Genome-Wide Analysis of *N*ucleotide-Binding Domain *L*eucine-Rich *R*epeat (NLR) Variation Patterns in *Arabidopsis thaliana*

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Darya Karelina MRes
aus der Ukraine

Tübingen
2019

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	25.03.2019
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Detlef Weigel
2. Berichterstatter:	Prof. Dr. Thorsten Nürnberger

Zusammenfassung

NLR (Nucleotide-Binding Domain Leucine-Rich Repeat)-Proteine spielen eine zentrale Rolle bei der Pflanzenimmunität, indem sie Pathogenproteine direkt kontrollieren oder die Auswirkungen von Pathogenen auf Pflanzenproteine überwachen. Um mit den oft raschen Veränderungen des Erregerspektrums fertig zu werden, haben Pflanzen typischerweise ein vielfältiges NLR-Gen-Repertoire. Darüber hinaus variieren sowohl die Kopienzahl der NLR-Gene als auch ihre Sequenzen innerhalb einer Pflanzenart stark, was vermutlich den sich ändernden Pathogen-Druck widerspiegelt. Viele offene Fragen sind mit NLR-Repertoires verbunden. Zum Beispiel verleihen zwei *Arabidopsis thaliana* NLR-Gene, *RPM1* (*RESISTANCE TO PSEUDOMONAS SYRINGAE PV. MACULICOLA 1*) und *RPS5* (*RESISTANCE TO P. SYRINGAE 5*), Resistenz gegen bestimmte Pathogene, aber es ist bekannt, dass sie Fitnesskosten für die Pflanze mit sich bringen, die bis zu 10% erreichen – für natürliche Fitness ein erstaunlich hoher Wert. Beide Gene weisen auch einen langanhaltenden Präsenz/Abwesenheits (P/A) Polymorphismus auf, wobei einige Pflanzen in einer Population das Gen tragen, aber es in den anderen fehlt. Die Frequenz von P/A-Genen unter allen Genen wurde auf etwa 9% im *A. thaliana*-Genom geschätzt. Dies wirft die Frage auf, wie sich Pflanzen so hohe Fitnesskosten leisten können und wie häufig Gene wie *RPM1* und *RPS5* in Pflanzengenomen und NLR-Repertoires einzelner Individuen vorkommen. NLR-Gene sind auch aus der Sicht der Genomik als die variabelste Genfamilie in Pflanzen interessant. Sie gehören zu den repetitivesten Familien, und sind oft als Cluster von Tandem-Duplikaten im Genom präsent. Aus diesen Gründen eignen sich NLR-Gene weder für eine normale Referenz-basierte Analyse noch für *de novo* Assemblierung. Trotz der Verfügbarkeit von großen Mengen an Sequenzierungsdaten für die Modellpflanze *A. thaliana* existiert somit keine detaillierte Bewertung der Variation des gesamten Repertoires von NLR-Genen innerhalb dieser Pflanzenart.

In dieser Studie visualisiere und analysiere ich Muster der Diversität, die für NLR-Gene in *A. thaliana* charakteristisch sind. Im ersten Kapitel, schlage ich eine Methode für die Profilierung komplexer hypervariabler Regionen des Genoms vor, die auf kurzen Reads basiert, auch wenn nur eine zuverlässige Referenz verfügbar ist. Im zweiten Kapitel wende ich die Methode auf die Referenzmenge von 163 NLR-Genen in 80 Akzessionen von *A. thaliana* an. Im dritten Kapitel führe ich einen Vergleich zwischen Arten durch, indem ich meine Methode auf 26 Akzessionen von *Arabidopsis lyrata* und 22 Akzessionen von *Capsella rubella* anwende, die die nächste Art und Gattung zu *A. thaliana* darstellen. Ich vergleiche diese Ergebnisse mit Artenpolymorphismen in *A. thaliana*. Ich fand heraus, dass NLR-Muster der Diversität in drei Kategorien fallen: konserviert (vorhanden), P/A, und Gene mit komplexer Variation. Ich identifizierte 53 konservierte NLR-Gene, von denen 24 auch in *A. lyrata* und *C. rubella* vorhanden sind, und 52 P/A-Gene, von denen welche, wie zum Beispiel *ADR1-L3* (*ACTIVATED DISEASE RESISTANCE 1-LIKE3*), auch ein P/A-ähnliches Muster in den beiden anderen Arten aufwiesen. Ich kombinierte Variationsmuster mit genomischen Kontextinformationen und Nukleotid-Diversitätsinformationen, um es möglich zu machen, genomweit P/A-Genen mit Diversitätsmustern zu identifizieren, die *RPM1*- und *RPS5*- ähnlich sind. Ich führte eine genomweite Assoziationsstudie (GWAS) mit dem genomweitem P/A-Polymorphismus durch und fand, dass *RPS5* zwischen den signifikantesten Genen über mehrere Phänotypen hinweg ist. *RPM1* und *RPS5* zeigen auch die zweit- und dritt-geringste Variationsrate in *A. lyrata* unter NLR P/A-Genen. Ich schließe daraus, dass Gene wie *RPM1* und *RPS5* unter den NLR-Genen und im gesamten Genom selten sind. Ich beobachtete keine statistisch signifikante Anreicherung für Domänenarchitekturtyp von NLR-Genen (TIR-/CC-NB-LRR) oder für die genomische Anordnung (einzeln oder in Clustern) im Vergleich zwischen Pflanzenakzessionen. Cluster-Gene waren jedoch in beiden Fällen variabler als einzelne Gene, und ich fand statistisch signifikante Anreicherung für einzelne *A. thaliana* NLR Gene die waren auch in *A. lyrata* und *C. rubella* anwesend. Unsere

Ergebnisse erlauben neue Einblicke in das NLR-Repertoire in *Arabidopsis*-Genomen.

Abstract

NLR (Nucleotide-Binding Domain Leucine-Rich Repeat) proteins have central roles in plant immunity, by directly detecting pathogen proteins, or by monitoring the effects of pathogens on plant proteins. To cope with the often rapid changes in the spectrum of pathogens they encounter, plants typically have a diverse NLR gene repertoire. Moreover, both the copy number of NLR genes and their sequences vary greatly within a species, presumably reflecting changing pathogen pressures. Many outstanding questions are associated with NLR repertoires. For example, two *Arabidopsis thaliana* NLR genes, *RPM1* (*RESISTANCE TO PSEUDOMONAS SYRINGAE PV. MACULICOLA 1*) and *RPS5* (*RESISTANCE TO P. SYRINGAE 5*), confer resistance to specific pathogens, but are known to carry fitness costs for the plant that can reach 10% - a surprisingly high value for natural fitness. Both genes also exhibit long-standing presence/absence (P/A) polymorphism, where some plants in a population will carry the gene, but it will be absent in others. The frequency of P/A genes has been estimated to be around 9% of all genes in the *A. thaliana* genome. This raises the question of how plants can afford such high fitness costs, and how common genes like *RPM1* and *RPS5* are in plant genomes and NLR repertoires of specific individuals. NLR genes are also interesting from a genomics perspective as the most variable gene family in plants. They are among the most repetitive families, often present as clusters of tandem duplicates in the genome. For these reasons, NLR genes do not lend themselves to regular reference-based analysis nor to *de novo* assembly. Thus, despite availability of large amounts of sequencing data for the model plant *A. thaliana*, no detailed evaluation of variation in the complete repertoire of NLR genes within this species exists.

In this study, I visualize and analyze patterns of diversity characteristic of NLR genes in *A. thaliana*. In the first chapter, I propose a method for the profiling of complex hypervariable regions of the genome based on short reads, even when

only one reliable reference is available. In the second chapter, I apply the method to the reference set of 163 NLR genes in 80 accessions of *A. thaliana*. In the third chapter, I carry out a between-species comparison by applying my method to 26 accessions of *Arabidopsis lyrata* and 22 accessions of *Capsella rubella*, which represent the closest species and genus to *A. thaliana*, respectively. I compare these results with within-species polymorphism in *A. thaliana*. I found that NLR patterns of diversity fall into three categories: conserved (present), P/A genes, and genes with complex variation patterns. I identified 53 conserved NLR genes, of which 24 are were also present in *A. lyrata* and *C. rubella*, and 52 P/A genes, of which several, such as *ADR1-L3* (*ACTIVATED DISEASE RESISTANCE 1-LIKE3*), also had P/A-like pattern in the two other species. I combined variation patterns with genomic context and nucleotide diversity information to make it possible to identify P/A genes with diversity patterns reminiscent of *RPM1* and *RPS5* genome-wide. I carried out a genome-wide association study (GWAS) on the P/A polymorphism genome-wide, and found that *RPS5* is among the most significant genes across multiple phenotypes. *RPM1* and *RPS5* also show the second and third highest conservation in *A. lyrata* among NLR P/A genes. I conclude that genes like *RPM1* and *RPS5* are rare both among NLR genes and in the whole genome. I found no statistically significant enrichment for domain architecture type of NLR genes (TIR-/CC-NB-LRR) nor for genomic arrangement (single/clustered) in within-species comparison. However, clustered genes were more variable than single genes and there was significant enrichment of single genes among *A. thaliana* NLR genes that were also present in *A. lyrata* and *C. rubella*. My results reveal new insights into the NLR repertoires in *Arabidopsis* genomes.

Acknowledgements

I would like to thank my supervisor, Prof. Dr. Detlef Weigel, for the opportunity to work on a very exciting topic in his lab. I am also very grateful for his supervision and guidance on project direction over multiple years, despite being constantly engaged in an enormous number of important projects.

I would like to thank Prof. Dr. Gunnar Rätsch, in whose lab I was based for the first year and a half of my PhD, for providing supervision and guidance on project direction, which were essential to the project, and for opportunities to meet many excellent and inspiring scientists and students in his lab.

I would like to thank my university supervisor, Prof. Dr. Thorsten Nürnberger, for discussing the project with me on multiple occasions and for offering valuable suggestions and new perspectives. Prof. Nürnberger also introduced me to Prof. Dr. Laura Rose, whom I would also like to thank for discussing the project with me, and for offering valuable advice.

I was fortunate to work with Dr. François Vasseur, who at that time was a postdoc in Detlef Weigel's lab, and want to thank him for suggesting a collaboration project and for setting a new standard for me in terms of quality and quantity of work that can be achieved in a short time. It permanently changed my way of thinking and shifted my perspective on what was possible. I would also like to thank Dr. Eunyoung Chae, then a postdoc at Detlef Weigel's lab, who introduced me to the project and to the relevant literature on the topic, as well as discussed the project with me on multiple occasions. As a co-author on one of her papers, I was greatly impressed by her attention to detail and great professionalism. I would also like to thank Dr. Rui Wu, also a postdoc in Detlef

Weigel's lab, for helping me frame specific, meaningful questions and for teaching me persistence and resilience by her example.

For valuable discussions and answering my questions, I would like to thank: Maricris Zaidem, Yi Zhong, Congmao Wang, Domink Grimm, Vipin Sreedharan, Korbinian Schneeberger, Jonas Müller, Jörg Hagmann, André Kahles, Eshita Sharma, Claude Becker, Ya-Long Guo, Daniel Koenig, Ignacio Rubio, Beth Rowan, Wangsheng Zhu, Dino Jolic, Euncheon Lim. I would like to thank André Noll, Dagmar Sigurdardottir and Hülya Wicher for computational and administrative help.

Table of Contents

1 Introduction

1.1 Overview

The early 21st century saw the widespread application of genomics to understanding issues of biological and medical relevance. The current work applies genomics methods to the study of evolutionary patterns in plant immunity. In particular, I look at genetic variation in a key family of plant immune genes called the *n*ucleotide-binding domain *l*eucine-rich *r*epeat (NLR) family.

NLR genes are found in all three multicellular kingdoms of life – animals, plants and fungi (reviewed in Uehling et al., 2017). Individual constitutive modules of NLR genes have even been found in unicellular organisms (Yue et al., 2012), placing them among the most ancient and fundamental gene families. Plant NLR genes are more numerous that animal NLR genes, typically numbering in the hundreds, compared to tens in animals (Uehling et al., 2017). In plants, NLR genes predominantly function in innate immunity, enabling defence from a myriad of potential pathogens constantly present in the plant environment. NLR genes constitute the largest plant immune receptor family responsible for the specific pathogen recognition and the activation of downstream signaling.

As the most variable gene family in plants (Clark et al., 2007; Bakker et al., 2006; Li et al., 2015; Jones et al., 2016; Zhang et al., 2017; Meunier and Broz, 2017; Białas et al., 2018; Maekawa et al., 2011), while also being among the largest and the most repetitive families, NLR genes offer a rich source of complex natural variation patterns, which can be used for investigating local adaptation and gene function. As one of the fastest evolving gene families in plants, NLR genes are also crucial to understanding plant evolution. NLR genes have also been found to underlie phenomena that could potentially contribute to reproductive isolation, such as hybrid incompatibility (Bomblies et al., 2007, 2010; Rieseberg and Blackman, 2010; Chae et al., 2014).

In this study, I focus on NLR genes in *Arabidopsis thaliana*. *Arabidopsis thaliana* has been an established model species for plant research for many decades. Research on *A. thaliana* has benefitted greatly from the recent advances and genome sequencing. Thus, currently, in addition to such advantages as ease of manipulability and a small genome, which made it the preferred plant model species, it currently benefits from a large number of sequence, phenotype and functional genomics resources as well as from a vast body of scientific knowledge that accumulated around its biology (e.g. 1001 Genomes Sequencing Consortium, 2016; Koornneef and Meinke, 2010; Weigel, 2012; Krämer, 2015; Weigel and Nordborg, 2015; Provart et al., 2016; Hehl, 2017; Lv et al., 2017; Seren et al., 2017; Chen et al., 2018; Togninalli et al., 2018; Woodward and Bartel, 2018). All this makes *A. thaliana* the ideal option for disentangling complex aspects of NLR biology and genomics, which can then inform studies in crop species, and possibly even animals, such as humans (Xu and Møller, 2011).

From an agricultural standpoint, NLR genes are at the core of increasing crop yield by preventing pathogen-driven losses. Plant yields are important as food for the increasing world population, as well as a source of biomass to produce packaging and fuels. Advances in NLR research also have the potential to reduce the use chemicals in agriculture, such as fungicides, which would benefit the environment.

NLR genes, however, present not only a rich source of natural variation and opportunity for genomic studies, but also a rich source of challenges for genomic assembly. From the standpoint of genomics, challenges of genome assembly and interpretation often converge on the challenges of NLR assembly. As a result of their variability and repetitiveness, for a long time polymorphism in NLR genes was difficult to unambiguously characterize, with consequent difficulties in applying methods such as Genome-Wide Association Studies (GWAS).

While there have been previous studies on NLR diversity in *A. thaliana* (e.g. Kuang et al., 2004; Bakker et al., 2006; Tan et al., 2016, 2012), these have been restricted to a subset of NLR genes, analyzed only a portion of each NLR gene, or investigated only a limited number of accessions. In order to deal with the genomic complexities characteristic of this family, I developed an intuitive visualization approach that enables categorization of NLR genes based on their patterns of diversity. I apply this approach to understanding how gene repertoires differ among individuals of the model plant species *A. thaliana*, as well as between *A. thaliana* and other species.

Since this study is positioned at the intersection of *A. thaliana* genomics and NLR gene biology, I begin by providing context on these two broad fields of research. A brief history of recent advances in *A. thaliana* genomics will thus be followed by an introduction to NLR structure, function and evolution. Within that section, I will also describe previous attempts to categorize NLR genes, which this study aims to extend, and provide background on the two types of gene-level variation patterns previously observed in the NLR family – Presence/Absence (P/A) genes and conserved genes. Finally, I define the objectives of the study within the framework of summarized literature.

1.2 A Brief History of Arabidopsis Genomics

1.2.1 Arabidopsis – the Model Species for Plant Genomics

Arabidopsis thaliana, a small crucifer flowering plant (Figure 1), is among the most widely used model species in modern plant biology. Due to colonization, the species today has a nearly worldwide range and exhibits many phenotypic differences among strains (accessions; Weigel et al., 2012). Due to its advantages of a small size, short generation time and ease of genetic manipulability, its biology has been studied to a great extent. From a genomics perspective, its small nuclear genome with relatively few repetitive sequences offers an additional advantage. Furthermore, its self-fertility allows one to readily obtain many homozygous individuals by the inbreeding of natural accessions. All these advantages have contributed to making *A. thaliana* an ideal candidate for genomic studies.

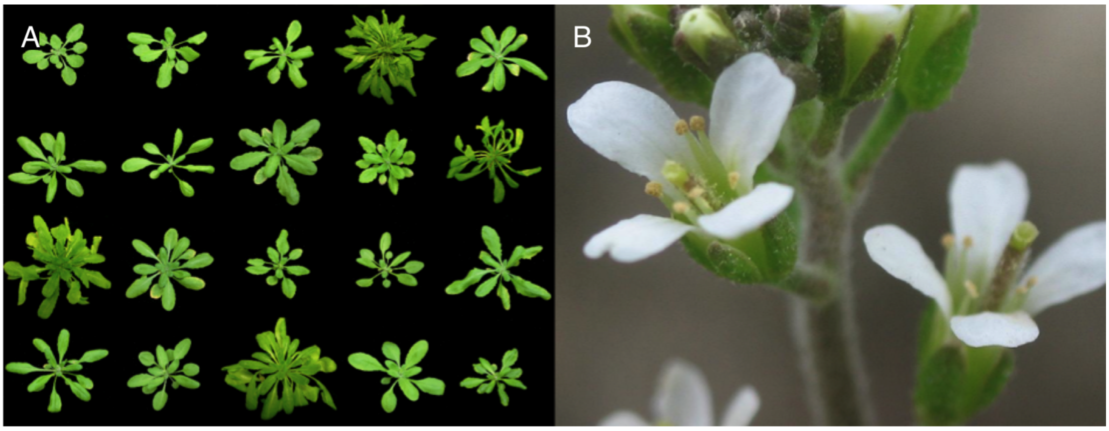

Figure 1. Variants of *Arabidopsis thaliana* (A). Source: Sureshkumar Balasubramanian and Janne Lempe, MPI for Developmental Biology. Inflorescences of *A. thaliana* (B). Source: es.wikipedia.org.

1.2.2 Sequencing of Arabidopsis - the First Plant Genome

Arabidopsis thaliana was the first plant whose genome was completely sequenced (Arabidopsis Genome Initiative, 2000). The *A. thaliana* nuclear genome is around 150 megabase in size and, according to the latest version of the *A. thaliana* genome annotation (TAIR10, http://arabidopsis.org), contains around 27,000 protein coding genes. It has been estimated that around 60% of the genome consists of large duplicated segments in varying degrees of conservation and homology (Arabidopsis Genome Initiative, 2000).

1.2.3 First Whole-Genome Resequencing Studies

Completion of the *A. thaliana* genome enabled whole-genome resequencing studies, with the aim of characterizing intraspecific (within-species) variation. In resequencing studies, the genome of interest is compared to a reference genome in order to identify new polymorphisms relative to the reference (variant analysis) or to test the status of known ones (genotyping). In the case of *A. thaliana* research, the reference genome commonly used is the Col-0 accession of *A. thaliana*, which was the original genome sequenced and for which therefore a high-quality assembly is available (Arabidopsis Genome Initiative, 2000).

In the first whole-genome study of intraspecific variation, Clark and colleagues (2007) used microarray technology to genotype single nucleotide substitutions (SNPs) - the most common type of polymorphism - and deletions. The high-density microarrays had at that time been applied to human and mouse, and the study applied this technology to 20 natural *A. thaliana* accessions. In the resequencing microarray technology, for each genomic position, on forward and reverse strands, a quartet of 25 nucleotide probes was designed, such that each of the four nucleotides was represented at the central position. The probes were then annealed to amplified genomic DNA, and the differences in the strength of annealing among the probe quartets were used to call SNPs. Microarray-based resequencing of 20 accessions allowed to identify 1 million nonredundant SNPs, and to conclude that about 4% of the reference genome was absent or highly diverged in the resequenced accessions. In that study, it was also found that gene families vary in their patterns of polymorphism, and that the three gene families with the most major effect changes were NLR genes, F-box genes and receptor-like kinase (RLK) genes.

In addition to Clark and colleagues (2007), multiple array-based resequencing studies have found that a substantial fraction of the reference genome was missing in the queried accessions (Borevitz et al., 2007; Zeller et al., 2008;

Plantegenet et al., 2009). This implied that, conversely, there were genomic sequences in other accessions missing from the reference genome, and lead to the quest to identify such sequences in non-reference accessions (reviewed in Weigel, 2012).

1.2.4 High Throughput Sequencing

Shortly after the study by Clark and colleagues (2007), SNP calling via remapping began to gain acceptance. This method relied on first obtaining genomic sequences for the accessions of interest, in the form of short overlapping reads (Metzker, 2010). These reads were then aligned, or mapped, to a high-quality reference accession genome *in silico*, to identify bases that differ (SNP calling) or identify other types of variants, such as insertions and deletions (indel calling). This approach was enabled by advances in high-throughput sequencing technologies, which were significantly less costly than earlier sequencing methods, and could provide higher genomic coverage, which can be defined as the number of short reads that include a given nucleotide position.

Originally developed for human genomics, these methods were quickly adopted by the pant genomics community as progress in human genomics was closely paralleled by progress in plant genomics. Methods such as Illumina short read sequencing facilitated intraspecific variation analyses and resulted in a number of whole-genome resequencing projects (e.g. Ossowski et al., 2008, Cao et al., 2011; Gan et al., 2011; Long et al., 2013; 1001 Genomes Sequencing Consortium, 2016).

There were, however, trade-offs to the lower cost of high-throughput sequencing technologies. These included a higher error rate and a shorter length of sequence reads. The higher error rate could in part be compensated by the increased genomic coverage. The shorter read lengths, however, characteristic of early high-throughput sequencing technologies, meant that the reads' relative positions and genomic arrangement had to be determined - a non-trivial assembly task.

1.2.5 Reference-Guided and *de novo* Assembly

Remapping, introduced in the previous section, became a standard approach for insertion, deletion and SNP calling in whole-genome resequencing projects. In addition to remapping, two main groups of methods emerged to deal with the assembly challenges. Reference-guided assembly relied on comparing the sequenced reads directly to an available high-quality reference genome in order to determine their genomic positions relative to the reference. In contrast, *de novo* assembly methods relied on sequence similarity and overlap between short reads themselves, without use of reference, in order to establish their positions in the genome relative to one another. Multiple software tools have been developed to deal with these tasks (reviewed in Basantani et al., 2017), such as BWA (Li and Durbin, 2009) and GenomeMapper (Schneeberger, 2009).

1.2.6 Assembly Challenges: High Divergence and Repetitiveness

An inherent limitation of reference-based assembly with short reads is the difficulty it presents for sequences that are highly diverged between the reference and the genome of interest, as well as for sequences that are repetitive or occur multiple times, either in the reference or in the resequenced genome.

For sequences that are highly diverged or entirely absent from the reference, no corresponding position in the reference genome may be determined and the sequences are usually discarded from further analysis. For sequences that are repetitive in either genome, no unique mapping position may be unambiguously determined, leading to the issue of cross-mapping. For example, a read from a paralogous gene may map to a gene of interest and a single nucleotide difference between the two paralogs may be erroneously assigned as an SNP in the gene of interest. Thus, for repetitive sequences, no unique genomic position can be determined.

Repetitiveness is unevenly distributed in the *A. thaliana* genome, and most of the genes in resequenced genomes can be effectively resolved using reference-based assembly. In certain instances, however, highly divergent and repetitive sequences can also be of high biological interest. Such is the case with the NLR family of *A. thaliana* immune genes, which are the subject of this study.

1.2.7 1001 Genomes Project and Future Outlook

In *A. thaliana*, resequencing efforts have culminated in the completion of the 1001 genomes sequencing project (1001 Genomes Consortium, 2016). Compared to earlier projects (Cao et al., 2011; Gan et al., 2011; Long et al., 2013), a larger and more representative sample of 1,135 inbred *A. thaliana* accessions was sequenced and made publicly available as a high-quality catalogue of *A. thaliana* intraspecific polymorphism and a resource to facilitate GWAS and forward genetic studies in the future.

Other plant species, including crops and wild species, are continually being sequenced and resequencing studies carried out (e.g. Wang et al., 2018; https://jgi.doe.gov/csp-2018-leebens-mack-open-green-genomes-initiative/; Cheng et al., 2018; reviewed in e.g. Li and Harkess, 2018; Chen et al., 2018; Koenig and Weigel, 2015). However, due to the accumulation of biological knowledge around *A. thaliana* and its convenience as a model organism, it is still the best candidate for elucidating NLR functional role and evolution.

1.2.8 Comparison of *A. thaliana* with Closely Related Species

Arabidopsis thaliana belongs to the family Brassicaceae, also known as the crucifers, which contains 338 genera and 3709 species (Warwick et al., 2006). Brassicaceae is an important group of angiosperms, and includes multiple economically important crops, such as cabbage, cauliflower, broccoli and rapeseed, the latter being used for edible oil production. Multiple Brassicaceae genomes have been sequenced, partially to provide context for understanding the *A. thaliana* genome and variation in it.

In this study, I compare *A. thaliana* to its close Brassicaceae relatives *Arabidopsis lyrata* and *Capsella rubella*. *Arabidopsis lyrata* is the closest extant species (sister species) to *A. thaliana*, and *C. rubella* belongs to the most closely related genus for *Arabidopsis*: *Capsella*. Both are model species and have been well studied.

Brassicaceae on average tend to have about 200 megabase genomes distributed over 8 chromosomes (Johnston et al., 2005; Oyama et al., 2008; Koenig and Weigel, 2015). *Arabidopsis thaliana*, however, is one of the exceptions, with 5 chromosomes only and a ca. 150 megabase genome (Yogeeswaran et al., 2005; Lysak et al., 2006). In terms of mating system, outcrossing is thought to be the ancestral state of the family (as in all flowering plants), as shown by studies of polymorphism at the self-incompatibility (SI) locus (Guo et al., 2009; Paetsch et al., 2006; 2010).

Arabidopsis lyrata and *A. thaliana* both belong to the same genus and separated about 13 million years ago (Beilstein et al., 2010). Unlike the self-compatible *A. thaliana*, *A. lytrata* is an outcrossing species. Its genome is significantly larger, and thought to be the ancestral state for the family (Hu et al., 2011). The size difference is due to the around 6,000 additional genes in *A. lyrata* over *A. thaliana*, and in deletions in noncoding DNA and transposons, which are ongoing in *A. thaliana* (Hu et al., 2011).

Capsella rubella and *A. thaliana* diverged about twice as long ago as did *A. lyrata* and *A. thaliana* (Koch et al., 2005). *Capsella rubella* has a closely related species within the *Capsella* genus: *Capsella grandiflora*. Although these species separated less than 100,000 years ago, their mating systems are different (Foxe et al., 2009; Guo et al., 2009; Brandvain et al., 2013; Slotte et al., 2013). *Capsella rubella* is self-compatible, like *A. thaliana*, but *C. grandiflora* is not. Together, they form a model system for investigating mating system shifts. *Capsella rubella* genome, like *A. lyrata* genome, is bigger than *A. thaliana* genome. In the case of *C. rubella*, however, this is attributed to expansion of centromeric repeats (Slotte et al., 2013).

Capsella rubella is also interesting in that it is thought to have originated through an extreme population bottleneck, perhaps even speciation by a single individual (Guo et al., 2009). It has thus been additionally suggested as a model species for understanding the initial stages of divergence and adaptation (Guo et al., 2009).

1.3 An Introduction to NLR Genes

1.3.1 Role of NLR genes in Plant Immunity and beyond

NLR genes are a subgroup of plant resistance (R) genes, known to be involved in plant immunity. In *A. thaliana*, the ca. 150 NLR genes comprise around three-quarters of *R* genes, making them the largest R gene family (Meyers et al., 2003). The other three families of R genes are intracellular kinases (such as the tomato *Pto* gene), receptor-like proteins (RLP) and receptor-like kinases (RLK) (Bent, 1996).

