S., Holland, J. B., Bradbury, P. J., et al. (2009). The genetic architecture of maize flowering time. *Science*, 325(5941):714–718.

Carlborg, O. and Haley, C. S. (2004). Epistasis: Too often neglected in complex trait studies? *Nature Reviews Genetics*, 5(8):618–625.

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., et al. (2017). The BioGRID interaction database: 2017 update. *Nucleic Acids Research*, 45(D1):D369–D379.

Chen, X. and Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, 99(6):323–329.

Clement-Ziza, M., Marsellach, F. X., Codlin, S., et al. (2014). Natural genetic variation impacts expression levels of coding, non-coding, and antisense transcripts in fission yeast. *Molecular Systems Biology*, 10(11):764–764.

Costanzo, M., Baryshnikova, A., Bellay, J., et al. (2010). The genetic landscape of a cell. *Science*, 327(5964):425–431.

Costanzo, M., VanderSluis, B., Koch, E. N., et al. (2016). A global genetic interaction network maps a wiring diagram of cellular function. *Science*, 353(6306).

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

Dong, C., Chu, X., Wang, Y., et al. (2008). Exploration of gene–gene interaction effects using entropy-based methods. *European Journal of Human Genetics*, 16(2):229–235.

Díaz-Uriarte, R. and De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1):3.

Ehrenreich, I. M., Gerke, J. P., and Kruglyak, L. (2009). Genetic dissection of complex traits in yeast: insights from studies of gene expression and other phenotypes in the BYxRM cross. *Cold Spring Harbor Symposia on Quantitative Biology*, 74:145–153.

Ehrenreich, I. M., Torabi, N., Jia, Y., et al. (2010). Dissection of genetically complex traits with extremely large pools of yeast segregants. *Nature*, 464(7291):1039–1042.

Evans, D. M., Marchini, J., Morris, A. P., and Cardon, L. R. (2006). Two-stage two-locus models in genome-wide association. *PLOS Genetics*, 2(9):e157.

Fabregat, A., Sidiropoulos, K., Garapati, P., et al. (2016). The Reactome pathway knowledgebase. *Nucleic Acids Research*, 44(D1):D481–487.

Fisher, R. A. (1925). *Statistical methods for research workers.* Genesis Publishing Pvt Ltd.

Flint, J. and Mackay, T. F. C. (2009). Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Research*, 19(5):723–733.

Freitag, C. M. (2007). The genetics of autistic disorders and its clinical relevance: a review of the literature. *Molecular Psychiatry*, 12(1):2–22.

García-Magariños, M., López-de Ullibarri, I., Cao, R., and Salas, A. (2009). Evaluating the ability of tree-based methods and logistic regression for the detection of SNP-SNP interaction. *Annals of Human Genetics*, 73(3):360–369.

Goudey, B., Rawlinson, D., Wang, Q., et al. (2013). GWIS — model-free, fast and exhaustive search for epistatic interactions in case-control GWAS. *BMC Genomics*, 14 Suppl 3:S10.

Grady, B. J., Torstenson, E. S., McLaren, P. J., et al. (2011). Use of biological knowledge to inform the analysis of gene-gene interactions involved in modulating virologic failure with efavirenz-containing treatment regimens in ART-naïve ACTG clinical trials participants. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 253–264.

Grossbach, J. (2015). *Detection of epistasis and eQTL mapping with Random Forest.* Master thesis.

Hannum, G., Srivas, R., Guénolé, A., et al. (2009). Genome-wide association data reveal a global map of genetic interactions among protein complexes. *PLOS Genetics*, 5(12):e1000782.

Harold, D., Abraham, R., Hollingworth, P., et al. (2009). Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nature Genetics*, 41(10):1088–1093.

Hill, W. G. (2010). Understanding and using quantitative genetic variation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 365(1537):73–85.

Johnsen, A. K., Valdar, W., Golden, L., et al. (2011). Genome-wide and species-wide dissection of the genetics of arthritis severity in heterogeneous stock mice. *Arthritis & Rheumatism*, 63(9):2630–2640.

