University of Cologne

Detection of Epistasis and eQTL Mapping with Random Forest

Master’s thesis in

Biological Sciences

Submitted by Jan Großbach

Supervisor: Prof. Dr. Andreas Beyer

March 2015

Abstract

Gene expression is a process that underlies all biological functions. Understanding regulatory relationships between genes furthers our understanding of biological systems and is crucial to many medical and industrial applications.

In this work, a machine learning algorithm called Random Forest (RF) is used to explain variation in gene expression. Different approaches to assess the influence of genetic variation at a locus on the expression levels were compared by using them to map quantitative trait loci (QTL).

Methods that attempt to quantify the amount of variance explained by a predictor (PI and RSS) led to the detection of more QTL than a method that describes the importance of the predictors relative to each other (SF). These additionally detected QTL appear to be biologically plausible.

To identify pairs of predictors that explain variance in a nonlinear manner, called epistasis, a method that exploits a phenomenon called split asymmetry was developed. The method was tested on *in silico* and real data benchmarks. The method outperformed a two-dimensional approach on the real data benchmark.

The benchmarks presented in this work can be employed to test additional methods. The approaches that were applied can be used to explain variation in a wide array of traits.

Acknowledgements

I want to thank those people who made it possible for me to work on this interesting project and to write this thesis. I am grateful to Andreas Beyer who gave me the opportunity to work on this fascinating topic, supported me at every step of the way and always took time out of his busy day to answer any question. I especially want to thank Mathieu Clément-Ziza who went to great lengths to help me understand and explore the finer details of this work and who offered valuable insights on the current scientific field and on scientific work in general.

I’m thankful to Achim Tresch for supporting me by giving me the opportunity to broaden my knowledge in his group and co-supervising this work. Many thanks go to Michael Rauer whose suggestions helped me immensely with the writing.

I want to thank Johannes Stephan for sharing his knowledge of Random Forest and Jacob Michaelson who did some great research that is fundamental to my work. I’m thankful to Michael Kuhn for providing me with his implementation of Random Forest.

I want to express my gratitude to all members of the Beyer- and Tresch-groups who gave great suggestions and made my time working on this thesis very enjoyable and memorable.

Contents

1. Introduction

1.1 Background

A central question in biology is how genomic variation influences and to which degree it determines complex biological traits. A large proportion of biologically important traits can be classified as quantitative traits (QTs). Unlike Mendelian traits, QTs are distributed on a continuous scale. Examples of QTs are phentoypes like body-height, yield of crops or growth rates of microorganisms.

Understanding these traits is crucial for many medical, agricultural and industrial applications (Raadsma *et al.*, 2009; Shpak *et al.*, 2014; Urbany *et al.*, 2011; Veyrieras *et al.*, 2008). The identification of genetic and environmental factors, that influence a trait, can further the understanding of the underlying biological system. A genetic locus that contributes to a QT is termed a quantitative trait locus (QTL). Genetic variation at a QTL affects the expression of the trait in a quantitative fashion.

The development of high-throughput sequencing technologies and high-throughput methods for the measurement of protein concentrations has decreased the costs and increased the quality of large scale measurements of some molecular traits like gene expression. A genetic locus that influences the expression rate of a specific gene is called an expression-QTL (eQTL, reviewed in Rockman and Kruglyak, 2006). eQTL indicate a regulatory relationship between the expressed transcript and the genomic region that the QTL is located in. Examples for possible eQTL are transcription factors that directly regulate the gene of interest, genes that are responsible for post-translational modifications of those transcription factors and the promoter region of the gene of interest itself. Therefore, identification of eQTL can further the understanding of genetic regulatory networks (Ackermann *et al.*, 2013; Clement-Ziza *et al.*, 2014; Hemani *et al.*, 2014; Michaelson *et al.*, 2009; Suthram *et al.*, 2008; Veyrieras *et al.*, 2008). QTL can be found for many molecular traits such as protein-concentrations (pQTL) or the concentrations of metabolites (mQTL)(Cyr *et al.*, 2011; Holdt *et al.*, 2013; Khan *et al.*, 2012).

In the case of genomic loci affecting mRNA or protein concentrations in the cell, QTL can be classified as *cis* or *trans*. A *trans*-eQTL affects the expression of a gene via an additional molecule while *cis*-eQTL affect the expression of a gene without any intermolecular signaling. Typically *cis*-regulatory loci are located in the proximity of the gene they regulate. They function allele-specific only affecting loci on the same chromosome as them. Transcription factor binding motifs in promoters are examples of *cis*-acting regulators. *Trans*-eQTL function through molecular signaling and independent of physical distance. Their effect is not allele-specific and not limited to one chromosome. Transcription factors are typical examples of *trans*-acting regulators (reviewed in Albert and Kruglyak, 2015).

To understand the biological relevance of a QTL, its effects on the QT have to be investigated.

If a quantitative trait is affected by more than one QTL, the QTL can have fundamentally different relationships to each other. If their effects are affecting the phenotype completely independent of each other the QTL have an additive relationship. If the effect is dependent on the interaction of different genetic loci, it is called epistatic. The relationship between these loci is referred to as epistasis (reviewed in Mackay, 2013).

The concept of epistasis can be used as a tool to identify common members of a pathway by discovering a functional relationship between the respective genes (Jiang *et al.*, 2009; Michaelson, 2010; Urbanowicz *et al.*, 2012).

QTL can be identified by considering if genetic variance at a locus has influence on the phenotypical trait value. QTL studies have traditionally been performed with univariate methods like composite interval mapping or Haley-Knott-regression (Haley and Knott, 1992; Zeng, 1994). Univariate methods only test for an association between the trait and one locus per time and are therefore unable to accurately account for the effects of genetic interactions on a trait. Multivariate methods like LASSO or sparse partial least squares, on the other hand, are designed to consider the relationship between variables (Chun and Keles, 2009; Tibshirani, 1996). The better the contributions of QTL to the phenotype are understood, the more accurately they can be used as predictors to explain the phenotype. Multivariate methods are generally more successful at the detection of eQTL (Michaelson *et al.*, 2010). Numerous current studies still employ univariate methods (Holdt *et al.*, 2013; Nagtegaal *et al.*, 2012; Nelson *et al.*, 2013; Pannebakker *et al.*, 2011; Shpak *et al.*, 2014).

A heuristic method for supervised machine learning that has gained popularity over the last decade is Random Forest (Breiman, 2001). This method has been successfully used in the context of eQTL-detection (Clement-Ziza *et al*., 2014; Francesconi and Lehner, 2013; Michaelson *et al*., 2010). Random Forest is an ensemble learning method for prediction that is based on decision trees. Each tree offers an independent solution to the task of predicting the phenotype. To produce a final prediction all predictions from individual trees are combined. Forests that predict a qualitative phenotype (e.g. eye color) perform classification while forests that predict a quantitative phenotype (e.g. body height) perform regression. As a heuristic method, Random Forest has several random sampling steps that ensure that the trees in a forest are different from each other. Trees are grown with a training-set of individuals that is sampled with replacement independently for each tree, called bootstrap-sampling. The initial group of individuals is then divided according to their status regarding a predictor. In Random Forest the predictor that minimizes the variance within the two resulting subgroups the most is chosen for the split. The subgroups, or nodes, are split further until they contain fewer individuals than a defined threshold. The sampling of predictors that are considered to split a node is independent for each split. This further diversifies the trees and bases the prediction on a larger number of predictors (Breiman, 2001; Liaw and Wiener, 2002; reviewed in Michaelson, 2010).

The forests that are created as described above serve as nonparametric models to explain trait variance. QTL can be identified by assessing the importance of each predictor in the model. The different methods that are used to score the importance of predictors in a model are referred to as importance measures. In this work three importance measures are used to assess the contribution of each predictor to the forest. The first one is the decrease in predictive accuracy of the model upon permutation of a predictor. Predictive accuracy is assessed with the help of individuals that were not used to build the tree that makes the prediction. They are considered "out-of-bag" (OOB). Their phenotypes are predicted through the decision tree and theses predictions are compared to the actual trait values. The more important a variable is, the more the predictive accuracy of the model will suffer if the variable is randomly permuted and its ability to explain variance is removed. This measure for variable importance is termed the permutation importance (PI). A second way to assess variable importance is to examine the average increase in node purity that can be observed when the variable was used to split a node. In this work this referred to as the reduction in node variance (RSS). RSS sis computed with OOB-samples. Another way to quantify the importance of a variable in the forest is to count how often it was used to split nodes. This is referred to as the selection frequency (SF). Recent work found SF to be more sensitive for eQTL-mapping than other importance measures (Michaelson *et al.*, 2010).

Random forest was first introduced by Breiman in 2001 (Breiman, 2001). The original Fortran-implementation was adapted to the R-environment (Liaw and Wiener, 2002).

The use of Random Forest is not limited to the identification of QTL. It is also possible to analyze interactions between predictors. This is especially relevant to the detection of epistasis. Epistasis describes a relationship between two loci whose contributions to a phenotype are dependent on each other. Different forms of epistasis can occur. In AND-epistasis, individuals with a specific allele-combination are the target of the epistatic effect, changing their trait values (Fig. 1.1a). XOR-epistasis in contrast requires either one or the other but not both loci to have a specific allele (Fig. 1.1b) (reviewed in Michaelson, 2010).

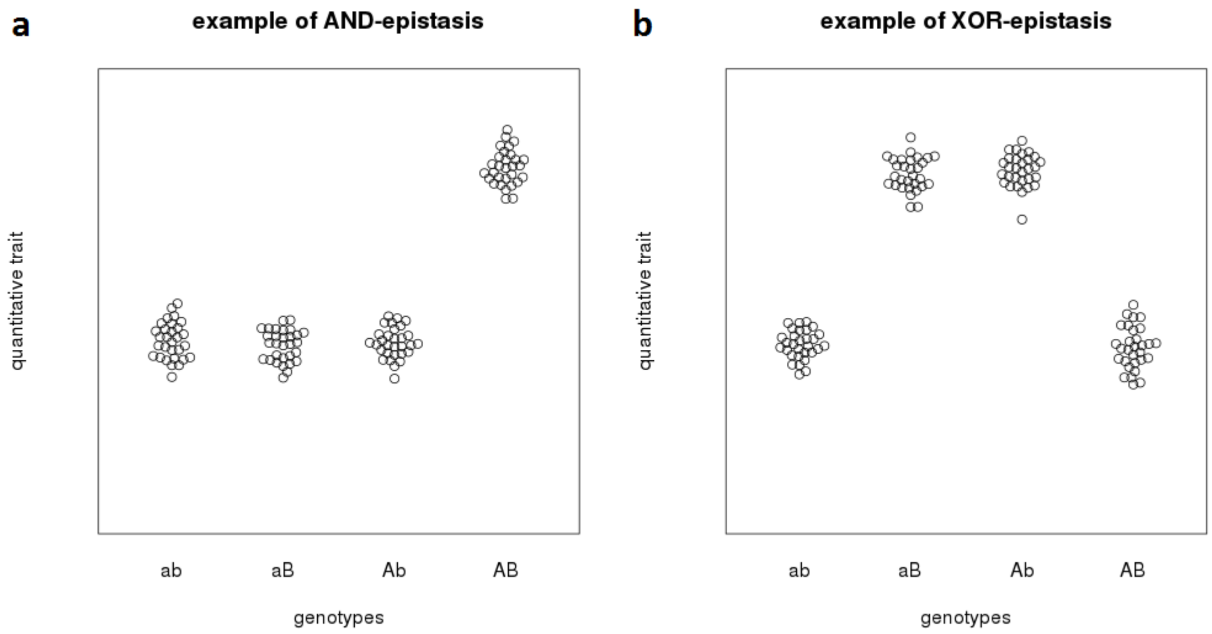

Figure 1.1: Simulated examples of AND-epistasis (a) and XOR-epistasis (b). Each dot represents one individual sample. The individuals are ordered according to their genotypes at two loci (a/A and b/B). In this example, the epistatic effect changes the trait values of individuals with the genotype AB in (a) and individuals with the genotypes aB and Ab in (b).

A concept used to identify epistatic interactions between predictors was introduced by Michaelson, termed split asymmetry (Michaelson, 2010). It aims at revealing predictors that interact by examining their performance when used on the same samples. If two predictors are in epistasis, the reduction in variance achieved by one would be dependent on the identity of the samples for the other predictor. In RF, this analysis is performed with the node-means of OOB-samples. It is performed on OOB- rather than IB-data because the latter is prone to overfitting. The score computed with the method described by Michaelson is dependent on factors including the correlation between predictors, main effects of the tested and not tested predictors and their allele frequencies. The pair-specific biases can be resolved with normalizations (Picotti *et al.*, 2013). The main effect of a predictor is its average effect on the outcome.

Epistasis between genomic loci affecting QTs is often investigated with two-dimensional methods. These methods only consider two predictors at a time and analyze the distributions of trait values for the different combinations of these genotypes for non-additive interactions. Contributions of any other predictor other than the two that are tested cannot be accounted for. This poses a problem in complex traits like body height in which a large number of loci contribute significantly to the phenotype (Willemsen *et al.*, 2004). Effects outside the tested predictors reduce the power of two-dimensional approaches, limiting their practical value.

1.2 Aim and Strategy

The first aim of this work is to propose a method for the detection of epistatic interactions that is robust to additive effects outside of the tested predictors. To be of high practical use the method does have to not be reliant on large null distributions. To achieve this goal we develop a workflow that collects the split performances of predictors for different subgroups of earlier splits with another predictor and compares them in an analytical way to detect pairs of predictors that show interdependence. We assess the performance of this method in simulated traits and real data and compare it to that of an exhaustive two-dimensional approach, which only considers two predictors at a time.

The second aim is to compare the suitability of different importance measures for eQTL-mapping. We use different importance measures and combinations of them to detect eQTL in a library of fission yeast strains. The identified eQTL are compared between the importance measures used for mapping them and evaluated for plausibility.

2. Methods

2.1 eQTL-mapping with a combination of permutation importance and reduction in node variance

2.1.1 Regression with Random Forest

Random Forest (RF) is a machine learning algorithm that attempts to predict outcomes based on a set of independent variables (predictors), using training data (Breiman, 2001). The training data is used to build decision trees that split the training set into increasingly homogeneous subgroups based on the available predictors. Each decision tree is constructed using an individual training-set that is a random subsample of all individuals in the data-set. The sampling is performed with replacement, resulting in trees that use on average about 63% of all samples.

