GLOBAL GENETIC HETEROGENEITY IN ADAPTIVE TRAITS

WILLIAM ANDRÉS LÓPEZ ARBOLEDA

Dissertation zur Erlangung des
naturwissenschaftlichen Doktorgrades
der Julius-Maximilians-Universität Würzburg
vorgelegt von

William Andrés López Arboleda
aus
Versalles

Würzburg, 2021

Eingereicht am: . . . .

Mitglieder der Promotionskommission:

Vorsitzender: . . . .

Gutachter: . . . .

Gutachter: . . . .

Tag des Promotionskolloquiums: . . . .

Doktorurkunde ausgehändigt am: . . . .

EIDESSTATTLICHE ERKLÄRUNGEN NACH §7 ABS. 2 SATZ 3, 4, 5 DER PROMOTIONSORDNUNG DER FAKULTÄT FÜR BIOLOGIE

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, die Dissertation: „Globale genetische Heterogenität in adaptiven Merkmalen", eigenständig, d. h. insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen, als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben. Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Weiterhin erkläre ich, dass bei allen Abbildungen und Texten bei denen die Verwer-tungsrechte (Copyright) nicht bei mir liegen, diese von den Rechtsinhabern eingeholt wurden und die Textstellen bzw. Abbildungen entsprechend den rechtlichen Vorgaben gekennzeichnet sind sowie bei Abbbildungen, die dem Internet entnommen wurden, der entsprechende Hypertextlink angegeben wurde.

Affidavit

I hereby declare that my thesis entitled: „Global genetic heterogeneity in adaptive traits" is the result of my own work. I did not receive any help or support from commercial consultants. All sources and / or materials applied are listed and specified in the thesis. Furthermore I verify that the thesis has not been submitted as part of another examination process neither in identical nor in similar form.

Besides I declare that if I do not hold the copyright for figures and paragraphs, I obtained it from the rights holder and that paragraphs and figures have been marked according to law or for figures taken from the internet the hyperlink has been added accordingly.

Würzburg, 2021

William Andrés López Arboleda

ABSTRACT

Genome Wide Association Studies (GWAS) have revolutionized the way on how genotype-phenotype relations are assessed. In the 20 years long history of GWAS, multiple challenges from a biological, computational, and statistical point of view have been faced. The implementation of this technique using the model plant species *Arabidopsis thaliana*, has enabled the detection of many association for multiple traits. Despite a lot of studies implementing GWAS have discovered new candidate genes for multiple traits, different samples are used across studies. In many cases, either globally diverse samples or samples composed of accessions from a geographically restricted area are used. With the aim of comparing GWAS outcomes between populations from different geographic areas, this thesis describes the performance of GWAS in different European samples of *A. thaliana*. Here, association mapping results for flowering time were compared. Chapter 2 describes the analyses of random resampling from this original sample. The aim was to establish reduced subsamples to later carry out GWAS and compare the outcomes between these subsamples. In Chapter 3, the European sample was split into eight equally-sized local samples representing different geographic regions. Next, GWAS was carried out and an attempt was made to clarify the differences in GWAS outcomes. Chapter 4 contains the results of a collaboration with Prof. Dr. Wolfgang Dröge-Laser, in which my mainly task was the analysis of RNAseq data from *A. thaliana* plants infected by pathogenic fungi. Finally, Appendix A presents a very short description of my participation in the GHP Project on Access to Care for Cardiometabolic Diseases (HPACC) at the university of Heidelberg.

ZUSAMMENFASSUNG

Die genomweiten Assoziationsstudien (GWAS) haben die Art und Weise revolutionierten, wie genotypische-phänotypische Zusammenhänge untersucht werden. In der 20-jährigen Geschichte dieser Analysen, gab es zahlreiche biologische, mathematische und statistische Herausforderungen. Die Anwendung dieser Methodik in der Modellpflanze *Arabidopsis thaliana* ermöglichte die Erkennung neuer Zusammenhänge für zahlreicher Merkmale. Obwohl viele Studien, die GWAS implementieren, neue Kandidatengene für verschiedene Merkmale entdeckt haben, werden in den verschiedenen Analysen oft unterschiedliche Populationen verwendet. Es werden entweder global unterschiedliche Accessionen oder alternative welche aus einem geografisch begrenzten Gebiet als Population für die Anaylsen verwendet. Mit dem Ziel, GWAS-Ergebnisse zwischen Populationen aus verschiedenen geografischen Gebieten zu vergleichen, beschreibt diese Arbeit die Eigenschaften der Analyse in verschiedenen europäischen Populationen von *A. thaliana*. Verglichen wurden die Ergebnisse der Assoziationskartierung für die Blütezeit. Kapitel 2 beschreibt die Analysen von zufälligen Populationen im Vergleich zur gesamten europäischen Population. Ziel war es, reduzierte Stichproben zu erstellen, um später GWAS durchzuführen und die Ergebnisse zwischen diesen Stichproben zu vergleichen. In Kapitel 3 wurde die europäische Population in acht gleich große lokale Subpopulationen aufgeteilt. Diese repräsentieren verschiedene geografische Regionen. Als nächstes wurde GWAS durchgeführt und die Unterschiede in den jeweilgen GWAS-Ergebnissen beschrieben. Kapitel 4 behinhaltet die Ergebnisse aus einer Zusammenarbeit mit Prof. Dr. Wolfgang Dröge-Laser: Hier war meine Hauptaufgabe die Analyse von RNAs Sequenzierungsdaten von mit pathogenen Pilzen befallenen *A. thaliana*-Pflanzen. Schließlich enthält Anhang A eine zusammenfassende Beschreibung meiner Mitarbeit am GHP-Projekt zum Zugang zur Versorgung bei kardiometabolischen Erkrankungen (HPACC) an der Universität Heidelberg.

PUBLICATIONS

Lopez-Arboleda, William Andres, Stephan Reinert, Magnus Nordborg, and Arthur Korte (2021). “Global genetic heterogeneity in adaptive traits.” In: *Molecular Biology and Evolution*. msab208. ISSN: 0737-4038. DOI: 10.1093/molbev/msab208. eprint: https://academic.oup.com/mbe/advance-article-pdf/doi/10.1093/molbev/msab208/38886451/msab208.pdf. URL: https://doi.org/10.1093/molbev/msab208

Fröschel, Christian, Jaqueline Komorek, Agnès Attard, Alexander Marsell, **Lopez-Arboleda, William.A**, Joëlle Le Berre, et al. (2021). “Plant roots employ cell-layer-specific programs to respond to pathogenic and beneficial microbes.” In: *Cell Host & Microbe* 29.2, pp. 299–310. DOI: https://doi.org/10.1016/j.chom.2020.11.014

Teufel, Felix, Jacqueline A.Seiglie, Michaela Theilmann, Maja-Emilia Marcus, Cara Ebert, **William.A Lopez-Arboleda**, et al. (2021). “Body-mass index and diabetes risk in 57 low-income and middle-income countries: a cross-sectional study of nationally representative, individual-level data in 685616 adults.” In: *The lancet* 398.10296, pp. 238–248. DOI: https://doi.org/10.1016/S0140-6736(21)00844-8

Science knows no country,
because knowledge belongs to humanity,
and is the torch which illuminates the world.

— Louis Pasteur

ACKNOWLEDGEMENTS

I am one of those who thing that feelings can not be expressed in a foreign language. For that reason, I will first thank the people who matter most to me in my mother tongue. Empezaré expresando mi más profundo agradecimiento a mis padres María Yolanda Arboleda Trujillo y William López Toro. Las dos persona que hicieron posible que un niño lleno de sueños nacido en lo más profundo de las montañas colombianas se formase como investigador. Agradezco a mi esposa Bettina María Reinholz por haber transformado mi vida y haberme dado el regalo más maravilloso, nuestra pequeña Rosa. A ti Rosa, te agradezco por tus sonrisas que llenan día a día de alegría mi corazon, por tus profundos ojos negros que cuando me miran me hacen olvidar la noción del tiempo. A mi hermana Beatriz Helena López Arboleda por las largas discusiones existenciales. Agradezco a toma mi familia, encabezada por mis tías maternas que siempre han creido ciegamente en mi. A mis suegros, por haberme acogido como si de un hijo se tratase. A Juan Pablo por su amistad y gran ayuda. I thank Arthur korte for giving me the opportunity to pursue my PhD. Thanks to Pascal Geldsetzer for hiring me as research assistant. I thank Michaela Theilmann for her guidance during my participation in the GHP Project on Access to Care for Cardiometabolic Diseases (HPACC). I thank my close friend Torsten, for his honesty and invaluable support. Thanks to COLCIENCIAS and DAAD for partially funding my PhD.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ALDH10A8	*ALDEHYDE DEHYDROGENASE*
AGs	Aliphatic glucosinaltes
AGL-20	*AGAMOUS-LIKE 20*
bp	base pairs
CDF5	*CYCLING DOF FACTOR 5*
CIR1	CIRCADIAN 1
CL	Cauline Leaf number
COL5	CONSTANS-LIKE 5
DHS	The Demographic and Health Surveys Program
DOG1	*DELAY OF GERMINATION*
eGWAS	expression Genome-Wide Asociation Study
EMB2739	*EMBRYO DEFECTIVE 2739*
EMB3127	EMBRYO DEFECTIVE 3127
EMMA	Efficient Mixed-Model Association
EMMAX	Efficient Mixed-Model Association eXpedited
FDR	False Discovery Rate
FLC	*FLOWERING LOCUS C*
FP	False Positive
FRI	*FRIGIDA*
FT	*FLOWERING LOCUS T*
GEMMA	Genome-Wide Efficient Mixed Model Analysis
GWAS	Genome-Wide Association Study

G6PD4	*NADP-dependent glucose-6-phosphato dehydrogenase*
HDA9	*HISTONE DEACETYLASE 9*
HPACC	GHP Project on Access to Care for Cardiometabolic Diseases
IGs	tryptophan-derived IndoleGlucosinolates
ICN	Indole-3-Carbonyl Nitrile
JMJ14	*JUMONJI 14*
KEGG	Kyoto Encyclopedia of Genes and Genomes
LIF2	*LHP1-INTERACTING FACTOR 2*
LMM	Linear Mixed Model
MAF	Minor Allele Frequency
MED12	*MEDIATOR 12*
MSI	*MULTICOPY SUPRESSOR OF IRA1*
NCD	Non-Communicable Diseases
NIP	Northern Iberian Peninsula
RCAR5	*REGULATORY COMPONENT OF ABA RECEPTOR 5*
RNA	RiboNucleic Acid
RPS5	*RESISTANT TO P. SYRINGAE 5*
SDH3-2	*succinate dehydrogenase*
SIP	Southern Iberian Peninsula
SMZ	*SCHLAFMÜTZE*
SNP	Single Nucleotide Polymorphism
ST	Stomata Size
STEPS	STEPwise approach to Surveillance
ST4B	*BRASSINOLIDE SULFOTRANSFERASE*

TRAP-seq	Translating Ribosome Affinity Purification followed by RNA sequencing
TP	True Positive
TPM	Transcripts Per Million
TSF	*TARGET OF FLC AND SVP1*
VIN3	*VERNALIZATION INSENSITIVE 3*
WHO	World Health Organization
ZTL	*ZEITLUPE*

1

THE RELATIONSHIP BETWEEN GENOTYPE AND PHENOTYPE

1.1 VARIABILITY IN BIOLOGY

Variability is an intrinsic property of all living organisms. During most of human history, species were seen as static and immutable groups of individuals. However, humankind has been taking advantages from this intrinsic variability to improve important traits in agronomy and animal breeding (Hickey et al., 2017). Multiple hypotheses trying to explain the origin and maintenance of this variability have been proposed, but it was not until the publication of "On the origin of species" that a single unifying theory could reasonably explain that this variability itself is responsible for the formation of new species (Darwin, 1859). Back then, this statement contradicted the notion of immutability of species, and deconstructed the idea of species as discrete entities, or even more that this interpretation of discrete entities is the result of the temporal scale in which they are being studied. This idea of slow, gradual, and continuous changes being under natural selection makes the experimental proof of species formation challenging, since a geological scale must be considered. Fortunately, fossil register has been supporting the theory of evolution (Prothero, 2007). Even though Darwin was not aware of genetics, it was clear for him that an inner force in all individuals was responsible for the modifications to be put under natural selection. Nowadays, 150 years later, we not only know where this variation come from, but also how it is inherited.

Although it is meanwhile known how this variability arises, after the publication of "On the origin of species", an extended explanation about this matter was demanded and at that time far for being elucidated. Initially, most of the first attempts were restricted to a philosophical discussion. However, multiple authors agreed that biology should move to more mechanistic explanations and that the main feature of biology is change. So, it was recognized that the only way biology might progress is through the clarification on how variability arises and fluctuates over time. In fact, some authors went even further : "...Meanwhile, we may safely predict that the biology of the immediate future will be the science of variation"[1]. Nowadays, we can entirely agree with this statement, as biology is still and will remain the science of variation.

Discussions about Variation and heredity held the attention of many researchers and triggered one of the most famous debates, namely, between Mendelians and Biometricians, and more precisely on a personal level between William Bateson and Karl Pearson (Farrall, 1975; Gillham, 2015; Morrison, 2002). Fortunately, R.A Fisher took advantage of both opposing points of view to reconcile biometry and mendelism (Fisher, 1918). This effort ratified the crucial role mathematics has in the integration of evolution and genetics (Cohen, 2004). At that point, genetics was in its infancy and was often seen as a discipline dealing with aberrant characters. However, among geneticists it was clear that genetics had to face critical challenges in order to survive (Punnett et al., 1950). Thomas Hunt Morgan stated for one of those challenges, which is still of high priority: "The relation of genes to characters. This is the explicit realization of the implicit power of the genes, and includes the physiological action of the gene on the rest of the cell. This is the gap in our knowledge to which I have referred already at some length"[2]. Currently, many of these "initial challenges" are still being addressed, nevertheless progress has been made due to technological developments. Some of these scientific advancements enables the precise mapping of genetic variability in populations using molecular markers. To summarize, we are still addressing some of the initial question of genetics but using different approaches due to the knowledge and technologies accumulated throughout the last century.

1.2 WHERE DOES THIS VARIABILITY COME FROM? GENOTYPE-PHENOTYPE RELATIONSHIP

At this point, I referred to variability as the spectrum of observable variation between individuals. This can be in fact documented as phenotypic variability. How this phenotypic variability arises has remained a major question in biology in the last century. The relationship and "opposite" characteristic of genotype and phenotype was first proposed by Wilhelm Johannsen back in 1909 (Johannsen, 1909, 1911). Initially, this distinction was not accepted in the new field of genetics,

[1] Fothergill, WE (1888). “The Biology of the Future.” In: *The Hospital* 3.68, p. 274

[2] Morgan, Thomas H (1932). “The rise of genetics.” In: *Science* 76.1969, pp. 261–267

however its importance was recognized later (Churchill, 1974) and since then a lot of effort has been made to elucidate this intricate relationship. Back then, it was not possible to characterize the genetic part of the genotype-phenotype relationship and it was not until the emergence of molecular biology that further progress in this matter was made. Human genetic studies using molecular markers as Restriction Fragment Length Polymorphisms (RFLP's) and microsatellites were able to link genetic variation in certain regions of the genome to specific diseases (Botstein et al., 1980; Wooster et al., 1995). However some of these initial reports raised doubts, whether these analyses actually revealed causation or were merely spurious. The fear that these associations might be spurious was mainly grounded in the fact that only short genomic regions were analyzed which can led to the detection of associations due to evolutionary history rather than recombination. That is, these associations would be the result of linkage disequilibrium which refers to the non-randomly assortment of alleles at two or more loci. Consequently, the magnitude of linkage disequilibrium of the associated maker would reflect the temporal position of a mutational event rather than its exact physical location (Templeton, 1998).

Later in the early 2000's, with the emergence of high-throughput sequencing (also known as Next Generation Sequencing-NGS), it was possible to produce an enormous amount of genomic data enabling the genotyping of Single Nucleotide Polymorphisms (SNPs) along the whole genome. This genotype information allows an direct association between phenotypes and genomic markers. However, this approach assumes that the phenotype is regulated only through additive genetic effects, as each marker is tested separately. Multiple studies have reported evidence that phenotypic variability is a result of intricate interactions including gene-environment interaction, gene-gene interaction (also known as epistasis), epigenetic effects and pleiotropy (Chiang et al., 2013; W. Huang et al., 2020; Huo, Wei, and Bradford, 2016; X. Li et al., 2018; Stinchcombe et al., 2004). Although this would contradict the basic ideas of the above described approach, many phenotypes are still governed by major polymorphisms of additive effect. In addition, for some studies it is enough to track down polymorphisms of large effect and the goal is not to explain the complete phenotypic variance. In this sense, Genome-Wide Association Studies (GWAS), an approach for testing association of a marker with a phenotype, have revolutionized the study of genotype-phenotype relationship. Since this was the principal methodology underlying these thesis, I will expand this topic in the next section, certainly with major focus on plant genetics.

1.3 CONNECTING GENOTYPES TO PHENOTYPES: GENOME-WIDE ASSOCIATION STUDIES (GWAS)

The idea behind GWAS was already discussed before the methodology was implementable, back in the middle-90s (E. Lander and Kruglyak, 1995; E. S. Lander and Schork, 1994; Risch and Merikangas, 1996), even though sequencing the genome of multiple individuals was unreachable at that time (the first platforms for high-throughput sequencing were fist available in 2000) (Kulski, 2016). Later, in 2002 the first GWAS paper was published (Ozaki et al., 2002), and since then this methodology has been successfully identifying associations in the field of human (Buniello et al., 2019) and plant genetics (X. Huang and Han, 2014). Before NGS was extended in the field of plant genetics, SNPs were initially genotyped in a population via arrays. This strategy allowed the genotyping of thousands SNPs. For species with a small genome, this methodology made SNP panels suitable for GWAS. More precisely, such SNP panels were used to implement GWAS on different traits of *A. thaliana*, which led to the successful detection of associated markers (Aranzana et al., 2005; Atwell et al., 2010).

At the same time as the implementation of GWAS were steadily growing, concerns about spurious associations began to emerge, primarily due to the inflated results most GWAS were delivering. GWAS test for association between SNPs with a phenotype each marker at a time under the null hypothesis of no effect of SNPs on the phenotype. This results in a high number of tested hypotheses making it necessary to correct for multiple testing. Under the null hypothesis pvalues are uniformly distributed, which means all are equally likely to be found. Assuming the null hypothesis is true, a type I error of 0.05 indicates that in 5% of the times (trials) a pvalue lower than 0.05 would be found just by chance, and so a false association would be detected. To better exemplify the effect of multiple testing, a GWAS run using 1 million SNPs would associate 50000 SNPs just by chance at a nominal $pvalue = 0.05$, because pvalues are normally distributed under the null hypothesis (Hung et al., 1997). The two most widely used methods to account for multiple testing are Bonferroni correction (Bonferroni, 1935) and False Discovery Rate (FDR) (Benjamini and Hochberg, 1995). Bonferroni correction controls for multiple testing by dividing the nominal pvalue by the number of tests to be made. In GWAS this number of tests corresponds to the number of markers for which association is evaluated. This method is considered to be very stringent since it controls for false positives among all tests and assumes that all markers are independent. However, in some occasions it would be useful to accept a small proportion of false positives in order to increase the total number of associations. This proportion of false positives in relation to the total number of associations is called the FDR. After performing GWAS, estimations of effect size and pvalue for each SNP are obtain. Commonly, pvalues for all SNPs are visualized using a manhattan plot. On this plot the SNP position (on the x-axis) is plotted against the negative logarithm (base 10) of the pvalue (on the y-axis) for each tested association. Usually a significant threshold is also plotted to indicated where the significant SNPs are located. Figure 1.1 is a manhattan plot representing GWAS results from a linear mixed model accounting for population structure (an extended explanation about the model will be later provided).

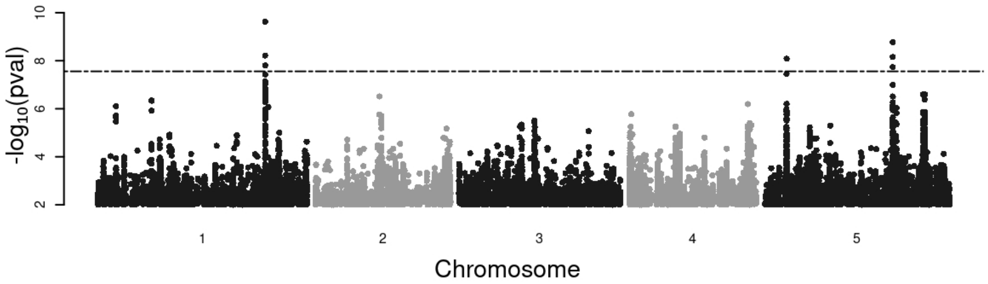


Figure 1.1: Manhattan plot representing GWAS results. Genomic location of SNPs on the x-axis is plotted against the negative logarithm (base 10) of the pvalue for each SNP on the y-axis. Horizontal dash-dotted line indicates the value for the Bonferroni correction

Although Bonferroni correction and FDR are widely used to define significant levels, other strategies can be implemented in order to control for multiple testing including permutation and Bayesian approaches. While Bonferroni correction and FDR are easily obtain, permutation test and Bayesian approaches are computational time consuming. Permutation testing allows to generate an empirical distribution of pvalues under the null hypothesis. To do so, phenotypes are randomly shuffle between individuals to unlink the genotype-phenotype relationship present in the data. This process is repeated N times and then the significance level is define according to the type I error (Bush and Moore, 2012). For the Bayesian approach both the null and the alternative hypothesis are addressed. The Bayes factor represents the ratio of the probability of the data under the null to the probability of the data under the alternative hypothesis. This ratio provides evidence of the data being better predicted by one of the hypotheses and is not affected by the number of test to be considered (Wakefield, 2012). The discussion about how to define a significance level is not only important to limit the type I error, but also to estimate the theoretical power of a GWAS experiment. Statistical power is calculated as $1 - \beta$, where β is the type II error. Type I and Type II errors have an inverse relationship and can never be avoid.

Once type I error is set, type II error can be reduced (and so statistical power increased) by enlarging the sample size (at a minimal detectable effect size). This is a very important issue when designing GWAS, since knowing the minimal sample size required to detect a minimal expected effect size can save time and money. Even knowing that high sample sizes can led to a sufficient statistical power, there are other factors affecting statistical power in GWAS. Allele frequency has a direct effect in statistical power, since the estimation of the effect size depends on the variance at each locus. For that reason estimations of effect size are more precise in common than in low-frequency variants. So, in order to detect low-frequency variants at the same statistical power as common variants, larger samples sizes are needed. GWAS basically test the association between a phenotype and the allele frequency of each marker genotyped in the sample, and this association is estimated using a linear model. In the first years of GWAS, a large number of associations were successfully reported even considering the stringent significance levels established by multiple testing correction. Concerns began to emerge about the validity of these associations and the possibility of detecting many of them due to inflated results. Later the effect of population structure on GWAS was demonstrated and the Q + K linear mixed model was proposed to account for relatedness between individuals (Yu et al., 2006). In fact, it has been shown that accounting for population structure reduces type II error dramatically (M. Wang and Xu, 2019). Figure 1.2 shows the effect on GWAS results when population structure is not taken into account. In this example, GWAS results based on a linear model detected a high number of association, while most of them were not significant when implementing the Q + K linear mixed model. This contrasting results exemplify how population structure can lead to false associations.

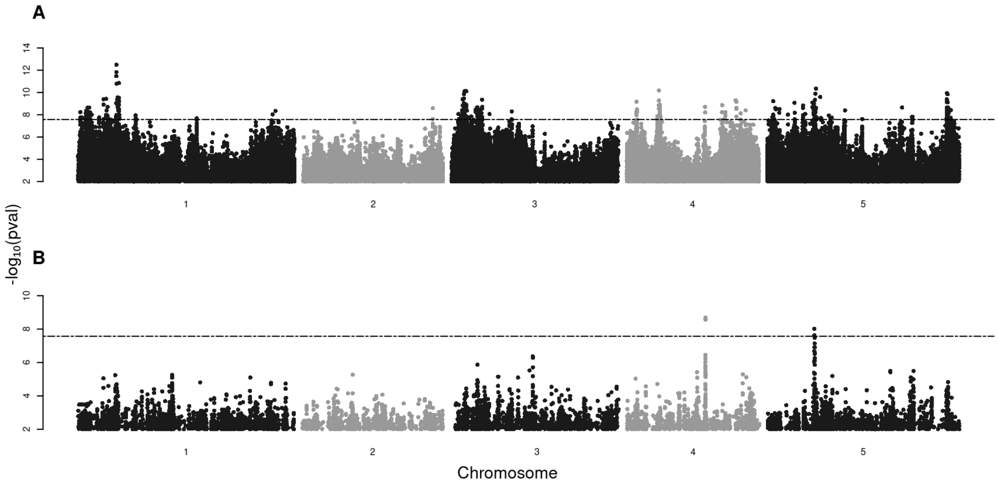


Figure 1.2: Manhattan plots showing the effect of population structure on GWAS results. (A) Manhattan plot displaying GWAS results based on a linear model. (B) Manhattan plot displaying GWAS results based on the Q + K linear mixed model.

This new approach in GWAS resulted in an increase of computational time making it impracticable for many experiments. Thus, the increasing amount of genomic data made it possible to design experiments with thousands of individuals genotyped for millions of SNPs, demanding the develop of new models to speed up the computational time of GWAS. There is an important number of models which facilitate the implementation of GWAS even for huge experiments. Efficient mixed-model association (EMMA) was the first method of this type to be proposed (M. Kang et al., 2008). In the subsequent years an important number of method were proposed, highlighting EMMAX (Efficient Mixed-Model Association eXpedited), a method build on EMMA which dramatically reduces its computational time (H. Kang et al., 2010) and GEMMA (Genome-wide Efficient Mixed Model Analysis) (Zhou and Stephens, 2012).

Soon after, the first GWAS in *A. thaliana* were published. There were an accelerated interest in such studies and the execution of the 1001 genomes project (Alonso-Blanco et al., 2016; Weigel and Mott, 2009) fueled the idea of creating a publicly available source for all GWAS using *A. thaliana*, the AraGWAS catalog (Togninalli et al., 2020). This source contains 462 GWAS (as of May 26th 2021) and is constantly growing. Even more interesting, the availability of this type of data has allowed the re-analysis. In addition, the 1001 genomes project (Alonso-Blanco et al., 2016; Weigel and Mott, 2009) enabled the calling of more than 10 million SNPs, which results in a high coverage with an average occurrence of SNPs nearly every 10-20 base pair (bp). This deep SNP coverage facilitates the detection of responsible makers via GWAS. The reason is the extend of linkage disequilibrium decay, which was firstly estimated to be around 250 kb (Nordborg et al., 2002), but current reports estimate it to be around 10 kb (S. Kim et al., 2007)).

In contrast with this characteristics, which facilitates the implementation of GWAS using *A. thaliana*, many agronomic important species are more challenging given that their genomes are more complex. Examples of this are: wheat (*Triticum aestivum*), canola (*Brassica napus*), potato (*Solanum tuberosum*), just to name some examples (polyploidy is common in plants with a high proportion of the angiosperm species displaying at least some ploidal level in their evolutionary history (Meyers and Levin, 2006)). On the other hand, most of the agronomic important traits are quantitative and therefore display a very complex genetic architecture with a lot of alleles having middle-to small effects (Yang et al., 2010). In fact, after the first years of GWAS implementation, it was clear that only a small proportion of the expected genetic variation in most of the studied organisms was being discovered (such observations were based on intensively validated heritability estimates). So the question arose: where is the missing heritability? (Gibson, 2010; Manolio et al., 2009). This question can be answer through several sources including: rare variants, gene-gene interactions (also known as epistasis), genetic heterogeneity, epigenetic effects (Brachi, Morris, and Borevitz, 2011) or gene-by-environment interactions (Thomas, 2010). GWAS on plant species can be challenging due to sample size. The detection of variants with low effect via GWAS might demand sample sizes not available in plant genetics studies. Despite this disadvantage, GWAS in plants have been able to explain a much greater proportion of the phenotypic variation than in humans. Although human height is a high heritable trait (heritability estimates around 80%), early GWAS found significant variants explaining only around 3.7% of the phenotypic variation (Gudbjartsson et al., 2008). In contrast, GWAS on flowering time have detected variants explaining up to half of the total heritable variation (Yan Li et al., 2010).

Over the past 20 years, GWAS have faced multiple challenges from whether a biological, statistical or computational perspective (Korte and Farlow, 2013; McCarthy et al., 2008; Moore, Asselbergs, and Williams, 2010; Visscher et al., 2017), as well as bitter criticism by researchers from diverse fields (Manolio et al., 2009; McClellan and King, 2010). Under powered GWAS, biased sampling, low heritable traits, among others are all problems which still occur. So, some of the criticism is based on poorly design studies whose conclusion are, of course, *a priori* invalid. On the other hand, it is important to keep in mind what GWAS can actually tell us, in other words that this methodology intends to statistically associate genetic markers (SNPs) with a trait under the assumption that the effect of all marker is independent and additive. But, how could we look into the dark side?

1.4 *Arabidopsis thaliana* as model organism

A. thaliana is probably the best known plant model organism. This species has been playing a decisive role in all fields of plant biology since it was suggested as an ideal model organism (Laibach, 1943). This section is not intended to give an extended description of the characteristics that took *A. thaliana* to its present position (see (Koornneef and D. Meinke, 2010)), but highlights those enabling its use for GWAS and evolutionary studies. Contrary to many angiosperms, *A. thaliana* has a small diploid genome (approx. 135 MB) and occurs predominantly as inbred lines, due to its self-compatibility, across its distribution range (D. W. Meinke et al., 1998). These two characteristics are probably the most important ones when implementing GWAS. As mentioned before, the available SNP panel for *A. thaliana* covers the genome at a larger extent than the linkage disequilibrium decay requests (in average at least one SNP per 10 kb (S. Kim et al., 2007)). In addition, its very low outcrossing rate (Abbott and Gomes, 1989) and self-fertilization allow to maintain accessions as inbred lines, making possible to phenotype genetically identical individuals (which have to be therefore genotyped only once) multiple times, in order to implement GWAS. This possibility of replicating enables less biased estimates of the phenotypes.

As already mentioned, the first GWAS in *A. thaliana* used a 250K SNP array (M. W. Horton et al., 2012) to test for associations and were able to detect significant association for diverse traits. The first GWAS, using this genotypic data showed that highly significant associations were most frequently located within or near to known candidate genes (Atwell et al., 2010). Later, after the execution of the 1001 Genomes project, it was possible to genotype more than 10 million SNPs (Togninalli et al., 2020), which enabled a even higher SNP density improving the chances to detect the exact physical position of responsible markers.

Flowering time is one of the most extensively studied traits in *A. thaliana*, with more than 300 genes reported to be involved in its regulation (Bouché et al., 2016). Contrary to other species in the genus, *A. thaliana* is an annual plant with some accessions displaying very short life cycles (Krämer, 2015). GWAS have successfully detected markers within or near to genes previously related to flowering time. This is in agreement with the idea that flowering time is regulated by relative stable major genes (Srikanth and Schmid, 2011). However, as aforementioned, markers detected in these studies could explain only half of the heritable variation. So, due to the complex genetic architecture of this trait, underlying phenomena (as those responsible for the "missing heritability") remain hidden after GWAS. Thus, since flowering time is an adaptive trait with wide natural variation across the *A. thaliana* distribution (Agren et al., 2017), one could hypothesize that this trait is regulated by a combination of multiple factor as epistasis, genetic heterogeneity, gene-by-environment interaction, among others. This might also be true for other traits, for example seed dormancy, a fitness-related trait which co-varies with flowering time (Debieu et al., 2013). The implementation of GWAS for such traits is challenging and solving the problem of the hidden phenomena requires much more than just increasing the sample size.

1.5 SIGNATURES OF LOCAL ADAPTATION IN *Arabidopsis thaliana*

Evolution is responsible for biological diversity on earth and the major mechanism shaping this biological diversity is natural selection (Arnegard et al., 2014; Schluter and Conte, 2009). Initially, natural selection was understood as a single mechanism operating over populations, however, we know to date that this term refers to a compendium of processes which we are still trying to understand (Hanson et al., 1999). The complete history of past natural selection is what we see today as adaptation. However, the comparison between ancestral and derived populations might require a geological time scale (Knoll and Nowak, 2017). In contrast, local adaptation can be detected between populations on a much short time scale. This evolutionary process refers at the outcome where individuals in their native population display a higher fitness when compare with individuals coming from a foreign population (Kawecki and D. Ebert, 2004). Local adaptation can be mainly explained by genetic trade-off (with alleles at a single locus contrary affecting fitness when compared on alternative environments) or conditional neutrality (with alleles having a positive effect on fitness on the adapted environment but non on a contrasting environment) (Anderson et al., 2013). Therefore, if a trait is under local adaptation, variation at major genes controlling the trait should cause phenotypic variation between populations.

