universität
wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Discovering nonlinear contributions from multi-omics to phenotypic variations“

verfasst von / submitted by

Michelle Binsfeld, B.Sc.

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2021 / Vienna, 2021

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 821
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Mathematik
Betreut von / Supervisor:	Univ.-Prof. Dr. Joachim Hermisson

Contents

List of Figures

List of Tables

Zusammenfassung

Das Ziel von Assoziationsstudien ist die Identifizierung von Polymorphismen, die im Zusammenhang mit einem komplexen Merkmal stehen. Die Standardmodelle hierfür sind lineare Modelle, die auf sogenannten “Single-Locus” Tests beruhen. Deswegen können Gene mit nicht-linearem Effekt, sowie Gene, die zwar mit anderen Genen interagieren, aber selbst keine marginale Effekte zeigen, nicht identifiziert werden. Die meisten komplexen Merkmale sind jedoch durch multiple Gene und ihre wechselwirkende Effekte bestimmt. Wir zeigen, dass ein nicht-lineares Modell, das auch Interaktionen und Heterogenität erfasst, mehr Gene identifizieren kann als das Standardmodell. Hierfür haben wir eine Pipeline entwickelt, die auf Random Forests beruht und die Anzahl der genetischen Marker (Prädiktorvariablen) in jedem Schritt reduziert. Dies ist eine vielversprechende Alternative zu den linearen Modellen in genomweiten (GWAS) und transkriptomweiten (TWAS) Assoziationsstudien. Weiterhin erlauben Random Forests es, eine neue Assoziationsstudie (GTWAS) zu betrachten, indem genetische und transkriptomische Variation in ein Modell integriert werden.

Abstract

The aim of association studies is to understand the genetic architecture of complex traits, which is a fundamental issue in genetics. The standard approaches are based on linear mixed models using single-locus tests. Therefore, they might miss nonlinear loci that are involved in interactions but do not show significant marginal effects. Many complex phenotypes however are determined by multiple loci and their interactive effects. We show that a model, which allows for nonlinear effects, epistatic interactions and heterogeneity, increases the association power. Due to the high-dimensionality of omics data, the search of epistatic interactions is challenging. We have developed a pipeline based on random forests that reduces the search space for significant associations and interactions in each step. It is a promising alternative to the traditional approaches in genome-wide (GWAS) and transcriptome-wide (TWAS) association studies and furthermore allows us to construct genome-and-transcriptome-wide-association studies (GTWAS), which integrate both genotypic and transcriptomic variation.

Acknowledgements

First and foremost, I want to thank my supervisor Haijun Liu for all his support. The completion of this thesis would not have been possible without his expertise and encouragement. He was always available when I needed his help and guided me through the research process. Furthermore, I also want to acknowledge Joachim Hermisson for providing feedback and for his valuable comments on the thesis. Then, I want to express my gratitude to Magnus Nordborg and the whole Nordborg group for the opportunity I was given to write my master thesis at the Gregor Mendel Institute of Molecular Plant Biology. I enjoyed being a part of the group and learned a lot during the time at the institute. Special thanks go to Yoav for providing his data and to Almudena for helping with the administrative part and for reading the manuscript. Finally, I would like to thank my family, my friends and mostly Vinzenz for supporting and encouraging me through the process of researching and writing this thesis.

Author contributions

H.L. designed the project and encouraged M.B. to investigate random forests in association studies. H.L. supervised the findings of the work while suggesting the direction of the research process. M.B. developed the pipeline with all the necessary data preprocessing steps and performed the computations. H.L. verified the statistical methods and commented on the pipeline. M.B. and H.L. discussed the results and H.L. helped with the biological interpretation. M.B. wrote the manuscript. J.H. provided feedback and comments on the manuscript, which has been read and approved by H.L. and J.H..

Chapter 1

Introduction

1.1 Biology

1.1.1 Phenotype, genotype and gene expression

A *phenotype* is an observable characteristic of an individual which can be morphological, physiological or behavioral [1, 2]. Phenotypes influence the viability or the reproductive success of an individual and may be more or less adapted to the environment [2]. Phenotypic variation between individuals of the same species is essential for natural selection and thus evolution. Understanding this variation is necessary in many areas as for example crop improvement, breeding or disease treatment strategies. Phenotypes are not directly inherited but are influenced by heritable and non-heritable factors. The non-heritable factors are external to the organism and are referred to as environmental factors. The *genotype* is the part of an individual which is directly heritable. It is encoded in its *genome* in the form of *deoxyribonucleic acid* (DNA) [2]. The amount of phenotypic variation that can be explained by the genotype is called the *heritability* of a trait [1].

DNA is organized in two strands twisted around each other forming a double helix. Each strand is a polymer that consists of smaller units called *nucleotides*, which differ in the four bases that they contain: adenine (A), guanine (G), thymine (T) and cytosine (C). The two DNA strands are held together by hydrogen bonds formed by A-T and G-C base pairings and are complementary to each other. In eukaryotes[1], the nucleus contains several DNA strands which are called *chromosomes* [2]. Each chromosome is represented as a sequence composed of the four letters A, G, T and C, which correspond to the bases of one of the complementary strands [1]. Given some individuals of the same species, one usually is only interested in positions of the DNA sequence differing between them as these represent a potential source of genetic variability which could explain phenotypic variation. Generally, most of the DNA sequence is identical between different individuals of a given species (e.g. 99.9% of the base

[1] Eukaryotes are organisms whose cells contain a nucleus enclosed with a nuclear envelope.

pairs are identical for humans). Genetic loci[1] showing differences between individuals are called *polymorphic* and the different variants which can be observed at these loci are called *alleles.* One example of genetic variants are *single-nucleotide polymorphisms* (SNPs), which are the result of point mutations in a single base pair [1]. Organisms can have different levels of *ploidy*, where haploid organisms have a single copy of each chromosome, diploids have two copies of each chromosome and some organisms even have higher levels of ploidy. For haploids, a single-locus genotype is determined by a single allele while for diploids, it is given by a pair of alleles [2]. An individual carrying two identical alleles is called *homozygous* while an individual that carries two different alleles is called *heterozygous* [1].

The genotype can be divided into functional units called *genes.* They carry the genetic information that can be turned into functional products, which usually are proteins. There is an important bridge between the genotype and the functional product which is *gene expression* [3]. Gene expression includes two main sequential steps. During *transcription*, DNA is transcribed into *ribonucleic acid* (RNA) and during *translation*, RNA is translated into proteins. Transcription is the key process which controls whether a gene is expressed or not, and also its quantification level. All tissues and cells in an individual contain the same DNA (ignoring somatic mutations) but different expression patterns create functional diversity [4]. Furthermore, given the same tissue, different gene expressions between individuals widely exist and can lead to phenotypic variation. High-throughput sequencing technologies have not only enabled the profiling of genetic variants but also of transcriptomes (gene expression levels) [5]. This allows to study omic variation and its effect on phenotypic variation.

1.1.2 Arabidopsis thaliana

Arabidopsis thaliana is a flowering plant which is widely used as a model organism in plant biology and genetics. The first plants to be experimented on were collected in 1905. *Laibach* discovered that *A. thaliana* carries 5 pairs of chromosomes (within the nucleus) and pointed out the benefits (amongst others a small genome size, a short lifecycle, diversity, selfing,...) to adopt it as a model organism [6]. *A. thaliana* is a diploid organism but it is believed to be at least 99% selfing. This makes it mainly homozygous [7] and thus it also naturally exists as inbred lines. Furthermore, its high phenotypic variation makes *A. thaliana* an ideal model system.

Flowering time is a locally adaptive trait which is subject to strong selection and which is well studied [8]. In *A. thaliana*, the pathways controlling flowering time have been characterized, thus it is well-understood molecularly. Also genes responsible for natural variation have been identified. Using new computational approaches, findings can be validated by the abundant

[1] A *locus* is the position of a single base or on a broader level the position of a gene [2].

information on the molecular networks and by using prior knowledge of the 306 flowering time genes that have been characterized in the FLOR-ID database [9]. However, we claim that it is also possible to obtain novel knowledge on the genetic basis of flowering time.

1.2 Association studies

1.2.1 Linkage disequilibrium

According to Mendel's second law (law of *independent assortment*), genes are inherited independently of each other. However, this is only true if two genes are *unlinked*, i.e. if they are located on different chromosomes. Genes that are located on the same chromosome, and particularly genes that are close to each other, tend to be inherited together as it is unlikely that some recombination event occurred between them [2].

Linkage disequilibrium (LD) is defined as the non-random association of two alleles in haplotypes. In order to make disequilibria between pairs of alleles comparable, one often uses the difference between the observed and expected frequency of haplotypes or the squared correlation coefficient r^2 as indicators of LD [2]. As a result of this, one usually observes a correlation structure between loci in a population. This correlation structure is a well defined function of the distance[1] between nearby loci in populations. Due to selfing, *A. thaliana* shows extensive linkage disequilibrium decaying approximately within 50 kb [10]. This brings another advantage of selfing in association studies that it increases LD without decreasing polymorphism that much [7].

1.2.2 Population structure

Population structure refers to any deviation from random mating, which leads to differences in allele frequencies between subpopulations of a species. It often arises from geographical separation (e.g. mountains and oceans) but also from phenomena as inbreeding or associative mating. Even when there are no geographical barriers to gene flow, organisms tend to not randomly mate across the species range. They rather mate with individuals close to where they have been born and thus genetic and phenotypic differences can accumulate between subpopulations of that species [11]. *A. thaliana* accessions have been clustered based on their genotype [10]. The clusters separate the accessions along geographic boundaries indicating that there is global population structure. This means that individuals are more similar to individuals that grow close to each other than to individuals from far away, both genotypically and phenotypically.

[1] The distance between two loci is given as the number of base pairs between them.

1.2.3 Genome-wide association studies

A major goal in biology is to understand the genetic architecture of complex traits [8]. The genetic architecture underlying a trait refers to the genetic variants that are associated to the trait, including their number, their effect size, their allele frequencies and all the interactions between them. Therefore the aim of genome-wide association studies (GWAS) is to identify genetic variants that are responsible for phenotypic variation by mapping single-nucleotide polymorphisms to the phenotype [8]. The power of GWAS to identify an association between a SNP and a phenotype depends on the phenotypic variation within the population which is determined by how strongly two alleles differ in their phenotypic effect (i.e. their effect size) and their frequency in the population. Hence both rare variants and variants of small effect present problems for GWAS [12].

GWAS has actually been initiated by human geneticists. Their aim was to find the genetic basis of common human diseases. Due to advances in sequencing technology, GWAS are considered as the standard approach to also understand the genetic basis of complex traits of agricultural importance [13]. A highly polygenic or a complex trait is a phenotype which is influenced by many genetic variants [14]. Although the standard models in GWAS are based on single-marker and additive effects, most complex traits are associated to genes involving interactions between them and also between genetic and environmental factors [15].

GWAS profits from linkage disequilibrium as it allows to identify genetic markers that tag causal variants when they are in LD [16]. However, linkage disequilibrium can also make it difficult to differentiate between causal variants and linked neutral markers [12]. Furthermore, it is important to account for population structure in GWAS as it might lead to an increased rate of false positives.

1.2.4 Expression quantitative trait loci studies

Expression quantitative trait loci (eQTL) studies identify genetic variants that are responsible for variation in gene expression levels by studying gene expression as an intermediate phenotype. A locus that explains variation in expression is called an eQTL. Depending on the physical distance from the gene that they regulate, eQTLs are characterized as either *cis* or *trans* acting. *Cis* eQTLs regulate the expression of nearby genes while *trans* eQTLs regulate the expression of distant genes [17]. A genetic variant does not necessarily lead to gene expression changes and it might not affect a phenotype, but studies have shown that most genetic variants that do influence a phenotype do so by regulating the expression of the genes [3, 18].

1.2.5 Transcriptome-wide association studies

Transcriptome-wide association studies (TWAS) emerged as a promising technique to integrate omics data from expression measurements into association studies. Whereas GWAS associate phenotypic variation directly with genetic variants, TWAS map gene expression levels to the phenotype allowing to get further insights into the intermediary effects on the phenotype [5]. While for GWAS, the associations can not always directly be linked with the causal gene due to linkage disequilibrium, TWAS do not have this limitation.

Studies have shown that most single-nucleotide polymorphisms influence the phenotype by regulating the expression of genes [18]. However, there are scenarios where TWAS can identify phenotype-associated genes that are missed by GWAS and the other way around. If a population does not contain any polymorphic marker nearby a causal gene (e.g. *trans* eQTL), GWAS will miss this gene. If in this case expression variation of this causal gene contributes to phenotypic variance, it can be identified by TWAS. Furthermore, variation in expression leading to phenotypic variation does not have to be caused by an eQTL, but could also be due to epigenetic variation in or near the causal gene or due to the complex regulation consequence of other regulators for example. In these cases, the gene can not be identified by GWAS. Also TWAS can miss causal genes that can be identified by GWAS, for example genes that are not expressed at a sufficient level or genes for which the causal variant is not associated with expression variation.

1.2.6 Current achievements of GWAS and TWAS in A. thaliana

As *A. thaliana* is highly selfing, it naturally exists as inbred lines. Inbred lines have the advantage that once they have been genotyped, they can be phenotyped multiple times which allows to study different traits in different environments [13]. Therefore *A. thaliana* is a well-suited organism to conduct genome-wide association studies.

Nowadays, the full genome sequences of more than 1000 natural inbred lines (accessions), which are products of natural selection, are available [19]. This allows to quantify genome variation in a large sample of accessions [20]. So far, more than $44,600$ significant associations have been identified while studying 462 phenotypes (AraGWAS database[1]). Generally, transcriptome-wide association studies are less common in the literature than genome-wide association studies, however *Li et al.* recently published a study that shows that the TWAS results are complimentary to GWAS results and much less affected by linkage disequilibrium [21].

[1] `https://aragwas.1001genomes.org/`

1.3 The role of epistasis and heterogeneity

The term *epistasis* has been used for the first time by *Bateson* in 1909 [22]. *Bateson* had in mind biomolecular interactions at the level of an individual when referring to epistasis ("biological epistasis") [23]. *Fisher* however had a rather statistical definition of epistasis, which is the deviation from an additive effect of two alleles at different loci ("statistical epistasis") [24]. This means that their contribution to a quantitative phenotype differs from the sum of their individual (or marginal) effect [25]. An example would be one locus preventing another locus from manifesting its effect, which could be interpreted as a multi-locus extension of the dominance effect for alleles within the same locus [26]. Generally, two genes interact if the effect of one gene depends on the state of the other gene [27]. Epistasis or gene-gene interactions have been discussed in detail in [28].

Although genome-wide association studies could identify many genetic variants associated with complex traits, the single identified genetic variants do not account for much of the heritability of phenotypes. The authors in [29] have discussed potential sources for this missing heritability. Epistatic interactions are assumed to account for a substantial fraction of missing heritability and to be important to understand the genetic architecture of complex traits [30, 28]. Also in *A. thaliana*, epistatic interactions are likely to make a contribution to the phenotype [31]. Generally, the presence of epistatic effects lowers the power of the standard models, which are based on single-marker tests. If for example the effect of one locus is masked by effects of another locus, the power to detect the first locus is reduced [26]. For more than two loci involved in an epistatic interaction, the situation is even more complicated.

Genetic heterogeneity describes the situation where different loci in different genes produce the same phenotypic effect [12]. As many phenotypes are involved in local adaptation (e.g. flowering time), the results of association studies depend on the origin of the samples. If the samples come from a wide range of environments, the power of GWAS might be decreased due to increased genetic heterogeneity. It has been shown that for local *A. thaliana* subpopulations, the phenotypes are associated with different genetic variants. Some SNPs are actually common (same minor allele frequency) in all subpopulations but only show a sign of association for some subpopulations. This is probably due to differences in the broader genetic background or to epistatic interactions with other loci [8]. So at the end, genetic heterogeneity actually falls within a definition which is interpretable as epistasis [26]. The fact that some genetic variants only show an association for some subpopulations or accessions that have certain alleles at other loci lowers the power of traditional models as they generally lump an entire sample into a single group to assess average effects [32].

Chapter 2

Theory

2.1 Linear mixed models

A linear mixed model (LMM) is an extension of the linear regression model where variables are divided into fixed and random effects [14]. While fixed effects are modelled as regression parameters that are constant across individuals, random effects are modelled as being drawn from a distribution and thus vary. This allows to account for random effects without having to estimate a value for them [14]. The LMM can be described by the following model

$$y = X\beta + Zu + \epsilon \tag{2.1}$$

where $y \in \mathbb{R}^n$ represents the outcome variable, $X \in \mathbb{R}^{n \times p}$ the intercept and the predictor variables, $\beta \in \mathbb{R}^p$ the regression coefficients and $\epsilon \in \mathbb{R}^n$ with $\epsilon \sim \mathcal{N}(0, \mathbb{I}\sigma_\epsilon^2)$ the error. While these terms are known from standard linear regression, $Z \in \mathbb{R}^{n \times m}$ represents another set of predictor variables. This time however, we are not interested in specified estimates of the coefficients but use the associated random effects $u \in \mathbb{R}^m$ with $u \sim \mathcal{N}(0, \mathbb{I}\sigma_u^2)$ instead [14]. Thus the model can be expressed as a conditional distribution of the outcome variable y

$$y|u \sim \mathcal{N}(X\beta + Zu, \mathbb{I}\sigma_\epsilon^2), \quad u \sim \mathcal{N}(0, \mathbb{I}\sigma_u^2). \tag{2.2}$$

Under the distributional conditions from above, we get the following unconditional model[1]

$$y \sim \mathcal{N}(X\beta, ZZ^T\sigma_u^2 + \mathbb{I}\sigma_\epsilon^2). \tag{2.3}$$

In equation (2.3), the fixed effects enter the mean while the matrix Z and the random effects enter as a component of the covariance matrix [14]. If we define $\sigma_g^2 = m\sigma_u^2$, $G = \frac{1}{m}ZZ^T$ and use $g \sim \mathcal{N}(0, G\sigma_g^2)$ instead of Zu, then the model can be rewritten as

$$y = X\beta + g + \epsilon. \tag{2.4}$$

[1] $\mathbb{E}[y] = \mathbb{E}[X\beta] = X\beta$ and $Var(y) = Var(Zu) + Var(\epsilon) = ZZ^T\sigma_u^2 + \mathbb{I}\sigma_\epsilon^2$ as $X\beta$ and Z are fixed and furthermore, the residuals ϵ and the random effects u are independent of each other.

If the specific values of Z are unknown but G can be calculated or estimated and if one is mainly interested in estimating β, this representation of the LMM is more useful [14].

The aim in association studies is to identify genetic markers (e.g. SNPs or genes' expression, as predictor variables) that are significantly associated with the phenotype (outcome variable). Assume that we want to test a single SNP x_1 for an association with the phenotype with some univariate model

$$y = \beta_0 + x_1\beta_1 + \epsilon \tag{2.5}$$

to estimate the effect β_1 and test its significance, while the true model is polygenic and thus it should be $y = u_0 + x_1u_1 + \sum_{k>1} x_k u_k + \eta$ [14]. If we run a univariate model as in (2.5), the effect of the other markers is ignored and is modelled as part of the error, i.e. $\epsilon = \sum_{k>1} x_k u_k + \eta$. Due to linkage disequilibrium and population structure, this results not only in a correlation between the tested SNPs and the error term but also between the errors of observations from the same subpopulation (ϵ_i and ϵ_j, i≠j are not independent if observation i and j belong to the same subpopulation), which results in false positives [14]. In order to solve this problem, linear mixed models proved to be a robust and reliable method for genome-wide association studies as they can correct for confounding effects [33]. LMMs treat the SNPs that are not directly tested for an association as random effects [14]. The effects u_k are treated as identically and independently distributed random variables drawn from $\mathcal{N}(0, \sigma_u^2)$ and thus are not modelled as fixed parameters. Therefore $g = \sum_{k>1} x_k u_k$ represents an additional genetic effect with variance σ_g^2 such that the model becomes

$$y = \beta_0 + x_1\beta_1 + g + \epsilon. \tag{2.6}$$

For genetically similar individuals, the genetic effects are positively correlated [14]. This is actually the same formulation than we got in (2.4) for the LMM. The matrix G is obtained by calculating a kinship matrix which measures the correlation between two individuals using the observed genotypes. However, any relationship matrix used to correct for population structure only represents an approximation for the truly underlying genetic background [34]. The linear mixed model-based method in TWAS includes all gene expression levels, except those in close physical proximity to the tested gene, as random effects to account for the confounding effects, including the correlations among distal genes [35]. The error term ϵ in a LMM is usually referred to as the environmental effect combining effects which are not measured and random noise [14]. The main use of linear mixed models in association studies is to estimate fixed and random effects (using maximum likelihood methods) and to test for significant associations. The most common approach to estimate the variances of random effects, i.e. the variance components is called restricted maximum likelihood (REML) as it can produce unbiased estimates of variance and covariance parameters [36]. These estimates are then used to test hypotheses about the fixed effect of the form $H_0 : \beta = 0$.