The decision which predictor is used for splitting is made individually for each group of individuals (node). At each split a random sample of all predictors is tested regarding the variance it would reduce by splitting the node if chosen. The parameter that specifies the amount of predictors that are tested for their performance can have a large influence on the forest (Díaz-Uriarte and De Andres, 2006; Liaw and Wiener, 2002). If fewer predictors are sampled, predictors that explain smaller amounts of variance are chosen more often.

The predictor that reduces the variance the most from the predictor-sample is chosen to split the group. A prediction is made by assigning a combination of predictors to a final node. RF can be used to predict categorical and quantitative data, which is referred to as classification and regression. In the case of classification the prediction is the mode of the final node, in the case of regression it is the mean of the final node. The properties of the forests are determined by several parameters such as the minimal number of individuals for splitting, the total number of trees or the number of predictor that are sampled per split (Liaw and Wiener, 2002).

Since not all predictors are equally informative regarding the outcome, it is crucial to assess their individual importance. In RF this can be accomplished in multiple ways. The overall accuracy with which the forests predict an outcome can be assessed by comparing the prediction that was made to the true outcome. The smaller the difference between prediction and truth, the better the Forest describes the data. In forests that perform classification the prediction is either identical to the true outcome or not. In this case the predictive accuracy refers to the rate with which samples were placed in a final node whose mode is identical to the true outcome for this sample. Forests that perform regression predict outcomes on a continuous scale. The average of the difference between the mean of the node and the true outcome relative to the total variance over all individuals constitutes the predictive accuracy of forests that perform regression.

One way to assess the importance of predictors is to permute the predictor in question for the OOB-samples and to observe the degradation in predictive accuracy. After permutation, splits with the predictor in question group samples in a random manner. The degradation in predictive accuracy is dependent on how much variance is explained with each split and on the number of samples whose outcome is predicted with the help of the predictor. This measure is referred to as permutation importance (PI). If more splits happened before the predictor was used, the groups of samples that are split with the predictor are smaller. If the prediction of fewer samples is affected by the permutation, the decrease in predictive accuracy is smaller.

Another method to evaluate the importance of a single predictor is to compute how the variance in the subgroups compares to the variance in the group that was split with the predictor in question. This referred to as the reduction in node variance (reduction of the sum of squares, RSS). This importance measure is assessed using the individuals that were not part of the training set. Using OOB-data to assess the importance of a predictor is necessary because all splits were made when no better predictor was available for a specific set of samples. If the noise was distributed differently among a population of a given node, another predictor might be chosen. Using OOB-data to calculate the RSS gives an estimate of the performance of the predictor that is not as limited to a specific combination of samples and therefore more informative about the predictor itself. This importance measure does not directly take into account the position at which a marker was used in a tree.

A third measure that is informative about the importance of a predictor is the amount of times it was used in the forest in total. This is referred to as the selection frequency (SF) (Michaelson *et al.*, 2010). This measure does not consider the actual increase in node purity by using the predictor for splitting or the position in the tree when the predictor was used. The split of the first node has the same weight as a final split which affects much fewer individuals.

Since only IB-data affects the choice of predictor for a specific split, SF does not account for overfitting. In this context, overfitting refers to predictors being chosen for a split because they explain noise in a small group of samples. Other importance measures account for overfitting by using OOB-samples to calculate the scores.

2.1.2 RandomForestExtended

QTL-mapping with Random Forest relies on the ability of an implementation to grow forests consisting of as many trees as possible given the computational resources. Kuhn *et al.* released an R implementation of RF that vastly improves on the earlier implementation in R by Liaw *et al.* by enabling parallelization and the storage of predictors and other features of the forest as binary data (Kuhn, 2013; Liaw and Wiener, 2002). The method for the detection of epistasis by Michaelson *et al.* and our method, presented below, rely on the use of OOB-data to make predictions that are robust against overfitting. OOB-data is not part of the output in the Random Forest implementation by Liaw *et al.* but it is returned by the R implementation *rfdev* by Jacob Michaelson (Michaelson, 2010). In *RandomForestExtended* we combined features from *parallelRandomForest* and *rfdev* to enable the resource-efficient creation of large forests that include OOB-data in the output.

In addition to combining those useful features, *RandomForestExtended* contains a number of bug fixes both in the R and C++ code of the package.

To detect possible unintended differences between these packages, three forests of 25000 trees each were grown using each package. The forests aim to predict the same simulated trait with no major additive or epistatic effects and normally distributed noise. As a direct source of information about the average architecture of the trees, the selection frequency was computed for all forests. The Pearson-correlation of the selection frequencies was computed for all forests. If the packages select predictors based on the same criteria, the selection frequencies are expected to correlate regardless of the package that was used to grow the forest they are extracted from.

2.1.3 Fission yeast strain library

To compare the performance of different importance measures for QTL-mapping, RF was used to map eQTL in a fission yeast library (Clement-Ziza *et al.*, 2014). Here two parental strains (968 x Y0036) of *Schizosaccharomyces pombe* were crossed to produce 44 F2-crosses. Two closely related strains of fission yeast were chosen as parents for the cross because the resulting low genetic complexity of the library facilitates more powerful QTL-mapping. The parental strains were sequenced and the genotype of the offspring was determined through RNA-sequencing. Out of 4570 polymorphisms between the parental strains a set of 708 markers was chosen as predictors for mapping purposes. The gene expression of all strains was measured using RNA-sequencing and the expression of 6464 transcripts was measured. The expression levels were quality-filtered and normalized as described in Clement-Ziza *et al.* (Clement-Ziza *et al.*, 2014).

2.1.4 Mapping procedure

In the context of explaining variation in the magnitude of transcription, the trait that is to be predicted is the concentration of specific mRNAs. The set of predictors is made up by information about the alleles at genetic loci for each sample in the data-set, often single nucleotide polymorphisms (SNPs). Using RF to predict the gene expression from the available genetic information allows the evaluation of the contributions of individual polymorphic loci to the phenotype.

The procedure for growing the Random Forests used in this work was largely adapted from Clement-Ziza *et al.* (Clement-Ziza *et al.*, 2014).

For each transcript, 100 forests with 160 trees each were grown. In each forest, missing genotypes were replaced by random assignments to the allele responding to one of the parental strains individually. This strategy avoids a bias that is associated with using a third category for the genotype in addition to the parental alleles that describes missing values. In addition to the 708 genotypic markers, eight covariates representing the population structure were included in the set of predictors. This was done to avoid attributing phenotypic differences that are due to the population structure to specific markers. The selection frequency, the permutation importance and the reduction in node variance were extracted from each forest and averaged over all forests for a single trait.

To generate empirical null distributions for the importance measures permutations of one randomly selected quarter (1616) of all traits were mapped as well. Each of these traits was permuted and mapped a thousand times. The permutation was performed by randomly distributing the real phenotypic values among the individuals and thereby removing all associations between genetic variation and the trait. For each permutation a score for each marker was extracted, resulting in a total of 1,616,000 values per marker.

The distribution for importance measures per marker is dependent on genomic features such as the correlation to other markers and allele frequencies (Michaelson *et al.*, 2010; Strobl *et al.*, 2008). This prevents the pooling of null distributions over different markers. The marker-specific difference in the null distribution can be observed for SF and to a lesser degree for PI and RSS (Fig. 2.1).

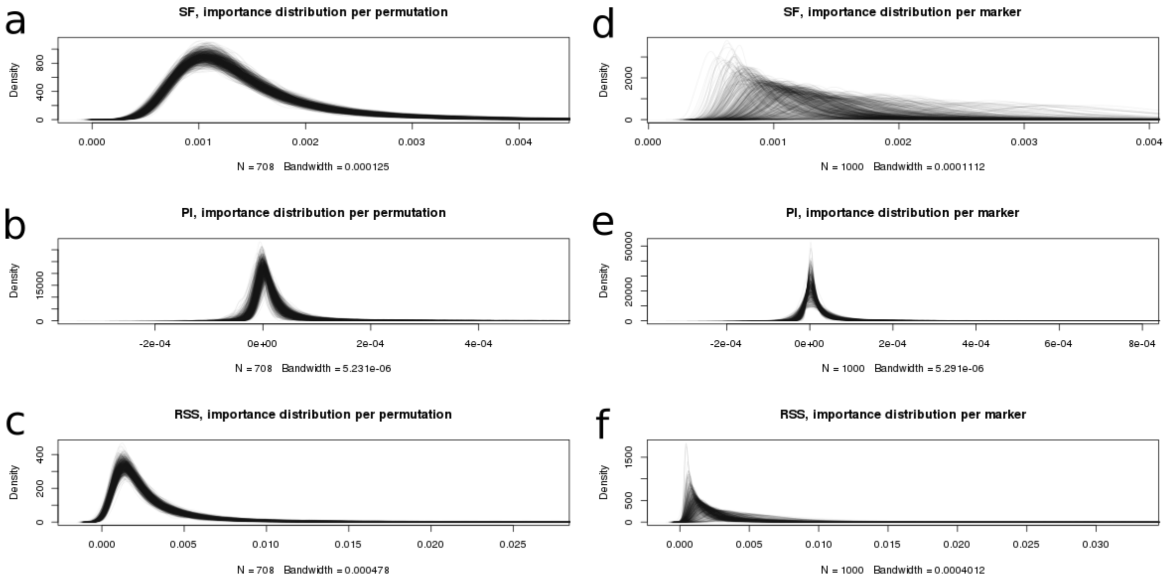

Figure 2.1: Distribution of importance measures plotted per permutation (a-c) or per marker (d-f) for a single trait. The distributions of SF (a and d), PI (b and e) and RSS (c and f) are shown separately. The plots in (a-c) each show the distribution of the importance measure across all predictors for a thousand permutations. One distribution describes the scores of all predictors in a single permutation. In (d-f), the same data is shown, with each distribution describing the scores of a specific predictor from a thousand permutations.

In recent work the selection frequency has often been used for QTL-mapping (Clement-Ziza *et al.*, 2014; Francesconi and Lehner, 2013; Heyn *et al.*, 2014; Picotti *et al.*, 2013). Since the selection frequency always sums up to one, its magnitude is not dependent directly on the total variance or the mean of the trait but on direct comparisons of the ability of predictors to reduce variance at a split. Since the distribution of selection frequencies is largely independent of the distribution of the actual trait values, null distributions of the selection frequency can be pooled across permutations and traits for single markers. The distribution of the RSS and PI on the other hand are highly dependent on the trait variance (Fig. 2.2a).

Pooling the null distributions from the permutation of different traits for these measures would result in false positive QTL in traits with more total variance and false negative results for those traits with smaller total variance. Normalizing the measures with the total trait variance produces null distributions that are more similar for different traits (Fig. 2.2b). This allows pooling the null distribution for the normalized RSS and PI over traits. Pooling of null distributions reduces the amount of permutations necessary to accurately determine empirical *p*-values for QTL.

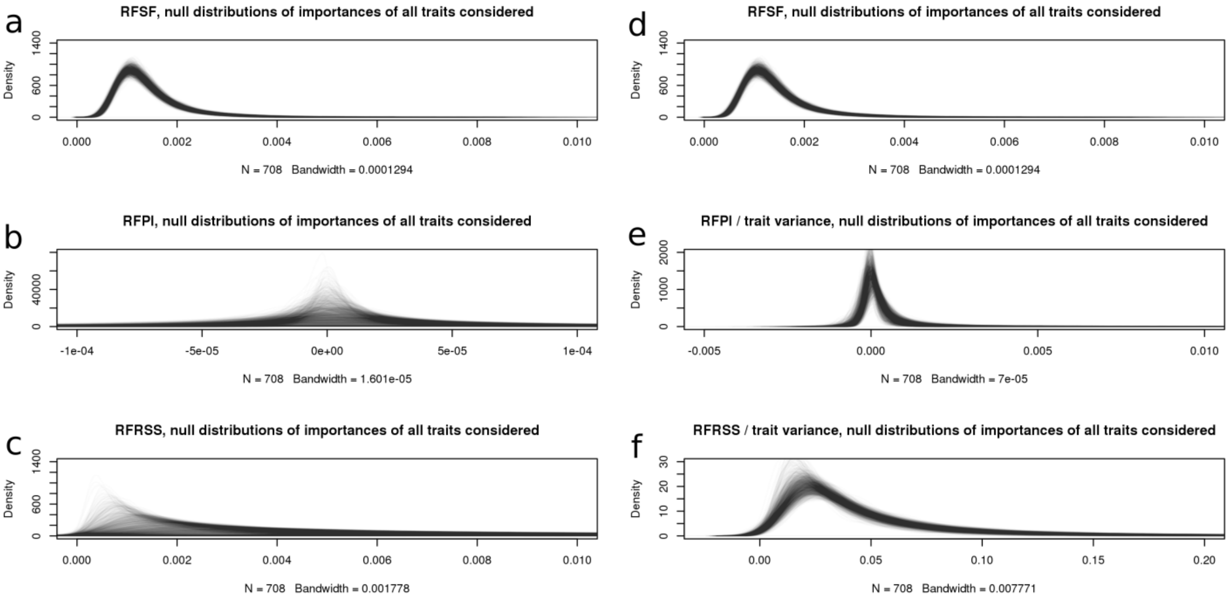

Figure 2.2: Distribution of importance measures across all 1616 permuted traits before (a-c) and after (e-f) normalization with the total trait variance. (a-c) show the distribution of the importance measures across all predictors of one permutation for 1616 randomly selected traits. (e-f) show the distributions from (b-c) after the scores were normalized with the total variance of the trait ffom which they originate. (a and d) depict the trait-specific distribution of SF. (b and e) show the PI-distributions per trait and (c and f) show the distributions of RSS in the same manner.

Empirical *p*-values for all marker-trait-associations are estimated by calculating the proportion of the null distribution for the marker in question that is larger than the importance measure for this marker in this trait. A global correction for multiple testing was performed using the FDR-procedure for all associations at once. Loci with FDR≤0.05 were considered as significant.

Significant markers in the same trait were considered to belong to a single QTL if they were consecutive or if they were in LD (Pearson's $r > 0.8$) and not separated by more than ten not significant markers (Clement-Ziza *et al.*, 2014).

2.1.5 Combinations of importance measures

The three basic importance measures (SF, corrected PI, corrected RSS) were combined in different ways to create new importance measures (Tbl. 2.1).