Kang, H. M., Zaitlen, N. A., Wade, C. M., et al. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723.

Li, W. and Reich, J. (2000). A complete enumeration and classification of two-locus disease models. *Human Heredity*, 50(6):334–349.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R news*, 2(3):18–22.

Mackay, T. F., Stone, E. A., and Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*, 10(8):565.

Mackay, T. F. C. (2014). Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nature Reviews Genetics*, 15(1):22–33.

Mackay, T. F. C., Richards, S., Stone, E. A., et al. (2012). The Drosophila melanogaster Genetic Reference Panel. *Nature*, 482(7384):173–178.

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature News*, 456(7218):18–21.

Manolio, T. A., Collins, F. S., Cox, N. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753.

Marchini, J., Donnelly, P., and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*, 37(4):413–417.

McCarthy, M. I. and Zeggini, E. (2009). Genome-wide association studies in type 2 diabetes. *Current diabetes reports*, 9(2):164–171.

McKinney, B. A., Crowe, J. E., Guo, J., and Tian, D. (2009). Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. *PLoS Genetics*, 5(3):e1000432.

McKinney, B. A., Reif, D. M., Ritchie, M. D., and Moore, J. H. (2006). Machine learning for detecting gene-gene interactions. *Applied Bioinformatics*, 5(2):77–88.

Michaelson, J. (2010). *Applications and extensions of Random Forests in genetic and environmental studies*. PhD thesis.

Michaelson, J. J., Alberts, R., Schughart, K., and Beyer, A. (2010). Data-driven assessment of eQTL mapping methods. *BMC Genomics*, 11(1):502.

Michaelson, J. J., Loguercio, S., and Beyer, A. (2009). Detection and interpretation of expression quantitative trait loci (eQTL). *Methods*, 48(3):265–276.

Moore, J. H., Gilbert, J. C., Tsai, C.-T., et al. (2006). A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of Theoretical Biology*, 241(2):252–261.

Moore, J. H. and Williams, S. M. (2009). Epistasis and its implications for personal genetics. *The American Journal of Human Genetics*, 85(3):309–320.

Nelson, M. R., Kardia, S. L. R., Ferrell, R. E., and Sing, C. F. (2001). A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Research*, 11(3):458–470.

Niel, C., Sinoquet, C., Dina, C., and Rocheleau, G. (2015). A survey about methods dedicated to epistasis detection. *Frontiers in Genetics*, 6.

Parks, B., Nam, E., Org, E., et al. (2013). Genetic control of obesity and gut microbiota composition in response to high-fat, high-sucrose diet in mice. *Cell Metabolism*, 17(1):141–152.

Pecanka, J., Jonker, M. A., Bochdanovits, Z., Vaart, V. D., and W, A. (2017). A powerful and efficient two-stage method for detecting gene-to-gene interactions in GWAS. *Biostatistics*, 00(00):1–18.

Phillips, P. C. (2008). Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11):855–867.

Picotti, P., Clément-Ziza, M., Lam, H., et al. (2013). A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature*, 494(7436):266–270.

Purcell, S., Neale, B., Todd-Brown, K., et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575.

Purcell, S. and Sham, P. C. (2004). Epistasis in quantitative trait locus linkage analysis: Interaction or main effect? *Behavior Genetics*, 34(2):143–152.

Ritchie, M. D. (2015). Finding the epistasis needles in the genome-wide haystack. *Methods in Molecular Biology*, 1253:19–33.

Ritchie, M. D., Hahn, L. W., Roodi, N., et al. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *American Journal of Human Genetics*, 69(1):138–147.

Robnik-Šikonja, M. and Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning*, 53(1-2):23–69.

Saito, T. and Rehmsmeier, M. (2017). Precrec: Fast and accurate precision–recall and ROC curve calculations in R. *Bioinformatics*, 33(1):145–147.

Schuldiner, M., Collins, S. R., Thompson, N. J., et al. (2005). Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell*, 123(3):507–519.

Shang, J., Zhang, J., Sun, Y., et al. (2011). Performance analysis of novel methods for detecting epistasis. *BMC Bioinformatics*, 12:475.