2.2 Machine learning

There are two major goals in mathematical modeling which include inference and prediction. Inference is about creating a model in order to understand the underlying mechanisms of the system or to test a hypothesis on how the system behaves. Prediction is about forecasting future behaviour or unobserved outcomes of the system [37]. Although statistical and machine learning models can be used for both, statistical models (e.g. linear mixed models) usually rather focus on making inference while machine learning models (e.g. random forests) rather concentrate on prediction.

Statistical models are based on making distributional assumptions and on restricting the complexity of the model in order to make it interpretable. Machine learning does not assume any model, which often results in better prediction accuracy [38]. However, machine learning models can also be used to make inference while offering a non-parametric alternative to traditional statistical models. They can capture nonlinear relationships between predictors and the outcome and furthermore they also capture interactions between predictors that contribute to the outcome. Thus they have more flexibility to learn the patterns from the data. In machine learning terminology, an association study would be a supervised regression problem (for quantitative traits) [39].

2.3 Random forests

A random forest is a fully non-parametric, tree-based machine learning algorithm which has been introduced by *Breiman* in 2001 [40]. It involves growing an ensemble of m_{tree} decision trees and can be used for both classification (discrete outcome) and regression (continuous/quantitative outcome). Each tree in the random forest is built on a bootstrap sample of the observations and at each node a random subset of variables is chosen as potential splitting candidates. In order to describe the theoretical aspects of the algorithm, we will rely on the two books [41] and [42].

2.3.1 The basic principle of random forests

Let us consider a regression problem and assume that the data consists of p predictor variables $x_1, ..., x_p$, one continuous outcome variable y and n observations, i.e. $(x_i, y_i)_{i=1}^{n}$ with $x_i = (x_{i1}, ..., x_{ip}) \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$. Starting at the root of the tree, a decision tree splits the p-dimensional space into two regions represented by nodes and in each node, the outcome is modelled by a constant which minimizes the squared error. This is done subsequently until a stopping condition is met and the leaf nodes have been reached. Therefore a decision tree is a model that partitions the p-dimensional variable space into K regions represented by the leaves of the tree.

Suppose that the p-dimensional variable space has already been partitioned into K distinct regions R_k where $k = 1, ..., K$ and $\cup_{k=1}^{K} R_k = \mathbb{R}^p$, $R_m \cap R_n = \emptyset, m \neq n$. A decision tree is defined such that in each region for some $x \in \mathbb{R}^p$, it fits a constant $c_k, k \in \{1, ..., K\}$ with

$$f(x) = \sum_{k=1}^{K} c_k \mathbb{1}(x \in R_k). \tag{2.7}$$

For regression problems, the goal is to minimize the loss function represented by the squared error $\mathcal{L}(y_i, \hat{y_i}) = \sum_{i=1}^{n} (y_i - \hat{y_i})^2$ with $f(x_i) = \hat{y_i}$ being the outcome which is predicted by the model. Assuming that the regions $R_1, ...R_K$ have already been identified, the squared error is minimized by choosing c_k as the mean outcome of all data points with $x_i \in R_k$, i.e.

$$\hat{c_k} = \frac{1}{n_k} \sum_{i:x_i \in R_k} y_i = \bar{y}_{R_k} \tag{2.8}$$

where $n_k = |\{i : x_i \in R_k\}|$. The proof can be found in appendix B.

Now the question is how in the first place the regions $R_1, ..., R_K$ are identified. At each node, the decision tree needs to decide on the splitting variable x_j, $j \in \{1, ..., p\}$ and the splitting point s which split the variable space into two smaller regions. The general objective function which needs to be minimized is $\sum_{k=1}^{K} \sum_{i:x_i \in R_k} (y_i - \hat{y_i})$. As this is computationally infeasible, decision trees use a Greedy algorithm, which is an algorithm that makes locally the best choice at each node and which is referred to as *recursive binary splitting*. Consider a splitting variable j and a splitting point s, and define the half-planes as

$$R_l = \{x_i | x_{ij} \leq s\} \quad \text{and} \quad R_r = \{x_i | x_{ij} > s\}. \tag{2.9}$$

At each node, the algorithm chooses the variable x_j and the splitting point s which solve the following minimization problem

$$min_{j,s}[\sum_{x_i \in R_l(j,s)} (y_i - \bar{y}_{R_l})^2 + \sum_{x_i \in R_r(j,s)} (y_i - \bar{y}_{R_r})^2] \tag{2.10}$$

where $\bar{y}_{R_l}$ is the mean outcome of the observations in $R_l(j, s)$ and $\bar{y}_{R_r}$ the mean outcome of those in $R_r(j, s)$. In other words, the variable x_j and the splitting point s which split the variable space into R_l and R_r are chosen such that the resulting split minimizes the squared error in each of the two resulting regions.

The determination of the best splitting point for each splitting variable is not computationally intensive. Hence the algorithm can scan through all possibilities in order to find the best pair (j, s). Once the best splitting pair has been found, the variable space is splitted into two distinct regions. This process is repeated on the two previously defined regions until some stopping criterion is fulfilled (see section 2.3.3).

Decision trees might perform well on the data which has been used to build the tree but they tend to overfit by building a tree which is too complex. Due to the hierarchical tree-building process, they are very sensitive to small changes in the data, especially for splitting variables at the top of the tree. The effect of some error at the root of the tree for example will affect all splits below it. In order to reduce the variance, random forests rely on bootstrap aggregation (or bagging), which is a general approach to reduce the variance of statistical learning models.

Assume that $X_1, ..., X_n$ are n independent and identically distributed random variables each with variance σ^2. Then $Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{\sigma^2}{n}$, thus the variance is reduced once the average over a set of independent and identically distributed random variables is taken. This idea can be applied to decision trees by taking multiple data sets from the population, building separate decision trees on each set and averaging the predictions from the trees. In random forests, m_{tree} bootstrap samples of size n are used to build m_{tree} decision trees. These are samples that are drawn uniformly with replacement from the n observations in the data. In average, each bootstrap sample consists of approximately two-third of the n observations[1]. The remaining one-third is called the out-of-bag sample.

Given that $m_{tree} = B$ trees have been grown on different boostrap samples, each with prediction function $f^{\star 1}(x), ..., f^{\star B}(x)$, the final prediction is made by averaging over all the trees

$$f_{ave}(x) = \frac{1}{B}\sum_{b=1}^{B} f^{\star b}(x). \tag{2.11}$$

Let the prediction of each decision tree $f^{\star b}$ be a random variable T_b. As the trees are trained on a sample from the same population, the random variables $T_1, ..., T_B$ are identically distributed with variance σ^2 but they are not independent. The correlation between them is given by $\rho = \frac{Cov(T_i,T_j)}{\sigma^2}$, i≠j. The variance of the random forest is then given by

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{2.12}$$

and the proof can be found in appendix B. For large B, the second term of the random forest variance vanishes. However, it still depends on the correlation between trees. Generally, decision trees in bagging tend to be highly correlated due to the fact that if there is for example one strong predictor x_j then most trees will split on x_j in early splits. The goal is to improve on bagging and to reduce the variance by decorrelating the trees. This is achieved by randomly choosing a subset of $m_{try} < p$ variables as splitting candidates from all of the p variables at each node. Moreover, this also reduces the search space for the best splitting variable. The algorithm can be summarised as follows [42]:

[1] Assume that there are n observations, then the chance of being drawn is given by $\frac{1}{n}$ and thus the probability of not being drawn among n trials is given by $(1 - \frac{1}{n})^n$. If n is big enough, one gets $lim_{n\to\infty}(1 - \frac{1}{n})^n = \frac{1}{e} \approx 0.37$. As each bootstrap sample is of size n, the other observations are duplicates.

Algorithm Random forest

```
1. For $b \in \{1, ..., B\}$:
    Draw a bootstrap sample with replacement of size $n$ from the observations.
    Grow a tree $T_b$ on the boostrap sample by repeating the following steps until a stopping criterion is reached:
        i. Select a random subset of $m_{try} < p$ from the p variables.
        ii. Pick the best splitting variable $x_j$ and splitting point $s$ among the $m_{try}$ variables.
        iii. Split the node into two daughter nodes $R_l$ and $R_r$.
2. Output the ensemble of decision trees $T_b$, $b \in \{1, ..., B\}$.
```

2.3.2 Out-of-bag sample

An important feature of random forests is that due to bagging only approximately two-third of the observations are used to train each tree, the other one-third being called the out-of-bag (OOB) sample. The predictive accuracy of the model can be assessed through the OOB sample, which removes the need of a test set and which is almost identical to n-fold cross-validation [40, 42]. In order to calculate the OOB error, all trees in the random forest calculate the predicted outcome for the observations in their OOB sample. For each observation in the data, the prediction of all the trees containing it in their OOB sample is averaged and the mean squared error is calculated. Finally, the OOB error is the mean of all the single observation errors. Also the variable importance (see section 2.3.4) is calculated using the OOB sample in order not to overestimate it.

2.3.3 Hyperparameters

Although random forests do not involve learning any parameters, there are some hyperparameters which are not learned during the tree-building process but have to be prespecified by the programmer. One of them is the number of trees m_{tree} that are built per random forest. *Breiman* proved that the out-of-bag error is bounded and converges asymptotically as the number of trees grows [40]. Generally, it is suggested to build around 500 to 1000 decision trees per random forest. Another hyperparameter is the size m_{try} of the random subset of potential splitting variables at each node of a tree. The recommendation for regression problems is to use $m_{try} = \frac{p}{3}$ and the effect of m_{try} has been discussed in the literature [43]. Finally, the programmer also has to define some criterion defining when to stop growing a tree. Some strategies have been proposed, as for example only splitting a node in a tree if the decrease in squared error exceeds some threshold. However, this strategy might not be appropriate when there is no candidate splitting variable in the random subset that decreases the squared error [42]. The most common strategy is to grow trees until a minimum node size has been reached. For regression problems, the standard is to stop the splitting process as soon as there are only 5 observations left in the node.

2.3.4 Univariate permutation importance

Random forests are non-parametric, which means that there is no predefined functional form between the predictors and the outcome, hence there is also no p-value or significance test for predictor variables [44]. Nevertheless, random forests can not only be used to make predictions on new data, but also to calculate variable importances and thus to choose the most predictive variables. Several variable importance measures have been proposed but the most frequently used is the permutation variable importance which has originally been described by *Breiman* [45]. In order to calculate the permutation variable importance (VIMP) of a variable x_j, the observed and the permuted mean squared errors are assessed. First, the OOB samples are passed down the trees and the mean squared error is calculated for each tree. Then, in order to calculate the permuted mean squared error of a tree, several strategies have been proposed. The most common one is to permute the values of the variable x_j in the OOB sample and to calculate the mean squared error for each tree again. Others rely on a random left-right daughter node assignment for nodes splitting on x_j when passing the OOB samples down the trees [46]. No matter which permutation strategy is used, the permutation variable importance of x_j is given by

$$VIMP_{x_j} = \frac{1}{B}\sum_{b=1}^{B}\left(MSE_{OOB_b}^{x_j,perm} - MSE_{OOB_b}\right) \tag{2.13}$$

with B the number of trees in the forest, MSE_{OOB_b} the observed OOB mean squared error of the bth tree and $MSE_{OOB_b}^{x_j,perm}$ its permuted mean squared error. If x_j is important (i.e. predictive), the relation between the outcome variable and x_j should be broken once x_j is permuted, leading to an increase in the mean squared error. The permutation importance can be used to rank the variables by their predictiveness as important variables have high permutation importance [30]. Another advantage of the permutation importance is that it does not only take marginal but also interaction effects into account. Therefore variables that have high interaction effects will also be retained [47]. This is an advantage when searching for potential interactions as the search space can be reduced by calculating the univariate permutation importance first. Thus random forests are well-suited as a variable selection tool.

2.3.5 Bivariate minimal depth importance

Several measures to identify interactions have been introduced [48]. The bivariate minimal depth variable is based on the original minimal depth, which relies on the observation that predictive variables frequently split close to the root [49]. *Ishwaran et al.* studied this concept by introducing maximal i-subtrees [46, 50]. A subtree of the original tree is an i-subtree if the root node of the subtree is split on x_i and it is maximal if it is not a subtree of a larger i-subtree. Variables with maximal subtrees closer to the root have a larger effect on the prediction accuracy and therefore are more important [51]. The univariate minimal depth

(MD) importance D_{x_i} involves measuring the distance of a maximal i-subtree to the root of the tree and averages this distance over all the trees. The smaller the minimal depth value, the more important the variable is. This concept can be extended for interactions as due to the hierarchical structure of the tree, a split in one node is conditional on all the previous splits and thus two interacting variables tend to split in the same branch of a tree. The bivariate minimal depth D_{x_i,x_j} indicates the minimal depth of a variable x_i with respect to the maximal subtree of x_j. As this measure is not symmetric, the interaction minimal depth (IMD) importance between two variables x_i and x_j is defined as

$$IMD_{x_i,x_j} = min(D_{x_i,x_j}, D_{x_j,x_i}). \tag{2.14}$$

Small values indicate a possible interaction between the variables x_i and x_j and thus suggests that the interaction is important in the sense that it contributes to the outcome. However, interactions identified by the bivariate minimal depth could also be additive when both variables have strong marginal effects. Therefore it can be used to filter candidate interactions.

2.3.6 Determination of significance

Random forests do not return significance p-values but they do return importance scores for predictors and interactions [52]. Although these are reasonable measures, there is still the issue of thresholding in order to determine significance [50]. Some strategies, as for example relying on a cut-off for the number of predictors assumed to be important, have been proposed [39]. However, this is not a reliable method as even in the absence of any association, this method will identify important varibles and usually there is no prior knowledge on how to set this cut-off [52]. Furthermore, the choice of the cut-off can not be statistically justified. The most robust method to identify significant associations is to rely on permutation p-values and to use a statistical cut-off as for example a 5% significance level [33]. Permutation p-values have been applied to the Gini index which is used to calculate variable importance in classification tasks in order to determine whether a variable's importance score is truly significant [53]. The same method can be applied to both the permutation variable importance and the interaction minimal depth.

The idea of permutation p-values is based on first calculating the observed importance and then permuting the outcome S times and assessing the importance each time. These S importances are called null importances as they represent the importance in a non-informative setting. Depending on the distribution of the null importances, either a parametric or a non-parametric/empirical p-value is calculated. If the null distribution fits a normal distribution, the p-value is calculated by fitting a normal distribution to the S null importances. If it does not fit a normal distribution, the p-value is calculated by counting the number of null importances as more extreme (higher for the permutation variable importance and lower for the interaction minimal depth) than the observed importance and dividing it by S.

Multiple testing refers to the problem that more than one hypothesis is tested simultaneously [1]. The tests that we perform are based on the null hypothesis H_0 that there is no association between a variable and the phenotype. If H_0 is true but the test rejects H_0, we make a type I error (false positive). If the alternative hypothesis H_A is true but the test accepts H_0, we make a type II error (false negative). We would like to have the two errors as small as possible but this is not possible as the two errors are related in the sense that a low type I error probability leads to a high type II error probability and vice versa. For a statistical test, the type I error probability is fixed at a level α (often 5%), which means that we can control the probability of making a type I error at α. Hence if a test is performed at the 5% level and the null hypothesis is true, there is a 5% probability of incorrectly rejecting the null hypothesis [1].

Multiple corrections have been suggested to control this error (e.g. false discovery rate [54]) but the standard in genome-wide association studies is to use a 5% Bonferroni threshold [12]. The Bonferroni correction ensures a control of the family wise error rate (FWER) which is the probability to make at least one type I error and is based on dividing the original p-values by the number of performed tests. For large m, this correction is rather conservative but in order to control the occurrence false positives, we also applied a Bonferroni correction to the univariate permutation p-values. As the interaction minimal depth only returns candidate interactions, we did not apply a Bonferroni correction to the bivariate p-values but filtered them for interesting patterns instead (see section 4.2).

2.4 Interpretable machine learning

Statistical models make distributional assumptions and restrict the complexity of the model in order to make it interpretable. Machine learning models however follow a non-parametric approach, which often results in less interpretable models [38]. They are usually referred to as “black boxes” as it is not always clear how results have been obtained. Although machine learning models generally focus on their predictive performance, there is a lot of research on the interpretability of the models. We have already introduced importance measures which are built-in features of random forests in order to make them more interpretable. However, these only return predictive variables (and interactions) but they do not reveal anything about the relationship between predictor variables and the outcome variable. Therefore we will introduce two methods which tackle this problem and which work for different types of machine learning models. They can be used to assess the effect a continuous variable has on the outcome, i.e. how a change in a variable changes the predicted outcome[1] [38].

[1] As the predictor variables in TWAS are continuous, the following methods can be used to visualize the effect of gene expression variation on the phenotype.

2.4.1 Partial dependence plots

The idea of partial dependence plots (PDP) is to visualize the predicted outcome $\hat{f}$ as a function of the continuous predictor variables [55]. Although PDPs can also be used to visualize the effect of two variables on the predicted outcome, we will focus on the one-dimensional case, i.e. visualizing $\hat{f}$ with respect to a variable x_k. Assume that there are p predictor variables and that the predicted outcome depends on all the variables, i.e.

$$\hat{f}(x) = \hat{f}(x_1, ..., x_p). \tag{2.15}$$

If we are interested in the effect of x_k on the predicted outcome $\hat{f}$, we can condition $\hat{f}(x)$ on the remaining variables $x_{\backslash k} = (x_1, ..., \hat{x}_k, ..., x_p)$ and consider it to be a function depending only on x_k with

$$\hat{f}_{x_{\backslash k}}(x_k) = \hat{f}(x_k | x_{\backslash k}). \tag{2.16}$$

$\hat{f}_{x_{\backslash k}}(x_k)$ will depend on the values of $x_{\backslash k}$, but if this dependence is not too strong, then the average or partial dependence function can represent the partial dependence of $\hat{f}$ on the variable of interest x_k [55]. The partial dependence function is given by

$$\bar{f}_k(x_k) = \mathbb{E}_{x_{\backslash k}}[\hat{f}(x)] = \int \hat{f}(x_k, x_{\backslash k}) p_{\backslash k}(x_{\backslash k}) dx_{\backslash k} \tag{2.17}$$

where $p_{\backslash k}(x_{\backslash k})$ is the marginal probability density of $x_{\backslash k}$ and $x = (x_1, ..., x_p)$. It is defined by

$$p_{\backslash k}(x_{\backslash k}) = \int p(x) dx_k \tag{2.18}$$

with $p(x)$ being the joint density of x. Of course, the probability density of the predictor variables is not known but it can be estimated from the data [55]. Assume that the data is given by $(x_i, y_i)_{i=1}^n$ with $x_i = (x_{i1}, ..., x_{ip}) \in \mathbb{R}^p$. Then the partial dependence function in (2.17) can be estimated by

$$\bar{f}_k(x_k) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}(x_k, x_{i,\backslash k}).$$

The $x_{i,\backslash k}$ represent the actual observed variable values from the training set. Hence the true model is approximated by $\hat{f}$ and the integral over $x_{\backslash k}$ is estimated by averaging over the n values of $x_{\backslash k}$ observed in the data [56]. One assumption of partial dependence plots is that the variable x_k is not highly correlated to any of the variables in $x_{\backslash k}$, otherwise the averages that are calculated for the partial dependence plots include data points which are not likely to be observed [57]. Having identified predictive variables the first place, PDPs can be used to get a visual representation of the relationship between the variable and the outcome which could be linear, monotonic or even more complex. A flat PDP indicates that the feature is not important [57]. The PDP is a global method in the sense that it returns plots showing the global, average relationship of a variable with the predicted outcome [57].

2.4.2 Individual conditional expectation plots

Partial dependence plots show the average relationship between a predictor variable x_k and the predicted outcome $\hat{f}$. However, if interaction effects are present, the relationship can be heterogeneous and might be different for each individual observation and thus the average curves of PDPs might be misleading. Individual conditional expectation (ICE) plots extend PDPs in that they return the relationship between a variable of interest x_k and the predicted outcome for each observation [56]. Thus they can show to what extent heterogeneous relationships exist and help to identify interactions making a contribution to the predicted outcome $\hat{f}$. We will use an example from [56] to illustrate the problem with PDPs. Consider the following data generating process which includes one interaction

$$\begin{gathered} y = 0.2x_1 - 5x_2 + 10x_2\mathbb{1}_{x_3\geq 0} + \epsilon \\ \epsilon \overset{\text{iid}}{\sim} \mathcal{N}(0,1), x_1, x_2, x_3 \overset{\text{iid}}{\sim} \mathcal{U}(-1,1). \end{gathered} \tag{2.19}$$

We generated 1000 observations from this model and fitted a random forest in order to get the predicted outcome $\hat{f}$. Interested in the effect of x_2 on the predicted outcome, we can get the true dependence of y on x_2 (Fig. 2.1A) and the PDP combined with an ICE plot (Fig. 2.1B). The PDP (straight yellow line) suggests that x_2 is not associated with the outcome. The ICE plot (grey lines) however, shows that the predicted outcome is related to x_2 by a linear relationship which is increasing or decreasing depending on the sign of the variable x_3 and thus suggesting that there are interaction effects.