For this phenomenon to occur, the absence of the constrain of some evolutionary forces like gene flow or genetic drift is necessary. These forces allow shuffling of local polymorphims between populations or can lead to the random loss or fixation of polymorphism (Kawecki and D. Ebert, 2004). Despite local adaptation has been extensively documented in animals and plants, the latter are best suitable to study this phenomenon since they can not migrate, at least actively, and therefore forced to face the local environmental conditions. In addition, plants with a wide geographic distribution represent a valuable source to study local adaptation. In fact, the most classic experiment to prove local adaptation was carried out by transplanting plants into different environments (Turesson, 1922). Since then, reciprocal transplant experiments have been used to appeal for local adaptation in a high number of plant species (Leimu and Fischer, 2008). Given its wide geographic distribution across contrasting environments, local adaptation for many traits in *A. thaliana* might be expected, highlighting flowering time (Agren et al., 2017; Fournier-Level et al., 2011) and seed dormancy (Kronholm et al., 2012; Postma and Ågren, 2016). These two traits have an enormous impact on fitness due to the importance of fine-tuning of germination to avoid desiccation and flower formation in order to ensure reproduction. Reciprocal transplant experiments in *A. thaliana* brought important evidence for local adaptation and the major role played by the climate gradients present along the geographic distribution of this species (Agren et al., 2017; Fournier-Level et al., 2011; Hancock et al., 2011; Postma and Ågren, 2016; Price et al., 2018).

The wide geographic distribution of *A. thaliana* across contrasting environments has finely tuned flowering time, which traduces not only in a latitudinal cline (Stinchcombe et al., 2004) across Europe but in an altitudinal one on the Iberian Peninsula (Méndez-Vigo et al., 2011). Ågren and Schemske, 2012 were the first to report evidence for local adaptation in *A. thaliana* by doing reciprocal transplant experiments between populations representing the extremes of *A.*

thaliana geographic distribution. In their experiments, in 80% of all possible site x year comparisons the native population displayed a higher fitness than the foreign. Along with this seminal work many other studies have highlighted the relevance of *A. thaliana* for ecological and evolutionary studies (Krämer, 2015). Given that this species is not of agronomic importance, and thus it has never been actively selected by humans, the phenotypes collected reflect its natural variation instead the variation of interest for breeders. In addition, the unparalleled availability of genomic data makes *A. thaliana* an ideal species to elucidate the genetic base of local adaptation. In this way, Fournier-Level et al., 2011 provided strong molecular evidence for local adaptation in *A. thaliana* by using a genome-wide approach. Fitness associated alleles at specific sites (four European sites) tend to distribute more locally (be more locally abundant) in comparison with genomic controls (Fournier-Level et al., 2011). In addition, multiple studies by Agren have addressed local adaptation in *A. thaliana* using different approaches (Ågren, Oakley, et al., 2017; Ågren and Schemske, 2012; Oakley et al., 2014; Postma and Ågren, 2016; Postma, Lundemo, and Ågren, 2016; Price et al., 2018)

Considering these facts, local adaptation appears as an important barrier to recovery the same GWAS results when using different samples coming from the same population. This characteristic is still considered as an important signal for reliability (Chanock et al., 2007). Under this assumption, strikingly different results between samples might invalidate a lot of experiments. However, considering the existing evidence of local adaptation for flowering time, it would not be precipitate to predict that GWAS on geographically distant samples should point to dissimilar associations.

Therefore, sampling might be decisive to find signature for local adaptation when implementing GWAS. Including a large number of individuals from very diverse populations in a single sample could hinder the detection of adaptation patterns closely dependent on particular alleles. In this case a more locally restricted sampling could enable the detection of local adapted alleles. However by extreme locally defined sampling one could neglect a wider picture of the genetic architecture of the trait. Thus, under this scenario, differences in GWAS results between geographically distant samples should be considered as evidence for such hidden processes, like local adaptation, rather than a lack of reliability. In fact, there is enough evidence for local adaptation even at smaller geographic scales (Frachon et al., 2018). Despite all these evidence for local adaptation, the genetic basis underlying this process remains hidden.

1.5.1 *Aims of the study*

GWAS are a well established methodology to map the genotype-phenotype relationship. By means of GWAS many associations have been detected for multiple traits in the model plant *A. thaliana*, however most of these studies have used samples either from a wide or a restricted geographical area. In the specific case of flowering time, multiple associations have been detected directly in or near to genes already related to flowering time. However, most of these candidate genes have been characterized using only the reference accession col-1. For that reason and considering that flowering time is a trait presumably under local adaptation, differences in the genetic architecture of this trait should be expected along the geographic distribution of *A. thaliana*. Therefore, the aim of this study was to performed GWAS on different populations of *A. thaliana* in order to see how the outcomes vary according to the geographic distribution. Additionally, multiple strategies were implemented in order to track the phenomena responsible for the differences in GWAS outcomes: gene-gene interaction, allelic heterogeneity and effect of genetic background. Finally, based on these assumptions, we used the compendium of results to point out the following: 1. The regulatory network of flowering time in *A. thaliana* based on col-1 should be reevaluated with the aim of including geographically-dependent regulations, 2. GWAS reliability, understood as the recovery of GWAS results when using different samples from the same population, is not necessary expected for adaptive traits, and 3. dissimilar GWAS outcomes between geographically distant populations constitute evidence for a more complex and geographically-dependent regulatory network for flowering time. Basically, we addressed a very fundamental question in GWAS design: how should a sample be define?, however we approached it not only from a statistical but also from a biological point of view.

2

GWAS AND THE CHALLENGE OF SAMPLING

2.1 INTRODUCTION

Sampling is a major issue when designing GWAS. Defining the appropriate number and origin of the individuals to be included in the analysis not only has important implication due to statistical power but also due to budget. Intuitively, to opt for a sample as large as possible would be the best choice from a statistical point of view, however in most cases economical limitations do not make this possible. On the other hand, the genetic composition of the sample can affect GWAS, as already mentioned in Chapter 1. In most cases samples are defined randomly, which could lead to under-representation of low-frequency variants in the experiment. Of course, this is an expected outcome during sampling and differences in GWAS results when comparing different subsamples could be explained through allele frequency. However, in cases when low-frequency variants are locally distributed or alleles of globally distributed variants have different effects depending on geographic location, important associations could be neglected.

In contrast to GWAS design in human genetics, plant geneticist might not have a huge amount of individuals suited for GWAS. Besides, in many cases well-known phenotypes are consider for GWAS. In this field, sufficient statistical power to detect meaningful effects has been reported using samples fewer than 100 individuals (Atwell et al., 2010). In fact, an important proportion of GWAS in the plant model species *Arabidopsis thaliana* have been performed using samples below 200 individuals (158 from a total of 462 up-to-date stored studies on AraGWAS Catalog (Togninalli et al., 2020)). In these cases, common variants with middle to large effects could be detected at a very high statistical power. On the other hand, low-frequency variants with large effects or common variants with small effects still represent a challenge for such studies (Korte and Farlow, 2013).

In order to practically see how subsampling affect GWAS results, we took advantage of the publicly available genomic and phenotypic data of *A. thaliana* stored in AraPheno (https://arapheno.1001genomes.org/). We selected flowering time at 10 °C. This adaptive trait has been broadly studied not only in *A. thaliana* but also in agronomic important species. In addition, this trait is of paramount importance for flowering plants once its exact timing determines the success of the next generation. To performed GWAS, we used 888 *A. thaliana* accessions distributed across Europe, and created random subsamples, coming from this original sample, varying in size from 800 to 110.

2.2 METHODOLOGY

Flowering time at 10 °C of 888 *A. thaliana* accessions, distributed across Europe, was used to run GWAS. This phenotypic data was filter from a global sample containing 1163 *A. thaliana* accessions, available on the public database AraPheno ((Seren et al., 2016), https://arapheno.1001genomes.org/phenotype/261/). Genotypic data, represented by more than 10 million SNPs, for these 888 *A. thaliana* accessions was obtain from the 1001 genomes project (Alonso-Blanco et al., 2016). Initially, GWAS were run on the complete sample represented by all European accessions and significant associations were defined using Bonferroni correction, which corrects the nominal pvalue (0.05) for multiple testing (Bonferroni, 1935). Next, the complete sample was randomly reduced under three scenarios: filtering randomly out 11, 44 and 88 accessions from the complete sample. For each scenario 100 subsamples were produced and GWAS were performed on all subsamples. As for the initial GWAS run, significant SNPs were defined using Bonferroni correction. Additionally to these 300 reduced subsamples we randomly generated subsamples containing 200 and 110 accession from the complete sample. Same as for the three mentioned scenarios 100 subsamples were generated and GWAS were performed on each of them. In all GWAS runs a linear mixed model accounting for population structure was used:

$$y = X\beta + Z\mu + e$$

y is a vector containing the observed phenotypes, X is a matrix (No. of Individuals x No. of SNPs) of fixed effects, β is a vector representing the effect sizes, Z is a matrix of random effect represented by the kinship matrix, μ is the random effect and e is a matrix containing the residuals. All Parameters were estimated using a customized R script (available at: https://github.com/arthurkorte/GWAS implementing a fast approximation of the mixed model as described in H. Kang et al. 2010. The kinship matrix was inferred from genome wide markers. Theoretical powers were estimated using a derived non-centrality parameter for the Wald test statistic (M. Wang and Xu, 2019):

$$\delta = n_0(\lambda + 1)\frac{h^2_{QTL}}{1 - h^2_{QTL}}$$

n_0 is the effective sample size estimated from the eigenvalues of the kinship matrix

$$n_0 = \sum_{i=1}^{n}(d_i\lambda + 1)^{-1}(\lambda + 1)$$

d are the eigenvalues of the kinship matrix and λ is the ratio of the genetic variance to the residual variance from the null model

$$\lambda = \frac{Var_g}{Var_e}$$

Finally theoretical power was define as:

$$Power = 1 - F_{\chi^2}(\chi^2_{\alpha-1}|1, \delta)$$

2.3 RESULTS AND DISCUSSION

A European sample of 888 *A. thaliana* accession that had been phenotyped for flowering time at 10°C was used for the purpose of assessing how the reduction in sample size affects GWAS outcomes. Therefore, GWAS were carried out in randomly reduced samples coming from the original European sample. This re-sampling was implemented by randomly filtering out 11, 44 and 88 *A. thaliana* accession until 100 samples per group were obtained. We refer to these three groups of re-sampling as RM11, RM44 and RM88. Later, we compared GWAS summary statistics coming from these three groups of re-sampling. Reduction in sample size resulted in a lower statistical power (Figure 2.1, Table 2.1). Standard deviation of pvalues for all SNPs increased as the samples size decreased (Figure 2.1 D). However, the three randomly reduced groups show a similar trend of GWAS results when compared with the complete sample (Figure 2.1). When looking at the highest pvalues for RM11, RM44 and RM88, only the two most significant association would be significant for RM11 and no significant association would be detected for RM44 and RM88 (Figure 2.1 C). In the same way, all reduced groups show extreme pvalues when looking at the lowest pvalues compared to the complete sample (Figure 2.1 B). These results are in agreement with the observed standard deviations (Figure 2.1 D), and suggest that more extreme pvalues can be found in a reduced sample albeit at a very low frequency. Conversely, as Figure 2.1 A shows, a clear trend with more reduced subsamples displaying higher pvalues, due to lower power, is expected.

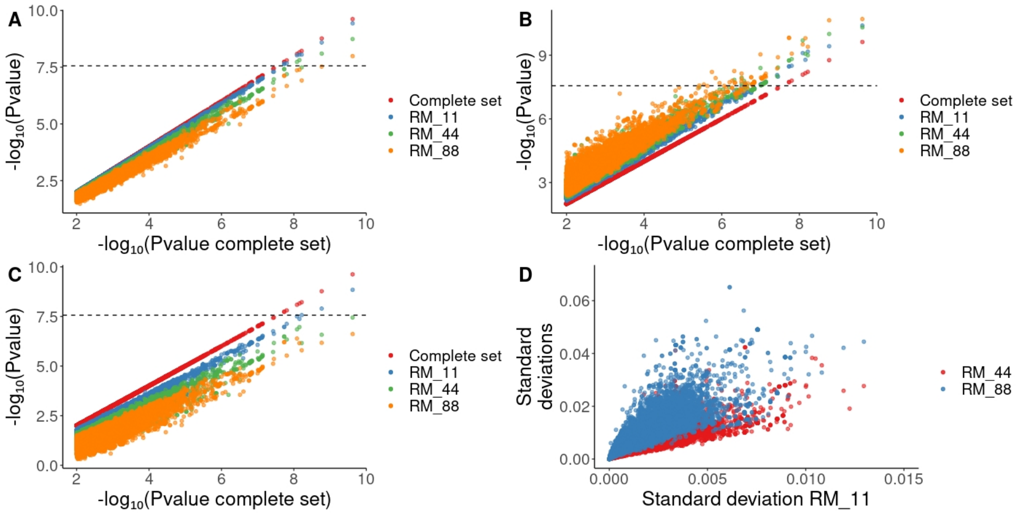


Figure 2.1: Summary of GWAS results for the reduced subsamples RM11, RM44 and RM88 (SNPs with a pvalue $\leq$ 0.01 were plotted). (A) Plot of the mean pvalue of each SNP for RM11, RM44 and RM88 against the pvalue in the complete sample. (B) Plot of the minimum pvalue of each SNP for RM11, RM44 and RM88 against the pvalue in the complete sample. (C) Plot of the maximum pvalue of each SNP for RM11, RM44 and RM88 against the pvalue in the complete sample. (D) Comparison of the standard deviation of pvalues between RM11, RM44 and RM88

GWAS on the original European sample (from now on complete sample) detected 8 significant associations. These associations were considered as those to be recover after implementing GWAS on the subsamples. Table 2.1 contains the empirical power to retrieve these associations in the subsamples. For some SNPs a 1% reduction from the original sample size was enough to decrease the frequency of detection by 25%. In addition, only two SNPs (1- 24339560 and 5- 18590501) were frequently detected in the RM88 subsamples (more than 85%), while for the rest the frequency of detection varied between 33% and 71%. To appropriately compare these results, the theoretical power for the complete sample and for the reduction of the complete sample was calculated (M. Wang and Xu, 2019). For these calculations the proportion of the variance explained for a SNP was set to 0.04, which was the mean of the proportion of variance explained from the significant SNPs in the complete sample. The theoretical power for all sample sizes ranged between 0.83 (complete sample) and 0.73 (RM88). For the reduced samples the power decreased 2% (RM11), 6% (RM44) and and 12% (RM88) in relation to the theoretical power for the complete sample. These proportions are very closed to the obtained proportion for the SNP 5-18590501, however most of the SNPs show a lower proportion of detection compared to the decrease of theoretical power.

SNP	RM11	RM44	RM88
1- 24337820	100	69	47
1- 24339560	100	99	92
1- 24342759	76	42	33
5- 3188327	97	78	66
5- 18590327	97	86	71
5- 18590501	99	95	87
5- 18590741	74	56	51
5- 18590743	74	56	51

Table 2.1: Significant SNPs in the complete sample and their percentage of detection in RM11, RM44, RM88.

Allele frequency is one of the factors affecting GWAS results (Tabangin, Woo, and L. J. Martin, 2009; Zan and Carlborg, 2019). Figure 2.2 shows the change in allele frequency for the significant SNPs in the complete sample for each reduced re-sampling. Allele frequency remained almost unchanged for all reduced re-sampling (all t-test were non-significant) (Figure 2.2 A), which suggests that allele frequency had no major role in the reduction of power in these cases. As showed in Table 2.1 these SNPs were not detected in all reduced samples. For that reason, allele frequency for each SNP was compared between subsamples were the SNP was either significant or non-significant (Figure 2.2 B-C). This comparison revealed no differences in allele frequency for all re-samplings. These results support the idea that the differences in flowering time observed in the subsamples could be explained due to alleles with a moderate effect and being distributed over a wide geographical range (Zan and Carlborg, 2019).

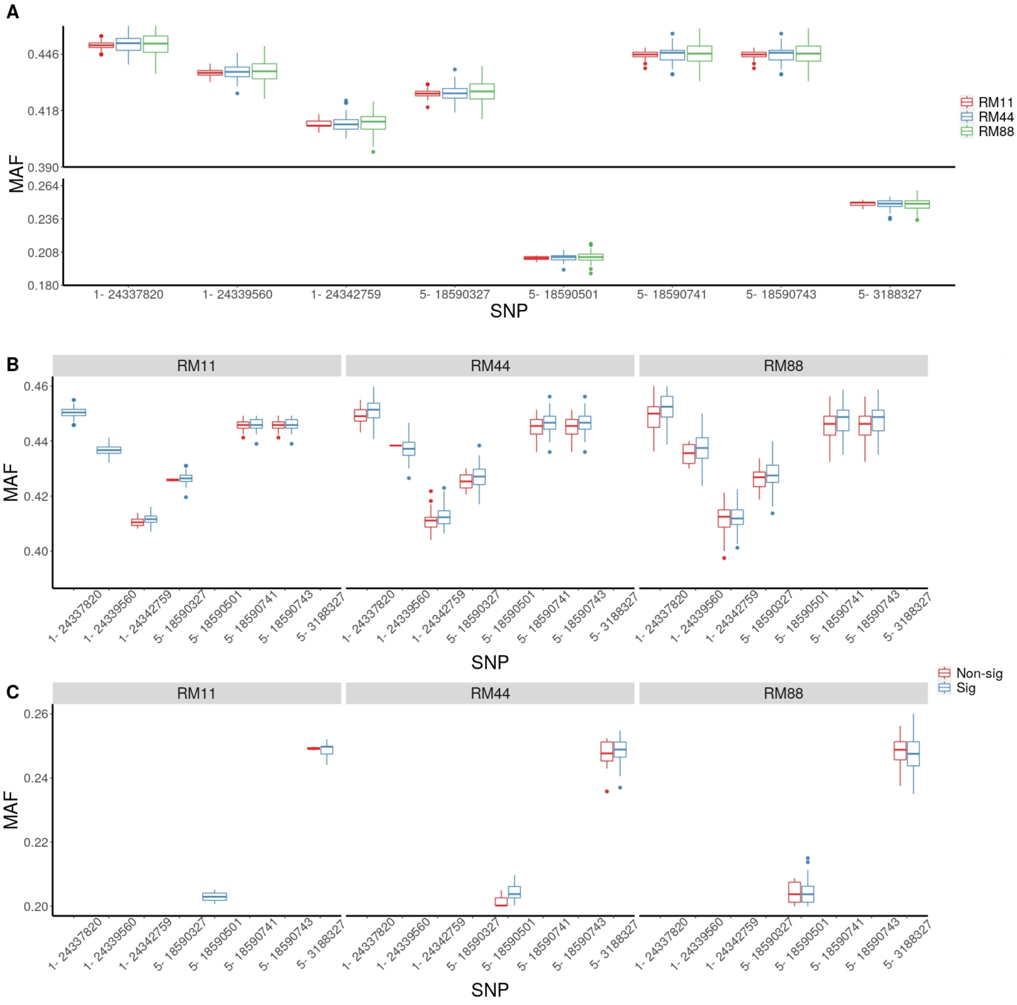


Figure 2.2: Allele frequency variation of significant SNPs in the subsamples RM11, RM44 and RM88. (A) Comparison of minor allele frequency (MAF) between all random subsamples for each significant SNP. (B-C) Comparison of MAF for GWAS results divided into two groups: SNPs being either significant or non-significant associated. To facilitate the comparison SNPs with MAF > 0.4 wer plotted in B while SNPs with MAF < 0.25 were plotted in C.

In contrast to these subtle differences in allele frequency, more noticeable differences in the proportion of the variance explained by a SNP are expected. When comparing the variance explained for each SNP between RM11, RM44 and RM88, no difference was observed for the SNPs on chromosome 1, however there was a slight increase for the SNPs on chromosome 5 (Figure 2.3 A). Nevertheless, this increase would not be enough to explain the low proportion of detection especially for 5- 3188327, 5- 18590741 and 5- 18590743. For that reason, we compared the proportion of variance explained by each SNP between significant and non-significant samples for RM11, RM44 and RM88 (Figure 2.3 B). In all cases, the proportion of variance explained in samples were the SNP was non-significantly associated, was lower. This reduction in the proportion of variance explained led to a lack of power in these samples. Such a lack of power has been extensively reported (Zhu and Zhou, 2020). Even in cases where the marker has a middle size effect on the phenotype, this could be significantly associated if it explains an important proportion of the variance of the phenotype. This scenario reinforces the need for larger samples in cases where the effect size and the proportion of variance explained are expected to be low. In this respect, case-control studies for rare deceases are even more sampling demanding (Momozawa and Mizukami, 2020; Nishino et al., 2018).

Although small sample sizes could be problematic due to reduced power, 158 (from a total of 462) studies reported on AraGWAS Catalog (https://aragwas.1001genomes.org/#/) performed GWAS on samples containing 200 accessions or less. Moreover, sufficient power when using reduced sample sizes, in studies looking at adaptive traits, has been already reported (Atwell et al., 2010). Defining the sample size can affect directly the time execution and budget of GWAS. For such reason, for some GWAS it would be helpful to use small sample sizes, but with sufficient power to detect the lowest expected effect. Taking into account these factors, we randomly produced subsamples with 110 and 200 accessions (each 100 times) from the complete sample and run GWAS. Unlike the results reported for RM11, RM44 and RM88, no similar trend of GWAS results was obtain when comparing the subsamples and the complete sample (Figure 2.4). Whereas, the significant associations found in the complete sample were detected in some randomly subsamples, a large number of makers were significantly associated in the subsamples but not in the complete sample (Figure 2.4 B). More interestingly, a closer look at the GWAS results for the significant associations in the complete sample shows that only

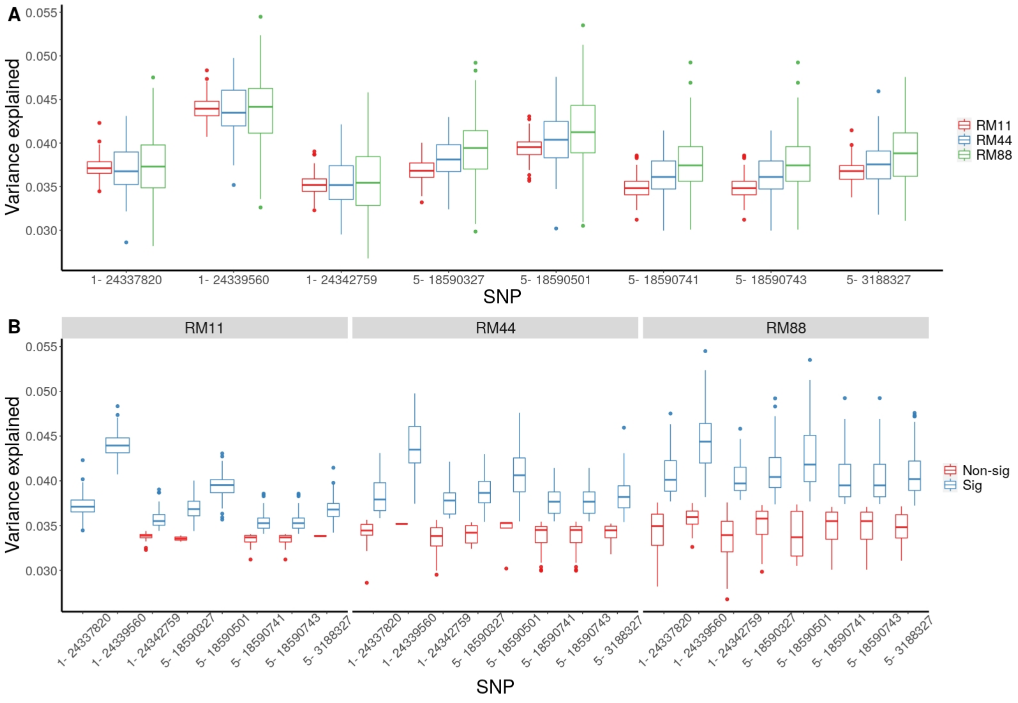


Figure 2.3: Differences in variance explained between random subsamples. (A) Comparison of variance explained between RM11, RM44, and RM88 for each significant SNP. (B) Comparison of variance explained for GWAS results divided into two groups: SNPs being either significant or not significant associated.

two of them were detected in the subsamples (5- 18590501 and 5- 3188237). Under a less stringent pvalue (10^{-4}) both were detected in 39% and 42% of the runs respectively (Table 2.2).

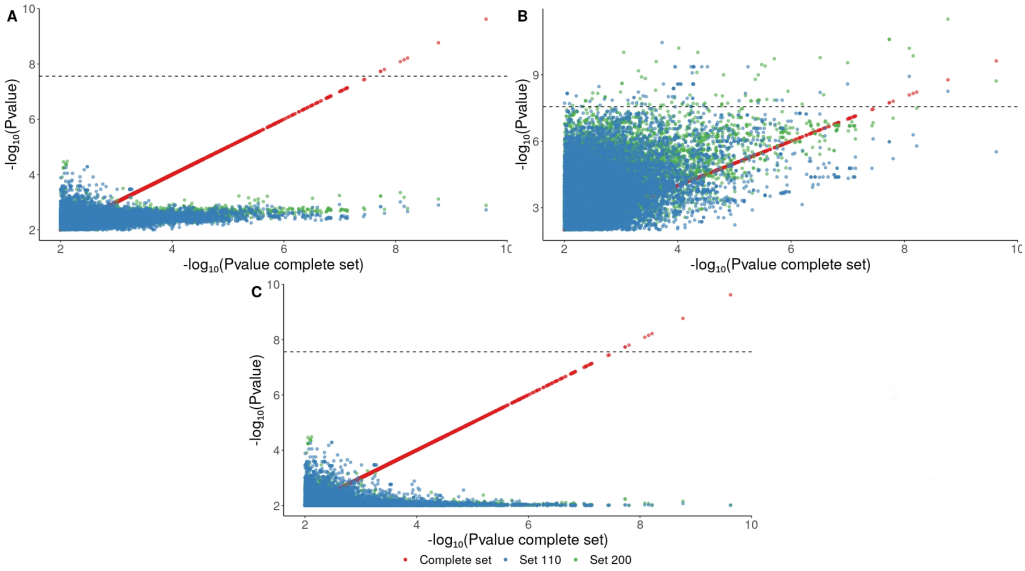


Figure 2.4: Summary of GWAS results for the random 110 and 200 subsamples (SNPs with a pvalue ≤ 0.01 were plotted). (A) Plot of the mean pvalue of each SNP against the pvalue in the complete sample. (B) Plot of the minimum pvalue of each SNP against the pvalue in the complete sample. (C) Plot of the maximum pvalue of each SNP against the pvalue in the complete sample

SNP	MAF ≥ 0.05	Significant	pvalue ≤ 10^{-4}
1- 24337820	100	0	9
1- 24339560	100	0	10
1- 24342759	100	0	11
5- 3188327	100	3	39
5- 18590327	100	0	24
5- 18590501	100	4	42
5- 18590741	100	0	30
5- 18590743	100	0	30

Table 2.2: Significant SNPs in the complete sample and their proportion of detection in the random 110 subsamples at two significance thresholds: Bonferroni correction and pvalue ≤ 10^{-4}

As previously shown for RM11, RM44 and RM88, we expected the variance explained of 5- 18590501 and 3- 3188237 to be higher in the runs in which these SNPs were significantly associated. Based on the parameters used to calculate the theoretical power for the complete sample, we estimated the lowest proportion of variance explained to be detected with a power of 0.8 for a sample size of 110 *A. thaliana* accessions. Based on this estimation, only SNPs explaining at least 0.27 of the variance would be detected at a power of 0.8. Figure 2.5 shows the differences in variance explained when SNPs are either significant or non-significant for all SNPs which were significant in the complete sample. As predicted, in the samples where the SNPs were significant they explained a high proportion of the variance, ranging from 0.25 to 0.29. In contrast, in the runs in which the SNPs were not significant we found values of variance explained as extreme as $2.2x10^{-5}$.

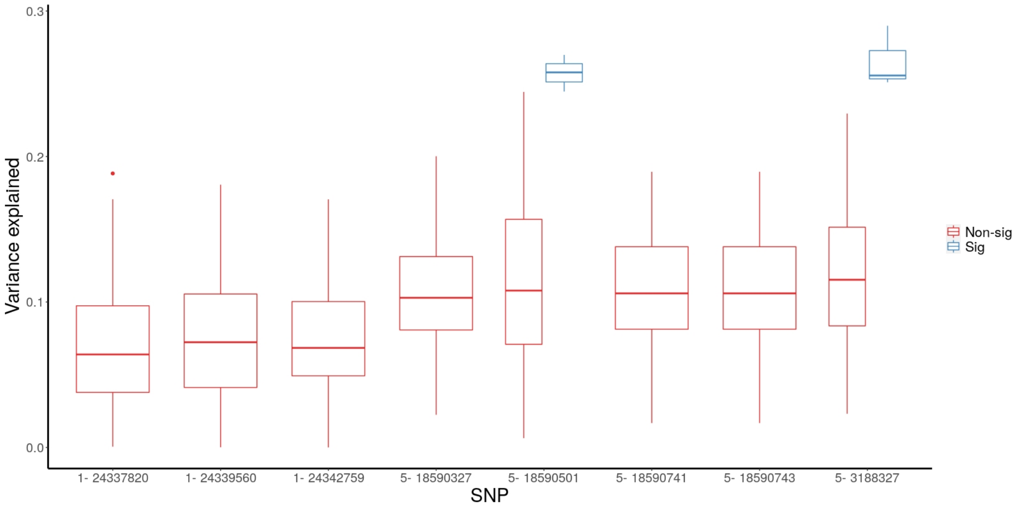


Figure 2.5: Comparison of variance explained for GWAS results from the random 110 subsamples. GWAS results are divided into two groups: SNPs being either significant or non-significant associated.

Flowering time in *A. thaliana* displays a latitudinal cline (Stinchcombe et al., 2004) with early flowering accession mainly distributed in latitudes closer to the Mediterranean Sea. Based on this trend it would be relevant to see if geographic patterns, that randomly appears, could have some effect on the variance explained. By comparing the geographic distribution of the accession in the random samples in which the SNP 5- 18590501 presented the highest and the lowest pvalue, we were not able to distinguish a specific geographic pattern as the accessions belonging to both samples were distributed all over Europe

(Figure 2.6 A). The same broad geographical distribution was observed for the subsamples where this SNP was significantly associated (Figure 2.6 B).

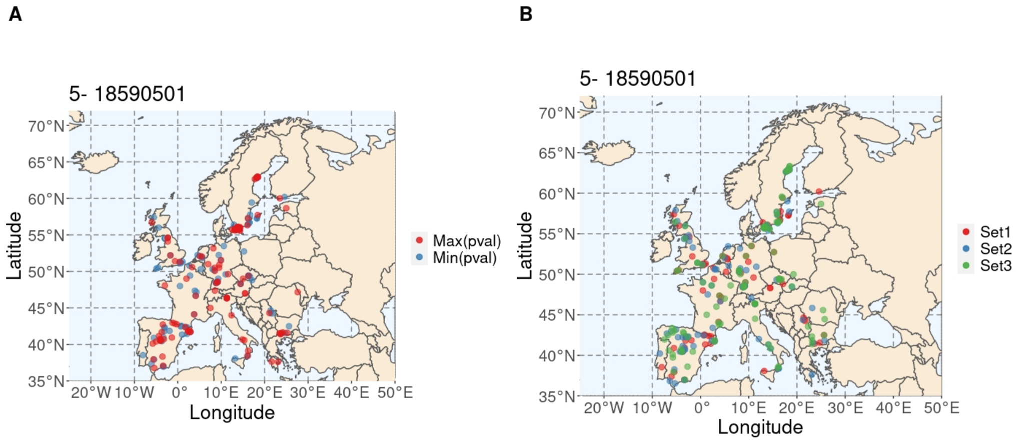


Figure 2.6: Geographic location of *A. thaliana* accessions in 5 random 110 subsamples. (A) Geographic distribution of accessions present in both the subsample with the lowest (red) and highest pvalue (blue) for 5- 18590501. (B) Geographic distribution of accessions representing the three subsamples where 5- 18590501 was significant.

Nonetheless, it would be more enlightening to compare the geographic distribution of the alleles for this SNP in the already mentioned subsamples. Such comparison could reveal a specific distribution pattern correlating with the geographic cline observed for flowering time. Through this comparison, we were able to recognize that the alternative allele in the significant subsamples is mainly present in southern Sweden and northern Spain. Even more remarkable was the fact that in the significant samples the alternative allele was predominantly present in northern Spain (Figure 2.7). By removing these accessions from the original subsamples (which has a negligible impact on the theoretical statistical power) the proportion of variance explained decreases by 63% and the SNP is no longer significantly associated.