The plant immune response can be divided into two stages (Dangl et al., 2006). The first stage is PAMP-triggered immunity (PTI), in which the plant recognises pathogen – associated molecular patterns (PAMPs) by means of RLP and RLK receptors on the plant surface, cumulatively known as pathogen recognition receptors (PRRs), and triggers the first line of defence (Figure 2). In response, pathogens secrete small molecules called effectors inside the plant cell in order to suppress the PTI. In the second stage, called effector-triggered immunity (ETI), plants recognise effectors via intracellular NLR receptors and trigger a signaling cascade that results in cell death and thereby suppression of pathogen growth, in the case of biotrophic pathogens (Dangl et al., 2006). Within this paradigm, NLR receptors typically act in the ETI stage of plant immune response (Figure 2).

This dichotomy between ETI and PTI, however, has been challenged as there is evidence for overlap between these two kinds of response (Thomma et al., 2011). In support of the ETI/PTI hypothesis there is some evidence that the PTI response is more basal and more ancient, based on comparative genomics studies of LRR-containing RLK genes (Yue et al., 2012). A systems view of plant immunity that encompasses the complexity of interactions between plant immune components, with plant immune receptors forming a single layer, has also been recently suggested (Wu et al., 2018).

In addition to acting in plant defence, NLR genes are known to underpin the phenomenon of hybrid incompatibility, in which a hybrid of two natural accessions of *A. thaliana* has constitutively activated immune defences through inappropriate recognition of self and an attendant autoimmune response (Bomblies et al., 2007, 2010; Alcázar et al., 2009, 2010, 2014; Yamamoto et al., 2010; Chae et al., 2014; Chen et al., 2014). It is not known whether these phenomena are a side effect of immunity or serve a separate function. In fungi, for example, it has been speculated that hybrid incompatibility-like phenomena that also involve NLR genes serve to maintain organismal integrity of mycelial individuals and prevent the spread of mycoviruses (Uehling et al., 2017).

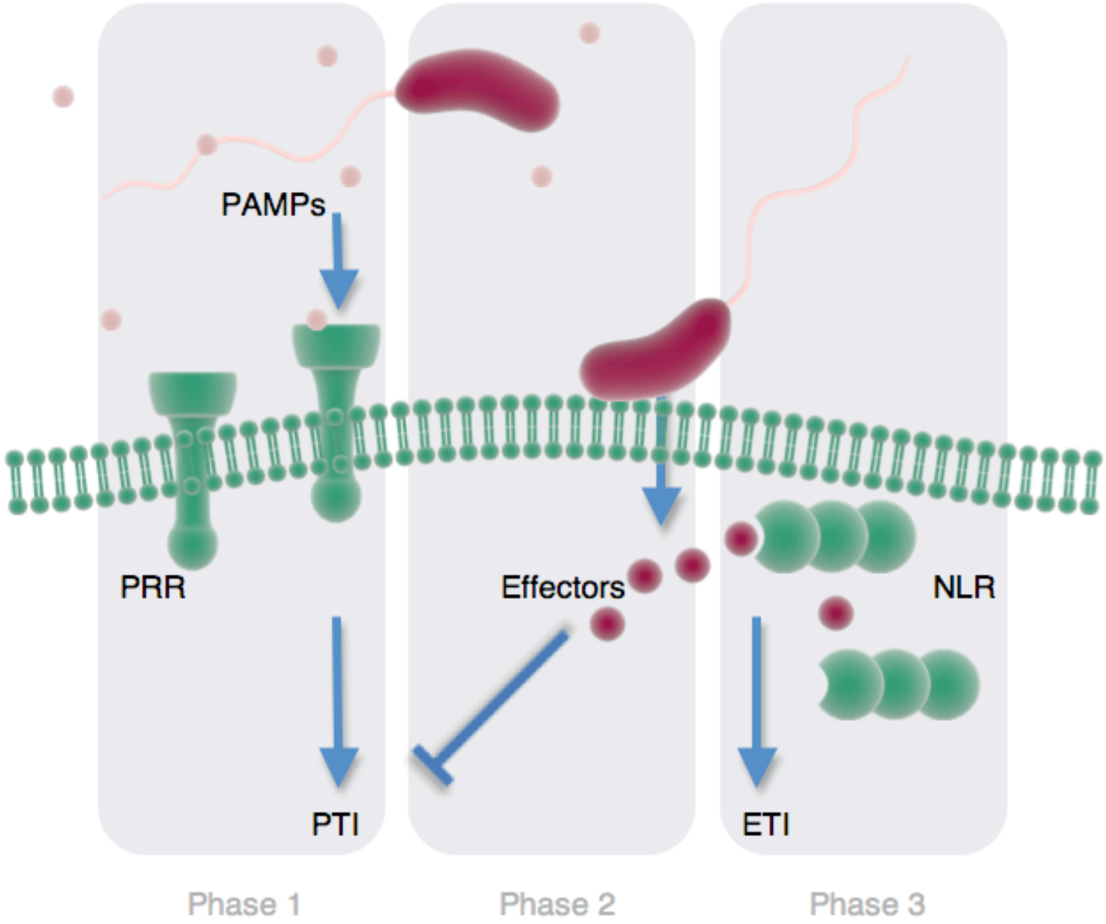

Figure 2. The traditional paradigm of ETI/PTI. Lower part of the diagram represents the plant cell. Pathogen- (or Microbe-) Associated Molecular Patterns (PAMPs/MAMPs) are recognized by the cell surface Pathogen Recognition Receptors (PRPs) in the plant. As a result of this recognition, Pathogen Triggered Immunity (PTI) response is initiated. Pathogens secrete effectors inside the plant cell to suppress PTI. However, if any of these are recognised by intracellular plant NLR receptors, an even stronger Effector Triggered Immunity (ETI) response is activated. Based on description in Jones and Dangl (2006). Image made using graphical elements from the Library of Science and Medical Illustrations (https://www.somersault1824.com/science-illustrations/).

Some NLR genes have also been assigned atypical functions, such as regulating abiotic stress. For example, *CHS1* (*CHILLING SENSITIVE 1*) regulates response to cold temperatures by limiting chloroplast damage and cell death (Zbierzak et al, 2013), although this conclusion typically comes from gain-of-function mutations, and it is not always clear what the normal function of the mutant genes is.

1.3.2 NLR Domain Architecture and Associated Function

NLR genes are characterized by the presence of two types of domains: a nucleotide-binding domain (NB) and, at the C terminus, a variable number of leucine-rich repeats (LRRs). They normally also contain either a coiled-coil (CC) domain or a Toll/Interleukin-1 receptor (TIR) domain at the N terminus (Meyers et al., 2003).

The NB domain is the most conserved of the three and is involved in ATP and GTP binding and hydrolysis (Meyers et al., 1999). The LRR domain contains a variable number of repetitive LRR modules, which normally determine the specificity in pathogen effector recognition. TIR and CC domains are thought to be primarily involved in signal transduction (Casey et al., 2016; Williams et al., 2016). Monocots, such as grasses, only have TIR-based domain architectures and tend to lack CC domains (Meyers et al., 1999; Tarr and Alexander, 2009).

The canonical plant NLR architecture, thus, consists of three building blocks: the N-terminal domain, which can be either TIR or CC; a central NB domain, which is NB-ARC in plants (named after proteins that contain it; see Abbreviations); and a C-terminal LRR domain (Figure 3). Thus, the two standard architectures in plants are TIR-NB-LRR (abbreviated TNL) and CC-NB-LRR (abbreviated CNL).

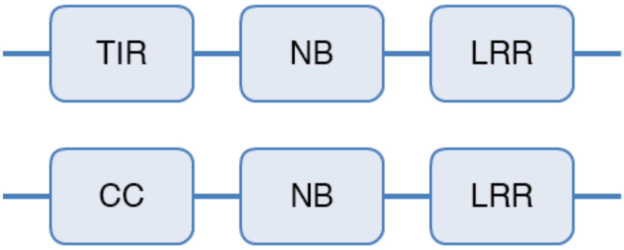

Figure 3. Schematic representation of canonical plant NLR architectures. Above is the TIR-NB-LRR (TNL) architecture and below is the CC-NB-LRR (CNL) architecture.

However, many variations on this standard architecture exist. These include “truncated” forms lacking either an N-terminal or the C-terminal domain. The *CHS1* gene mentioned above (Section 1.3.1), for example, encodes a TN protein. More complex domain combinations such as TNTNL or TTNL also exist in *A. thaliana* (Meyers et al., 2003).

Additional domains have also been found within NLR architectures. One example is the WRKY domain at the C-terminal in *RRS1* (*RESISTANCE TO RALSTONIA SOLANACEARUM 1*; Deslandes et al., 2002). In this case, the WRKY domain is thought to act as an ”integrated decoy” (ID) domain by mimicking pathogen effector targets and activating downstream signaling after binding such effectors (Cesari et al., 2014; Jones et al., 2016). Integrated decoy domains are widespread, with an estimated 10% of NLR genes carrying novel integrated domains (Ellis, 2016), and seem to derive from defence-related proteins that pathogen effectors originally targeted (reviewed in Kroj et al., 2016).

The NB domains in TNL and CNL proteins are distinguishable and segregate as monophyletic clades (Meyers et al., 1999). Other differences exist: for example, TNL type genes frequently contain multiple introns, while CNL type genes frequently encode a single exon (Meyers et al., 2003). TNL genes are rare in monocots (Meyers et al., 1999; Tarr and Alexander, 2009), despite being considered more ancient than CNL genes (Yue et al., 2012). Furthermore, TNL and CNL genes differ in their downstream signaling pathways and thus possibly in function as well (Aarts et al., 1998; Tarr and Alexander, 2009). Whether and how these genes differ in their variation patterns or evolutionary modes is not fully elucidated.

1.3.3 NLR Evolutionary History

NLR proteins belong to the superfamily of signal transduction ATPases with numerous domains (STAND), which in turn belong to the AAA+ superfamily

(Leipe et al., 2004; Maekawa et al., 2011; Rairdan and Moffett, 2007). STAND type P-loop ATPases often have a tripartite domain architecture, with a central ancient ATPase domain, which acts as a regulatory switch, surrounded by a signal-generating N-terminal effector domain and a sensory C-terminal repeat domain (Maekawa et al., 2011; Rairdan and Moffett, 2007). Studies suggest that the canonical domains of NLR proteins, such as NB-ARC, TIR and LRR, are ancient in origin and already existed in the genomes of eubacteria and archaebacteria (Yue et al., 2012).

There are striking parallels between plant immunity and animal innate immunity (Nürnberger and Brunner, 2002). Plant transmembrane PRR proteins have their equivalent in animal Toll-like receptors (TLRs) and TLR-recruited kinases (Staal and Dixelius, 2007). As for NLR proteins, not only do animals contain them but also fungi, which share the tripartite architecture of plant NLR proteins (reviewed in Uehling et al., 2017). The domains themselves, however, differ and the tripartite architectures are thought to be a result of convergent evolution (Ausubel et al., 2005; Rairdan and Moffett, 2007; Yue et al., 2012). This might reflect the fact that interacting domains have a tendency towards being fused into a single gene in the course of evolution (Marcotte et al., 1999, Enright et al., 1999).

To provide a context for plant NLR domains and domain architectures, I can draw a comparison to animal and fungal NLRs. The central NB domain in animals is NACHT (after NAIP, CIITA, HET-E and TP1), which is structurally similar to NB-ARC, but thought to have evolved independently (Urbach and Ausubel, 2017). In fungi, the central domain can be either NACHT as in animals, which is most frequent; NB-ARC as in plants; or a fungi-specific domain (Uehling et al., 2017). The C-terminal domain, both in plants and animals, is LRR. In fungi, however, three other types of domains can be present at that location. All of these are, however, repeat domains like LRR.

There is more variety in terms of N-terminal domains in animals and fungi than just the two options (TIR and CC) in plants. Animals have five described N-terminal NLR domains, and fungi have at least twelve, some of which show enzymatic activity directly, rather than just having a signaling function as in plants and animals (reviewed in Uehling et al., 2017).

1.3.4 NLR Genomic Distribution

Genomic arrangement of NLR genes is non-random, with many residing in clusters (Guo et al., 2011) (Figure 4). Some clusters consist of simple tandem arrays of closely related genes, usually of the same type, thus creating homogeneous TNL or CNL clusters (Figure 5) and others consist of a mixture of the two types. Thus, in *A. thaliana*, of the total of 34 NLR clusters, 7 consist of tandem duplicates and 27 are mixed (Guo et al., 2011). Not only NLR genes, but around 17% of all *A. thaliana* genes exist as tandem arrays (Arabidopsis Genome Initiative, 2000). However, NLR genes with sequence similarity but located on different chromosomes also exist (Figure 5).

Other NLR clusters contain a heterogeneous mixture of distantly related sequences (McDowell and Simon, 2006). These may be a result of gene or large-scale segmental chromosome duplications followed by local rearrangements (Baumgarten et al., 2003; Leister, 2004). Transposable element (TE)-mediated rearrangement might also play a role in generating NLR clusters: it has been observed that the size of clusters correlates positively with the number of transposable elements in the same chromosome (Li, J. et al., 2010; Ameline-Torregrosa et al., 2008).

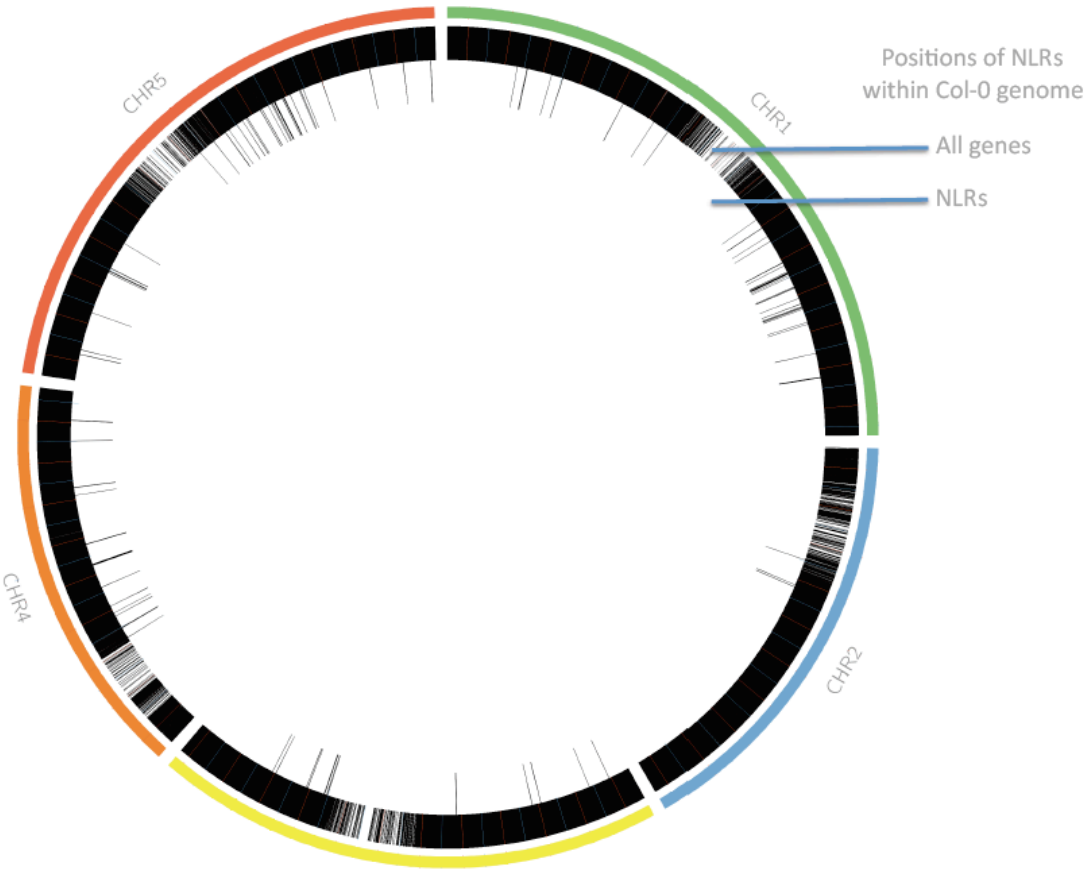

Figure 4. An overview of the distribution of NLR genes within the Col-0 genome, compared to other genes.

Clustered arrangement may be beneficial in promoting the creation of new NLR specificities by mechanisms such as non-homologous recombination and gene conversion. In ectopic (non-allelic) gene conversion, a sequence is replaced by a homologous sequence from a paralog, which can serve as a template. For both of these processes, the clustered arrangement of NLR genes can be important, as it offers more possibilities for recombination between linked loci (Meyers et al., 2003; Hulbert et al., 2001).

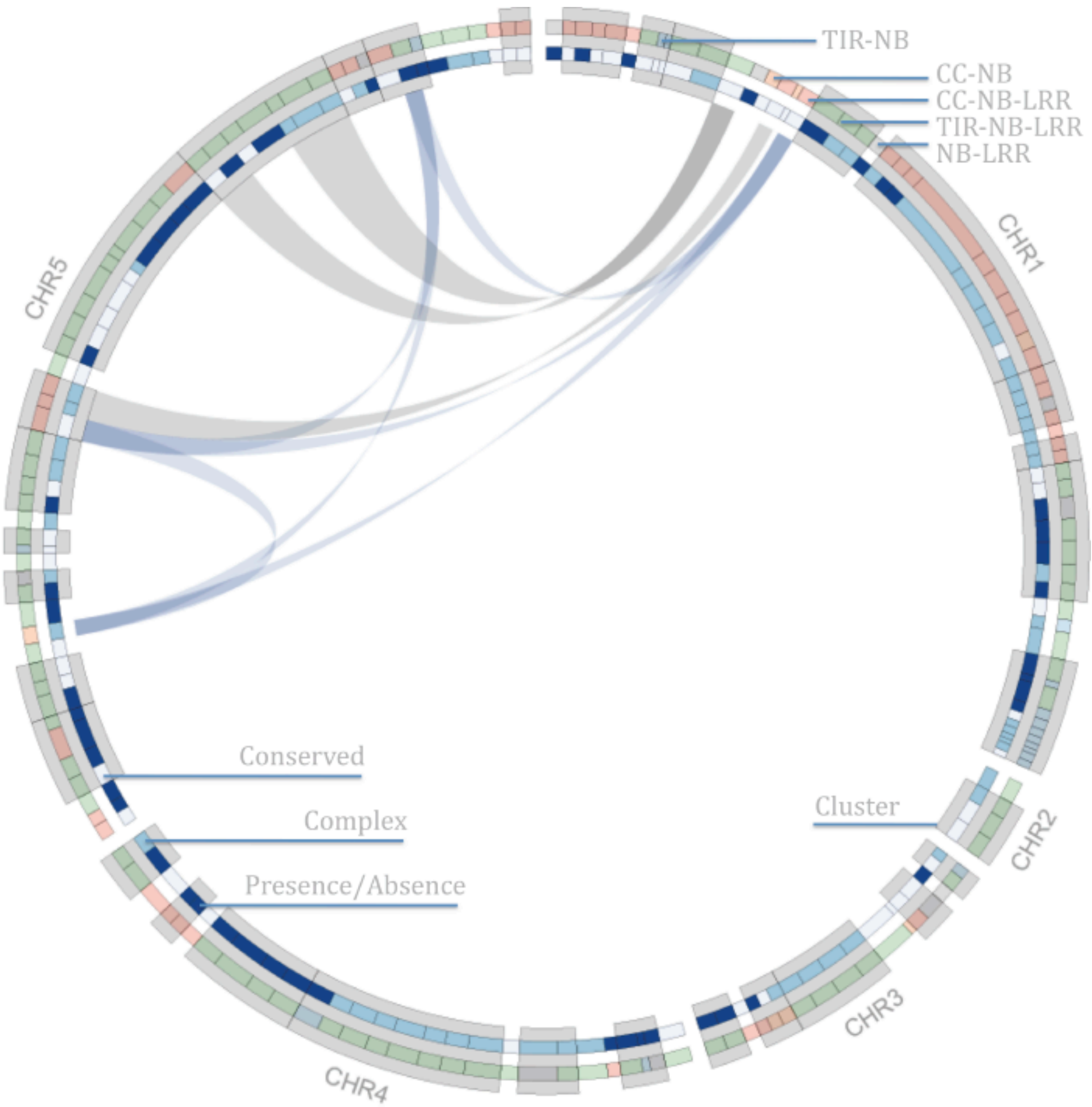

Figure 5. Concatenated NLR genes from the Col-0 genome. Grey boxes indicate clusters with TNL clusters shown in green and CNL clusters in orange (data from Guo et al., 2011). Long-range sequence similarity between sequences is shown with ribbon bands inside the circle. Inner track indicates the categorization of NLR genes described in this paper, which will be discussed in future chapters.

In some cases, genomic co-occurrence of NLR genes might be connected to their function. Protein products of several pairs of NLR genes located in head-to-head orientation in the genome are known to function together. For example, the TNL pair *RPS4* and *RRS1*, which are co-localized in the genome and whose protein products interact (Narusaka et al., 2009; Williams et al., 2014).

Not all NLR genes are clustered, however, and some exist as single copies. Accordingly, NLR genes in *A. thaliana* are a complex mixture of tandem duplicates, ohnologs retained from ancient polyploidization events, and unique copies (Hofberger et al., 2014), either arranged in complex homo- and heterogeneous clusters, or present as single genes.

1.3.5 Evolutionary Mechanisms that Generate NLR Diversity

NLR genes can exist in complex chimeric and repeptitive architectures. Characteristic NLR features that contribute to generating excessive diversity include clustered genomic arrangements and internally repeated structure (LRRs). Clustered arrangement has been correlated with gene conversion (Mondragón-Palomino and Gaut, 2005; Xu et al., 2008; Guo et al., 2011). Repeated structure within LRR domains can cause mispairing and recombination between different regions of the same gene (Hulbert et al., 2001; Kuang et al., 2004; Wicker et al., 2007). About 10% of reference NLR genes are estimated to be pseudogenes, and have been suggested to serve as reservoirs of variation for nonhomologous recombination and gene conversion (Meyers et al., 2003).

In terms of evolutionary mechanisms, excessive diversity could result either from rapid evolution and strong directional selection, or maintenance of many old alleles by balancing selection (discussed in Sections 1.4.3 and 1.4.6). Strong directional selection has been well documented for NLR genes (Mondragón-Palomino et al., 2002; Chen et al., 2010). Specifically, directional selection has been detected in certain residues of the LRR domain, and purifying selection in the NB domain (reviewed in Jacob et al., 2013).

Transposable elements can contribute to both generation and reassortment of variation. For example, transposons were associated with the deletion in the well-studied case of *RPS5* (Henk et al., 1999), which features a P/A polymorphism. Transposable element presence can also contribute to clustered arrangement (discussed in Section 1.3.4) and transposition has also been linked to the introduction of new domains into NLR genes, such as integrated decoy domains (Bailey et al., 2018).

Polyploidization followed by diversification can be another mechanism contributing to NLR diversity in plants. Arabidposis lineage is estimated to have undergone at least five polyploidization events (Jiao et al., 2011). Apart from tandem duplication and polyploidization events, segmental and transpositional duplication can generate paralogous copies at ectopic locations (Freeling, 2009) that provide raw material for diversification. Together, these processes make NLR genes one of the most diverse gene families within a species (Clark et al., 2007).

1.4 Polymorphism in NLR Genes

NLR genes exhibit complex patters of diversity. Polymorphism in NLR genes does not merely encompass allelic diversity, but also gene copy number variation (reviewed in Baggs et al., 2017). This includes the limiting case of P/A polymorphism, where the whole gene is either present or absent in multiple accessions.

1.4.1 Studies of Single Genes

Early NLR gene papers were based on a forward genetics approach, in which genes were identified by mapping loci that segregate for susceptibility and resistance alleles (Nishimura and Dangl, 2010). The respective genes were then amplified using polymerase chain reaction (PCR) and sequenced. Polymorphism analysis was then carried out and their evolutionary dynamics inferred. Different levels of polymorphism have been reported (Table 1).

Gene	Polymorphism	Clustered	Reference
RPM1	P/A	No	Grant et al. 1998; Stahl et al. 1999
RPP1	High	Yes	Botella et al. 1998
RPP5	High	Yes	Noël et al. 1999
RPP8	High	Yes	McDowell et al. 1998
RPP13	High	No	Bittner-Eddy et al. 2000; Rose et al. 2004
RPS2	High	No	Caicedo et al. 1999; Mauricio et al. 2003
RPS4	Low	Yes	Gassmann et al. 1999
RPS5	P/A	Yes	Henk et al. 1999; Tian et al. 2002

Table 1. Levels of polymorphism and genomic distribution (clustered/non clustered) of some characterized *A. thaliana* genes. For Presence/Absence (P/A) genes, type of polymorphism is indicated instead of level.

This initial subset of genes did not necessarily provide a complete or representative picture of the type of polymorphism patterns that exist in the NLR family. First, PCR amplification step in these studies would not work for loci with truly high levels of variation, thus limiting the sample to genes with moderate to low divergence. However, the fact that these loci were originally identified as segregating for resistance and susceptibility alleles had biased the sample towards genes that were, at least to a certain extent, variable, as pointed out by Bakker and colleagues (2006). The bias could have led to an overrepresentation of patterns such as P/A polymorphism in the characterized set of genes, while their true prevalence in the NLR complement remained unknown. This motivated the need for whole-genome studies that would characterize the whole NLR complement.

As a transition to that stage, larger though still partial and PCR-based studies sought to sample NLR variation more evenly by including a larger subset of 27 NLR genes (Bakker et al., 2006; 2008). These studies resequenced individual genes or domains in 96 accessions of *A. thaliana* and relied on phylogenetic trees and population genetic statistics to provide an overview of NLR variation.

1.4.2 Genome-Wide Analyses

Following the sequencing of the first *A. thaliana* genome and the subsequent advent of resequencing projects, genome-wide analyses of NLR variation started to become increasingly widespread. Such projects sought to sample variation in all genes, not just the ones that segregated for resistance and susceptibility alleles.

A key paper (Guo et al., 2011) compared genome-wide NLR complements of *A. thaliana* with its closest extant species *A. lyrata*. It found that the number of NLR genes is similar in the two species. The study compared interspecific (between-species) variability in two families of R genes, NLR and RLP. Genes of the NLR family were found to be the more variable of the two (Guo et al., 2011).

In potato and tomato, targeted enrichment for NLR genes followed by high-throughput sequencing was proposed to study diversity in NLR genes (Jupe et al., 2013; Andolfo et al., 2014). Multiple studies on NLR variation in species other than *A. thaliana* have been carried out (Table 2; also reviewed in Monteiro and Nishimura, 2018; Borrelli et al., 2018; Baggs et al., 2017), and a study compared NLR repertoires in five Brassicaceae genomes, including *A. thaliana*, *A. lyrata* and *C. rubella* (Zhang et al., 2016). However, these studies tend to consider single or very few accessions of each species (e.g. two rice accessions in Yang et al., 2006), with few exceptions (e.g. 80 *A. thaliana* accessions in Guo et al., 2011), and no definitive study characterizing the full extent of NLR polymorphism in multiple accessions of *A. thaliana* exists.

Species	Common name	Reference
Arabidopsis thaliana	thale cress	(Guo et al., 2011; Zhang et al., 2016)
Arabidopsis lyrata	-	(Guo et al., 2011; Zhang et al., 2016)
Gossypium hirsutum	cotton	(Shi et al., 2018)
Solanum tuberosum	potato	(Jupe et al., 2013)
Solanum lycopersicum	tomato	(Andolfo et al., 2014)
Oryza sativa	rice	(Yang et al., 2006)
Sorghum bicolor	sorghum	(Yang et al., 2016)
Brachypodium distachyon	stiff brome	(Tan and Wu, 2012)
Phaseolus vulgaris	bean	(Richards et al., 2018)
Setaria italica	foxtail millet	(Zhae et al., 2016)
Triticum aestivum	wheat	(Bouktila et al., 2014)
Lotus japonicus	lotus	(Li, X. et al., 2010)
Glycine max	soybean	(Kang et al., 2012)
Eucalyptus grandis	eucalyptus	(Christie et al., 2016)

Table 2. Selected published genome-wide analyses of NLR genes in plants.