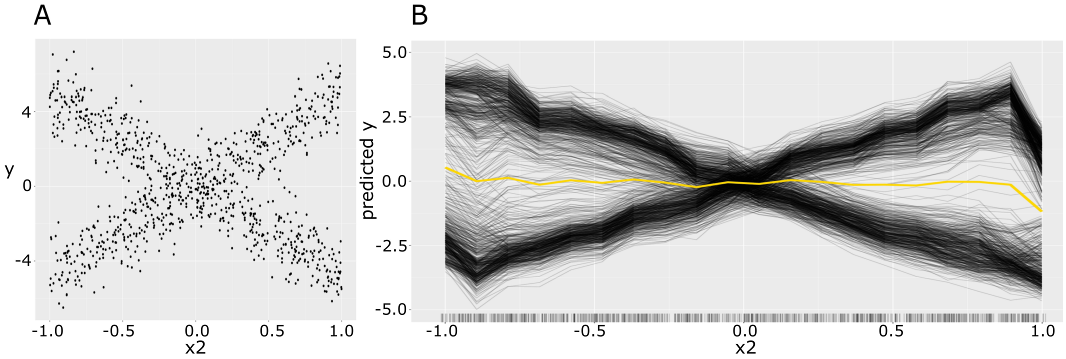

Figure 2.1: An example of a partial dependence plot (PDP) incorrectly suggesting that there is no relationship between a predictor variable and the outcome. (A) Scatter plot of y as function of x_2. (B) PDP (yellow line) and ICE (grey lines) of the data generating process described in equation (2.19). The predicted outcome $\hat{f}$ has been predicted by fitting a random forest and the plot has been created with the package `iml`[1]. The plots have been modified from [56].

[1] https://cran.r-project.org/web/packages/iml/index.html

Chapter 3

Motivation

Association studies have successfully identified genetic loci that are associated with complex diseases or traits. However, many of the identified loci only explain a small proportion of the heritability, hence the question arises how the remaining, “missing heritability” can be explained [29]. The standard approach in genome-wide (GWAS) and transcriptome-wide association studies (TWAS) are linear mixed models. They assume linear relationships between the genetic markers and the phenotype and furthermore the popular single-locus based hypothesis tests assess the potential association of each marker without considering the other markers [39]. In this study, we will focus on nonlinear individual effects, epistatic interactions and genetic heterogeneity, from either or both of genetic and transcriptomic variation, and their contribution to the “missing heritability”.

Identifying epistatic interactions is both statistically and computationally demanding due to the high-dimensionality of omics data and hence the high number of tests that need to be performed [12]. Several strategies to identify interactions have been proposed. The most straight-forward would be an exhaustive search, addressed by constructing linear models while allowing for pairwise interactions [58]. Due to the definition of epistasis being the deviation from a model accounting for additive effects only, these tests of interaction seem to be the most natural approach. However, it quickly becomes computationally infeasible as the number of loci increases. Some other current strategies in GWAS consist in reducing the number of tests that have to be performed by only testing loci that previously have been shown to be important in marginal GWAS or by reducing the dimensionality beforehand [59, 60]. The problem with the first strategy is that marginal effects might eventually be absent and hence loci that are involved in interactions but do not show strong marginal effects might be missed.

Machine learning models offer a non-parametric alternative to traditional statistical models. A review on several machine learning algorithms that have been applied to genetic data has been published [23]. The algorithms include neural networks, multifactor dimensionality reduction (has been developed to also detect higher-order interactions [61]) and random forests.

Random forests are particularly interesting as they allow to incorporate a large number of markers by handling high-dimensional data and have further advantages [52]. While linear mixed models make the assumption that each phenotype is affected by simple linear, additive effects of individual loci, random forests do not make any assumption about the underlying genetic model. It is a non-parametric model, which learns patterns from the data and thus it can capture nonlinear relationships and interactions. Another advantage of random forests is that due to importance measures, it can be used as a variable selection tool. As the permutation importance also takes interaction effects into account, variables with high interaction effects but only small marginal effects will be retained. Besides nonlinear effects and epistatic interactions, an additional concern that traditional methods are confronted with is the presence of genetic heterogeneity. Linear models assess average effects by lumping an entire sample into a single group and thus have reduced power to detect associations if there is genetic heterogeneity [32]. Random forests have been shown to be able to handle genetic heterogeneity as early splits in the tree actually define separate models for separate subsets of the data [47]. Therefore, random forests might actually perform better than single-locus based linear models. The performance of random forests to identify interactions in GWAS has been studied extensively, however it has mostly been applied to case-control data without using measures to specifically identify interactions between genetic markers [48, 47, 32, 62]. Using the bivariate minimal depth importance, candidate interactions which potentially contribute to the phenotype can be identified.

It has been shown that interactions are likely to contribute to the phenotype and that phenotypes are influenced by genetic heterogeneity (take *A. thaliana* as an example: [31, 8]), thus allowing for interactions and heterogeneity might increase the power of association studies and reveal more loci with complex biology (partially) filling in the hidden heritability. The aim of this thesis is to apply an improved random forest pipeline to *A. thaliana* multi-omics data in order to identify complex, potentially interacting loci making a contribution to the flowering time phenotype and improve the power of GWAS and TWAS. Furthermore, random forests can handle different data types as predictors, thus we can combine discrete SNPs and continuous gene expression in one model. This opens a new study, which we call genome-and-transcriptome-wide association study (GTWAS). Integrating heterogeneous data provides an opportunity to better understand the genetic architecture and thus the biology behind associations that can not be identified when analyzing either genetic or transcriptomic variation.

Chapter 4

Methods

Except if stated otherwise, all steps have been implemented in *R* statistical programming language [63] and the plots have been created with the package `ggplot2`[1].

4.1 Data collection and preprocessing

The phenotypic data has been obtained from the public *A. thaliana* phenotype database Ara-Pheno [64]. Phenotypic traits represented in this study include flowering time at 10°C and 16°C (FT10[2] and FT16[3]), two flowering time measurements in different environments, which are highly correlated ($r = 0.88$, p $<$ 2.2e-16). The phenotypic values are scored as days until the first open flower can be observed in the respective temperature. The genotypic data (only SNPs are used here) has been obtained from the 1001 Genomes Consortium and is available for $1,135$ accessions [65]. Missing values have been imputed by Haijun Liu with `Beagle 4.1` [66], which is common in GWAS and necessary for the random forest to ensure all accessions have genotype data for all the sets of variants. The RNA-seq data (gene expression) has been generated and cleaned for population structure and additional batch effects by Yoav Voichek.

The general preprocessing steps involved removing accessions for which phenotypic values were missing, leaving $1,003$ accessions for the flowering time at 10°C and 970 accessions for the flowering time at 16°C. Then, the genotypic data has been filtered for a minor allele frequency[4] of $< 2\%$ (in `PLINK` [67]). The RNA-seq data has been cleaned by removing genes that were expressed in less than 10% of the accessions, leaving $19,554$ genes' expression. The imputation of other missing values has been done using the minimum observed value for that particular gene as missing values were probably due to low expression. After these steps, we turned to preprocessing that is specific to random forests.

[1] `https://cran.r-project.org/web/packages/ggplot2/index.html`

[2] `https://arapheno.1001genomes.org/phenotype/261`

[3] `https://arapheno.1001genomes.org/phenotype/262/`

[4] The minor allele frequency is the frequency at which the second most common allele occurs in the population.

4.1.1 SNP filtering

The effect of correlated variables in random forests results in a reduction of the importance for predictive variables that are correlated with many other variables [39]. Therefore, SNPs have been pruned both in GWAS and GTWAS and additionally those associated with any gene's expression have been filtered out in GTWAS.

Due to linkage disequilibrium, there is some correlation structure between SNPs and thus neutral SNPs that are linked to causal SNPs can act as surrogates of each other in the tree. Assume that two linked SNPs tagging a causal SNP are assigned to the same tree. The prediction accuracy might not decrease when permuting the values of one of the SNPs if the values of the second SNP are unchanged [68]. This decreases the variable importance of both SNPs and the more highly correlated the variables are, the more they can serve as surrogates of each other (see 5.1.1 for the impact on this real dataset) [47]. Hence they might retain undetected although they do tag a causal SNP and thus are predictive. Several approaches have been proposed to handle this problem, including conditional variable importance [69]. However, this alternative importance is computationally intensive and thus not feasible for high-dimensional data. Therefore, we pre-filtered the SNPs based on a correlation threshold [70]. We used overlapping sliding windows of 50 kb and within each window, we filtered out SNPs having a correlation higher than $r^2 = 0.8$ to one of the other SNPs (implemented in `PLINK` [67]). The filtered genotypic data was used both in GWAS and GTWAS as predictors of the random forest. An additional filtering has been applied to the SNPs that are significantly associated to genes' expression (similar to the high linkage between SNPs, this indicates a strong correlation between the SNP and the genes' expression) in GTWAS. Yoav Voichek generated SNPs passing the significance threshold in the eQTL study for each of the genes that were available in the expression data. These SNPs act as eQTLs and thus are strongly correlated with at least one of the genes' expression, therefore they have been removed in order to avoid potentially high correlations among the predictors and to improve the power of random forest GTWAS.

4.1.2 Correcting for population structure

Various methods have been proposed to account for population structure in association studies and thereby reducing false positives. *Price et al.* suggested a regression-based correction [71]. The idea relies on applying principal component analysis to the genotype data in order to reduce the dimensionality of the data to a smaller number of axes of variation. These axes being defined as the eigenvectors to the largest eigenvalues of the covariance matrix between samples describe a maximum of the variability and are called principal components. For organisms displaying ancestry differences between individuals, principal components often have a geographic interpretation (Fig. 4.1).

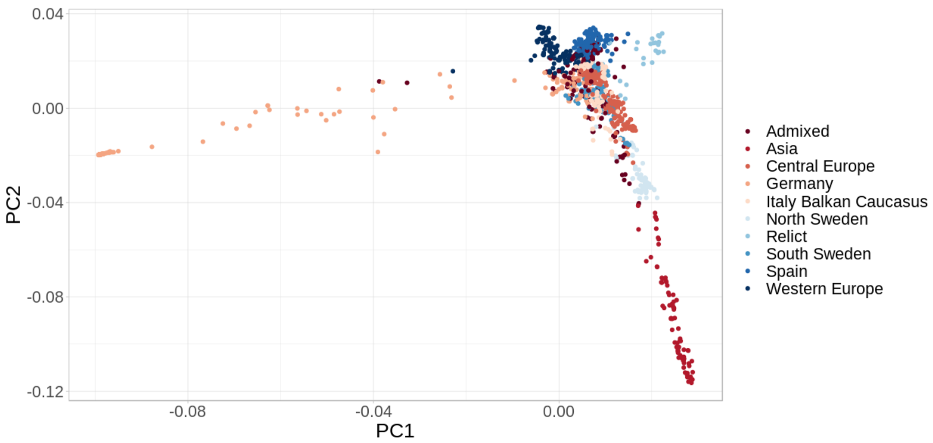

Figure 4.1: Clustering $1,135$ *A. thaliana* accessions along the two first principal components shows that the principal components have a geographic interpretation.

In order to correct for population structure in random forests, it has been recommended to apply the two-step approach suggested by *Price et al.* and based on regression to both genotypes and the phenotypes [72, 73]. Then, the adjusted genotypes and phenotypes are used as predictor and outcome variables in the random forest analysis, respectively. *Zhao et al.* have studied the sensitivity of the correction to the number l of principal components that are being used [73]. As the results were almost identical for $l = 10$ and $l = 20$, we have followed their approach and used $l = 10$ principal components in order to correct the genotype and phenotype for population structure.

The approach can be summarized as follows:

- First, a principal component analysis was applied to the genotype matrix and the l ($= 10$ in the present analysis) first principal components (PCs) were returned. The l first PCs represent the eigenvectors to the l largest eigenvalues of the covariance matrix between the samples.

- Then, each SNP and the phenotype data vector were transformed such that they are orthogonal to the l selected principal components. This was done by regressing the data vector on the l first principal components and calculating the residuals while fitting a generalized linear model (GLM). As the genotype data is binary, a logit link function has been used when regressing each SNP while a linear regression model has been used for the continuous phenotype.

- Finally, the residuals of these models were used as predictor and outcome variables in the random forest analysis, respectively.

4.2 Pipeline

We optimized the hyperparameters on a subset of the variables (SNPs and genes' expression) in order to fix the hyperparameters that we would use for all of the random forest analyses. We observed that the out-of-bag (OOB) error converged already after 100-200 trees and thus we followed the general suggestion of building m_{tree}=500 trees per random forest. Furthermore, the OOB error was not very sensitive to the size m_{try} of the candidate splitting variables at each node of the trees. Thus we used the (widely accepted) recommendation for regression problems of m_{try}=$\frac{p}{3}$ where p is the total number of predictor variables in the analysis. Finally, we stopped growing trees as soon as only 5 observations were left in a node.

After the preprocessing step, two panels of $970/1,003$ accessions were available for the flowering time at 16°C/10°C in GWAS. TWAS has been applied to a subset of the accessions for which gene expression was available including 453/470 accessions for the flowering time at 16°C/10°C and the same subset of accessions was available for GTWAS. Typically, the number of predictor variables included in association studies is huge. Although random forests can handle high-dimensional data, dimensionality reduction and filtering are necessary to capture associations and interactions. The pipeline that we have developed is based on a multi-step approach which reduces the dimensionality of the data in each step.

The first step of the pipeline is reducing the number of variables for which univariate p-values are calculated. As computing permutation p-values includes building $S + 1$ random forests (in order to calculate S permuted importances and one observed importance), the number p of predictor variables in the analysis has to be reduced in advance in order to reduce the computational load. As the package `ranger`[1] is particularly well suited for high-dimensional data, the permutation importance was computed for all the p predictor variables in order to rank them according to their importance. We have retained the k most important variables (k = 200 for TWAS and k = 500 for GWAS and GTWAS). As the permutation importance also takes interaction effects into account, interacting variables without marginal effects are retained too.

The second step of the pipeline involves identifying significant trait-associated genetic markers while relying on permutation p-values. The `vita`[2] package has been used to calculate univariate permutation p-values for the k pre-selected variables as there is an implementation of the approach described in section 2.3.6 [53]. In order to get reliable results, we computed S = 1000 permuted importances to get the permutation p-values. Due to multiple testing, a Bonferroni correction and a 5% significance threshold have been applied to the univariate p-values in order to identify whether a variable's permutation importance is truly significant.

[1] `https://cran.r-project.org/web/packages/ranger/index.html`

[2] `https://cran.r-project.org/web/packages/vita/index.html`

The third step in the pipeline aims to identify interactions making a contribution to the phenotype. In order to reduce the search space for interactions, we again only retained a subset of the previous variables (SNPs or genes' expression). As a Bonferroni corrected 5% significance threshold would be too strict and we might miss many potential interactions, we tested variables that have a raw univariate p-value < 0.05 for interactions as variables with interaction effects also have high univariate permutation importance. Candidate interactions have been identified by computing the interaction minimal depth. We used the "find.interaction" function with the option "max.subtree" from the `randomForestSRC`[1] package. In order to identify reliable interactions, we again relied on permutation p-values and have implemented the approach described in section 2.3.6 for the interaction minimal depth in order to get bivariate p-values.

The last step included filtering the candidate interactions identified by minimal depth for interesting non-additive patterns. The interactions returned by minimal depth might be additive if for example two variables have strong independent marginal effects and therefore happen to split nodes more often than other variables and thus have low interaction minimal depth. We did not apply a Bonferroni correction to the bivariate p-values in order to reduce both false positives and false negatives. Instead, we interpreted the interactions with raw p-values < 0.05 as either additive or non-additive based on pairwise t-tests and based on the patterns in figure 4.2, which has been modified from [25]. The advantage is that the interactions are very interpretable in that way.

In GWAS, each node split is based on a 0/1 split as SNPs are binary, thus random forests do not have to choose a splitting point s. In TWAS and partially also GTWAS (if the node is split by a genes' expression) however, the variables (genes' expression) are continuous and random forests do search for the best splitting point s which maximizes the variance in each resulting node. Assume that the two variables 1 and 2 have been retained as their bivariate p-value is < 0.05. Taking all of the accessions, the data is separated into 4 distinct haplotypes (or nodes in random forest terminology) based on their values (alleles for SNPs and lower or higher median for genes' expression, as med(expr); Fig. 4.2) for the two interacting variables. In order to identify interesting interactions, we first test whether a significant difference in mean phenotype between the accessions in the left, respectively the right branch of the interaction tree can be observed, i.e. how steep the slope is (represented by the black lines in Fig. 4.2). This is tested with a Welch two-sample t-test (allowing for different population variances [74]). Also, the signs of the slopes are of interest as different signs reveal very heterogeneous effects (Fig. 4.2C).

[1] `https://cran.r-project.org/web/packages/randomForestSRC/index.html`

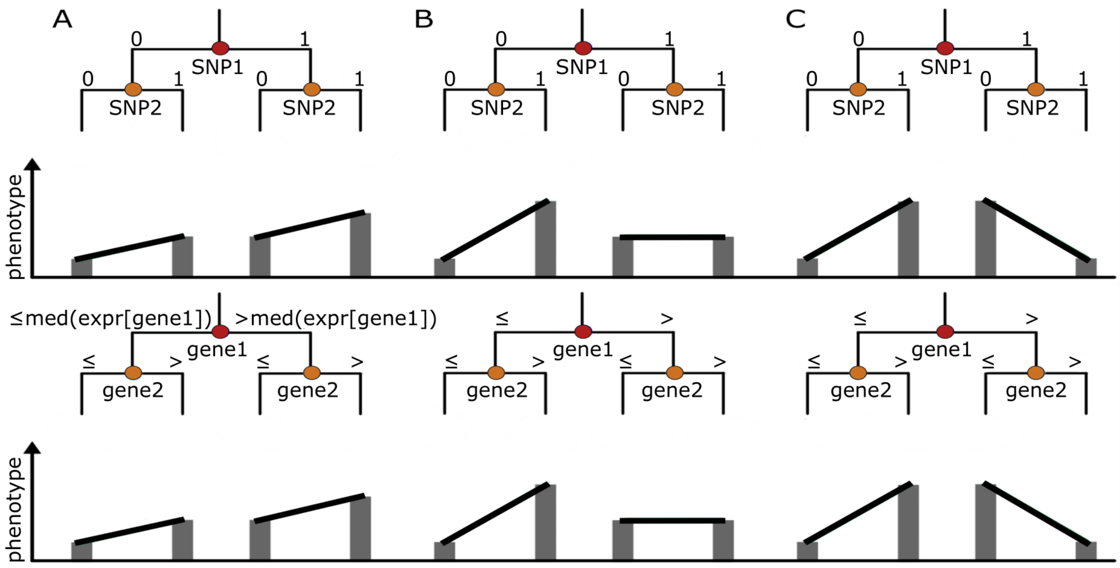

Figure 4.2: Interaction patterns that can be identified in GWAS (above) and TWAS (below). (A) is an example of an additive interaction while (B) and (C) are non-additive. This figure has been modified from [25].

The first interaction pattern (Fig. 4.2A) would be classified as additive as each variable has a marginal effect on the phenotype that does not depend on the state of the other. The other two patterns would be classified as non-additive. The second pattern (Fig. 4.2B) describes the situation where an allele of the first variable masks the effect of the second variable and can be seen as a multi-locus extension of the dominance effect for alleles within the same locus [26]. In the third interaction pattern, the effect direction of the second variable depends on the first variable. For both of the last patterns, a split on the second variable is more probable given a previous split on the first variable. Interactions in GTWAS are classified similarly including 0/1 splits for SNPs and ≤ / > split for genes' expression. A summary of how interesting interactions have been identified using pairwise t-tests can be found in the supplemental data A.

Random forests profit from cases where at least one of the variables has some marginal effect as only then it will be selected for a split. In figure 4.2B for example, it might happen that the second variable does not have any marginal effect but based on a prior split on the first variable, random forest will probably split on the second variable in the left branch of the tree where it shows some marginal effect [25]. Generally, for this interaction pattern, it might happen that both variables would be retained by a single-locus test if both show some marginal effect, but it might also happen that the power to detect the second variable is reduced [26]. Furthermore, whether random forests or single-locus approaches detect such cases of course also depends on the frequencies in the population. In figure 4.2C however, it might happen that none of the variables show any marginal effect and thus an interaction like this might even be hard for random forests to capture [25].