These results suggests that markers globally distributed can display a mixture of middle to low effect in a global context, and a more pronounced effect in more locally contexts. These differences could be attributed to the effect of the genetic background. We can think of this as epistasis with the genetic background. Looking more closely at the accessions carrying the alternative allele for 5- 18590501, we found that these accessions tend to flower later (Figure 2.8 A), especially those distributed on the Iberian peninsula. Even though these findings

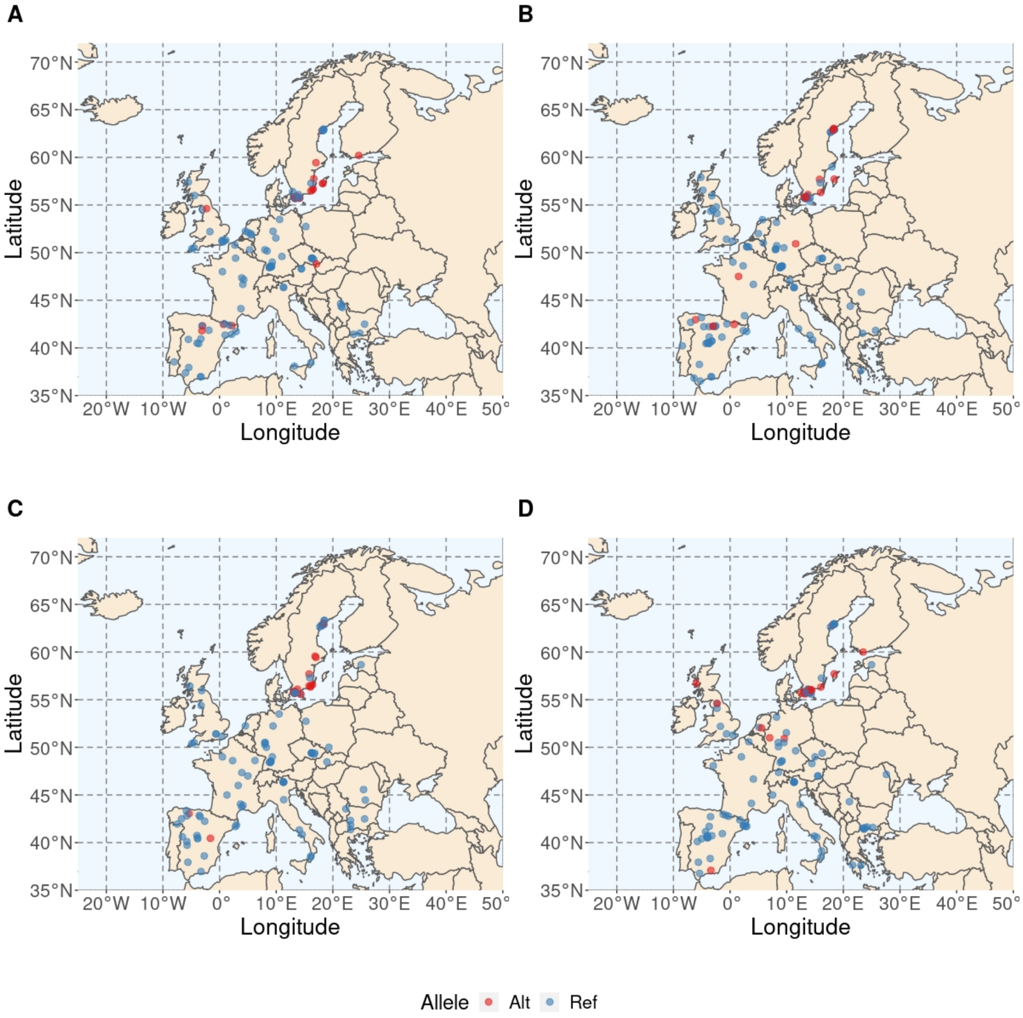


Figure 2.7: Geographic location of *A. thaliana* accessions in 4 random 110 subsamples. (A-C) Geographic distribution of alternative and reference allele for 5-18590501 in subsamples where this SNP was significant. (D) Geographic distribution of alternative and reference allele for 5-18590501 in a subsample where this SNP was non-significant.

seems unusual once the latitudinal cline for flowering time would predict an earlier flowering in Iberian accessions, the Iberian peninsula has a heterogeneous climate ranging from regions with hot and dry summers and mild winters with variable rainfall peaks, through regions where temperatures can fall to 0 °C in winter and rainfall peaks occurring in spring and fall, to regions at higher altitudes where the winters are colder and *A. thaliana* populations require strict vernalization to flower, which resembles conditions from more northern latitudes (Exposito-Alonso, 2020). Also, there is a correlation between flowering time and elevation when filtering for Iberian accessions. To see if the effect of the alleles differ depending on geographic location, we compared the flowering time of the accessions carrying the alternative allele between these two separated locations (Figure 2.8 B). On average, accessions carrying the alternative allele in Spain flower later than those located in southern Sweden. These differences in allele effect between distant locations have been already reported in *A. thaliana* (Agren et al., 2017).

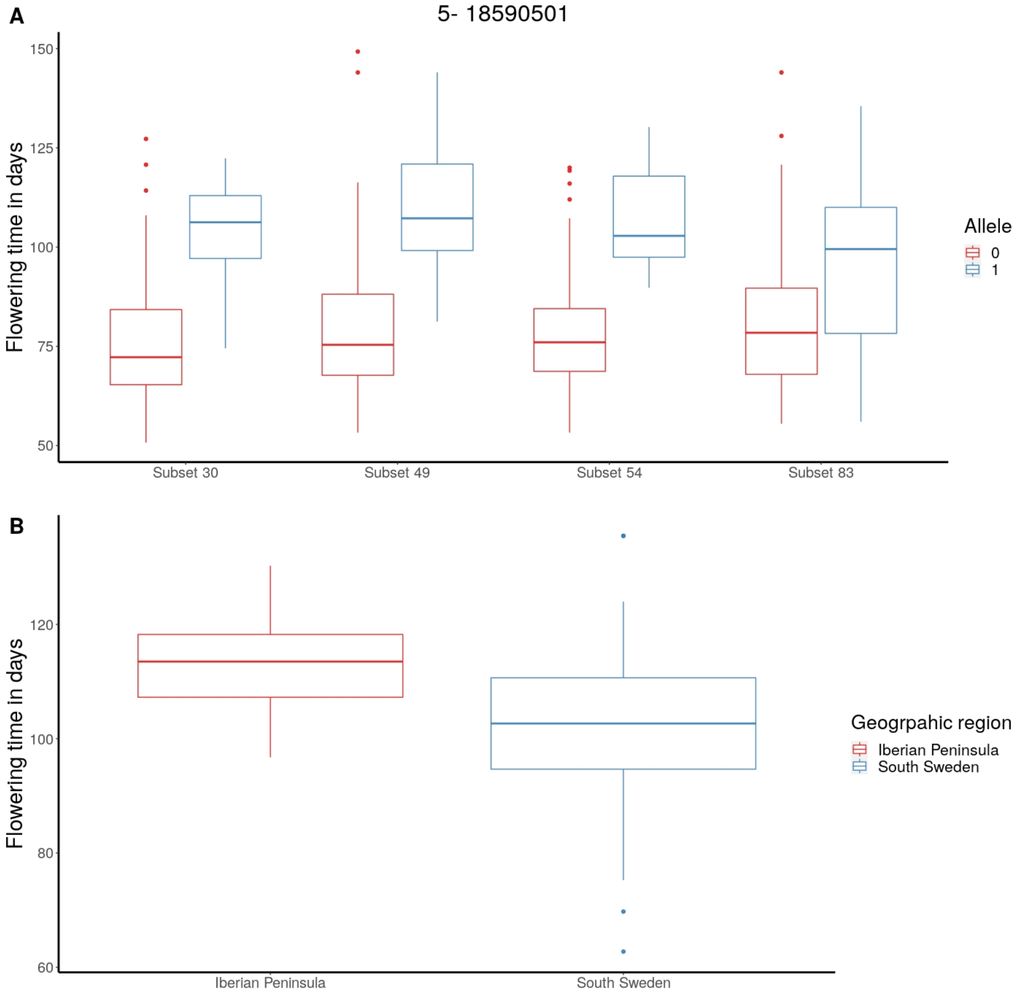


Figure 2.8: Differences in flowering time between accession carrying the alternative or reference allele of 5-18590501. (A) Comparison of the effect of the alternative allele in four 110 random subsamples. (B) Differences in flowering time between Iberian accessions and south Swedish accessions carrying the alternative allele of 5-18590501.

The above reported phenomenon for 5- 18590501 seems not to be true for the SNPs on chromosome 1. First, for these SNPs the overall proportion of variance explained was less variable compared to 5- 18590501, in subsamples where the pvalue were lower than 10^{-4}. Second, the differences in pvalue between subsamples were mainly explained due to changes in MAF. Summarizing, the results for the SNPs on chromosome 1 suggest a trend more oriented to a marker with a stable overall allele frequency and a moderate effect on global contexts. The differences between these two groups of markers are even more relevant considering that both are located in candidate genes already associated to flowering time. The SNP 5- 18501590 is located in *DOG1* (*DELAY OF GERMINATION*), a major gene affecting both germination and flowering time in *A. thaliana* (Huo, Wei, and Bradford, 2016; Kerdaffrec et al., 2016). On the other hand, 1- 24339560 is located in the extensively studied gene *FT* (*FLOWERING LOCUS T*) (Corbesier et al., 2007), which mediates floral transition at the shoot apical meristem through the formation of the florigen activation complex (Kinoshita and Richter, 2020).

Taking in mind the genetic complexity of adaptive traits, it would be straightforward to anticipate that this phenomenon is not restricted to these loci for flowering time, or even for this trait. These results reinforce how challenging the selection of a sample to implement GWAS can be. In experiments with thousands of individuals, a lack of power due to sample size would not be expected, however through random selection the effect of some locally adapted alleles might be undetectable. But in cases where local effects are of interest, it would be advisable to consider the life history of the species. In addition to the classic random sampling strategy, other approaches have been successfully implemented, such as the extreme study design (Berndt et al., 2013; Padmanabhan et al., 2010). In this design, the sampling process is focused on the extremes of the phenotypic distribution, resulting in two extreme samples which are consider as case-control when later implementing GWAS. As well as for the classical sampling strategies, theoretical and empirical power have been estimated under different conditions for this design (Yi Li et al., 2019; Schork et al., 2000). However, despite the availability of alternative sampling strategies, some GWAS still demand large sample sizes to obtain a sufficiently high statistical power (Momozawa and Mizukami, 2020; Nishino et al., 2018). An example for that would be the analysis of rare variants association.

Finally, the results presented in this chapter highlight the necessity of rethinking how to judge GWAS results. The reliability of GWAS does not necessarily rely on the recovery of identical results when comparing different samples from the same population. According to our results, even small sample reductions (10% of the total individuals) can have a strong effect on GWAS outcomes. In more extreme cases, when new small samples are derived from a global sample, completely different GWAS results can be expected. These differences might be detected even if these new small samples have enough theoretical statistical power to detect polymorphisms with middle-to strong effect. Although random sampling is regularly implemented for most GWAS, our results highlight that this might not be the best strategy when implementing GWAS on complex traits. This strategy can effectively catch overall variation but ignore markers with more local effect. So, besides taking into account the theoretical statistical power, it would be even more enlightening to consider additional biological information related to the species and the trait. In cases when strictly Mendelian traits are considered, ensuring enough statistical power should be sufficient to implement GWAS. However when studying more complex traits possible scenarios like gene-gene interaction, allelic heterogeneity, pleiotropy and local adaption, or even a combination of those, should be contemplated. Summarizing, designing a robust sample to implement GWAS depends not only one the size.

3

GLOBAL GENETIC HETEROGENEITY IN ADAPTIVE TRAITS

3.1 INTRODUCTION

This chapter is based on results reported in the following manuscript "Global genetic heterogeneity in adaptive traits". This manuscript is available at: https://www.biorxiv.org/content/10.1101/2021.02.26.433043v1 and currently under review at *Molecular Biology and Evolution*. The introduction is meant to contextualize how challenging the implementation of GWAS on adaptive traits can be.

Elucidating the genetic architecture of complex traits by means of GWAS presents multiple challenges since hidden phenomena, as gene-gene interaction, pleiotropy, genetic heterogeneity that have been described in more detail in the previous chapter, can affect GWAS results. Thus, many complex traits are under local adaption. This will not only affects allele frequency but also the effect of the alleles depending on geographic location. This scenario is certainly expected when assessing the genetic architecture of flowering time, in fact multiple studies reported sound evidences for local adaptation in this trait using the model organism *A. thaliana* (Agren et al., 2017; Ågren and Schemske, 2012; Fournier-Level et al., 2011; Postma and Ågren, 2016). Adding to these evidences the fact that *A. thaliana* occurs predominantly as inbred lines across its geographic distribution (D. W. Meinke et al., 1998) and that flowering time has a huge impact on fitness (Price et al., 2018), the comparisons of GWAS results from populations geographically distant should point to a dissimilar genetic architecture. Such differences should be seen as pieces of a much more complex regulatory network rather than be considered as lack of power or reliability.

In order to asses to which extent local adaptation affects GWAS results, we used publicly available data for flowering time at 10 °C of a dense European sample. This sample was split into 8 almost equally-sized geographic subsamples. Later, GWAS were performed on the whole sample as well as on the subsamples and summary statistics were compared. On the assumption of local adaptation, differences in allele frequency are expected between the mentioned subsamples for loci under selection. In the same way, differences in allele frequency in populations without migration can result from genetic drift. So, also neutral loci can display differences in allele frequency just by chance. It has been assumed that these differences account for most of the deviation in GWAS results when comparing different populations (Zan and Carlborg, 2018, 2019). Nonetheless, taking into account the most likely mechanisms explaining local adaptation, a combination of other factors should be considered to be responsible for these differences as well. GWAS results were used to compare differences in allele effects between subsamples, to detect allelic heterogeneity in major polymorphisms and to assess the effects of the genetic background. Also, GWAS on two additional traits which are presumably under local adaptation as well, namely cauline leaf number and stomata size, were carried out to see if varying GWAS results between populations are the rule rather than the exception. Finally two complementary data sets were use to carried out GWAS. First, simulated phenotypes using accessions from two geographically distinct samples were used in order to see if responsible markers with middle-to-high global effect can be detected via GWAS regardless of the population. Next, GWAS using gene expression as phenotype (eGWAS) were implemented in two geographically distant samples with the aim of testing if gene expression is globally or locally regulated.

3.2 METHODOLOGY

3.2.1 *Plant material and phenotypic data*

Phenotypic and genotypic data were obtained as described in Chapter 2. Phenotypic traits used in the present study include flowering time at 10°C (FT10 ,`https://arapheno.1001genomes.org/phenotype/261/`), flowering time at 16°C (FT16, `https://arapheno.1001genomes.org/phenotype/262/`), stomata size (ST, `https://arapheno.1001genomes.org/phenotype/750/`) and cauline leaf number (CL, `https://arapheno.1001genomes.org/phenotype/705/`). AraPheno stores 1,163 *Arabidopsis thaliana* ecotypes, distributed around the world, for which FT10 has been measured. Taking advantage of the dense sampling in Europe, we resized the original data set to 888 *Arabidopsis thaliana* ecotypes distributed across the European continent. This sample of 888 ecotypes was split into eight, approximately equally-sized (103-119 ecotypes) subsamples, ranging longitudinally from the Iberian Peninsula to Russia and latitudinal from Southern Italy to Northern Sweden (Table 3.1, Figure 3.1). These subsamples are named according to their geographic location as follow: Southern Iberian Peninsula (SIP), Northern Iberian Peninsula (NIP), Germany, France and UK, Central Europe, Skane, North Sweden and eastern Europe. For ST and CL, the total number of ecotypes used in our analyses was 240. For both traits, the initial group of 240 ecotypes was split into two geographic subsamples, one containing Iberian ecotypes (IP, 109 ecotypes) and the other containing Scandinavian ecotypes (SW, 131 ecotypes).

In addition to these traits, we used publicly available RNA expression data ((Kawakatsu et al., 2016), also available via AraPheno (https://arapheno.1001genomes.org/study/52/)) that contain gene expression data for 24,175 genes measured in 727 different ecotypes. We filtered for ecotypes, where also full sequence exist (665) and created two distinct-subsamples following the logic applied before. One subsample is from Scandinavia (termed SW, which contains 70 ecotypes from Sweden, 2 ecotypes from Denmark and 2 ecotypes from Norway), while the second subsample is from the Iberian Peninsula (termed IP, containing 83 ecotypes from Spain and 8 ecotypes from Portugal). The RNAseq data have been generated in two distinct batches, but ecotypes from both subpopulations were predominantly present in the second batch, minimizing potential batch effects in the analyses of the two subsamples[1]. As a control, we also performed GWAS in two random, non-local samples of 91 and 74 ecotypes, respectively, that were sampled from the 165 ecotypes used

3.2.2 *GWAS*

GWAS were performed on the entire European sample, as well as in all subsamples using a linear mixed model (MLM) to account for population structure. The genotype-phenotype correlation was estimated as described in Chapter 2. Significance thresholds were defined using both, Bonferroni and permutation-based threshold. The Bonferroni threshold was obtained dividing the significance level ($\alpha = 0.05$) by the number of SNPs with minor allele count greater than five in each GWAS run. Permutation-based thresholds were derived from running 100

[1] Arthur Korte, personal communication, February 2021

MLMs per phenotype with a random reordering of the phenotypic values, using a fast implementation of the above mentioned model (Freudenthal et al., 2019).

3.2.3 *Candidate gene enrichment*

To look for an enrichment of potential candidate genes, the regions identified for being significantly associated with flowering time have been cross-referenced with a list of 306 known flowering time genes (Bouché et al., 2016). All genes that are within 10 kb of an associated regions have been considered. This analysis was conducted with the 74 regions that are associated in at least two subsamples with flowering time. 22 of these overlap with known flowering time genes. Permutation analysis by re-sampling random regions of the same size across the genome leads to 18.9 + 3.2 regions that overlap the candidate gene list. The slightly higher enrichment of the shared region is not statistically significant. Both, changing the window size between the regions and the candidate genes or restricting the analysis to region that are shared in three or more subsamples had no effect on these results.

3.2.4 *Simulations*

In order to obtain simulated data that mimic local and global effects, we chose the same subsamples used for the anaylsis of ST and CL and established three different scenarios: 1. A single marker explaining a certain amount of variance in the sample containing all 240 ecotypes, 2. A single marker explaining the respective amount of variance only in the IP subsample (109 ecotypes) and 3. A single marker explaining the respective amount of variance only in the SW subsample (131 ecotypes). In each scenario, the responsible marker was chosen randomly from all markers having a minor allele count greater five and set to explain 20%, 15% and 10% of the phenotypic variance, respectively. To mimic population structure, 1,000 random markers were additionally assigned random small effects that are zero-centered. For each possible combination 1,000 simulated phenotypes with different causative markers were generated. This resulted in a total of 9,000 simulated phenotypes that were analyzed in all three different populations. All simulated data were generated using a custom R script (https://github.com/arthurkorte).

When the simulated causative marker explained 20% of the phenotypic variation, GWAS performed using all ecotypes resulted in the detection of this causative marker in 96.4% of the cases, albeit at a high false discovery rate (FDR) of 18.9%. Here, we consider an association as false, if it is more then 100 kb apart from the simulated causal marker. This high FDR dropped dramatically when a more stringent threshold of 10^{-9} was applied. Even with this more stringent threshold, a power of 87.6% was reported, while the FDR dropped to 8.4 %. We observed a reduced power in GWAS when using the two different subsamples (24.6% in IP and and 39% in SW). The reduced rate of detection of the responsible marker in IP and SW is caused by a reduced power due to the smaller sample size. If the simulations mimics a scenario of a marker having a local effect, it was only detected in the respective local subsample (42% in SW and 27.4% in IP) and - with a reduced power - in the analyses using all available ecotypes (6.5% and 27%, respectively) (Table 3.7). Representative GWAS results of the simulated phenotypes are presented in Figure 3.11. The analyses of simulations with a reduced effect size of the causative marker, led to similar results, albeit at a reduced power (Table 3.7).

3.2.5 *Polygenic overlap*

First, we estimated the polygenic overlap among all subsamples by comparing lists of significant SNPs. Since the comparison of significant SNPs between subsamples showed no shared signals, we set a less stringent pvalue threshold (10^{-4}) and generated a new list of SNPs for comparing subsamples. Additionally, we looked at shared significant genomic regions. For this, we summarized all SNPs (10^{-4}) with either r^2>0.9 or located within a 10 kb window for each subsample and compared significant genomic regions. The same procedure has been performed for the respective GWAS results of the subsamples, as well as with GWAS results from permutations within the respective subsample to compare the overlap to the expected overlap in a scenario where no causal markers are present. Next, we estimated the polygenic overlap using the statistical tool *MiXeR* (Frei et al., 2019), which overcomes the intrinsic problem of detecting the exact location of shared causal variants. In short, a summary table containing SNP information, genomic location, beta estimates, and z-scores for each subpopulation was created and used to estimate the proportion of shared causal SNPs between subsamples based on their beta and z-score distributions.

3.2.6 *RNA expression data*

The available RNAseq data contain transcription levels of 24,175 genes. Before performing GWAS on the RNA expression data, we removed transposable elements and genes that are encoded by the organelle genomes, leaving 23,021 nuclear genes for the further analyses. Next we used only genes where the heritability estimate was above 0.5 and a statistical power analysis indicates the power to be above 0.9 in all three samples analyzed. Heritability was estimated for all genes using the above mentioned implementation of the mixed model. The power of each data set was calculated using the *pwr.p.test* function implemented in R package *pwr* (Champely et al., 2017). This filtering led to a set of 2,237 genes for which GWAS were performed in both subsample and the combined sample. We only considered markers with a minor allele count of more than five in the respective subsample. Given the amount of tests we performed, we used a very stringent multiple-testing threshold of 10^{-10} to term an association as significant, but similar results have been reproduced with threshold ranging from 10^{-8} to 10^{-12}.

Significant associations were grouped into regions, if they occur within 50 kb of each other. Genes showing inflated GWAS results (which quite often co-occurs with a non-normal distribution of the expression values), have been filtered out, if the number of associated genomic region was greater than three in either the Iberian (IP) or the Scandinavian (SW) subsample. This procedure left us with a set of 1,982 genes for the analyses. From this set, significant genome-wide associations, at least in one of the subsample, were detected for 780 genes at a significance threshold of 10^{-10}. For genes, where the same region was associated, we defined genes having a global genetic regulation as genes displaying the same significant marker in both subsamples (110), while we excluded genes where the same region but different markers are associated in the subsamples (potential allelic heterogeneity). Genes, that show a local genetic regulation were defined as genes having a significant association either in IP or SW, but not in the other subsamples. This led to the identification of 377 genes displaying an association only in IP and 176 genes displaying an association only in SW. We filtered these genes for genes, where the respective pvalue was lower in the analysis of the respective subsample compared to the results of the combined sample. We argues that a true local association should be more significantly associated in the local subsample.

Additionally, we also excluded genes, where different regions had been associated in the analysis of the combined sample then in the respective local subsample. This procedure led to a set of 118 genes displaying a local association only in IP and 64 gens for SW. Next, we took all significant association in these three groups of genes, having the same association in both subsamples, a local association only in IP or a local association only in SW, and verified, if the associated SNP was in *cis*, aka the same genomic region (defined by a maximum distance of the associated region of 100 kb to the respective gene) where the gene is located, or if the associated marker is in *trans*. As a control, we also performed the same analysis described above with two random, non-local samples of 91 and 74 accessions, respectively, that had been sampled from the merged sample of 165 accessions.

3.3 RESULTS AND DISCUSSION

3.3.1 *Complex genetic architecture and local adaptation affect GWAS results*

Since *A. thaliana* was proposed as a suitable model plant for genetics (Laibach, 1943), scientists have been trying to elucidate the underlying mechanisms controlling flowering time. To date, more than 300 candidate genes involved in flowering time have been reported (Bouché et al., 2016). For our study, we took publicly available data of flowering time measured in growth chamber at 10 °C using a globally distributed sample. In order to compare flowering time between more locally defined samples, we restricted our analysis to an European sample containing 888 *A. thalina* accessions. As the latitudinal cline for flowering time in *A. thaliana* predicts (Stinchcombe et al., 2004), this European sample shows a phenotypic distributions with both early flowering accessions mostly distributed at lower latitudes (near to the Mediterranean sea) and later flowering accessions distributed at more northern latitudes (Scandinavia) (Figure 3.1 C). From this European sample, we defined eight approximately equally-sized (n = 103 - 109 subsamples (Southern Iberian Peninsula, Northern Iberian Peninsula, Germany, France UK, Central Europe, Skane, North Sweden and Eastern Europe) using geographic information (Figure 3.1 A). The most variation in flowering time was observed in the Iberian subsamples (SIP nad NIP) and France/UK, while Germany and Central Europe displayed the lowest variation (Figure 3.1 B). As already mentioned in Chapter 1, this higher variation of flowering time in

Iberian subsamples is in agreement with the heterogeneous climatic regions present on the Iberian peninsula, in addition to a clear correlation between flowering time and elevation (Figure 3.1 D).

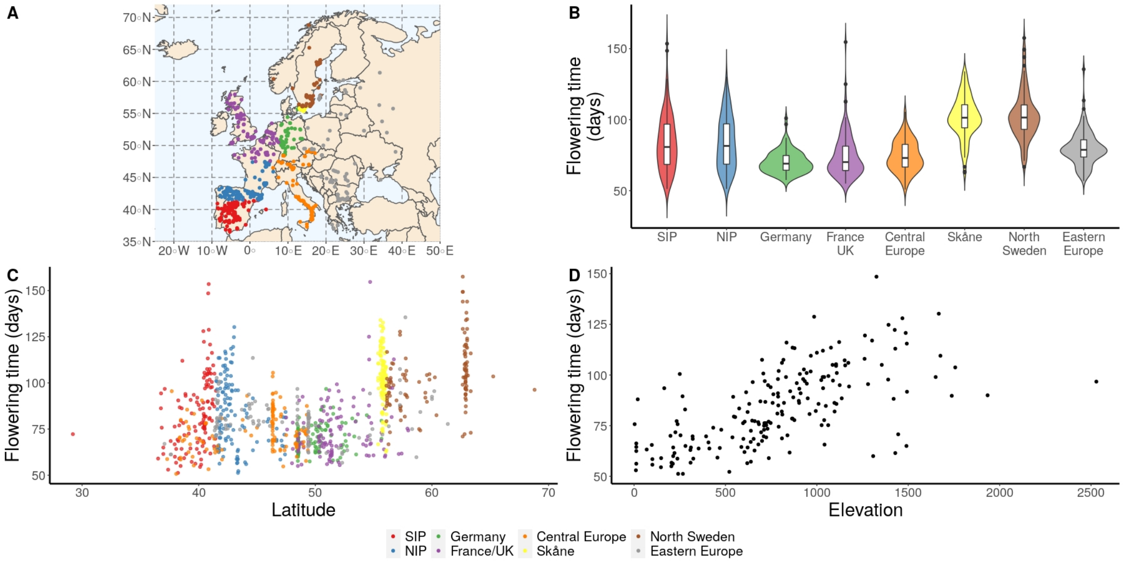


Figure 3.1: Flowering time and geographic distribution of 888 *A. thaliana* accessions across Europe. (A) Geographic distribution of the eight European subsamples. (B) Phenotypic distribution across the eight European subsamples. (C) Correlation between latitude and flowering time. (D) Correlation between Elevation and flowering time of accessions from the Iberian Peninsula

Since the results presented in Chapter 2 showed strong variation in GWAS results when comparing randomly defined subsamples, a similar or even stronger trend when comparing GWAS results between these eight European subsamples could be expected. These subsamples were geographically defined and therefore with a greater tendency to display non-shared local associations. To initially verify to what extend GWAS results differ, manhattan plots displaying GWAS results for the European sample and the eight subsamples were compared (Figure 3.2). This visual comparison uncover strikingly different associations between subsamples and even when comparing with the European sample. Genome-wide significant associations (or just below the significance level) were detected in the European sample, SIP, NIP, North Sweden and East Europe. These associations were defined using both a permutation based threshold (Freudenthal et al., 2019) and Bonferroni correction (Bonferroni, 1935). The first explanation that comes to mind for these differences is the reduced statistical power in the subsamples compared to the European sample. However, flowering time is a trait with high heritability (Table 3.1) and it is believed that major polymorphisms are common (Mouradov, Cremer, and Coupland, 2002). Additionally, enough statistical power to detect meaningful associations has been reported in samples just over 100 individuals (Atwell et al., 2010).

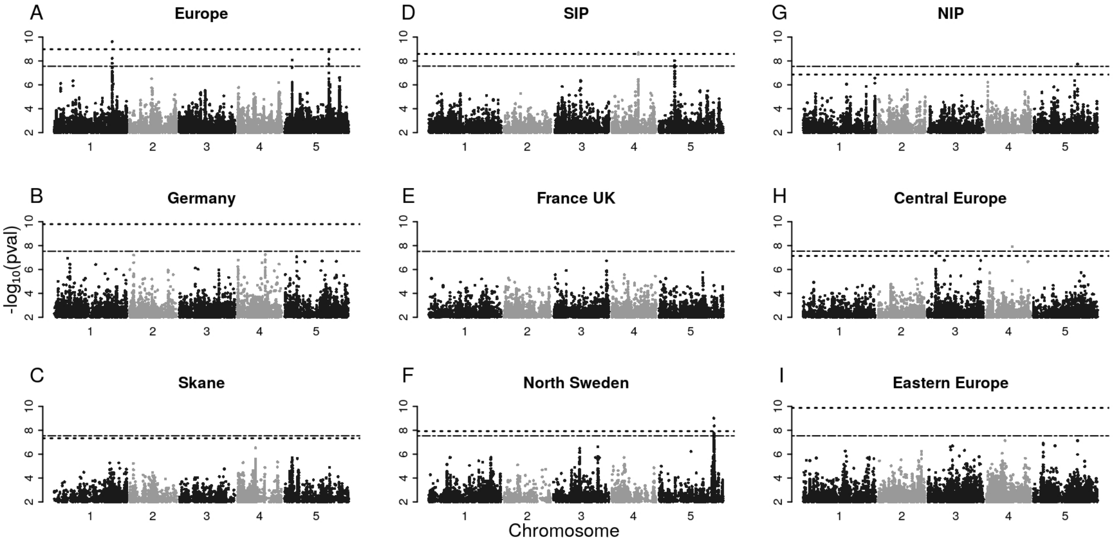


Figure 3.2: Manhattan plots showing GWAS results obtained from the European sample and from its eight derived subsamples. Dashed lines and dash-dotted lines indicate 5% permutation-based and Bonferroni threshold, respectively.

Estimates of pseudo-heritability and power to detect major effects (markers explaining at least 10% of the phenotypic variation) for the eight subsamples are present in Table 3.1. All European subsamples have sufficient statistical power to detect major polymorphisms. Based on this, similar GWAS results could be expected across subsamples. However, as already illustrated, manhattan plots showed a dissimilar distribution of association when comparing across subsamples and with the complete sample (Figure 3.2). These dissimilar results can not be explain through subtle differences in statistical power. In such a case it would be expect a similar distribution of association with some of them being below the significance threshold due to a lower proportion of variance explained in the smaller subsamples.

Beyond the statistical power, it will be necessary to consider other factors that could enlighten the source for the observed differences. Taking in mind the latitudinal cline and being aware of the extreme phenotypic variation for flowering time when comparing accessions from different geographic locations, it would not be wrong to think that strong differences in allele frequency between subsamples might be one of these factors. In contrast to other studies, where allele frequency was the more relevant factor to explain differences in

Table 3.1: Geographic location of the European subpopulations.

Subsets	No. accessions	min_lat	max_lat	min_lon	max_lon	$\widehat{h_2}$[a]	power
SIP	107	36.52	41.48	-8.54	4.25	0.99	1.00
NIP	108	41.50	47.45	-7.80	6.13	0.91	0.99
Germany	107	48.39	55.67	8.00	13.73	0.99	1.00
France/UK	107	47.50	57.97	-5.98	7.50	0.95	0.99
Central Europe	106	37.30	49.37	6.08	17.31	0.66	0.99
Skåne	119	55.38	56.10	13.10	14.78	0.91	0.99
North Sweden	118	56.10	68.80	6.19	18.52	0.75	0.99
Eastern Europe	116	37.07	61.36	38.28	38.28	0.91	0.99
Europe	888	36.52	68.80	-8.54	38.28	0.86	1.00

[a] $\widehat{h_2}$: pseudo-heritability estimate

GWAS results using flowering time as a trait (Zan and Carlborg, 2019), here it is considered as one more player affecting GWAS results in different samples. In order to provide a more clear example, the allele frequency of the significant associations was compared between subsamples. The SNP 5- 23100540 represents a clear case with allele frequency being the most grounded explanation for the observed differences. This SNP was significant associated in the European sample and in North Sweden, with a marginal proportion of the alternative allele in SIP and NIP. This extreme case shows how the alternative allele of this marker is almost restricted to north Sweden (Figure 3.3).