1.4.3 Presence/Absence (P/A) Polymorphism

Presence/Absence polymorphism has long been known in NLR genes. It has also been observed that the presence or absence of an NLR gene at a given locus could be mapped almost perfectly to resistance and susceptibility phenotypes in a plant (Stahl et al., 1999, Tian et al., 2002).

RPM1 and *RPS5* are well known examples of P/A NLR genes (Grant et al., 1998; Henk et al., 1999). From the point of view of variation patterns, NLR P/A genes have been reported to have a relatively low level of polymorphism within the NLR gene itself when accessions carrying the gene are compared, notably also in the otherwise quickly evolving LRR domain (Ding et al., 2007; Shen et al., 2006; Mondragón-Palomino, 2002). However, the level of polymorphism is increased around the deletion junctions, which is consistent with the theory of balancing selection (Stahl et al., 1999; Tian et al., 2002).

Studies of P/A NLR genes (Shen et al., 2006) as well as P/A genes across the whole genome (Tan et al., 2012) based on 80 genomes (Cao et al., 2011) data have been carried out. In Tan and colleagues (2012), P/A genes have been identified genome-wide, allele frequency analysis carried out, and gene ontology (GO) functional category enrichment described. Studies looking at gene expression data have also suggested that the majority of P/A genes in the genome might be under relaxed selective constraints (Bush et al., 2014).

1.4.4 Conserved Genes

Perhaps the first main task of *A. thaliana* research in the post-genomics era was assigning functions to the ca. 27,000 identified genes. In the case of NLR genes, historically, the first characterized gene set was biased toward the more variable rather than conserved types, since they were first identified by mapping loci that would differentially segregate for resistance and susceptibility to specific pathogens (reviewed in Nishimura and Dangl, 2010). This resulted in a lag in the study of NLR conserved genes. Thus, while the role and fitness cost of P/A genes such as *RPM1* and *RPS5* has been well established (e.g. Tian et al., 2003; Karasov et al., 2015), little is known of conserved NLR genes as a group.

In addition, unlike for P/A genes (e.g. Tan et al., 2012), no genome-wide studies of conserved genes as a group have been attempted. Conservation, which is typical of most of the genome, is not a given in the case of NLR genes. Understanding the functional importance of conserved NLR genes would offer insights into how NLR gene repertoires are assembled. It would allow to address the question of how plants cope with carrying large numbers of NLR genes, considering that some of them have high fitness costs (Tian et al., 2003; Karasov et al., 2015), and enable the estimation of costs of plant NLR repertoires (see Section 1.4.5).

The existence of conserved NLR genes has been known for a long time. Previous studies in lettuce subdivided NLR genes into two broad classes based on their evolutionary history: quickly evolving Type I, characterized by frequent sequence exchanges and present as chimeras, and the more conserved Type II, for which obvious allelic relationships can be established (Kuang et al., 2004). Some studies have also been carried out in *A. thaliana*. A subset of 27 NLR genes has been sorted based on the extent of exhibited allelic divergence (Bakker et al., 2006). Based on between-species comparisons, Hofberger and colleagues (2014), for example, identified 4 NLR loci present across twelve sampled plant species, suggesting a basic housekeeping function. *ADR1* family in *A. thaliana*; which includes *ADR1* and its homologs *ADR1-L1*, *ADR1-L2* and *ADR1-L3*; are CNL type genes of the *RPW8* (*RESISTANCE TO POWDERY MILDEW 8*) type that act as helper NLRs, and are very highly conserved across plant species (Collier et al., 2011; Chini and Loake, 2005, Zhang et al., 2016). However, no definitive list of conserved NLR genes in *A. thaliana* exists.

Originally, it was expected that NLR genes, when present, would be similar based on within-species comparisons, but show divergence when compared between species. This was based on the “arms race” theory (Flor, 1956; Holub, 2001; reviewed in Bergelson et al., 2001). The term “arms race” originally comes from military terminology and denotes a dynamic where people and groups are forced to arm themselves because others are doing so, without any intrinsic benefit of the process. A classical “arms race” dynamic between plants and pathogens would thus entail a series of “selective sweeps” in which newly discovered variants that offer advantage would propagate thorough populations and sweep to fixation. Thus, when comparing individual accessions, they would be expected to carry the same allele of an NLR gene that most recently swept to fixation. However, this was not found to be the case in practice as polymorphism in NLR genes is common. *RPS4* is a rare example of an NLR gene thought to have undergone a recent selective sweep (Bergelson et al., 2001; Bakker et al., 2006).

1.4.5 NLR Fitness Costs and NLR Complement Size

There are about 150 NLR genes in *A. thaliana* (Meyers et al., 2003) and plant NLR repertoires are usually in the hundreds (Uehling et al., 2017). Relating to this, two questions have been asked. First, how can such relatively small repertoires produce specific defence against a myriad of pathogens present in the environment. Second, how can plants afford to maintain such relatively extensive repertoires, considering that to maintain multiple copies of immunes gene for a plant can be costly: two P/A NLR genes, *RPM1* and *RPS5*, are known to carry fitness costs of up to around 10% (Tian et al., 2003, Karasov et al., 2015).

How limited NLR complements provide specific defence against a wide variety of pathogens, and an even greater number of effectors, may be explained in part by the guard hypothesis. Rather than monitoring the presence of pathogen effectors directly, certain NLR receptors monitor (guard) the integrity of their own host proteins that would be compromised as a result of effector presence (Chisholm et al., 2006; Jones and Dangl, 2006). As a result, multiple effectors that compromise the same host protein could be monitored by the same NLRs. In addition, it has been shown that a single NLR can bind more than one effector (Cesari et al., 2013), thus contributing to the explanation of how limited NLR complements can monitor significantly larger numbers of effectors. Studies using yeast two-hybrid (Y2H) technology to probe protein interactions have found that a large number of pathogen effectors converged onto a common set of plant proteins (Mukhtar et al., 2011; Weßling et al., 2014).

The high fitness cost of NLR genes might be mitigated, from the evolutionary perspective, by balancing selection. In the case of frequency-dependent balancing selection, the frequency of a gene is regulated within a population based on selective need. Presence/Absence NLR genes with high fitness costs, *RPM1* and *RPS5*, are known to be under balancing selection (Stahl et al., 1999; Tian et al., 2002), as are several other NLR genes.

In the case of the birth-and-death evolutionary model, extra copies of NLR genes may be removed through deletion or pseudogenization (Michelmore and Meyers, 1998). Alternatively, the high costs of NLR genes might also be mitigated by microRNA (miRNA)-based control, whereby NLR transcripts are degraded in bulk based on the targeting of conserved sequences in the NB-ARC domain (Fei et al., 2013; Li et al., 2012; Shivaprasad et al., 2012).

1.4.6 Evolutionary Dynamics of NLR genes

The highly polymorphic nature of NLR genes is probably a result of selective pressure imposed by pathogens (Meyers et al., 2003; Yue et al., 2012). The originally dominant hypothesis on the evolution of NLR genes was the coevolutionary arms race model (Flor, 1956). According to this hypothesis, the plant R genes and pathogen effectors interact in a “gene-for-gene” manner, in which the protein produced by a specific NLR gene in a plant recognizes an effector protein produced by a specific pathogen. Such recognition results in a hypersensitive response and cell death in the plant.

It would thus be beneficial for the pathogen to retain novel variants that avoid detection by the plant, and it would offer competitive advantage for the plant to discover NLR alleles that would enable detection of the new variants. Such alleles in the plant would then quickly increase in frequency to replace all the other alleles in a population (selective sweep). Based on population genetic theory, such a model would predict intraspecific polymorhism to be rare and transient in nature.

While it seems to be the case that the LRR regions of many NLR genes are under positive selection (Meyers et al., 1998; Bittner-Eddy et al., 2000; Mondragón-Palomino et al., 2002), as predicted by the coevolutionary arms race model, the study by Bergelson and colleagues (2001) highlighted the prevalence of long-lived polymophism in NLR genes in *A. thaliana* populations. This observation has been partially explained by balancing selection in NLR genes, where two or more allelic forms of a gene are maintained in a population over long periods of time - in stark contrast to selective sweep model suggested previously.

From an evolutionary perspective, NLR genes showing variable patterns might be under directional selection for local adaptation, or might be under the influence of genetic drift or balancing selection. Some studies have indeed addressed the question of the relative prevalence of these types of selection (Bakker et al., 2006).

NLR genes are subject to a complex mixture of positive and balancing selection. There also seems to be an interplay between different types of selection, as P/A NLR genes under balancing selection tend to lack strong signatures of positive selection in LRR regions, which are otherwise prevalent in NLR genes (Shen et al., 2006).

NLR genes are not uniform in their evolutionary mode, probably as a result of varying selective pressures that act on them. Previous studies in lettuce subdivided NLR genes into two broad classes based on their evolutionary history: quickly evolving Type I, characterized by frequent sequence exchanges and present as chimeras, and the more conserved Type II, for which obvious allelic relationships can be established (Kuang et al., 2004).

1.5 Summary

NLR genes have been categorized by variation pattern, evolutionary mode (fast/slow), evolutionary origin (ohnologs/paralogs), genomic distribution (single/clustered) and domain architecture (TNL/CNL). However, which functional roles correspond to these categories is not known.

Variation is distributed unequally among NLR genes. Previous studies (Bakker et al., 2006; Kuang et al., 2004) attempted to classify NLR genes according to their variation profiles. However, no systematic study to characterize variation across all NLR sequences has been undertaken; nor has it been studied in depth how sequence variation in NLR genes differs from background genes.

My objective is to look at diversity patterns in NLR genes in detail. I provide classification, visualization and enrichment in functional categories for NLR genes, as well as a GWAS analysis on P/A polymorphism. My method is useful for studies that wish to target a particular subgroup of NLR genes or to assess allelic diversity in an identified set of genes.

2 Chapter 1. Technology

In this chapter, I describe an approach for profiling complex hypervariable regions of the genome, exploiting short read high-throughput sequencing data. This approach works even when whole genome *de novo* assembly is not feasible and when only one reliable reference is available. I begin by describing the basic workflow briefly. I then describe in detail data preparation, choice of parameters for mapping and visualization. Subsequently, I address the issues of validation and interpretation. Finally, I carry out a statistical analysis of the obtained data.

2.1 Approach

To investigate natural variation in the complete complement of NLR genes in the *A. thaliana* reference genome, I used the annotation of Guo and colleagues (2011). The main data set consisted of short Illumina reads from 80 accessions from the first phase of the 1001 Genomes project, chosen to represent both local and global genomic diversity in *A. thaliana* (Cao et al., 2011).

The strategy of aligning short Illumina reads from a resequenced genome to a gene of interest provides information on which segments of the reference gene are present in the interrogated individual. I visualize this information, which I call presence-of-coverage profiles, and use it to compare accession-specific coverage along the length of the gene of interest. What follows is a brief overview of the workflow (Figure 6). Parameter and procedure choices for each step are discussed in detail in Section 2.3.

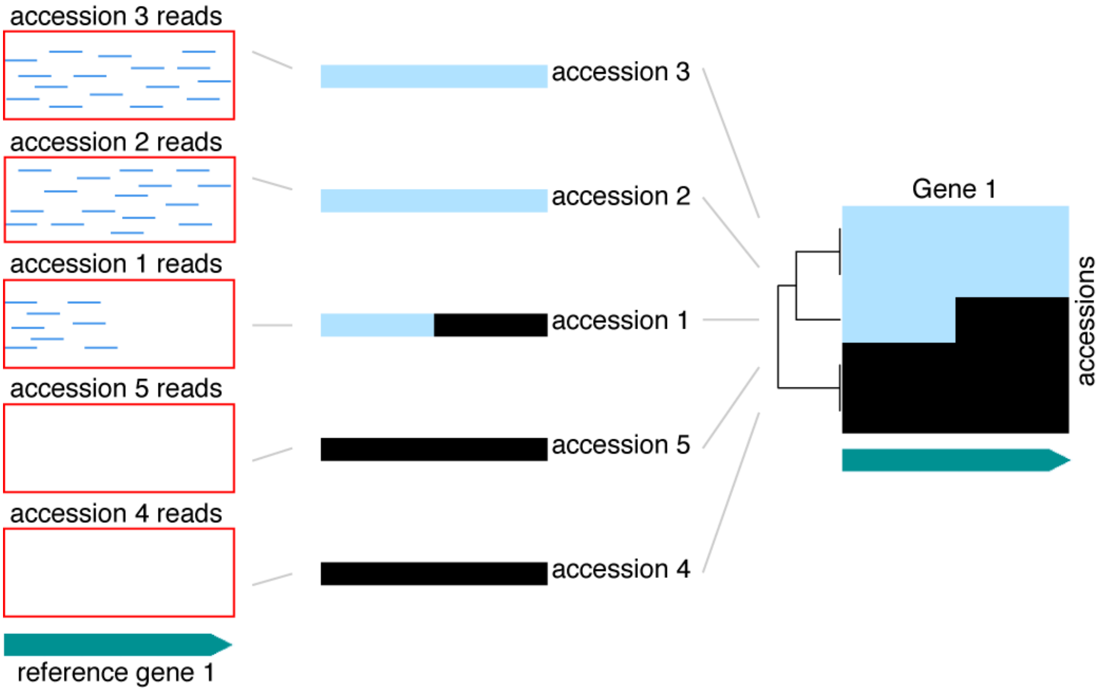

Figure 6. Schematic representation of how coverage profiles are obtained from mapped reads data.

To apply this approach, it is necessary to have a reference sequence of interest and multiple resequenced genomes as short reads or in any other format. To prepare data for mapping (Step 1), the short reads were cut into uniform segments (*k* mers), of 36 base pairs each, to ensure conisistency when comparing accessions and datasets. I used non-overlapping *k* mers to maximize mapping efficiency. Subsequently, these *k* mer reads were aligned (mapped) to each of the reference NLR genes separately.

To prepare reference NLR sequences for mapping, these were extracted from the TAIR10 assembly of the *A. thaliana* Col-0 reference genome (https://arabidopsis.org), while retaining small segments of genomic context on either side of the genes to allow the mapping of reads at the edges of the genes. These additional sequences were removed after the mapping was accomplished. Likewise, I had originally kept intronic sequences of the genes for the mapping stage. Once the mapping was complete, I kept for further processing only the positions that correspond to the coding (CDS) portions of the gene.

The list of 163 reference NLR genes was based on the Supplemental Table I in Guo and collegues (2011), with 4 additional genes (AT1G63860, AT1G72920, AT1G72930, AT5G45230) added by manual curation (courtesy of Eunyoung Chae).

In the second step, the mapping was carried out using BWA (Li and Durbin, 2009) software. I chose conservative mapping parameters with an edit distance of one, thus allowing one mismatch for each 36 base pair segment (see Section 2.3.3).

In the third step, presence of coverage for each accession along the length of each gene was recorded and visualized. If one or more reads mapped to a position in the reference gene, it was assigned “Presence” status (1); otherwise it was given “Absence” status (0). In simplified terms, presence-of-coverage status indicates whether sequences similar to the reference exist or not in each of the accessions or genomes of interest, in a position-specific manner. These vectors can then be visualized by color-coding the presence of coverage as blue and the absence of coverage as black, for example. Thus, for a single gene, each resequenced genome generates a distinct color-coded profile.

Also in the third step, the coding sequences of the gene (CDS) were extracted and used for all further processing. For this I used the overlap or union of all the CDS models of the gene based on the TAIR10 annotation (https://arabidopsis.org).

In the final fourth step, the previously obtained color-coded coverage profiles were clustered, grouping accessions with similar patterns, and displayed as a heatmap. The coverage profiles were thus aligned by position in the reference gene, and the sequential order of positions in the gene was preserved as 5’ to 3’. Thus, I obtained a position-specific coverage profile for each of the NLR genes.

2.2 Visualization of Variation at Gene Level (Blue-Black Plots)

In Figure 7, I present examples of coverage profiles for three representative classes of NLR genes, based on their variation across accessions. With my approach, one can easily discern that the gene in Figure 7A is highly conserved in all accessions and that the gene in Figure 7B is conserved in some accessions, but absent in the others (a phenomenon known as presence/absence [P/A] polymorphism), and, finally, that the genes in Figure 7C and Figure 7D have more complex patterns or variation.

Genes with complex patterns of variation that do not fall into either the conserved or the P/A group can be further classified. For instance, a gene with a complex overall pattern may (Figure 7C) or may not (Figure 7D) be missing in individual accessions. In addition, alleles may fall into a smaller number of distinct groups (Figure 7C), or there may be a very large number of such groups, with almost every accession appearing to have a distinct pattern of variation (Figure 7D).

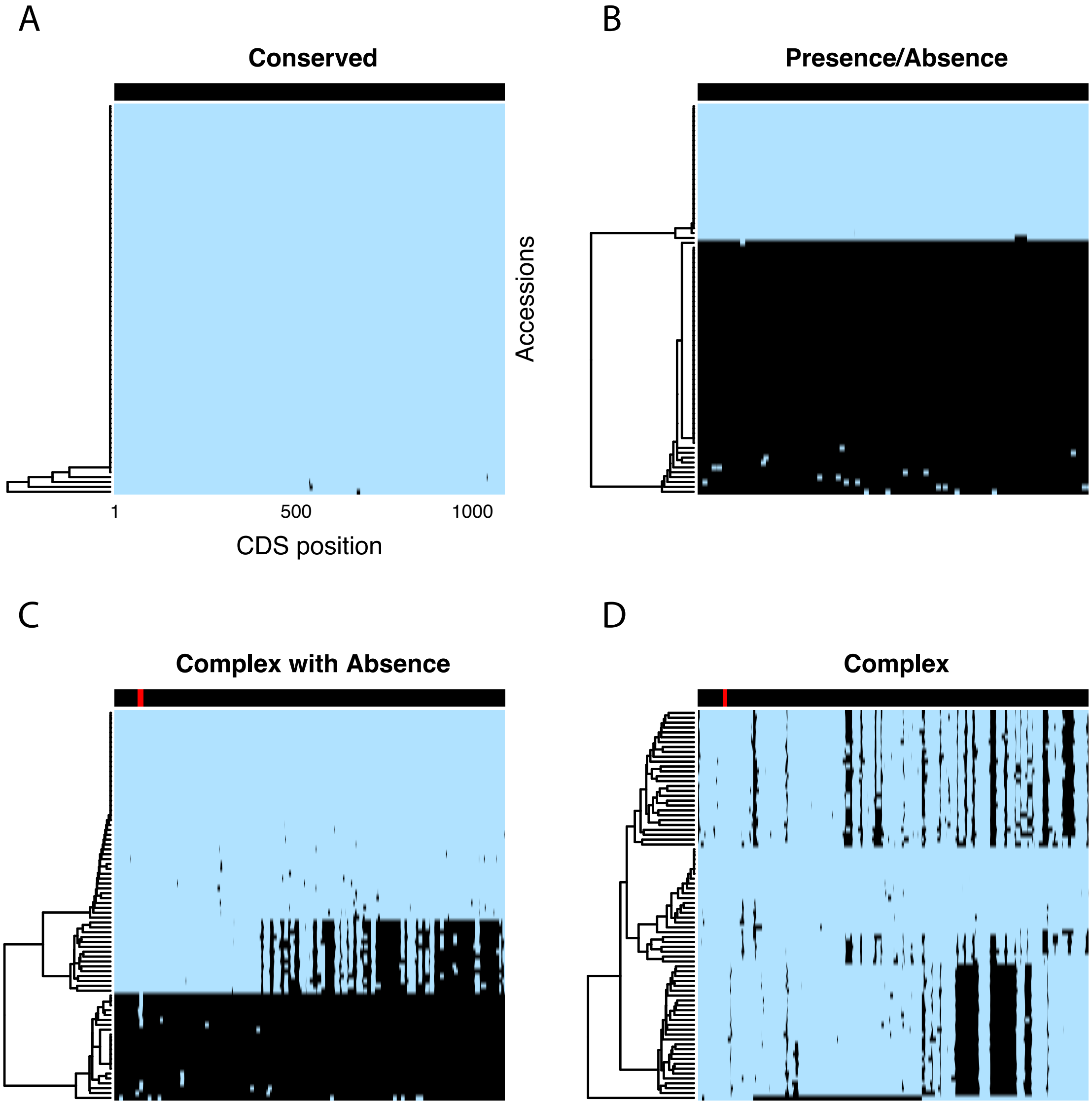

Figure 7. Examples of coverage profiles for four genes representative of the broad classes discussed in text. Presence of short read coverage along concatenated genomic CDS positions is indicated in blue, absence of coverage – in black. The 80 represented accessions are clustered by their coverage profiles. Top panel indicates repetitiveness control. Red bands indicate positions where repetitiveness within the reference genome might make mapping from non-orthologous sequences possible. The genes shown are AT1G17615, AT4G27220, AT5G51630 and AT4G11170.

2.3 Choice of Parameters and Validation

2.3.1 Choosing Read Length

Using reads (or *k* mers) as short as 36 bases offers the advantage of higher resolution coverage profiles compared to longer reads. This can improve the ability to resolve small deletions and hypervariable regions. In addition, this allows for the use of datasets that may contain very short reads.

The initial high-throughput sequencing technologies produced reads as short as 36 bp. Such reads can still be part of larger datasets and can be exploited with this approach. The obvious disadvantage in using shorter reads is ambiguity in mapping, where reads coming from non-strictly-orthologous sequences may map to the gene of interest (cross-mapping). It has been observed that read lengths starting at 36 bp can make little difference to the ability to distinguish sequences between genes (Chhangawala et al., 2015).

To investigate the effect that read length can have on cross-mapping in the case of NLR genes in *A. thaliana*, I calculated expected cross-mapping based on tiled *k* mer segments from the assembled Col-0 genome, mapped back to the NLR reference (Figure 8). It can be seen that while having longer reads offers an advantage, the increase affects only a fraction of the total NLR CDS length. Recent tandem duplicates, in particular, may be difficult to distinguish with increasing gene length.

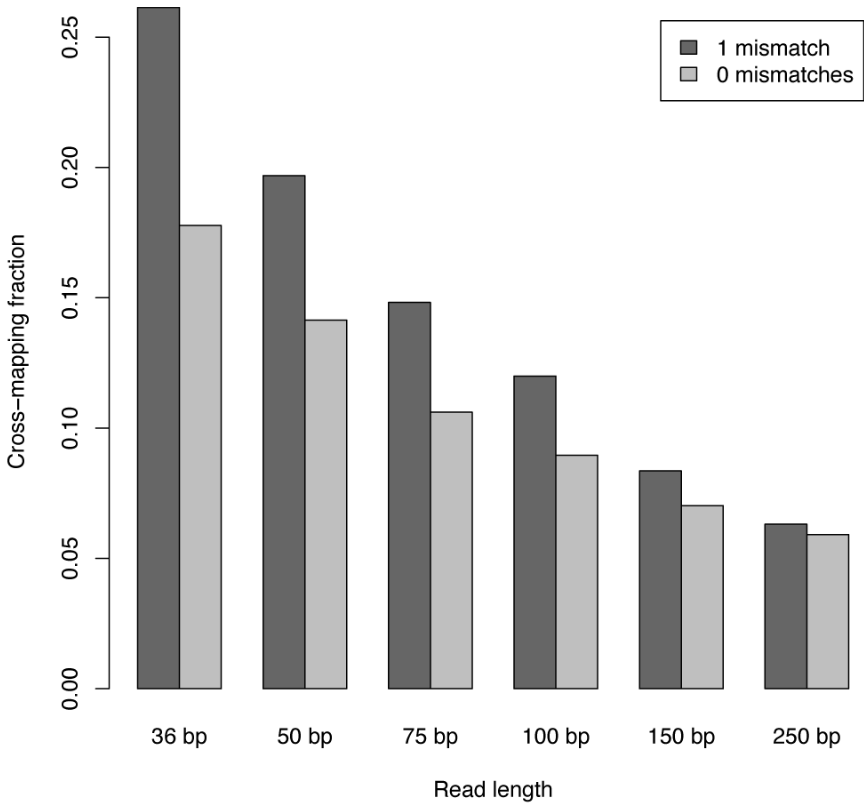

Figure 8. Fraction of total NLR coding sequence (CDS) length in Col-0 accession with mapping outside of the original position for different length of simulated reads, obtained based on tiled Col-0 data. Results for the edit distances of one and of zero are shown.

Thus, in choosing read lengths, there is a tradeoff between more detailed coverage profiles and ability to distinguish orthologous sequences. Therefore, *k* mer choice will be a function of available data and desired resolution. The issue of cross-mapping controls and the approaches to deal with it are further discussed in Section 2.3.2.

2.3.2 Cross-Mapping

In addition to divergence, NLR genes are also characterized by high levels of similarity: within an eudicot genome, an average of 50% of NLR genes correspond to tandem duplicates and a further 22% are copies retained from poliploidization events (Hofberger et al., 2014). In *A. thaliana*, an estimated 7 out of 38 NLR clusters consist of tandem duplicates (Guo et al., 2011) and sequence identities above 60% are not uncommon among NLR genes (Meyers et al., 2003). Such high levels of repetitiveness can lead to reads mapping outside of their genomic locations of origin. Thus, I note that coverage may originate from sequences that are not strictly orthologous.

To assess coverage potentially contributed by sequences outside the gene of interest, I simulated 36 bp reads from the assembled reference genome (TAIR10), masked reads that came from the alignment target itself, and mapped the remaining reads back to the reference. The top panel track in Figure 7 shows the extent to which this type of out-of-place mapping would be expected to occur in Col-0, as proportion of gene length, to provide a control for this effect. The amount of ectopic mapping also depends on the number of mismatches allowed when mapping reads to the reference. I discuss the effect of varying the number of mismatches in Section 2.3.3.

Three approaches to deal with the issue of cross-mapping can be used. One would be to provide a control for positions susceptible to cross-mapping by marking such positions in the plot, as described above. The second solution would be to mask out such positions, as is commonly done in gene expression studies through short read sequencing. However, in this particular application this would result in an unnecessary loss of information and discriminative power, compared to the other two solutions. The third approach would be to understand the interpretation of the plot to include non-orthologous sequences. This makes biological sense as NLR clusters are highly complex in nature, and orthologous relationships cannot always be determined (Chae et al., 2014; Kuang et al., 2004).

Understood in this way, the profiles would reflect sequence conservation on the whole-genome level, and local deletions or divergence might be obscured where the sequence is lost locally relative to the reference, but is still present at one of the other locations. To identify sequence similarity in the non-contiguous sequences would thus be of advantage. See Section 5.11 for more details.

2.3.3 Edit Distance

I chose conservative mapping parameters allowing zero gaps and one mismatch. My aim was to make mapping criteria as stringent as possible so as to avoid cross-mapping from non-orthologous sequences, but at the same time to be flexible enough to avoid losing the bulk of the data due to sequencing errors in the reads which would prevent them from mapping with too stringent criteria. In the data similar to the one I used, sequencing errors varied, with an estimate of about 67% of reads being error-free, with the rest containing one or more errors (Ossowski et al., 2008). To be able to exploit information contained in those reads, I thus chose to allow up to one mismatch, although using a higher number of mismatches increases cross-mapping to a certain extent (Figure 8). Thus, allowing for one mismatch as here, 26% of NLR CDS was covered by possibly cross-mapping reads, whereas decreasing the number of allowed mismatches to zero, decreases the estimated cross-mapping to 18% of NLR CDS.

Higher mismatch numbers can be leveraged for mapping more divergent genomic regions. For example, iteratively increasing the allowed numbers of mismatches for reads that do not map at lower values has also been proposed (Ossowski et al., 2008). However, I used a constant number of mismatches for consistency and ease of interpretation.

2.3.4 Effect of Coverage

Coverage by short Illumina reads is non-uniform, and gaps in coverage may arise by chance. To prevent this from significantly affecting the coverage profiles, I recommend using datasets with coverage values above 7x. When coverage is low, using overlapping segments when dividing reads into uniform sizes can maximize the use of existing coverage. The effects of low coverage and how to deal with them are further discussed in Section 3.4.1.