Finally, the pipeline is based on a multi-step approach which reduces the dimensionality of the data in each step and can be summarized as follows:

Algorithm Multi-step random forest

```
1. Reduce the dimensionality of the data by calculating the permutation importance.
     Retain the $k$ most important variables.
2. Compute univariate permutation p-values based on the permutation importance.
     Retain variables that have raw univariate p-values $< 0.05$.
3. Compute bivariate permutation p-values based on the interaction minimal depth.
     Retain interactions that have raw bivariate p-values $< 0.05$.
4. Filter the candidate interactions based on pairwise t-tests.
```

In order to compare random forest results to single-locus approaches, we used the linear mixed models (LMM) implemented in `LIMIX` and in `OSCA` for GWAS and TWAS, respectively [75, 35]. We also filtered the SNPs for a minor allele frequency of 2%, however we did not filter the SNPs for correlations but used all the available SNPs for LMM GWAS. In LMM TWAS, we used the same genes' expression that were also available in the random forest analysis and corrected the phenotype for population structure by regressing out the top 10 PCs beforehand. A kinship matrix has been used to estimate the covariance matrix of the random effects in `LIMIX`. `OSCA` (with the --moa mode) includes all gene expression levels as random effects to account for confounding effects and correlations between distal genes and thus to control for false positives [35].

Chapter 5

Results

We have conducted three types of association studies (GWAS, TWAS and GTWAS) with univariate and bivariate analyses, leading to six different studies for each phenotype. First, we will emphasize some properties of the pipeline and show how we optimized the pipeline to make the results reliable and comparable. Then, we will get to the associations to show that we can improve GWAS and TWAS by a non-parametric and non-linear model, which furthermore allows us to combine both approaches into a GTWAS model. We will take specific examples to show how the new findings enhance our understanding on genetics and molecular biology.

5.1 Properties and optimization of the pipeline

5.1.1 Effect of SNP filtering

The effect of correlated variables in random forests (RF) results in a decrease of the importance for predictive variables that are correlated with many other variables (see section 4.1.1). By focusing on a flanking region around *FLOWERING LOCUS C* (*FLC*) (±100 kb; chromosome 5), we will show that important variables are being returned lower variable importances when big linkage groups exist. *FLC* is one of the main flowering time genes involved in flowering time regulation and encompasses SNPs which lead to variation in flowering time. We will use flowering time at 16°C as phenotype and use both non-pruned (2343 SNPs) and pruned (methods described in section 4.1.1; leaving 1089 SNPs) genotypes to test the effect of LD filtering.

We have computed the permutation importance for both data sets, normalized them and compared the importances on the subset of SNPs that are contained in both (1089 SNPs). The SNPs in the *FLC* gene, that have been ranked as important (> 0.1) in both data sets, have generally been returned higher variable importance once the SNPs have been filtered for LD (Fig. 5.1A; mean pruned vs. non-pruned: 0.37 vs. 0.25). For the group of SNPs that are not associated to the phenotype and thus have been returned lower variable importance in general (< 0.1), there are no significant differences between the importance scores (Fig. 5.1A; mean pruned vs. non-pruned: 0.04 vs. 0.05), which is also the case for the likely non-functional flanking region of *FLC* (Fig. 5.1B). These findings suggest that filtering highly correlated variables increases the power to identify potential positive targets while maintaining unchanged importance levels (compared to including all correlated variables) for irrelevant variables.

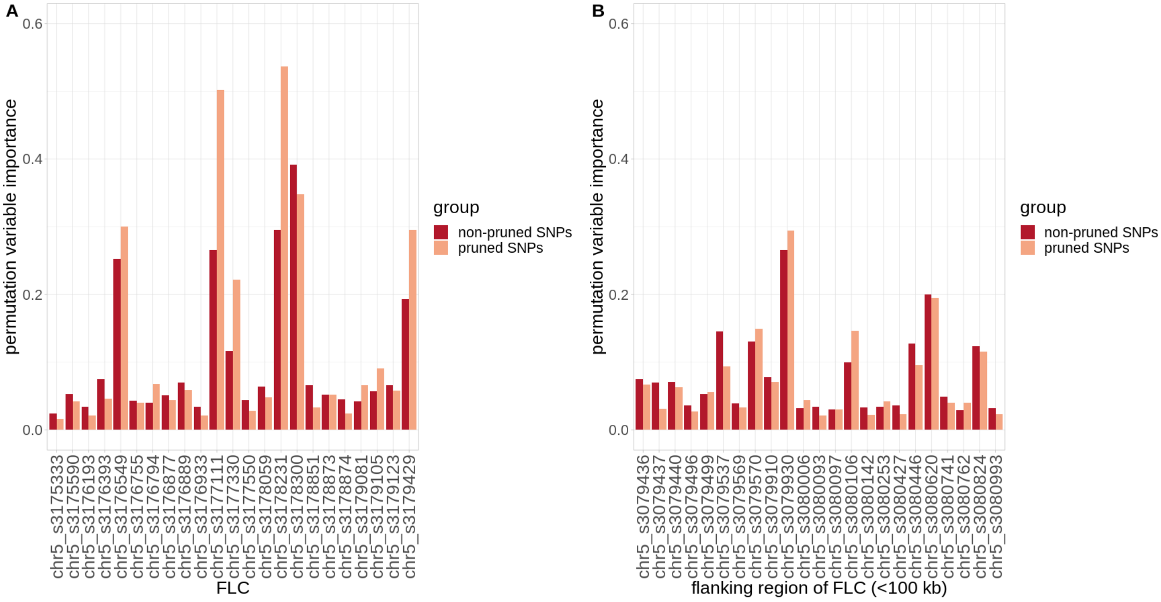

Figure 5.1: SNP filtering improves the power of random forests. Comparing the permutation importances of SNPs located in the flowering time regulator *FLC* (A), respectively in a likely non-functional flanking region of *FLC* (B) before and after SNP pruning.

We thus conclude that for predictive variables, big linkage groups result in a reduction of the permutation variable importance. Furthermore, the computational load can also be reduced by pruning the SNPs for highly correlated SNPs. Therefore, we will always filter the SNPs based on linkage disequilibrium in the further analysis. The point to note in this way is that although the significant SNPs will be highlighted in the following, they represent linkage groups (loci-based) rather than individual variants.

5.1.2 Effect of population structure correction on both genotype and phenotype

Zhao et al. studied the effect of population structure correction in random forest analysis [73]. The authors simulated case-control data by modeling random, differentiated and causal SNPs. Differentiated SNPs show different allele frequencies and different rates of disease among subpopulations. As they are indirectly correlated with the disease, differentiated SNPs might be as good predictors as causal SNPs. When both differentiated and causal SNPs are used in random forests, they compete with each other to be selected as the splitting variable in each node.

Zhao et al. found that if population structure is not reasonably corrected for, random forests might return low importances to causal SNPs and high importances to differentiated, but disease-unrelated SNPs. They propose to correct the genotype and the phenotype by the information of population structure from principal component analysis (as in section 4.1.2) and to use the adjusted genotype and phenotype as the predictors and outcome in random forest analysis, respectively. Population structure is a well-known confounding variable, and in linear mixed models it usually is corrected by adding it as a fixed effect to the dependent (predictor) variables. Here, we will show that population structure acts both on the dependent and the independent (outcome) variables and therefore both genotype and phenotype should be corrected for population structure. In addition to the previous simulation study [73], the flowering time phenotype (16°C) has been used as an example to demonstrate this in detail.

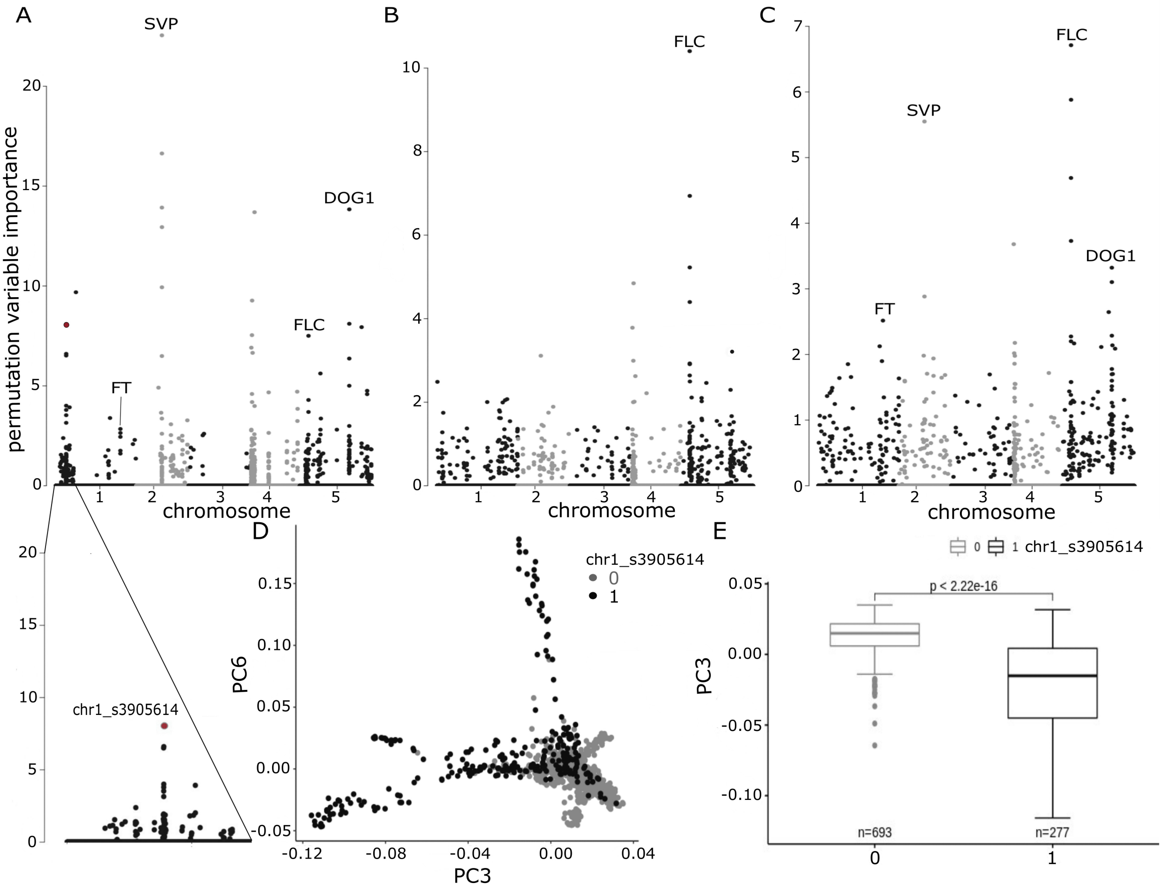

Figure 5.2: Population structure (PS) correction on both genotype and phenotype reduces false positives and improves the power to detect true associations. (A) Random forest GWAS using the raw genotype and phenotype, with the peak at chromosome 1 highlighted below. (B) Random forest GWAS correcting only the phenotype for PS. (C) Random forest GWAS correcting both the genotype and the phenotype for PS. Peaks have been annotated if a SNP within 20 kb of a known flowering time gene exists. (D) Clustering and (E) distribution of the alleles from the peak at chromosome 1 (chr1_s3905614) that are most highly correlated with PS, measured with the principal components.

Three flowering time genes *FLC*, *DELAY OF GERMINATION 1* (*DOG1*) and *FLOWERING LOCUS T* (*FT*) have been associated with the flowering time phenotype by the linear mixed model (see section 5.2 Fig. 5.5A), which provides a baseline for further comparison. We first used the raw genotype and phenotype without any correction for population structure (Fig. 5.2A), then the corrected phenotype only (Fig. 5.2B) and finally both the corrected genotype and phenotype (Fig. 5.2C) as predictors and outcomes of the random forest analysis, respectively. In addition to the three well-known genes identified in the LMM, another known flowering time regulator *SHORT VEGETATIVE PHASE* (*SVP*), was observed in analyses where either none or both genotype and phenotype have been corrected for population structure (Fig. 5.2A,C), while remaining non-significant in phenotype-only correction (Fig. 5.2B). Compared to both genotype and phenotype corrected, the analysis with the raw data identified a new peak on chromosome 1, at which the alleles show high correlation with population structure (PC3 $r^2 = 0.61$, $p < 2.2e\text{-}16$; PC6 $r^2 = 0.32$, $p < 2.2e\text{-}16$), suggesting confounding and thus a potentially false-positive candidate (Fig 5.2 A,D,E). This peak is neither significant when correcting both genotype and phenotype for population structure, nor in the linear mixed model, and furthermore it could not be related to any flowering time gene. With all these findings, we conclude that phenotype-only correction in random forest is not sufficient and lowers the association power, while additional SNP correction could not only improve the power to find true associations, but also is necessary to reduce false positives.

Consistent with the study of *Zhao et al.*, we have shown that the power to detect causal genes increases and possibly spurious associations are well controlled when correcting both the genotype and the phenotype, although correcting the genotype for population structure means fitting p (= number of SNPs) linear models, which is computationally expensive to some extent. Thus for all of the random forest analyses, we correct both genotype and phenotype for population structure.

5.1.3 P-value estimation for univariate statistics

As seen in section 2.3.4, variables with high permutation importance are predictive and thus associated to the trait while variables with low permutation importance are not. Since the permutation importance is not that interpretable and straightforward in comparison to significance levels in linear models, we wanted to compute individually trait-associated variables (SNPs and genes' expression, as univariate) p-values. Although univariate might not be the right term as the permutation importance also takes interaction effects into account, we will stick with this term to differentiate from the below bivariate interaction p-values. In order to get the univariate p-values, we first computed the observed permutation importance for each tested variable in a non-permuted setting and then $S = 1000$ permuted importances. We next applied a *Kolmogorov-Smirnov* test to each null distribution of the individually tested variables to test whether the importance score follows a normal distribution.

As more than 95% of the null distributions did fit a normal distribution in univariate analyses, for both GWAS and TWAS (Fig.5.3B,C), the univariate p-values have been calculated from a normal distribution by fitting a Gaussian distribution to the S null importances for each tested variable.

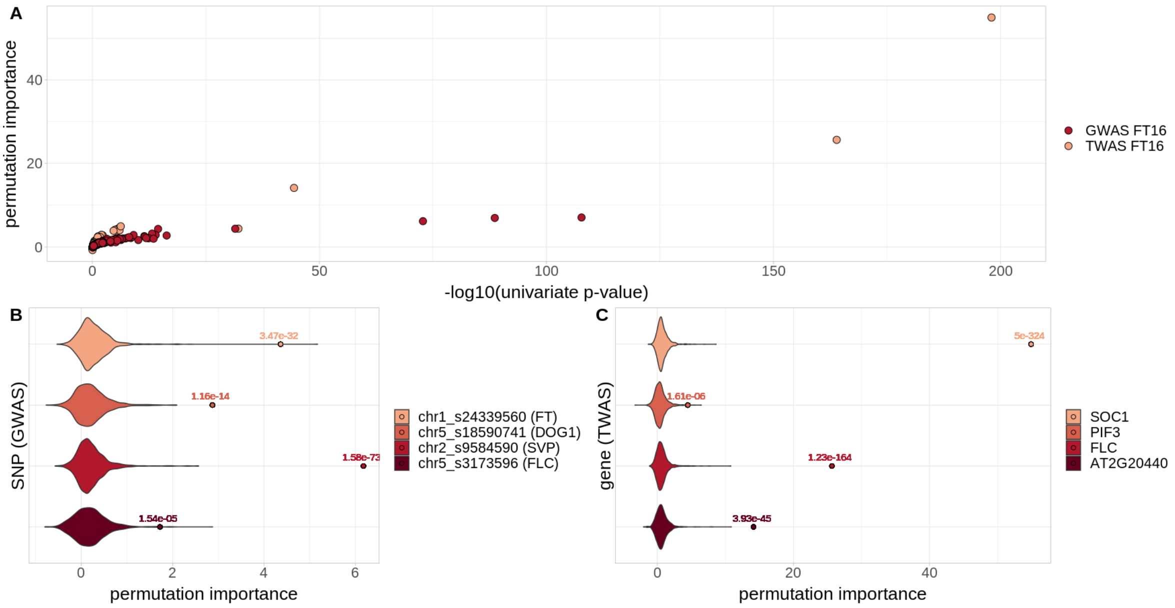

Figure 5.3: Univariate p-values improve the interpretation of the model. (A) Relation between the univariate p-values and the univariate permutation importances from random forests. (B,C) Examples of null distributions (colored areas), their observed permutation importance (black circles) and the associated univariate p-value of highly significant SNPs from GWAS (B), respectively of highly significant genes from TWAS (C). All SNPs are located within a window <6 kb around the annotated flowering time genes.

The p-values estimated in this way are highly correlated with the importance scores, both for GWAS and TWAS (Fig. 5.3A; p < 2.2e-16 and p = 4.19e-08, respectively). Variables with high permutation importance do, while variables with low importance do not receive significant p-values. Thus univariate p-values have been used in the following analyses.

5.1.4 P-value estimation for bivariate statistics

As seen in section 2.3.5, small values of the bivariate minimal depth importance indicate that there is a possible interaction between two variables. Again, we wanted to compute bivariate p-values to define a significance level as the raw minimal depth importance scores are not that interpretable. In order to get the bivariate p-values, we first computed the observed minimal depth interaction importance for each tested interaction in a non-permuted setting and then $S = 1000$ permuted importances. As for the univariate analyses, we next applied a *Kolmogorov-Smirnov* test to each null distribution to test whether the importance score of the tested interaction follows a normal distribution.

However, as more than 95% of the null distributions did not fit a normal distribution (Fig. 5.4B,C), bivariate p-values have been computed by using the empirical distribution of the S null importances for each tested interaction.

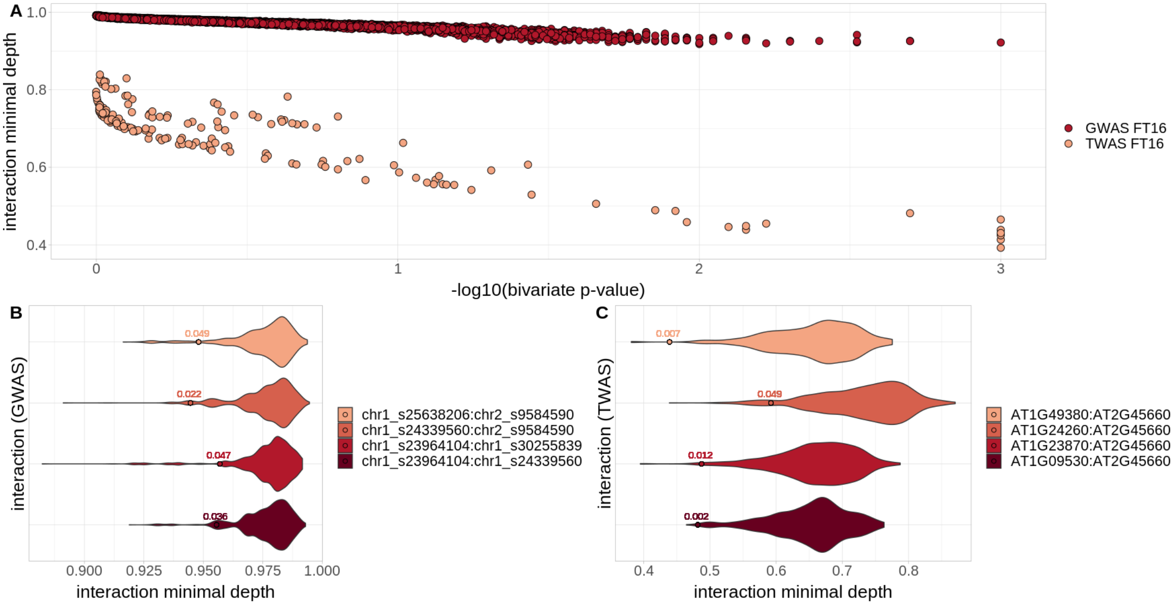

Figure 5.4: Bivariate p-values improve the interpretation of interactions. (A) Relation between the bivariate p-values and the minimal depth interaction importances from random forests. (B,C) Examples of null distributions (colored areas), their observed permutation importance (black circles) and the associated bivariate p-value of candidate SNP-SNP interactions from GWAS (B), respectively of candidate gene-gene interactions from TWAS (C).

The bivariate p-values estimated in this way are highly correlated with the importance scores from minimal depth, both for GWAS and TWAS (Fig. 5.4A; p < 2.2e-16 for both). Interactions with small minimal depth do receive significant p-values while interactions with large minimal depth do not. Another advantage using bivariate p-values is that we do not have to rely on some threshold of an absolute minimal depth value, which could not be chosen universally for GWAS and TWAS due to the different ranges of the importance scores (Fig. 5.4). Due to the large number of SNPs in GWAS, it is less probable for a SNP to be chosen as one of the candidate splitting variables as this probability decreases with the number of variables increasing. Hence also the probability for two variables to have low interaction minimal depth decreases. As there are fewer variables in TWAS than in GWAS ($\approx 20,000$ genes' expression compared to > 1 million SNPs), SNP-SNP interactions generally have higher interaction minimal depth than gene-gene interactions (Fig. 5.4A). Permutation-based bivariate p-values adapt to these ranges, are thus more robust to the number of variables used and are better interpretable than the raw importance scores. This makes the direct comparison possible not only between different omics, but also between machine learning and statistical models. Therefore, bivariate p-values have also been used in the following analyses.