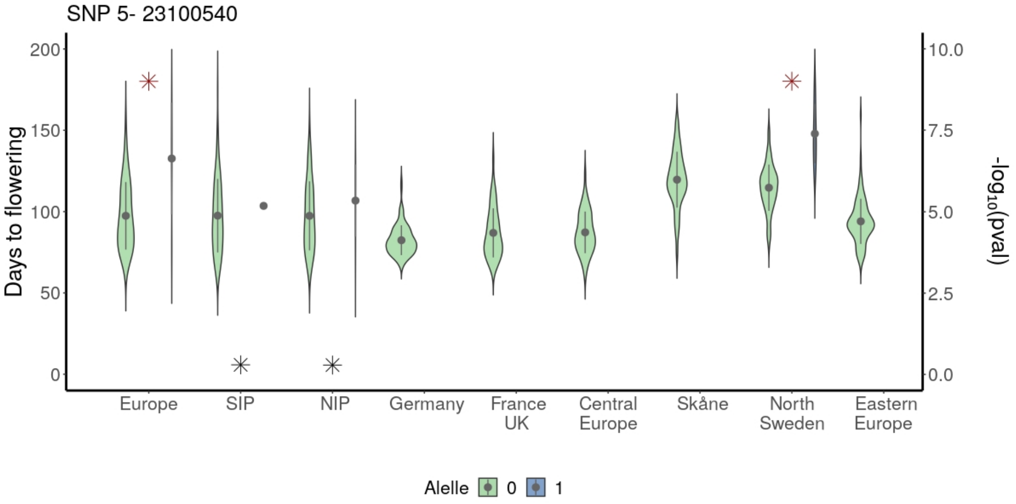


Figure 3.3: Violin plots comparing flowering time between accessions carrying the reference or alternative allele for SNP 5:23100540. Stars represent the -$\log_{10}$ of the pvalue and a red colored star indicates a significant association.

This scenario is in accordance with previous observations that key genes for local adaptation are mostly local and specific to certain environments (Fournier-Level et al., 2011). Additionally, it is important to note that this SNP is located in *VIN3* (*VERNALIZATION INSENSITIVE 3*), a gene already reported to be involved in the control of flowering time in *A. thaliana* (D.-H. Kim and Sung, 2013; Sung and Amasino, 2004). Table 3.2 contain the location of the significant associations and the closest gene already reported affecting flowering time in *A. thaliana*. Despite this clear example of allele frequency affecting GWAS results, in most cases allele frequency failed as an explanation for the dissimilar association peaks between subsamples.

Table 3.2: Significant SNPs (chromosome:position) in the GWAS of different subpopulations. Entries are pvalue (minor allele frequency), with genome-wide significance using a 5%-permutation-based threshold shown in red. Candidate genes were assigned to the SNPs from a list of 306 flowering time genes (Bouché et al., 2016) using 10 kb window.

SNP	1:24339560	3:3458977	4:10949262	4:11016778	5:18590501	5:23100540	5:23234243
Candidate gene	*FT*[a]		*TSF*[b], *JMJ14*[c]	*TSF, JMJ14*	*DOG1*[d]	*CIR1*[e], *VIN3*[f]	*CIR1, VIN3*
Europe	2.4e-10 (0.44)	4.3e-03 (0.18)	4.1e-01 (0.34)	1.4e-04 (0.25)	1.7e-09 (0.20)	9.9e-10 (0.03)	1.4e-06 (0.07)
SIP	1.7e-02 (0.45)	5.1e-02 (0.47)	9.3e-01 (0.19)	2.0e-09 (0.12)	2.9e-02 (0.03)	5.2e-01 (0.01)	7.4e-01 (0.04)
NIP	1.2e-02 (0.35)	1.2e-01 (0.35)	8.2e-01 (0.32)	6.8e-02 (0.22)	1.8e-08 (0.14)	5.2e-01 (0.04)	5.8e-01 (0.10)
Germany	3.1e-02 (0.28)	9.3e-01 (0.05)	6.4e-01 (0.49)	9.4e-01 (0.28)	2.6e-02 (0.06)		2.4e-01 (0.06)
France/UK	7.0e-04 (0.41)	2.9e-01 (0.08)	6.4e-01 (0.50)	9.5e-01 (0.17)	1.4e-01 (0.05)		8.2e-01 (0.08)
Central Europe	7.4e-02 (0.46)	4.1e-08 (0.22)	1.2e-08 (0.37)	1.1e-01 (0.12)	8.2e-02 (0.05)		
Skåne	5.1e-02 (0.24)	2.1e-01 (0.12)	5.7e-01 (0.36)	8.5e-02 (0.49)	1.5e-01 (0.33)		2.7e-01 (0.01)
North Sweden	3.1e-01 (0.20)	9.7e-01 (0.13)	9.1e-01 (0.22)	4.2e-01 (0.37)	1.0e-01 (0.48)	9.8e-10 (0.20)	4.3e-09 (0.24)
Eastern Europe	2.6e-01 (0.44)	6.2e-01 (0.09)	7.4e-01 (0.26)	2.8e-01 (0.14)	7.4e-08 (0.08)		4.1e-01 (0.01)

[a] *FT* (*FLOWERING LOCUS T*, Corbesier et al. 2007)

[b] *TSF* (*TARGET OF FLC AND SVP1*, Yamaguchi et al. 2005)

[c] *JMJ14* (*JUMONJI 14*, Lu et al. 2010)

[d] *DOG1* (*DELAY OF GERMINATION 1*, Huo, Wei, and Bradford 2016)

[e] *CIR1* (*CIRCADIAN 1*, X. Zhang et al. 2007)

[f] *VIN3* (*VERNALIZATION INSENSITIVE 3*, Sung and Amasino 2004)

This trend of an alternative allele being almost restricted to a geographic region was only found in a handful of markers. The significant associations on chromosome 5 (5- 18590501 and 5- 18590247) represent a clear example of an opposite trend. Both SNPs are located in *DOG1* (*DELAY OF GERMINATION*), an extensively studied gene involved in the regulation of seed dormancy (Huo, Wei, and Bradford, 2016; Kerdaffrec et al., 2016), which has been also reported to affect flowering time (Alonso-Blanco et al., 2016). The alternative alleles of these SNPs ares present across Europe, mainly on the Iberian Peninsula, Sweden and Eastern Europe (Figure 3.4, Figure 3.5). In this case, allele frequency seems not to be the main factor affecting GWAS results. Here, it is more important the differences of allele effect on flowering time when comparing NIP and eastern Europe with Skane and Norht Sweden. This type of differences in allele effect have been already reported (Agren et al., 2017), and this case in particular was also detected using random subsamples (see Chapter 2). The extreme differences in allele effect between these distant subsamples suggest a strong effect of the genetic background on this locus. In addition to these differences between distant subsamples, more local differences in seed dormancy between accessions distributed on the Iberian Peninsula have been reported (Exposito-Alonso, 2020; Martınez-Berdeja et al., 2020). The more late flowering accessions carrying the alternative allele on the Iberian Peninsula are mainly distributed in climate regions which resemble more northern latitudes and some of them require strict vernalization to flower (especially those located over 900 meter above the see level). Considering some evidence that flowering time and seed dormancy co-vary (Debieu et al., 2013), the differences in allele frequency observed for these SNPs could be the result of the fine-tuning of both traits instead of just one of them, suggesting a pleitropic effect of this gene on flowering time (Auge, Penfield, and Donohue, 2019).

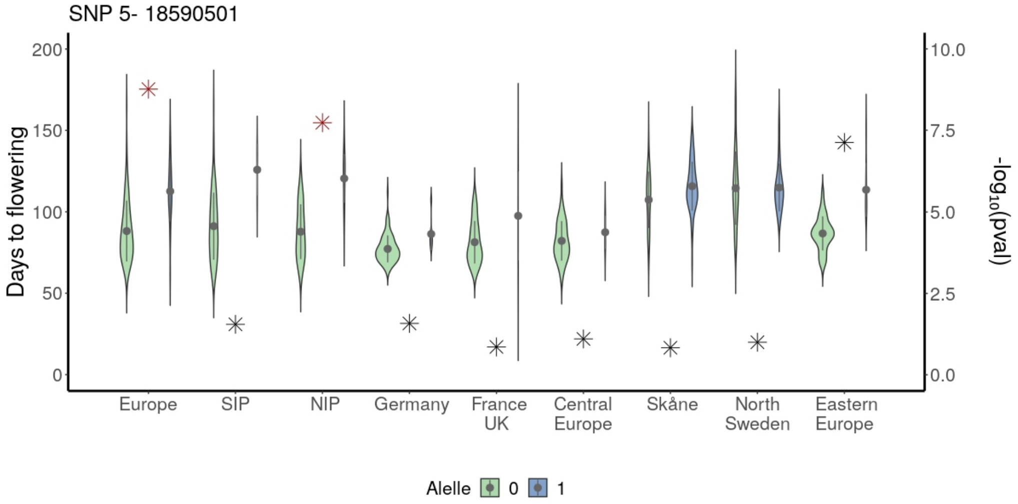


Figure 3.4: Violin plots comparing flowering time between accessions carrying reference and alternative allele for the SNP 5:18590501. Stars represent the -$\log_{10}$ of the pvalue and a red colored star indicates a significant association.

The SNP 4- 11016778 represents an example, where both alleles are present in all subsamples. This SNP displays slightly differences in allele frequency between

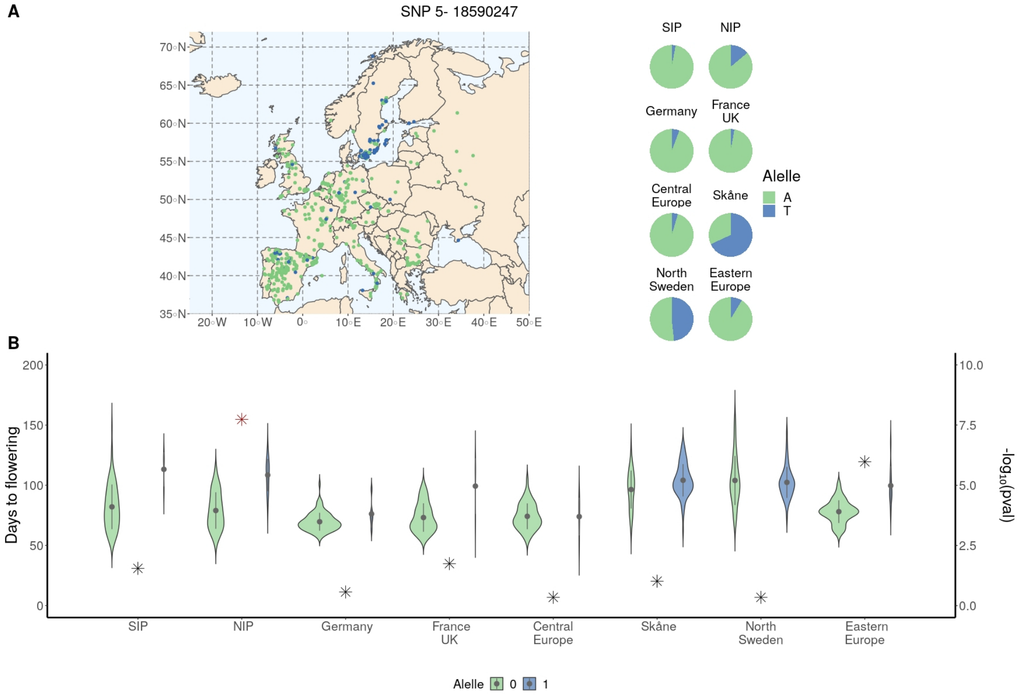


Figure 3.5: (A) Geographic distribution of the alleles for the SNP 5:18590247 and their proportion in the different subsamples. (B) Violin plots comparing flowering time between accessions carrying reference and alternative allele for the SNP 5:18590247. Stars represent the -log$_{10}$ of the pvalue and a red colored star indicates a significant association.

subsamples and was significant associated only in SIP (Figure 3.6). As well as the aforementioned SNPs, it locates near to genes reported to affect flowering time, namely *TSF* (*TARGET OF FLC AND SVP1*, Yamaguchi et al. 2005) and *JMJ14* (*JUMONJI 14*, Lu et al. 2010). At this point all associations were located directly in (or near to) a gene already reported to affect flowering time. These results support the idea that the reported associations are true discoveries. However, when comparing all significant association across subsamples none of them were found to be shared between subsamples. This lack of shared associations could be the result of loci with middle-to-low effect being undetectable due to small sample size (Korte and Farlow, 2013). To uncover these possible hidden shared associations, SNPs with pvalues below 10^{-4} were compare between all subsamples. Although this less stringent significance level could go along with it a high number of false associations, the number of truly shared associations should be higher than the number of shared association expected only by chance, since false positives at this significance level are uncorrelated, as shown by comparing permutation results.

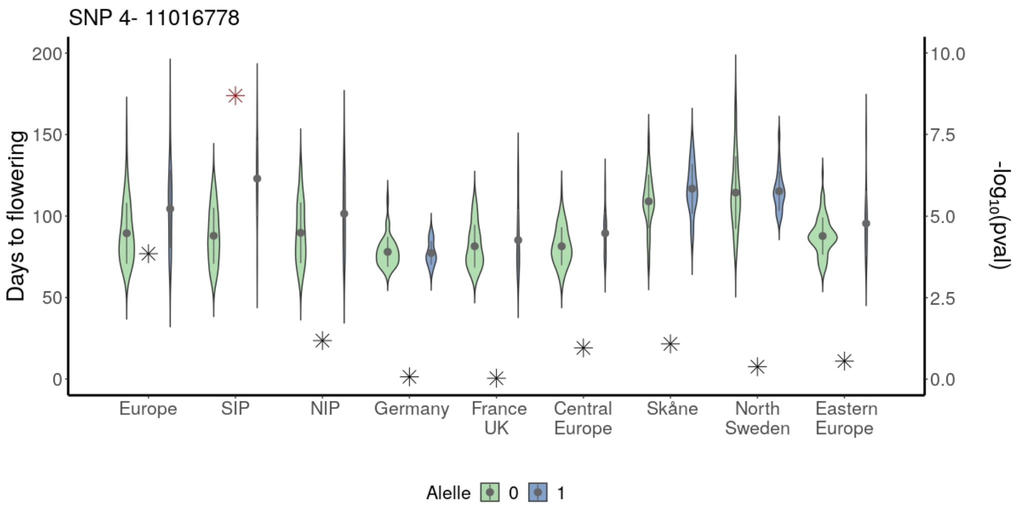


Figure 3.6: Violin plots comparing flowering time between accessions carrying reference and alternative allele for the SNP 4:11016778. Stars represent the -log$_{10}$ of the pvalue and a red colored star indicates a significant association.

Table 3.3: Shared SNPs between subpopulations at significance level of $p < 10^{-4}$

Associated marker[a]	1:29186215	1:29199833	3:20379636	4:6781375	5:18589998	5:18590247	5:18590501	5:18590591
Candidate gene[b]			*SMZ*[c]		*DOG1*[d]	*DOG1*[d]	*DOG1*[d]	*DOG1*[d]
SIP	3.75e-05 (0.06)	3.75e-05 (0.06)	6.22e-01 (0.04)	1.15e-02 (0.02)	1.38e-01 (0.04)	2.92e-02 (0.03)	2.92e-02 (0.03)	2.92-02 (0.03)
NIP	4.80e-01 (0.12)	4.80e-01 (0.12)	2.51e-01 (0.06)	9.45e-01 (0.06)	1.83e-08 (0.14)	1.83e-08 (0.14)	1.83e-08 (0.14)	1.83e-08 (0.14)
Germany	1.55e-01 (0.10)	1.04e-01 (0.10)	9.06e-05 (0.07)	3.88e-01 (0.07)	8.28e-03 (0.05)	2.73e-01 (0.05)	2.58e-02 (0.05)	6.30e-01 (0.03)
France/UK	7.34e-01 (0.29)	7.34e-01 (0.29)	4.52e-01 (0.20)	7.75e-05 (0.08)	1.74e-03 (0.019)	1.86e-02 (0.03)	1.43e-01 (0.05)	5.53e-03 (0.03)
Central Europe	5.90e-01 (0.03)	5.9e-01 (0.03)		2.70e-02 (0.13)	2.60e-01 (0.05)	4.60e-01 (0.05)	8.15e-02 (0.05)	
Skåne	8.55e-01 (0.25)	8.55e-01 (0.25)	4.35e-02 (0.10)	4.44e-01 (0.45)	8.00e-01 (0.23)	9.86e-02 (0.32)	1.51e-01 (0.33)	1.80e-01 (0.37)
North Sweden	4.61e-01 (0.5)	4.78e-01 (0.5)	4.49e-01 (0.43)	7.04e-01 (0.36)	5.12e-01 (0.34)	4.59e-01 (0.48)	1.03e-01 (0.48)	5.75e-10 (0.36)
Eastern Europe	2.85e-05 (0.05)	2.85e-05 (0.05)	9.29e-05 (0.05)	9.62e-05 (0.06)	7.43-08 (0.08)	1.05e-06 (0.09)	7.43e-08 (0.08)	7.43e-08 (0.08)

[a] represented as "chromosome:position"

[b] known flowering time gene within a 10 kb window around the associated marker using a list of 306 flowering time genes from Bouché et al. 2016

[c] *SMZ (SCHLAFMÜTZE)*, Mathieu et al. 2009

[d] *DOG1 (DELAY OF GERMINATION)*, Huo, Wei, and Bradford 2016

Even under this relaxed significance level, the same trend was observed, with most of the associations being unique in a subsample and only eight associations being shared out of a total of more than 5700 associations (Figure 3.7 A). More interesting, all shared associations were present only between two subsamples, and the higher overlap size matches with the already mentioned associations in *DOG1* (Table 3.3). Although the number of shared associations remains low, it is higher than the shared associations expected only by chance (in this case no shared associations are expected) (Figure 3.8 A), which supports the idea that the observed shared associations were not produce due to false positives. One might hypothesize that the low number of shared associations could be explained through different SNPs tagging the same causal polymorphism in the subsamples, however the high SNP density used in this study, with approximately one SNP per 10 bp, does not support this explanation. In fact, if this phenomenon occurs in such a dense SNP panel, it would be more likely as an effect of allelic heterogenity which is consistent with local adaptation (Atwell et al., 2010; Kerdaffrec et al., 2016; P. Li et al., 2014; L. Zhang and Jiménez-Gómez, 2020). In order to address this explanation and at the same time because of the reduced number of shared association detected after comparisons on a single marker level, a strategy based on associated genomic regions was implemented.

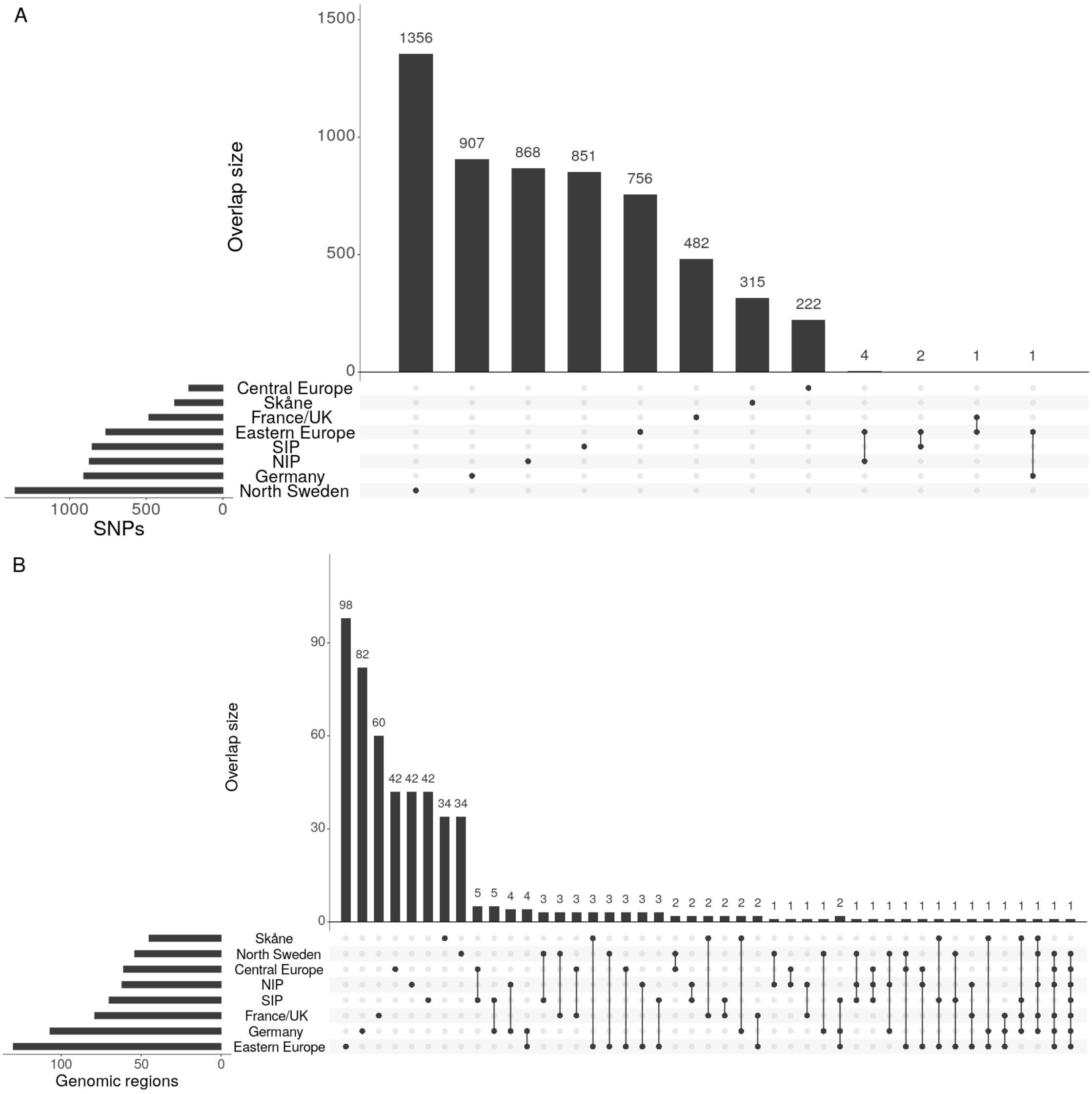


Figure 3.7: Sharing of sub-significant ($p < 10^{-4}$) associations. (A) Histogram of the number of associated SNPs in each subpopulation and shared between subpopulations. (B) Histogram of the number of associated genomic regions in each subpopulation and shared between subpopulations.

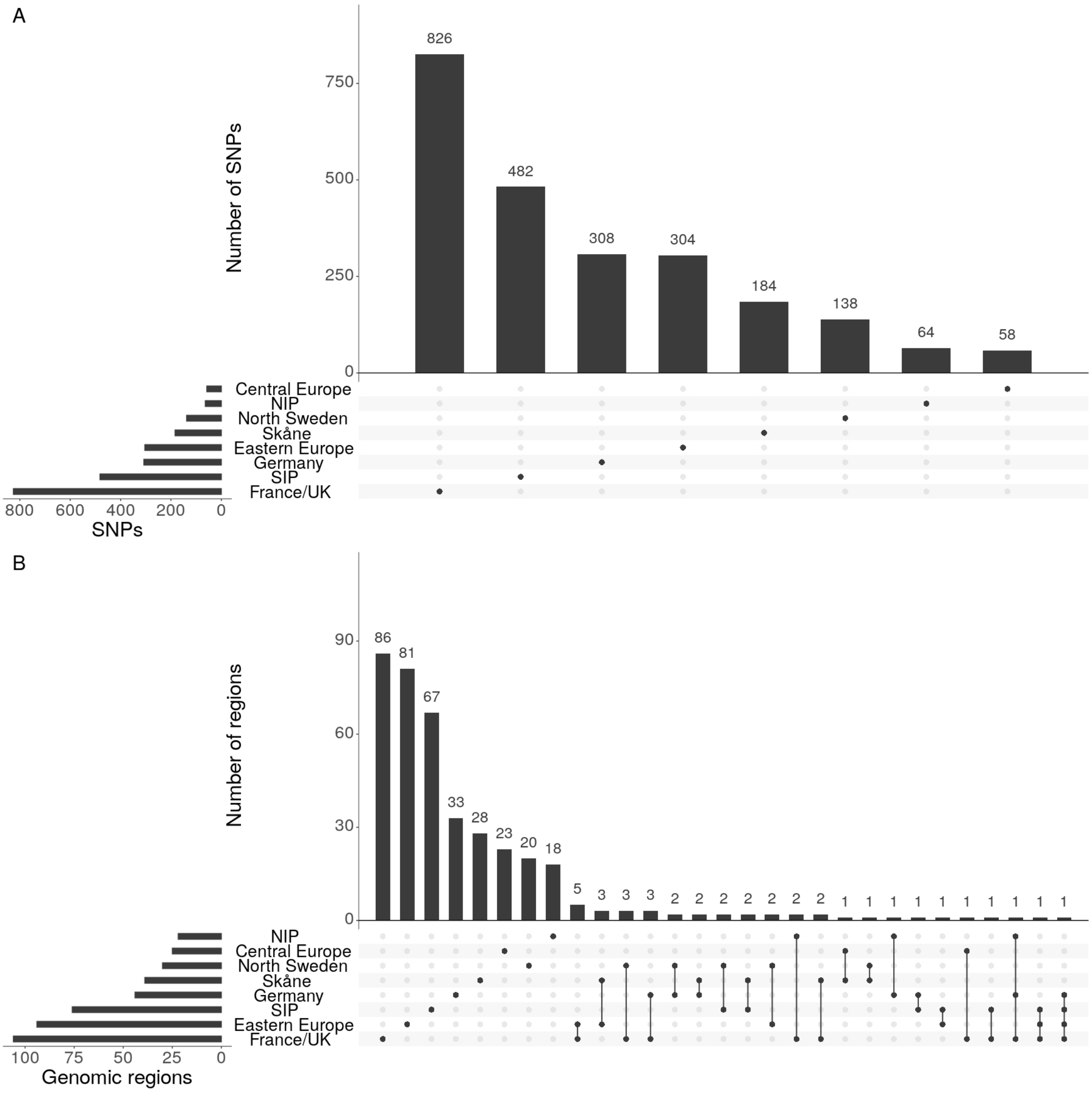


Figure 3.8: Sharing of sub-significant ($p < 10^{-4}$) associations for permuted phenotypes. (A) Histogram of the number of associated SNPs in each subpopulation and shared between pairs of subpopulations. (B) Histogram of the number of associated regions in each subpopulation and shared between pairs of subpopulations.

The implementation of this strategy leaded to a slightly increased number of shared genomic regions between subsamples, but more remarkable was the fact that shared genomic regions were not only found being shared between two subsamples (as was the case for the above mentioned shared associations), but also between three and up to seven subsamples. However from a overall view, the trend of most genomic regions being unique in a subsample remained (Figure 3.7 B). As in the comparison at single marker level, shared genomic regions between at least four subsamples contained genes previously associated with flowering time (Table 3.4).

Table 3.4: Overlap of candidate genes with shared genomic regions

Region[a]	1:(26151612–26570642)	4:(192421–572878)	5:(22786643 23605491)
Candidate gene[b]	*CDF5*[c]	*FRI*[d]*LIF2*[e]*MED12*[f]	*COL5*[g]*MSI1*[h]*VIN3*[i]*ZTL*[j]
SIP	0	0	1
NIP	1	1	1
Germany	1	1	1
France/UK	1	1	1
Central Europe	1	0	1
Skåne	0	1	0
North Sweden	1	1	1
East Europe	1	0	1

[a] represented as "chromosome:(start–stop)

[b] known flowering time gene within the detected region using a list of 306 flowering time genes from Bouché et al. 2016

[c] *CDF5 (CYCLING DOF FACTOR 5)*, Fornara et al. 2009

[d] *FRI (FRIGIDA)*, Stinchcombe et al. 2004

[e] *LIF2 (LHP1-INTERACTING FACTOR 2)*, Latrasse et al. 2011

[f] *MED12 (MEDIATOR 12)*, Imura et al. 2012

[g] *COL5 (CONSTANS-LIKE 5)*, Hassidim et al. 2009

[h] *MSI (MULTICOPY SUPRESSOR OF IRA1)*, Bouveret et al. 2006

[i] *VIN3 (VERNALIZATION INSENSITIVE 3)*, Sung and Amasino 2004

[j] *ZTL (ZEITLUPE)*, W.-Y. Kim et al. 2007

Again, the number of observed shared regions was higher than expected only by chance (Figure 3.8 B), which supports the thought that these shared regions were not originated by false associations. The detected overlaps highlight the amount of genetic heterogeneity within different loci, where different alleles of the same gene are present in different local subsamples. The finding of this type of region containing the gene *FRI* (*FRIGIDA*) is in agreement with previous reports where the effect of different natural alleles on flowering time was already extensively studied (P. Li et al., 2014; L. Zhang and Jiménez-Gómez, 2020). Beyond *FRI*, this seems to be truth for the other genomic regions as well.

The comparison of shared genomic regions has the advantage of not being restricted to the exact position of the associated SNP. Continuing with this strategy,

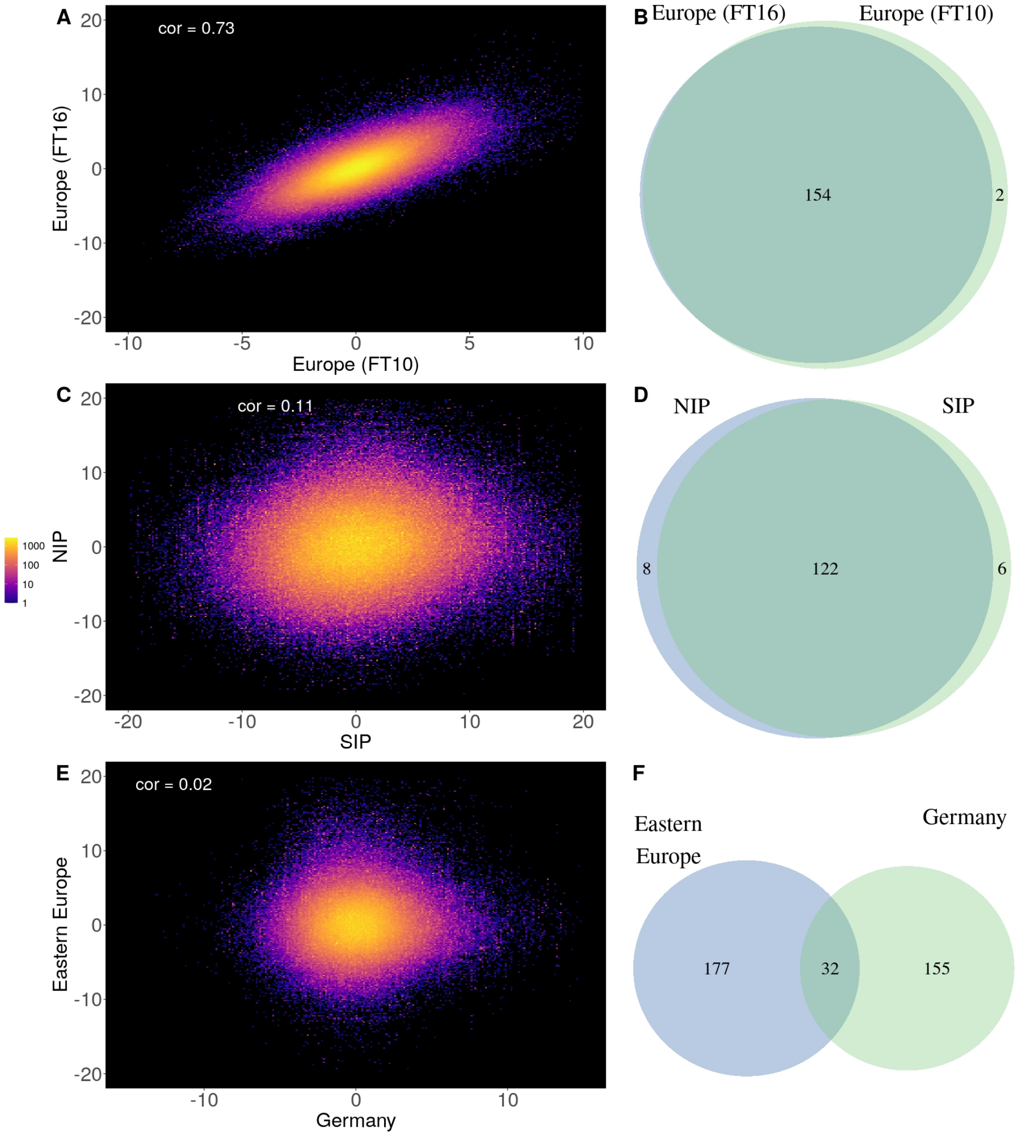


Figure 3.9: Density plots comparing effect sizes estimates and Venn diagrams showing estimated polygenic overlap using *MiXeR*. (A-B) The comparison between flowering time at 10°C and 16 °C in the complete European population. (C-D) Comparison of FT10 between the Northern Iberian Peninsula (NIP) and Southern Iberian Peninsula (SIP) subsamples. (E-F) Comparison of FT10 between the Eastern Europe and German subsamples.