2.4 Comparison with Known Examples

Information contained in the coverage profiles compares well with what is already known from genes that have been extensively characterized. For instance, previous studies have shown that *RPM1* alleles, when present, are very similar, although a substantial number of accessions lack the gene (Ding et al., 2007; Stahl et al., 1999). This is consistent with the picture one can infer from Figure 9B. Likewise, *RPS2* alleles are known to fall into two distinct clades, and deletion of the entire gene seems to be rare (Mauricio et al., 2003; Caicedo et al., 1999). This is also reflected in the pattern seen in Figure 9A. More examples are available in Section 3.1.

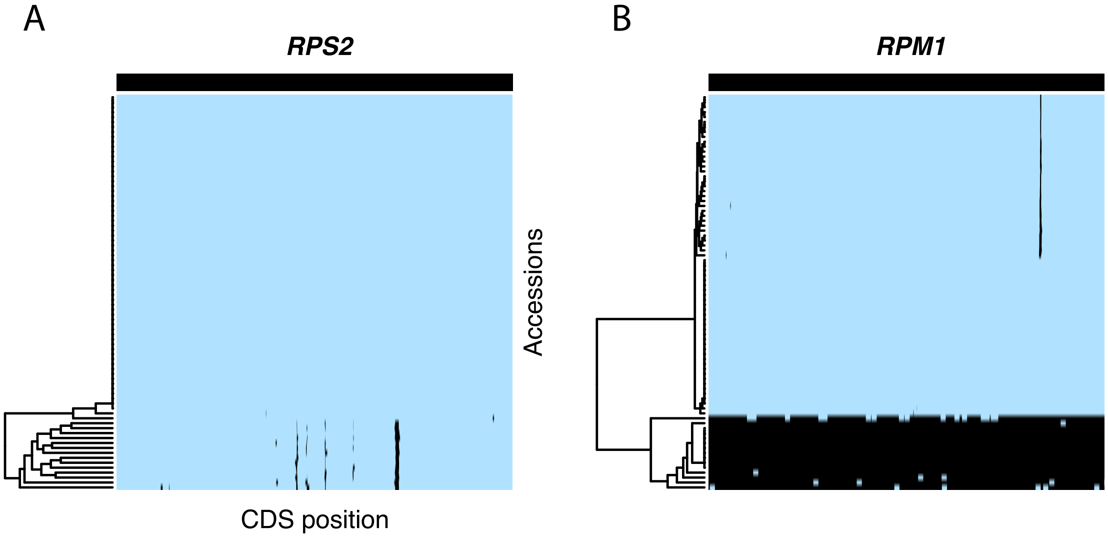

Figure 9. Examples of coverage profiles for two well-studied genes, *RPS2* (AT4G26090) and *RPM1* (AT3G07040). Presence of short read coverage along genomic CDS position is indicated in blue, absence of coverage – in black. The 80 represented accessions are clustered by their coverage profiles. Top panel indicates repetitiveness control. Red bands would indicate positions where repetitiveness within the reference genome might make mapping from non-orthologous sequences possible (none visible).

2.5 Visualization of NLR Variation at Genome Level (Orange Plots)

To be able to visualize an overview of all NLR genes and accessions in one plot, I represented each gene and accession combination by the overall degree of conservation, that is, the fraction of total gene length covered by reads, regardless of the positions of variants. An overview schematic of how the plots were generated and their relation to the previously discussed position-specific coverage profiles is shown in Figure 10. In this representation, the absence of coverage is shown as red, and the presence of coverage is shown as yellow.

This representation provides a comparative overview for how absence is distributed among accessions in different genes. For example, a gene which is absent in half the accessions and present in the other half (a P/A gene), will be represented as a mix of yellow and red elements. A gene of which a half is missing in every accession, in contrast, will be represented by uniformly orange elements.

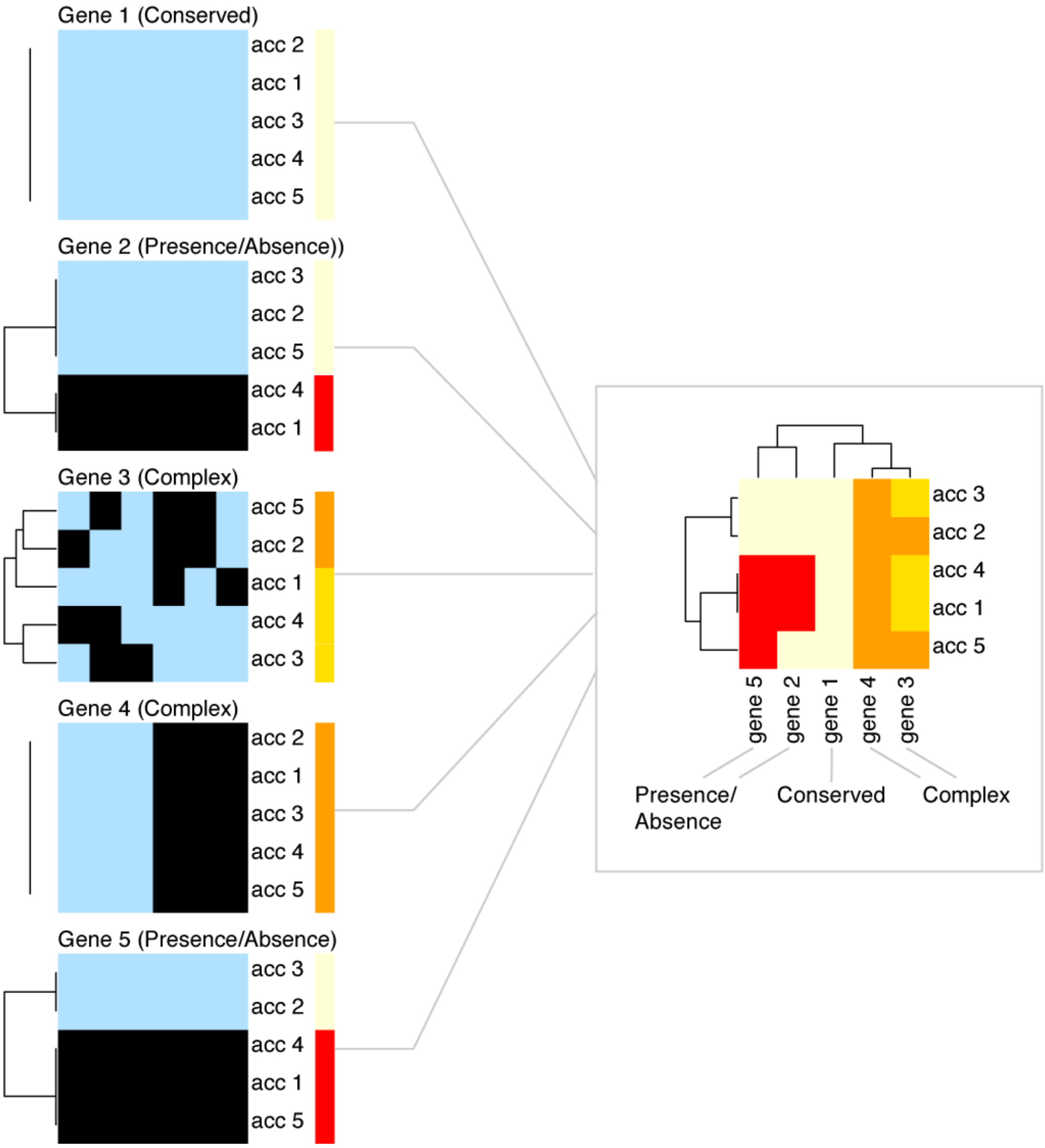

Figure 10. Schematic of overview conservation plot for multiple accessions and genes and how it is generated based on coverage profiles of individual genes.

Figure 11 represents how variation is distributed in the 163 considered NLR genes. From the figure, it is visible that conservation as defined above is not randomly distributed, but rather falls into distinct, gene-specific patterns. Thus, variation among genes, or columns, is much more pronounced than variation among accessions or rows. Visually, three categories can be discerned – P/A genes on the left, where the coverage is either present over most of the gene or absent almost entirely; conserved genes on the right, where coverage is present over most of the gene; and complex genes in the middle, which appear as various patterns of orange to indicate partial gene coverage (further analysis in Section 3.1.1).

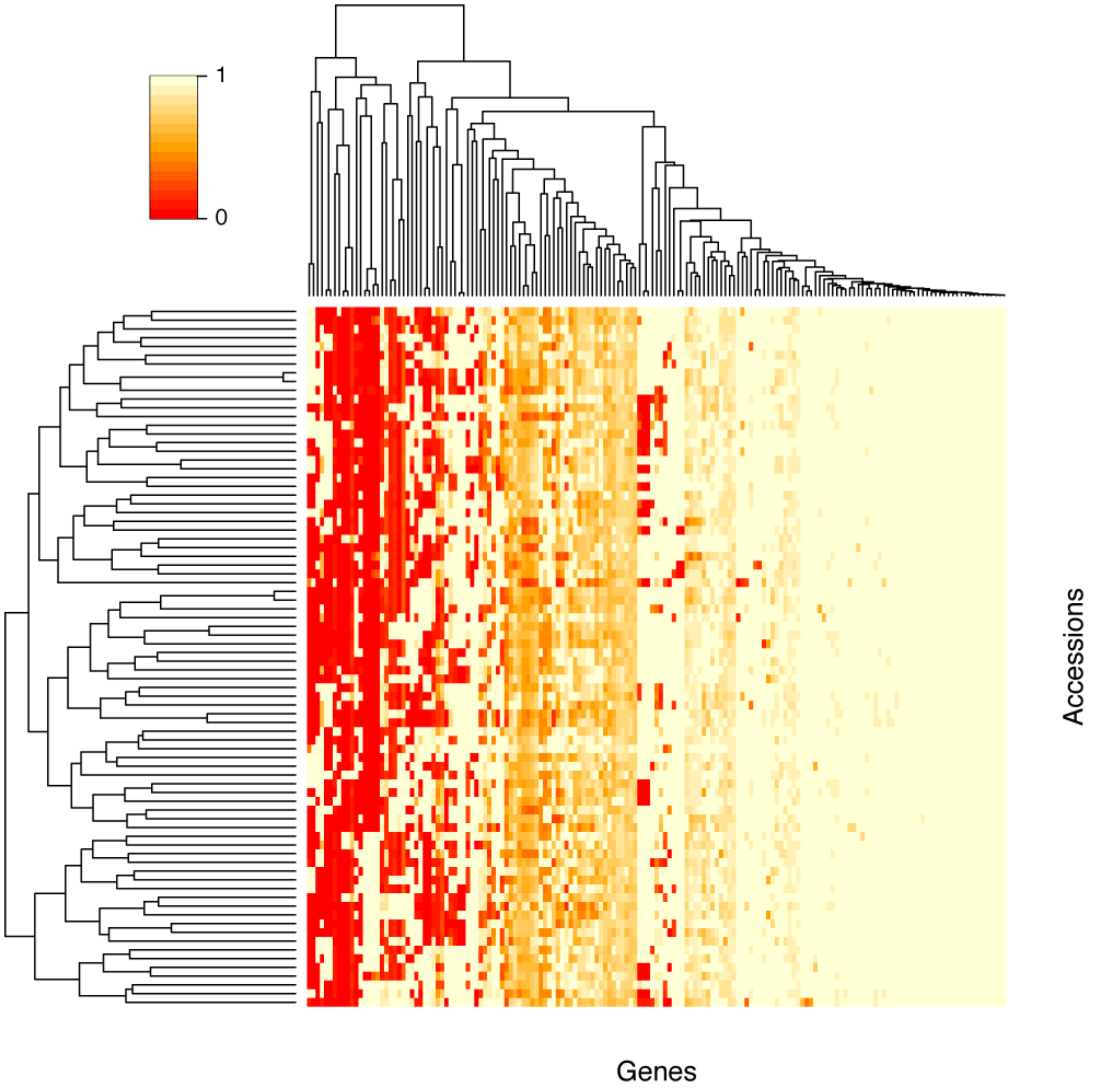

Figure 11. Overview of all gene (NLR) and accession conservation data represented by the fraction of reference CDS length covered by reads from each accession.

2.6 Population Structure

Accessions used in this study were selected from eight geographic regions (Cao et al., 2011). The clustering of accessions based on the pattern of polymorphism from Figure 11 does not show a strong population structure characteristic of the clustering based on whole-genome SNPs (Figure 12).

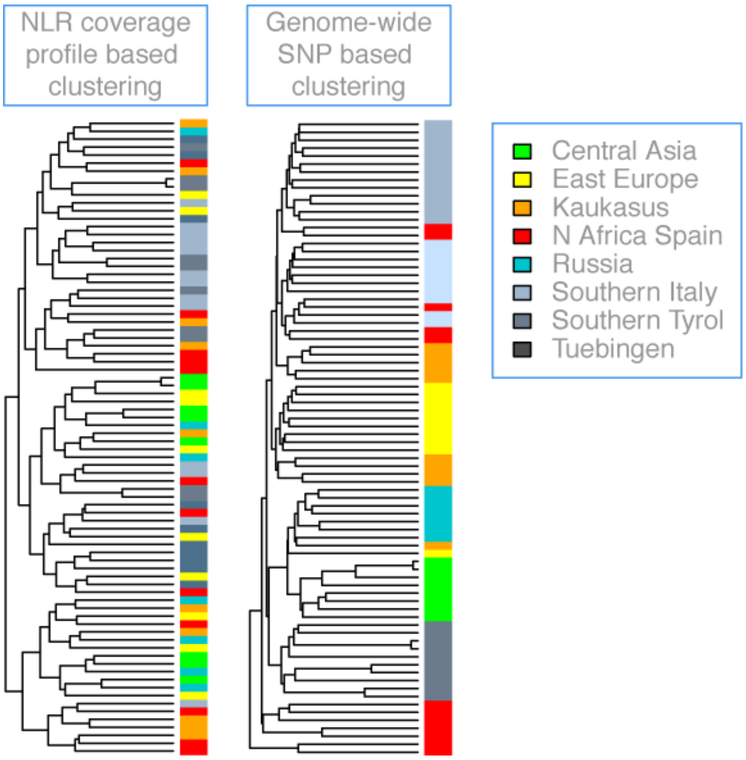

Figure 12. Clustering of accession by geographic region based on NLR coverage profiles (Figure 11), and based on 9045 randomly selected SNP from the whole genome (Cao et al., 2011).

2.7 Summary Statistics

Table 1S summarizes conservation data for NLR genes. Results for the ten highest-scoring genes are shown below (Table 3). Interestingly, some of the most conserved genes are co-located in the genome, as is the example for

AT1G17600 (*SOC3*), AT1G17610 (*CHS1*) and AT1G17615 (*TN2*). *CHS1* has been linked to cold stress response. *SOC3* (*SUPRESSOR OF CHS1-2 3*) is listed on TAIR (https://arabidopsis.org) as a known modifier of *CHS2* mutant phenotype. Thus, it is plausible that these genes might be functionally related and form part of the same network. Among these highest-scoring candidates, both TNL and CNL domain architectures are represented, as are partial architectures like CN and TN.

Gene ID	Other Names	Domain Architecture	Average	Minimum	Maximum
AT3G15700	-	CN	1.00	1.00	1.00
AT1G17615	*TN2*	TN	1.00	0.99	1.00
AT1G17600	*SOC3*	TNL	1.00	0.99	1.00
AT5G22690	-	TNL	1.00	0.99	1.00
AT1G12290	-	CNL	1.00	0.99	1.00
AT1G17610	*CHS1*	TN	1.00	0.95	1.00
AT5G04720	*ADR1-L2*	CNL	1.00	0.98	1.00
AT1G12280	*SUMM2*	CNL	1.00	0.98	1.00
AT3G04220	-	TNL	1.00	0.99	1.00
AT5G18370	*DSC2*	TNL	1.00	0.99	1.00

Table 3. Ten NLR genes with the highest conservation scores, based on within-species comparison in *A. thaliana*. Conservation values were calculated as described in text based on the number of non-absent calls as fraction of the total CDS length of the gene. The last three columns represent average over 80 accessions, based on which the table is sorted, as well as range, or minimum and maximum values, over 80 accessions. Domain Architecture column is based on the data in Guo and colleagues (2011).

3 Chapter 2. Classification within *Arabidopsis thaliana*

A number of previous studies have sought to categorize NLR genes based on the patterns of natural variation that they exhibit. It has been long known that polymorphism in NLR genes can follow a P/A pattern, where whole gene is either present or absent (Grant et al., 1998; Henk et al., 1999; Shen et al., 2006). However, other NLR genes have been known to be highly conserved, and present across multiple land plant species (Hofberger et al., 2014). In addition, studies in lettuce have shown that NLR genes in plants can be classified into the quickly evolving Type I and the more conserved Type II (Kuang et al., 2004).

These studies were, however, limited in scope and do not provide complete classification or *A. thaliana* NLR gene set. This work expands upon these studies to provide a classification of the complete set of NLR genes in *A. thaliana* into three categories: Presence/Absence (P/A), Conserved and Complex. After describing my classification approach, I visualize and compare diversity, discuss validation and provide statistics on the resulting sets of genes.

I then carry out gene enrichment analysis in gene categories for domain architecture and genomic distribution, and subsequently extend my categorization to whole-genome data set and carry out Gene Ontology analyses on the resulting lists of genes. I then carry out a Genome-Wide Association Study (GWAS) on the P/A polymorphism in the genome-wide set of *A. thaliana* P/A genes. To identify the set of genes for GWAS, I extend P/A classification to the whole genome. Finally, I filter the resulting gene lists for reliable candidates based on the surrounding genomic regions and nucleotide diversity patterns.

3.1 Classification of 163 NLR genes

3.1.1 Approach

As described in Section 2.5 with reference to Figure 11, there are three groups of NLR genes clearly visible based on their conservation profiles. To assign NLR genes into the three categories, I used a threshold-based approach. As the first step, I used k means to identify thresholds for category assignments. To do this, I applied k means to the distribution of coverage values for NLR genes in all 80 considered accessions – in other words, to all the color-coded values shown in Figure 11. Starting with three clusters, thresholds at 0.37 and 0.81 were obtained (Figure 13).

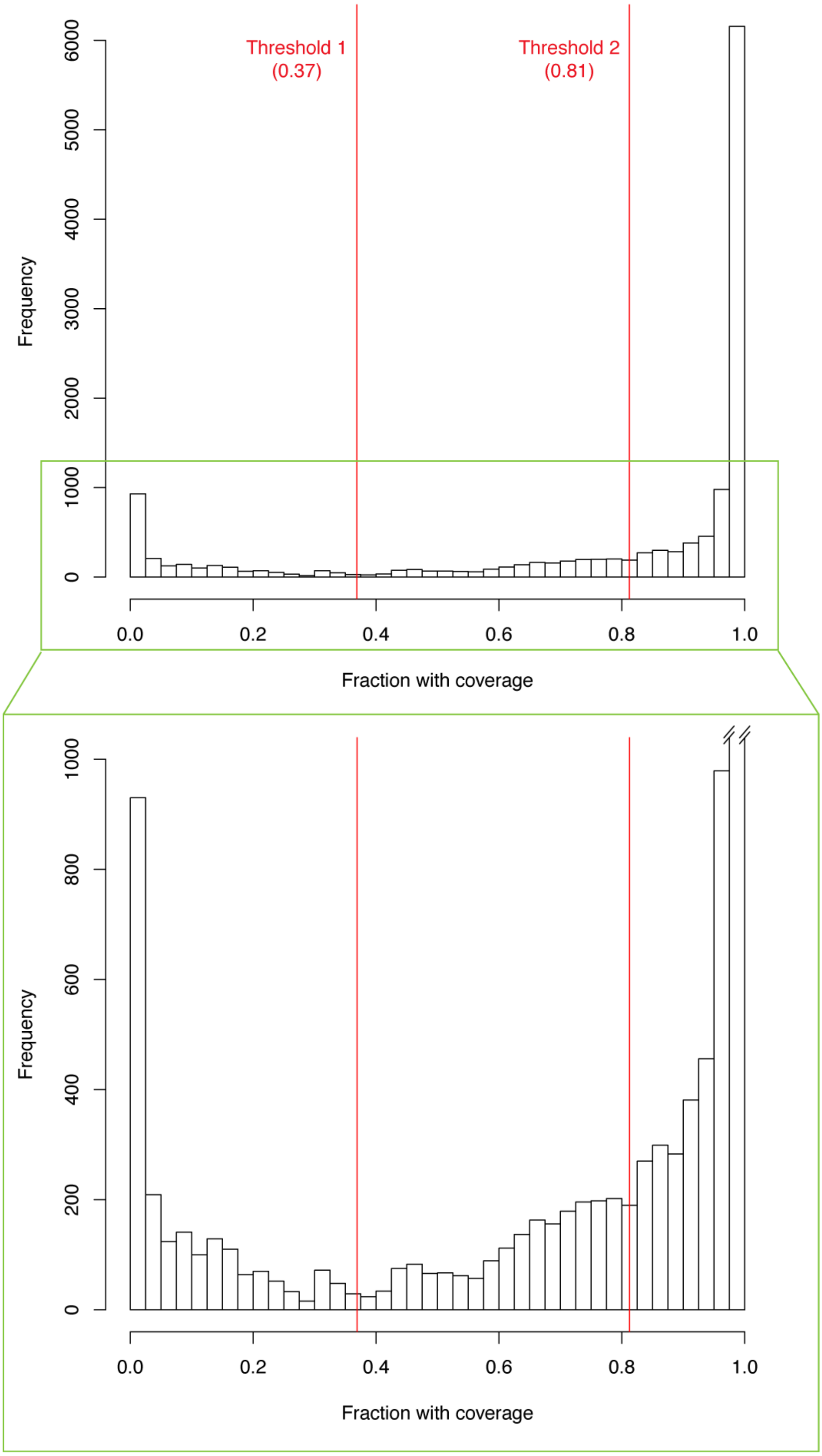

Figure 13. Thresholds for conservation value generated using *k* means.

Threshold values were used to assign each accession for each gene into one of three groups: absent if its corresponding conservation value was less than the first threshold, intermediate or complex if the conservation value was between the two thresholds, and present if the conservation value was greater than the second threshold.

Based on these assignments, each gene was represented by a set of labels, one for each accession, which were then summarized to classify each gene as either conserved, complex or P/A. If a gene contained more than 5% intermediate values, it was classified as Complex; if it contained any absent values, it was classified as P/A; otherwise it was classified as Conserved. Values labeled as intermediate and present at less than 5% frequency in a gene were disregarded to account for the possibility of random variation in coverage or noise. The value of 5% was chosen since it is a commonly used minor allele frequency (MAF) threshold. For absence, no MAF threshold was used, as it is less likely to obtain absence through noise in the data for an accession where the gene is present, and for consistency with existing definitions of P/A genes (Tan et al., 2013). Introducing a 5% MAF for absence would result in 9 P/A genes (5.52% of total NLR set) being assigned to the Conserved category.

The resulting groups were of similar size, with 52 NLR genes classified as P/A (32%), 53 NLR genes classified as Conserved (32%), and 58 NLR genes classified as Complex (36%) (Figure 14).

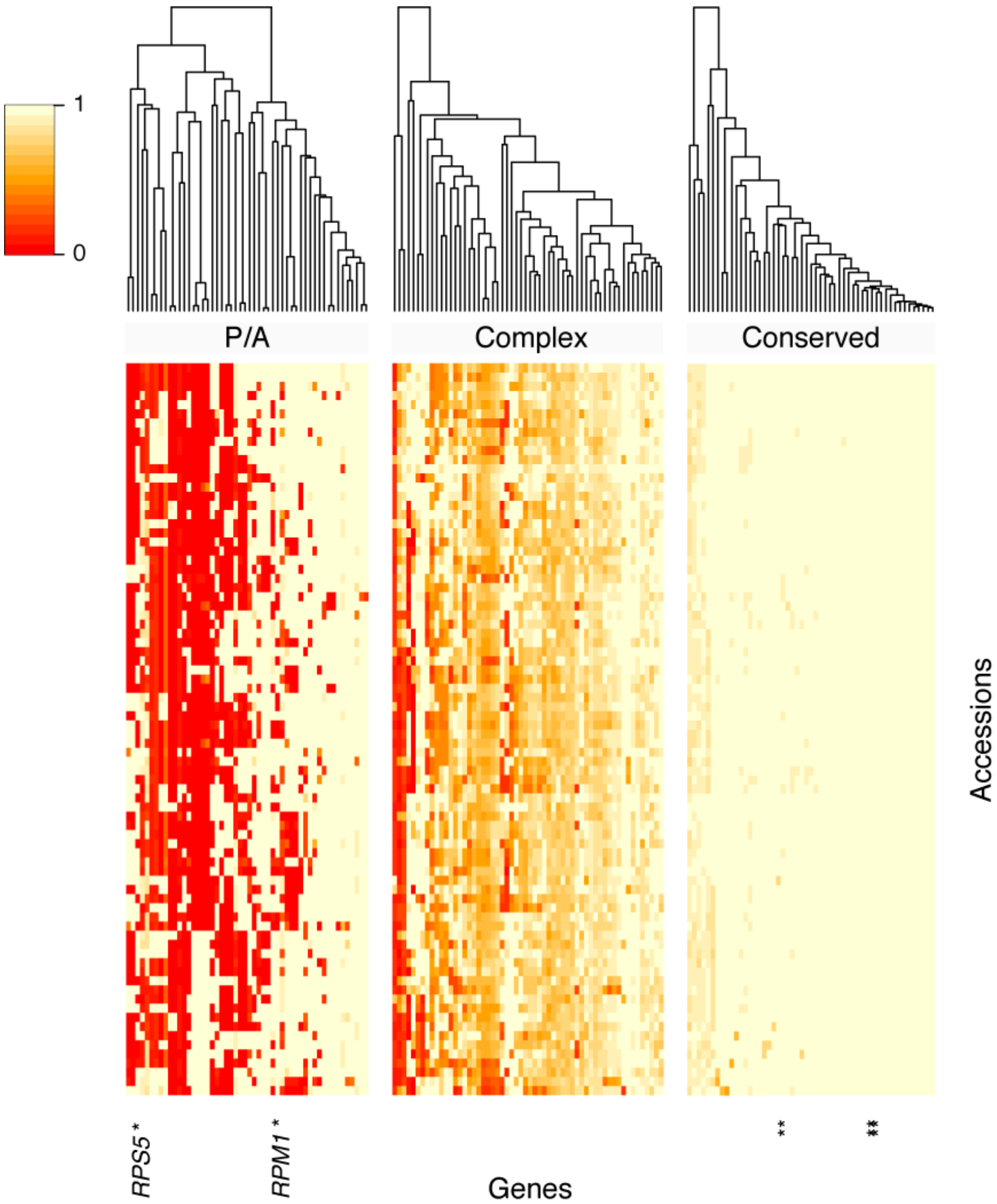

Figure 14. Heatmap of NLR gene presence. Double asterisks (**) indicate highly conserved genes identified by Hofberger and colleagues (2014). Single asterisks indicate known P/A genes (Grant et al., 1998; Henk et al., 1999).

3.1.2 Conserved Genes

All three NLR loci known to have high interspecies conservation (Hofberger et al., 2014) that were part of my NLR dataset - AT3G14470 (unknown), AT3G50950 (*ZAR1*) and AT4G33300 (*ADR1-L1*) – were classified as Conserved (marked with double asterisks in Figure 14) based on within-species variation.

3.1.3 Presence/Absence (P/A) Genes

RPS5 and *RPM1*, which are both well known P/A genes and which have been introduced in Sections 1.4.3 and 1.4.5, classify as P/A in this analysis (marked with single asterisks in Figure 14). Furthermore, of the nine P/A genes identified by Shen and colleagues (2006), all nine were classified here as P/A.

3.1.4 Complex Genes

RPP13 (*RESISTANCE TO PERONOSPORA PARASITICA 13*) is an example of a known NLR gene that is extremely diverse, and thought to be under both balancing and diversifying selection (Bittner-Eddy and Beynon, 2001; Rose et al., 2004). It can thus be considered an example of a known gene with a Complex pattern of diversity. *RPP13* falls within the Complex category as expected.