5.2 RF improves GWAS power by incorporating interactions

5.2.1 Individually significant SNPs

First, we repeated previously published GWAS results for flowering time at 16°C in a panel of 970 *A. thaliana* accessions by fitting a linear mixed model to the data (Fig. 5.5A; implemented in `LIMIX` [75]; see section 4.2). There is only one peak on chromosome 5 near the flowering time regulator *DOG1* that reaches genome-wide significance when using a Bonferroni-threshold of 5%. However, two remaining peaks can not be ignored and are identified near the flowering time regulators *FT* and *FLC*, which both have been associated to flowering time variation before. We can compare the results from random forest univariate analyses to the results from the linear mixed model and show that random forests improve the power of GWAS by allowing interaction effects. Significant associations using a Bonferroni-threshold of 5% identified by univariate random forest analyses have been cross-referenced with the FLOR-ID database [9]. Although linkage disequilibrium is too extensive to find the causal variants with GWAS alone, we have considered all candidate genes within 20 kb of a significant SNP association. This left us with 6 significant candidate flowering time genes (Fig. 5.5B). The three genes *FT*, *FLC* and *DOG1* revealed with the LMM have been identified as well. Furthermore, we have identified three flowering time regulators, which are not significant in the linear mixed model (Fig. 5.5B; *SVP*, *FY* and *LATE FLOWERING* (*LATE*); most significant p-values from the LMM in a 20 kb window of the respective loci of 1.57e-05, 7.73e-04 and 1.04e-04). Another peak on chromosome 4 (whose significance comes only after *FLC* and *SVP*) could not be related to any flowering time gene when using a 20 kb window.

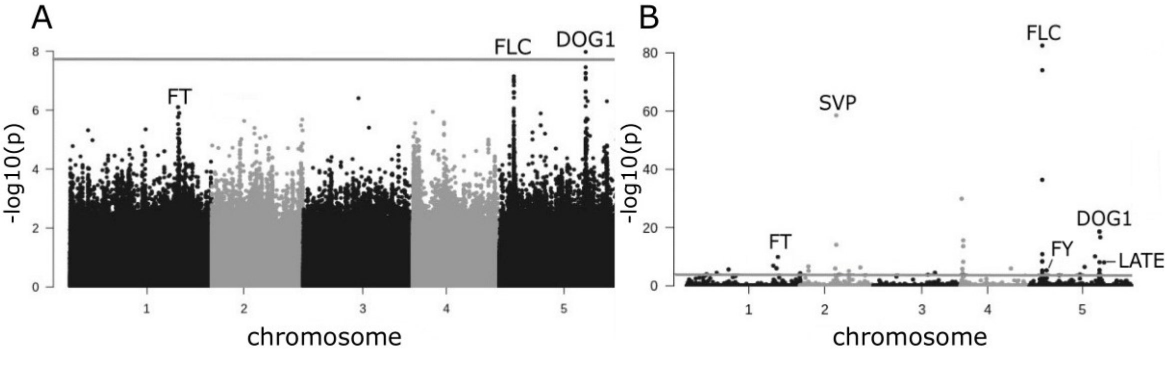

Figure 5.5: Random forests improve the power of GWAS; more loci can be identified. (A) Linear mixed model GWAS Manhattan plot for flowering time at 16°C where the highest peaks have been annotated. (B) Random forest GWAS Manhattan plot for flowering time at 16°C where candidate genes have been assigned to the SNPs from the FLOR-ID database using a 20 kb window [9]. The lines indicate Bonferroni-significance thresholds for both plots. The plots have been created with the package `qqman`[2].

We have been interested in why these loci have been retained in individual random forest

[2] `https://cran.r-project.org/web/packages/qqman/index.html`

analyses while they are not significant in the linear mixed model and suggest that epistasis effects contribute to individual effects. The univariate permutation importance does not only take marginal effects but also interaction effects into account. Thus even if a SNP has been retained in univariate RF analyses, it might not show any significant marginal effect on its own. We will take these specific examples to show that the power to detect loci can be improved with random forests. Due to non-additive interaction effects, significant SNPs linked to these loci show very heterogeneous effects depending on the broader genetic background. Linear mixed models can not account for such heterogeneity as they fit average effects to SNPs, which are fixed for the mapping population.

The flowering time gene *SVP* acts as floral repressor and interacts with many other flowering-related genes[3]. When using a nonlinear model and allowing for interactions, *SVP* significantly contributes to the flowering time phenotype and is found to be involved in many interactions that have been retained in bivariate analyses. Its interaction with a SNP on chromosome 5 (chr5_s22627545-chr2_s9584590 (*SVP*); $p < 0.05$) is an example of an interaction where the effect of the alleles is diminished once they occur together (Fig. 5.6A). An analysis of variance (ANOVA) yielded significant variation among the four haplotypes ($p < 2.2e\text{-}16$) and a post hoc Tukey test showed that the effect of *SVP* is significant in the left branch (allele 0 of chr5_s22627545) of the tree ($p_{\mathrm{adj}} < 2.2e\text{-}16$). Although the effect in the right branch (allele 1 of chr5_s22627545) of the tree is not significant, the effect direction changes depending on the allele of the interacting SNP and thus is very heterogeneous. Furthermore, the minor allele of *SVP* (allele 1 of chr2_s9584590) shows a high variance in flowering time (0/1 and 1/1 haplotypes are significantly different in their mean flowering time; $p_{\mathrm{adj}} < 2.2e\text{-}16$). However, it does not show any significant difference to the other allele without considering the heterogeneous effects (allele 0 of chr2_s9584590; 72.14 ± 16.55 vs. 78.45 ± 20.9, $p = 0.0001$ in LMM), explaining why this locus remains undetected in traditional LMM after correcting for multiple testing.

The same situation occurs at *FY*, which has been described to be involved in flowering time regulation and which significantly contributes to the flowering time phenotype through interactions. Its interaction with a SNP on chromosome 4 (chr4_s5252410-chr5_s4334594 (*FY*); $p < 0.05$) is an example where the effect of *FY* is masked for some accessions (Fig. 5.6B). Also here an analysis of variance (ANOVA) yielded significant variation among the four haplotypes ($p = 5.092e\text{-}09$). A post hoc Tukey test showed that in the left branch (allele 0 of chr4_s5252410) of the tree, *FY* significantly contributes to the phenotype ($p_{\mathrm{adj}} < 2.2e\text{-}16$) while its effect is masked in the right branch (allele 1 of chr4_s5252410) of the tree ($p_{\mathrm{adj}} = 0.99$). Also here the minor allele shows a high variance in flowering time ($p_{\mathrm{adj}} = 0.01$).
Although a SNP linked to the flowering time regulator *LATE* is significant in univariate

[3] FLOR-ID database describing the flowering time network (`http://www.phytosystems.ulg.ac.be/florid/`).

random forest analyses, we could not identify any interaction involving it, which might be due to its low minor allele frequency of 6%. We still suggest that this locus has been retained in univariate random forest analyses due to epistasis effects.

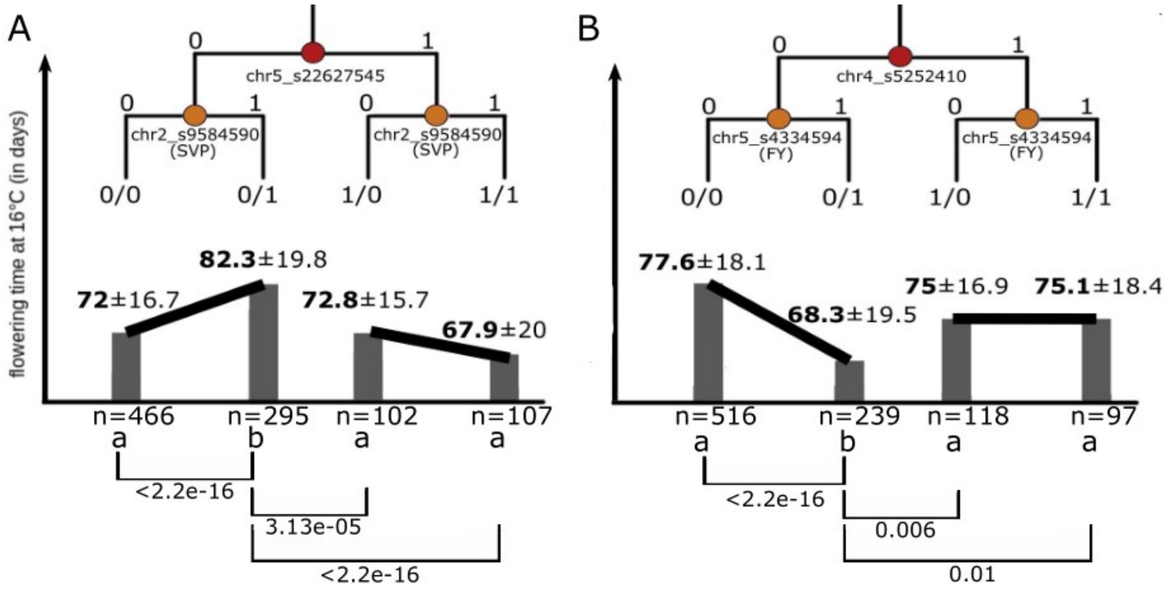

Figure 5.6: Trees of an interaction involving *SVP* (A), respectively *FY* (B) with the mean ± the standard deviation of the flowering phenotype, the sample size of the given haplotype and adjusted p-values from a multiple comparison Tukey test. Significant p-values are given and haplotypes with the same lowercase letter indicate no significant difference ($p_{adj} > 0.05$).

Manhattan plots for flowering time at 10°C with a panel of 1003 accessions including random forest and linear mixed model annotations can be found in the supplemental figures A (Fig. A.1). One additional locus has been identified in random forest univariate analyses (*GIBBERELLIN 3-OXIDASE 2* [*GA3OX2*]). While it is not significant in the LMM (most significant p-value from the LMM in a 20 kb window of the respective locus of 1.48e-04), it has been shown to be involved in interactions. All these cases together indicate that, in addition to the same power on linear effects, random forests could be further used to discover loci underlying nonlinear effects, which are most likely caused by epistatic interactions with other (hidden) candidate(s).

5.2.2 Interaction effects genome-wide

As the univariate GWAS analyses suggested that nonlinear interactions could be masked in the LMM and being recovered by random forest, we next wanted to identify non-additive interactions, while particularly focusing on new loci that possibly are not significant in univariate analyses. The interactions returned by minimal depth are candidate interactions and might actually involve additive effects of two SNPs. Thus we have not applied any Bonferroni-correction to the bivariate p-values but have considered all interactions with raw bivariate p-values < 0.05 and filtered them for interesting non-additive patterns using pairwise t-tests (methods described in section 4.2). SNPs involved in interesting interactions have been cross-referenced with the FLOR-ID database using a 20 kb window [9]. This allowed us to identify new loci, but first we will concentrate on two interactions involving loci that have also been retained in univariate analyses (*FT*-*SVP*; *DOG1*-*FLC*).

The interaction involving *FT* and *SVP* revealed one of the lowest p-values (chr1_s24339560 and chr2_s9584590; p = 0.022). It has been shown that the effect of *SVP* significantly depends on the genetic background and thus epistasis effects are crucial for this gene [76]. *SVP* is significant in individual random forest analyses, however it can not be identified in the linear mixed model that only considers marginal effects. We have already shown an interaction involving this same *SVP* SNP in figure 5.6A in univariate GWAS results. However, the SNP on chromosome 5 involved in the previous interaction is not in linkage with the *FT* SNP. The interaction between *FT* and *SVP* reveals that for accessions carrying the minor allele of *FT* (allele 1 of chr1_s24339560), the effect of *SVP* is masked ($p_{\text{adj}} = 0.17$) while for accessions carrying the major allele (allele 0 of chr1_s24339560), *SVP* represses flowering and thus there is an effect ($p_{\text{adj}} = 1.54\text{e-}04$). The interaction tree can be found in the supplemental figures A (Fig. A.3). We have only analyzed pairwise interactions but considering the common effect of the two loci acting on *SVP* (chr5_s22627545 and *FT*), we found that their effect is rather additive and thus independent.

Furthermore, the pairwise GWAS analyses revealed a nonlinear interaction between *DOG1* and *FLC* involving chr5_s18590741 and chr5_s3172910 (p = 0.031; chr5_s18590741 located in the *DOG1* gene; chr5_s3172910 located < 1 kb downstream of the *FLC* gene). An analysis of variance (ANOVA) yielded significant variation among the four haplotypes (p < 2.2e-16) and a post hoc Tukey test showed that the 1/1 haplotype is significantly different in mean flowering time from all the others (Fig. 5.7A; $p_{\text{adj}} < 2.2\text{e-}16$). Although the interaction is interesting in both directions, we will concentrate on the effect of *DOG1* on *FLC*. This firstly shows that for some accessions the effect of *FLC* is masked and secondly that the minor allele of *FLC* shows a high variance (0/1 and 1/1 haplotypes are significantly different in their mean flowering time; $p_{\text{adj}} < 2.2\text{e-}16$), suggesting heterogeneity due to epistasis effects. The geographical map reveals that most of the haplotypes carrying the *DOG1* allele for which the *FLC* effect is significantly larger (haplotype 1/1) are located in Spain (Fig. 5.7B).

FLC has been described to be the main flowering time regulating gene. While in the linear mixed model, the *DOG1* peak is higher than the *FLC* peak, in random forests it is the other way around as *FLC* has been returned higher permutation importance. As random forests take interaction effects into account, it may reflect the biology behind the flowering time network better: *FLC* is one of the major hubs in the network and the highest peak in the random forest GWAS (Fig. 5.5B).

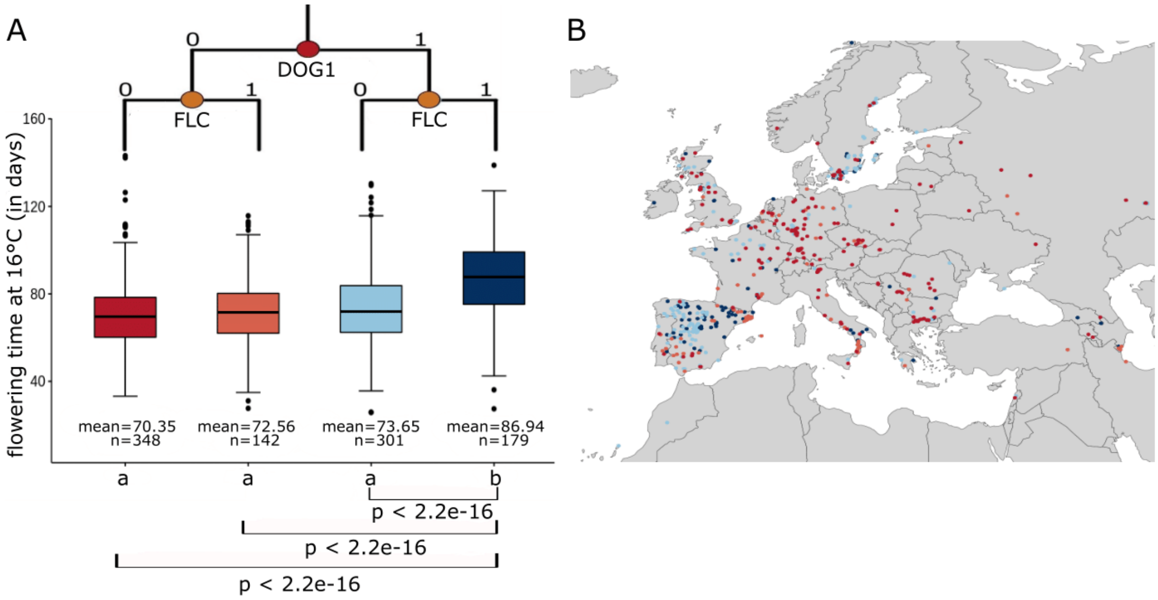

Figure 5.7: Non-additive interaction involving *DOG1* and *FLC*. (A) The boxplots are annotated with the mean and the sample size of the given haplotype. The p-values are adjusted p-values from a multiple comparison Tukey test. Significant p-values are given and haplotypes with the same lowercase letter indicate no significant difference. (B) Geographical distribution of the four haplotypes.

In addition to these interactions including loci that have been individually significant too, the pairwise analyses reported other interactions involving loci that are not individually significant in RF and that can not be identified in the LMM including *ARABIDOPSIS ASH2 RELATIVE* (*ASH2R*), *ARABIDOPSIS THALIANA CENTRORADIALIS* (*ATC*), *GA REQUIRING 1* (*GA1*), *GOLDEN2-LIKE 1* (*GLK1*) and *MICRORNA156E* (*MIR156E*). Although all of these new genes identified in pairwise interactions are related to flowering time, there is no further evidence in the FLOR-ID that they interact with specific genes. Furthermore, many of the SNPs involved in interactions could not be linked to any known flowering time gene, which complicates the interpretation of SNP-SNP interactions. We have listed all interactions including the above loci contributing to the flowering time phenotype at 16°C in the supplemental data A (Table A.1), providing a valuable resource for further explorations.

Finally, we want to emphasize how interactions identified in RF analyses could increase our understanding on previous findings. The *TWIN SISTER OF FT* (*TSF*) flowering time gene has been shown to interact with *DOG1* in bivariate RF analyses for the flowering time phenotype at 10°C (chr4_s10983356 and chr5_s18590741; Fig. 5.8A; n = 1003 accessions). Interestingly, *TSF* is shown only significant in one subpopulation (Southern Iberian Peninsula, SIP) with regular GWAS [8]. The authors have shown that GWAS results depend on where your samples come from and that sampling across a wide range of environments increases genetic heterogeneity and thus reduces the power of GWAS [8]. They have divided accessions (a subset of 888 accessions of the 1003 accessions including only European accessions) into 8 subpopulations using geographic information to identify more peaks specific to local populations. The most significant *DOG1* SNP (chr5_18590501) has a very low minor allele frequency for the SIP population (3%) while relatively intermediate in the others (5% to 48%), which leads to the alternative explanation that the effect of *TSF* is only detectable for accessions carrying the major allele of that *DOG1* SNP. For the interaction involving chr5_s18590741 (*DOG1*; $r^2 = 0.16$ to chr5_18590501) and chr4_s10983356 (*TSF*), we observed that the effect of *TSF* is only visible for accessions carrying the 0 allele of chr5_s18590741 (*DOG1*) ($p_{adj} = 2.2\text{e-}06$), while it is masked for accessions carrying the other allele (allele 1 of chr5_s18590741; $p_{adj} = 0.93$; Fig. 5.8A for all 1003 accessions). While on a global level, only ≈ 51% carry the 0 allele of the *DOG1* SNP (chr5_s18590741), for the SIP population it is ≈ 82% (Fig. 5.8A,B). Thus the effect on the flowering time phenotype is significant for the SIP subpopulation (p = 7.2e-06; Fig. 5.8D) while it remains undetected on a global level (p = 0.1; Fig. 5.8C). This shows that random forests can handle genetic heterogeneity and might be more interpretable without building separate models for separate subpopulations.

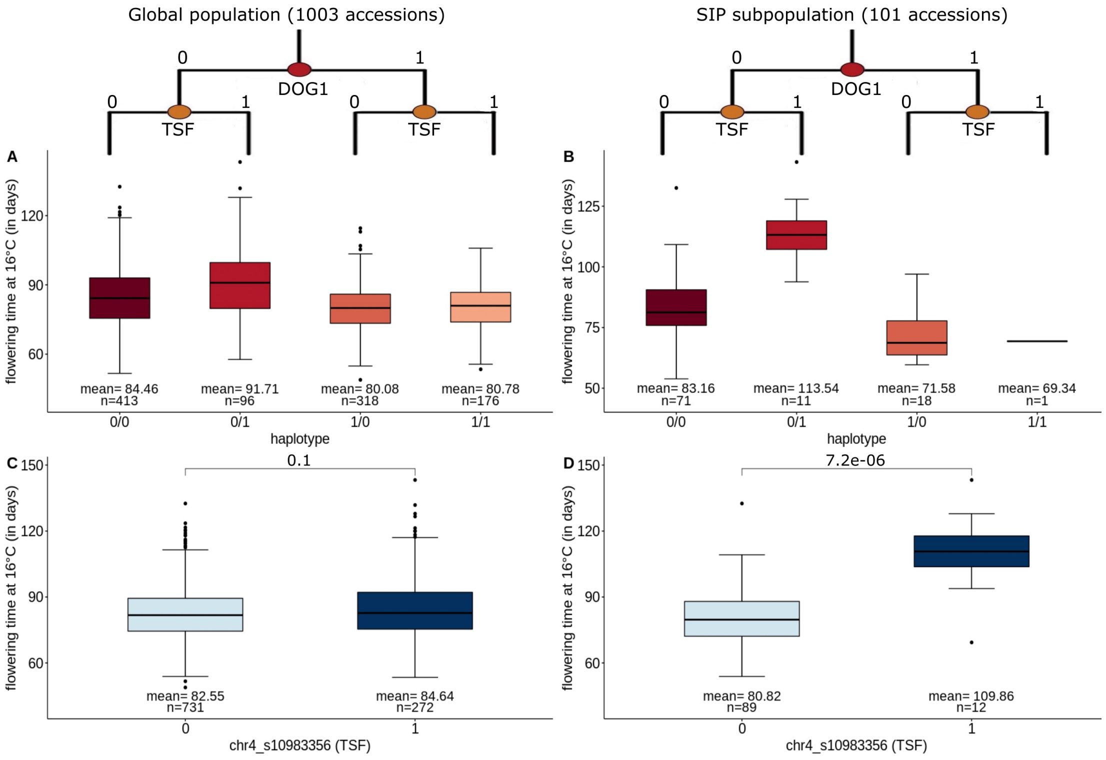

Figure 5.8: Random forest can handle genetic heterogeneity by allowing for epistatic interactions. (A,B) Interaction involving *DOG1* and *TSF* for the global and the SIP (South Iberian Peninsula) subpopulation, respectively. The boxplots are annotated with the mean and the sample size of the given haplotype. (C,D) Effect of *TSF* on the flowering time phenotype at 10°C for the global and the SIP (South Iberian Peninsula) subpopulation, respectively. The p-values are from pairwise t-tests.