MIXeR was implemented to estimate shared causal variants between subsamples, since this statistical tool overcomes the intrinsic problem of detecting the exact locations of such associations (Frei et al., 2019). Initially, the proportion of shared causal variants between to highly correlated traits (flowering time at 10 °C and 16 °C) in the European sample was estimated. As expected most of the estimated causal variants were shared between these traits with both a high genetic and effect size correlation (Figure 3.9 A, Table 3.5).

Table 3.5: Shared causal variants estimation using *MIXeR*

Subsample 1	Subsample 2	s1_s2[a]	cv_1[b]	cv_2[c]	rho1_2[d]	rg[e]
FT16	FT10	154	0	2	0.95	0.94
Europe	SIP[f]	45	131	88	0.75	0.22
Europe	NIP[g]	29	133	95	0.39	0.08
Europe	Germany	19	12	231	0.03	0.01
Europe	North Sweden	35	70	4	0.55	0.3
Europe	Eastern Europe	103	59	111	0.55	0.3
SIP	NIP	122	6	8	0.81	0.77
SIP	Germany	29	107	185	0.42	0.07
SIP	France + UK	1	178	3	-0.02	-0.01
SIP	North Sweden	26	114	11	-0.26	-0.09
SIP	Eastern Europe	52	87	194	-0.20	-0.06
NIP	Germany	116	11	81	-0.85	-0.62
NIP	France UK	1	135	1	0.79	0.06
NIP	North Sweden	29	105	6	-0.13	-0.05
NIP	Eastern Europe	76	56	153	0.31	0.14
Germany	France UK	14	318	1	0.89	0.17
Germany	North Sweden	31	164	2	0.90	0.35
Germany	Eastern Europe	32	155	177	-0.59	-0.10
France UK	North Sweden	27	113	17	-0.01	-0.01
France UK	Eastern Europe	1	0	238	0.99	0.07
North Sweden	Eastern Europe	26	7	195	0.58	0.18

[a] s1_s2: Estimated number of shared causal variants

[b] cv_1: Estimated causal variants only present in subsample 1

[c] cv_2: Estimated causal variants only present in subsample 2

[d] rho1_2: Estimated effect size correlation between shared causal variants

[e] rg: Estimated genetic correlation

[f] SIP: Southern Iberian Peninsula

[g] NIP: Norhtern Iberian Peninsula

This high degree of shared effects is consistent with the fact that the effect of major markers affecting flowering time is stable (Srikanth and Schmid, 2011).

On the other hand, the reduced number of shared causal variants between geographically distant subsamples indicates the importance of markers with smaller effects in different geographic regions. Besides the geographic location, the genetic background plays an important role in defining the effect of these markers too. When comparing the correlation of shared causal variants between SIP and the remaining subsamples it is possible to recognized how the correlations of effect sizes decreased when the subsamples are more distant (Table 3.5). This indicates a high genetic heterogeneity within subsamples and highlights that markers can have a different or even opposite effect depending on the geographic location and genetic background of the analyzed populations. In summary, this analysis confirms the GWAS results shown before, without explicitly looking at distinct markers.

All the aforementioned results are in agreement with the findings presented in Chapter 2, but at a more complex level since these subsamples are defined using geographically restricted areas, which increases the probability of detecting markers with local effect. The complex architecture of flowering time and its suggested genetic heterogeneity leaded to high dissimilar association between European subsamples, reinforcing the idea that replicating genome-wide associations in different samples when analyzing adaptive traits is unlikely or indeed unexpected. This scenario would be more likely when analyzing traits, in which only a few markers explain most of the phenotypic variation and are not under local adaptation.

3.3.2 *GWAS on further adaptive traits*

Assuming that the above described phenomenon is not restricted to flowering time but detectable in different adaptive traits, we carried out the same analyses of performing GWAS in distinct local subsamples for two additional traits, stomata size and cauline leaf number. The adaptive importance of stomatal traits for the fine-tuning of water-use efficiency has also been suggested previously (Dittberner et al. 2018). In addition, the number of cauline leaves is linked to photomorphogenesis (Pouteau and Albertini 2009) and therefore it could be under local adaptation too and being optimized by distinct genes or pathways in different subsamples. In this case, two geographically far apart subsamples, namely Iberian Peninsula (IP) and Sweden (SW), which overlap with those used for flowering time, were utilized (Figure 3.10 A). For both traits, only slight phenotypic differences were detected (Figure 3.10 B).

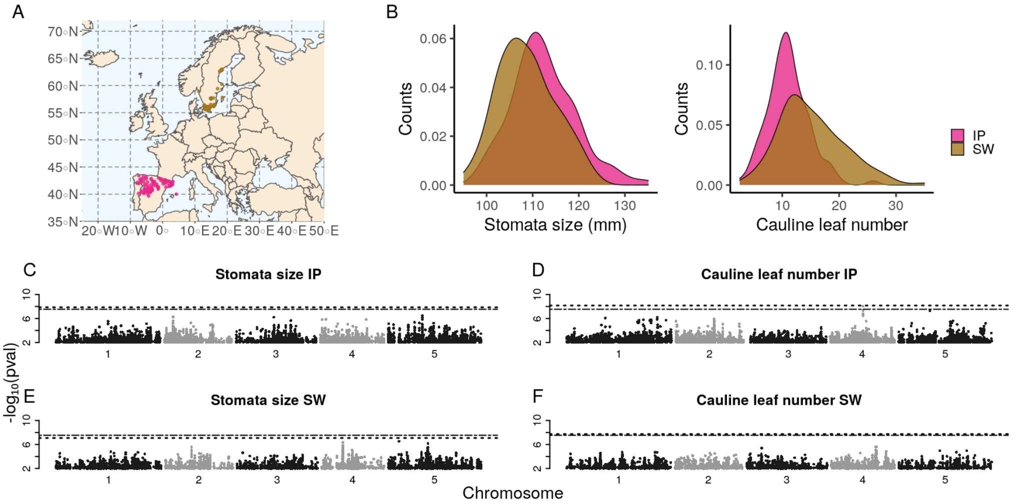


Figure 3.10: Analyses of stomata size and cauline leaf number. (A) Geographic distribution of the used 240 *Arabidopsis thaliana* ecotypes. (B) Phenotypic distribution of stomata size and cauline leaf number in the Iberian (IP) and Scandinavian (SW) subsamples. (C-F) Manhattan plots of GWAS results from the different subsamples. Dashed lines and dash-dotted lines indicate permutation-based threshold and Bonferroni threshold, respectively.

Similar as the GWAS results for flowering time, manhattan plots of GWAS results for these subsamples show dissimilar association peaks for both traits. Despite the fact that heritability estimates and theoretical statistical power to detect major polymorphisms suggested that these subsamples were suitable for GWAS (Table 3.6), only one significant association was detected (Figure 3.10 C-F). Again, presuming that poylmorphisms with middle-to-low effect could be undetectable, SNP overlap at a less stringent pvalue (10^{-4}) was estimated. In this case, there was no overlap between subsamples for both traits. This repetitive scenario suggests that replication of GWAS results in different samples using traits with complex genetic architecture or even involved in local adaptation, as is the case for adaptive traits, is unlikely. Additionally, it has to be noted that the fact of GWAS being under-power to detect causal variants suggests that both traits might not be regulated for major polymorphisms but for a combination of variants with middle-to-low effect.

Table 3.6: Geographic limits, pseudo-heritability and power estimation of the IP and SW subpopulation used for the analyses of stomata size (ST) and cauline leaf number (CL).

Trait	Source	No. accessions	min_lat	max_lat	min_lon	max_lon	$\widehat{h_2}$	Power
ST	ALL	240	39.66	63.02	-7.80	18.52	0.56	1.00
	IP	109	39.66	43.40	-7.80	4.25	0.27	0.81
	SW	131	55.38	63.02	11.2	18.52	0.54	0.99
CL	ALL	240	39.66	63.02	-7.80	18.52	0.78	1.00
	IP	109	39.66	43.40	-7.80	4.25	0.9	1.00
	SW	131	55.38	63.02	11.2	18.52	0.85	1.00

3.3.3 *Can Polymorphisms with global effects be detected regardless of the population ?*

With the aim of verify if major polymorphisms can be detected with the aforementioned subsamples (IP and SW) used for stomata size and cauline leaf number, phenotypes under control of major polymorphisms were simulated. GWAS were performed on a total of 27000 simulated phenotypes and power to detected the simulated polymorphism as well as the proportion of false positives based on Bonferroni correction were estimated. Results presented in Table 3.7 support the thought that these subsamples have power enough to detect major polymorphisms.

More exactly, global polymorphisms explaining 20% of the phenotypic variation were detected in 96.4% of the cases in the whole sample and as expected in a lower proportion in the subsamples due to reduced power produced by a smaller sample size (Korte and Farlow, 2013). However a high false discovery rate was detected. Because the number of simulations made it impracticable to calculate permutation based threshold (it would have been 27 million GWAS runs), this false discovery rate was based on Bonferroni correction. Due to this high discovery rate, a more stringent significance threshold of 10^{-9} (just below the Bonferroni correction $2.7x10-8$) were considered. This slight reduction in the significance threshold resulted in a dramatic decrease of the false discovery rate, while the empirical power still remained high (Table 3.8).

For the scenario where major polymorphisms with local effect were simulated a reduced empirical power was observed, however more importantly, these simulated polymorphisms were detected only in the respective subsample for

Table 3.7: Overview of the sensitivity of the different simulation scenarios.

Variance explained		Sensitivity		
		ALL	IP	SW
20%	ALL	0.964	0.264	0.390
	IP	0.027	0.274	0.000
	SW	0.065	0.000	0.420
15%	ALL	0.265	0.130	0.250
	IP	0.050	0.000	0.230
	SW	0.020	0.140	0.000
10%	ALL	0.170	0.000	0.20
	IP	0.010	0.000	0.020
	SW	0.000	0.000	0.000

Table 3.8: Summarized GWAS results from simulated data. True positives (TP), false positives (FP) and false discovery rate (FDR) are reported for 1,000 simulations per scenario with GWAS performed either in all 240 accessions (ALL) or only in the IP or SW subpopulation.

Variance explained		TP[a]			FDR[a]			TP[b]			FDR[b]		
		ALL	IP	SW	ALL	IP	SW	All	IP	SW	All	IP	SW
20%	ALL	964	264	390	0.19	0.22	0.27	876	118	217	0.08	0.01	0.03
	IP	27	274	0	0.14	0.07	0.04	5	101	0	0.02	0.02	0.00
	SW	65	0	420	0.19	0.00	0.15	20	0	226	0.04	0.00	0.03

[a] Bonferroni correction

[b] $\alpha = 10e - 9$

which the local effect was simulated (Table 3.7). Figure 3.11 presents GWAS results for the three scenarios in all samples. Simulated phenotypes with a major global effect generated repetitive GWAS results only differing in a higher pvalue of the causal variant in the subsamples, which constitute a clear example of reduced power (but still significant at Bonferroni correction) (Figure 3.11). These results suggests that replication of GWAS results in different samples for traits under such type of regulation would be expected. In contrast to these results, divergent GWAS results were observed for simulated phenotypes with a major local effect (Figure 3.11). Unlike to the above mentioned trend, differences in GWAS results can neither be explain through statistical power nor through allele frequency. All the describe results are based on simulated phenotypes with a causal variant explaining 20% of the phenotypic variance. Results for the remaining simulations show how the empirical power goes sharply down when causal variants explain only a low proportion of the phenotypic variance, which is in agreement with the results observed for stomata size and cauline leaf number, where a presumed polygenic regulation (with a lot of causal variant explaining only a marginal proportion of the phenotypic variation) might have produced under-powered GWAS results. All the results presented indicate that GWAS in these subsamples have enough power to identify major polymorphisms with both global and specific local effects, in addition to reinforce the presumed local adaption for flowering time.

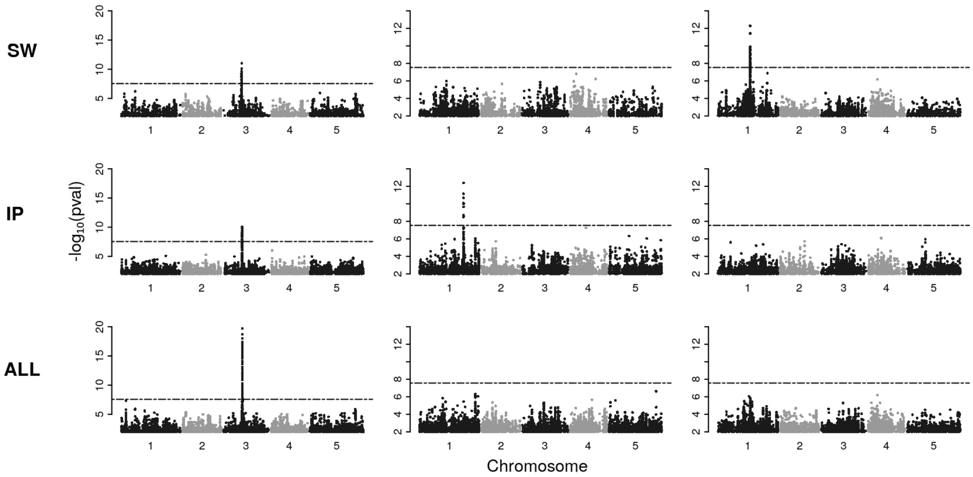


Figure 3.11: Manhattan plots of GWAS results from three different simulations. The causative markers were simulated to have an effect in all accessions (left panel), only in IP (middle panel) or only in SW (right panel). The respective population for GWAS are displayed in the different rows, where for the results in the top row, the SW subpopulation has been used, the IP subpopulation has been used to generate the results in the middle row and the bottom row displays the results in the merged population of 240 accessions. Dashed lines indicate the Bonferroni threshold used in the simulations.

3.3.4 *Detection of global and local regulation patterns via eGWAS*

After having successfully simulated phenotypes regulated either by major global or local effects, the next step was to find real traits under these regulation patterns. However, finding phenotypic data for such traits is not an easy task, even for *A. thaliana*. First, it would be necessary to find phenotypes whose variation is mainly explained by few or even a single common causal variant. But most of the data available on AraPheno consists of phenotypes presumably under polygenic regulation or even involved in local adaptation, as in the case of the phenotypes already analyzed. Nevertheless, expression data (Kawakatsu et al., 2016) available on AraPheno (https://arapheno.1001genomes.org/study/52/) emerged as a candidate data set to reproduce patterns of global and local regulation. It would not be wrong to hypothesize that expression levels of more "structural" genes should be consistent between accessions regardless of their geographic origin, moreover one might predict that major polymorphisms directly located in the gene would be responsible for its regulation (Signor and Nuzhdin, 2018). On the other hand, genes involved in pathways finely tuned through environmental cues or even implicated in secondary metabolism, might tend to show a more local and polygenic regulation.

Table 3.9: Overview of RNA expression data. This table shows the filters applied to the whole data. GWAS were performed on 2,483 genes.

Filter	Sample size
Total RNA expression data	24,175
Nuclear genes	23,021
$\widehat{h_2} > 0.5$	4,873
$\widehat{h_2} > 0.5$ & power > 0.9	2,483
$\widehat{h_2} > 0.5$ & power > 0.9 and not inflated	1,982

Grounded on these hypotheses, eGWAS were carried out on 2483 molecular traits for the Iberian, the Scandinavian and the whole sample (Table 3.9). This number represents about 10% of the total of genes for which expression data is available (Kawakatsu et al., 2016), and was obtain after filtering out genes with low estimated heritability and insufficient statistical power to detected major effects (Table 3.9). Significant genome-wide associations, at least in one of the subsample, were detected for 780 genes at a significance threshold of 10^{-10}. Evidence for global and local regulation patterns were found in 282 genes and typical GWAS results are exemplify in Figure 3.12. The tendency of these manhattan plots is completely comparable with that obtain from GWAS results of simulated phenotypes (Figure 3.11). These results confirm the occurrence of global and local genetic architecture in real traits, in addition that these samples are powered to detect major polymorphisms.

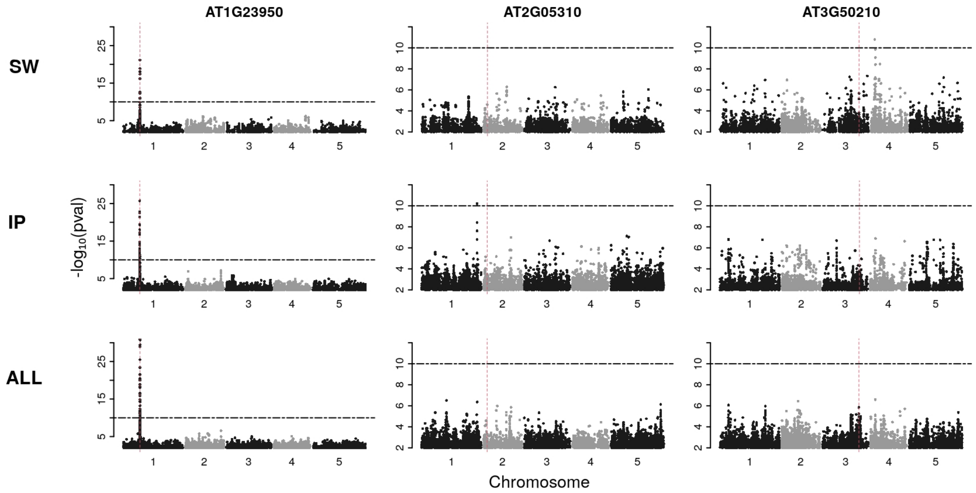


Figure 3.12: Manhattan plots from GWAS of expression levels for three different genes. The Columns show the results from genes representing different scenarios. The rows display the GWAS results of the analysis in the two subpopulations (SW and IP, respectively), or in the merged population (ALL). Horizontal dash-dotted lines indicate the significance threshold of $p < 10^{-10}$. Vertical dashed lines show the position of the gene whose expression is being used as a molecular phenotype.

As shown in Figure 3.12, some of the associated markers are directly found in (or close to) the gene whose expression level was used as phenotype. This kind of regulation is known as *cis*. Conversely, *trans* regulation refers to genes whose expression levels are regulated by markers that are not located in close proximity to the respective gene. A pattern of *cis* and *trans*-regulation for the above mentioned genes was found. Almost all associations (about 99%) shared between IP and SW presented *cis*-regulation, whereas mainly *trans*-association were detected for unique associations both in IP and SW (Figure 3.13). Assuming that globally regulated genes under a *cis*-regulation pattern should be persistent regardless of the subsamples to be compare and, in contrast, that local regulated genes under trans-regulation should be highly dependent of the subsamples to be consider, equally-sized (91 accessions as IP and 74 accessions as SW) random subsamples (by randomizing the subsample lables) were produced. eGWAS were carried out on these random subsamples and the resulting regulation patterns were compared (Figure 3.14). As expected, shared association were mostly under *cis*-regulation, with almost 80% of the shared genes between IP and SW being recovered between these random subsamples. In a similar way, non-shared associations were predominantly *trans*-regulated, and in a considerably smaller number than the detected in IP and SW.

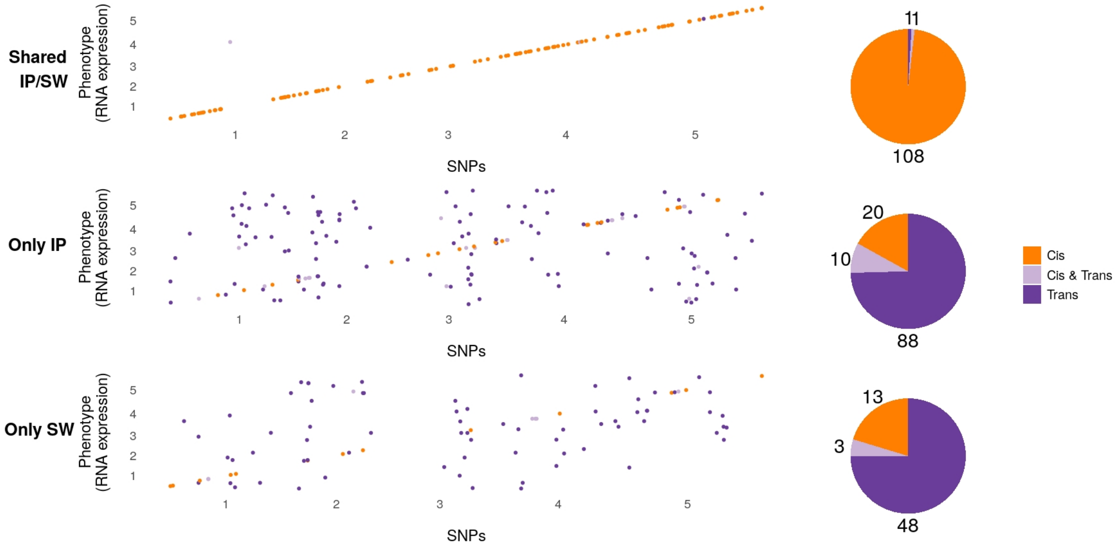


Figure 3.13: Summary of the difference between shared and non-shared GWAS results for expression data. Plots show the position (x-axis) of significant associations for each expressed gene (y-axis). Associations shared between subpopulations (top panel) are almost all in *cis*, whereas associations specific to one subpopulation (bottom panels) are mostly found in *trans*. Pie-charts show the number of genes in each category.

Moreover, GO enrichment analysis using the shared genes between IP and SW detected significant enrichment for molecular functions associated with primary metabolism (Table 3.10). Additionally, highly conserved genes as those involved in glycolysis (*G6PD4* (NADP-dependent glucose-6-phosphate dehydrogenase)), tricarboxylic acid cycle (*SDH3-2* (succinate dehydrogenase)) and the shikimato pathway (*MEE32*) were detected. In the same way, genes required for growth and development in *A. thaliana* were found too (EMB2739 (EMBRYO DEFECTIVE 2739) and EMB3127 (EMBRYO DEFECTIVE 3127) (D. W. Meinke, 2020)). Added to these genes, genes like RPS5 (*RESISTANT TO P. SYRINGAE 5*), which is linked to bacterial and downy mildew resistance (Warren et al. 1998), and which is likely to be under global balancing selection (Tian et al. 2002) were also detected (a complete list can be found in: `https://www.biorxiv.org/content/10.1101/2021.02.26.433043v1`). In contrast, the group of genes under presumably local adaptation are related for example to flowering time (*AGL-20* (AGAMOUS-LIKE 20) (H. Lee et al., 2000)), stress response (*RCAR5/PYL11* (REGULATORY

COMPONENT OF ABA RECEPTOR 5/ PYRABACTIN RESISTANCE-LIKE 11) (Lim and S. C. Lee, 2020), and *HDA9* (HISTONE DEACETYLASE 9) (Zheng et al., 2016)), secondary metabolism (ST4B (brassinolide sulfotransferase)(Hashiguchi et al., 2014)) and salt tolerance (ALDH10A8 (ALDEHYDE DEHYDROGENASE) (Jacques et al., 2020)) (a complete list can be found in: `https://www.biorxiv.org/content/10.1101/2021.02.26.433043v1`).

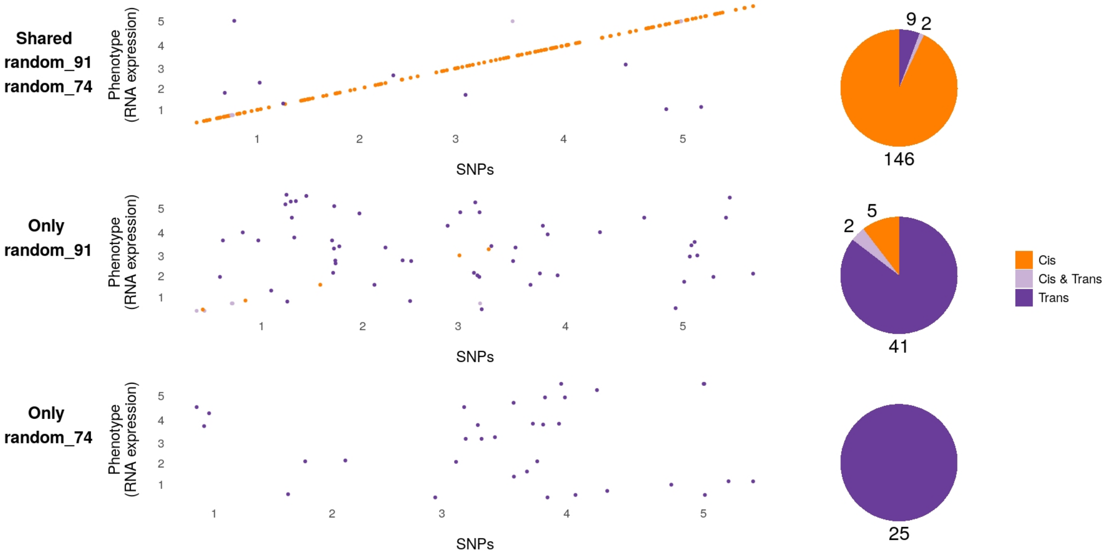


Figure 3.14: Summarized GWAS results for the analyses of RNA expression data in *A. thaliana* in random subpopulations. Genes are grouped in three categories: 1) Shared random_91/random_74, where the same association for a gene is recapitulated in the GWAS of both subpopulations. 2) Only random_91, where a significant association is only found using the random subpopulation containing 91 accessions. 3) Only random_74, where a significant association is only found using the random subpopulation containing 74 accessions. Scatter plots show the genomic location of the respective associated markers per gene for each class, where *cis*-regulatory variants are colored in orange, while variants in *trans* are shown in purple. Pie charts display the amount of genes per class that have *cis*, *cis* and *trans* or only *trans*-associations.

Our finding of more structural genes being under *cis*-regulation is in agreement with multiple studies (A. Martin and Orgogozo, 2013; Romero, Ruvinsky, and Gilad, 2012). Changes in this regulation pattern should be less pleiotropic, since these genes do not directly affect the expression of other genes (as would be the case of transcription factors for example), but rather their own expression level (Prud'homme, Gompel, and Carroll, 2007). Additionally, the fact to having found most *cis*-regulation in both subsamples supports the idea that *cis*-regulated loci are less affected by genetic background, moreover the expected higher additivity of *cis*-regulated loci makes it easier to detect them by means of GWAS (Lemos et al., 2008). On the other hand, Clauw et al. 2016 found that *cis*-regulatory variants had the same effect in different environments, suggesting that they could only be involved in local adaptation through highly polygenic allele-frequency changes.

Table 3.10: GO enrichment analysis using shared genes between the Iberian and Scandinavian subsamples.

Ontology	ID	Description	p.adjust	qvalue
BP	GO:0006730	one-carbon metabolic process	0.04	0.04
BP	GO:0035999	tetrahydrofolate interconversion	0.19	0.18
MF	GO:0043531	ADP binding	0.00	0.00
MF	GO:0032559	adenyl ribonucleotide binding	0.00	0.00
MF	GO:0016616	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	0.03	0.02
MF	GO:0019238	cyclohydrolase activity	0.03	0.03
MF	GO:0016614	oxidoreductase activity, acting on CH-OH group of donors	0.03	0.03
MF	GO:0016646	oxidoreductase activity, acting on the CH-NH group of donors, NAD or NADP as acceptor	0.04	0.03
MF	GO:0050661	NADP binding	0.05	0.04

In contrast to *cis*-regulatory patterns, *trans*-regulations tend to explain a less proportion of the variance (Lemos et al., 2008). This makes their detection through GWAS more challenging. However we were able to detect such a regulation pattern using two geographically wide separated subsamples. The important role of *trans*-regulation in local adaptation has been already reported (Clauw et al., 2016). Having found these type of regulation majorly as unique depending of the subsample is consistent with the idea that this type of regulation involve a more polygenic architecture (Signor and Nuzhdin, 2018), which is more affected by genetic background.

The presented results supports the idea that replicating GWAS results in different samples as a signal of reliability has to be reconsidered. In fact, our results suggest that when considering adaptive traits, GWAS results can extremely vary between samples, mostly due to local adaptation. In General, more widely distributed samples might uncover associations with a more global effect, but neglect associations affecting the phenotype more locally, and the contrary is true when considering geographically more restricted samples. In both cases conclusions should reflect the nature of the samples used. According to the differences observed between the Iberian subsamples, a strikingly different profile of associations might be obtain even when comparing geographically close samples. Additionally, we could exemplify the effect of the genetic background on a significant association when comparing to geographically distant samples (5 – 18590501). Finally, we were able to detect *cis* and *trans* regulation via eGWAS. These regulation patterns followed a very specific trend with shared associations being mostly in *cis* and unique associations being mostly in *trans*. It would be interesting to test whether genes under *cis*-regulation present different epigenetic regulation between subsamples, or even whether these genes could be under a combination of *cis*-trans regulation (associations in trans with a very low effect, might not have been detected in our study).

4

RNA-SEQ DATA ANALYSIS

4.1 INTRODUCTION

The content presented in this chapter is the result of a collaboration with the group of Prof. Dr. Wolfgang Dröge-Laser from the department of pharmaceutical biology of the university of Würzburg. This introduction describes the biological question underlying the analysis of the RNA sequencing (RNAseq), which I performed. Although my part was restricted to the analysis of the data, I highlight why the obtained results are relevant to understand host-pathogen interaction. These results are published in: Fröschel, Christian, Jaqueline Komorek, Agnès Attard, Alexander Marsell, **Lopez-Arboleda, William.A**, Joëlle Le Berre, et al. (2021). "Plant roots employ cell-layer-specific programs to respond to pathogenic and beneficial microbes." In: *Cell Host & Microbe* 29.2, pp. 299–310. DOI: https://doi.org/10.1016/j.chom.2020.11.014.

Roots play multiple functions in plants, as anchoring, storage and for the uptake of water and nutrients. Roots face both biotic and abiotic stress (Eshel and Beeckman, 2013). A part of the biotic stresses is caused by the interaction with pathogens. Some of the most abundant pathogens in soil are fungi. Fungi display a broad spectrum of lifestyles including biotrophic (interacting with plants and utilizing their living tissue), necrotrophic (killing plants and feeding on the resulting dead tissue) and saprophytic (surviving and feeding on dead plants). According to the different fungal lifestyles, fungal pathogens produce diverse virulence factors, which could be involved in host penetration, suppression of host defense and nutrient acquisition (Doehlemann et al., 2017). In the same way, in the presence of a pathogen plants produce an immune reaction in response to the infection (Ryan et al., 2016). Depending on the pathogen's lifestyle, the colonization of roots could take place at different cell-layers, and allows to track specific modifications in gene expression. Since roots are build of concentric cell layers performing specific functions, it could be hypothesized that these layers differ in their responses towards microorganisms.

In order to prove such a hypothesis, three fungal species (more precisely one fungus-like organism and two fungal species) displaying different lifestyles were used to inoculate *Arabidopsis thaliana* roots:

- *Verticillium longisporum* (Ascomycota): *V. longisporum* is a soil-borne vascular pathogen whose infection produce symptoms like wilting, chlorosis, vascular discoloration and early senescence. The life cycle of *V. longisporum* begins as microsclerotia, which are resistant structures waiting for a new host (sclerotia can survive up to 10 years in soil). Microsclerotia germination is stimulated through exudates from plants, resulting in hyphae growth and colonization of the root surface. As a result, hyphae growth across all root cell layers until reaching the xylem. At this stage, *V. longisporum* acts as an endophytic-biotrophic fungus, changing to a more saprophytic style with microsclerotia formation due to plant senescence (Depotter et al., 2016)
- *Phytophthora parasitica* (Ooomycota): *P. parasita* is a soil-borne pathogen which causes root and stem rot in over 70 species. Zoospores play a major role during infection reaching plant surfaces and becoming immobile cyst to subsequently germinate. After root colonization, by means of the secretion of a range of degradative enzymes that break down physical barriers to infection, and an initial biotrophic state, *P. parasitica* switches to a necrotrophic state resulting in severe plant damage (Y. Meng et al., 2014). The *Arabidopsis-Phytophtora* pathosystem has been extensively described. In this species, infection process occurs in a similar way as in the natural host, however disease severity can vary depending on *A. thaliana* ecotypes and *P. parasita* strains, suggesting variation in host specificity (Y. Wang et al., 2011).
- *Serendipita indica* (Basidiomycota): *S. indica* is an endophytic, mutualistic species which can act as growth promoter, immune modulator, phytoremediator among others. The root colonization takes place after a biotrophic growth phase, and develops in a cell dead-dependent phase. This cell dead-dependent phase is a result of the suppression of the root innate immune system and the induction of endoplasmic reticulum stress, whose adaptive pathway response is at the same time inhibited. This root colonization is limited to the rhizodermis and cortex, and rarely extended at the root meristematic and elongation zones (Qiang et al., 2012).