Table 2.1: Overview of the importance measures and combinations thereof.

Importance measure	description
SF	Selection frequency
PI	Average permutation importance divided by trait variance
RSS	Average reduction in node variance divided by trait variance
C1	$PI * RSS$
C2	$SF * PI * RSS$

C1 is a combination of the corrected PI and RSS measures. This score is calculated by multiplying the corrected PI- and RSS-scores for a specific marker-trait-association. Since OOB-data is used to calculated the original PI- and RSS-scores it is possible for these values to be negative, that is for the use of the marker to increase variance or to decrease predictive accuracy. These negative values are set to zero before taking the product.

C2 is the product of the scores of the SF, corrected PI and corrected RSS. Negative scores were set to zero in this measure as well. Null distributions were created for the basic importance measures and both combinations and the QTL-mapping was performed as described above.

Like SF, PI and RSS, the combined scores increase with increasing importance of a predictor.

2.1.6 Definition of *cis*- and *trans*-loci

All QTL that were located within 10kb of the coding region of the gene on either side were considered as acting in *cis*. All other QTL were considered as acting in *trans* for this transcript.

2.2 Detecting epistasis in quantitative data

2.2.1 Split asymmetry

Epistasis is a term that describes a relationship between two or more genetic loci that produces a contribution to a trait differing from the expectation based on the single additive contributions. In Random Forest, individuals are progressively separated into subgroups starting with a random subset of the whole dataset in the first node. If two predictors are in epistasis regarding a specific trait they are expected to influence the performance of each other when used consecutively on the same samples (Fig. 1.1). This difference in splitting behavior was used by Michaelson *et al.* to detect epistatic interactions and termed split asymmetry (Michaelson, 2010). Split asymmetry can be observed in both AND-epistasis and XOR-epistasis. For two markers (A and B) four combinations are possible: A is used first and B is used on the left or on the right side of the first split (ABl or ABr) or B is used first and A is used on the left (BAl) or on the right side (BAr).

Michaelson *et al.* consider only the asymmetries in splits that are consecutive. The difference between the means of two child-nodes is also referred to as the slope (S). These slopes are summed up separately for the side of the first split they occur (Mr and Ml). If no instances of this marker combination are present in the forest, the entry for that side of the combination is zero.

$$M_r = \sum_{i=1}^{n} Sr_i$$

$$M_l = \sum_{i=1}^{m} Sl_i$$

For each pair of predictors the absolute average difference (D) between the side-specific means of the slopes (Mr and Ml) is determined.

$$D = |M_r - M_l|$$

The average magnitude of the side-specific slopes of ABl and ABr (S) is then subtracted from the absolute difference of these slopes (D).

$$S = \frac{|M_r| + |M_l|}{2}$$

$$D' = D - S$$

For a pair of markers two D' values are calculated, responding to A->B and B->A. The smaller of both values is chosen for the marker-pair. Through these operations D' is only positive when the absolute difference between the slopes is larger than the absolute of the mean of the summed up slopes from both sides. After this step the value of D' still reflects how many slopes were summed up initially, i.e. the selection frequencies of the predictors and also the asymmetry in selection frequencies. D' is a score that has been successfully used to identify epistatic relationships between loci (Michaelson, 2010; Picotti *et al.*, 2013). The practical use of the method is limited by the enormous forest sizes necessary to produce stable scores for marker-pairs (Picotti *et al.*, 2013).

Here we present a method that also uses split asymmetry in Random Forest to detect epistatic interactions. Instead of summing up the side-specific slopes, they are collected and their distributions are tested for significant differences.

The method presented here grows forests predicting the trait in question. The side specific slopes are collected across the forest and stored. As in the method by Michaelson *et al.*, OOB-data is used for the calculation of the slopes, making the method less prone to overfitting compared to the use of IB-data. In contrast to the method of Michaelson *et al.* the method presented in this work also considers marker-pairs for which the markers in question are not used consecutively but on the same individuals in the same tree. The nodes that are split by the predictors of interest can be separated by one or more nodes vertically. This increases the amount of slopes that can be extracted from a single tree. Since the side-specific distributions of the slopes are tested for qualitative differences, it is not necessary to collect all slopes for frequently used predictor-pairs to detect epistatic interactions. Therefore the maximal size of the distributions can be limited, easing the requirements on computational resources necessary to run the method. After all necessary slopes are extracted from the trees the marker-pairs that have enough slopes on each side are tested for significant differences using an unpaired two-sample two-sided Student's t-test. Requiring pairs of markers to have a defined amount of slopes preselects interesting marker-pairs for explicit testing. Marker-pairs that are seldom used in the same tree are also unlikely to explain variation in an interdependent manner.

Pairs consisting of markers that have a Pearson-correlation-coefficient with a magnitude of 0.3 or higher are not tested. These marker-pairs are excluded to avoid false positive results that are the product of the distribution of noise on a smaller number of individuals and allelic incompatibilities. The probability that a large majority of individuals with a combination of two alleles is affected by noise in the same direction is directly linked to the amount of individuals with that genotype. Even the use of OOB-samples would not protect against overfitting if the noise affects a large enough portion of the samples with a specific allele-combination in the same manner.

The *p*-values for both directions are combined using the Fisher's method (Fisher, 1934).

The method is implemented in R. The forests are grown using *RandomForestExtended* with the parameters affecting the properties of the forests set to the default values. All parameters including the minimal size of distributions for testing and the maximal amount of slopes that is stored per side can be freely chosen. The *t.test* function is used with default parameters.

2.2.2 Simulation based benchmark

The performance of the split asymmetry approach presented in this work was assessed using two benchmarks, one entirely based on simulated traits and one based on the recovery of experimentally validated interactions.

For the *in silico* benchmark traits were simulated, using real genotypes that were previously published (Brem and Kruglyak, 2005). Brem *et al.* created a library of *Saccharomyces cerevisiae* crosses with the RM11-1a- and BY4716-strains as parents (RM and BY). A set of 1275 polymorphic markers was used for this benchmark. The selected markers were not identical to direct neighbors. Missing data in the genotype was substituted by zero representing the RM-allele. There were 124 missing genotypic values over 114 samples and 1275 markers. The traits were simulated by adding additive and/or epistatic effects and noise to a base value. All effects were simulated on the BY-allele, i.e. purely additive effects were added to the trait-values of the individuals with the BY-allele at the causal marker, AND-epistatic effects were added to the individuals with the BY-allele at both or all markers, depending on the complexity of the epistatic effect, and XOR-epistatic effects were added to the trait-values of those individuals with the BY/RM- or RM/BY-allele-combination at the markers of interest. Different combinations of additive and epistatic effects were simulated. These traits differ in the number of additive QTL that were simulated, the size of the additive effects and the amount, size, kind and complexity of epistatic interactions (Tbl. 2.2). The direction of the effects was chosen randomly. Each scenario was simulated 32 times.

Table 2.2: Simulated traits for the *in silico* benchmark.

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

The method was assessed by growing forests of 30,000 trees; all other parameters remained at the default. As a comparison to our method, all marker-combinations were also tested with a two-way ANOVA, including an interaction term. The results from both methods were ordered by the p-value. A simulated interaction was considered as recovered if all epistatic interactions were located in the top half of the 99th percentile of the entirety of all marker-pairs regarding significance.

2.2.3 Benchmark based on real data

In an effort to create a functional network of budding yeast genes, Costanzo *et al.* used the synthetic gene array (SGA) approach to test gene-pairs for interactions (Costanzo *et al.*, 2010). Double mutants for specific pairs of genes were created and used for growth assays. Their growth-rate was compared to the expected growth-rate based on the single mutants. All gene-pairs with an interaction score of $|\epsilon| \geq 0.08$ and $\rho \leq 0.05$ were considered as epistatic. All genes were mapped to their closest marker from the Brem-dataset, based on the R64.1.1-reference genome for *S. cerevisiae* (Brem and Kruglyak, 2005; Engel *et al.*, 2014). The closest marker was defined as the marker whose midpoint is the closest to the midpoint of the gene. The gene-gene-interactions from Costanzo *et al.* were mapped to marker-pairs from the genotypes from Brem *et al.*. Based on the assumption that epistatic interactions that affect the expression levels of essential genes can also affect growth rates of budding yeast, we used the expression data of essential genes from Brem *et al.* to predict epistatic interactions of marker-pairs. Out of expression data for 6203 genes, 1107 transcripts of essential genes were chosen (Saccharomyces Genome Deletion Project, accessed in June 2014). Genes were considered essential if a haploid null-mutant for the gene in question was not viable.

A Random Forest consisting of 30,000 trees was grown modeling the expression of each essential gene using the same genotype as for the simulation-based benchmark above. All other parameters remained at the default settings. For each pair of markers the smallest p-value from any of these traits was used for further analysis.

As an example of a widely-used method and as a comparison to the method presented in this work, a two-way ANOVA was performed for all marker-pairs for the expression of all essential genes. The smallest p-value for a marker-pair from all traits was used.

For both methods ROC-curves and precision-recall-curves were generated. Marker-pairs to which no interactions from Costanzo (epistatic or purely additive) were mapped were excluded.

3. Results

3.1 Randomforest extended

In order to benefit from both the inclusion of OOB-node-means in the output of Random Forest and improved usage of computational resources we combined features of *rfdev* and *parallelRandomForest* (Kuhn, 2013; Michaelson, 2010).

rfdev is based on *randomForest*, which is the original Random Forest implementation in the R-environment by Liaw *et al.*. *randomForest* and *rfdev* perform expensive computations in the C-environment while *parallelRandomForest* uses C++ for these tasks (Kuhn, 2013; Liaw and Wiener, 2002; Michaelson, 2010). *RandomForestExtended*, presented in this work, is largely based on *parallelRandomForest* with the addition of features from *rfdev*.

These four implementations of Random Forest in R (*randomForest*, *rfdev*, *parallelRandomForest* and *RandomForestExtended*) were used to grow forests predicting a simulated trait containing no major effects. Since Random Forest involves several sampling events, results are only reproducible by setting a computational seed. Because the implementations do not sample in identical orders we could not use a seed in the R-environment to produce identical forests with these packages. The selection frequencies were computed for the forests and the correlation of forests created with the same package was compared to the correlation of forests created with different packages.

Overall the selection frequencies extracted from all forests correlated strongly with one another (Fig. 2.1). The selection frequencies of forests generated with packages that use C correlate more strongly with one another and the forests created with packages that use C++ are more similar to those created with the other package that uses C++. This is due to slight biases in marker selection that play a role when two markers that were preselected by random sampling explain the same amount of variance. Predictors are tested in the same order as they are sampled for a specific split. If two predictors explain the same amount of variance, the predictor that was sampled first is prioritized. Regardless of the sampling order, *randomForest* and *rfdev* chose some predictors over others. When a minimal threshold, ensuring that a marker does improve on the previously chosen predictor, is introduced to *randomForest* and *rfdev* the bias is not observed anymore. The lack of this bias in *parallelRandomForest* and *RandomForestExtended* can be viewed as an improvement on the original Random Forest implementation by Liaw *et al.*.

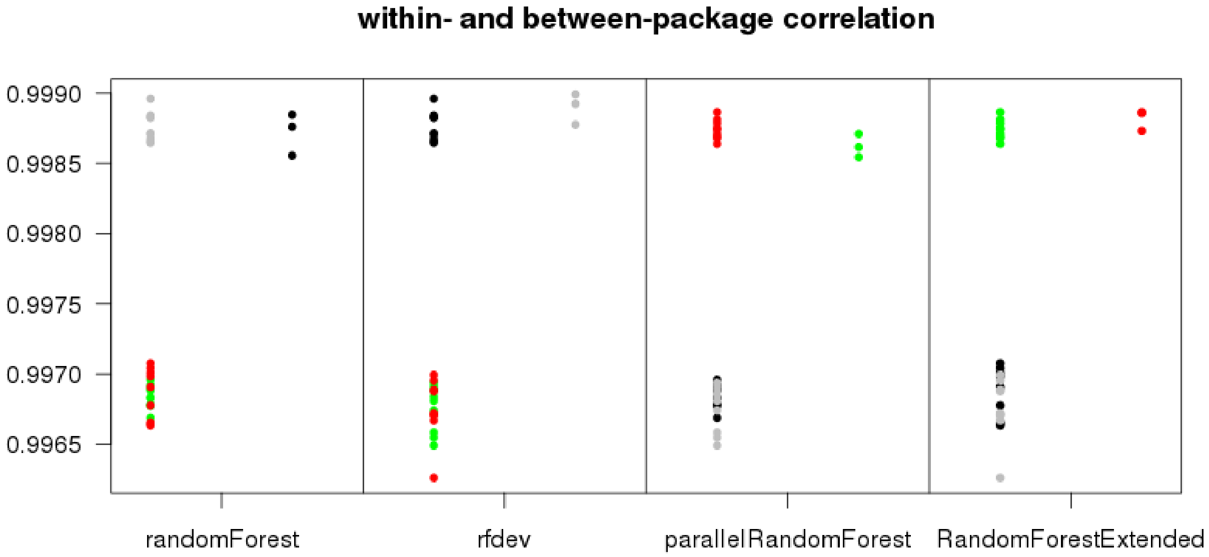

Figure 3.1: Correlation between the selection frequencies of Random Forests modelling a random distribution created with different packages. Points represented the correlation of the selection frequency between a forest created with the package indicated below and another forest. The points are colored according to which package was used to create the second forest. Correlation between replicates of the package indicated at the bottom and *randomForest* are colored in black, while grey represents *rfdev*, green represents *parallelRandomForest* and red represents *RandomForestExtended*. Correlation of forests created with the same package is plotted to the right of each subplot and correlation to forests created with other packages is shown on the left of the subplots.

We validated the OOB-means that are returned by our implementation of Random Forest manually by recomputing them with an external function based on the information which data was OOB in which tree.

3.2 eQTL-mapping with Random Forest

When mapping QTL with Random Forest several options of importance measures are available. The selection frequency (SF), the permutation importance (PI) and the average reduction in variance (RSS) all score the performance of a predictor. We mapped QTL with these importance measures for the expression data of 6,464 transcripts in 65 cultures of budding yeast, including 46 different genotypes. By permuting the phenotype-vectors and computing the importance measures we created empirical null distribution and used those to estimate the significance of the associations between genetic markers and transcripts. A global correction for multiple testing was performed using the FDR-procedure.