5.3 RF improves TWAS power by incorporating nonlinear effects

5.3.1 Individually significant genes

In the previous section, we showed that random forests improve the power of GWAS as nonlinear interactions can be recovered. Our first interest in RF TWAS is whether we can identify individual genes whose expression have a more complex relationship to the flowering time phenotype and remain undetected by a linear model. Therefore, we first obtained TWAS results for flowering time at 16°C by fitting a linear mixed model to the data (Fig. 5.9A; implemented in `OSCA` --moa [35]; see section 4.2). We used a panel of 453 *A. thaliana* accessions, a subset of the 970 GWAS accessions for which genes' expression levels were available. After Bonferroni-correction, there are two significant associations including *SUPPRESSOR OF OVEREXPRESSION OF CO 1* (*SOC1*) and *AT4G33625*. *SOC1* is one of the major hubs in the flowering network [77], however it can not be identified by GWAS. We will elaborate on this later and show that there is no correlated *SOC1* polymorphism (Fig. 5.10). The second most significant gene *AT4G33625* was identified in a previously published *A. thaliana* TWAS (using a LMM, $n = 690$ accessions [21]), however, there is no evidence in the literature for it to be flowering time related. We have considered the 10 most significant genes and cross-referenced these with the FLOR-ID database [9] and the previously published TWAS results [21]. We found that 2 genes are included in the FLOR-ID database (*SOC1*, *FRUITFULL* [*FUL*]), 5 genes have been identified in the TWAS study too (*SOC1*, *RNA HELICASE 30* [*RH30*], *PHYTOCHROME INTERACTING FACTOR 3* [*PIF3*], *AT4G33625*, *AT5G49360*) and *TREHALOSE-PHOSPHATASE/SYNTHASE 9* (*TPS9*) has also been linked to flowering time before [78]. As there is no clear evidence of a link to flowering time for the two genes *AT4G33625* and *AT5G49360*, the linear mixed model identified 5 known flowering time regulators among the 10 most significant genes.

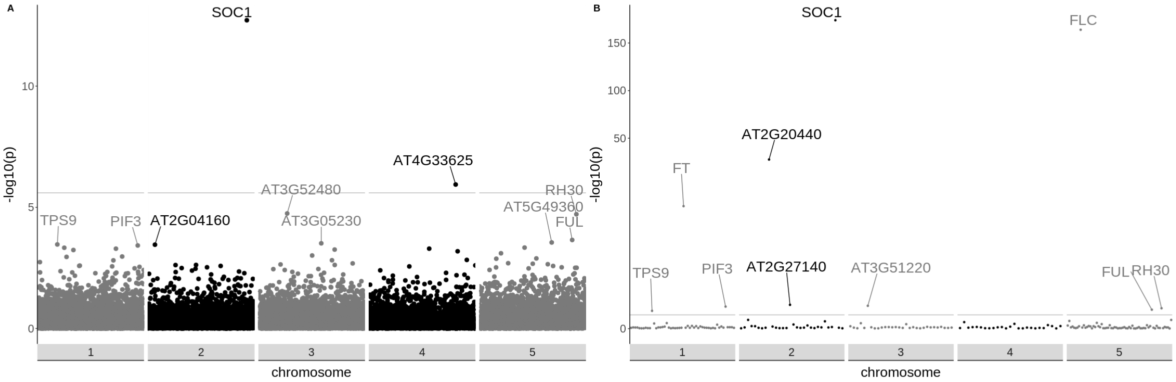

Figure 5.9: Random forests improve the power of TWAS; more loci can be identified. (A) Linear mixed model TWAS Manhattan plot for flowering time at 16°C. (B) Random forest TWAS Manhattan plot for flowering time at 16°C. For both plots, the 10 most significant genes have been annotated. The lines indicate Bonferroni-significance thresholds.

In random forest univariate analyses on gene expression, we could identify 10 significant associations with the same flowering time phenotype, including all the 5 flowering time regulators discovered by LMM TWAS (Fig. 5.9B), and another 3 new trait-associated genes (*FT*, *FLC* and *AT2G20440*) whose expression levels have been associated with flowering time [9, 21]. *FLC* and *FT* are the only genes that have been identified in both random forest GWAS and TWAS analyses and it actually has been observed before that both *FLC* polymorphism and *FLC* expression are correlated with flowering time independently [79]. Compared to GWAS analyses, we could identify new candidate genes in TWAS, thus the associations are complementary to each other, which is consistent with the results in [21]. *SOC1* has been ranked to be the most predictive and thus the most flowering time related gene in TWAS. Its gene expression levels and the phenotype are highly correlated (r = -0.48, $p < 2.2e\text{-}16$), leading it to be the most significant gene in the LMM as well. In *SOC1* eQTL[4] however, no SNPs passed the significance threshold and there is also no significant GWAS peak (Fig. 5.10). Still, variation in expression level of this gene contributes to phenotypic variation, and thus it can be identified by TWAS. As we have already mentioned in section 1.2.5, variation in gene expression levels could be due to unrevealed genetic variants, the complex regulation consequence of the other regulators, or epigenetic variation in or near the gene.

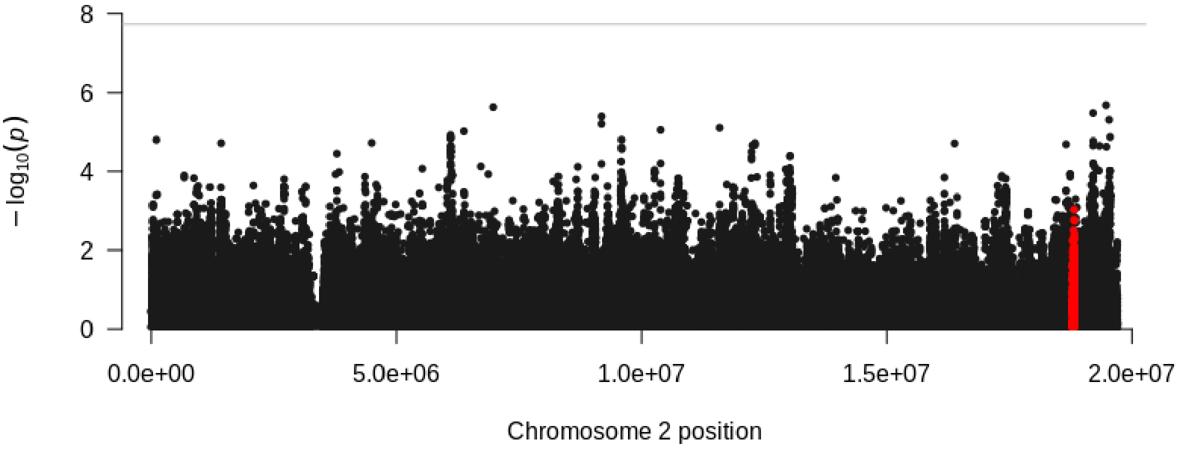

Figure 5.10: GWAS Manhattan plot for flowering time at 16°C zooming in on chromosome 2. SNPs in a 20 kb window around *SOC1* have been highlighted in red. The line in grey indicates the genome-wide Bonferroni-significance threshold.

Among the 10 most significant associations, random forest TWAS could identify 8 genes that have been related with flowering time before, including the 5 genes from the linear mixed model[5]. The advantage of random forests is that they can capture complex, nonlinear relationships between the predictor variables and the outcome. Partial dependence plots can be used to visualize the predicted phenotype as a function of the continuous gene expression

[4] The expression quantitative trait loci (eQTL) results have been generated by Yoav Voichek.

[5] Also among the 20 most significant associations from the linear mixed model, neither of the remaining 3 genes (*FLC*, *AT2G20440*, *FT*) from random forests can be identified. They have been ranked as 23rd, 35th and 290th by their p-value, respectively.

levels, i.e. to visualize the average effect of gene expression on the phenotype (see section 2.4.1). As a control, we added the partial dependence plot of *SOC1*, where the average effect is mostly linear (Fig. 5.11). The effect of the three genes that can not be identified by the linear model (*FLC*, *AT2G20440*, *FT*), however, is only partly linear (Fig. 5.11). *FLC* for example, although its expression levels are high in late flowering and low in early flowering accessions, is only linearly related to flowering time for high expression levels (r = 0.27, p = 8.91e-07) while it does not show any significant effect on the flowering phenotype for low levels (r = 0.06, p = 0.49[6]). *AT2G20440* has a rather sigmoidal effect on the phenotype and *FT* only shows an effect for low expression levels. For these nonlinear cases, random forests can clearly improve the power of TWAS.

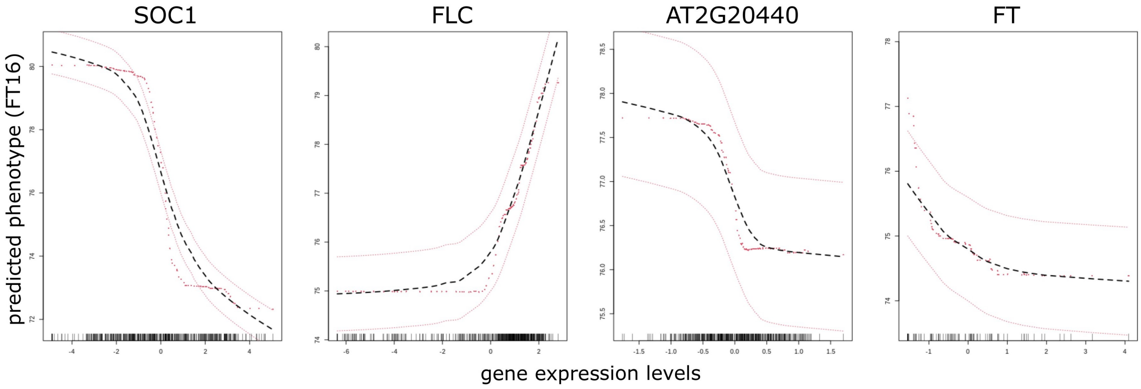

Figure 5.11: Partial dependence plots of *SOC1* and of the 3 genes that can not be identified in the linear mixed model (*FLC*, *AT2G20440*, *FT*). The effect of *SOC1* on the phenotype is clearly linear while for the others it is only partly linear. The genes are sorted by their variable importance. The red points are used to indicate partial values (realisations of the partial dependence function), the dashed red lines indicate a smoothed error bar of ± two standard errors and the black dashed line is a smoothed function of the partial values. The density bars at the bottom represent the distribution of the observations. The larger the range on the y-axis, the larger the effect on the flowering time phenotype (16°C). The plots have been created with the package `randomForestSRC`[8].

Finally, beside the 8 genes that have been related to flowering time before, random forest TWAS identified two genes without any evidence in the literature to link them to flowering time (*AT2G27140* and *AT3G51220*). For both *AT2G27140* and *AT3G51220*, the gene expression levels are correlated with the phenotype (r = -0.19, p = 6.28e-05; r = -0.18, p = 8.47e-05, respectively). However, interaction effects could also contribute to individual effects as they are not significant in the LMM analysis, and the permutation variable importance in random forest does take interaction effects into account. We highlight how random forest might have identified *AT3G51220* and how it can be interpreted with the help of tools

[6] When dividing the accessions into two subpopulations based on where an effect of *FLC* is visible.

[8] `https://cran.r-project.org/web/packages/randomForestSRC/index.html`

from interpretable machine learning (see section 2.4). When considering the partial dependence plot (PDP) to visualize the effect of *AT3G51220* on the phenotype, there is a rather straight line indicating that there is no effect on the flowering time phenotype (yellow line in Fig. 5.12A). Individual conditional expectation plots extend the results from PDPs as they return the relationship between a gene of interest and the predicted phenotype for each individual observation and thus can take interaction effects into account. Indeed, the individual effects of *AT3G51220* are heterogeneous (grey lines in Fig. 5.12A). For some accessions, there is an effect on the phenotype while for other accessions the effect seems to be masked.

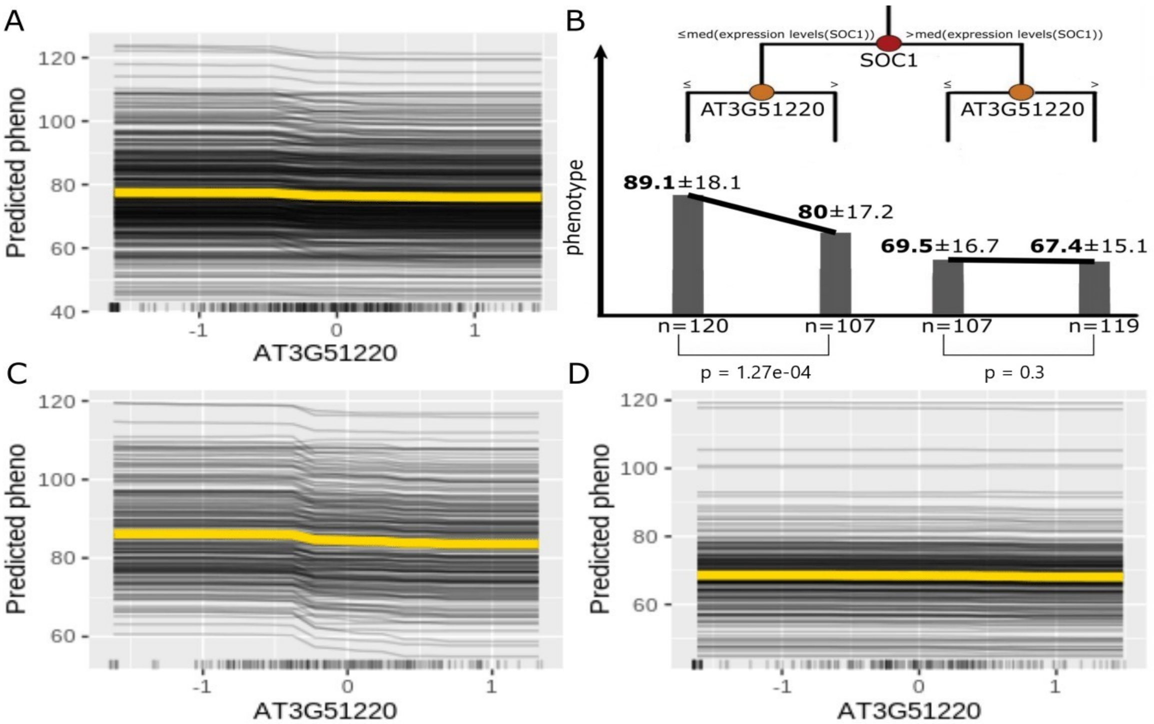

Figure 5.12: Partial dependence plots (PDP), individual conditional expectation plots (ICE) and interaction involving the significant gene *AT3G51220*. (A) PDP (yellow line) and ICE (grey lines) plot of *AT3G51220*. (B) Non-additive interaction with *SOC1*. The p-values are from pairwise t-tests. (C,D) Partial dependence and individual conditional expectation plots of a subset of accessions from the left (C) and the right (D) branch of the interaction tree, respectively. The plots in (A), (C) and (D) have been created with the package `iml`[10].

In bivariate analyses, we identified an interaction involving *AT3G51220* and *SOC1* which shows that there are heterogeneous effects on the phenotype depending on the broader background (Fig. 5.12B). The effect of *AT3G51220* seems to be masked once *SOC1* shows high expression levels. In a next step, we can consider partial dependence plots and individual conditional expectation plots when taking only the subset of accessions from the low and high *SOC1* expression (left and right branch of the tree; n = 227 in Fig. 5.12C; n = 226 in Fig. 5.12D), respectively. Correlations between the phenotype and gene expression levels

[10] `https://cran.r-project.org/web/packages/iml/index.html`

of *AT3G51220* change when dividing the accessions into two subpopulations based on the interaction with *SOC1.* While accessions in the left branch have a higher correlation with the phenotype (r = -0.33, p = 2.94e-07) and the individual conditional expectation plot is more homogeneous (Fig. 5.12C; however, there still are accessions without any effect on the phenotype), there is no effect on the phenotype for accessions in the right branch (Fig. 5.12D; r = 0, p = 0.95). Similar observations apply to *AT2G27140.* Confirming the potential links of these novel candidates to flowering time might be worth to be studied experimentally. Manhattan plots for flowering time at 10°C with a panel of 470 accessions including random forest and linear mixed model annotations can be found in the supplemental figures A (Fig. A.2). In random forest analyses, 5 of the 10 genes overlap with the associations from the phenotype at 16°C while the others are novel candidates.

5.3.2 Interaction effects of gene-pair expressions

After univariate TWAS analyses, and inspired by the interaction finding between *AT3G51220* and *SOC1*, we wanted to identify non-additive interactions involving genes that possibly are not significant in univariate analyses. The interactions returned by minimal depth are candidate interactions and might actually involve additive effects of two genes. Thus we have not applied any Bonferroni-correction to the bivariate p-values but have considered all interactions with raw bivariate p-values < 0.05 and filtered them for interesting non-additive patterns using pairwise t-tests (methods described in section 4.2). Consistent with the observation that *SOC1* is one of the major hubs in the regulatory networks that underlie floral timing and flower development by controlling other genes [80], *SOC1* has been found to be involved in many non-additive interactions.

Although we could not identify any new candidate gene from the FLOR-ID database, we did identify novel genes involved in non-additive interactions, including *SEPALLATA3* (*SEP3*), which is a potential flowering time gene targeted by *SOC1* [80]. *SEP3* has no average effect on the phenotype (Fig. 5.13B). However, it has been described that *SOC1* binds to the regulatory regions of *SEP3* [80]. By binding to *SEP3*, *SOC1* activates *SEP3* and promotes its expression, and thus also its effect on flowering time. Accessions that have low *SOC1* expression seem to have a masked (or at least not very significant due to a low sample size) effect of *SEP3* while for accessions with high *SOC1* expression, *SEP3* has a significant effect on the flowering phenotype (p = 5.8e-04) as *SOC1* promotes its expression (Fig. 5.13A). Similar to *SOC1*, *SEP3* also promotes flowering. Furthermore, we could retain interesting interactions between known flowering regulators that have been confirmed experimentally. *FLC* inhibits flowering (high expression leads to late flowering) while *SOC1* promotes flowering (high expression leads to early flowering). It has been shown that *FLC* interacts with *SOC1* and delays flowering by repressing *SOC1* via direct binding to the *SOC1* promoter region, thus high expression of *FLC* reduces the promotive response of floral signals [81]. First, there is a notable difference in sample sizes; when *FLC* expression levels are low, many accessions have high *SOC1* expression levels (n = 180); however, when *FLC* expression levels are high, only a few accessions have high *SOC1* expression levels (n = 46) as *FLC* represses the expression of *SOC1* (Fig. 5.13C). Also the mean *SOC1* expression is significantly different depending on *FLC* expression (Fig. 5.13D; p < 2.2e-16). This also leads to the fact that the effect size of *SOC1* is different depending on whether *FLC* expression is high or low (Fig. 5.13C).