In order to compared gene expression across cell layers after the inoculation with the different pathogens, cell-layer-specific translatomes were obtain (described as an infection-based TRAP-seq). Briefly, TRAP-seq (translating ribosome affinity purification followed by RNA sequencing) allows the isolation of ribosome-mRNA complexes via immunoprecipitaion, resulting in a subsequent data set of expression at a cell-layer resolution (Mustroph et al., 2009; Sorenson and Bailey-Serres, 2015). In contrast to the nuclear RNAseq, TRAP-seq separates the actually translating mRNA from the total mRNA. The cell-layer resolution is achieved by using transgenic lines expressing FLAG-tagged ribosomal proteins. It is expected that the divergent lifestyles of pathogens might trigger differential profile expressions at root cell layer. Therefore, expression across cell layers between mock and infected plants was compared. Through these comparisons, clusters of genes according to pathogen lifestyle were found. In summary, after the expression data analysis, reported plant responses were confirmed and hypotheses related to endodermis as a barrier against pathogen invasion could be tested at gene expression level and later confirmed via fluorescence confocal microscopy and *A. thaliana* mutant infection (these later confirmation was motivated by the results from the expression analysis and was carried out by Christian Fröschel, department of Pharmaceutical Biology, University of Würzburg). .

4.2 METHODOLOGY

4.2.1 *Transcript quantification*

To experimentally detect differential expression, Transcripts Per Million (TPM) were compared between mock and infected plants for each cell layer at the presence of each pathogen. In average 10-20 million reads were yielded after RNAseq for each sample. Reads coming from each samples were mapped to the reference genome of *A. thaliana* (TAIR 10 genome release) using Salmon v0.7.2 (Patro et al., 2017). Only reads, which could be unambiguously mapped to the reference genome were used to estimate the TPM. To do so, the transcriptome of *A. thalina* was downloaded and indexed using the flag *index* of salmon tools. After the index was built, quantification of samples for each pathogen was carried out using a bash script (a detailed description can be found in Appendix B).

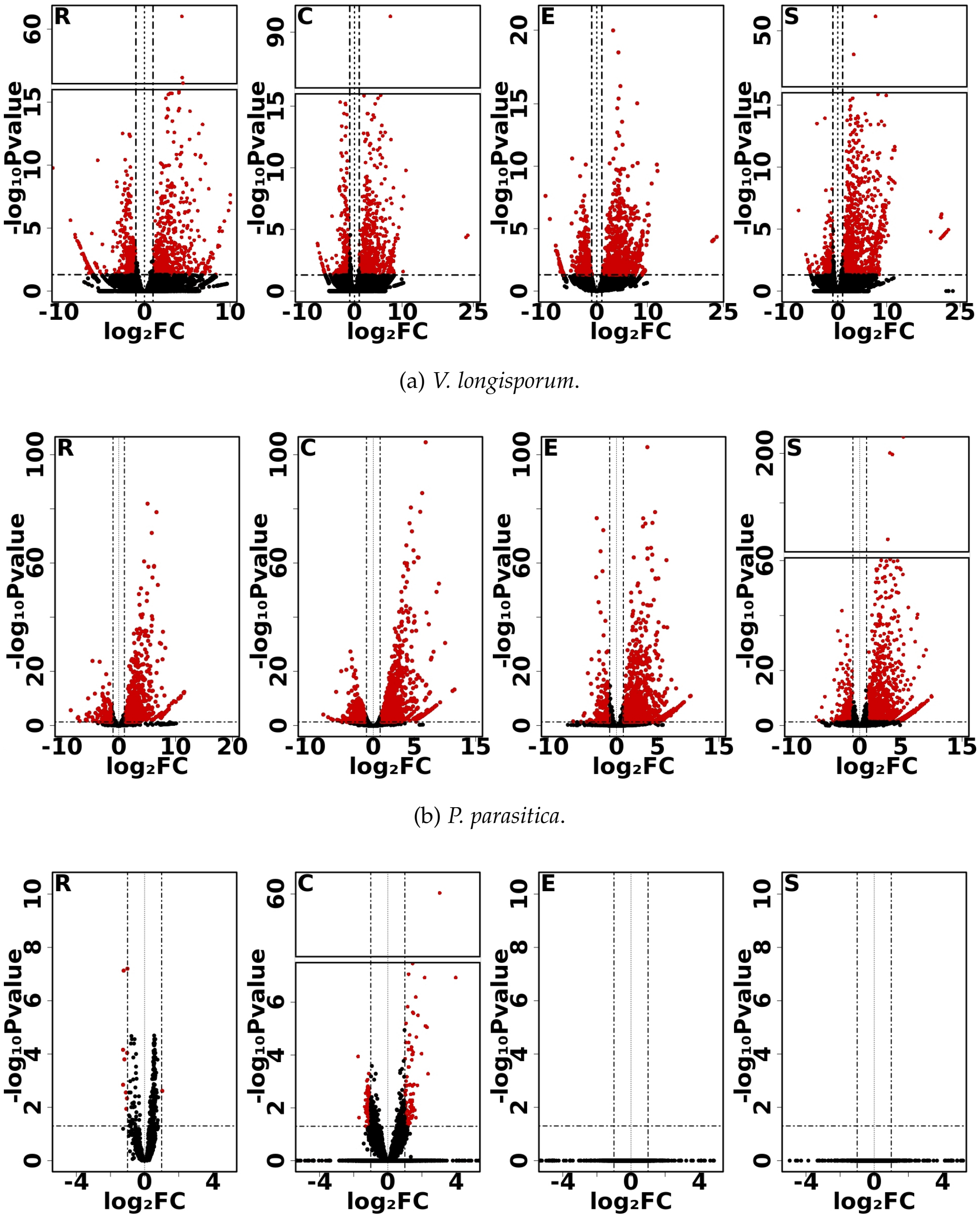
4.2.2 *Differential expression analysis*

The resulting data frame containing TPM estimation for all genes across samples was used to assess gene differential expression between mock and infection treatments for each pathogen in each cell type. In order to do so, columns were filtered according to pathogen and cell type, and the resulting data frame was formatted applying the `DESeqDataSetFromMatrix` function and before to being passed to the `DESeq` function (both from the `DESeq2` package (Love, Huber, and Anders, 2014)), all genes with counts lower than 10 were filtered out. A gene was considered as differentially expressed if both the pvalue based on *independent hypothesis weighting* (Ignatiadis et al., 2016) was lower than 0.05 and the absolute value of $log_2FoldChange$ was bigger than 1. After applying this workflow for all possible mock-infected comparisons, volcano plots (Figure 4.1) were created to visualize the magnitude and distribution of expressed genes. To better compare up-und-down regulated genes either between pathogen in each cell layer or across cell layers of plants infected by each pathogen Venn diagrams (using custom functions of the `VennDiagram` package (Chen and Boutros, 2011)) were generated. For an overall visualization of the differentially expressed genes, a heatmap was created. This heatmap represented the total number of diffentially expressed genes across cell layer for each pathogen. Next, Hierarchical cluster (applying functions from the `pheatmap` package (R. Kolde and M. R. Kolde, 2015)) was implemented in order to find patterns of common or unique responses of these genes. Finally, Gene Ontology (GO) and KEGG (Kyoto Encyclopedia of Genes and Genomes) enrichment analysis were carried out (using the clusterProfiler package).

4.3 RESULTS AND DISCUSSION

4.3.1 *Comparison of differentially expressed genes*

The magnitude of differential expression was initially visualized using volcano plots. On these plots, differentially expressed genes are colored to separate them from no differentially expressed ones. Volcano plots showed a major proportion of up-regulated genes in all cell layers ($log_2FoldChange > 1$) under treatment with *V. longisporum* as well as under treatment with *P. parasitica* (Figure 4.1 A-B).



(a) *V. longisporum*.

(b) *P. parasitica*.

(c) *S. indica*.

Figure 4.1: Volcano plots highlighting differentially expressed genes (DEGs) in each root cell layer in the presence of three fungal (one fungus-like) species.

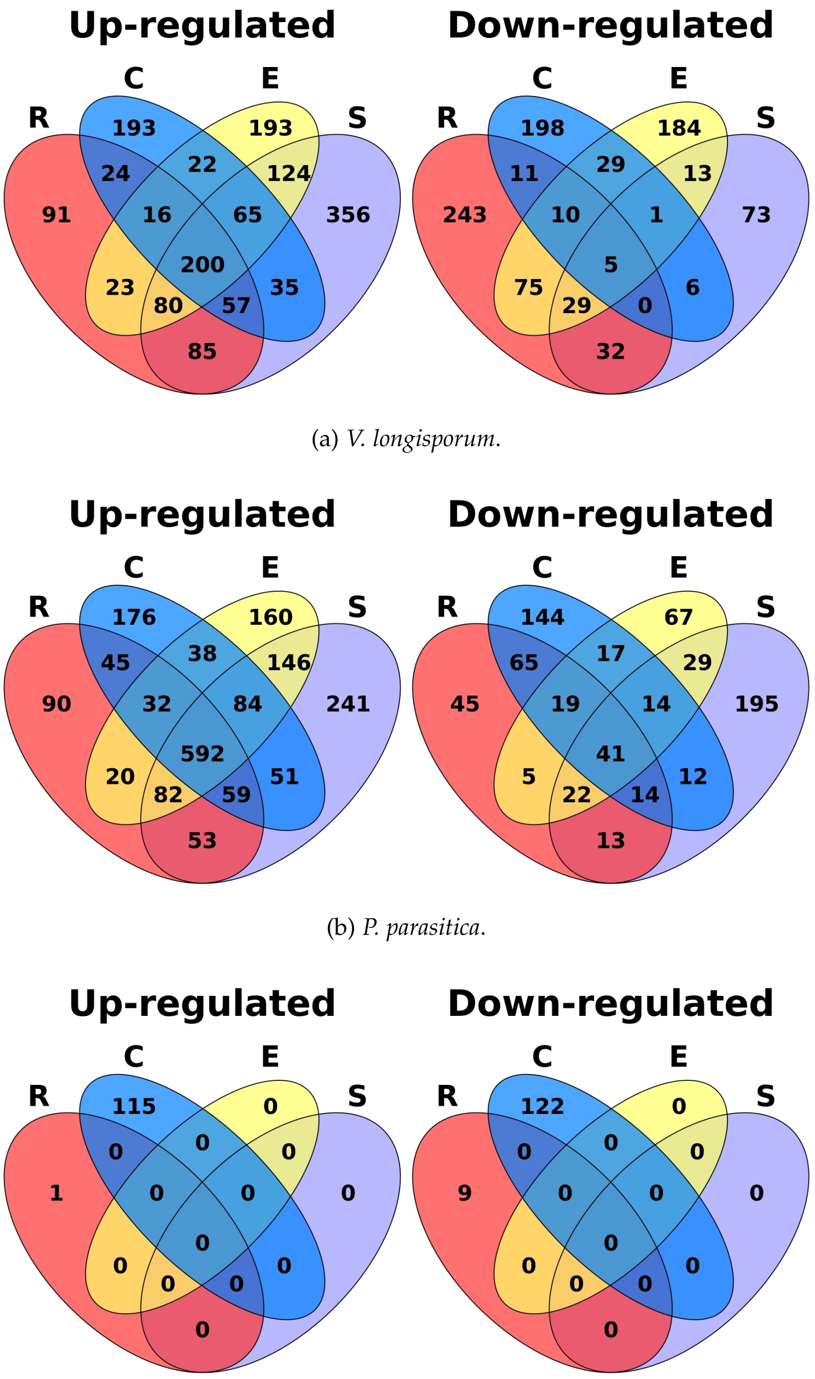
As expected a marginal proportion of differentially expressed genes was present in both the rhizodermis and the cortex of plants treated with *S. indica* (Figure 4.1 C). Most studies indicate that this mutualistic endophytic fungus does not colonize beyond the cortex, with rare cases reporting colonization of meristematic and elongation zones (Qiang et al., 2012). The high proportion of differentially expressed genes in all cell layer of plants treated with *V. longisporum* and *P. parasitica* suggests an active dynamic with plants responding to the attack and pathogens inhibiting these responses (Qi et al., 2018). This overview is in agreement with the three different lifestyles of the pathogens, hence the next step was to compared differentially expressed genes across layers for each pathogen and across pathogens for each cell layer.

Venn diagrams were generated to visualize the aforementioned comparisons. Infection by *V. longisporum* seems to trigger a core reaction in all cell layers, with 200 up-regulated genes shared between cell layers (Figure 4.2 A). However most of the differentially expressed genes were cell-dependent. Additionally, the high number of up-regulated genes in the stele might be related to vascular colonization (Depotter et al., 2016). Unlike *V. longisporum*, *P.parasitica* seems to trigger a more overall reaction with most of the up-regulated genes (Figure 4.2 B) being shared between cell layers. This is in agreement with a necrotrophic lifestyle, resulting in plant damage across root cell layers (Y. Meng et al., 2014). Plants treated with *S. indica* display all of the differential expression in the cortex with a marginal proportion in the rhizodermis. This set of up- and down-regulated genes might be the result of the suppression of the root innate immune response and the establishment of the cell dead-dependent phase (Qiang et al., 2012).

Comparison across pathogens in each cell layer reveled a markedly differential induction with a lower number of genes being shared between pathogens (Figure 4.3). Infection by *P. parasitica* led to the major number of up-regulated genes in all cell layers, whereas down-regulation was higher in the rhizodermis and the endodermis of plants infected by *V. longisporum*. In general, both comparisons supported the hypothesis that roots display a cell-layer and pathogen-specific response. Grounded on this evidence, the following question was: which genes are involved in common or unique responses.

4.3.2 *Gene clustering and hypotheses formulation*

Heatmaps and hierarchical clustering were generated in order to separate genes displaying common or unique responses. Figure 4.4 present 15 demarcated clusters with either up- or -down regulated genes. Cluster 1 contains genes involved in common response to pathogens (with an insignificant induction in



(a) *V. longisporum*.

(b) *P. parasitica*.

(c) *S. indica*.

Figure 4.2: Venn diagrams comparing the distribution of differentially expressed genes (DEGs) across root cell layers.

the presence of *S. indica*), some of these genes are involved in the biosynthesis of secondary antimicrobial compounds and ethylene biosynthesis. Cluster 7 lists genes down-regulated in the endodermis of roots infected by *V. longisporum*.

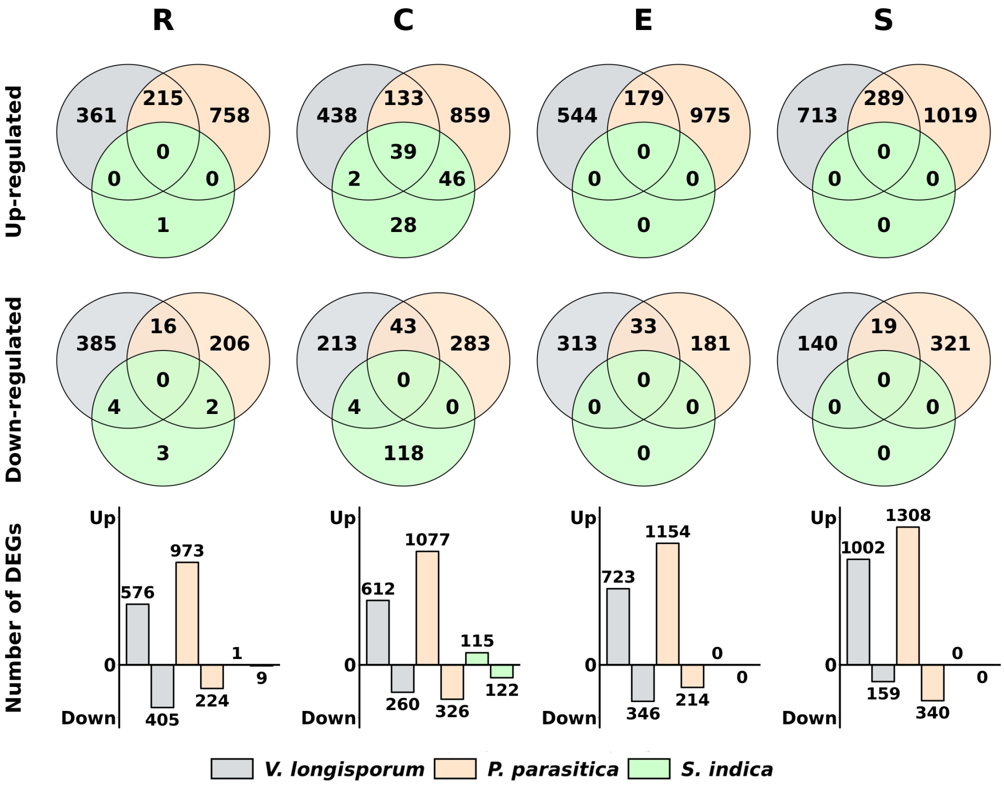


Figure 4.3: Venn diagrams comparing the response of each root cell layer against the fungal (one fungus-like) species

This induction could be related with pathogen mechanisms that enable the colonization of root vascular system.

Together with this hierarchical clustering KEGG enrichment analysis was carried out. These analyses provided an overall view of the type of genes being differentially expressed in the presence of each microbe. KEGG enrichment of down-regulated genes in roots infected by *V. longisporum* point to the hypothesized function of genes present in cluster number 7 (Figure 4.5). In a similar way, genes grouped in cluster 1 probably correspond to the pathogen defense response pathway enriched in the presence of *V. longisporum* and *P. parasitica* (Figure 4.5, Figure 4.6). Additionally, pathways involved in pathogen responses were significantly enriched for up-regulated genes in all root cell layers infected by *P. parasitica*, which was expected due its necrotrophic life style (Figure 4.6). The large amount of dead cells resulting from this life style triggers an overall defense response across cell layers. On the other hand, as anticipated up-regulated genes in the cortex of root colonized by *S. indica* point to pathways related to pathogen defense response. This response is related to the endoplasmic reticulum (ER) stress induced in the early stages of the colonization

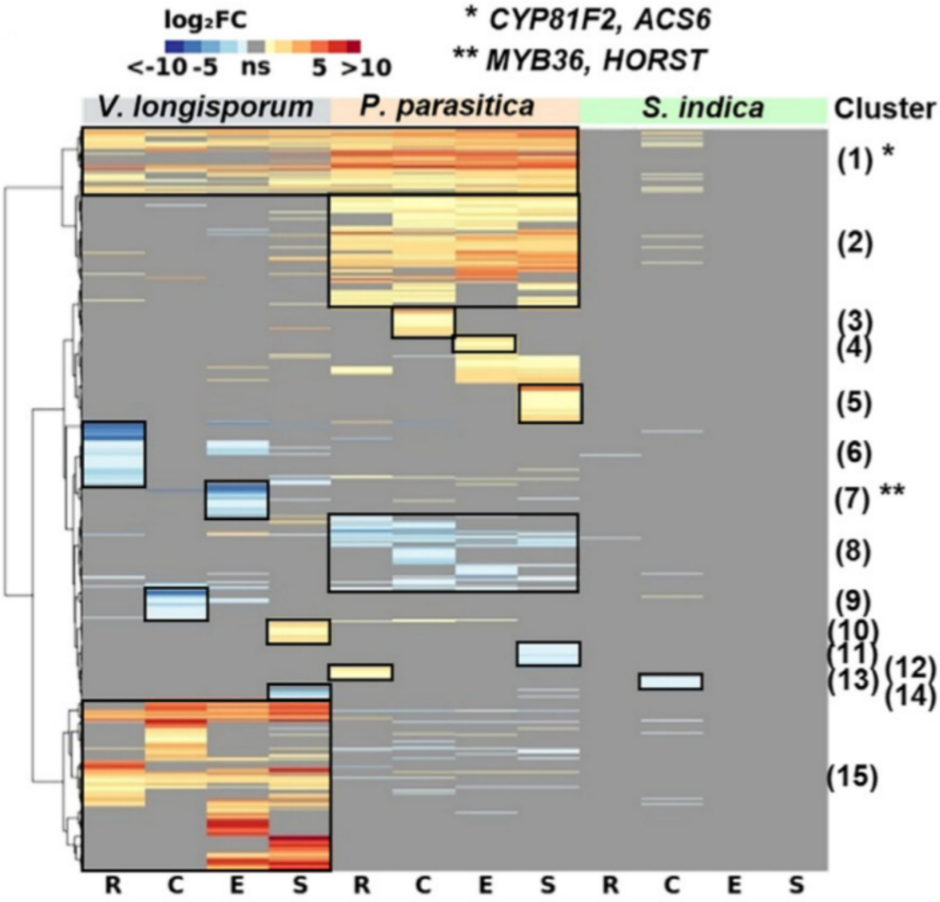


Figure 4.4: Hierarchical clustering of differentially expressed genes (DEGs). Blue and rot color scales represent down-and up-regulation respectively. Asterisks mark genes to be studied in more detail.

(after three days). Later, *S. indicia* suppresses the adaptive ER stress response in order to facilitate root colonization (Qiang et al., 2012). Therefore, since the gene induction was evaluated after a short time colonization, no genes related to this posterior adaptive ER stress response were found among the down-regulated genes (Figure 4.7).

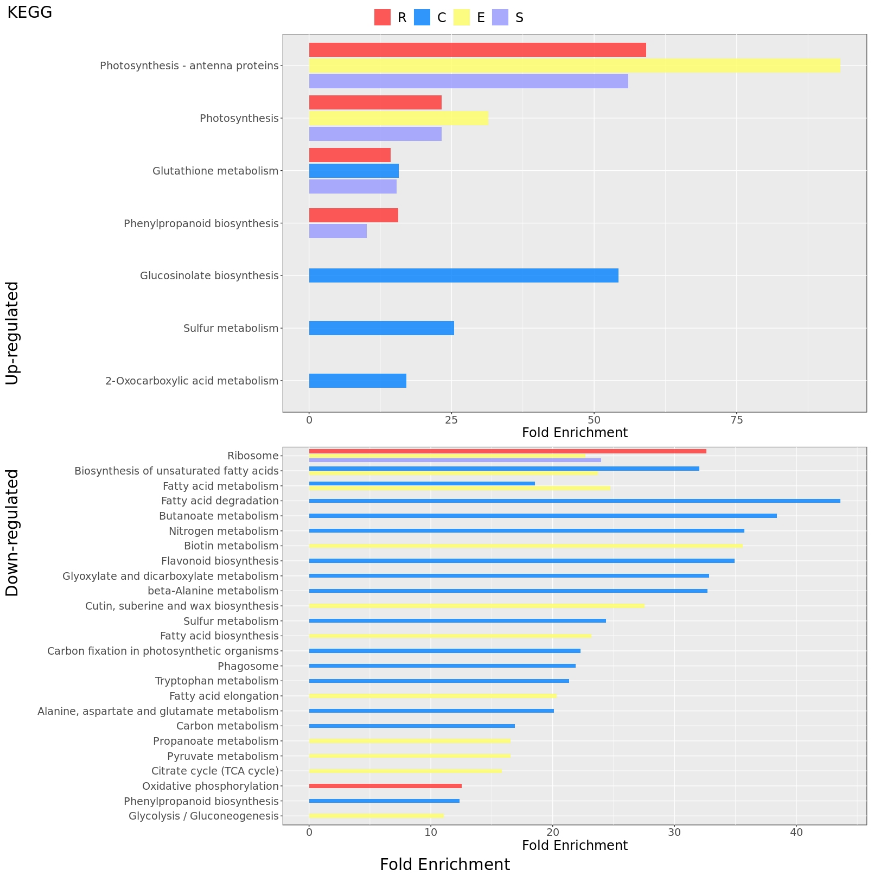


Figure 4.5: KEGG enrichment of down-and up-regulated genes across root cell layers of plants infected by *V. longisporum*. Letter R, C, E and S refer to rhizodermis, cortex, endodermis and stele respectively

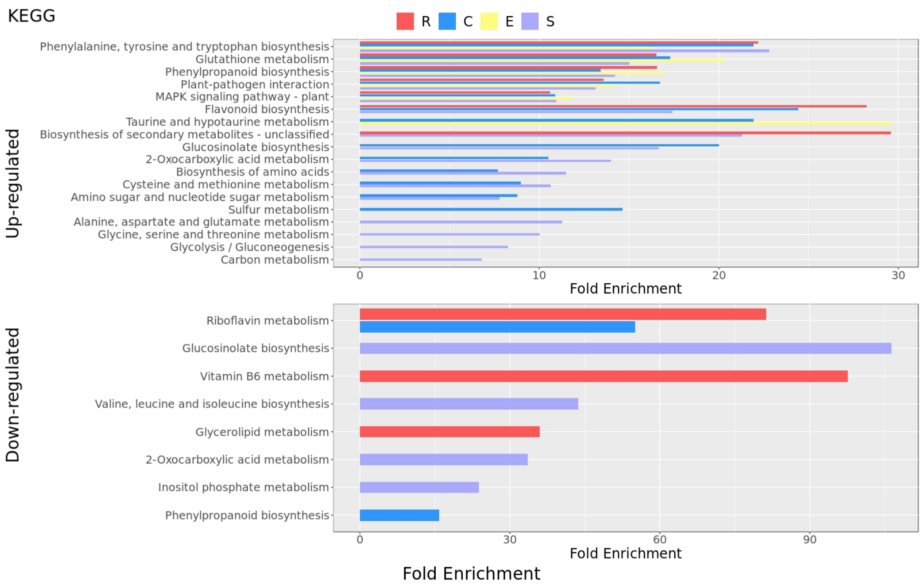


Figure 4.6: KEGG enrichment of down-and up-regulated genes across root cell layers of plants infected by *P. parasitica*. Letter R, C, E and S refer to rhizodermis, cortex, endodermis and stele respectively

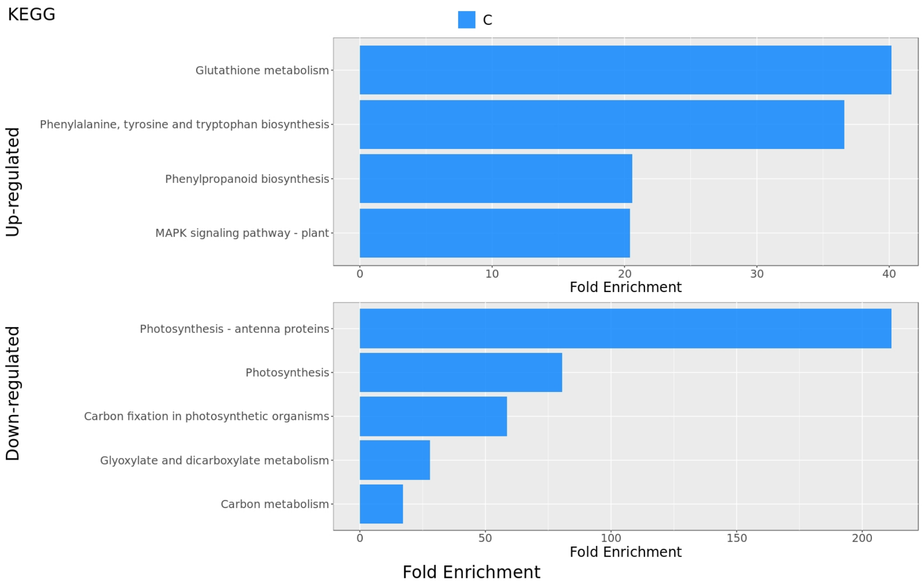


Figure 4.7: KEGG enrichment of down-and up-regulated genes across root cell layers of plants infected by *S. indica*. Letter C refers to cortex

For a better visualization of the pathways in which differentially expressed genes were located, path viewers were generated. It is important to note that down-regulated genes in the endodermis of roots colonized by *V. longisporum* matched with both, suberin formation pathway and acid fatty elongation. The latter supply the initial blocks for suberin deposition, which plays an important role as a barrier against pathogens. These results highlight the strong effect on this process by *V. longisporum* colonization (Figure 4.8, Figure 4.9). Contrasting with this lifestyle, and as noted through the KEGG analysis, reaction towards *P. parasitica* took place across root cell layers, with groups of genes being up-regulated in all cell layers (Figure 4.10, Figure 4.11, Figure 4.12, Figure 4.13). All these previous explorations gave us clues to select the clusters to be examined in more detail.

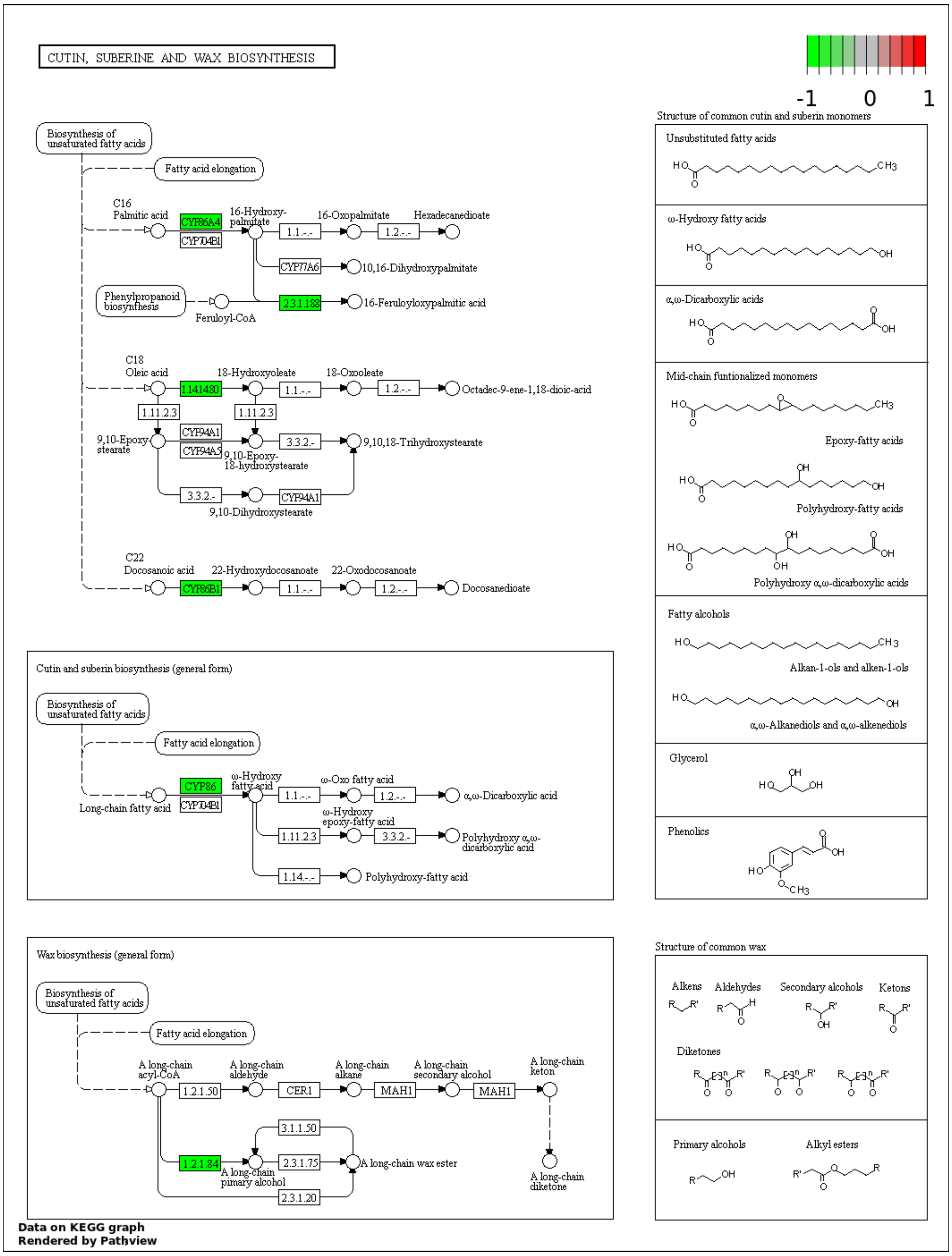


Figure 4.8: Inductions of genes involved in cutin, suberin and wax biosynthesis in the endodermis of root colonized by *V. longisporum*. Green and rot boxes indicate down-and up-regulated genes respectively.