3.2 Gene Enrichment Analysis of NLR Classification Results

3.2.1 TNL and CNL Variation Patterns

As mentioned in the introduction, multiple differences between TNL and CNL genes exist. Although both TNL and CNL genes contain an NB domain, these domains segregate as monophyletic clades (Meyers et al., 1999). TNL type genes often encode multiple introns, while CNL type genes frequently encode a single exon (Meyers et al., 2003). In addition TNL genes are thought to be absent in monocots, such as grasses (Meyers et al., 1999). Whether the evolutionary modes and variation patterns of TNL and CNL genes differ, has not been established, though some studies suggest a difference (Yang et al., 2008; Chen et al., 2010).

To investigate whether there is enrichment of TNL and CNL genes in the Conserved, Complex and P/A categories, I carried out gene enrichment analysis on my variation pattern categorization and domain architecture categorization of 126 NLR genes with full TNL/CNL architectures, based on the study by Guo and colleagues (2011). There was no significant enrichment ($P = 0.63$; two-sided Fisher's Exact Test), with both TNL and CNL architectures represented in all three categories (Table 4).

The counts of TNL and CNL genes in each category were within two of the expected value, based on the total percentages of TNL and CNL genes (Table 2). For Complex genes the numbers matched exactly those expected from overall ratios. The ratios for TNL to CNL genes were 1.81, 1.63 and 2.55 for Complex, Conserved and P/A genes, respectively. The greatest difference in ratio was for P/A genes, where TNL genes were marginally overrepresented (Table 2).

	CNL	TNL	*Total*
Complex	16 (expected 16)	29 (expected 29)	*33*
Conserved	16 (expected 14)	26 (expected 28)	*64*
Presence/Absence	11 (expected 13)	28 (expected 26)	*29*
Total	*43 (34% of Total)*	*83 (66% of Total)*	*126*

Table 4. Contingency table of variation pattern (Conserved, Presence/Absence, Complex) and domain architecture (TNL, CNL) of 126 NLR genes with canonical full domain architectures based on data in Guo and colleagues (2011). Values indicate the number of genes in each category. Expected values were calculated based on the percentage of TNL and CNL genes out of total considered genes.

3.2.2 Clustered and Single Genes

I used the assignment of genes as single or clustered in Guo and colleagues (2011) to uncover differences in evolutionary pattern depending on the type of NLR gene considered. There was no significant enrichment of the categories (P = 0.052; two-sided Fisher's Exact Test), with both genomic arrangements represented in all three categories. A greater number of P/A and Complex genes was observed in the clustered category than would be expected based on the total proportion of clustered genes (Table 5).

	Clustered	Single	*Total*
Complex	43 (expected 39)	12 (expected 16)	*55*
Conserved	31 (expected 38)	22 (expected 15)	*53*
Presence/Absence	39 (expected 36)	12 (expected 15)	*51*
Total	*113 (71% of Total)*	*46 (29% of Total)*	*159*

Table 5. Contingency table of variation pattern (Conserved, Presence/Absence, Complex) and genomic arrangement (clustered, single) of 159 genes annotated in Guo and colleagues (2011). Values indicate the number of genes in each category. Expected values were calculated based on the percentage of clustered and single genes out of total considered genes.

3.3 Genome-Wide Classification to Identify Presence/Absence Genes for GWAS

3.3.1 Approach

To identify P/A genes genome-wide for subsequent P/A-based GWAS analysis, I prepared and mapped read data as described for 163 NLR genes (Section 2.1), using the entire *A. thaliana* genome as the reference (TAIR10; https://arabidopsis.org), and subsequently extracting CDS of the TAIR10 annotated 27206 genes, and classifying genes into the three categories using the same parameters as described previously (Section 2.1).

This initial assignment identified 23958 Conserved genes (88%), 1161 P/A genes (4%) and 2087 Complex genes (8%). My estimate of 4% P/A genes is lower than the previous estimate of 9% by Tan and colleagues (2012). This is partially due to the fact that I distinguish between Complex and Conserved, whereas the mentioned paper classified as P/A any gene which was absent in one or more accessions. For comparison, I calculated the total number of genes in which at least one accession had absence status. There were 1,988 such genes, constituting 7% of all genes, which is closer to the estimate by Tan and colleagues.

Assignment of genes from the whole genome into the three categories (Figure 15) is visually consistent with the expected results and with the previously obtained assignment of NLR genes (Figure 14). As expected from the decision rule used, the Complex category contains accessions where the gene appears to be absent. In all cases, however, respective genes also contain a substantial number - above 5% - of accessions with intermediate coverage values, explaining their assignment to the Complex category.

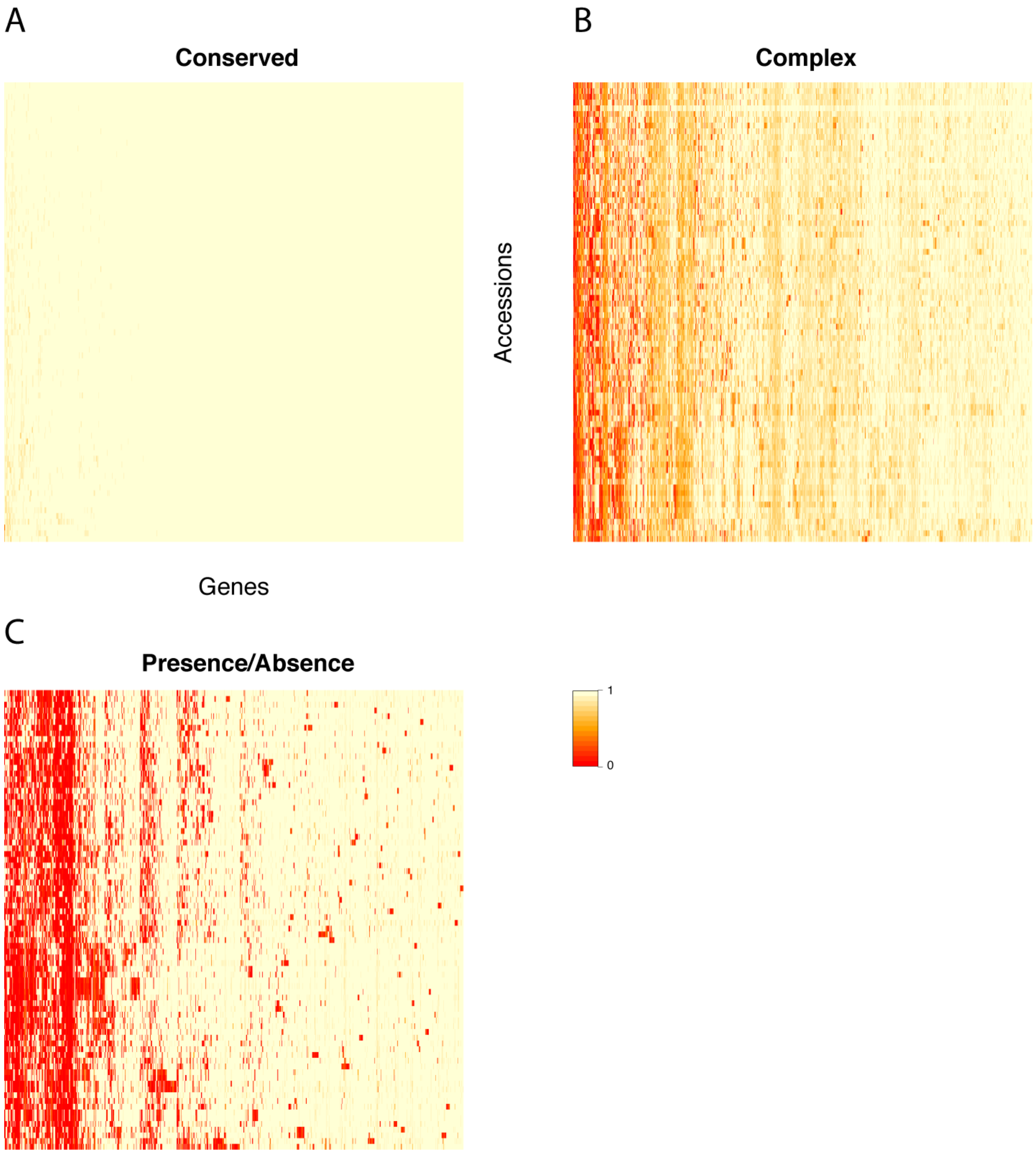

Figure 15. Overview of whole-genome classification results. For Conserved genes (Panel A), only the first 1000 genes were used, as the overall number was too large to plot.

3.3.2 Gene Ontology Analysis

Gene Ontology (GO) offers annotation of genes by Biological Process, Molecular Function and Cellular Component categories. I compared my three gene categories to GO category sets for overrepresentation (http://geneontology.org/; Ashburner et al., 2000; The Gene Ontology Consortium, 2017; Mi et al., 2017). A summary of analysis parameters is reported in Supplementary Methods.

Conserved genes were very substantially underrepresented in the Unclassified genes category for all three GO Annotation sets – Biological Process, Molecular Function and Cellular Component. The most significant value was a 0.68 fold underrepresentation in the Cellular Component set, with a false discovery rate (FDR) of 3.69e-33. In contrast, there was an overrepresentation across a wide range of categories, such as “binding” (1.09 fold enrichment; FDR 4.62e-13) for the Molecular Function GO set, and a 1.10 fold enrichment in “organic substance metabolic process” category (FDR 3.51e-12) for the Biological Process GO set.

Complex genes, in contrast, were substantially underrepresented across a very large number of categories. They were enriched, however, for “killing of cells of other organism” (4.47 fold enrichment; FDR 6.45e-24) and several other defence related categories in terms of Biological Process GO set. They were also enriched for “regulation of fertilization” (6.56 fold enrichment; FDR 6.56e-09). In terms of Molecular Function GO set, there was a 4.39 fold enrichment for ADP binding (FDR 4.48e-09). The most significant underrepresentation was in the “binding” category (0.53 fold; FDR 1.38e-56). In terms of cellular component, the most significant overrepresentation was in the “extracellular region” (1.45 fold; FDR 2.45e-08). Of the many underrepresented categories, the most significant was “intracellular organelle part” (0.30 fold; FDR 2.43e-56). Together, several of these results seem characteristic of plant defence components.

Presence/Absence genes were overrepresented for “defence response” (1.68 fold enrichment; FDR 9.95e-05), and underrepresented for many categories, such as “cellular process” (0.52 fold; FDR 2.87e-42), in terms of GO Biological Process set. For Molecular Function set, these genes were overrepresented for ADP binding (7.70 fold; FDR 2.52e-15); and underrepresented across multiple general categories, such as “binding” (0.58 fold; FDR 4.44e-26). There was an underrepresentation at the “intracellular organelle part” localization in terms of Cellular Component GO (0.18 fold; 4.47e-50). See also Section 3.4.3.

3.3.3 Enrichment of LRR Immune Receptor Families and F-box Genes

Apart from NLR genes, there are two further groups of LRR domain immune receptors – Receptor-Like Proteins (RLP) and Receptor-Like Kinases (RLK). These cumulatively are known as pathogen recognition receptors (PRR) and are localized to the plant membrane. RLP is a smaller LRR-carrying family with about 50 members (Wang et al., 2008). RLK genes are a very large family with upward of 600 members in *A. thaliana*, whose protein products perform various functions, including disease resistance, and can contain various domains, including LRRs (Shiu et al., 2001).

I looked for enrichment of my categories in 57 RLP (Wang et al., 2008), 605 RLK (Shiu et al., 2004) and 681 F-box genes (Xu et al., 2009). For consistency, I included only identifiers annotated as genes in the TAIR10 (https://arrabidopsis.org) release, thus excluding pseudogenes. Both RLP and F-box genes had fewer than expected Conserved genes, and more Complex and P/A ones, relative to the background. Thus, RLP and F-box genes were enriched in the variable categories relative to the expected values (Table 6). Overall, the distribution for RLP (*P* = 3.44e-09; two-sided Fisher’s Exact Test) and F-box (*P* = 1.93e-12; two-sided Fisher’s Exact Test) genes was significantly different from the background, although not for RLK genes (*P* = 0.57; two-sided Fisher’s Exact Test).

	RLP	RLK	F-box
Complex	13 (expected 4)	41 (expected 47)	82 (expected 52)
Conserved	31 (expected 49)	543 (expected 540)	536 (expected 600)
Presence/Absence	11 (expected 2)	29 (expected 26)	63 (expected 29)
Total	*55*	*613*	*681*

Table 6. Counts of genes from three selected families (RLP, RLK and F-box) in two variable categories (Presence/Absence and Complex) and one conserved (Conserved) category. The selected gene families include immune receptors (RLP and RLK) and F-box genes, which are involved in targeting proteins for degradation via ubiquitination.

Around 40% of RLK receptors carry LRR domains (Shiu et al, 2004). Several RLK receptors with LRR domains have been associated with PAMP recognition in plant immunity (reviewed in Nürnberger and Kemmerling, 2006; Böhm et al., 2014). To investigate whether the LRR-containing RLK genes are more variable than the other RLKs, we compared conservation values in the two groups. We found that gene conservation values were very similar, with 96% conservation in LRR-RLK genes, on average, and 97% in other RLK genes, as defined here. There was no statistically significant difference in conservation values between the two groups (P = 0.18; Wilcoxon rank sum test), suggesting that LRR-RLKs have similar conservation levels as the background RLK genes.

3.4 Genome-Wide Association Study Analysis

Presence/absence information for the P/A genes identified in the previous section was used to carry out a GWAS analysis of phenotypic traits related to fitness, including growth. Statistical power in a typical GWAS study depends on the number of variants and the number of accessions. Increasing the number of accessions has a positive effect on statistical power. However, increasing the number of variants – in this case, P/A genes, - decreases statistical power. In a typical set up like the one used in this study, each variant is tested for association separately, and thus the total number of variants has no effect on each individual test. However, at the stage of multiple testing correction, significance values are penalized for the total number of tests carried out, and thus the more variants or P/A genes have been tested, the heavier this penalty would be. For this reason, it was important to have a high quality set of P/A candidates to carry out the GWAS analysis.

Phenotypes for the plants were collected by Vasseur and colleagues (Vasseur et al., 2018a, 2018b, 2018c). For additional traits, unpublished phenotypic data were courtesy of Vasseur and colleagues. To carry out the GWAS study, I first needed to assign presence or absence status to individual accessions from the phenotyped set of accessions, which was a subset of the 1001 genomes data set (1001 Genomes Sequencing Consortium, 2016). I will briefly describe how this was carried out and then proceed to GWAS analysis and interpretation.

3.4.1 Assigning Presence and Absence Status to Individual Accessions

In order to assign presence or absence status to individual accessions from the 1001 genomes dataset in previously identified P/A genes, I used an existing mapping of reads (1001 Genomes Sequencing Consortium, 2016). While this mapping was generated for another purpose and does not follow the methods outlined in this study, it is nevertheless sufficient to assign either presence or absence status to accessions in genes already identified as P/A. I discretized coverage as descried above (Section 2.1) and calculated fraction of gene length with non-zero coverage for each accession-gene combination (as described in Section 2.5). I than applied a threshold of 0.5, corresponding to half the gene length with non-zero coverage, to assign presence and absence status to accessions for each gene.

To explore the effect of varying this threshold to above and below 0.5, I constructed a plot of how the number of accessions classified as present changes based on this value (Figure 16A). There is little difference for choices of threshold in the middle range of the plot. Thus, a simple choice of threshold of 0.5 is sufficient for this dataset.

To explore the effect of coverage on the number of alleles classified as absent, I plotted the number of gene alleles classified as present for each accession against coverage values for those accessions, based on the identified P/A genes and the 407 phenotyped accessions that were used for the GWAS (Figure 16B). I used coverage from the 1001 Genomes project (1001 Genomes Sequencing Consortium, 2016). There was a small correlation between the two quantities (R^2=0.030), meaning that less than 3% of the variance in the number of accessions classified as present can be accounted for by the effect of coverage. For this dataset, thus, the use of a simple threshold like 0.5 is justifiable. For datasets where the effect of coverage is large, threshold can be determined separately for each accession (adaptive threshold, suggested by François Vasseur).

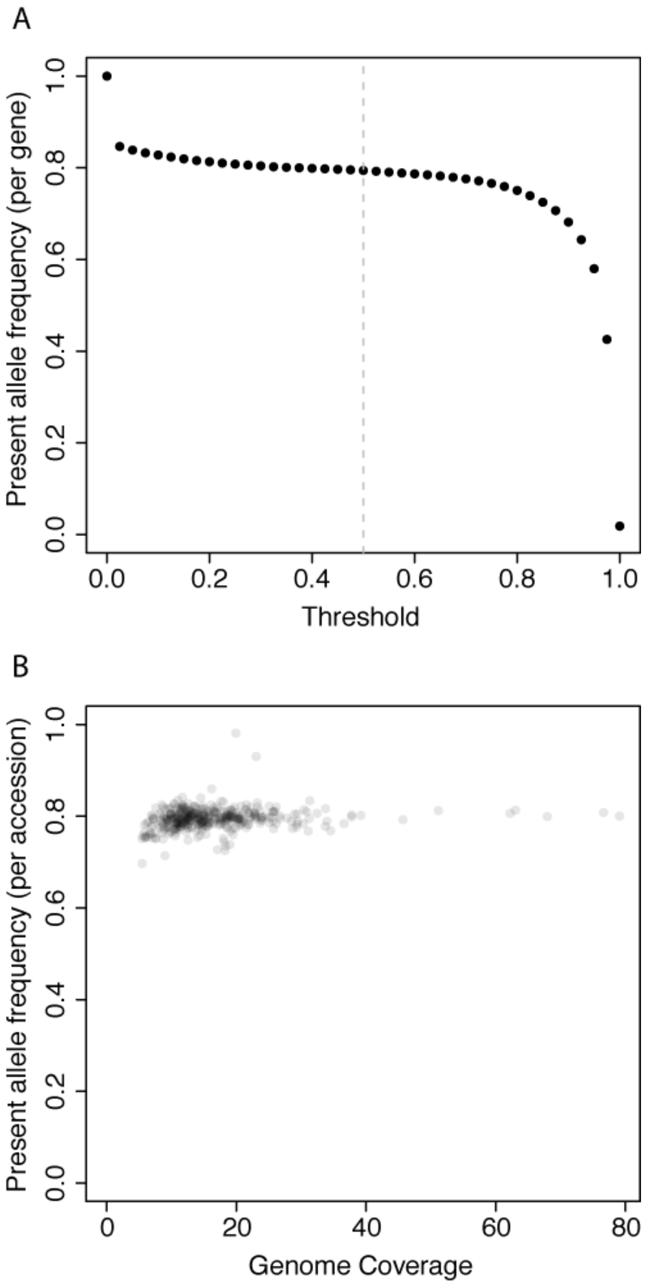

Figure 16. The effect of the choice of threshold and of genome coverage on the number of alleles classified as absent in the set of genes and accessions used for GWAS analysis. Panel A shows the effect of varying the threshold. Vertical dashed grey line corresponds to a choice of threshold of 0.5. Panel B shows the fraction of genes out of P/A candidates that were assigned presence status for each accession, and how this value varies with genome coverage for the corresponding accessions.

3.4.2 GWAS Analysis and Interpretation

I carried out a GWAS analysis on P/A polymorphism in the selected P/A genes using Single Variant EMMAX association analysis (see Supplementary Methods Section 7.3 for details) with population structure correction. Manhattan plots for four commonly used phenotypes out of 43 are shown in Figure 17.

Although none of the genes showed significant *P*-values at 0.05 level after Bonferroni correction for multiple testing, an interesting pattern emerged: three of the ten most significant tests correspond to *RPS5* (AT1G12220), for three different phenotypes (Table 7). While some phenotypes were correlated, consistency among multiple phenotypes could suggest correlation with a latent growth-related variable common to numerous phenotypes. All three *P*-values for *RPS5* are significant prior to multiple testing correction, with the most significant being 0.00011. The two most significant results correspond to the same F-box gene, AT1G67455. A second F-box gene, AT1G59680, is also in the list. Two further genes are of unknown families, annotated simply as "hypothetical protein" and "transmembrane protein" (AT5G49640 and AT2G18938, respectively; Araport11; http://arabidopsis.org). The final gene in the table, AT1G77150, is annotated as corresponding to Pentatricopeptide Repeat (PPR) superfamily member of uncertain function.

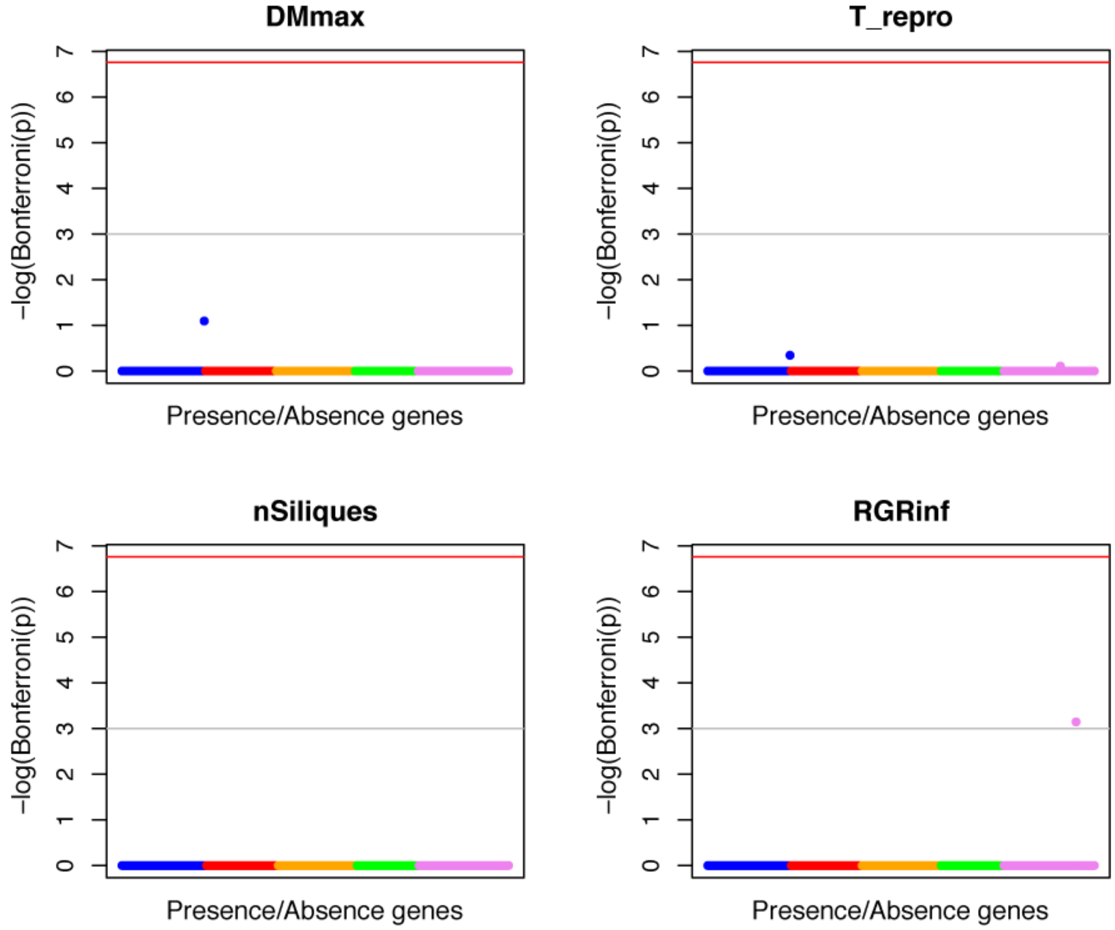

Figure 17. Manhattan plots for the Genome-Wide Association Study (GWAS) with P/A genes, showing four commonly used phenotypes out of the 43 tested. Point colors indicate chromosomes from which the P/A genes came for chromosomes one to five. Phenotypes shown are courtesy of Vasseur and colleagues (see also Vasseur et al., 2018a, 2018b, 2018c). Abbreviations are: DMmax=maximum dry mass, T-repro=time of reproduction (flowering), sSiliques=number of siliques, RGRinf=relative growth rate (mg d^{-1} g^{-1}) at inflection point. Grey lines indicate significance threshold after Bonferroni correction for the number of genes only; red lines indicate significance threshold after correction for the number of genes and the total number of phenotypes, 43, tested.

In all cases, for top-scoring *RPS5* phenotypes, the effect size was negative (see Beta column in (Table 7). This means that accessions where *RPS5* was absent had lower growth-related phenotype values (Figure 18A). These results were not directly consistent with previously published observations of a high fitness cost associated with carrying *RPS5* gene, including lower plant biomass (Karasov et al., 2014; discussed in Section 5.8).

Gene ID	MAF	Other Names	*P*-values	Beta	SE (Beta)	R^2	Phenotypes	*P* (Bonferroni)
AT1G67455	0.10	- (F-box)	0.000051	-0.860	0.210	0.040	nLeaves14d	1
AT1G67455	0.10	- (F-box)	0.000066	15.290	3.791	0.039	RGR14d	1
AT5G49640	0.19	-	0.000070	7.808	1.943	0.038	RGRinf	1
AT1G12220	0.42	*RPS5* (NLR)	0.000110	-0.380	0.097	0.036	RA14d	1
AT1G12220	0.42	*RPS5* (NLR)	0.000121	-0.116	0.030	0.036	ER14d	1
AT5G38180	0.06	-	0.000138	25.270	6.565	0.035	DMinfRA	1
AT1G59680	0.41	- (F-box)	0.000182	-0.538	0.142	0.034	rosDM14d	1
AT1G12220	0.42	*RPS5* (NLR)	0.000227	-4.343	1.167	0.033	RER14d	1
AT2G18938	0.22	-	0.000365	1.089	0.303	0.031	RootAlloc	1
AT1G77150	0.09	-	0.000389	-3.694	1.033	0.031	ReproAlloc	1

Table 7. The ten most significant GWAS test results, sorted by *P*-value. *P*-value in the last column is corrected for multiple testing using Bonferroni correction. Phenotypes column represents unique identifiers of various growth scaling and fitness trait measurements, courtesy of Vasseur and colleagues (see also Vasseur et al., 2018a, 2018b, 2018c). Briefly: A=area, GR=growth rate, 14d=measured at 14 days stage, inf=measured at inflection point, DM=dry mass, ros=rosette, T=time, germn=germination. SE (Beta) stands for standard error of Beta.

Although *RPM1* had neither significant *P*-values based on GWAS analysis, nor was among the highest-scoring genes, however, it is a P/A NLR gene that, like *RPM1*, is known to carry fitness costs. I compared effect size direction for this gene to *RPS5* for the most significant phenotype in *RPS5,* and found that in *RPM1* the effect size had the reverse direction (Figure 18B). In fact, the seven GWAS results with the highest significance that correspond to *RPS5* all have negative effect size values, while those same phenotypes for *RPM1* all have positive effect size values. This suggests that there is further unexplored complexity associated with fitness effects of these genes.