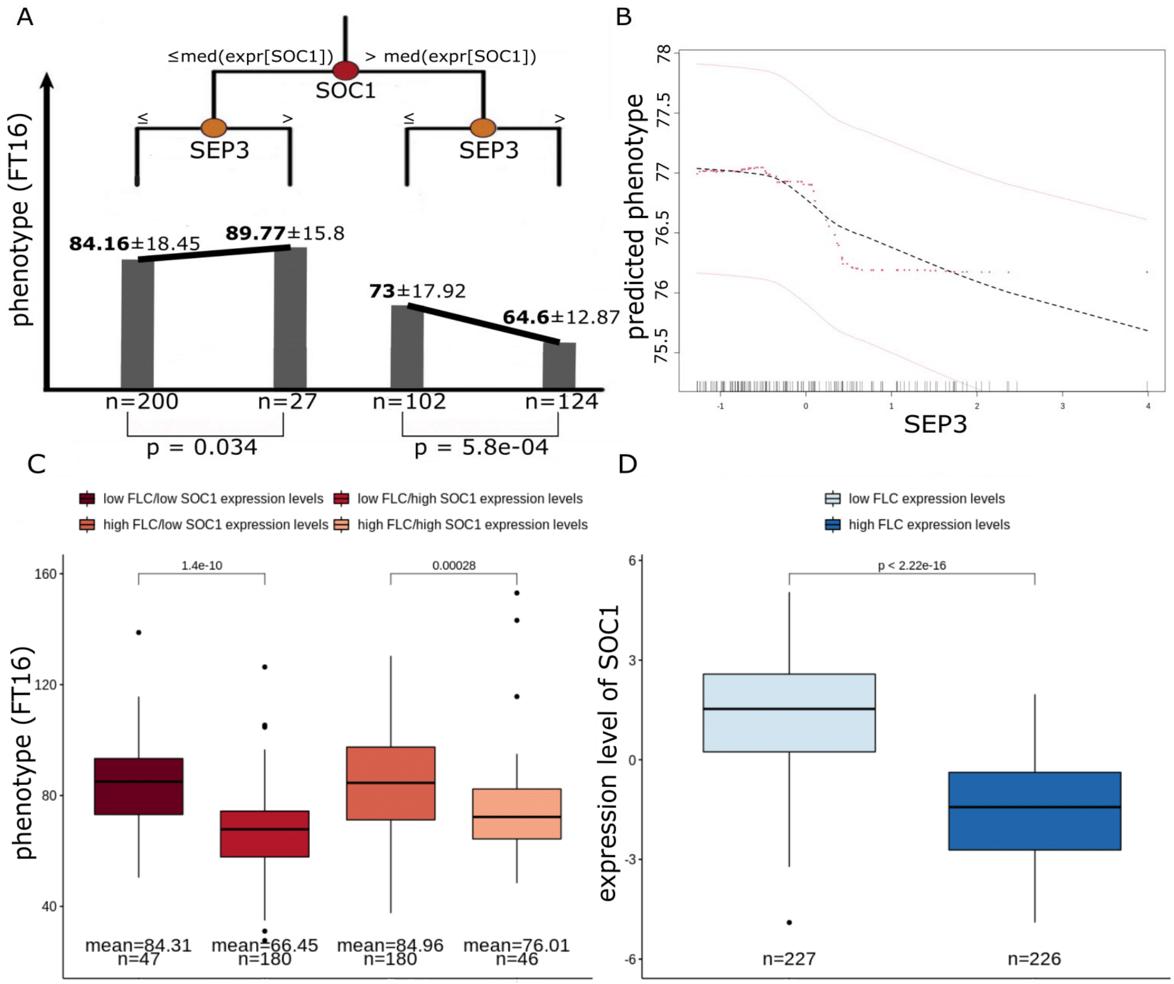

Figure 5.13: Examples of non-additive interactions that have been identified in bivariate TWAS analyses. (A) Interaction involving *SOC1* and *SEP3*. (B) Partial dependence plot of *SEP3*. (C) Interaction involving *FLC* and *SOC1*. The effect size of *SOC1* depends on expression levels of *FLC* as *FLC* has been shown to repress *SOC1*. (D) *SOC1* shows significantly different expression levels depending on *FLC* expression levels. The p-values are from pairwise t-tests.

Random forests improve the power of TWAS by allowing nonlinear relationships. Furthermore, interactions allow us to better understand the biology behind the flowering network. We proved that experimentally evidenced observations can be confirmed by the data, hence novel genes involved in non-additive interaction might be target genes. The identified interactions involving novel genes can be found in the supplemental data A (Table A.2).

5.4 RF leverages GTWAS by combining multi-omics

5.4.1 Individually significant SNPs and genes' expression

Random forests allow to integrate different data types (binary, multi-categorical or continuous) in one single model and hence to construct GTWAS by integrating omics data from different sources: in our present case SNPs and gene expression levels. We wanted to find out whether we could detect more loci/relationships than in individual GWAS and TWAS and get new insights into the flowering time network. For flowering time at 16°C, a panel of 453 accessions, for which both SNPs and gene expression levels were available, have been used.

In univariate analysis of GTWAS, we could identify a subset of the loci that have been identified in GWAS and TWAS, including *FLC* and *DOG1* in the form of SNPs and *SOC1*, *FLC*, *AT2G20440*, *FT* and *PIF3* in the form of expression levels (Fig. 5.14). As eQTLs have been filtered out due to correlations between the two data types (see section 4), the *FLC* polymorphism and *FLC* expression are not correlated and thus are identified independently of each other.

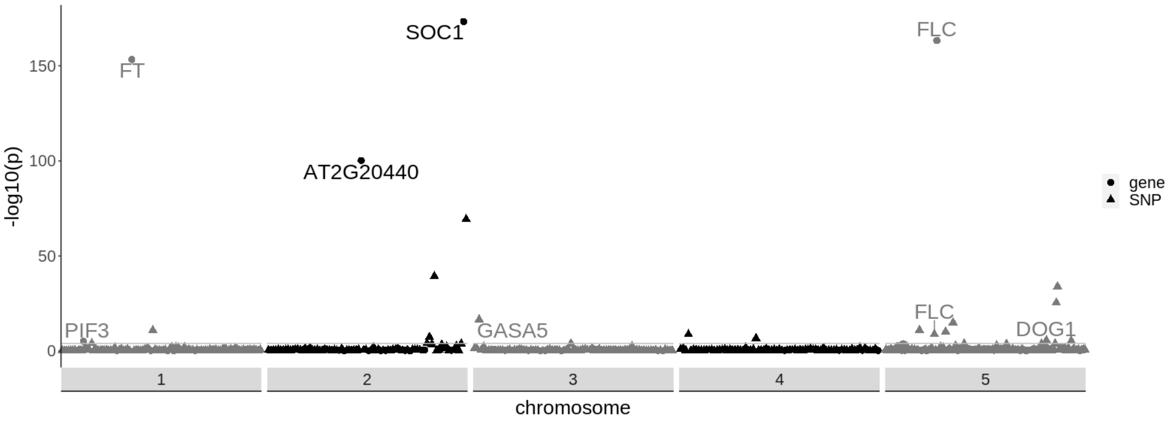

Figure 5.14: Random forests open GTWAS and allow the identification of one new locus *GASA5*. GTWAS Manhattan plot for flowering time at 16°C where candidate genes have been assigned to the SNPs from the FLOR-ID database using a 20 kb window [9]. The line indicates the Bonferroni-significance threshold.

Furthermore, a SNP (chr3_646496) near *GAST1 PROTEIN HOMOLOG 5* (*GASA5*), a gene included in the FLOR-ID database, has been identified. This gene has been described to delay flowering by enhancing *FLC* expression and by repressing the expression of the two flowering time genes *FT* and *LEAFY* (*LFY*) [82]. Probably due to its low minor allele frequency (3%), no SNP interaction involving *GASA5* was retained in bivariate analyses. However both for *FLC* and *FT*, we observed significant differences in mean expression for accessions carrying the major respectively the minor allele of *GASA5* (p = 1.64e-04 for *FLC*, p = 1.6e-13 for *FLC*). For accessions carrying the minor allele, mean expression is higher for *FLC* and lower for *FT*, consistent with the observation that *GASA5* inhibits flowering by promoting *FLC* and repressing *FT* expression. Important to mention is that the genetic variant linked to *GASA5* is not highly correlated with *FLC* or *FT* (r = 0.07 and r = -0.06, respectively) expression as eQTLs have been filtered out.

5.4.2 Interaction effects between SNPs and genes' expression

In bivariate GTWAS, we identified three new flowering time genes including *AGAMOUS-LIKE 42* (*AGL42*), *AT2G32870* and *PHYTOCHROME INTERACTING FACTOR 4* (*PIF4*; chr2_17892882). *AT2G32870* has not been linked to flowering time before but *AGL42* has been described to be involved in the floral transition and to promote flowering [83]. Both were not identified in TWAS but have been found to be involved in many non-additive interactions with SNPs (when also considering interactions with p-values > 0.05), however only the interaction with *SOC1*, respectively *FLC* remained significant (Table A.3). This still shows that adding omic data from other sources helps to identify new loci due to possible interactions.

The flowering time gene *PIF4* (chr2_17892882) has been found to be involved in a SNP-gene interaction with *SOC1*. The SNP linked to *PIF4* does not show any significant effect on flowering time on its own (Fig. 5.15A; p = 0.064). However, when considering the interaction with the expression of *SOC1*, its effect on the phenotype is very heterogeneous and even shows different effect directions with significant differences in mean flowering time depending on *SOC1* expression (Fig. 5.15B). As in *SOC1* and *PIF4* eQTL no SNP passed the significance threshold, we can rule out that *PIF4* is an eQTL for *SOC1* or vice versa, suggesting that SNP-gene interactions are largely nonlinear.

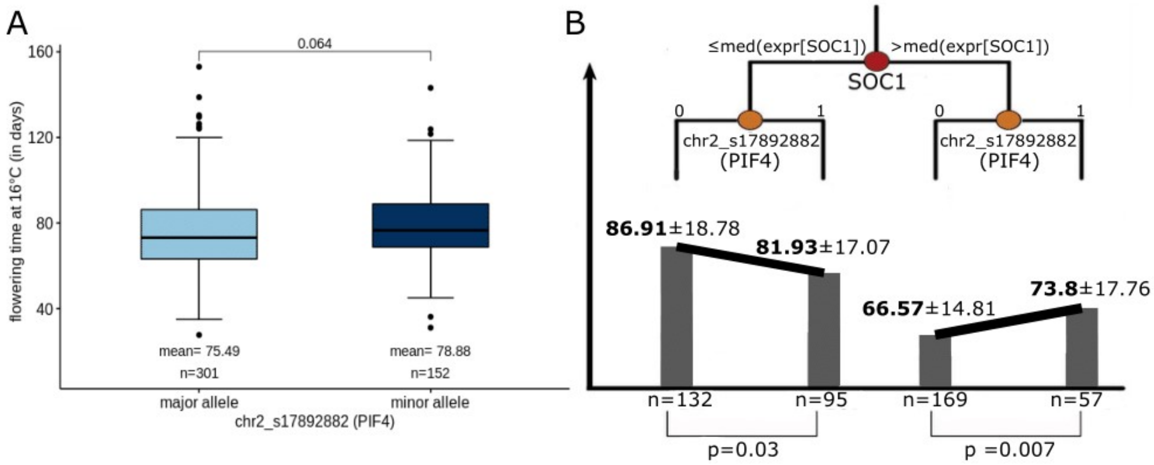

Figure 5.15: Non-additive interaction that has been identified in bivariate GTWAS involving *SOC1* and *PIF4*. (B) Interaction involving *SOC1* and *PIF4* with their mean ± standard deviation of the flowering phenotype and the sample size of the given haplotype. The p-values are from pairwise t-tests.

Although the biological evidence for this interaction is not clear, the results show that GTWAS can identify new loci and new interactions that are not identified in single GWAS or TWAS analyses. Thus adding multi-omics in one model results in new peaks that can lead to new and more complex biological discoveries. Interactions that have been retained in the bivariate GTWAS analyses can be found in the supplemental data A (Table A.3).

One important point to mention is that in univariate and bivariate GTWAS, we have identified almost all flowering time genes (except from *TPS9*) that were significant in univariate RF TWAS, while only two genes from univariate RF GWAS have been identified (*FLC* and *DOG1*). However, one has to take into account here that only a subset ($n = 453$) of the 970 accessions that were available in GWAS, have been used in GTWAS.

Chapter 6

Conclusion and discussion

Association studies aim to identify loci that are significantly associated to the phenotype, which is a very fundamental issue in genetics. Both the power of identification of candidate loci and the understanding of how underlying genetic effects contribute to phenotypic variation (the interpretability of the mathematical model) are important in practice. The goal of this thesis was to show that a model, which allows for nonlinear effects, epistatic interactions and heterogeneity, increases the power of association studies. Although machine learning models are often accused to be “black boxes” where it is not clear how the results have been obtained, we were able to develop an improved pipeline based on random forests which is well interpretable and statistically justified. This pipeline allowed us to identify both individual, possibly nonlinear loci and potentially interacting loci (with the interacting locus being identifiable or hidden) through the univariate permutation importance and the bivariate minimal depth. Both in genome-wide (GWAS) and transcriptome-wide association studies (TWAS), random forests outperformed the standard linear mixed model by the identification of more, known and novel, flowering time genes. Combining an ensemble approach of both GWAS and TWAS while opening genome-and-transcriptome-wide association studies (GTWAS), we showed that we could increase the association power by identifying new loci that are not significant in individual analyses. Generally, this shows the feasibility of integrating machine learning models in association studies.

Random forests have the advantage that they are non-parametric and hence do not assume any genetic model. A very important finding was that the main hub in the flowering network *FLC* has been identified as the most important gene in random forest GWAS. Moreover, the flowering time regulator *SVP*, which is known to be involved in epistatic interactions, has been identified in random forests while remaining undetected in the linear mixed model. Therefore, random forests might learn the true underlying model better than parametric models when the genetic architecture is complex. They also outperformed the standard approach in TWAS as they do not assume linear effects on the phenotype and thus could identify more known flowering time genes. Furthermore, random forests facilitate the identification of epistatic interactions as the search space can be reduced through the univariate permutation importance. Regarding this, we identified epistatic interactions that involve loci contributing in a non-additive way to the phenotype. While some of the statistical interactions identified in the data have been confirmed experimentally and have been described in the literature, others need to be evaluated at a biological level in order to study their real effects and mechanisms on influencing flowering time, and thus to exclude the possibility that they are confounded with some hidden factors (e.g. the environment), although these have been carefully controlled in the current pipeline.

Although random forests performed better than the standard linear mixed model, they do have some limitations. First, the computational load is much larger compared to the linear mixed model. This is due to the data preprocessing steps, and mainly the correction for population structure of the genotypic data, but also the introduction of significance levels increases computations since multiple models have to be built. Then, a further disadvantage of random forests is that the results might not be 100% reproducible due to the stochasticity. Although we improved the reliability of the associations by introducing significance levels, we realized that the pipeline is very sensitive to randomness, especially as the dimensionality of the data is reduced in each step. Moreover, the removal of high linkage groups in GWAS complicates the interpretability of the associations as in contrast to linear mixed models, the number of SNPs tagging the causal variants is reduced. Finally, some loci that have been identified in RF GWAS, could not be identified in the ensemble approach GTWAS. As we have already mentioned, this might simply relate to the fact that only a subset ($n = 453$) of the 970 accessions from RF GWAS, have been used. However, it might also partially be due to the phenomenon that random forests prefer splitting on continuous variables if both data types are available in the subset of candidate splitting variables [84]. This would lead to SNPs being more unlikely to be returned a high variable importance and SNP associations being harder to be identified in GTWAS. We would expect that all these aspects could be overcome by further improvements.

As the results of our pipeline are encouraging, we propose a wide application to more traits and/or more species, with some future directions that could be considered to improve. At the data preprocessing step, we only filtered the SNPs for local linkage disequilibrium (LD). However, long-distance LD does exist due to population structure and might lead to false negatives. Therefore, the performance of the pipeline might be even better if long-distance LD was corrected. Furthermore, we only focused on pairwise interactions and higher-order interactions were beyond the scope of this work. However, they are likely to play an important role in complex traits and thus new biological discoveries might be possible when taking them into account. Considering these and potentially other improvements on the aforementioned limitations, random forests provide a powerful alternative to the traditional approaches in association studies, and therefore are very promising and worth studying in the future.

Appendix A

Supplemental figures and data

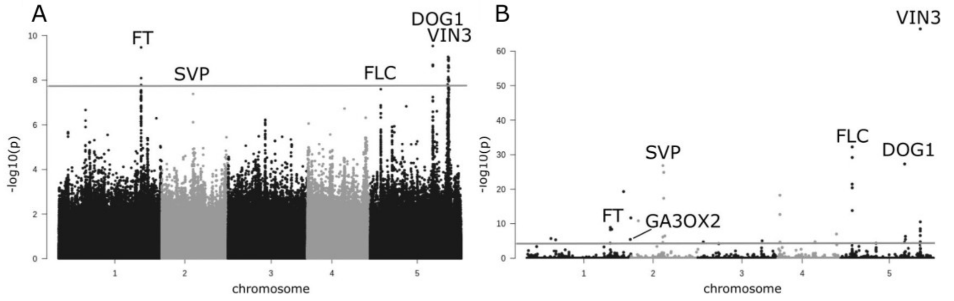

Figure A.1: Linear mixed model (A) and random forest (B) GWAS results for flowering time at 10°C. Candidate genes have been assigned to the SNPs from the FLOR-ID database using a 20 kb window [9]. The lines indicate Bonferroni-significance thresholds for both plots.

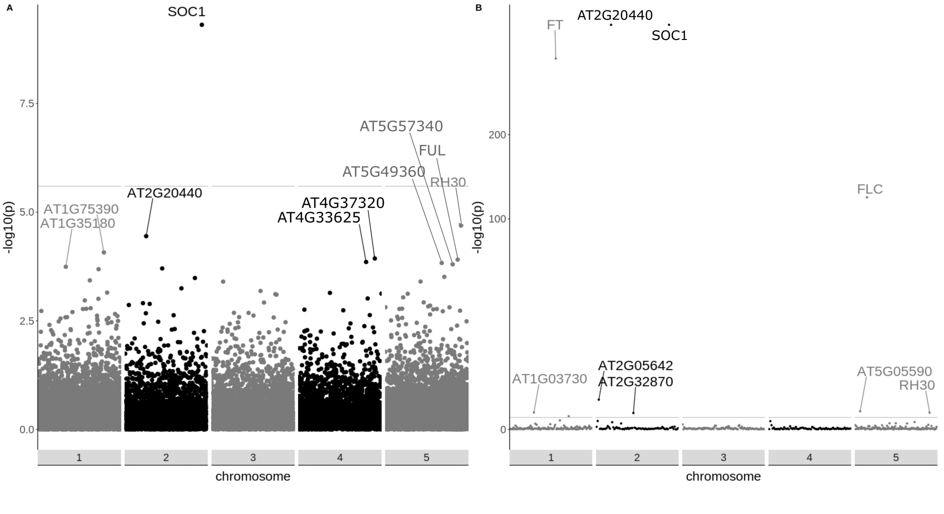

Figure A.2: Linear mixed model (A) and random forest (B) TWAS results for flowering time at 10°C. The lines indicate Bonferroni-significance thresholds for both plots.

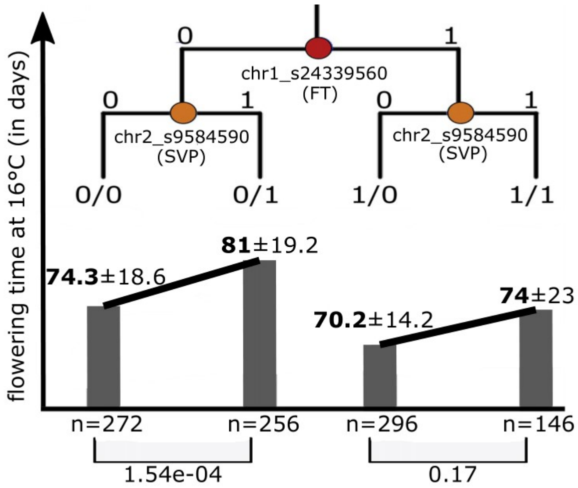

Figure A.3: Tree of the interaction involving *FT* and *SVP* with the mean ± the standard deviation of the flowering phenotype, the sample size of the given haplotype and adjusted p-values from a multiple comparison Tukey test.