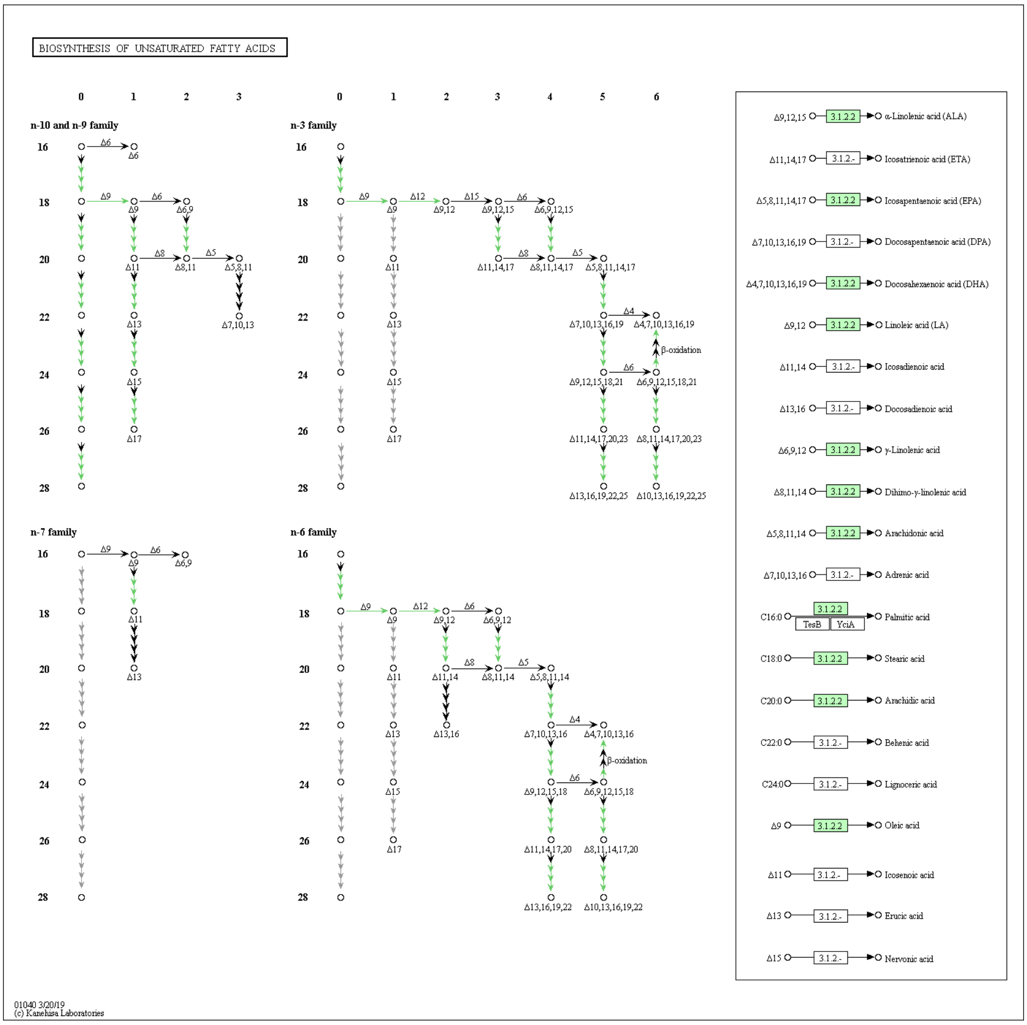

BIOSYNTHESIS OF UNSATURATED FATTY ACIDS
0 1 2 3
0 1 2 3 4 5 6
n-10 and n-9 family
16 18 20 22 24 26 28
Δ6
Δ9
Δ6,9
Δ11
Δ8
Δ8,11
Δ5
Δ5,8,11
Δ7,10,13
Δ13
Δ15
Δ17
n-3 family
Δ12
Δ9,12
Δ15
Δ9,12,15
Δ6,9,12,15
Δ11,14,17
Δ8,11,14,17
Δ5,8,11,14,17
Δ4
Δ7,10,13,16,19
Δ4,7,10,13,16,19
β-oxidation
Δ9,12,15,18,21
Δ6,9,12,15,18,21
Δ11,14,17,20,23
Δ8,11,14,17,20,23
Δ13,16,19,22,25
Δ10,13,16,19,22,25
n-7 family
n-6 family
Δ11,14
Δ8,11,14
Δ5,8,11,14
Δ13,16
Δ7,10,13,16
Δ4,7,10,13,16
Δ9,12,15,18
Δ6,9,12,15,18
Δ11,14,17,20
Δ8,11,14,17,20
Δ13,16,19,22
Δ10,13,16,19,22
Δ9,12,15 3.1.2.2 α-Linolenic acid (ALA)
Δ11,14,17 3.1.2.- Icosatrienoic acid (ETA)
Δ5,8,11,14,17 3.1.2.2 Icosapentaenoic acid (EPA)
Δ7,10,13,16,19 3.1.2.- Docosapentaenoic acid (DPA)
Δ4,7,10,13,16,19 3.1.2.2 Docosahexaenoic acid (DHA)
Δ9,12 3.1.2.2 Linoleic acid (LA)
Δ11,14 3.1.2.- Icosadienoic acid
Δ13,16 3.1.2.- Docosadienoic acid
Δ6,9,12 3.1.2.2 γ-Linolenic acid
Δ8,11,14 3.1.2.2 Dihimo-γ-linolenic acid
Δ5,8,11,14 3.1.2.2 Arachidonic acid
Δ7,10,13,16 3.1.2.- Adrenic acid
C16:0 3.1.2.2 TesB YciA Palmitic acid
C18:0 3.1.2.2 Stearic acid
C20:0 3.1.2.2 Arachidic acid
C22:0 3.1.2.- Behenic acid
C24:0 3.1.2.- Lignoceric acid
Δ9 3.1.2.2 Oleic acid
Δ11 3.1.2.- Icosenoic acid
Δ13 3.1.2.- Erucic acid
Δ15 3.1.2.- Nervonic acid
01040 3/20/19
(c) Kanehisa Laboratories

Figure 4.9: Inductions of genes involved in the biosynthesis of unsaturated fatty acids in the endodermis of root colonized by *V. longisporum*. Green and rot boxes indicate down-and up-regulated genes respectively.

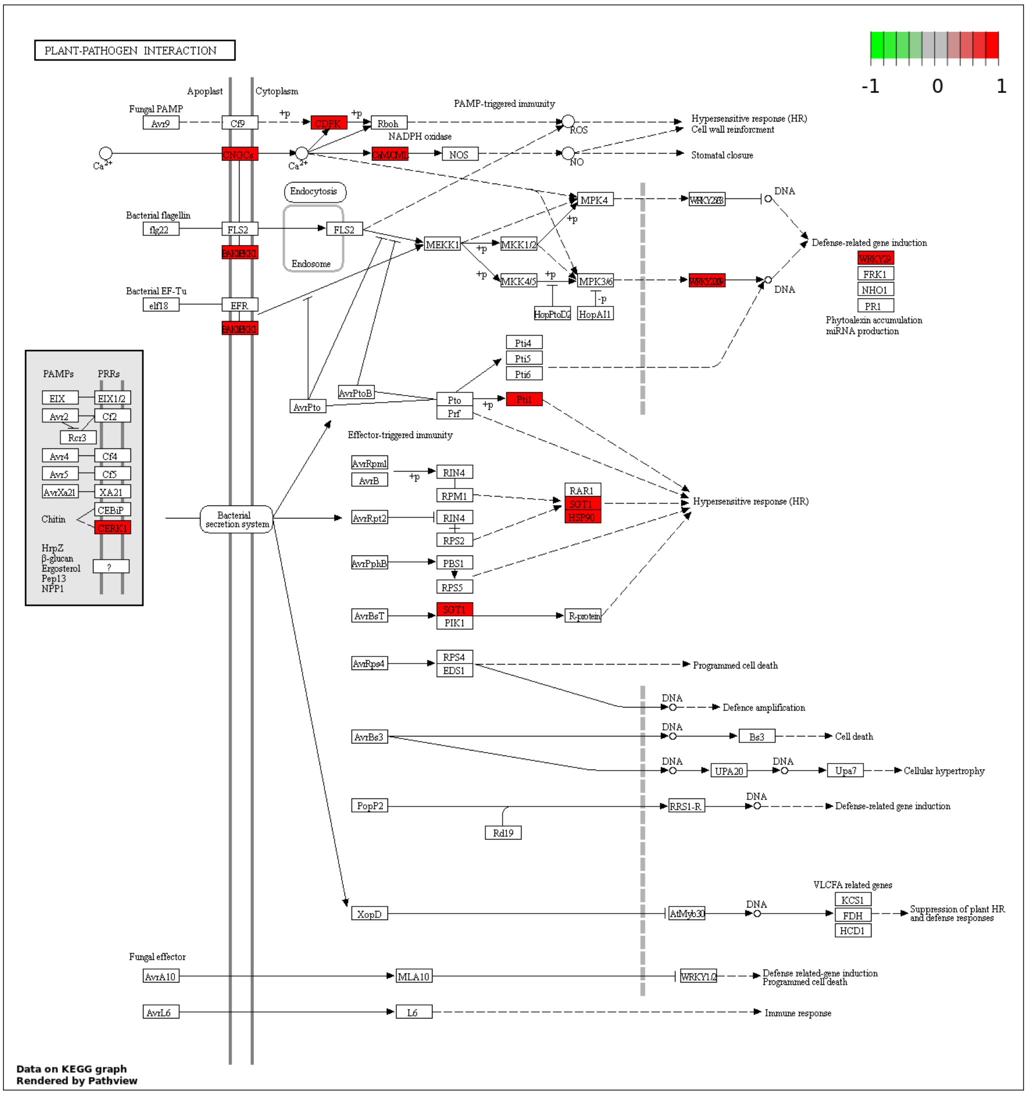


Figure 4.10: Inductions of genes involved in plant-pathogen interaction in the rhizodermis of root colonized by *P. parasitica*. Green and rot boxes indicate down-and up-regulated genes respectively.

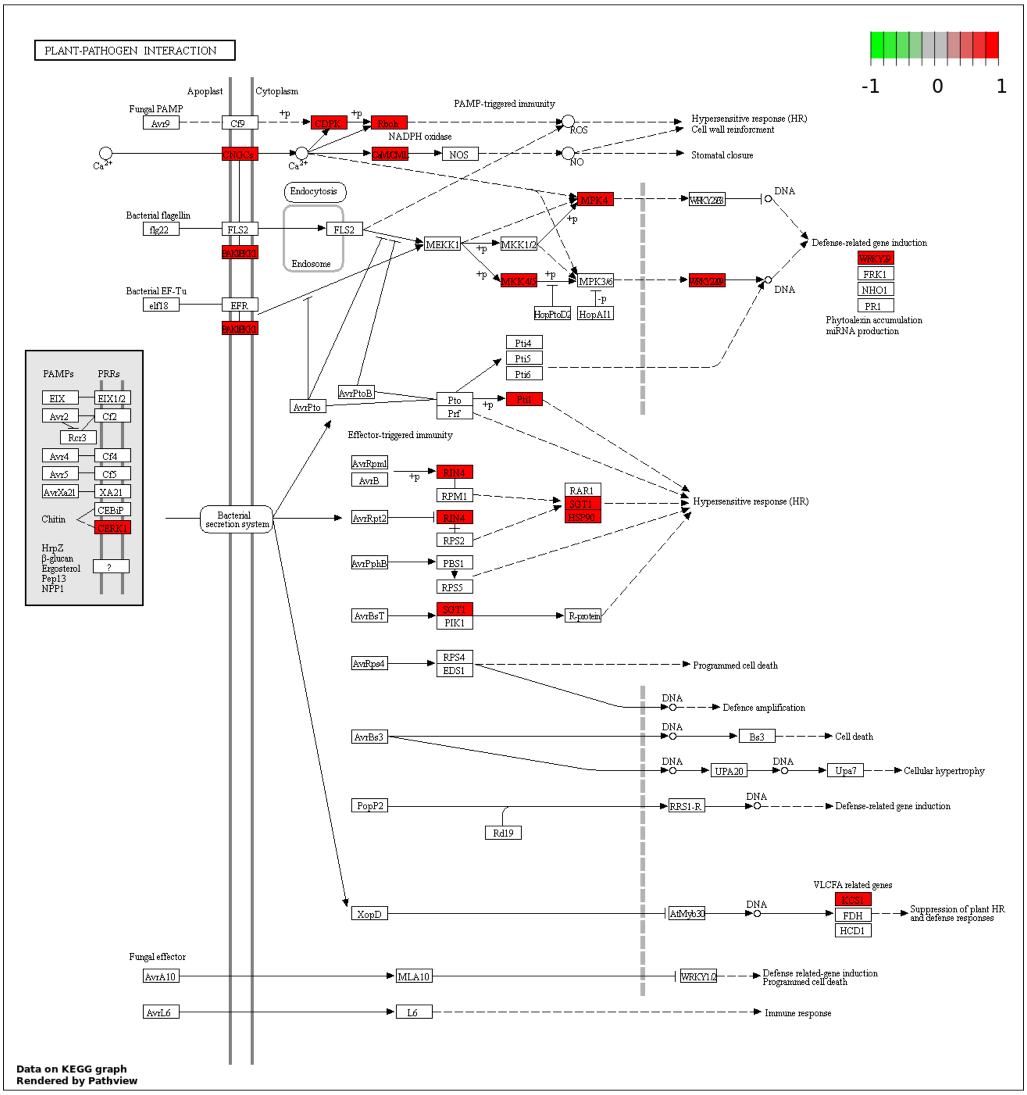


Figure 4.11: Inductions of genes involved in plant-pathogen interaction in the cortex of root colonized by *P. parasitica*. Green and rot boxes indicate down-and up-regulated genes respectively.

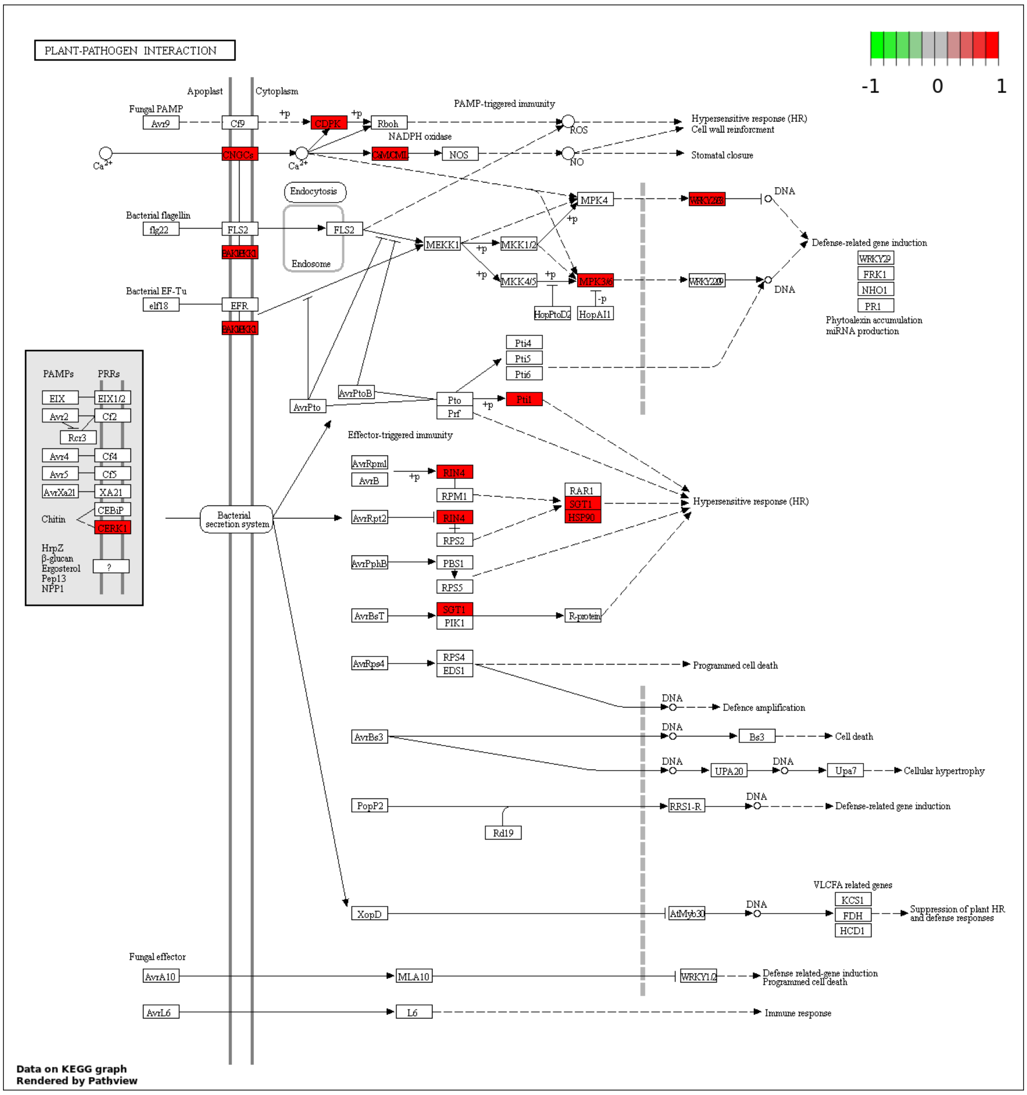


Figure 4.12: Inductions of genes involved in plant-pathogen interaction in the endodermis of root colonized by *P. parasitica*. Green and rot boxes indicate down-and up-regulated genes respectively.

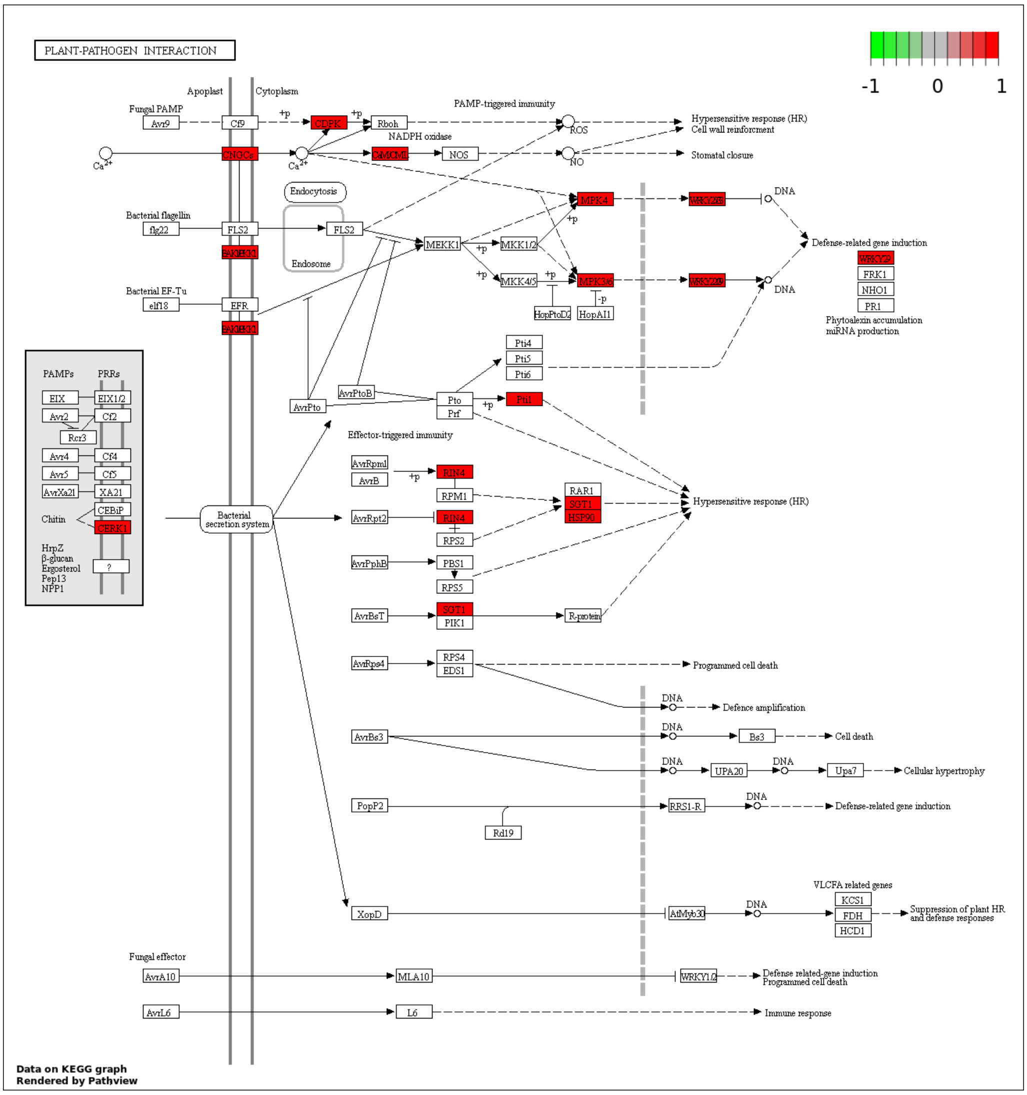


Figure 4.13: Inductions of genes involved in plant-pathogen interaction in the stele of root colonized by *P. parasitica*. Green and rot boxes indicate down-and up-regulated genes respectively.

Zooming into the cluster 7 (Figure 4.14), it was possible to confirm that *V. longisporum* actively down-regulates genes implicated in the casparian strip formation and suberin deposition. This induction was only detected in the infection by this pathogen, which is supported by the fact that it penetrates across all cell layers until reaching the pericycle. Casparian strips and suberin deposition might act as barrier against this pathogen, if so one could hypothesize that mutants of *A. thaliana* unable of establishing this barrier should be more susceptible to *V. longisporum* infection. This hypothesis was later tested and confirmed by Christian Fröschel using confocal microscopy and *A. thaliana* mutants in suberin (*horst-1*, *horst-2*, and the double mutant *pCASP1:CDEF1*).

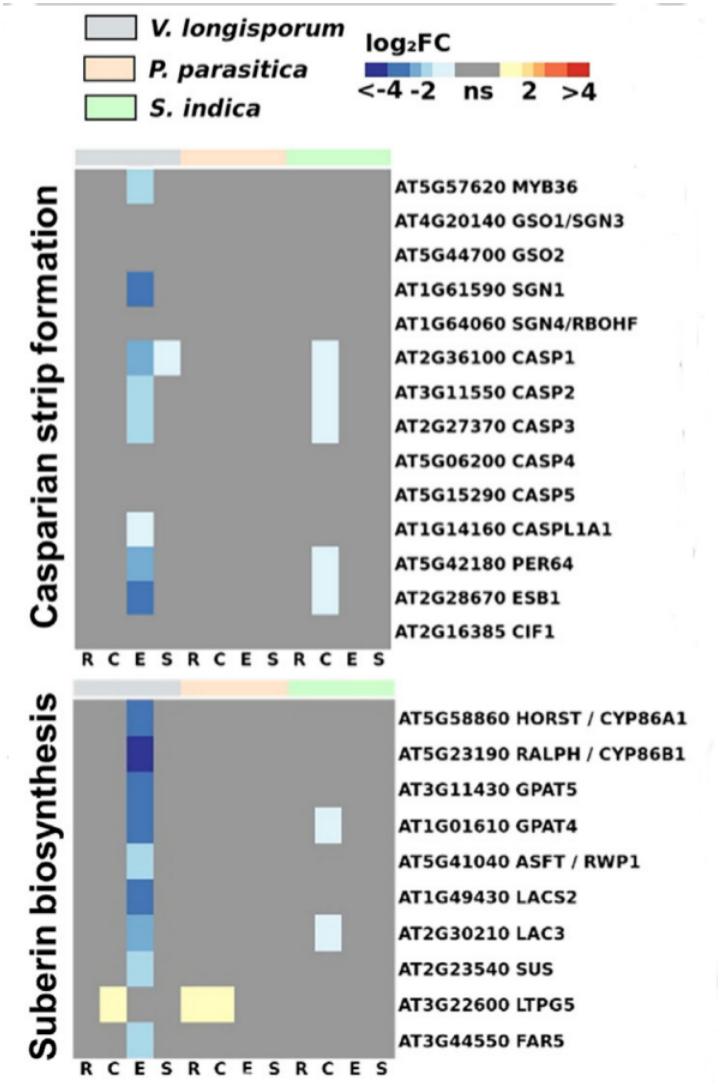


Figure 4.14: Heatmap comparing gene expression of genes related to casparian strip formation and suberin formation across root cell layers and pathogens. Blue and rot color scales represent down-and up-regulation respectively.

In the presence of pathogens plants activated complex metabolic networks associated with the aminoacids methionine and tryptophan which lead to the production of secondary metabolites with antimicrobial properties (Ahuja, Kissen, and Bones, 2012). Figure 4.15 contains differential expressed genes involved in some of these metabolic networks. A common induction of genes related to tryptophan-derived metabolites (Figure 4.15 left side) was detected in the presence of both *V. longisporum* and *P. parasitica* in all root cell layers suggesting an active defense response throughout the root. In a similar way, genes related to sulfur assimilation and indole-glucosinolate (methionine-derived metabolites) were up-regulated in the presence of both pathogens, however a contrasting pattern of induction was observed in genes involved in aliphatic-glucosinolates biosynthesis between these pathogens, with up-regulation in the cortex triggered by *V. longisporum* and down-regulation in the stele triggered by *P. parasitica* (Figure 4.15 right side). The hypothesis of susceptibility to infection by these pathogens was later tested and confirmed by Christian Fröschel using *A. thaliana* mutants and confocal microscopy.

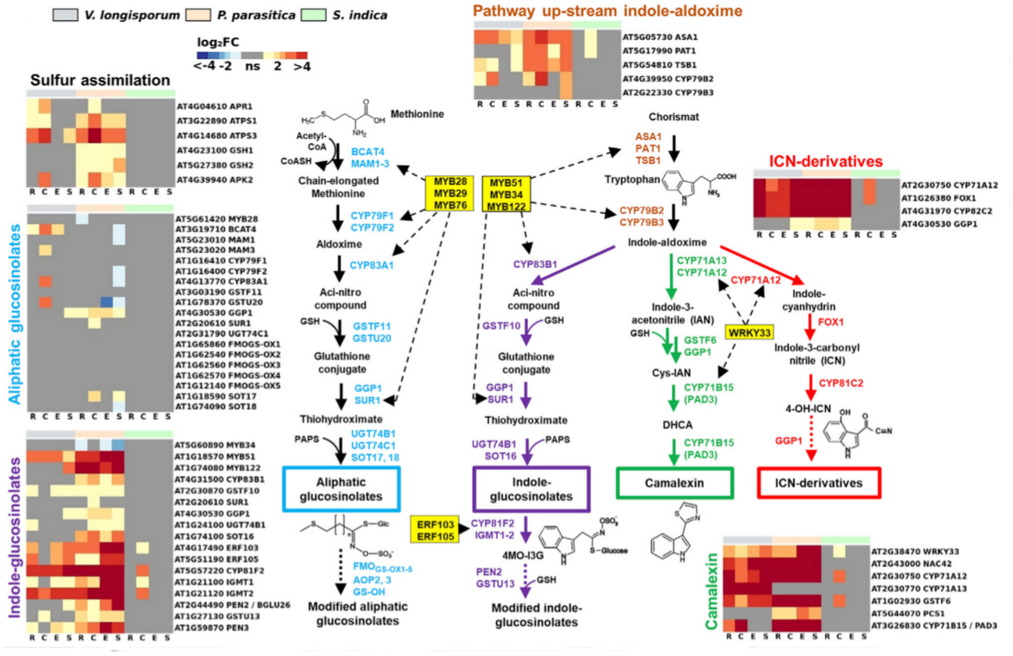


Figure 4.15: Simplified schematic illustration of the biosynthetic pathways and enzymes leading to aliphatic glucosinolates (AGs, blue) and the tryptophan-derived indoleglucosinolates (IGs, purple), camalexin (green), and ICN derivatives (red). Taken from: Fröschel et al., 2021

4.4 FINAL COMMENTS

Results coming from RNAseq data analysis provided a general overview of gene induction across root cell layers of *A. thaliana* plants infected by one mutualistic and two pathogenic fungi. These results do not only recover previously identified candidates genes, but shed light on the pathogen response at a cellular resolution. The data allow the comparison of these responses across distinct root cell layers. The results related to casparian strips formation and secondary antimicrobial compounds provided important evidence to propose subsequent hypotheses and design further experiments. In this respect, GO and KEGG enrichment analysis were the first explorations to be carried out in order to gain a general view about how the contrasting fungal lifestyles induce differential responses throughout the root.

A DATA HARMONIZATION

A.1 INTRODUCTION

The content presented in this appendix is the result of my participation in the GHP Project on Access to Care for Cardiometabolic Diseases (HPACC). This project is not scientific related to my PhD, but of relevance since represented an important source of funding during my PhD. Under the supervision of Dr. Pascal Geldsetzer I was responsible for the data harmonization of surveys made available by the World Health Organization (WHO), mainly based on the STEPS strategy. This introduction is intended to give an overview about data harmonization and its importance during the execution of a project with the participation of multiple collaborators, and how data coming from the STEPS strategy can be used to propose new health policies.

Digitalization of medicine has been steadily growing in the last years, which has improved clinical, research and public health databases. This increase has taken public health research to the level of big data and made worldwide studies with sufficient theoretical statistical power possible. Such studies allow testing of hypotheses in a global scale in order to drive new health polices (Auffray et al., 2016). However, a high proportion of this data is not consistent across countries making it necessary to adopt strategies intended to combine them in an unified and comparable data set. In this sense, data harmonization aims to combine data from different sources and to make it comparable and accessible for researches, often from distinct fields. In particular for public health studies, governments and institutions have been focusing on international comparisons, which demands harmonized cross-national data sets (Granda, C. Wolf, and Hadorn, 2010).

Data Harmonization can be achieved through the use of identical data collection tools and procedures, which refers to the stringent approach. On the contrary, the flexible approach does not require identical data collections, as long as methodology ensures inferential equivalence of the harmonized data. In terms of implementation, data harmonization can be either prospective or retrospective.The first implicates that researchers agree on a core set of variables and data collection tools to ensure standard operability, while the second focuses on information already collected by existing studies (Fortier, Doiron, et al., 2011). The implementation of the prospective approach is in many cases not suitable due to the difficulty of foreseeing future harmonization requirements when planning a new study. For that reason, retrospective harmonization is in most cases the only practicable strategy to be implemented (Fortier, Raina, et al., 2017). In this project, we implemented a retrospective harmonization approach based on a codebook. In data harmonization, codebooks contain accurate information regarding the variables to be consider in a study. Highly specific and detailed codebooks enable researchers to finely distinguish variables and labels and to deliver consistent results. This characteristics can be fulfilled by implementing a six-component codebook including: code name/label, brief definition, full definition,inclusion criteria, exclusion criteria, and examples (MacQueen et al., 1998).

Data harmonization in this project was intended to clean data coming mainly from STEPS (STEPwise approach to surveillance) surveys made available by the WHO. STEPS is a simple, standardized method for collecting, analyzing and disseminating data in WHO member countries. Non-communicable diseases (NCD) are the major cause of death and disability in a lot of countries and the burden generated by this type of diseases are rapidly growing in developing countries. Without control policies, NCD might put in check the already overwhelmed health services in these countries (Bonita et al., 2003). In this way, WHO STEPwise approach to noncommunicable disease (NCD) risk factor surveillance is aimed to help countries to collect consistent data in order to assist health services in the planification and determination of public health priorities (World Health Organization, 2005). In this sense, the first phase of the HPACC is to collate and analyze existing data from nationally representative population-based surveys under the STEPwise approach. During my participation on this project, I was able to include more than 40 cleaned data sets to our merged data set. Thanks to this fast production of harmonized data, one paper was already accepted (*Body mass index and diabetes risk in fifty-seven low- and middle-income countries: a cross-sectional study of nationally representative individual-level data*, at: *The Lancet*) and a second is under review (*Patterns of tobacco use prevalence and frequency in 70 low- and middle-income countries*, at: *Nature Medicine*).

A.2 methods and results

Data harmonization on 40 surveys coming from the STEPwise approach and 2 coming from the Demographic and Health Surveys (DHS) Program was carried out by using a codebook with approx 280 variables in R (R Core Team, 2021) . This variables included demographic information, tobacco and alcohol consume, physical activity, domestic violence, blood pressure and biochemical parameters like blood glucose level and lipid profile. The first step was to look for comparable variables between the codebook and the raw data of each country data set. To do that, regular expressions were used in order to search for compatible patterns between the question in the original surveys and the codebook. Matches were principally defined throughout frequency of use for tobacco and alcohol consume (for example *grep("[D-d]aily")* or directly by using the name's variable. To confirm the matches, variables codes in the raw data were compared to the respectively question in the surveys. Once all possible variables were matched, labels of each variable were transform according to those contained in the codebook. Figure A.1 summarizes the process from raw data to harmonized data. Difficulties were especially found for educational labels since in most cases each country defined its on labels or grades, making them not comparable between countries.