A combined total of 6,531 significant marker-trait-associations were mapped using SF (Tbl. 3.1). The normalized PI- and RSS-measures mapped a substantially larger amount of significant marker-trait-associations (19,270 and 20,125). Using the product of PI and RSS, termed C1, 22,221 significant associations were mapped to traits, more than with any of other measures tested. C2, the product of SF, PI and RSS, detected 17,573 significant associations, more than SF but less than PI, RSS and C1.

Table 3.1: Results of eQTL-mapping with different importance measures. Given are the markers that affect a trait with FDR≤5%, QTL after marker joining, affected traits and information about the location of the loci affecting the trait in relation to the gene which encodes for the transcript whose level is predicted.

	SF	PI	RSS	C1	C2
Significant predictors	6531	19270	20125	22221	17573
Significant cis-markers	354	881	942	1007	827
Proportion of cis-markers in all significant predictors	0.054	0.046	0.047	0.045	0.047
Significant QTL	2676	5135	5142	5683	4874
Significant cis-QTL	160	271	287	303	261
Traits with at least one QTL	2286	3855	3739	4081	3694
Proportion of traits with a cis-QTL in all traits with at least one QTL	0.070	0.070	0.077	0.074	0.071

Following the criteria described in the methods for joining significant associations of several markers with the same trait to consecutive QTL-regions, 2,676, 5,135, 5,142, 5,683 and 4,874 QTL were mapped for SF, PI, RSS, C1 and C2 respectively for all transcripts combined (Fig. 3.2).

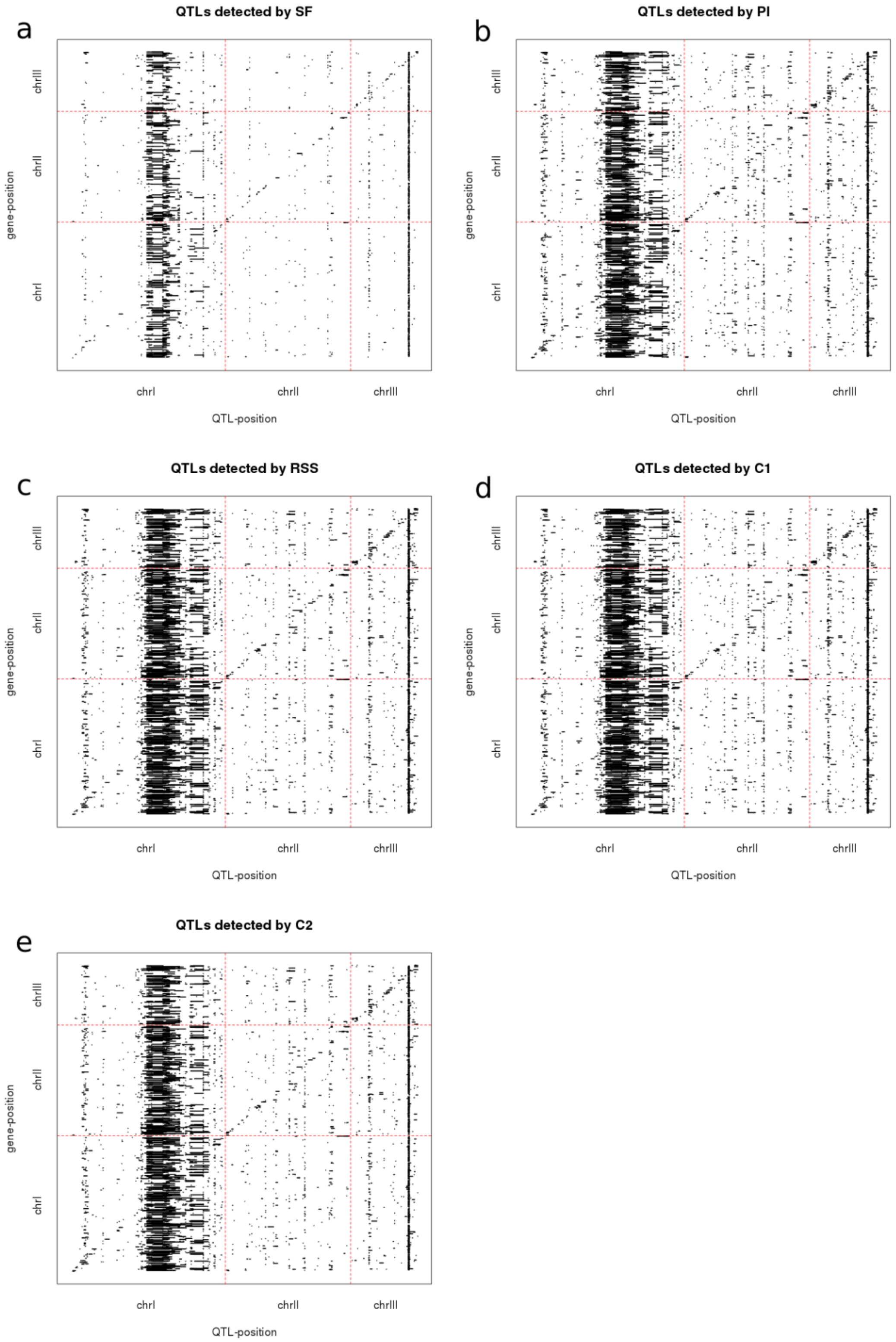

Figure 3.2: QTL with FDR≤5%. The significance of predictors was determined with empirical null distributions of importance scores. Predictors with FDR≤5% after a global correction for multiple testing are joined to trait-specific QTL as described in the methods. The detected QTL are plotted separately for SF (a), PI (b), RSS (c), C1 (d) and C2 (e). The y-axis represents the genomic region where the affected transcript is encoded and the x-axis represents where the QTL that affects the trait is located.

There was a large overlap between the results derived from the use of different importance measures (Fig. 3.3, Tbl. 3.2 and Tbl. 3.3).

Table 3.2: Proportion of significant results shared between importance measures as percentages. Each row shows which proportion of the significant markers detected with that measure was also detected in the same trait with other measures.

	SF	PI	RSS	C1	C2
SF	-	92.1	94.3	94.1	95
PI	47.8	-	84.5	95.5	87.4
RSS	48.8	83.8	-	96	89.4
C1	44	85.7	86.8	-	84.7
C2	51.9	92.1	94.9	99.5	-

Table 3.3: Proportion of significant results shared between importance measures as percentages. Each row shows which proportion of the significant QTL detected with that measure was also detected in the same trait with other measures.

	SF	PI	RSS	C1	C2
SF	-	95.3	96.8	96.7	97.3
PI	32.3	-	84.8	95.3	85
RSS	31.4	81.2	-	96.2	83.8
C1	28.4	82.7	87.1	-	78.9
C2	36.2	93.3	96	99.8	-

Most marker-trait-associations that were significant in the SF-mapping were also discovered using all other measures. The other two fundamental measures, PI and RSS, also showed an agreement of 81.2% and 84.8%, depending on which measure was used as the query. As combinations of the fundamental importance measures, C1 and C2 agree strongly with each other and the fundamental importance measures.

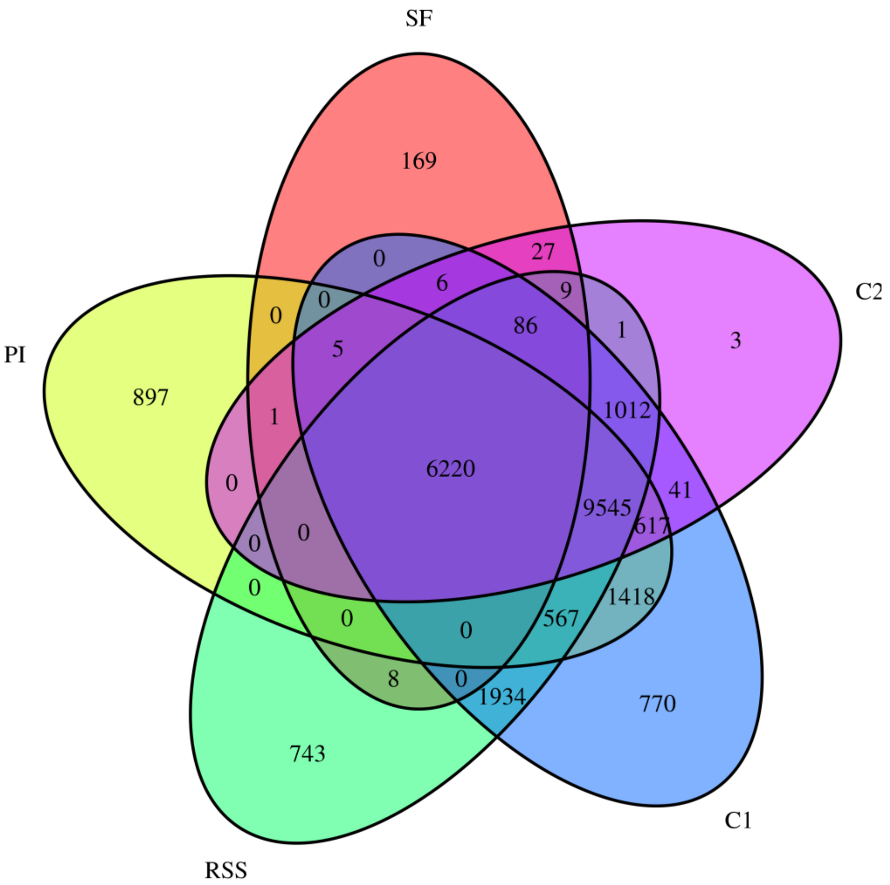

Figure 3.3: Overlap between predictors that are significant in the same traits in different importance measures. Shown are predictors with an FDR≤5%. The size of the area displayed is not proportional to the overlap.

On the log scale, all measures were highly correlated (Tbl. 3.4). Among the basic importance measures RSS and PI showed the most similarity with a correlation-coefficient of $r_{\log(\rho PI),\log(\rho RSS)} = 0.87$. The combined measures C1 and C2 correlate well with each other and the measures that were used to create them. C1 and C2 correlate better than SF and C2, indicating that the influence of SF on C2 is rather small in comparison of the combined influence of PI and RSS.

Table 3.4: Correlation of *p*-values of markers on the log-scale determined with different importance measures.

	PI	**RSS**	**C1**	**C2**
SF	0.676	0.794	0.748	0.764
PI	-	0.87	0.969	0.944
RSS	-	-	0.955	0.973
C1	-	-	-	0.991

The agreement of the measures after joining of significant markers to eQTL was similar, although the proportion of significant results that were not recovered with SF was substantially smaller (Tbl. 3.2, Tbl. 3.3). When comparing the number of markers that are included in QTL that were discovered with all measures considered, QTL were found to include on average more than twice as much markers in PI, RSS, C1 and C2 as in SF (Fig. 3.4).

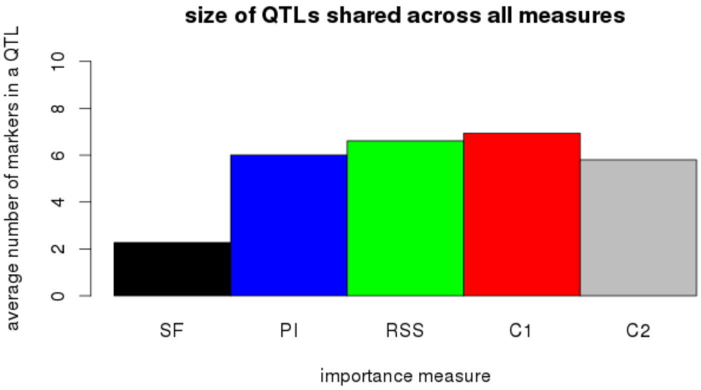

Figure 3.4: Average number of predictors in QTL that were mapped with all importance measures. Predictors with FDR≤5% were joined to QTL as described in the methods. QTL that were mapped with all importance measures were analyzed for the amount of predictors they incorporate.

When examining the traits with significant QTL in C1 but not SF we found that although SF did not find significant QTL for these traits, the highest scoring markers in C1 scored high in SF as well. In 66% of the traits where significant marker-trait-associations were mapped using C1 but not SF, the top marker in C1 was the best scoring marker using SF as well. In 99.4% of the traits with significant QTL in C1 but no significant QTL in SF, the top scoring marker in C1 was among the top ten scoring markers in SF.

When considering all traits with a significant marker-trait-association detected by C1, 99.7% of all top scoring markers in C1 were also placed among the ten best scoring markers in SF.

Although the selection frequency maps less single significant markers and less QTL after marker-joining than all other importance measures considered, the measures agreed on top-scoring results. In traits where SF mapped no significant marker-trait-associations, the marker that scored best using C1 also scored among the top ten results from SF in almost all cases.

Clement-Ziza et al detected both *cis*- and *trans*-eQTL in this data with the *trans*-eQTL outnumbering the cis-eQTL by a wide margin (Clement-Ziza *et al.*, 2014). Some *trans*-eQTL spanned a large number of traits in many cases qualifying them as *trans*-hotspots. *Trans*-hotspots affect the expression of many genes that are physically distant from the polymorphic loci. Among the hotspots Clement-Ziza *et al.* found an inversion on chromosome I and a locus on chromosome III, *swc5*, which regulates the deposition of histone H2A.Z. This polymorphism affects the expression of many transcripts by modulating the probability with which RNA-polymerase II initiates sense- or antisense-transcription. These eQTL-hotspots were found to contain growth QTL as well, further supporting their biological relevance.

Among the traits for which no QTL were mapped with SF, the significant results in C1 were highly enriched for both the hotspot on chromosome I (one-sided Fisher's exact test, *p*<2^{-16}) and the hotspot around *swc5* on chromosome III (one-sided Fisher's exact test, *p*<2^{-16}). This *trans*-hotspot globally increases antisense expression and decreases sense expression (Clement-Ziza *et al.*, 2014).

Cis-loci often have a strong influence on nearby transcripts, which arguably makes *cis*-eQTL easier to detect (Loguercio *et al.*, 2010). The significant markers are highly enriched for *cis*-loci for all importance measures considered (*p*<$10^{-50}$for all measures, one-sided Fisher's exact test). When analyzing only the significant markers detected with the use of C1 in traits with no QTL detected with SF, *cis*-markers are enriched among the markers that significantly affect a specific trait (*p*<10^{-12}, one-sided fisher test).