Algorithm Analysis based on pairwise t-tests in order to interpret candidate interactions that have been identified by the interaction minimal depth.

```
For i = 1,...,n (number of candidate interactions):
1. Separate the accessions in 4 distinct haplotypes $h_1$, $h_2$, $h_3$ and $h_4$ based on the values of the variables involved in $interaction_i$ with their mean phenotypes $m_1$, $m_2$, $m_3$ and $m_4$.
2. Perform two-sample Welch t-tests $test_1$ = t.test($h_1$, $h_2$) and $test_2$ = t.test($h_3$, $h_4$).
   If p-value($test_1$) $<$ 0.05, p-value($test_2$) $<$ 0.05 and sign($slope_1$) = sign($slope_2$):
      $interaction_i \rightarrow$ additive interaction
   If p-value($test_1$) $<$ 0.05, p-value($test_2$) $>$ 0.05 and sign($slope_1$) = sign($slope_2$):
      $interaction_i \rightarrow$ non-additive interaction
   If p-value($test_1$) $>$ 0.05, p-value($test_2$) $<$ 0.05 and sign($slope_1$) = sign($slope_2$):
      $interaction_i \rightarrow$ non-additive interaction
   If p-value($test_1$) $<$ 0.05, p-value($test_2$) $<$ 0.05 and sign($slope_1$) $\neq$ sign($slope_2$):
      $interaction_i \rightarrow$ non-additive interaction
```

SNP 1	**SNP 2**	**p-value**
chr5_s3190675 (*FLC*)	chr2_s8854032 (*GLK1*)	0.007
chr1_s19043372	chr5_s3192727 (*FLC*)	0.007
chr2_s13022480	chr5_s3192727 (*FLC*)	0.009
chr5_s18590741 (*DOG1*)	chr5_s3397038	0.015
chr2_s13024062	chr5_s3192727 (*FLC*)	0.016
chr4_s18590741 (*DOG1*)	chr4_s1224865 (*GA1*)	0.02
chr1_s24339560 (*FT*)	chr2_s9584590 (*SVP*)	0.022
chr1_s25638206	chr5_s18590741 (*DOG1*)	0.024
chr5_s3192727 (*FLC*)	chr5_s3875146 (*MIR156E*)	0.024
chr1_s25641251	chr5_s18590741 (*DOG1*)	0.026
chr2_s9584590 (*SVP*)	chr4_s172773	0.029
chr5_s18590741 (*DOG1*)	chr5_s3172910 (*FLC*)	0.031
chr4_s5252410	chr5_s4334594	0.031
chr5_s17415554	chr5_s3173596 (*FLC*)	0.032
chr5_s18590741 (*DOG1*)	chr1_s19081255 (*ASH2R*)	0.034
chr2_s11770118 (*ATC*)	chr5_s3192727 (*FLC*)	0.034
chr2_s13022480	chr5_s3197574 (*FLC*)	0.034
chr4_s5252410	chr5_s4334594	0.034
chr4_s470748	chr5_s3192727 (*FLC*)	0.035
chr1_s5194758	chr5_s18590741 (*DOG1*)	0.036
chr2_s1094522	chr4_s435252	0.043
chr2_s9584590 (*SVP*)	chr4_s682738	0.044
chr4_s1224865 (*GA1*)	chr5_s17415554	0.045
chr2_s13024062	chr5_s3197574 (*FLC*)	0.046
chr1_s25638206	chr5_s4334594	0.048

Table A.1: SNP-SNP interactions that have been identified in pairwise GWAS analyses for flowering time at 16°C, cross-referenced with the FLOR-ID database using a 20 kb window [9]. Multiple SNPs could not be related with any known flowering time gene.

gene 1	**gene 2**	**p-value**
AT2G45660 (*SOC1*)	*AT3G47050*	0.001
AT2G45660 (*SOC1*)	*AT3G51220*	0.001
AT2G45660 (*SOC1*)	*AT5G22860*	0.001
AT2G45660 (*SOC1*)	*AT5G676000* (*WINDHOSE 1*)	0.001
AT2G45660 (*SOC1*)	*AT4G01800* (*ALBINO OR GLASSY YELLOW 1*)	0.006
AT2G45660 (*SOC1*)	*AT5G10140* (*FLC*)	0.008
AT2G45660 (*SOC1*)	*AT5G63120* (*RH30*)	0.022
AT2G45660 (*SOC1*)	*AT5G60910* (*FUL*)	0.036
AT1G09530 (*PIF3*)	*AT5G01980* (*BCA2A ZINC FINGER ATL 16* [*BTL16*])	0.434
AT3G51220	*AT5G01980* (*BTL16*)	0.27
AT2G45660 (*SOC1*)	*AT1G24260* (*SEP3*)	0.049

Table A.2: Gene-gene interactions involving novel and known genes that have been identified in pairwise TWAS analyses for flowering time at 16°C.

variable 1	**variable 2**	**p-value**
AT2G45660 (*SOC1*)	*AT5G60910* (*FUL*)	0.001
AT2G45660 (*SOC1*)	*AT5G63120* (*RH30*)	0.001
AT2G45660 (*SOC1*)	*AT2G32870*	0.001
AT2G45660 (*SOC1*)	chr1_s25638206	0.001
AT2G45660 (*SOC1*)	chr1_s27905306	0.001
AT2G45660 (*SOC1*)	chr2_s17892882 (*PIF4*)	0.001
AT2G45660 (*SOC1*)	chr2_s6917250	0.001
AT2G45660 (*SOC1*)	chr4_s10489410	0.001
AT2G45660 (*SOC1*)	chr4_s5464442	0.001
AT2G45660 (*SOC1*)	chr5_s18607960 (*DOG1*)	0.001
AT2G45660 (*SOC1*)	chr5_s18761369	0.001
AT2G45660 (*SOC1*)	chr5_s3067861	0.001
AT2G45660 (*SOC1*)	chr5_s3188552 (*FLC*)	0.001
AT2G45660 (*SOC1*)	chr5_s3220602	0.001
AT2G45660 (*SOC1*)	chr5_s6388228	0.001
AT2G45660 (*SOC1*)	*AT5G10140* (*FLC*)	0.002
AT5G10140 (*FLC*)	*AT5G60910* (*FUL*)	0.002
AT5G10140 (*FLC*)	chr2_s17277841	0.002
AT5G10140 (*FLC*)	chr5_s18761369	0.002
AT5G10140 (*FLC*)	chr5_s6388228	0.002
AT5G10140 (*FLC*)	chr5_s3220602	0.003
AT5G10140 (*FLC*)	chr4_s15567763	0.004
AT5G10140 (*FLC*)	chr1_s27905306	0.009
AT5G10140 (*FLC*)	chr2_s6917250	0.017
AT5G10140 (*FLC*)	chr5_s18606182 (*DOG1*)	0.017
AT5G10140 (*FLC*)	*AT5G62165* (*AGL42*)	0.023
AT1G65480 (*FT*)	chr1_s27905306	0.025
AT1G65480 (*FT*)	chr5_s18918834	0.029
AT2G20440	chr5_s3188552 (*FLC*)	0.048

Table A.3: Interactions that have been retained in bivariate GTWAS analyses for flowering time at 16°C. The SNPs have been cross-referenced with the FLOR-ID database using a 20 kb window [9]. Many SNPs could not be related with any known flowering time gene.

Appendix B

Proofs

Proof. Equation (2.8).

$$
\begin{aligned}
\mathcal{L}(y_i, \hat{y_i}) &= \sum_{i=1}^{n}(y_i - \hat{y_i})^2 \\
&= \sum_{i=1}^{n}\left(y_i - \sum_{k=1}^{K} c_k \mathbb{1}(x_i \in R_k)\right)^2 \\
&= \sum_{i=1}^{n}\left(y_i^2 - 2y_i \sum_{k=1}^{K} c_k \mathbb{1}(x_i \in R_k) + \left(\sum_{k=1}^{K} c_k \mathbb{1}(x_i \in R_k)\right)^2\right) \\
\frac{\partial \mathcal{L}}{\partial c_k} &= \sum_{i=1}^{n}\left(-2y_i \mathbb{1}(x_i \in R_k) + 2\Big(\sum_{k=1}^{K} c_k \mathbb{1}(x_i \in R_k)\Big)\mathbb{1}(x_i \in R_k)\right) \\
&= \sum_{i:x_i \in R_k} -2y_i + 2\sum_{i:x_i \in R_k} c_k \\
&= \sum_{i:x_i \in R_k} -2y_i + 2n_k c_k \\
\frac{\partial \mathcal{L}}{\partial c_k} &= 0 \iff \hat{c_k} = \frac{1}{n_k}\sum_{i:x_i \in R_k} y_i
\end{aligned}
$$

□

Proof. Equation (2.12).

$$
\begin{aligned}
Var\Big(\frac{1}{B}\sum_{b=1}^{B} T_b\Big) &= \frac{1}{B^2}\Big(\sum_{b=1}^{B} Var(T_b) + \sum_{i \neq j} Cov(T_i, T_j)\Big) \\
&= \frac{1}{B^2}\Big(B\sigma^2 + \sum_{i=1}^{B}\sum_{j=1, j\neq i}^{B} \rho\sqrt{Var(T_i)Var(T_j)}\Big) \\
&= \frac{1}{B^2}\Big(B\sigma^2 + B(B-1)\rho\sigma^2\Big) \\
&= \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2
\end{aligned}
$$

□

Bibliography

[1] M. Hagmann. “A comparison of Bayesian Model Selection Methods for the Analysis of Genome Wide Association Studies”. MA thesis. 2016.

[2] J. Hermisson. “Lecture notes in Mathematical Population Genetics” (2018).

[3] B. Ding. “Power Analysis of Transcriptome-Wide Association Studies (TWAS)”. MA thesis. Science, 2020.

[4] J. Guo. *Transcription: the epicenter of gene expression.* 2014.

[5] C. Cao et al. “Power analysis of transcriptome-wide association study: Implications for practical protocol choice”. *PLoS genetics* 17.2 (2021), e1009405.

[6] M. Somssich. *A short history of Arabidopsis thaliana (L.) Heynh. Columbia-0.* Tech. rep. PeerJ Preprints, 2019.

[7] M. Nordborg et al. “The extent of linkage disequilibrium in Arabidopsis thaliana”. *Nature genetics* 30.2 (2002), pp. 190–193.

[8] W. A. Lopez-Arboleda et al. “Global genetic heterogeneity in adaptive traits”. *bioRxiv* (2021).

[9] F. Bouché et al. “FLOR-ID: an interactive database of flowering-time gene networks in Arabidopsis thaliana”. *Nucleic Acids Research* 44.D1 (2016), pp. D1167–D1171.

[10] M. Nordborg et al. “The pattern of polymorphism in Arabidopsis thaliana”. *PLoS biology* 3.7 (2005), e196.

[11] G. McVean. “Population structure” (2001).

[12] A. Korte and A. Farlow. “The advantages and limitations of trait analysis with GWAS: a review”. *Plant methods* 9.1 (2013), pp. 1–9.

[13] S. Atwell et al. “Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines”. *Nature* 465.7298 (2010), pp. 627–631.

[14] D. Golan, S. Rosset, and D.-Y. Lin. “Mixed models for case-control genome-wide association studies: major challenges and partial solutions”. *Handbook of Statistical Methods for Case-Control Studies.* Chapman and Hall/CRC, 2018, pp. 495–514.

[15] Z. Chen and W. Zhang. “Integrative analysis using module-guided random forests reveals correlated genetic factors related to mouse weight”. *PLoS computational biology* 9.3 (2013), e1002956.

[16] M. Joiret et al. "Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies". *BioData mining* 12.1 (2019), pp. 1–23.

[17] A. C. Nica and E. T. Dermitzakis. "Expression quantitative trait loci: present and future". *Philosophical Transactions of the Royal Society B: Biological Sciences* 368.1620 (2013), p. 20120362.

[18] D. L. Nicolae et al. "Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS". *PLoS genetics* 6.4 (2010), e1000888.

[19] *1001genomes*. URL: `www.1001genomes.org`.

[20] C. Alonso-Blanco et al. "1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana". *Cell* 166.2 (2016), pp. 481–491.

[21] D. Li, Q. Liu, and P. S. Schnable. "TWAS results are complementary to and less affected by linkage disequilibrium than GWAS". *Plant Physiology* (2021).

[22] W. Bateson. "Mendel's Principles of Heredity: Cambridge University Press". *März 1909; 2nd Impr* 3 (1909), p. 1913.

[23] B. A. McKinney et al. "Machine learning for detecting gene-gene interactions". *Applied bioinformatics* 5.2 (2006), pp. 77–88.

[24] R. A. Fisher et al. "009: The Correlation Between Relatives on the Supposition of Mendelian Inheritance." (1918).

[25] C. L. Schmalohr et al. "Detection of epistatic interactions with Random Forest". *bioRxiv* (2018), p. 353193.

[26] H. J. Cordell. "Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans". *Human molecular genetics* 11.20 (2002), pp. 2463–2468.

[27] R. Hornung and A.-L. Boulesteix. "Interaction Forests: Identifying and exploiting interpretable quantitative and qualitative interaction effects" (2021).

[28] P. C. Phillips. "Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems". *Nature Reviews Genetics* 9.11 (2008), pp. 855–867.

[29] T. A. Manolio et al. "Finding the missing heritability of complex diseases". *Nature* 461.7265 (2009), pp. 747–753.

[30] J.-E. Dazard et al. "Ensemble survival tree models to reveal pairwise interactions of variables with time-to-events outcomes in low-dimensional setting". *Statistical applications in genetics and molecular biology* 17.1 (2018).

[31] I. M. Ehrenreich, P. A. Stafford, and M. D. Purugganan. "The genetic architecture of shoot branching in Arabidopsis thaliana: a comparative assessment of candidate gene associations vs. quantitative trait locus mapping". *Genetics* 176.2 (2007), pp. 1223–1236.

[32] M. Yoshida and A. Koike. “SNPInterForest: a new method for detecting epistatic interactions”. *BMC bioinformatics* 12.1 (2011), pp. 1–10.

[33] J. Stephan, O. Stegle, and A. Beyer. “A random forest approach to capture genetic effects in the presence of population structure”. *Nature communications* 6.1 (2015), pp. 1–10.

[34] B. J. Vilhjálmsson and M. Nordborg. “The nature of confounding in genome-wide association studies”. *Nature Reviews Genetics* 14.1 (2013), pp. 1–2.

[35] F. Zhang et al. “OSCA: a tool for omic-data-based complex trait analysis”. *Genome biology* 20.1 (2019), pp. 1–13.

[36] H. D. Patterson and R. Thompson. “Recovery of inter-block information when block sizes are unequal”. *Biometrika* 58.3 (1971), pp. 545–554.

[37] H. IJ. “Statistics versus machine learning”. *Nature methods* 15.4 (2018), p. 233.

[38] C. Molnar, G. Casalicchio, and B. Bischl. “Interpretable machine learning–a brief history, state-of-the-art and challenges”. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* Springer. 2020, pp. 417–431.

[39] V. Botta et al. “Exploiting SNP correlations within random forest for genome-wide association studies”. *PloS one* 9.4 (2014), e93379.

[40] L. Breiman. “Random forests”. *Machine learning* 45.1 (2001), pp. 5–32.

[41] G. James et al. *An introduction to statistical learning.* Vol. 112. Springer, 2013.

[42] J. Friedman, T. Hastie, R. Tibshirani, et al. *The elements of statistical learning.* Vol. 1. 10. Springer series in statistics New York, 2001.

[43] M. Hatz. “Der Einfluss von mtry auf Random Forests”. PhD thesis. 2018.

[44] J. Ehrlinger. “ggRandomForests: Exploring random forest survival”. *arXiv preprint arXiv:1612.08974* (2016).

[45] L. Breiman et al. “Classification and Regression Trees (Wadsworth and Brooks: London)”. *Cité en* (1984), p. 109.

[46] H. Ishwaran. “Variable importance in binary regression trees and forests”. *Electronic Journal of Statistics* 1 (2007), pp. 519–537.

[47] K. L. Lunetta et al. “Screening large-scale association study data: exploiting interactions using random forests”. *BMC genetics* 5.1 (2004), pp. 1–13.

[48] A. Bureau et al. “Identifying SNPs predictive of phenotype using random forests”. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 28.2 (2005), pp. 171–182.

[49] C. Strobl et al. “Bias in random forest variable importance measures”. *Workshop on Statistical Modelling of Complex Systems.* Citeseer. 2006.

[50] H. Ishwaran et al. “Random survival forests for high-dimensional data”. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 4.1 (2011), pp. 115–132.

[51] M. N. Wright, A. Ziegler, and I. R. König. “Do little interactions get lost in dark random forests?” *BMC bioinformatics* 17.1 (2016), pp. 1–10.

[52] M. S. Brieuc et al. “A practical introduction to Random Forest for genetic association studies in ecology and evolution”. *Molecular ecology resources* 18.4 (2018), pp. 755–766.

[53] A. Altmann et al. “Permutation importance: a corrected feature importance measure”. *Bioinformatics* 26.10 (2010), pp. 1340–1347.

[54] Y. Benjamini and Y. Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[55] J. H. Friedman. “Greedy function approximation: a gradient boosting machine”. *Annals of statistics* (2001), pp. 1189–1232.

[56] A. Goldstein et al. “Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation”. *journal of Computational and Graphical Statistics* 24.1 (2015), pp. 44–65.

[57] C. Molnar. *Interpretable Machine learning.* URL: `https://christophm.github.io/interpretable-ml-book/`.

[58] H. J. Cordell. “Detecting gene–gene interactions that underlie human diseases”. *Nature Reviews Genetics* 10.6 (2009), pp. 392–404.

[59] S. D. Turner et al. “Knowledge-driven multi-locus analysis reveals gene-gene interactions influencing HDL cholesterol level in two independent EMR-linked biobanks”. *PloS one* 6.5 (2011), e19586.

[60] S. Oh et al. “A novel method to identify high order gene-gene interactions in genome-wide association studies: gene-based MDR”. *BMC bioinformatics.* Vol. 13. 9. BioMed Central. 2012, pp. 1–9.

[61] D. Brassat et al. “Multifactor dimensionality reduction reveals gene–gene interactions associated with multiple sclerosis susceptibility in African Americans”. *Genes & Immunity* 7.4 (2006), pp. 310–315.

[62] S. J. Winham et al. “SNP interaction detection with random forests in high-dimensional genetic data”. *BMC bioinformatics* 13.1 (2012), pp. 1–13.

[63] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria, 2017.

[64] Ü. Seren et al. “AraPheno: a public database for Arabidopsis thaliana phenotypes”. *Nucleic Acids Research* (2016), gkw986.

[65] 1001 Genomes Consortium. “1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana”. *Cell* 166.2 (2016), pp. 481–491.

[66] B. L. Browning and S. R. Browning. “Genotype imputation with millions of reference samples”. *The American Journal of Human Genetics* 98.1 (2016), pp. 116–126.

[67] S. Purcell et al. “PLINK: a tool set for whole-genome association and population-based linkage analyses”. *The American journal of human genetics* 81.3 (2007), pp. 559–575.

[68] X. Chen and H. Ishwaran. “Random forests for genomic data analysis”. *Genomics* 99.6 (2012), pp. 323–329.

[69] C. Strobl et al. “Conditional variable importance for random forests”. *BMC bioinformatics* 9.1 (2008), pp. 1–11.

[70] B. A. Goldstein et al. “An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings”. *BMC genetics* 11.1 (2010), pp. 1–13.

[71] A. L. Price et al. “Principal components analysis corrects for stratification in genome-wide association studies”. *Nature genetics* 38.8 (2006), pp. 904–909.

[72] J. A. Holliday, T. Wang, and S. Aitken. “Predicting adaptive phenotypes from multilocus genotypes in Sitka spruce (Picea sitchensis) using random forest”. *G3: Genes| genomes| genetics* 2.9 (2012), pp. 1085–1093.

[73] Y. Zhao et al. “Correction for population stratification in random forest analysis”. *International journal of epidemiology* 41.6 (2012), pp. 1798–1806.

[74] B. L. Welch. “The generalization of ‘STUDENT’S’problem when several different population varlances are involved”. *Biometrika* 34.1-2 (1947), pp. 28–35.

[75] C. Lippert et al. “LIMIX: genetic analysis of multiple traits”. *BioRxiv* (2014).

[76] J. H. Lee et al. “Role of SVP in the control of flowering time by ambient temperature in Arabidopsis”. *Genes & development* 21.4 (2007), pp. 397–402.

[77] J. Lee and I. Lee. “Regulation and function of SOC1, a flowering pathway integrator”. *Journal of experimental botany* 61.9 (2010), pp. 2247–2254.

[78] H. Tian et al. “Photoperiod-responsive changes in chromatin accessibility in phloem companion and epidermis cells of Arabidopsis leaves”. *The Plant Cell* 33.3 (2021), pp. 475–491.

[79] E. Sasaki, F. Frommlet, and M. Nordborg. “GWAS with heterogeneous data: estimating the fraction of phenotypic variation mediated by gene expression data”. *G3: Genes, Genomes, Genetics* 8.9 (2018), pp. 3059–3068.

[80] R. G. Immink et al. “Characterization of SOC1’s central role in flowering by the identification of its upstream and downstream regulators”. *Plant physiology* 160.1 (2012), pp. 433–449.

[81] S. R. Hepworth et al. “Antagonistic regulation of flowering-time gene SOC1 by CONSTANS and FLC via separate promoter motifs”. *The EMBO journal* 21.16 (2002), pp. 4327–4337.

[82] S. Zhang et al. "GASA5, a regulator of flowering time and stem growth in Arabidopsis thaliana". *Plant molecular biology* 69.6 (2009), pp. 745–759.

[83] C. Dorca-Fornell et al. "The Arabidopsis SOC1-like genes AGL42, AGL71 and AGL72 promote flowering in the shoot apical and axillary meristems". *The Plant Journal* 67.6 (2011), pp. 1006–1017.

[84] C. Strobl et al. "Bias in random forest variable importance measures: Illustrations, sources and a solution". *BMC bioinformatics* 8.1 (2007), pp. 1–21.