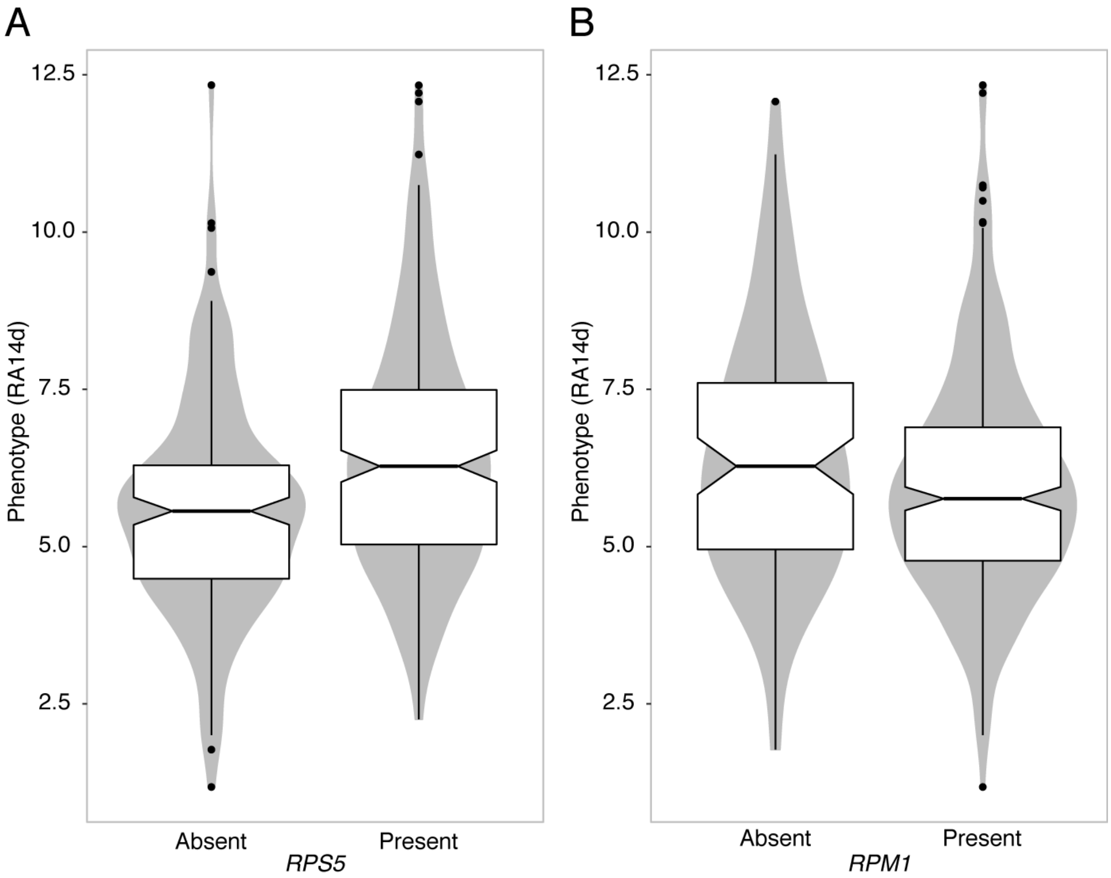

Figure 18. Boxplot of median phenotypes in the highest-scoring phenotype in GWAS analysis for the highest scoring gene *RPS5* (A). (B) displays the same phenotype in *RPM1* for comparison.

3.4.3 Genomic Context and Nucleotide Diversity Plots

The approach presented in this study provides an opportunity to examine not only genes themselves but also surrounding genomic regions. Two known and well-studied P/A genes are *RPS5* and *RPM1* (introduced in Sections 1.4.3 and 1.4.5). These genes are known to be under balancing selection (Stahl et al., 1999; Tian et al., 2002) and to carry fitness costs (Tian et al., 2003, Karasov et al., 2015). When represented within their genomic context using my conservation profiles, these genes show a clearly identifiable and separable deletion region (Figure 19).

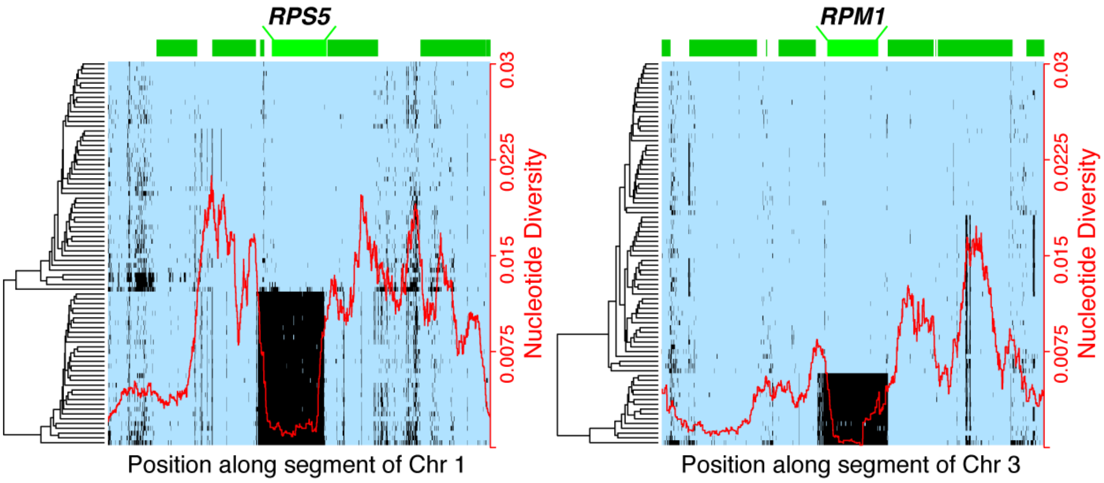

Figure 19. Coverage profiles showing *RPS5* and *RPM1* genes within their genomic context of 10 kb on either side. Nucleotide diversity for these regions is superimposed in red (averaged over 1000 bp windows). *RPS5*, *RPM1* and the surrounding genes are shown as green bars.

In terms of their variation pattern, it is known that these genes have a relatively low level of nucleotide diversity within the genic region among accessions in which the gene is present (Bergelson et al., 2001; Ding et al., 2007) compared to the high levels of nucleotide diversity surrounding the deletion junction, which then decrease over the adjacent 10kb (Stahl et al., 1999; Tian et al., 2002; Shen et al., 2006). These patterns are also clearly visible in my plots, consistent with the expectation (Figure 19). The genes also display relatively clear surrounding regions, without other deletions, although one possible contributing factor might be that *RPM1* is single and not part of a cluster of tandem duplicates. Finally, the loci show intermediate absence frequencies.

To further investigate the P/A gene candidates in the highest scoring GWAS associations list obtained in the previous step (Table 7), I generated plots showing conservation patterns of the seven genes (Figure 20), except *RPS5*, which has already been shown in Figure 19. For these plots, I used existing mapping (Cao et al., 2011) and included introns and the surrounding genomic regions of ten thousand nucleotides on either side. I then combined my profiles with nucleotide diversity plots for those regions.

When interpreting these plots, it is important to note that nucleotide diversity can only be calculated where coverage is present, and thus provides information on diversity within the blue areas of the plot only. This could potentially lead to an underestimate of nucleotide diversity where coverage is absent in a biased way - for example, in highly diverse regions. Displaying nucletide diversity and presence of coverage for the same positions allows to visually control for this effect and to interpret nucleotide diversity plots. Together, nucleotide diversity and presence of coverage offer complementary sources of information on polymorphism at different scales. Nucleotide diversity was calculated using vcftools version 0.1.12b (Danecek et al., 2011). In the plot, nucleotide diversity was averaged over one thousand base pair windows.

I subsequently visually examined the plots to select highest-quality candidates with patterns of variation similar to those of *RPS5* and *RPM1*, which are known to be under balancing selection and to carry high fitness costs. Out of the screened six genes, none showed the clear deletion pattern and a characteristic decrease in nucleotide diversity values, with only two genes approximating this pattern, AT5G49640 and AT1G59680 (Figure 20), which provides additional information on the quality of the selected candidates.

Furthermore, out of the 1,161 genome-wide identified P/A genes, only 35 (3% of P/A genes, 0.1% of total 27206 genes used in the study) showed patterns similar to *RPM1* and *RPS5*. These 35 genes correspond to 30 deletions, as 5 of the deletions spanned 2 genes each (see Table S2). Of these, only 7 genes, corresponding to 6 deletions, showed clear patterns comparable to *RPM1* and *RPS5* (Figure 21, except for *ADR1-L3*, which is shown in Figure 26).

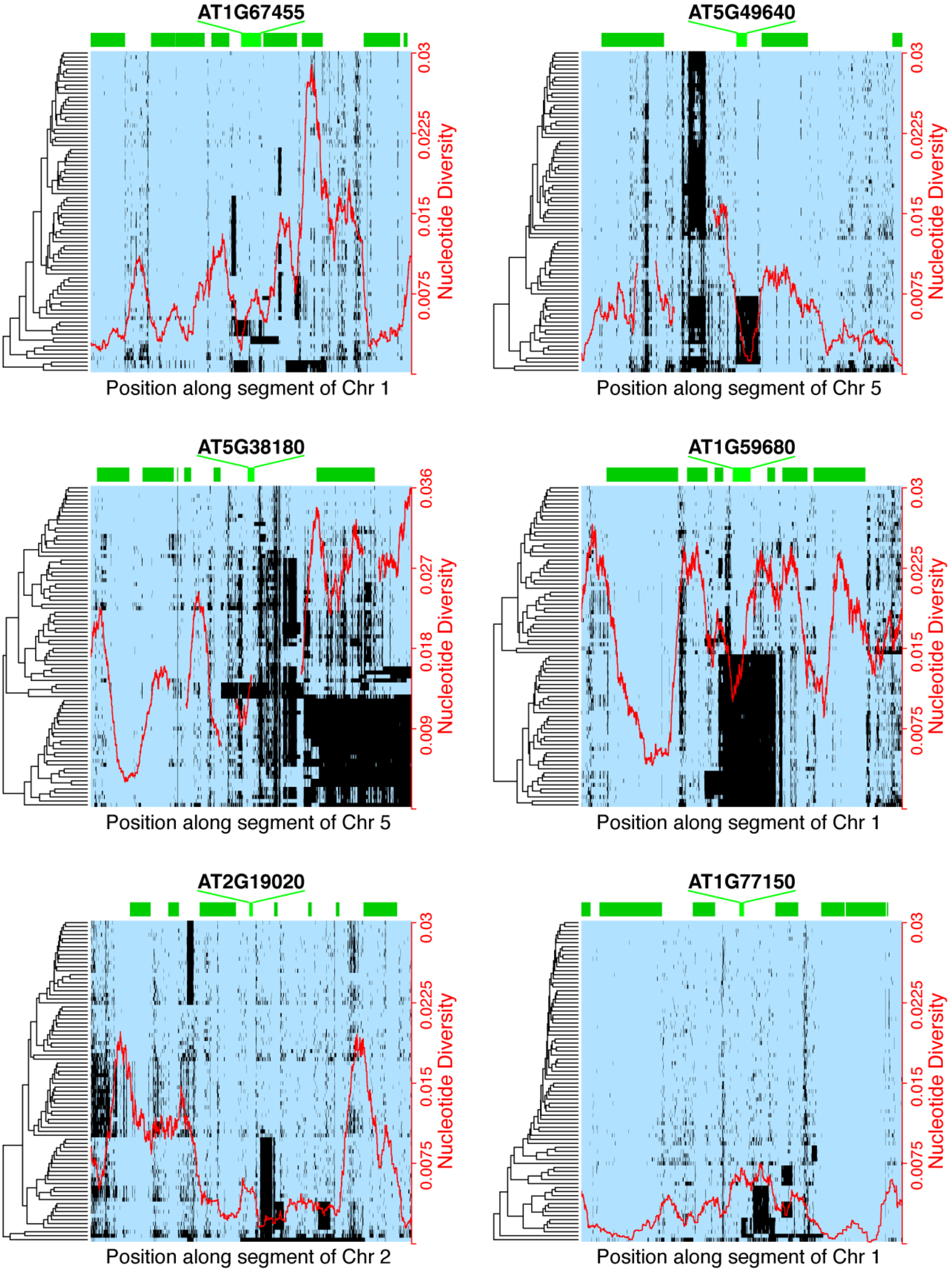

Figure 20. Coverage profiles of the highest-scoring P/A genes in GWAS analysis within their genomic context of 10 kb on either side. Nucleotide diversity for these regions is superimposed in red (averaged over 1000 bp windows). Title genes and the surrounding genes are shown as green bars.

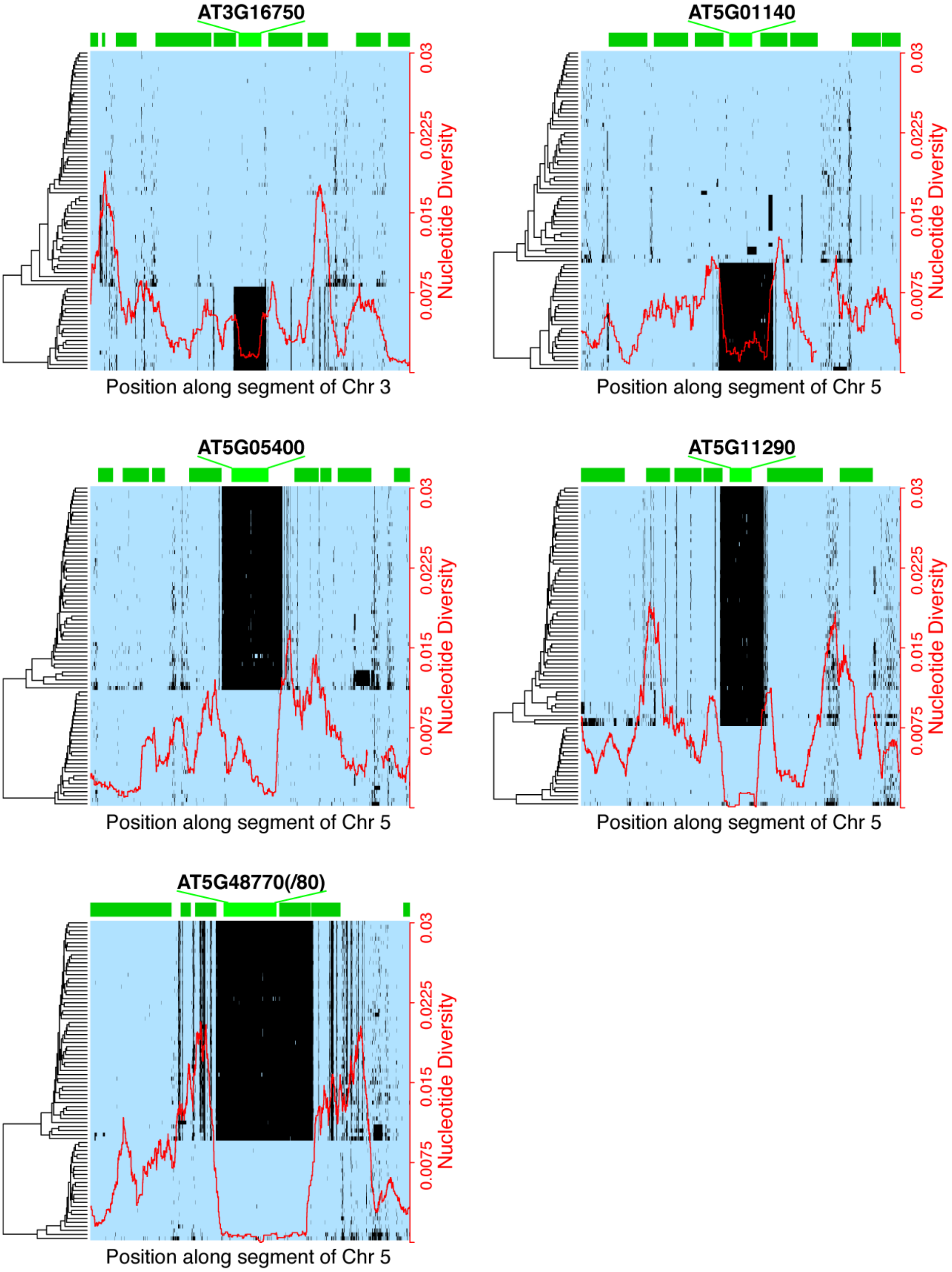

Figure 21. Coverage profiles of the visually selected genome-wide P/A genes with patterns similar to *RPM1* and *RPS5* within their genomic context of 10 kb on either side. Nucleotide diversity for these regions is superimposed in red (averaged over 1000 bp windows). Title genes and the surrounding genes are shown as green bars. In the last plot, the deletion encompasses two genes.

4 Chapter 3. Comparison with *Arabidopsis lyrata* and *Capsella rubella*

Arabidopsis lyrata and *C. rubella* represent the closest species and the closest genus to *A. thaliana*, respectively, and provide a context for understanding variation in *A. thaliana*. Comparisons between *A. thaliana* and its relatives have been used to understand the evolution of genome size and structure, polyploidy, and mating system shifts (reviewed in Koenig and Weigel, 2015).

Capsella rubella has a very low standing variation, as it is thought to have originated through an extreme population bottleneck, potentially by speciation from a single individual (Guo et al., 2009). It is thus undergoing initial stages of divergence and adaptation, and has been proposed as a model for understanding these processes (Guo et al., 2009) and well as for understanding mating system shifts (Slotte et al., 2013; reviewed in Koenig and Weigel, 2015).

Arabidopsis lyrata is the sister species of *A. thaliana*. Its NLR complement of genes has been compared to that of *A. thaliana* (Guo et al., 2011). For both *A. lyrata* and *C. rubella*, reference accession genomes have been sequenced (Hu et al., 2011 and Slotte et al., 2013, respectively).

In this study, I used 26 accessions of *A. lyrata* and 22 accessions of *C. rubella* to assess between-species levels of polymorphism and how they relate to within-species levels of polymorphism. I compared inter- and intra- specific variation patterns in the set of NLR genes from *A. thaliana*. First, I provide an overview of variation patterns in the three species. Subsequently, I compare variation patterns in *A. lyrata* and *C. rubella* and how these correspond to the three previously identified gene categories based on patterns of variation in *A. thaliana*: Conserved, P/A and Complex. I identify a subset of highly conserved genes in all three species and test this subset for enrichment in genomic distribution and domain architecture categories. I then compare interspecific conservation in NLR genes to other variable gene families. Finally, I survey intraspecific variation in NLR genes of the two species by mapping their reads to their own reference genomes.

4.1 Overview of Between-Species Conservation Patterns

To compare conservation in NLR genes in *A. thaliana* to multiple *A. lyrata* and *C. rubella* accessions, I searched the Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) for reads from these accessions. I retrieved reads for the 22 accessions of *C. rubella* from the first stage of *C. rubella* sequencing project (SRA project PRJEB6689), sequenced by Daniel Koenig from Detlef Weigel laboratory (Ågren et al., 2014; see Supplementary Methods). *Arabidopsis lyrata* reads for 26 accessions were also downloaded from SRA (Project PRJNA459481; Hämälä et al., 2018). Reads were pre-processed and mapped, and CDS were extracted, as described in Section 3.3.1.

Arabidopsis lyrata shows overall higher values and thus more similarity to *A. thaliana* than does *C. rubella*, as expected from them being more closely related and part of the same genus based on species phylogeny. Concordant with this, more genes have high coverage values and the highest values are higher in *A. lyrata* than in *C. rubella*, which is particularly visible in the Conserved genes category (Figure 22). Not surprisingly, conservation in accessions of *A. thaliana* was very visibly higher than among different species.

Overall, Conserved genes show clearly higher coverage values than any other category in both *A. lyrata* and *C. rubella* (Figure 22). However, not all genes in the Conserved category display high coverage. This demonstrates that NLR genes conserved in *A. thaliana* can have varying degrees of conservation in the other two species.

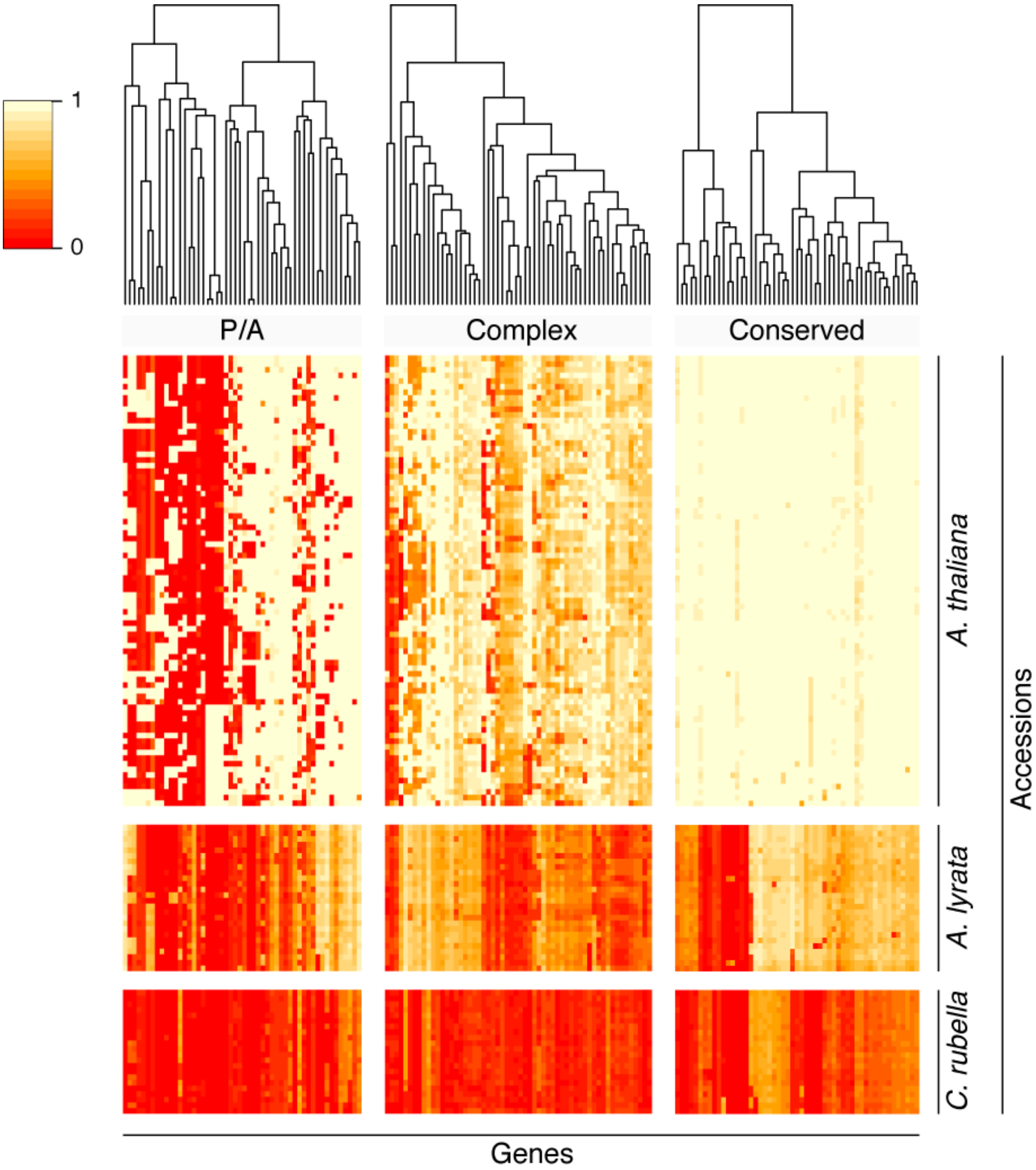

Figure 22. Heatmap representation of the fraction of reference CDS length in 163 NLR genes covered by reads from *A. thaliana*, *C. rubella* and *A. lyrata* accessions. Clustering of genes is done separately for each conservation category in *A. thaliana*.

4.2 Comparison of Between-Species Conservation to Within-Species Polymorphism Pattern

4.2.1 Comparison with *A. lyrata*

To see how the between-species conservation values relate to within-species polymorphism, I looked at the distribution of conservation values in the three previously identified categories of NLR genes based on within-species polymorphism patterns in *A. thaliana*: Conserved, P/A and Complex (Figure 23).

Conserved genes in *A. thaliana* were clearly also the most conserved category in *A. lyrata*, with both median and mode above 50% (Figure 23A). P/A and Complex genes had distributions with both median and mode below 50%. P/A genes showed the lowest median and mode, both below 25%, when average conservation values were considered (Figure 23A). This was changed, however, when the maximum conservation of any accession was used to represent each gene (Figure 23B), rather than average (Figure 23A).

The switch to using maximum conservation values of any accession rather than the average had almost no effect on the overall shape of the Conserved NLR distribution, and little effect on the shape of the Complex NLR distribution. In contrast, the shape of the P/A NLR distribution was inverted, with the mode rising far above 50%. Thus, the reason for the lowest average conservation values in the P/A category was likely the absence of genes in some accessions, which indicates that P/A polymorphism is maintained across speciation barrier at least for some P/A genes (explored in more detail in Section 4.3).

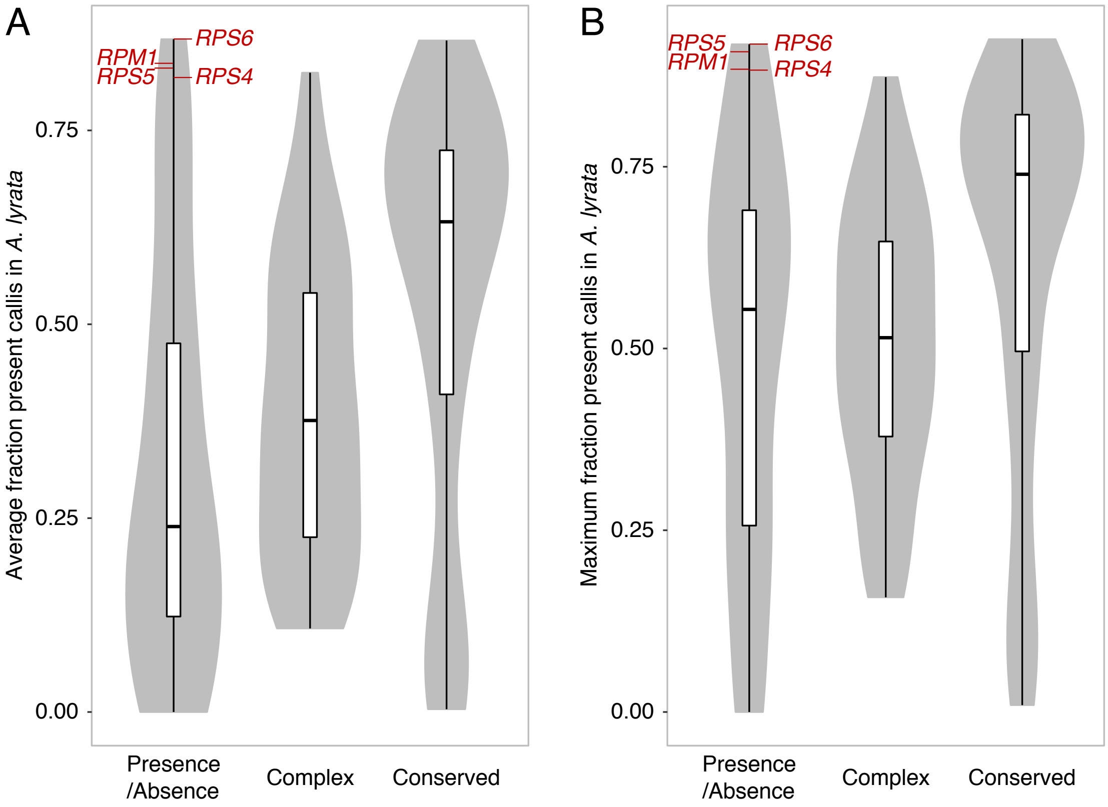

Figure 23. Violin plot of average (A) and maximum (B) conservation of 163 NLR genes from *A. thaliana* across 26 *A. lyrata* accessions. Boxplot is in black and white with the highest raking P/A genes labeled in red. Density distributions were trimmed to fit the range of data.

Although P/A genes overall show the lowest average conservation values based on between-species comparison, this gene group also contains several highly conserved genes (Figure 23A), two of which were found to correspond to well-known NLR P/A genes *RPM1* and *RPS5*, ranking as second and third most conserved, respectively. Two further NLR genes, *RPS6* and *RPS4*, ranked as first and fourth by conservation. Maximum conservation values also mirrored this pattern (Figure 23B), with the four most conserved genes remaining in the same order.

RPM1 and *RPS5* were also identified as outliers in other sections of this study, and the present finding reinforces their uniqueness among NLR P/A genes. The fact that these genes do not change their ranking when maximum values are used instead of average values points to the fact that these genes might be fixed and no longer segregate as P/A polymorphism in *A. lyrata* populations (further explored in Section 4.3). Thus, some genes under P/A polymorphism in *A. thaliana* seem to also display P/A polymorphism in *A. lyrata*, but others appear to be fixed.