We mapped the expression of 6464 different transcripts in *S. pombe* to a total of 708 markers. As a criterion to assess the degree of influence of a locus on a trait we used the selection frequency, permutation importance, the reduction in variance, a combination of the latter two and a combination of all three fundamental measures. Strong correlation on the log scale of the corrected p-values from all approaches indicates that they score marker importance in a similar way. SF recovers the least significant marker-trait-associations and C1 the most. PI, RSS and the measures they influence map QTL that include more predictors than those mapped with SF (Fig. 3.4). SF also maps QTL in drastically fewer traits than the other measures, indicating that the differences between the measures are not solely due to sharper peaks in SF.

In traits with no significant markers detected with SF, the significant marker-trait-associations found by C1 are highly enriched around two previously published eQTL-hotspots and for *cis*-acting QTL.

3.3 Detecting epistasis with a split asymmetry approach

3.3.1 *In silico* benchmark

We attempt to detect epistatic interactions between genetic loci in quantitative traits by analyzing the interdependent performance of the predictors. When a predictor performs significantly different on the subgroups generated by another predictor it can be assumed that this pair of predictors has a relationship that is not purely additive, i.e. the effects of one predictor on the trait are dependent on the other predictor. The absolute difference in the means of the subgroups created by the split with the second predictor is collected specific for the side of the first split it is on. For a pair of predictors (A and B) four distinct distributions (ABl, ABr, BAl and BAr) are collected. To compare ABl to ABr and BAl to BAr a two-sided Student's t-test is used. Given large enough distributions we expect to detect the interaction regardless of which marker is used first, but since the number of trees and therefore the total number of splits are limited, we cannot expect to detect all interactions in both directions. We combine both *p*-values with the Fisher-method (Fisher, 1934).

In order to evaluate the performance of the split asymmetry approach of analyzing predictor-interdepended splitting behavior to detect epistasis, both simulated and previously published experimental data was used as a benchmark.

Scenarios containing epistatic interactions were simulated for the genotype of a previously published library of yeast crosses (Brem and Kruglyak, 2005). Additive, AND-epistatic and XOR-epistatic effects in different combinations, with different effect sizes

and with and without overlap of markers between the purely additive and the epistatic markers were combined. The traits were simulated by assigning a base value to all individuals, adding additive and epistatic effects to the phenotypic values of those individuals that are targeted by the effect and by finally adding normally distributed values with a mean of zero and a standard deviation proportional to the trait mean to all individuals at different levels. The direction of the simulated effects was randomly chosen for each level of noise but consistent across all individuals. All simulations were repeated 32 times. As a representative of a two-dimensional approach we chose to perform a two-way ANOVA allowing for main effects of the tested predictors and an interaction term.
Simulations were considered as recovered if all simulated interactions scored in the top half of the 99th percentile of all possible interactions. In the case of three-way interactions, all possible two-way interactions between the interacting loci had to be recovered.

In the traits that only contained one two-way AND-epistatic interaction, both methods recovered most interactions at low levels of noise (Fig. 3.5). The recovery decreased with increasing levels of noise. The interactions were detected at a higher rate if the simulated effect was larger, i.e. if the signal-to-noise ratio was bigger. In the simulations of three-way AND-epistasis the rate with which the simulations were recovered was smaller. The simulations for three-way interactions were more difficult to recover by definition, since they are comprised of three two-way-interactions instead of one. The split asymmetry approach performs better than the ANOVA when recovering three-way interactions, although the overall recovery-rate is much lower than that of single two-way interactions. Three-way AND-epistasis is recovered at a very low rate by both methods if the epistatic contribution to the trait is small.

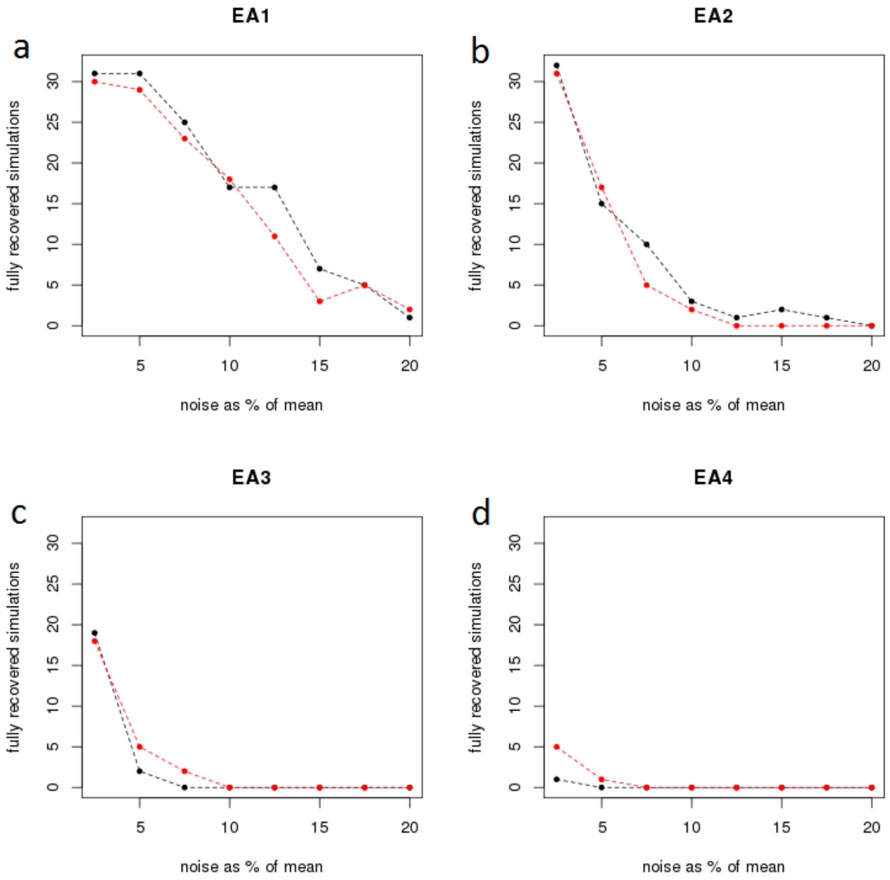

Figure 3.5: Recovery of simulated interactions in different AND-epistatic scenarios. The amount of fully recovered simulations by the exhaustive approach are depicted as black points, the simulations recovered by the split asymmetry approach are shown in red. EA1, the scenario in (a), includes one AND-epistatic interaction with a big effect-size and no purely additive effects. (b) shows the recovery of EA2, which combines one AND-epistatic interaction with a smaller effect-size and no purely additive effects. EA3 is shown in (c) and includes a big AND-epistatic effect based on three predictors. All interactions between the three predictors have to be recovered. (d) shows EA4, which is designed like EA3, the scenarios differ only in the effect size which is smaller in EA4.

The split asymmetry method misses some interactions at the lowest noise-level. In most cases the reason for these false negatives are distributions of side specific splitting performance that are too small for testing.

The ANOVA outperforms the split asymmetry approach by a wide margin when recovering the simulations of XOR-epistasis with both big (Fig. 3.6a) and small effect sizes (Fig. 3.6b).

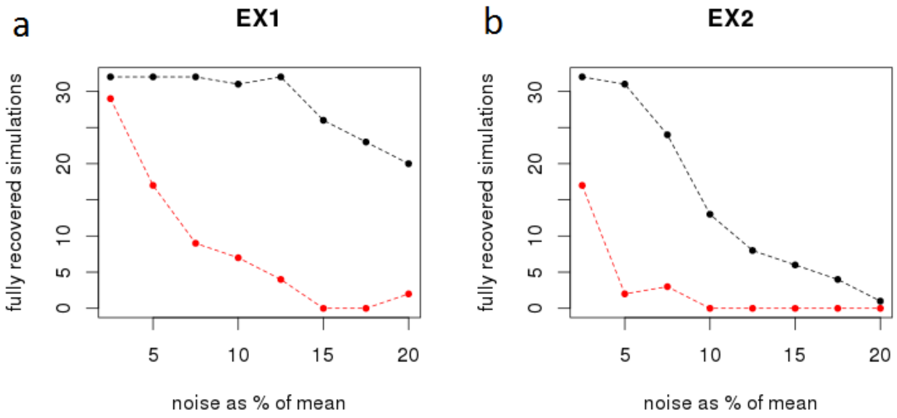

Figure 3.6: Recovery of simulated interactions in XOR-scenarios. Red points represent the amount of fully recovered simulations with the split asymmetry-approach while the recovery with the exhaustive approach is depicted in black. Depicted is the performance for the EX1 (a) and EX2 (b) scenarios which consist on one XOR-epistatic interaction with big and small effect-sizes respectively.

To investigate the performance of the approaches at recovering epistatic interactions in the presence of purely additive effects, traits with two and-epistatic interactions and one to four predictors with additive effects were simulated. The predictors for whom the purely additive effects were simulated did not overlap with the interacting predictors.

Both methods performed better with fewer purely additive effects (Fig. 3.7). While the ANOVA outperformed the split asymmetry approach in each of these four simulations at lower levels of noise, both methods performed in a similar fashion for the traits with more simulated noise.

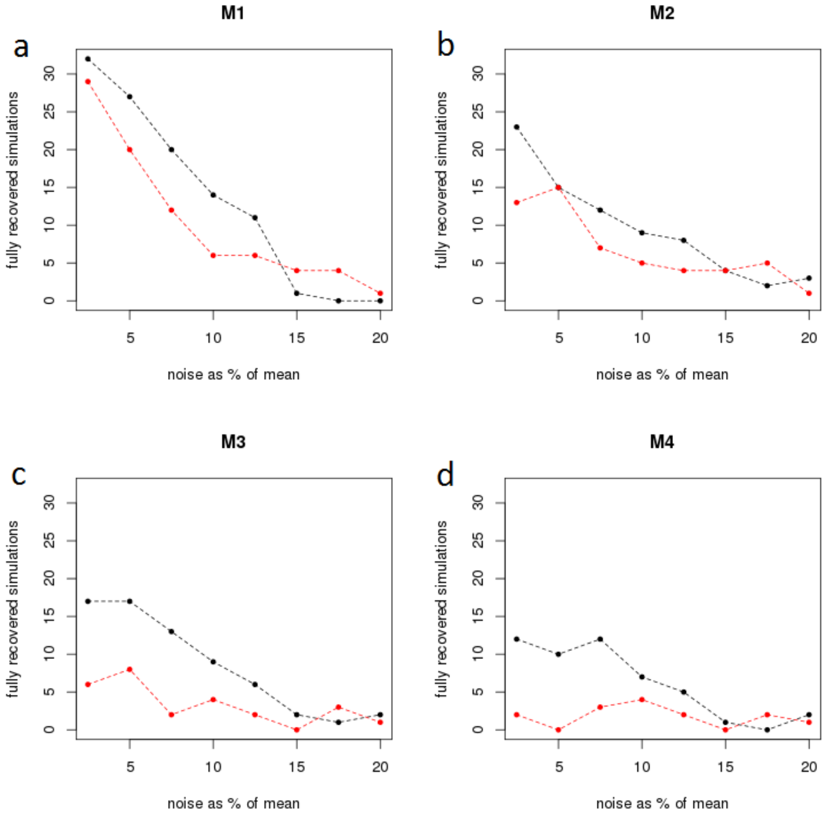

Figure 3.7: Recovery of simulations of AND-epistatic interactions in scenarios that also contain predictors with purely additive effects. The additive effects are simulated for predictors that are different from those with epistatic effects. One AND-epistatic interaction was combined with one (a), two (b), three (c) or four (d) separate additive effects on distinct predictors of the same size as the epistatic interaction. The recovery with the split asymmetry approach is plotted in red and the recovery with the exhaustive approach is shown in black.

In order to assess the performance of the split asymmetry approach and the ANOVA in traits that combine additive effects and interactions on the same loci, scenarios that address this question were simulated. In four traits two AND-epistatic interactions were combined with two additive effects each.

The ANOVA performed better when the simulated epistatic effects were larger (Fig. 3.8b and d). When the interactions contained loci for which a purely additive effect has been simulated as well, the ANOVA detected more epistatic interactions than in the simulations with distinct causal loci (Fig. 3.8a and b). The recovery rate for these interactions was still lower than in the simulation with no purely additive interaction present (Fig. 3.5a). In the scenario that combines small epistatic effects with additive effects on other markers, the ANOVA recovered no simulated traits at all (Fig. 3.8c). When the epistatic effects are bigger, some simulations were recovered (Fig. 3.8d). In the scenarios in which the additive effects each share a locus with a epistatic interaction, the split asymmetry approach clearly outperformed the ANOVA especially with smaller epistatic effect sizes and higher noise levels (Fig. 3.8a and b). Surprisingly the split asymmetry approach performed better if the epistatic effect sizes are smaller (Fig. 3.8a). For all simulated traits in this scenario that the split asymmetry did not detect the collected distributions were either too small or the correlation between the epistatic markers was too high. All interactions for which big enough distributions were available and the correlation was not too strong, were detected as highly significant.

Like the ANOVA, the split asymmetry approach did not successfully recover any simulations when small epistatic effects were simulated for markers that were distinct from those with purely additive effects (Fig. 3.8c). A small amount of simulations was recovered if the epistatic effects were stronger (Fig. 3.8d).