Shao, H., Burrage, L. C., Sinasac, D. S., et al. (2008). Genetic architecture of complex traits: Large phenotypic effects and pervasive epistasis. *Proceedings of the National Academy of Sciences*, 105(50):19910–19914.

Steinmetz, L. M. and Davis, R. W. (2004). Maximizing the potential of functional genomics. *Nature Reviews. Genetics*, 5(3):190–201.

Stephan, J., Stegle, O., and Beyer, A. (2015). A random forest approach to capture genetic effects in the presence of population structure. *Nature Communications*, 6:7432.

Sullivan, P. F. (2005). The genetics of schizophrenia. *PLOS Medicine*, 2(7):e212.

Szklarczyk, D., Morris, J. H., Cook, H., et al. (2017). The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research*, 45(D1):D362–D368.

Tong, A. H. Y., Lesage, G., Bader, G. D., et al. (2004). Global mapping of the yeast genetic interaction network. *Science*, 303(5659):808–813.

Visscher, P. M. (2008). Sizing up human height variation. *Nature Genetics*, 40(5):489–490.

Wan, X., Yang, C., Yang, Q., et al. (2010). BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *American Journal of Human Genetics*, 87(3):325–340.

Wang, M., Chen, X., and Zhang, H. (2010). Maximal conditional chi-square importance in random forests. *Bioinformatics*, 26(6):831–837.

Wei, C. and Lu, Q. (2014). GWGGI: software for genome-wide gene-gene interaction analysis. *BMC Genetics*, 15.

Wilkening, S., Lin, G., Fritsch, E. S., et al. (2014). An evaluation of high-throughput approaches to QTL mapping in Saccharomyces cerevisiae. *Genetics*, 196(3):853–865.

Wooley, J. C., Lin, H. S., and on Frontiers at the Interface of Computing and Biology, N. R. C. U. C. (2005). Computational modeling and simulation as enablers for biological discovery. In *Catalyzing Inquiry at the Interface of Computing and Biology*, pages 117–204. National Academies Press (US), Washington (DC).

Wright, M. N. and Ziegler, A. (2015). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *arXiv:1508.04409 [stat]*. arXiv: 1508.04409.

Wright, M. N., Ziegler, A., and König, I. R. (2016). Do little interactions get lost in dark random forests? *BMC Bioinformatics*, 17:145.

Yates, A., Akanni, W., Amode, M. R., et al. (2015). Ensembl 2016. *Nucleic Acids Research*, 44(D1):D710.

Yoshida, M. and Koike, A. (2011). SNPInterForest: A new method for detecting epistatic interactions. *BMC Bioinformatics*, 12:469.

Appendix

Abbreviations

ANOVA	ANalysis Of VAriance
AUPR	area under precision-recall curve
AUROC	area under receiver operating characteristic
CART	classification and regression tree
DKO	double knock-out
DNA	desoxy-ribonucleic acid
eQTL	expression quantitative trait locus
FDR	false discovery rate
FPR	false positive rate
GWAS	Genome-wide association study
LASSO	Least Absolute Shrinkage And Selection Operator
LD	linkage disequilibrium
MAF	minor allele frequency
mRNA	messenger ribonucleic acid
OOB	out-of-bag
pairedSF	paired selection frequency
pQTL	protein quantitative trait locus
PR	precision-recall
QT	quantitative trait
QTL	quantitative trait locus
RF	Random Forest
RNA	ribonucleic acid
ROC	receiver operating characteristic
selA	selection asymmetry
splitA	split asymmetry
TPR	true positive rate
YPD	yeast extract peptone dextrose

List of Figures

List of Tables

Statutory Declaration

I hereby declare that I have developed and written the enclosed Master Thesis completely by myself, and have not used sources or means without declaration in the text. Any thoughts from others or literal quotations are clearly marked. The Master Thesis was not used in the same or in a similar version to achieve an academic grading. I assert that the enclosed electronic version of the Master Thesis corresponds to the printed version completely.

Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Schriften entnommen wurden, sind als solche kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.

Köln, 16. Nov, 2017

Corinna Lewis Schmalohr