4.2.2 Comparison with *C. rubella*

For *C. rubella*, analysis analogous to *A. lyrata* was carried out. Similarly, genes identified as Conserved based on within-species comparisons were also the most conserved based on between-species comparisons (Figure 24). Similarly to *A. lyrata*, P/A genes had the lowest mode and median of the three categories, with Conserved genes showing on average the most presence of coverage calls and P/A genes the least.

Using maximum conservation value rather than the average had less drastic effect on the overall shape of the P/A gene distribution in the case of *C. rubella* (Figure 24B) than it did for *A. lyrata* (Figure 23B). This possibly reflects the fact that P/A polymorphism was less well maintained across the genus barrier than it was across the species barrier.

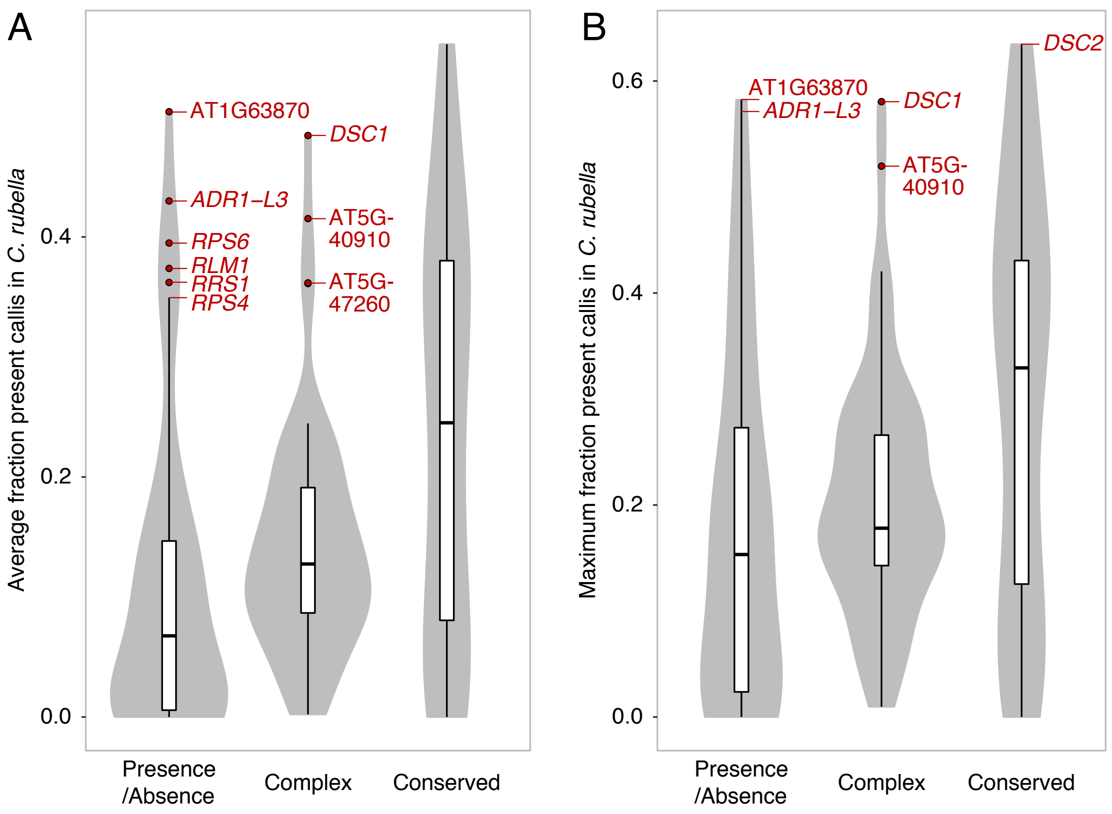

Figure 24. Violin plot of average (A) and maximum (B) conservation of 163 NLR genes from *A. thaliana* across *C. rubella* accessions. Boxplot is in black and white with outliers shown as dots and both outliers and the highest-ranking genes labeled in red. Density distributions were trimmed to fit the range of data.

Most highly conserved P/A genes in *C. rubella* were AT1G63870 and *ADR1-L3*, likewise for maximum and average conservation (Figure 24A, B). Polymorphism pattern for *ADR1-L3* is clearly visible in heatmap representation in Section 4.3. Additionally, a Complex gene *DSC1* (*DOMINANT SUPRESSOR OF CAMTA3 NUMBER 1*) was detected as an outlier with unusually high conservation values (Figure 24A-B). A Conserved gene, *DSC2*, had the highest maximum conservation value of any NLR gene.

4.3 Maintenance of Presence/Absence genes across species barrier

Out of P/A genes in *A. thaliana*, a significant number show conservation in *A. lyrata*, and fewer in *C. rubella* (Figure 25). The gene showing the most clear pattern across all three species is AT5G47280 or *ADR1-L3*, which has been classified as a P/A gene in *A. thaliana*. Its nucleotide diversity and genomic context profile (Figure 26) shows patterns characteristic of experimentally well characterized P/A genes *RPM1* and *RPS5* (Section 3.4.3), including clear genomic context, increase in nucleotide diversity around deletion junctions and intermediate frequencies of absence alleles. It was also highlighted as the P/A gene with the second highest conservation in *C. rubella* in Section 4.2.2. This evidence makes it a candidate gene for having maintained P/A polymorphism across genera.

A total of 21 P/A genes appear to have P/A-like patterns in *A. lyrata* and three further genes in addition to *ADR1-L3* appear to have a P/A-like pattern in both *A. lyrata* and *C. rubella*, although for two of these the pattern is less clear (Figure 25B). An additional five genes appear to be under P/A polymorphism in *C. rubella*, although not in *A. lyrata*.

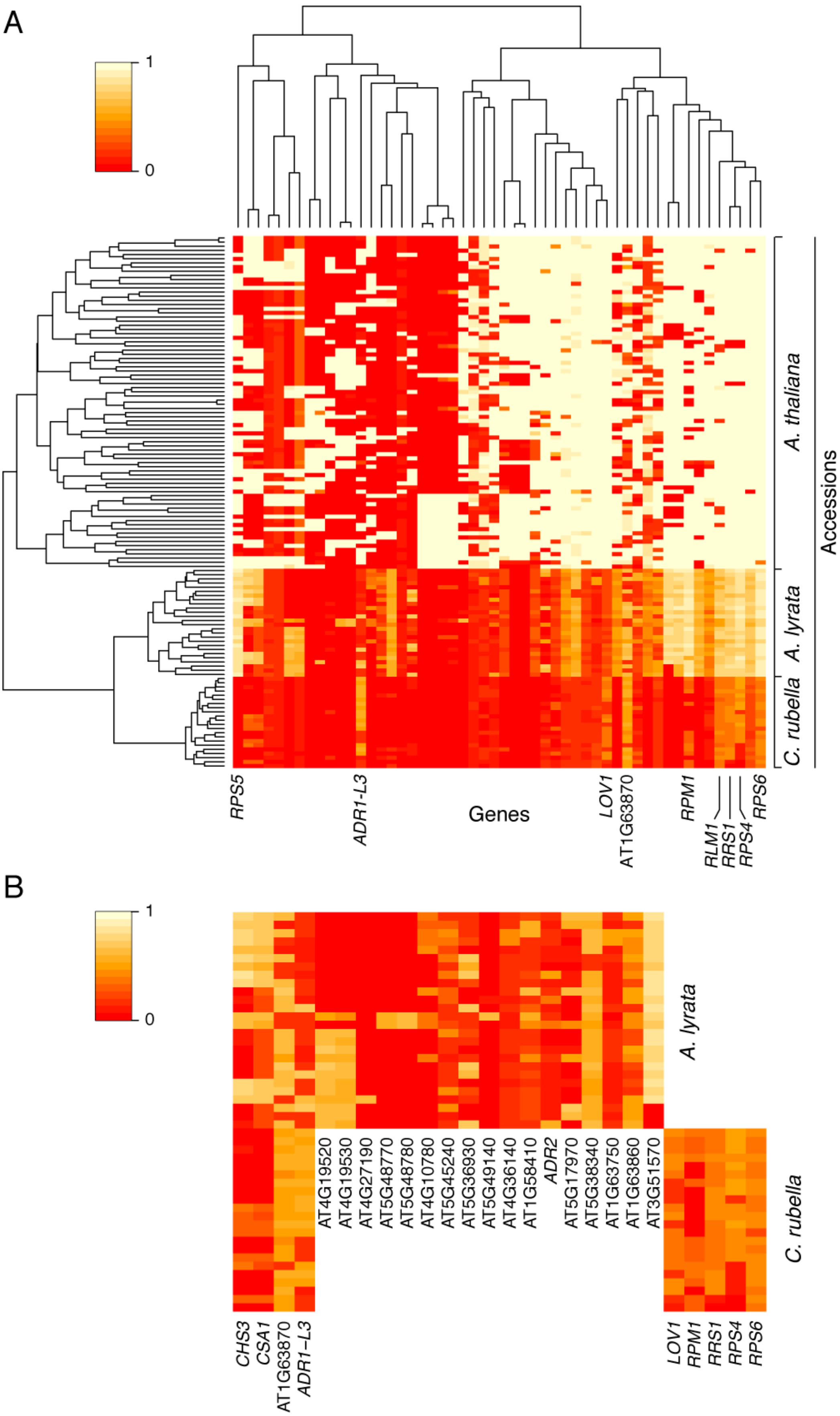
A
1
0
RPS5
ADR1-L3
Genes
LOV1
AT1G63870
RPM1
RLM1
RRS1
RPS4
RPS6
A. thaliana
A. lyrata
C. rubella
Accessions
B
1
0
CHS3
CSA1
AT1G63870
ADR1-L3
AT4G19520
AT4G19530
AT4G27190
AT5G48770
AT5G48780
AT4G10780
AT5G45240
AT5G36930
AT5G49140
AT4G36140
AT1G58410
ADR2
AT5G17970
AT5G38340
AT1G63750
AT1G63860
AT3G51570
LOV1
RPM1
RRS1
RPS4
RPS6
A. lyrata
C. rubella

Figure 25. NLR Presence/Absence (P/A) genes clustered based on *C. rubella* and *A. lyrata* data, as described for Figure 22 (A) and selected genes that appear to have P/A-like patterns in *A. lyrata* or *C. rubella* (B).

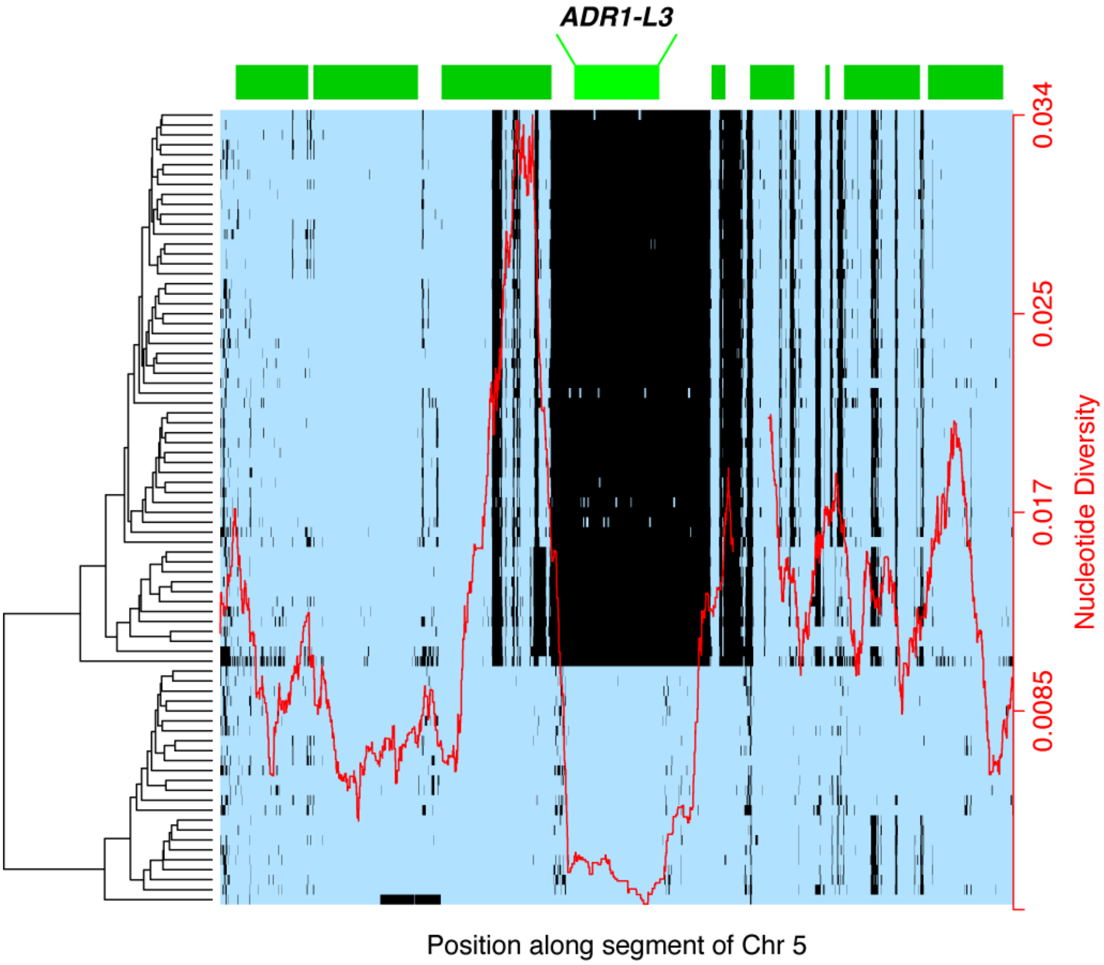

Figure 26. Coverage profiles showing *ADR1-L3* NLR gene within its genomic context of 10 kb on either side. Nucleotide diversity for these regions is superimposed in red (averaged over 1000 bp windows). *ADR1-L3* and surrounding genes are shown as green bars.

Apart from *ADR1-L3*, *RPM1* and three further genes appear to have patterns reminiscent of P/A polymorphism in *C. rubella* (Figure 25). Table 8 shows the five NLR P/A genes that are most conserved in *C. rubella* based on maximum values in any accession. In *A. lyrata*, both *RPM1* and *RPS5* are highly conserved, although they do not appear to show P/A-like patterns.

Rank	Gene ID	Other Names	Maximum fraction of presence calls in *C. rubella* accessions
1	AT1G63870	-	0.58
2	AT5G47280	*ADR1-L3*	0.57
3	AT5G45250	*RPS4*	0.51
4	AT5G46470	*RPS6*	0.46
5	AT5G45260	*RRS1*	0.44

Table 8. Presence/Abesence NLR genes from *A. thaliana* with the highest maximum conservation in *C. rubella*.

In *A. lyrata*, multiple genes appear to have P/A-like patterns, some of which are very conserved. Table 9 shows five NLR P/A genes which were most conserved in *A. lyrata* based on maximum value in any accession. Two NLR P/A genes with high conservation values in both *A. lyrata* and *C. rubella* were *RPS4* and *RPS6*, suggesting an important role for these gene across multiple species related to *A. thaliana*. These genes, like *RPS5*, are named after their involvement in resistance to the bacterial pathogen *Pseudomonas syringae*.

Rank	Gene ID	Other Names	Maximum fraction of presence calls in *A. lyrata* accessions
1	AT5G46470	*RPS6*	0.92
2	AT1G12220	*RPS5*	0.91
3	AT3G07040	*RPM1*	0.88
4	AT5G45250	*RPS4*	0.88
5	AT3G51560	-	0.88

Table 9. Presence/Absence NLR genes from *A. thaliana* with the highest maximum conservation in *A. lyrata*.

4.4 Conservation Across Species

Both *C. rubella* and *A. lyrata* accessions cluster together and form monophyletic clades (Figure 27). Over a half of NLR genes appear to be conserved in *A. lyrata*, of which 48 show high conservation values (A*; Figure 27). There seems to be a small but clearly identifiable cluster of highly conserved genes in *C. rubella* (B* highlighted in red; Figure 27). This cluster consists of 34 genes, out of which all appear highly conserved in *A. lyrata* and 24 have been classified as Conserved in *A. thaliana*. I have thus obtained a list of 24 *A. thaliana* NLR genes that are classified as Conserved based on both within-species and between-species comparisons.

There is no perfect correlation between conservation in *A. lyrata* and *C. rubella*. While a substantial number of genes show high conservation numbers across both species, approximately a third of genes highly conserved in *A. lyrata* seem absent or highly diverged in *C. rubella* (C* highlighted in green; Figure 27).

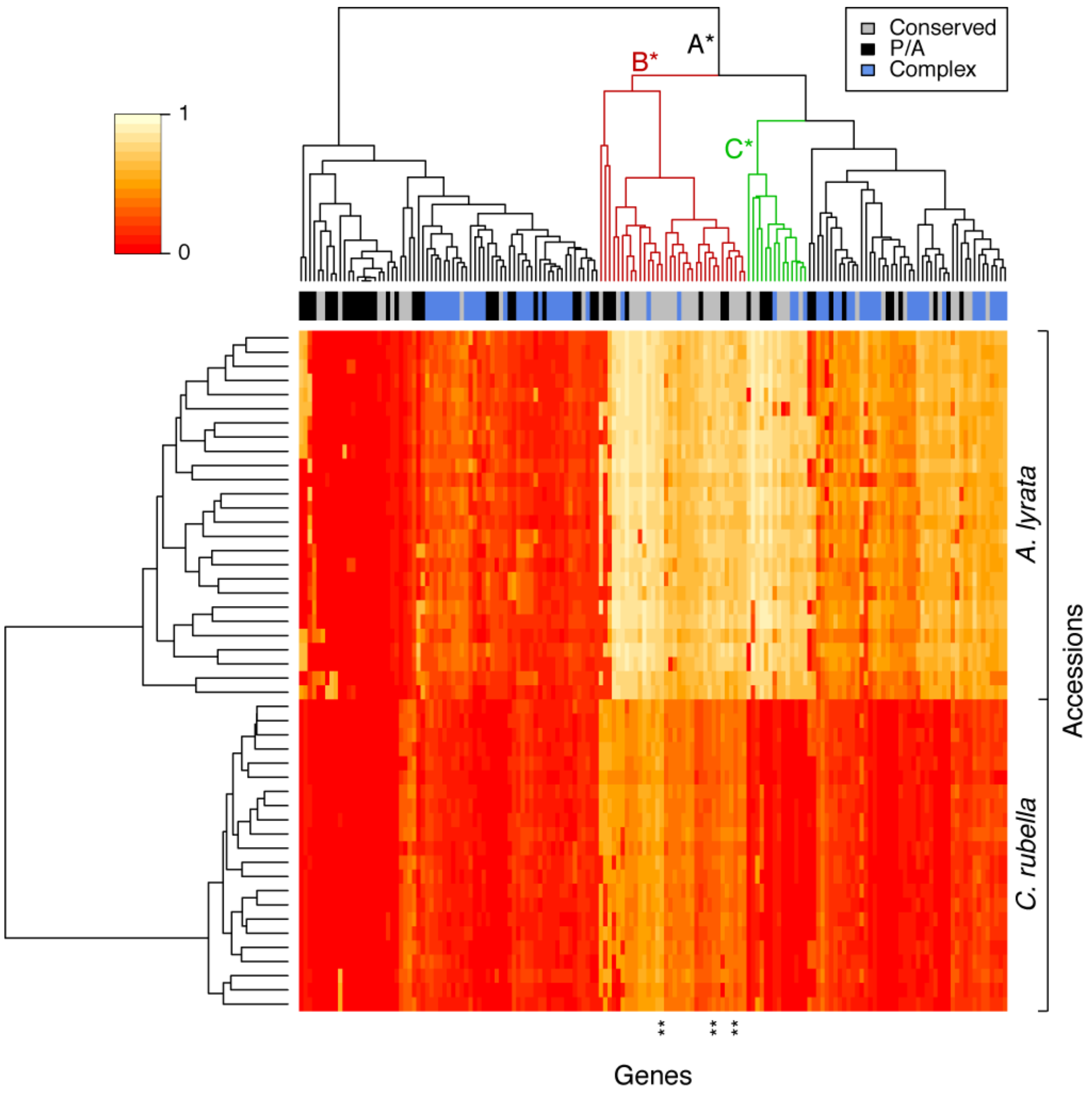

Figure 27. Heatmap representation of the fraction of reference CDS length in 163 NLR genes covered by reads from *C. rubella* and *A. lyrata* accessions. Clade A* represents genes conserved in *A. lyrata*; Clade B* represents genes that are conserved in both *A. lyrata* and *C. rubella*; Clade C* represent genes that are highly conserved in *A. lyrata*. Double asterisks (**) represent genes known to be conserved across multiple accessions from literature (Hofberger et al., 2014).

I found a highly significant enrichment of Conserved genes in both *A. lyrata* (*P* = 3.22e-07; two-sided Fisher's exact test) and *C. rubella* highly conserved clusters (*P* = 5.11e-07; two-sided Fisher's exact test). In *A. lyrata*, of the 48 highly conserved genes, 30 belonged to the Conserved category in *A. thaliana*, compared to the expected 16 (Table 10). In *C. rubella*, 24 of the 34 conserved genes belonged to the Conserved category in *A. thaliana*, compared to the expected 7 (Table 11). Of the 24 genes, both TNL and CNL domain architectures, as well as partial architectures CN, TN, NL were represented (Table 12).

	Highly conserved in *A. lyrata*	Background	*Total*
Conserved in *A. thaliana*	30 (expected 16)	23 (expected 37)	*53*
Background	18 (expected 32)	92 (expected 78)	*110*
Total	*48 (29% of Total)*	*115 (71% of Total)*	*163*

Table 10. Contingency table of *A. thaliana* NLR genes highly conserved in *A. lyrata* and NLR genes belonging to the Conserved category in *A. thaliana*. Values indicate the number of genes in each category. Expected values were calculated based on the percentage of clustered and single genes out of the total considered genes.

	Conserved in *C. rubella*	Background	*Total*
Conserved in *A. thaliana*	24 (expected 11)	29 (expected 42)	*53*
Background	10 (expected 23)	100 (expected 87)	*110*
Total	*34 (21% of Total)*	*129 (79% of Total)*	*163*

Table 11. Contingency table of *A. thaliana* NLR genes conserved in *C. rubella* and NLR genes belonging to the Conserved category in *A. thaliana*. Values indicate the number of genes in each category. Expected values were calculated based on the percentage of clustered and single genes out of the total considered genes.

Gene ID	Other Names	Domain Architecture	Average
AT1G12280	*SUMM2*	CNL	1.00
AT1G12290	-	CNL	1.00
AT1G50180	-	CN	0.99
AT1G52660	-	CN	1.00
AT1G53350	-	CNL	0.99
AT1G63730	-	TNL	0.99
AT1G72950	-	TN	1.00
AT3G14460	-	NL	1.00
AT3G14470	-	CNL	0.99
AT3G15700	-	CN	1.00
AT3G50950	*ZAR1*	CNL	0.99
AT4G26090	*RPS2*	CNL	1.00
AT4G33300	*ADR1-L1*	CNL	0.99
AT5G04720	*ADR1-L2*	CNL	1.00
AT5G17680	-	TNL	0.99
AT5G18360	-	TNL	0.99
AT5G22690	-	TNL	1.00
AT5G40090	*CHL1, CHS1-L1*	TN	0.99
AT5G40100	-	TNL	1.00
AT5G43470	*RPP8*	CNL	0.92
AT5G45060	-	TNL	0.95
AT5G47250	-	CNL	0.98
AT5G48620	-	CNL	0.88
AT5G66900	-	CNL	1.00

Table 12. Eighteen genes from the Conserved category in *A. thaliana* that were found to be also present in *A. lyrata* and *C. rubella*, sorted by their gene ID. Average column displays conservation values in *A. thaliana* averaged over 80 accessions, calculated as fraction of total CDS length that had non-zero read coverage. Domain architecture data is from Guo and colleagues (2011).

4.5 Enrichment in Gene Categorizations for Between-Species Conservation

4.5.1 Clustered and Single NLR Genes

To assess whether clustered and single NLR genes follow different patterns of between-species conservation, I looked for enrichment of the two categories in the subsets of NLR genes highly conserved in *A. lyrata* and *C. rubella*. I found a statistically significant enrichment for genomic arrangement in the 34 highly conserved NLR genes in *C. rubella* (P = 0.011; two-sided Fisher's Exact Test), all of which were also present in *A. lyrata*, and an almost significant enrichment in the 48 highly conserved NLR genes in *A. lyrata* (P = 0.059; two-sided Fisher's Exact Test). Both genomic arrangements were represented in both interspecific conservation categories. However, a greater number of single genes was observed in the highly conserved subsets based on comparisons with both *A. lyrata* (Table 13) and *C. rubella* (Table 14). This is consistent with the previously described within-species comparisons.

	Clustered	Single	*Total*
Highly conserved in *A. lyrata*	29 (expected 34)	19 (expected 14)	*48*
Background	84 (expected 79)	27 (expected 32)	*111*
Total	*113 (71% of Total)*	*46 (29% of Total)*	*159*

Table 13. Contingency table of conservation in *A. lyrata* and genomic arrangement (clustered, single) of 159 NLR genes annotated in Guo and colleagues (2011). Values indicate the number of genes in each category. Expected values were calculated based on the percentage of clustered and single genes out of the total considered genes.

	Clustered	Single	*Total*
Conserved in *C. rubella*	18 (expected 24)	16 (expected 10)	*34*
Background	95 (expected 89)	30 (expected 36)	*125*
Total	*113 (71% of Total)*	*46 (29% of Total)*	*159*

Table 14. Contingency table of conservation in *C. rubella* and genomic arrangement (clustered, single) of 159 NLR genes annotated in Guo and colleagues (2011). Values indicate the number of genes in each category. Expected values were calculated based on the percentage of clustered and single genes out of the total considered genes.

4.5.2 NLR Genes with TNL and CNL Domain Architectures

Previously, I found no significant differences in the conservation patterns of CNL and TNL genes based on within-species comparisons in *A. thaliana*. To investigate whether between-species comparisons might show a significant correlation, I assessed enrichment of TNL and CNL genes in the two subsets of genes highly conserved in *A. lyrata* and *C. rubella*. I found no significant enrichment for either *A. lyrata* (*P* = 0.22; two-sided Fisher's Exact Test) or *C. rubella* (*P* = 0.058; two-sided Fisher's Exact Test). However, CNL genes were present slightly more frequently than expected among the genes conserved in both *A. lyrata* (Table 15) and *C. rubella* (Table 16).

	CNL	TNL	*Total*
Highly conserved in *A. lyrata*	16 (expected 13)	21 (expected 24)	*37*
Background	27 (expected 30)	62 (expected 59)	*89*
Total	*43 (34% of Total)*	*83 (66% of Total)*	*126*

Table 15. Contingency table of conservation in *A. lyrata* and domain architecture (TNL, CNL) of 159 genes annotated in Guo and colleagues (2011). Values indicate the number of genes in each category. Expected values were calculated based on the percentage of clustered and single genes out of the total considered genes.

	CNL	TNL	*Total*
Conserved in *C. rubella*	13 (expected 9)	12 (expected 16)	*25*
Background	30 (expected 34)	71 (expected 67)	*101*
Total	*43 (34% of Total)*	*83 (66% of Total)*	*126*

Table 16. Contingency table of conservation in *C. rubella* and domain architecture (TNL, CNL) of 159 genes annotated in Guo and colleagues (2011). Values indicate the number of genes in each category. Expected values were calculated based on the percentage of clustered and single genes out of the total considered genes.

4.6 Comparison of Immune Receptor Families and F-box Genes

I compared conservation of NLR, RLP, RLK and F-box genes. All these families are known to be variable and their protein products - apart from those of F-box genes, which are involved in targeting proteins for degradation - all include groups of immune receptors. I obtained gene lists for these families as described in Section 3.3.3, removing pseudogenes and genes that were no longer annotated in TAIR10 (http://arabidopsis.org). I then took an average of the presence of coverage calls as a fraction of total CDS length over all accessions for each gene. The resulting values were compared among the gene families.