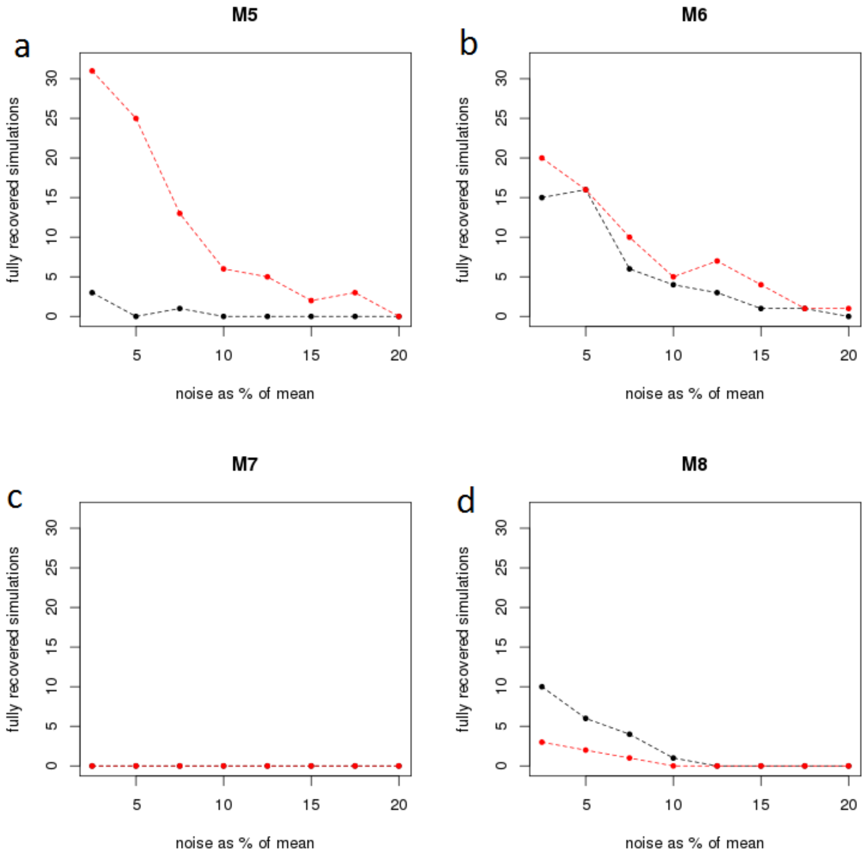

Figure 3.8: Detection of AND-epistatic interactions in traits with additive loci. M5-8 contain two additive loci and tow AND-epistatic interactions each. The performance of the split asymmetry approach is shown in red and the performance of the exhaustive approach is shown in black. The interactions in M5 (a) and M6 (b) each have one predictor that also has a purely additive effect while M7 (c) and M8 (d) do not. The effect size in M5 and M7 is smaller than that in M6 and M8.

When XOR-epistatic effects were simulated in the presence of two purely additive effects on a distinct marker-pair and a marker-pair the split asymmetry approach failed to recover more than one full simulation at each noise level (Fig. 3.9a). The ANOVA recovered almost all simulations at low levels of noise and a substantial amount of full simulations at high levels of noise. To further evaluate the performance of the approaches the recovery of the interaction that shares a marker with a purely additive effect was examined separately from the interaction that shared no markers with any purely additive effect (Fig. 3.9b). It should be stressed that this scenario contains two purely additive effects, ensuring that there is always at least one additive effect that is not explained by a single marker-pair. The ANOVA managed to recover most interactions at low levels of noise and about half of the interactions at high levels of noise for interactions that share a marker with an additive effect and also for those that do not (Fig. 3.9b). The split asymmetry approach on the other hand performed distinctly better if one of the interacting loci was shared with a purely additive effect. If a marker was shared, the split asymmetry approach performed slightly worse than the ANOVA, if no markers were shared almost no interactions were recovered. In all but one case where the interaction was detected, it was detected with higher significance with the additive marker being used to split first.

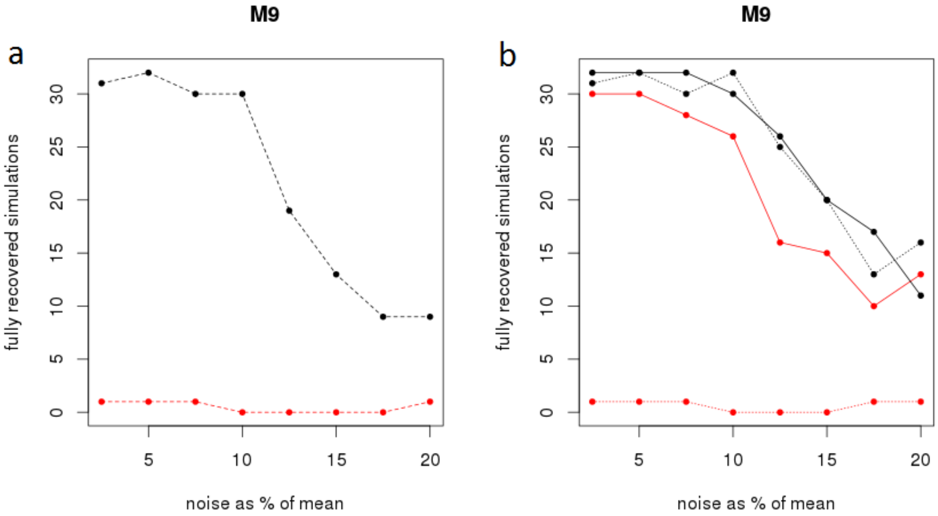

Figure 3.9: Recovery of simulations with XOR-epistatic interactions and additive effects. M9 combines two XOR-epistatic interactions on different predictors with two additive effects. The additive and epistatic effect sizes are the same. One of the additive effects is simulated on one of the epistatic loci. Recovery of the interactions is shown in red for the split asymmetry approach and in black for the exhaustive approach. The recovery of all epistatic interactions from the simulations is shown in (a). The recovery of the separate interactions is shown in (b). The points indicating the recovery of the interaction that does not share a predictor with a purely additive effect are connected with a dotted line. The points indicating the performance of the methods regarding the interaction that has a predictor with an additional additive effect are connected with a solid line.

To further examine the relationship between the performance of both approaches and the presence of additive effects and to lessen potential effects of limited tree-depth, scenarios in which there is only one purely additive effect with differing effect sizes and only one interaction were simulated. In these scenarios the rate with which the two-dimensional approach recovered the interactions was dependent on both the size of additive effect and on sharing of markers between the effects. If the additive effect was not simulated for one of the epistatic markers and the effect size of the additive contribution was large the ANOVA recovered very little interactions, even at low levels of noise (Fig. 3.10d). If the effect size of the additive contribution was smaller, the ANOVA performed better (Fig. 3.10b). While a large majority of simulations was recovered at low levels of noise, the recovery-rate dropped tremendously with increasing levels of noise. If a marker was shared with the additive contribution, the additive effect size had little influence on the performance of the ANOVA. At low levels of noise, the rate with which the interactions were recovered was high and at high levels of noise most interactions were not recovered.

The performance of the split asymmetry approach was best if the additive effect was simulated on one of the epistatic markers (Fig. 3.10a and c). With a smaller additive effect the recovery was high at low levels of noise and a considerable amount of simulations was still recovered at high levels of noise (Fig. 3.10a). If the additive effect was larger the approach recovered more simulations at medium and high levels of noise (Fig. 3.10c). If the additive effect was simulated on an epistatic predictor the split asymmetry approach clearly outperforms the ANOVA. In all but one case where the interaction was detected, it was detected with higher significance with the additive marker being used to split first. When the additive effect was smaller and simulated on a distinct marker, the split asymmetry approach performed slightly worse than the ANOVA (Fig. 3.10b). If the additive effect was larger the performance of the split asymmetry approach was very similar. In this scenario the split asymmetry approach vastly outperformed the ANOVA at lower to medium levels of noise (Fig. 3.10d).

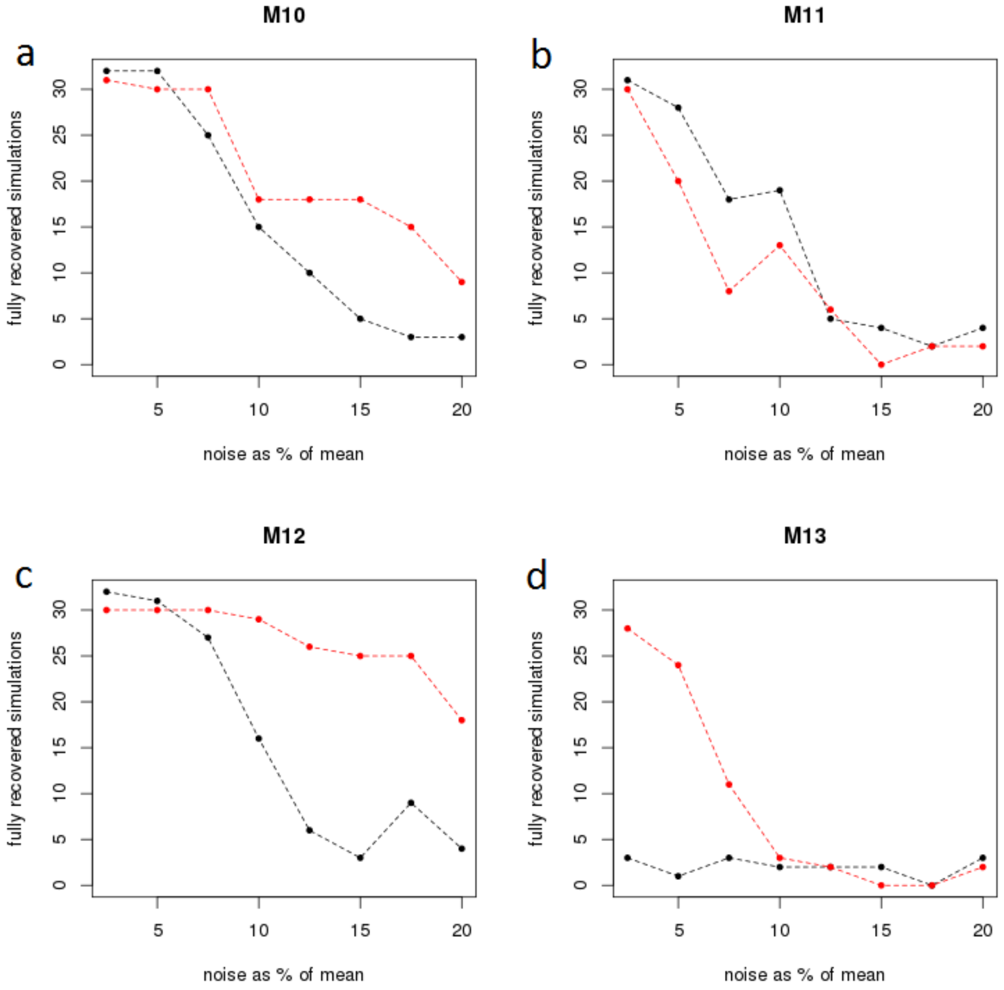

Figure 3.10: Recovery of interactions from simulations with larger additive effects. The performance of the split asymmetry approach is shown in red and the performance of the exhaustive approach is shown in black. M10 (a) contains one AND-epistatic interaction and one additional purely additive effect on one of the epistatic predictors. In M11 (b) the additive is on a predictor outside of the interaction. The additive effect in M12 (c) is larger than that in M10. Like in M11, the additive effect in M13 (d) is not on one of the epistatic predictors. The effect size of the purely additive effect is larger in M13.

The performance of both the split asymmetry and the two-dimensional approach were highly dependent on the presence and size of the additive contributions and on the kind of epistasis that was simulated. The ANOVA recovered XOR-epistatic interactions at a considerably higher rate compared to AND-epistatic interactions. In the simulated traits the presence of additive effects affected the performance in a negative way if the effect was simulated outside the epistatic marker pair but not if it could be explained by one of the epistatic markers. If the additive effect was simulated outside of the interaction, the performance of the exhaustive approach responded negatively to increasing additive effects. The rate of recovery for interactions with bigger effect sizes was higher than that for interactions with smaller effect sizes.

The split asymmetry approach largely failed at identifying XOR-epistatic interactions if additive effects outside the interacting predictors existed. If the additive contribution was simulated on an interacting predictor the method detected much more interactions but was still outperformed by the ANOVA. If no purely additive contribution existed, the recovery of XOR-epistatic interactions was smaller than that of AND-epistatic interactions.

In the scenarios that contain purely additive contributions outside of the interaction the performance of the split asymmetry approach unlike the performance of the ANOVA was largely independent of the size of the additive effect but very sensitive to the total number of contributing loci (Fig. 3.7). The rate of recovery for interactions from scenarios with a lot of additive effects was much smaller than that from scenarios with single purely additive contributions. If interacting loci also had independent additive contributions, the performance of the split asymmetry approach was better than in scenarios with no additive contributions at all (Fig. 3.10a and c, Fig. 3.5a).

If one of two interacting markers also had a purely additive contribution to the outcome, the interaction was usually detected with more significance for the combination of predictors that uses the predictor with the additive effect first.

3.2.2 Recovery of experimentally validated gene-interactions in *Saccharomyces cerevisiae*

Simulations can help in analyzing the functional properties of a method but the overall performance can be assessed in a more fashion that is more relevant to biological problems by using a benchmark of experimentally validated data. Gene expression data from budding yeast published by Brem and colleagues and growth data from double knockouts in budding yeast by Costanzo and colleagues (Brem and Kruglyak, 2005; Costanzo *et al.*, 2010) were used for this benchmark. Costanzo *et al.* created double mutants for which growth traits were measured using the SGA-approach. Under the null hypothesis, assuming additivity, the difference of the growth rate of the double mutant was expected to be the product of the growth rates of the single mutants in relation to the wild type. A growth rate that is higher than expected (i.e. the double mutant grows less slow compared to the wild type as expected) constitutes positive epistasis while a growth rate that is smaller than expected constitutes negative epistasis. For this benchmark those gene-pairs whose interaction was both significant and strong ($p \leq 0.05$, $|\varepsilon| \geq 0.08$) were considered as epistatic. 187,751 pairs of genes satisfied these requirements.

The expression data was used to detect epistatic interactions between predictors from a set of 1275 genomic loci (Brem and Kruglyak, 2005). For a total of 114 individuals the expression of 1107 transcripts that were annotated as essential by the Saccharomyces Genome Deletion Project was predicted (Saccharomyces Genome Deletion Project). Null-mutants for these genes were not viable in a haploid background. The detection of interactions that effect the expression of essential genes was chosen as a benchmark the expression changes of these transcripts can be assumed to be more likely to affect fitness traits like the growth rate. While epistasis can occur in non-essential genes as well, these interactions are expected to have more subtle effects on fitness measurements.

For each pair of markers the most significant p-value from the results of the split asymmetry method from all transcripts was used.

As an example of widely used exhaustive approaches, the ANOVA was used to predict interactions from the expression data. As for the split asymmetry for each pair of markers the smallest p-value achieved over all transcripts was used.

The marker-pairs were ordered by their smallest p-values and ROC-curves and precision-recall curves were created separately for both the performance of the ANOVA and the split asymmetry approach.

The rate with which epistatic interactions were recovered by the ANOVA is close to that achieved by randomly picking marker-pairs and does not deviate significantly from random sampling without replacement (one sided Wilcoxon rank-sum test, *p* > 0.05) (Fig 3.11a). Despite peaking early the precision of the ANOVA quickly drops to levels expected of random sampling (Fig. 3.11b). The *p*-value that the ANOVA determines for a pair of markers based on the expression data was not informative about the interactions from Costanzo *et al.* in the vast majority of cases.