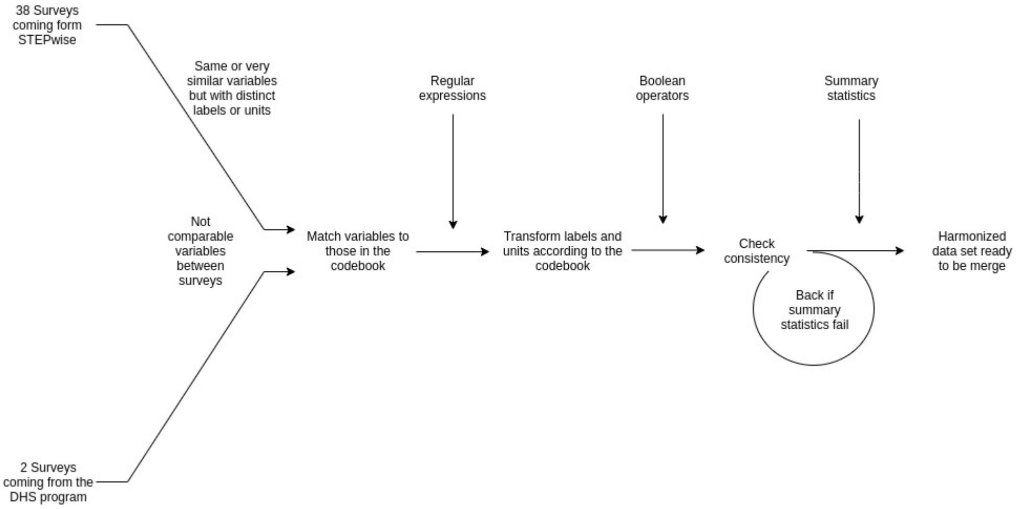

38 Surveys coming form STEPwise
Same or very similar variables but with distinct labels or units
Regular expressions
Boolean operators
Summary statistics
Not comparable variables between surveys
Match variables to those in the codebook
Transform labels and units according to the codebook
Check consistency
Harmonized data set ready to be merge
Back if summary statistics fail
2 Surveys coming from the DHS program

Figure A.1: Schematic representation of the steps followed during data harmonization

Additionally, an important part of the harmonization process was the assignation of especial labels, namely "don't know", "refused" and the skip patterns. While the first two were clearly defined in the raw data, the latter were always store as NA and therefore mixed with real NA representing lack of information. They were separated using Boolean operators based on the conditions contained in the surveys. It was very important to distinguish between them given that passing skip patterns as NA would inflate the proportion of lacking data making difficult for latter statistical analysis. Finally, for each country a semi-automated R script was created in order to allow the recovery of the cleaned data by all researchers. Considering that part of the statistical analysis was carried out using STATA, the *haven* package (Wickham and Miller, 2020) was used in order to read the raw data and export the cleaned data. Finally, all cleaned data sets were merge with the already existing master data set implementing *collation* in STATA (StataCorp, 2019).

A.3 FINAL COMMENTS

Digitalization of medicine has put within reach an unimaginable amount of data, however in most cases this data is not consistent making it impracticable for the use in research. Data harmonization refers to a compendium of strategies to reconcile these inconsistencies, allowing the use of data from different sources in order to test common hypotheses. During my participation in this project, I was able to harmonized data coming from different types of surveys, countries and years, with the aim of creating a master data set to be used in testing hypotheses related to the impact of Non-communicable diseases in low-and middle-income countries. This type of data set is of relevant importance since such diseases are responsible for approximately two out of three deaths worldwide. Although prolonged discussions were necessary in order to establish a final version of our codebook, even this version was updated multiple times, which showed me that data harmonization, at least from the retrospective approach, is dynamic and required a constantly rethinking of the consensus in relation to meaningful variables and labels to be taking into account in a project.

B

TRANSCRIPT QUANTIFICATION

B.1 TRANSCRIPT QUANTIFICATION

To quantify TPM in each sample a customized bash script was used:

```
#!/bin/bash
Files='/mnt/volume/Seq_Daten_AK/Verticillium_2dpi/Verti_
   fastq/*'
for f in $Files
do
samp=$f
echo "Processing sample ${samp}"
 Name="$(echo $Files | cut -d'/' -f7)"
salmon quant -i athal_index -l A \
        -r $f \
        -p 8 -o Verticillium/${Name}_quant
done
```

The flag *quant* invokes the quantification function and the arguments *-i, -l A, -r, -p* and *-o* tell Salmon where the index file is stored, to automatically define the library type of the sequencing reads, to create a new folder, to use a selected number of threads and where to save the outputs, respectively. After quantification of all samples, *quant.sf* files were read into R and TPM columns were merged between all samples to obtain a data frame containing all quantification results without splicing variants. This data frame was later used to performed the differential expression analysis.

BIBLIOGRAPHY

Abbott, Richard J and Mioco F Gomes (1989). "Population genetic structure and outcrossing rate of Arabidopsis thaliana (L.) Heynh." In: *Heredity* 62.3, pp. 411–418.

Agren, J. et al. (Mar. 2017). "Adaptive divergence in flowering time among natural populations of Arabidopsis thaliana: Estimates of selection and QTL mapping." In: *Evolution* 71.3, pp. 550–564.

Ågren, Jon, Christopher G Oakley, et al. (2017). "Adaptive divergence in flowering time among natural populations of Arabidopsis thaliana: Estimates of selection and QTL mapping." In: *Evolution* 71.3, pp. 550–564.

Ågren, Jon and Douglas W Schemske (2012). "Reciprocal transplants demonstrate strong adaptive differentiation of the model organism Arabidopsis thaliana in its native range." In: *New Phytologist* 194.4, pp. 1112–1122.

Ahuja, Ishita, Ralph Kissen, and Atle M Bones (2012). "Phytoalexins in defense against pathogens." In: *Trends in plant science* 17.2, pp. 73–90.

Alonso-Blanco, C et al. (2016). "1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*." In: *Cell* 166.2, pp. 481–491.

Anderson, Jill T et al. (2013). "Genetic trade-offs and conditional neutrality contribute to local adaptation." In: *Molecular ecology* 22.3, pp. 699–708.

Aranzana, Marıa José et al. (2005). "Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes." In: *PLoS Genet* 1.5, e60.

Arnegard, Matthew E et al. (2014). "Genetics of ecological divergence during speciation." In: *Nature* 511.7509, pp. 307–311.

Atwell, Susanna et al. (2010). "Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines." In: *Nature* 465.7298, pp. 627–631.

Auffray, Charles et al. (2016). "Making sense of big data in health research: towards an EU action plan." In: *Genome medicine* 8.1, pp. 1–13.

Auge, Gabriela A, Steven Penfield, and Kathleen Donohue (2019). "Pleiotropy in developmental regulation by flowering-pathway genes: is it an evolutionary constraint?" In: *New Phytologist* 224.1, pp. 55–70.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." In: *Journal of the Royal statistical society: series B (Methodological)* 57.1, pp. 289–300.

Berndt, Sonja I et al. (2013). "Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture." In: *Nature genetics* 45.5, pp. 501–512.

Bonferroni, Carlo E (1935). "Il calcolo delle assicurazioni su gruppi di teste." In: *Studi in onore del professore salvatore ortu carboni*, pp. 13–60.

Bonita, Ruth et al. (2003). "The WHO Stepwise approach to surveillance (STEPS) of non-communicable disease risk factors." In: *Global behavioral risk factor surveillance*. Springer, pp. 9–22.

Botstein, David et al. (1980). "Construction of a genetic linkage map in man using restriction fragment length polymorphisms." In: *American journal of human genetics* 32.3, p. 314.

Bouché, Frédéric et al. (2016). "FLOR-ID: an interactive database of flowering-time gene networks in Arabidopsis thaliana." In: *Nucleic Acids Research* 44.D1, pp. D1167–D1171.

Bouveret, Romaric et al. (2006). "Regulation of flowering time by Arabidopsis MSI1." In: *Development* 133.9, pp. 1693–1702.

Brachi, Benjamin, Geoffrey P Morris, and Justin O Borevitz (2011). "Genome-wide association studies in plants: the missing heritability is in the field." In: *Genome biology* 12.10, pp. 1–8.

Buniello, Annalisa et al. (2019). "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019." In: *Nucleic acids research* 47.D1, pp. D1005–D1012.

Bush, William S and Jason H Moore (2012). "Genome-wide association studies." In: *PLoS Comput Biol* 8.12, e1002822.

Champely, Stephane et al. (2017). "pwr: Basic functions for power analysis." In:

Chanock, S. J. et al. (June 2007). "Replicating genotype-phenotype associations." In: *Nature* 447.7145, pp. 655–660.

Chen, H and PC Boutros (2011). "VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R." In: *BMC bioinformatics* 12.1, pp. 1–7.

Chiang, George CK et al. (2013). "Pleiotropy in the wild: the dormancy gene DOG1 exerts cascading control on life cycles." In: *Evolution: International Journal of Organic Evolution* 67.3, pp. 883–893.

Churchill, Frederick B (1974). "William Johannsen and the genotype concept." In: *Journal of the History of Biology* 7.1, pp. 5–30.

Clauw, Pieter et al. (2016). "Leaf growth response to mild drought: natural variation in Arabidopsis sheds light on trait architecture." In: *The Plant Cell* 28.10, pp. 2417–2434.

Cohen, Joel E (2004). "Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better." In: *Plos biol* 2.12, e439.

Corbesier, Laurent et al. (2007). "FT Protein Movement Contributes to Long-Distance Signaling in Floral Induction of Arabidopsis." In: *Science* 316.5827, pp. 1030–1033.

Darwin, Charles (1859). *On the Origin of Species by Means of Natural Selection.* or the Preservation of Favored Races in the Struggle for Life. London: Murray.

Debieu, Marilyne et al. (2013). "Co-variation between seed dormancy, growth rate and flowering time changes with latitude in Arabidopsis thaliana." In: *PloS one* 8.5, e61075.

Depotter, Jasper RL et al. (2016). "Verticillium longisporum, the invisible threat to oilseed rape and other brassicaceous plant hosts." In: *Molecular plant pathology* 17.7, pp. 1004–1016.

Dittberner, Hannes et al. (2018). "Natural variation in stomata size contributes to the local adaptation of water-use efficiency in *Arabidopsis thaliana*." In: *Molecular ecology* 27.20, pp. 4052–4065.

Doehlemann, Gunther et al. (2017). "Plant pathogenic fungi." In: *The fungal kingdom*, pp. 701–726.

Eshel, Amram and Tom Beeckman (2013). *Plant roots: the hidden half*. CRC press.

Exposito-Alonso, Moises (2020). "Seasonal timing adaptation across the geographic range of Arabidopsis thaliana." In: *Proceedings of the National Academy of Sciences* 117.18, pp. 9665–9667.

Farrall, Lyndsay A (1975). "Controversy and conflict in science: A case study—The English biometric school and Mendel's laws." In: *Social Studies of Science* 5.3, pp. 269–301.

Fisher, Ronald Aylmer (1918). "The correlation between relatives on the supposition of mendelian inheritance." In: *Transactions of the Royal Society of Edinburgh* 52, pp. 899–438.

Fornara, Fabio et al. (2009). "Arabidopsis DOF transcription factors act redundantly to reduce CONSTANS expression and are essential for a photoperiodic flowering response." In: *Developmental cell* 17.1, pp. 75–86.

Fortier, Isabel, Dany Doiron, et al. (2011). "Invited commentary: consolidating data harmonization—how to obtain quality and applicability?" In: *American journal of epidemiology* 174.3, pp. 261–264.

Fortier, Isabel, Parminder Raina, et al. (2017). "Maelstrom research guidelines for rigorous retrospective data harmonization." In: *International journal of epidemiology* 46.1, pp. 103–105.

Fothergill, WE (1888). "The Biology of the Future." In: *The Hospital* 3.68, p. 274.

Fournier-Level, A. et al. (Oct. 2011). "A map of local adaptation in Arabidopsis thaliana." In: *Science* 334.6052, pp. 86–89.

Frachon, Léa et al. (2018). "A genomic map of climate adaptation in Arabidopsis thaliana at a micro-geographic scale." In: *Frontiers in plant science* 9, p. 967.

Frei, O et al. (2019). "Bivariate causal mixture model quantifies polygenic overlap between complex traits beyond genetic correlation." In: *Nature communications* 10.1, pp. 1–11.

Freudenthal, Jan A et al. (2019). "GWAS-Flow: A GPU accelerated framework for efficient permutation based genome-wide association studies." In: *BioRxiv*, p. 783100.

Fröschel, Christian et al. (2021). "Plant roots employ cell-layer-specific programs to respond to pathogenic and beneficial microbes." In: *Cell Host & Microbe* 29.2, pp. 299–310. DOI: https://doi.org/10.1016/j.chom.2020.11.014.

Gibson, Greg (2010). "Hints of hidden heritability in GWAS." In: *Nature genetics* 42.7, pp. 558–560.

Gillham, Nicholas W (2015). "The battle between the biometricians and the Mendelians: How Sir Francis Galton's work caused his disciples to reach conflicting conclusions about the hereditary mechanism." In: *Science & Education* 24.1, pp. 61–75.

Granda, Peter, Christof Wolf, and Reto Hadorn (2010). "Harmonizing survey data." In: *Survey methods in multinational, multiregional, and multicultural contexts*, pp. 315–332.

Gudbjartsson, Daniel F et al. (2008). "Many sequence variants affecting diversity of adult human height." In: *Nature genetics* 40.5, pp. 609–615.

Hancock, Angela M et al. (2011). "Adaptation to climate across the Arabidopsis thaliana genome." In: *Science* 334.6052, pp. 83–86.

Hanson, Brooks et al. (1999). "The diversity of evolution." In: *Science* 284.5423, pp. 2105–2105.

Hashiguchi, Takuyu et al. (2014). "Identification of a novel flavonoid glycoside sulfotransferase in Arabidopsis thaliana." In: *The Journal of Biochemistry* 155.2, pp. 91–97.

Hassidim, Miriam et al. (2009). "Over-expression of CONSTANS-LIKE 5 can induce flowering in short-day grown Arabidopsis." In: *Planta* 230.3, pp. 481–491.

Hickey, John M et al. (2017). "Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery." In: *Nature genetics* 49.9, p. 1297.

Horton, Matthew W et al. (2012). "Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel." In: *Nature genetics* 44.2, pp. 212–216.

Huang, Wen et al. (2020). "Genotype by environment interaction for gene expression in Drosophila melanogaster." In: *Nature communications* 11.1, pp. 1–10.

Huang, Xuehui and Bin Han (2014). "Natural variations and genome-wide association studies in crop plants." In: *Annual review of plant biology* 65, pp. 531–551.

Hung, HM James et al. (1997). "The behavior of the p-value when the alternative hypothesis is true." In: *Biometrics*, pp. 11–22.

Huo, Heqiang, Shouhui Wei, and Kent J Bradford (2016). "DELAY OF GERMINATION1 (DOG1) regulates both seed dormancy and flowering time through microRNA pathways." In: *Proceedings of the National Academy of Sciences* 113.15, E2199–E2206.

Ignatiadis, Nikolaos et al. (2016). "Data-driven hypothesis weighting increases detection power in genome-scale multiple testing." In: *Nature methods* 13.7, pp. 577–580.

Imura, Yuri et al. (2012). "CRYPTIC PRECOCIOUS/MED12 is a novel flowering regulator with multiple target steps in Arabidopsis." In: *Plant and Cell Physiology* 53.2, pp. 287–303.

Jacques, F et al. (2020). "Roles for Arabidopsis ALDH10 enzymes in γ-butyrobetaine synthesis, seed development, germination and salt tolerance." In: *Journal of Experimental Botany*.

Johannsen, Wilhelm (1909). *Elemente der exakten Erblichkeitslehre*. Gustav Fischer.

— (1911). "The genotype conception of heredity." In: *The American Naturalist* 45.531, pp. 129–159.

Kang, HM et al. (2010). “Variance component model to account for sample structure in genome-wide association studies.” In: *Nature genetics* 42.4, pp. 348–354.

Kang, MH et al. (2008). “Efficient control of population structure in model organism association mapping.” In: *Genetics* 178.3, pp. 1709–1723.

Kawakatsu, Taiji et al. (2016). “Epigenomic diversity in a global collection of Arabidopsis thaliana accessions.” In: *Cell* 166.2, pp. 492–505.

Kawecki, Tadeusz J and Dieter Ebert (2004). “Conceptual issues in local adaptation.” In: *Ecology letters* 7.12, pp. 1225–1241.

Kerdaffrec, Envel et al. (2016). “Multiple alleles at a single locus control seed dormancy in Swedish Arabidopsis.” In: *elife* 5, e22502.

Kim, Dong-Hwan and Sibum Sung (2013). “Coordination of the vernalization response through a VIN3 and FLC gene family regulatory network in Arabidopsis.” In: *The Plant Cell* 25.2, pp. 454–469.

Kim, Sung et al. (2007). “Recombination and linkage disequilibrium in Arabidopsis thaliana.” In: *Nature genetics* 39.9, pp. 1151–1155.

Kim, Woe-Yeon et al. (2007). “ZEITLUPE is a circadian photoreceptor stabilized by GIGANTEA in blue light.” In: *Nature* 449.7160, pp. 356–360.

Kinoshita, Atsuko and René Richter (2020). “Genetic and molecular basis of floral induction in Arabidopsis thaliana.” In: *Journal of Experimental Botany* 71.9, pp. 2490–2504.

Knoll, Andrew H and Martin A Nowak (2017). “The timetable of evolution.” In: *Science advances* 3.5, e1603076.

Kolde, Raivo and Maintainer Raivo Kolde (2015). “Package ‘pheatmap’.” In: *R package* 1.7, p. 790.

Koornneef, Maarten and David Meinke (2010). “The development of Arabidopsis as a model plant.” In: *The Plant Journal* 61.6, pp. 909–921.

Korte, A and A Farlow (2013). “The advantages and limitations of trait analysis with GWAS: a review.” In: *Plant methods* 9.1, p. 29.

Krämer, Ute (2015). “The natural history of model organisms: Planting molecular functions in an ecological context with Arabidopsis thaliana.” In: *Elife* 4, e06100.

Kronholm, Ilkka et al. (2012). “Genetic basis of adaptation in Arabidopsis thaliana: local adaptation at the seed dormancy QTL DOG1.” In: *Evolution: International Journal of Organic Evolution* 66.7, pp. 2287–2302.

Kulski, Jerzy K (2016). "Next-generation sequencing—an overview of the history, tools, and "omic" applications." In: *Next generation sequencing-advances, applications and challenges*, pp. 3–60.

Laibach, Frederick (1943). "Arabidopsis thaliana (L.) Heynh. als Objekt für genetische und entwicklungsphysiologische Untersuchungen." In: *Bot. Archiv* 44, pp. 439–455.

Lander, Eric and Leonid Kruglyak (1995). "Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results." In: *Nature genetics* 11.3, pp. 241–247.

Lander, Eric S and Nicholas J Schork (1994). "Genetic dissection of complex traits." In: *Science* 265.5181, pp. 2037–2048.

Latrasse, David et al. (2011). "Control of flowering and cell fate by LIF2, an RNA binding partner of the polycomb complex component LHP1." In: *PloS one* 6.1, e16592.

Lee, Horim et al. (2000). "The AGAMOUS-LIKE 20 MADS domain protein integrates floral inductive pathways in Arabidopsis." In: *Genes & development* 14.18, pp. 2366–2376.

Leimu, Roosa and Markus Fischer (2008). "A meta-analysis of local adaptation in plants." In: *PloS one* 3.12, e4010.

Lemos, Bernardo et al. (2008). "Dominance and the evolutionary accumulation of cis-and trans-effects on gene expression." In: *Proceedings of the National Academy of Sciences* 105.38, pp. 14471–14476.

Li, Peijin et al. (2014). "Multiple FLC haplotypes defined by independent cis-regulatory variation underpin life history diversity in Arabidopsis thaliana." In: *Genes & Development* 28.15, pp. 1635–1640.

Li, Xin et al. (2018). "Genomic and environmental determinants and their interplay underlying phenotypic plasticity." In: *Proceedings of the National Academy of Sciences* 115.26, pp. 6679–6684.

Li, Yan et al. (2010). "Association mapping of local climate-sensitive quantitative trait loci in Arabidopsis thaliana." In: *Proceedings of the National Academy of Sciences* 107.49, pp. 21199–21204.

Li, Yi et al. (2019). "Extreme sampling design in genetic association mapping of quantitative trait loci using balanced and unbalanced case-control samples." In: *Scientific reports* 9.1, pp. 1–9.

Lim, Chae Woo and Sung Chul Lee (2020). “ABA-dependent and ABA-independent functions of RCAR5/PYL11 in response to cold stress.” In: *Frontiers in plant science* 11.

Lopez-Arboleda, William Andres et al. (2021). “Global genetic heterogeneity in adaptive traits.” In: *Molecular Biology and Evolution*. msab208. ISSN: 0737-4038. DOI: 10.1093/molbev/msab208. eprint: https://academic.oup.com/mbe/advance-article-pdf/doi/10.1093/molbev/msab208/38886451/msab208.pdf. URL: https://doi.org/10.1093/molbev/msab208.

Love, Michael I, Wolfgang Huber, and Simon Anders (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” In: *Genome biology* 15.12, pp. 1–21.

Lu, Falong et al. (2010). “JMJ14 is an H3K4 demethylase regulating flowering time in Arabidopsis.” In: *Cell research* 20.3, pp. 387–390.

MacQueen, Kathleen M et al. (1998). “Codebook development for team-based qualitative analysis.” In: *Cam Journal* 10.2, pp. 31–36.

Manolio, Teri A et al. (2009). “Finding the missing heritability of complex diseases.” In: *Nature* 461.7265, pp. 747–753.

Martin, Arnaud and Virginie Orgogozo (2013). “The loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation.” In: *Evolution* 67.5, pp. 1235–1250.

Martınez-Berdeja, Alejandra et al. (2020). “Functional variants of DOG1 control seed chilling responses and variation in seasonal life-history strategies in Arabidopsis thaliana.” In: *Proceedings of the National Academy of Sciences* 117.5, pp. 2526–2534.

Mathieu, Johannes et al. (2009). “Repression of flowering by the miR172 target SMZ.” In: *PLoS Biol* 7.7, e1000148.

McCarthy, Mark I et al. (2008). “Genome-wide association studies for complex traits: consensus, uncertainty and challenges.” In: *Nature reviews genetics* 9.5, pp. 356–369.

McClellan, Jon and Mary-Claire King (2010). “Genetic heterogeneity in human disease.” In: *Cell* 141.2, pp. 210–217.

Meinke, David W (2020). “Genome-wide identification of EMBRYO-DEFECTIVE (EMB) genes required for growth and development in Arabidopsis.” In: *New Phytologist* 226.2, pp. 306–325.

Meinke, David W et al. (1998). “Arabidopsis thaliana: a model plant for genome analysis.” In: *Science* 282.5389, pp. 662–682.

Méndez-Vigo, Belén et al. (2011). "Altitudinal and climatic adaptation is mediated by flowering traits and FRI, FLC, and PHYC genes in Arabidopsis." In: *Plant physiology* 157.4, pp. 1942–1955.

Meng, Yuling et al. (2014). "Phytophthora parasitica: a model oomycete plant pathogen." In: *Mycology* 5.2, pp. 43–51.

Meyers, Lauren Ancel and Donald A Levin (2006). "On the abundance of polyploids in flowering plants." In: *Evolution* 60.6, pp. 1198–1206.

Momozawa, Yukihide and Keijiro Mizukami (2020). "Unique roles of rare variants in the genetics of complex diseases in humans." In: *Journal of Human Genetics*, pp. 1–13.

Moore, Jason H, Folkert W Asselbergs, and Scott M Williams (2010). "Bioinformatics challenges for genome-wide association studies." In: *Bioinformatics* 26.4, pp. 445–455.

Morgan, Thomas H (1932). "The rise of genetics." In: *Science* 76.1969, pp. 261–267.

Morrison, Margaret (2002). "Modelling populations: Pearson and Fisher on Mendelism and biometry." In: *The British journal for the philosophy of science* 53.1, pp. 39–68.

Mouradov, Aidyn, Frédéric Cremer, and George Coupland (2002). "Control of flowering time: interacting pathways as a basis for diversity." In: *The Plant Cell* 14.suppl 1, S111–S130.

Mustroph, Angelika et al. (2009). "Profiling translatomes of discrete cell populations resolves altered cellular priorities during hypoxia in Arabidopsis." In: *Proceedings of the National Academy of Sciences* 106.44, pp. 18843–18848.

Nishino, Jo et al. (2018). "Sample size for successful genome-wide association study of major depressive disorder." In: *Frontiers in genetics* 9, p. 227.

Nordborg, Magnus et al. (2002). "The extent of linkage disequilibrium in Arabidopsis thaliana." In: *Nature genetics* 30.2, pp. 190–193.

Oakley, Christopher G et al. (2014). "QTL mapping of freezing tolerance: links to fitness and adaptive trade-offs." In: *Molecular Ecology* 23.17, pp. 4304–4315.

Ozaki, Kouichi et al. (2002). "Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction." In: *Nature genetics* 32.4, pp. 650–654.

Padmanabhan, Sandosh et al. (2010). "Genome-wide association study of blood pressure extremes identifies variant near UMOD associated with hypertension." In: *PLoS Genet* 6.10, e1001177.

Patro, Rob et al. (2017). "Salmon provides fast and bias-aware quantification of transcript expression." In: *Nature methods* 14.4, pp. 417–419.

Postma, Froukje M and Jon Ågren (2016). "Early life stages contribute strongly to local adaptation in Arabidopsis thaliana." In: *Proceedings of the National Academy of Sciences* 113.27, pp. 7590–7595.

Postma, Froukje M, Sverre Lundemo, and Jon Ågren (2016). "Seed dormancy cycling and mortality differ between two locally adapted populations of Arabidopsis thaliana." In: *Annals of Botany* 117.2, pp. 249–256.

Pouteau, Sylvie and Catherine Albertini (2009). "The significance of bolting and floral transitions as indicators of reproductive phase change in Arabidopsis." In: *Journal of Experimental Botany* 60.12, pp. 3367–3377.

Price, Nicholas et al. (2018). "Combining population genomics and fitness QTLs to identify the genetics of local adaptation in Arabidopsis thaliana." In: *Proceedings of the National Academy of Sciences* 115.19, pp. 5028–5033.

Prothero, Donald R (2007). *Evolution: what the fossils say and why it matters*. Columbia University Press.

Prud'homme, Benjamin, Nicolas Gompel, and Sean B Carroll (2007). "Emerging principles of regulatory evolution." In: *Proceedings of the National Academy of Sciences* 104.suppl 1, pp. 8605–8612.

Punnett, Reginald Crundall et al. (1950). "Early days of genetics." In: *Heredity* 4, pp. 1–10.

Qi, Huan et al. (2018). "PlaD: a transcriptomics database for plant defense responses to pathogens, providing new insights into plant immune system." In: *Genomics, proteomics & bioinformatics* 16.4, pp. 283–293.

Qiang, Xiaoyu et al. (2012). "Piriformospora indica—a mutualistic basidiomycete with an exceptionally large plant host range." In: *Molecular plant pathology* 13.5, pp. 508–518.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Risch, Neil and Kathleen Merikangas (1996). "The future of genetic studies of complex human diseases." In: *Science* 273.5281, pp. 1516–1517.

Romero, Irene Gallego, Ilya Ruvinsky, and Yoav Gilad (2012). "Comparative studies of gene expression and the evolution of gene regulation." In: *Nature Reviews Genetics* 13.7, pp. 505–516.

Ryan, Peter R et al. (2016). *Plant roots: understanding structure and function in an ocean of complexity*.

Schluter, Dolph and Gina L Conte (2009). "Genetics and ecological speciation." In: *Proceedings of the National Academy of Sciences* 106.Supplement 1, pp. 9955–9962.

Schork, Nicholas J et al. (2000). "Linkage disequilibrium analysis of biallelic DNA markers, human quantitative trait loci, and threshold-defined case and control subjects." In: *The American Journal of Human Genetics* 67.5, pp. 1208–1218.

Seren, Ü et al. (2016). *AraPheno: a public database for Arabidopsis thaliana phenotypes*. Vol. 45. D1. Oxford University Press, pp. D1054–D1059.

Signor, Sarah A and Sergey V Nuzhdin (2018). "The evolution of gene expression in cis and trans." In: *Trends in Genetics* 34.7, pp. 532–544.

Sorenson, Reed and Julia Bailey-Serres (2015). "Rapid immunopurification of ribonucleoprotein complexes of plants." In: *Plant Functional Genomics*. Springer, pp. 209–219.

Srikanth, Anusha and Markus Schmid (2011). "Regulation of flowering time: all roads lead to Rome." In: *Cellular and molecular life sciences* 68.12, pp. 2013–2037.

StataCorp, LLC (2019). *Stata Statistical Software. Release 16.[software]. College Station, TX*.

Stinchcombe, John R et al. (2004). "A latitudinal cline in flowering time in Arabidopsis thaliana modulated by the flowering time gene FRIGIDA." In: *Proceedings of the National Academy of Sciences* 101.13, pp. 4712–4717.

Sung, Sibum and Richard M Amasino (2004). "Vernalization in Arabidopsis thaliana is mediated by the PHD finger protein VIN3." In: *Nature* 427.6970, pp. 159–164.

Tabangin, Meredith E, Jessica G Woo, and Lisa J Martin (2009). "The effect of minor allele frequency on the likelihood of obtaining false positives." In: *BMC proceedings*. Vol. 3. 7. Springer, pp. 1–4.

Templeton, Alan R (1998). "The complexity of the genotype-phenotype relationship and the limitations of using genetic "markers" at the individual level." In: *Science in context* 11.3-4, pp. 373–389.

Teufel, Felix et al. (2021). "Body-mass index and diabetes risk in 57 low-income and middle-income countries: a cross-sectional study of nationally representative, individual-level data in 685616 adults." In: *The lancet* 398.10296, pp. 238–248. DOI: https://doi.org/10.1016/S0140-6736(21)00844-8.

Thomas, Duncan (2010). "Gene–environment-wide association studies: emerging approaches." In: *Nature Reviews Genetics* 11.4, pp. 259–272.

Tian, Dacheng et al. (2002). "Signature of balancing selection in Arabidopsis." In: *Proceedings of the National Academy of Sciences* 99.17, pp. 11525–11530.

Togninalli, Matteo et al. (2020). "AraPheno and the AraGWAS Catalog 2020: a major database update including RNA-Seq and knockout mutation data for Arabidopsis thaliana." In: *Nucleic acids research* 48.D1, pp. D1063–D1068.

Turesson, Göte (1922). "The species and the variety as ecological units." In: *Hereditas* 3.1, pp. 100–113.

Visscher, Peter M et al. (2017). "10 years of GWAS discovery: biology, function, and translation." In: *The American Journal of Human Genetics* 101.1, pp. 5–22.

Wakefield, Jon (2012). "Commentary: Genome-wide significance thresholds via Bayes factors." In: *International journal of epidemiology* 41.1, pp. 286–291.

Wang, Meiyue and Shizhong Xu (2019). "Statistical power in genome-wide association studies and quantitative trait locus mapping." In: *Heredity* 123.3, pp. 287–306.

Wang, YAN et al. (2011). "Infection of Arabidopsis thaliana by Phytophthora parasitica and identification of variation in host specificity." In: *Molecular Plant Pathology* 12.2, pp. 187–201.

Warren, Randall F et al. (1998). "A mutation within the leucine-rich repeat domain of the Arabidopsis disease resistance gene RPS5 partially suppresses multiple bacterial and downy mildew resistance genes." In: *The Plant Cell* 10.9, pp. 1439–1452.

Weigel, Detlef and Richard Mott (2009). "The 1001 genomes project for Arabidopsis thaliana." In: *Genome biology* 10.5, pp. 1–5.

Wickham, Hadley and Evan Miller (2020). *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*. R package version 2.3.1. URL: https://CRAN.R-project.org/package=haven.

Wooster, Richard et al. (1995). "Identification of the breast cancer susceptibility gene BRCA2." In: *Nature* 378.6559, pp. 789–792.

World Health Organization (2005). *WHO STEPS surveillance manual: the WHO STEPwise approach to chronic disease risk factor surveillance*. Tech. rep. World Health Organization.

Yamaguchi, Ayako et al. (2005). “TWIN SISTER OF FT (TSF) acts as a floral pathway integrator redundantly with FT.” In: *Plant and Cell Physiology* 46.8, pp. 1175–1189.

Yang, Shujun et al. (2010). “Narrowing down the targets: towards successful genetic engineering of drought-tolerant crops.” In: *Molecular plant* 3.3, pp. 469–490.

Yu, Jianming et al. (2006). “A unified mixed-model method for association mapping that accounts for multiple levels of relatedness.” In: *Nature genetics* 38.2, pp. 203–208.

Zan, Yanjun and Örjan Carlborg (2018). “A multilocus association analysis method integrating phenotype and expression data reveals multiple novel associations to flowering time variation in wild-collected Arabidopsis thaliana.” In: *Molecular ecology resources* 18.4, pp. 798–808.

— (2019). “A polygenic genetic architecture of flowering time in the worldwide *Arabidopsis thaliana population*.” In: *Molecular biology and evolution* 36.1, pp. 141–154.

Zhang, Lei and José M Jiménez-Gómez (2020). “Functional analysis of FRIGIDA using naturally occurring variation in Arabidopsis thaliana.” In: *The Plant Journal*.

Zhang, Xiangbo et al. (2007). “Constitutive expression of CIR1 (RVE2) affects several circadian-regulated processes and seed germination in Arabidopsis.” In: *The Plant Journal* 51.3, pp. 512–525.

Zheng, Yu et al. (2016). “Histone deacetylase HDA9 negatively regulates salt and drought stress responsiveness in Arabidopsis.” In: *Journal of experimental botany* 67.6, pp. 1703–1713.

Zhou, Xiang and Matthew Stephens (2012). “Genome-wide efficient mixed-model analysis for association studies.” In: *Nature genetics* 44.7, p. 821.

Zhu, Huanhuan and Xiang Zhou (2020). “Statistical methods for SNP heritability estimation and partition: A review.” In: *Computational and Structural Biotechnology Journal* 18, pp. 1557–1568.