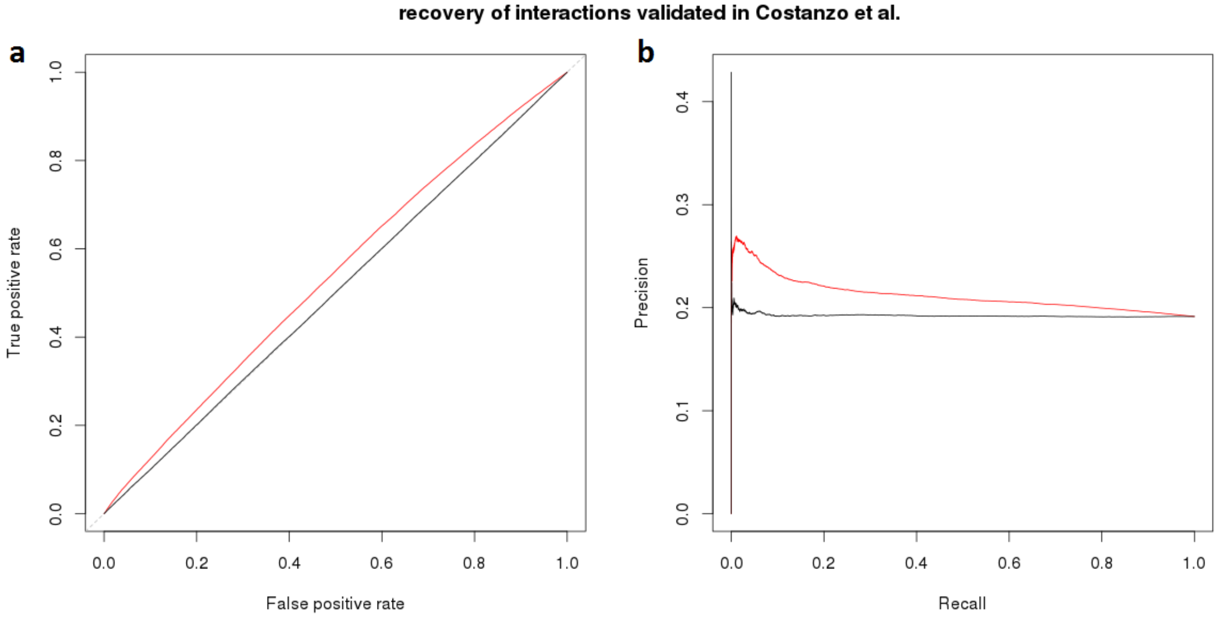

Figure 3.11: Detection of epistatic interactions in *Saccharomyces cerevisiae* from gene expression data. ROC-curves (a) and precision-recall-curves (b) describe the performance of the split asymmetry approach and the exhaustive approach, depicted in red and black respectively, at identifying epistatic interactions between markers, that were validated in Costanzo *et al.*.

The split asymmetry approach clearly outperformed the exhaustive approach when applied to this gold standard. Pairs of markers that rank high among all results from the split asymmetry approach were more likely to be mapped to genes that are interacting in Costanzo *et al.* (one sided Wilcoxon rank sum test, $p < 10^{-91}$).

4. Discussion

4.1 eQTL-mapping with Random Forest

Explaining variation in the expression of genes with differences in genotypes can improve our understanding of the underlying regulatory architecture of the cell (Ackermann *et al.*, 2013; Albert and Kruglyak, 2015; Brem *et al.*, 2002; Clement-Ziza *et al.*, 2014; Maurano *et al.*, 2012). As an intermediate trait, gene expression does not affect complex metabolic traits directly but through its influence on protein concentrations and states in the cell. Although their effects on complex phenotypes such as disease traits have the potential to be attenuated through other regulatory processes like feedback-loops and regulation of the proteasome, eQTL are enriched in disease-causing SNPs (Nicolae *et al.*, 2010). Aside from improving the underlying knowledge of biological processes the correct identification of eQTL can help to prioritize mutations among a set of possible causes of a disease (Lappalainen *et al.*, 2013).

Although univariate methods are still used in some eQTL-studies, multivariate methods like Random Forest were shown to be more powerful at identifying causal loci (Holdt *et al.*, 2013; Michaelson *et al.*, 2010; Shpak *et al.*, 2014). To further investigate how Random Forest can be used best for QTL-mapping, expression data of *S. pombe* was mapped using genotypes as predictors with different importance measures. The traits were mapped using the selection frequency, the permutation importance and the reduction in variance in addition to the product of PI and RSS (C1) and the product of SF, PI and RSS (C2). The measures generally agreed on which markers were important to explain variation in a trait (Fig. 3.3). Almost all QTL that were detected with the use of SF were also detected with the use of PI, RSS, C1 or C2. PI, RSS, C1 and C2 detected causal loci in over 50% more traits and about twice as many QTL overall compared to SF (Tbl 3.1). The amount of QTL that were detected with the use of SF is similar to the number of QTL detected by Clément-Ziza *et al.* (Clement-Ziza *et al.*, 2014). The strategy employed to deal with missing genotypes was different between this work and the study by Clément-Ziza *et al.* which could explain minor differences in the number of detected QTL.

Markers that were scored as significant with C1 in traits to which no QTL were mapped with SF also scored high in SF without reaching significance. Proportional to the increase in traits with significant predictors, C1 also identified additional *cis*-loci as having a significant influence on a trait (Tbl. 3.1). The *trans*-acting QTL in these traits were highly enriched for *trans*-regulatory hotspots. Associations between these hotspots and additional traits compared to those mapped with SF are highly plausible. The *swc5*-hotspot for example was shown to affect the regulation of sense- and antisense-transcription globally (Clement-Ziza *et al.*, 2014).

A larger percentage of the QTL mapped with SF was considered *cis*-acting compared to the other measures. These *cis*-QTL were mostly found with the other measures as well, in addition to other QTL. The assessment that *cis*-QTL can be detected easier would be consistent with less statistical power in SF *(Loguercio et al*., 2010).

The large overlap between the results achieved with the different measures and the general agreement on the most important predictors per trait suggest that the difference in the total amount of significant QTL might be due to a difference in statistical power.

SF, PI and RSS score different properties of the predictors. RSS and PI both describe how much variance is reduced in a split with the predictor of interest. While RSS is an average of the reduction in variance itself, PI combines this information with the proportion of individuals that were grouped using the predictor. When the predictor is permuted, only the predictions for those individuals which are in a node following a split with that predictor change. The magnitude of that change for these individuals is proportional to the variance that was explained by the predictor. In contrast to PI and RSS, SF only reflects how often a predictor was the best suited one for a specific split in a sample of predictors. Regardless of how well Random Forest describes the variance in a trait the sum of all SF-scores always equals one. An increase in the selection frequency of one predictor automatically decreases the frequencies of all other predictors regardless how they explain the trait variance. PI and RSS on the other hand can score relative to the variance they explain, i.e. their scores are not limited to a total sum. If many markers explain large amounts of variance they do not decrease like SF.

Aside from this interdependency of scores between predictors, SF also limits the scores of predictors that explain large amounts of variance. Each use of a predictor for a split is valued equally regardless of the position of the node in the tree it was used on. Typically nodes closer to the root of the tree contain more samples and are therefore less prone to overfitting. If a predictor is informative about the outcome, it is more likely to be used earlier. Samples in nodes that were grouped by the predictor already will not be split with this predictor again. This limits the scores of these predictors. In comparison to PI and RSS nodes with many samples have less weight and nodes with few samples have more weight. SF might lose statistical power because of these inherent properties.

The smaller number of QTL detected with SF is consistent with this potential explanation. Additional benchmarking with real and simulated traits would be helpful to further address this issue.

SF has the advantage of producing null distributions that do not have to be normalized for trait variance (Fig. 2.2). A potential approach to combine this property with predictor-scores that are not limited in the fashion described above and are not interdependent would be to compute how many samples were grouped with the predictor across the whole forest. The number of samples that are used as a training-set per tree is completely independent of their variance. The total sum the scores across all predictors would therefore also be independent of the variance. Since samples are grouped more than once in the process of growing the decision tree, they could increase the score for several predictors per tree. This would reduce the interdependency of the scores. The proposed importance measure would not limit the scores of informative predictors since it is proportional to the number of samples in the node that were split by the predictor. The problem of overfitting might also be lessened since small nodes which are more prone to overfitting would have less weight in the proposed importance measure.

A study by Michaelson *et al.* concluded that SF is more suited than PI or RSS for eQTL-mapping (Michaelson *et al.*, 2010). This conclusion is not supported by the results presented in this work. Michaelson *et al.* analyzed both simulated traits and gene-expression data from mouse- and yeast-datasets with the use of SF, PI and RSS in addition to other mapping approaches including composite interval mapping, Haley-Knott regression, LASSO and the elastic net. In the case of the simulated traits the top percentile of loci of each score was analyzed for the presence of the loci for which effects were simulated. The top percentile of loci for each method in the expression traits was analyzed for enrichment in *cis*-eQTL and enrichment in loci that share a functional annotation with the trait they influence. SF was normalized by subtracting the expected deviation from the mean under the null hypothesis. This was done to account for a marker-specific bias which influences the selection frequencies of markers under the null hypothesis depending on the correlation between markers. This bias is most prominent in SF but can also be observed in RSS and PI (Fig 2.1b). In this work the marker-specific biases are accounted for by the generation of marker-specific null-distributions for all importance measures.

Michaelson *et al.* found the corrected SF to score loci that a share a functional annotation with their target more often in the top percentile of all results. The performance of RSS and PI could potentially be improved by accounting for the marker-specific bias which in turn might change the performance of these measures. The comparably poorer performance of PI and RSS could be the result of the lack of marker-bias correction for these measures.

The results presented in this work suggest that SF might be less suited for eQTL-mapping than PI or RSS. The product of PI and RSS detected the most eQTL affecting the expression of *S. pombe* genes in this cross. To ensure that differences in mapping-performance are not due to the normalization of extracted PI- and RSS- scores with the trait variance remapping of a selection of traits with trait-specific null-distributions would be necessary. The null-distributions of PI- and RSS-scores are highly dependent on the variance in the trait and they can only be pooled to generate larger null-distributions if these trait-specific differences are accounted for. The pooling of the null-distributions reduces computation time drastically.

Like other methods for QTL-mapping, RF is highly dependent on the amount of samples. To investigate the relationship between the mapping results and sample-size different subsets of a larger dataset could be used for mapping. Suitable data for growth traits that could be analyzed in this fashion is readily available (Bloom *et al.*, 2013).

4.2 Detecting epistasis with a split asymmetry approach

Contributions to a quantitative trait like body height can be dependent on epistatic interactions between genetic loci (Fisher, 1918). As a multivariate approach to QTL-mapping, Random Forest is better suited to account for these interactions than univariate approaches (Michaelson *et al.*, 2010). Identifying these interactions not only helps in explaining trait variance but also improves our understanding of the underlying biological system (Costanzo *et al.*, 2010; Michaelson, 2010; Picotti *et al.*, 2013; Schuldiner *et al.*, 2005; Urbanowicz *et al.*, 2012).

We developed a new approach of detecting epistatic interactions based on a previously published method (Michaelson, 2010; Picotti *et al.*, 2013). The method presented in this work compares the splitting behaviors of a predictor in Random Forest depending on earlier splits to detect interactions in an analytical fashion.

To assess the performance of this approach, a benchmark of simulated traits was developed. This *in silico* benchmark addresses several factors that have the potential to influence the performance of a method that aims to detect epistasis. The benchmark includes both AND- and XOR-epistatic interactions of different effect sizes, purely additive loci of different effect sizes, combinations of these contributions and noise. The traits were simulated for real genotypes including highly correlated markers, missing allele combinations and replicates. The benchmark can be used to test the performance of any method that aims to detect epistatic interactions affecting QTs.

As an example of a widely used two-dimensional method the performance of a two-way ANOVA was assessed.

The split asymmetry approach outperformed the ANOVA in simulated scenarios that combined AND-epistasis with the presence of additive effects on one of the interacting loci (Fig. 3.8a and b, Fig. 3.10a and c). Scenarios combining an AND-epistatic interaction with large additive contributions to the trait on a predictor other than epistatic ones were recovered at a considerably higher rate by the split asymmetry approach (Fig. 3.10d). XOR-epistasis was only recovered at a high rate if for one of the interacting loci a purely additive effect was simulated as well (Fig. 3.9b).

The performance of the ANOVA was affected by both the effect size of the epistatic interaction and also the presence of additive effects. If additive effects were on the predictor-pair that was tested the ANOVA was able to better account for those effects and the additive effect size had a minor influence on its performance (Fig. 3.10a and c). Additive effects outside the tested pair of predictors decreased the performance of the ANOVA depending on their size. While small effects outside the tested pair of predictors or combinations of several small effects decreased the performance of the ANOVA to a lesser extent(Fig 3.7a-d), large additive effects on predictors other than the pair that was tested were much more problematic for the ANOVA to handle (Fig. 3.10d). XOR-epistasis was generally detected at a higher rate than AND-epistasis (Fig. 3.5, Fig 3.6 and Fig 3.9).

In the *in silico* benchmark the ANOVA outperformed the split asymmetry approach in scenarios addressing XOR-epistasis. When one of the interacting loci had an additive effect as well the performance of the split asymmetry approach improved (Fig. 3.9b). The lower rate of recovery for XOR-simulations most likely stems from the small main effects of the interacting predictors. In the case of AND-epistasis the predictors have a main effect proportional to the epistatic effect and the amount of samples that are affected. XOR-epistasis does not necessarily increase the main effect any predictor involved. Since Random Forest uses the predictor among a sample that explains the most variance by splitting, predictors with small main effects will rarely be chosen close to the root of the tree. Although the difference between the means of the child-nodes of the split with the other epistatic predictor will be large on both sides the two predictors will rarely be used in the same tree because of their small main effects. If one of the epistatic predictors also has a large additive effect, XOR-epistatic interactions are detected much easier. Preliminary tests suggest that the recovery of the XOR-simulations with the split asymmetry approach can be improved by increasing the forest size.

Although XOR-epistasis is often included in simulations, its biological relevance appears to be much smaller than that of AND-epistasis, which is frequently detected (Bloom *et al.*, 2013; Clement-Ziza *et al.*, 2014; Wan *et al.*, 2013).

Although the *in silico* benchmark helps to investigate the influence of different factors on the performance of these methods separately it differs from real data in some aspects. The simulated traits are much less complex than some biological traits in that they only include a small amount of different additive and/or epistatic effects. The performance in the simulated traits suggest that the split asymmetry approach is especially suited to handle large additive effects in a trait both on the epistatic loci and on loci other than the epistatic ones. Quantitative traits can contain large and numerous effects and often do (Ackermann *et al.*, 2013; Bloom *et al.*, 2013; Clement-Ziza *et al.*, 2014; Picotti *et al.*, 2013; Willemsen *et al.*, 2004). The part of the trait variance that is due to epistatic contributions is typically smaller than the additive portion (Bloom *et al.*, 2013; Crow, 2010; Maki-Tanila and Hill, 2014).

The *in silico* benchmark was complemented with a benchmark of validated interactions in real traits. The approaches were benchmarked by using them to detect interactions in expression traits of essential genes measured in a library of *S. cerevisiae* crosses (Brem and Kruglyak, 2005). These results were compared to experimentally validated interactions affecting growth traits that were used as a gold standard (Costanzo *et al.*, 2010). Although not all interactions that affect the expression of an essential gene affect a growth trait and not all epistatic interactions that affect growth traits also affect gene expression, it is reasonable to assume that a substantial amount of interactions affecting the expression of essential genes also affect growth traits. When predictor-pairs were ordered by their *p*-value determined by the ANOVA, the interactions from Costanzo *et al.* were distributed across the interactions at a nearly random rate (Fig 3.11). Ordering predictor-pairs by their *p*-value determined with the split asymmetry approach resulted in interactions that significantly affect growth traits being on average ranked higher.

The higher rate of recovery for epistatic interactions validated by Costanzo *et al.* with the split asymmetry approach suggests that this method is more suitable to detect epistatic interactions in quantitative traits than exhaustive approaches.

Although the split asymmetry approach outperformed the ANOVA when applied to expression data, it has its limitations. The power with which interactions between predictors can be detected depends on the number of times the second marker is applied to each subgroup of individuals separated by the first predictor. A challenge to the detection of AND-epistasis is the frequency at which the second marker of a pair is used on that subgroup that does not contain individuals who are targeted by the epistatic effect. If individuals with the allele-combination *ab* are targets of an AND-epistatic effect and predictor A is used first to split, the phenotypical difference between *ab* and *aB* will be much larger than that between *Ab* and *AB*. If the distribution of differences between *Ab* and *AB* is too small it cannot be tested against the distribution of *ab* and *aB*. In our simulations the recovery of AND-epistatic interactions was higher if one of the markers had an additive effect as well. This can be attributed to the overall greater number of splits with and the earlier use of the predictor with the bigger main effect. The second predictor may be used in nodes that are further away from the root of the tree containing fewer individuals. These nodes are more influenced by noise than nodes with many samples and overfitting can lead to predictors with main effects that are overall small being chosen for the split. The slopes that are collected are those of OOB-samples and are therefore not influenced by the overfitting. Because predictors with large main effects are used early on in the tree, more small nodes follow their splits than splits by predictors with smaller main effects. If one predictor of a pair has a large main effect it is easier to collect distributions with sufficient size for testing.

While the split asymmetry approach tests marker-pairs in a qualitative manner, AND-epistatic interactions could potentially also be detected by analyzing on which side of the split with one marker the other marker was used more often. The proposed approach would exploit the fact that AND-epistasis only contributes to the trait values of one allele-combination. Approaches that test split asymmetry in a qualitative and in a quantitative manner could complement each other.

While the method presented in this work is similar to the one by Michaelson *et al.*, it offers several improvements. The scores derived from the method of Michaelson *et al.* are highly dependent on selection frequencies of predictors and the effect size of the interaction in relation to the main effect of the second predictor that is used. Michaelson *et al.* only use the information of splits directly following each other, while the method presented in this work also uses splits that are separated by one or more other splits. This greatly increases the number of splits that can be examined per tree.

By providing an analytical framework to examine differences in splitting behavior of predictors dependent on other predictors, the need of large scale permutations of quantitative data to gain null distributions for the predictor-pairs is relieved. This reduces the runtime of the method.

4.3 Conclusions

The comparison of different importance measures in Random Forest for eQTL-mapping revealed a general agreement on the most important predictors for a trait between the different measures. Although the measures prioritized predictors in a similar fashion, PI, RSS and their combinations C1 and C2 appear to be more powerful than SF to significantly detect predictors that explain trait variance. Contrasting conclusions gained from earlier work might be due to not accounting for a predictor-specific bias in PI and RSS but doing so for SF. The differences between SF and the other measures could be further investigated using a simulation based benchmark.

The approach to detect epistatic interactions from quantitative data presented in this work outperformed the ANOVA when applied to expression data from a library of yeast crosses. The split asymmetry approach also performed better in simulated scenarios that include large additive effects, which is often observed in real data. The method presented here improves on an earlier implementation that detects split asymmetry by Michaelson *et al.* by providing an analytical framework in the form of Student’s t-tests applied to distributions of differences between the means of child-nodes. The practical use of the method could be further demonstrated by applying it to available data-sets (Bloom *et al.*, 2013; Clement-Ziza *et al.*, 2014; Picotti *et al.*, 2013; Smith and Kruglyak, 2008). The *in silico* benchmark used to assess the performance of the approaches can be employed for any method that aims to detect epistasis from quantitative traits.

5. References

Ackermann, M., Sikora-Wohlfeld, W., and Beyer, A. (2013). Impact of Natural Genetic Variation on Gene Expression Dynamics. PLoS Genet. *9*, e1003514.

Albert, F.W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. Nat. Rev. Genet.

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*, 234–237.

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

Brem, R.B., and Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc. Natl. Acad. Sci. U. S. A. *102*, 1572–1577.

Brem, R.B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic Dissection of Transcriptional Regulation in Budding Yeast. Science *296*, 750–752.

Chun, H., and Keles, S. (2009). Expression Quantitative Trait Loci Mapping With Multivariate Sparse Partial Least Squares Regression. Genetics *182*, 79–90.

Clement-Ziza, M., Marsellach, F.X., Codlin, S., Papadakis, M.A., Reinhardt, S., Rodriguez-Lopez, M., Martin, S., Marguerat, S., Schmidt, A., Lee, E., et al. (2014). Natural genetic variation impacts expression levels of coding, non-coding, and antisense transcripts in fission yeast. Mol. Syst. Biol. *10*, 764–764.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L.Y., Toufighi, K., Mostafavi, S., et al. (2010). The Genetic Landscape of a Cell. Science *327*, 425–431.

Crow, J.F. (2010). On epistasis: why it is unimportant in polygenic directional selection. Philos. Trans. R. Soc. B Biol. Sci. *365*, 1241–1244.

Cyr, D.D., Lucas, J.E., Thompson, J.W., Patel, K., Clark, P.J., Thompson, A., Tillmann, H.L., McHutchison, J.G., Moseley, M.A., and McCarthy, J.J. (2011). Characterization of Serum Proteins Associated with IL28B Genotype among Patients with Chronic Hepatitis C. PLoS ONE *6*, e21854.

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

Engel, S.R., Dietrich, F.S., Fisk, D.G., Binkley, G., Balakrishnan, R., Costanzo, M.C., Dwight, S.S., Hitz, B.C., Karra, K., Nash, R.S., et al. (2014). The Reference Genome Sequence of Saccharomyces cerevisiae: Then and Now. G3amp58 GenesGenomesGenetics *4*, 389–398.

Fisher, R.A. (1934). Biological Monographs And Manuals Vol-5.

Francesconi, M., and Lehner, B. (2013). The effects of genetic variation on gene expression dynamics during development. Nature *505*, 208–211.

Haley, C.S., and Knott, S.A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity *69*, 315–324.

Hemani, G., Shakhbazov, K., Westra, H.-J., Esko, T., Henders, A.K., McRae, A.F., Yang, J., Gibson, G., Martin, N.G., Metspalu, A., et al. (2014). Detection and replication of epistasis influencing transcription in humans. Nature *508*, 249–253.

Heyn, H., Sayols, S., Moutinho, C., Vidal, E., Sanchez-Mut, J.V., Stefansson, O.A., Nadal, E., Moran, S., Eyfjord, J.E., Gonzalez-Suarez, E., et al. (2014). Linkage of DNA Methylation Quantitative Trait Loci to Human Cancer Risk. Cell Rep. *7*, 331–338.

Holdt, L.M., von Delft, A., Nicolaou, A., Baumann, S., Kostrzewa, M., Thiery, J., and Teupser, D. (2013). Quantitative Trait Loci Mapping of the Mouse Plasma Proteome (pQTL). Genetics *193*, 601–608.

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

Khan, S.A., Chibon, P.-Y., de Vos, R.C.H., Schipper, B.A., Walraven, E., Beekwilder, J., van Dijk, T., Finkers, R., Visser, R.G.F., van de Weg, E.W., et al. (2012). Genetic analysis of metabolites in apple fruits indicates an mQTL hotspot for phenolic compounds on linkage group 16. J. Exp. Bot. *63*, 2895–2908.

Kuhn, M. (2013). Introducing parallelRandomForest: faster, leaner, parallelized.

Lappalainen, T., Sammeth, M., Friedländer, M.R., ‘t Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest. R News *2*, 18–22.

Loguercio, S., Overall, R.W., Michaelson, J.J., Wiltshire, T., Pletcher, M.T., Miller, B.H., Walker, J.R., Kempermann, G., Su, A.I., and Beyer, A. (2010). Integrative Analysis of Low- and High-Resolution eQTL. PLoS ONE *5*, e13920.

Mackay, T.F.C. (2013). Epistasis and quantitative traits: using model organisms to study gene–gene interactions. Nat. Rev. Genet. *15*, 22–33.

Maki-Tanila, A., and Hill, W.G. (2014). Influence of Gene Interaction on Complex Trait Variation with Multilocus Models. Genetics *198*, 355–367.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

Michaelson, J. (2010). Applications and extensions of Random Forests in genetic and environmental studies (Dresden, Techn. Univ., Diss., 2010).

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

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

Nagtegaal, A.P., Spijker, S., Crins, T.T.H., Neuro-Bsik Mouse Phenomics Consortium, and Borst, J.G.G. (2012). A novel QTL underlying early-onset, low-frequency hearing loss in BXD recombinant inbred strains: Novel gene locus for early-onset hearing loss. Genes Brain Behav. n/a–n/a.

Nelson, R.M., Pettersson, M.E., Li, X., and Carlborg, Ã. (2013). Variance Heterogeneity in Saccharomyces cerevisiae Expression Data: Trans-Regulation and Epistasis. PLoS ONE *8*, e79507.

Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. PLoS Genet. *6*, e1000888.

Pannebakker, B.A., Watt, R., Knott, S.A., West, S.A., and Shuker, D.M. (2011). The quantitative genetic basis of sex ratio variation in Nasonia vitripennis: a QTL study: Nasonia sex ratio QTLs. J. Evol. Biol. *24*, 12–22.

Picotti, P., Clément-Ziza, M., Lam, H., Campbell, D.S., Schmidt, A., Deutsch, E.W., Röst, H., Sun, Z., Rinner, O., Reiter, L., et al. (2013). A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. Nature *494*, 266–270.

Raadsma, H.W., Jonas, E., McGill, D., Hobbs, M., Lam, M.K., and Thomson, P.C. (2009). Mapping quantitative trait loci (QTL) in sheep. II. Meta-assembly and identification of novel QTL for milk production traits in sheep. Genet. Sel. Evol. *41*, 45.

Rockman, M.V., and Kruglyak, L. (2006). Genetics of global gene expression. Nat. Rev. Genet. *7*, 862–872.

Ronald Aylmer Fisher (1918). The correlation between relatives on the supposition of Mendelian inheritance.

Saccharomyces Genome Deletion Project http://www-sequence.stanford.edu/group/yeast_deletion_project/deletions3.html.

Schuldiner, M., Collins, S.R., Thompson, N.J., Denic, V., Bhamidipati, A., Punna, T., Ihmels, J., Andrews, B., Boone, C., Greenblatt, J.F., et al. (2005). Exploration of the Function and Organization of the Yeast Early Secretory Pathway through an Epistatic Miniarray Profile. Cell *123*, 507–519.

Shpak, M., Hall, A.W., Goldberg, M.M., Derryberry, D.Z., Ni, Y., Iyer, V.R., and Cowperthwaite, M.C. (2014). An eQTL analysis of the human glioblastoma multiforme genome. Genomics *103*, 252–263.

Smith, E.N., and Kruglyak, L. (2008). Gene–Environment Interaction in Yeast Gene Expression. PLoS Biol. *6*, e83.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional Variable Importance for Random Forests. BMC Bioinformatics *9*, 307.

Suthram, S., Beyer, A., Karp, R.M., Eldar, Y., and Ideker, T. (2008). eQED: an efficient method for interpreting eQTL associations using protein networks. Mol. Syst. Biol. *4*.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. J. R. Stat. Soc. *58*, 267–288.

Urbanowicz, R.J., Kiralis, J., Sinnott-Armstrong, N.A., Heberling, T., Fisher, J.M., and Moore, J.H. (2012). GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. BioData Min. *5*, 1–14.

Urbany, C., Stich, B., Schmidt, L., Simon, L., Berding, H., Junghans, H., Niehoff, K.-H., Braun, A., Tacke, E., and Hofferbert, H.-R. (2011). Association genetics in Solanum tuberosum provides new insights into potato tuber bruising and enzymatic tissue discoloration. BMC Genomics *12*, 7.

Veyrieras, J.-B., Kudaravalli, S., Kim, S.Y., Dermitzakis, E.T., Gilad, Y., Stephens, M., and Pritchard, J.K. (2008). High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation. PLoS Genet. *4*, e1000214.

Wan, X., Yang, C., Yang, Q., Zhao, H., and Yu, W. (2013). The complete compositional epistasis detection in genome-wide association studies. BMC Genet. *14*, 7.

Willemsen, G., Boomsma, D.I., Beem, A.L., Vink, J.M., Slagboom, P.E., and Posthuma, D. (2004). QTLs for height: results of a full genome scan in Dutch sibling pairs. Eur. J. Hum. Genet. *12*, 820–828.

Zeng, Z.-B. (1994). Precision Mapping of Quantitative Trait Loci. Genetics *136*, 1457–1468.