I found highly significant differences among the gene families for both *A. lyrata* (P < 2.2e-16; Kruskal-Wallis rank sum test) and *C. rubella* (P < 2.2e-16; Kruskal-Wallis rank sum test). RLK genes appeared substantially more conserved than the other three families, followed by RLP genes, in both species (Figure 28). In both species, NLR genes were more variable than RLP genes. However, while in *A. lyrata* NLR genes were slightly more variable on average than F-box genes (Figure 28A), in *C. rubella* there was no clear difference in the median values for the two families, with the median even marginally lower for F-box genes (Figure 28B), suggesting that over longer evolutionary distances, F-box genes might be equally or more variable than NLR genes.

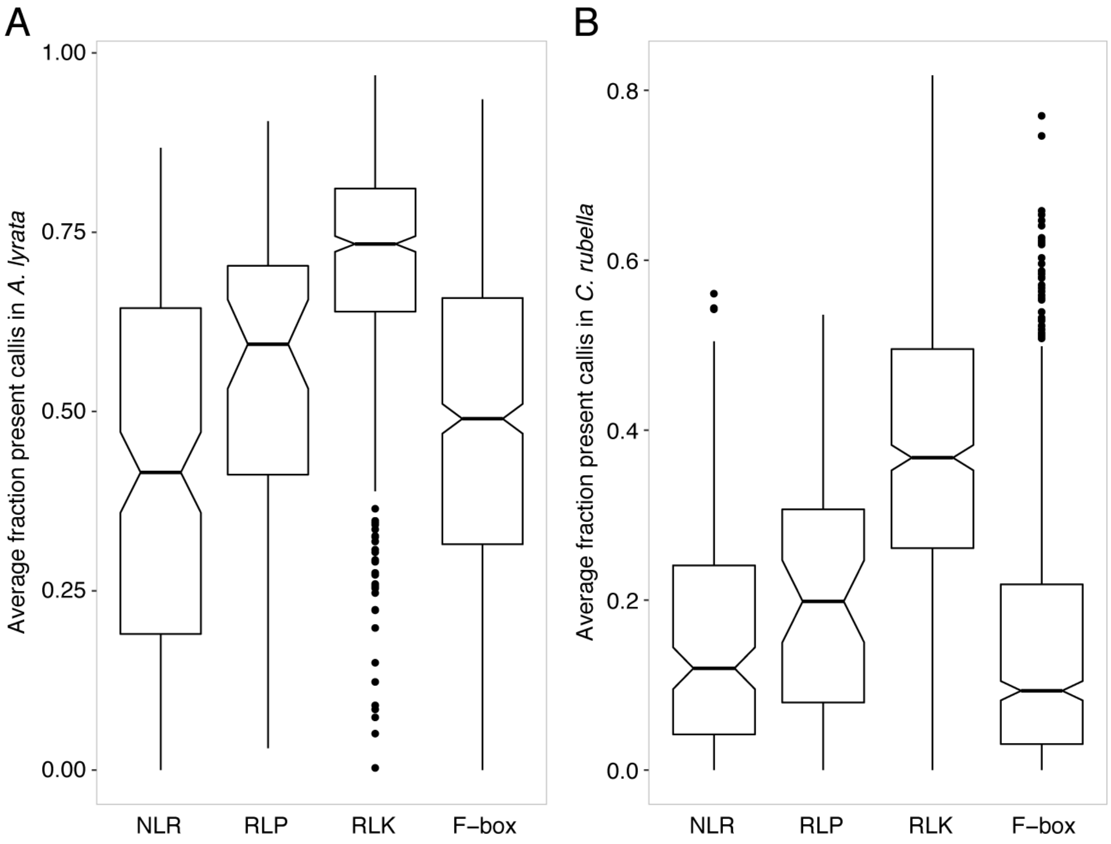

Figure 28. Conservation values of four families of genes in *A. lyrata* (A) and *C. rubella* (B), averaged over accessions.

4.7 NLR genes in *A. lyrata*

I used *A. lyrata* genome assembly version 1.0 (Hu et al., 2011) and annotation version 2.1 (updated by Rawat et al., 2015), which are part of the phytozyme release 2.1, downloaded from https://genome.jgi.doe.gov. I selected genes annotated as "Disease resistance protein" in the release, all of which were also annotated with NLR architectures, for a total of 198 NLR genes. All three categories of genes were represented (Figure 29).

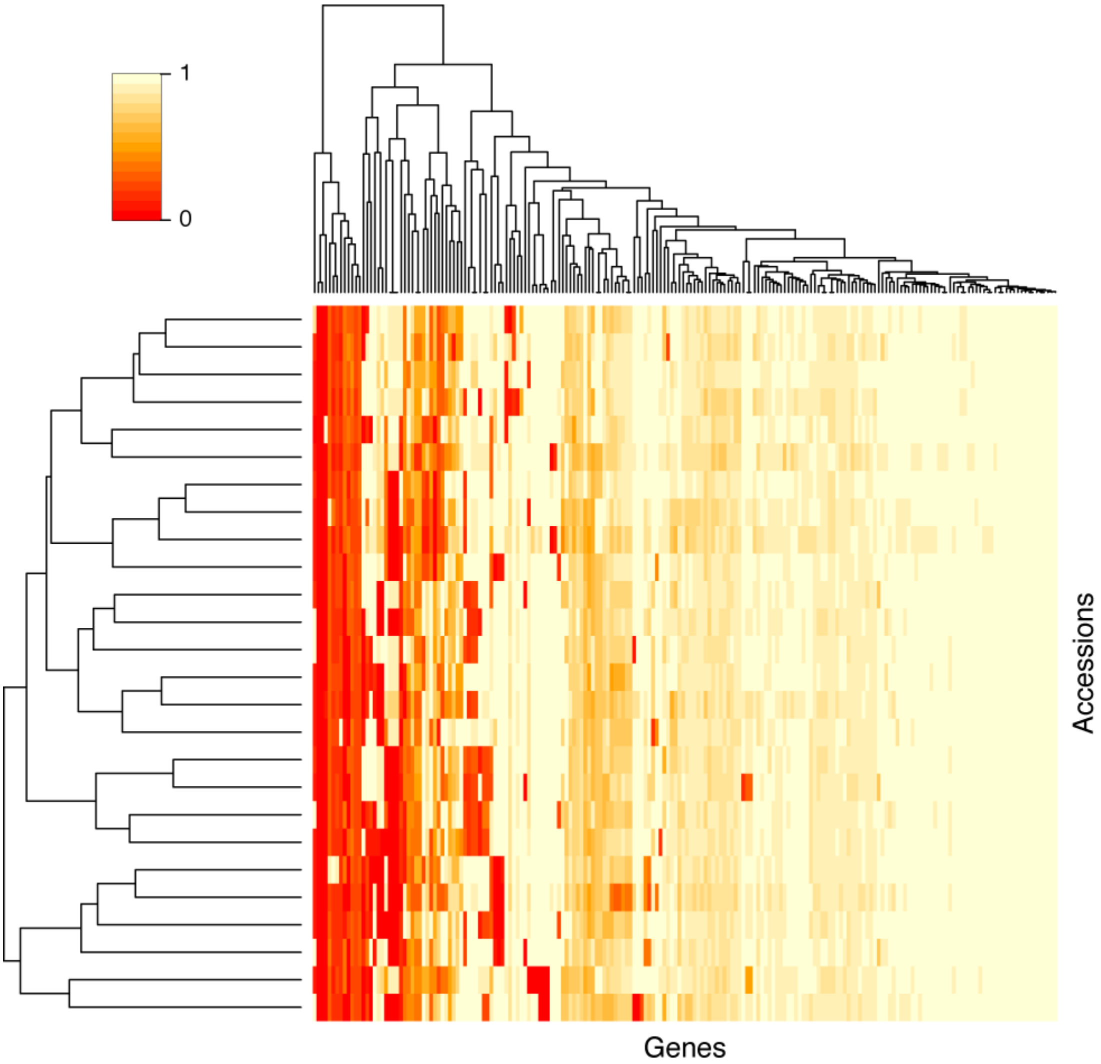

Figure 29. Heatmap representation of the presence of coverage in 198 NLR genes in accessions of *A. lyrata*.

4.8 NLR genes in *C. rubella*

I used *C. rubella* genome assembly version 1.0 and annotation version 1.1, which are part of the phytozyme release 1.1, downloaded from https://genome.jgi.doe.gov. NLR genes were selected following the same procedure as for *A. lyrata* (Section 4.7), for a total of 160 NLR genes. One outlier *C. rubella* accession was excluded from the analysis (See Section 7.3.5). Mapping 36 bp segments of the *C. rubella* reference genome back to itself produced no zero coverage values in any of the NLR CDS regions, as was also the case with *A. thaliana* and *A. lyrata* genomes, thus confirming the quality of mapping. The resulting heatmap has all three categories of genes represented (Figure 30).

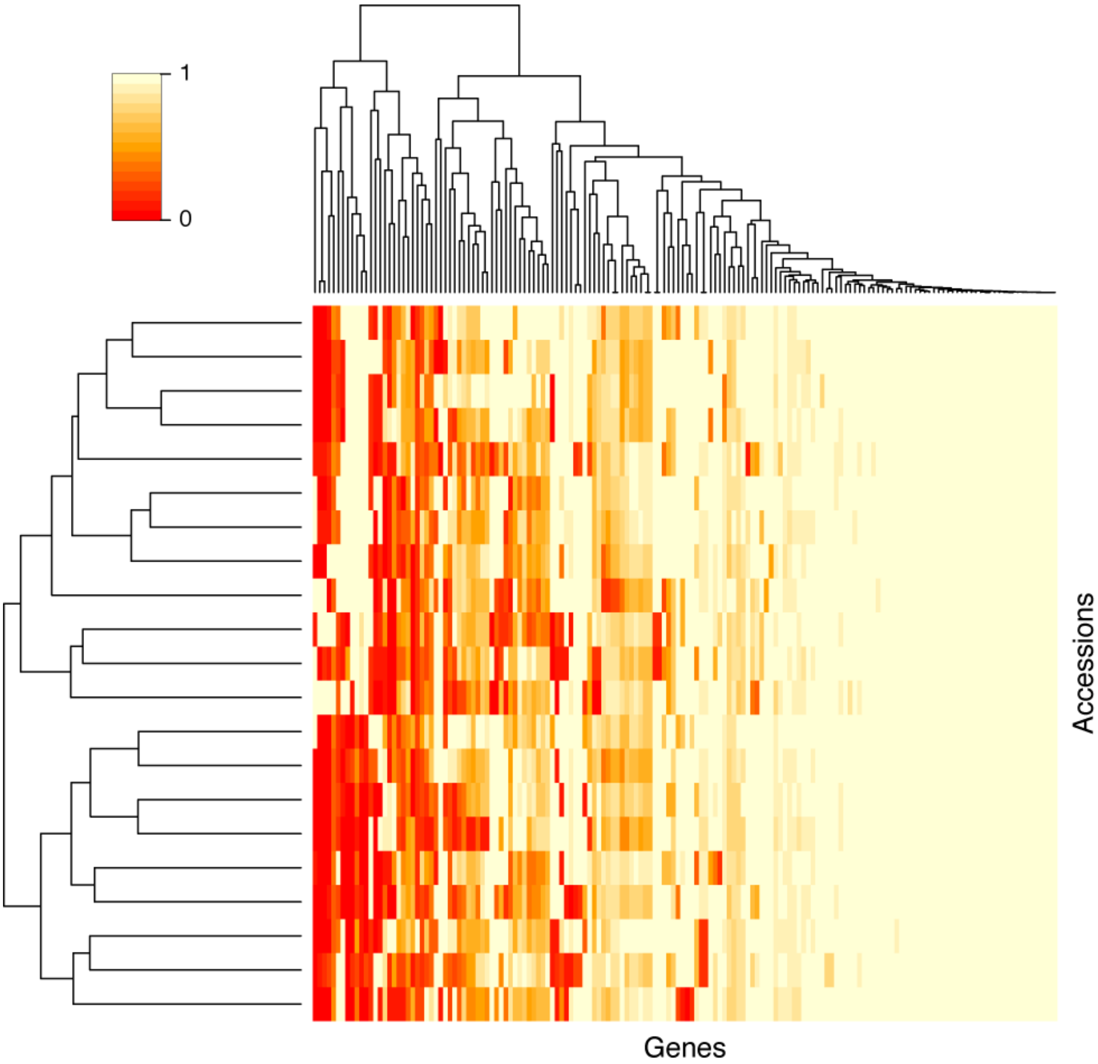

Figure 30. Heatmap representation of the presence of coverage in 160 NLR genes in accessions of *C. rubella*.

Considering that *C. rubella* went through an extreme population bottleneck (Guo et al., 2009), one would expect to see less variation among accessions in *C. rubella* than in *A. thaliana*, visible as consistency between values in the columns of Figure 30. However, considerable variation was present. Similarly, despite accessions of *A. lyrata* considered in this study proceeding from a single geographic location – Norway (Hämälä et al., 2018), considerable variation in NLR genes was apparent (Figure 29). Thus, extreme variability characteristic of NLR genes in *A. thaliana* also applies to its closely related species.

Conservation value distributions for NLR genes in the three species were significantly different (*P* < 2.2e-16; Kruskal-Wallis rank sum test). All three comparisons were significant (*P* = 3.0e-06 for *A. lyrata* to *C.rubella* comparison, and *P* < 6.0e-16 for the other comparisons; Nemenyi post hoc test, Bonferroni corrected). Average conservation values for NLR genes were similar in the three species: 0.78 for *A. thaliana* and 0.80 for both *A. lyrata* and *C. rubella*. The distributions of conservation values were also similar (Figure 31), indicating similar levels of NLR variability in the three species, with slighly more absence apparent in *A. thaliana* (peak near zero), possibly pointing to a greater P/A polymorphim. There was also a heavier tail in the distribution near the presence peak in *A. lyrata*, suggesting multiple genes with small to moderate levels of allelic divergence.

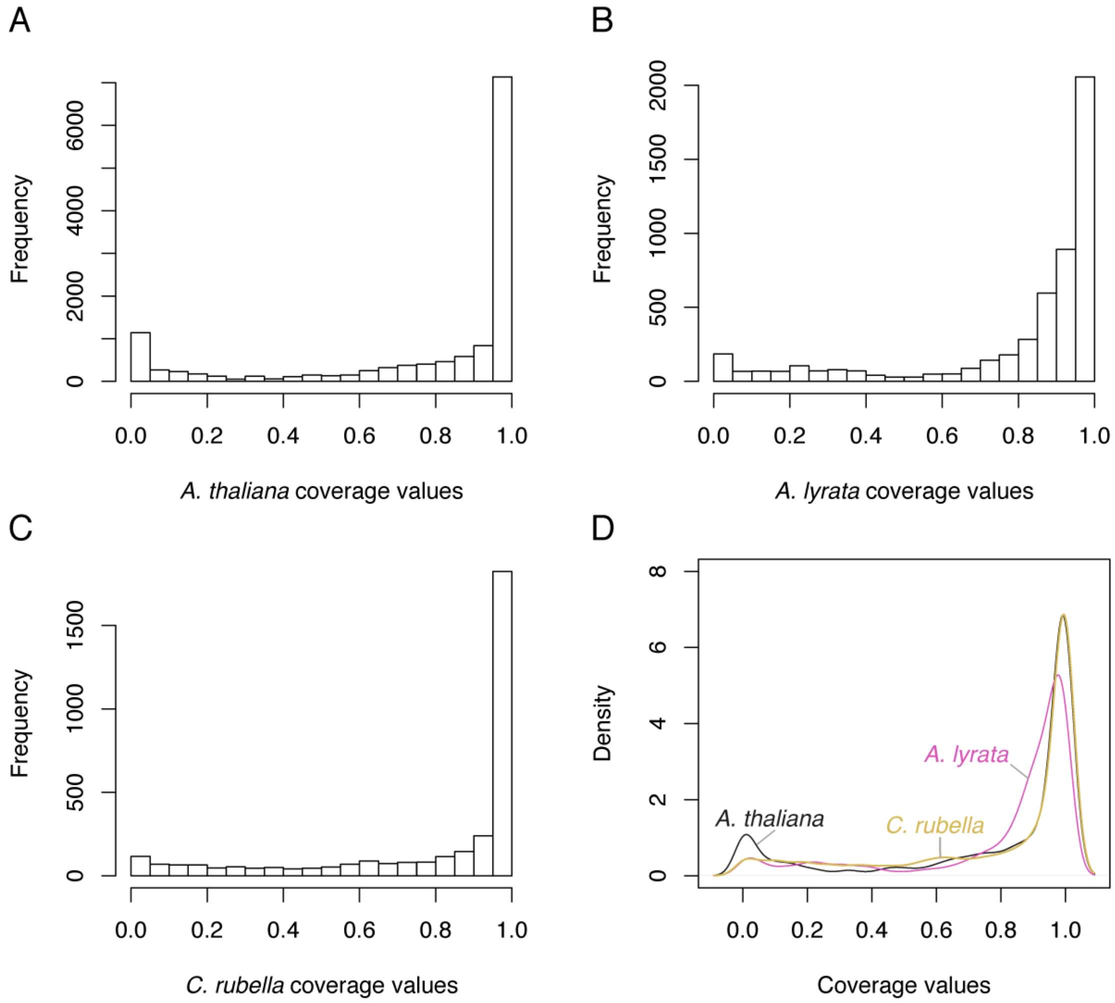

Figure 31. Distribution of coverage values for NLR genes in *A. thaliana* (A), *A. lyrata* (B) and *C. rubella* (C). Coverage values represent the fraction of reference CDS length covered by reads from each accession. Density representation with a bandwidth of 0.03 for all three species (D) summarizes the data.

5 Discussion

5.1 Overview

I start with a brief summary and interpretation of the main findings, followed by a comparison to existing literature, highlighting what is new in this study. I proceed with a critical assessment of the approach and new questions that the research has raised. A conclusion ends the section.

5.2 A Curated List of Presence/Absence Genes

Some P/A genes can be costly for the plant to carry, with fitness costs reaching 10%, as is the case with NLR genes *RPM1* and *RPS5* (Tian et al., 2003, Karasov et al., 2015). This leads to the questions of how many P/A genes are there in the *A. thaliana* genome, and whether all of these are NLR genes. In this study, I obtained an estimate of 4% P/A genes. Out of these, only a small fraction might share the patterns characteristic of *RPM1* and *RPS5*. I provide a way to further profile P/A genes genome-wide based on coverage variation pattern, nucleotide diversity and genomic context.

These estimates of the number of P/A genes are substantially lower than the previous estimate of 9% by Tan and colleagues (2012). The difference can be accounted for by the use of different methods for identifying P/A genes and by using a different definition of a P/A gene. My classification relies on the presence of read coverage and threshold-based clustering, whereas Tan and colleagues used paired end read mapping information. I distinguish between the P/A and Complex gene categories, thus making use of the whole variation pattern in each gene, rather than classifying genes with absence in at least one accession as P/A, which is the definition used by Tan and colleagues.

GO analysis is a standard approach to functional annotation, and I found that it bears out my genome-wide assignment of categories from the functional perspective. Of the the P/A genes identified, there was enrichment for defence response and ADP binding, and an underrepresentation of the cellular process category. This set of assignments is reminiscent of NLR proteins. In comparison, Tan and colleagues (2012) found enrichment in stress response and binding function, as well as in membrane localization and unknown annotation. These two assignments are consistent, except for the membrane localization.

5.3 A New Complex Category of NLR Genes

In this study, I propose to further stratify variable NLR genes into Complex as well as P/A and Conserved, based on visual assessment. Gene Ontology results support the distinction between the P/A and Complex categories from a functional perspective, as enrichment profiles of the two categories differ in several aspects, although they are similar in others. Based on the analysis of whole-genome Complex and P/A gene candidates, both categories show enrichment for ADP binding and defence response – categories characteristic of NLR genes; but there is a high overrepresentation of cell killing function and extracellular localization for Complex genes, which was absent for P/A genes. This description seems to match a role in plant immunity for Complex genes, though a distinct one from the function of NLR genes. This suggests that high degree of variation is not restricted to NLR genes, but is also present among other immune components.

5.4 A List of Conserved NLR Genes

The predominant majority of genes in the *A. thaliana* genome are conserved. However, this is not the case with NLR genes, where high degrees of variation seem to be the norm, with about two-thirds of NLR genes classified as either P/A or Complex, and the Conserved genes thus being a minority. This raises the question of whether their function may be distinguishable from that of other NLR genes.

I provide a list of conserved NLR genes based on natural variation within *A. thaliana*, representing a set of genes which can be used for functional genomics analysis or gene enrichment studies. I also provide a list of conserved genes based on between-species comparisons (discussed in Section 5.5). My results based on conservation in *A. thaliana* align well with previous studies based on between-species conservation. Of the four "gatekeeper" NLR loci present across a wide range of plant species (Hofberger et al., 2014), all four were classified as Conserved in my analysis (three as a part of the NLR set, and one based on whole-genome assignment). I thus confirm that these NLR genes are conserved across multiple time scales and may carry an important function.

In terms of GO annotation, Conserved genes were highly underrepresented in the Unclassified category and overrepresented across a wide range of categories, including organic substance metabolic processes and binding, meaning that the variable categories Complex and P/A have an overrepresentation of Unclassified genes. This suggests that Conserved genes tend to have more easily discoverable functions, making the identified list of NLR conserved genes prime candidates for functional genomics analysis, and a gateway to understanding NLR function. It is also possible that Complex and P/A NLR gene groups have a higher fraction of non-functional or non-expressed genes, and a part of the variation might be due to relaxed selective constraints.

5.5 Comparison between *A. thaliana*, *A. lyrata* and *C. rubella*

A previous study comparing *A. lyrata* and *A. thaliana* found a clear correlation between within-species and between-species levels of polymorphism in NLR genes (Guo et al., 2011). Based on comparisons of three species, *A. thaliana*, *A. lyrata* and *C. rubella*, I found that conservation between and within species is correlated, and I identified a list of genes that have high conservation scores based on both between-species and within-species comparisons. Complex and P/A genes overall have noticeably lower conservation values than Conserved genes in both *A. lyrata* and *C. rubella*. However, known P/A genes *RPM1* and *RPS5* are clear outliers in this pattern, having one of the highest conservation values in *A. lyrata* of all the NLR genes.

Previous studies identified very low levels of allelic diversity in *C. rubella*, proposing that it went through an extreme population bottleneck, possibly even speciation by a single selfing individual (Guo et al., 2009). Based on my comparison, however, there are still several genes that appear to have P/A-like patterns. A very clear example is *ADR1-L3*.

I identified *ADR1-L3* gene as a candidate for having maintained P/A polymorphism across genera. It shows clearly visible P/A pattern across accessions of the three species, is one of the most conserved P/A genes in *C. rubella*, and shows patterns of polymorphism, in terms of genomic context and nucleotide diversity, which resemble *RPM1* and *RPS5* – two well known NLR P/A genes exhibiting high fitness costs and maintained by balancing selection. The fact that it is not the most conserved P/A gene in *A. lyrata*, however, could indicate its lesser importance in that species.

While previous studies have found no correlation between the intraspecific and interspecific polymorphism in NLR P/A genes based on comparisons between *A. lyrata* and *A. thaliana* (Guo et al., 2011), my results demonstrate that polymorphism in some genes with clear P/A patterns can be conserved across species and genera.

In *A. lyrata*, both *RPM1* and *RPS5* are highly conserved, and do not display an obvious absence typical of P/A-like patterns in any accession. The absence of a P/A pattern in this species could also be due to a small sample size (26 accessions). It remains unknown, however, whether the high fitness costs characteristic of *RPM1* and *RPS5* genes in *A. thaliana* are also present in *A. lyrata*. *RPM1* possibly maintains a P/A pattern in *C. rubella*, while *RPS5* appears absent.

Two further P/A genes, *RPS4* and *RPS6*, were identified as highly conserved in both *A. lyrata* and *C. rubella*. Since both relate to resistance to *P. syringae*, as does *RPS5*, their conservation across species and genera highlights the importance of this pathosystem.

I carried out a comparison of between-species conservation among five gene families. There were significant differences among the gene families in both *A. lyrata* and *C. rubella*. RLK genes appeared as the most conserved, while NLR, RLP and F-box genes showed lower values of conservation. Previous studies based on both between-species and within-species comparisons in *A. thaliana* and *A. lyrata* have shown that NLR genes are more variable than RLP genes (Guo et al., 2011). In our study, this pattern was reproduced for *A. lyrata*, and was also found in *C. rubella*.

Previous studies have shown that the most variable gene family in *A. thaliana*, following NLR genes, was F-box genes (Clark et al., 2007). There have also been follow up studies on variable F-box genes in *A. thaliana,* which stratified F-box genes into subgroups (Common, Lineage-Specific and Pseudogenized) with different evolutionary histories and polymorphism levels (Hua et al., 2011; 2013). I found that NLR genes are clearly more variable than F-box genes based on comparisons among *A. thaliana* accessions, and also based on comparison of

A. thaliana and *A. lyrata*. However, a comparison with *C. rubella* showed no difference between the conservation values in NLR and F-box genes, with the median even slightly lower for F-box genes. This suggests that while NLR genes are the most variable family in *A. thaliana* based on within-species comparisons, F-box genes might be similarly variable, or even more variable, based on between-species comparisons. Comparisons with more distant species are needed to confirm this finding.

RLK genes were the least variable gene family based on both *A. lyrata* and *C. rubella* comparisons. However, RLK genes are a very large family, only a subset of which are involved in plant immunity. Analysis of a subset of RLK genes which both contain LRR domains and have a role as immune receptors would be needed to determine whether this pattern is a general one.

NLR gene sets within *A. lyrata* and *C. rubella* show almost identical levels of divergence to *A. thaliana*. However, their distributions are different and *A. lyrata* appears to be enriched for alleles with low to medium levels of divergence from the reference. *A. lyrata* is an outcrossing species - unlike *A. thaliana* and *C. rubella*, which are self compatible. This pattern might thus be explained by heterozygosity in *A. lyrata* accessions.

5.6 No Significant Difference by Conservation Pattern between TNL and CNL Genes within *A. thaliana*

A long-standing question is whether TNL and CNL genes follow distinct evolutionary trajectories, and whether one of these groups is more variable than the other. TNL genes have shown within species expansion (Yang et al., 2008), a pattern that was not observed in CNL genes. Comparisons between *A. thaliana* and its sister species *A. lyrata* have also shown that copy number variation (duplication/loss) is more common in TNL genes than in CNL genes (Chen et al., 2010). This was interpreted as possibly arising from differences in evolutionary pressures exerted by distinct pathogen sets corresponding to TNL and CNL genes (Chen et al., 2010).

In this study, I looked at both within- and between-species variation in TNL and CNL genes. Based on within-species comparison, I found TNL and CNL genes represented in each of the three categories, and could identify no significant stratification by conservation pattern (Conserved, P/A, Complex) based on comparison between accessions. In fact, the genes were surprisingly uniformly distributed among categories. No significant overrepresentation of highly conserved genes for either type of domain architecture was identified based on comparison among multiple accessions of either *A. lyrata* or *C. rubella*. If anything, CNL genes were slightly overrepresented among the genes conserved in *C. rubella*, which contradicts the previously mentioned expectation of them being more susceptible to copy number variation (Chen et al., 2010), in the sense of deletion.

These results are consistent with Shen and colleagues (2006), who identified seven NLR P/A genes, based on within-species comparison, out of which three and four were CNL and TNL, respectively. Discussing this apparent contradiction with the expectation based on between-species comparisons, Chen and colleagues (2010) suggested that intra- and inter-specific maintaining frequencies differ in TNL and CNL genes. As an alternative explanation, Chen and colleagues (2010) suggested the possibility of biased sampling in the Shen and colleagues' (2006) study. My results, looking at the complete set of NLR genes, would not be subject to the biased sampling issues mentioned, and thus dismiss this explanation.

The original argument for different evolutionary patterns among NLR genes is based on a study of a single sub-family of NLR genes in lettuce called RGC2 (Kuang et al., 2004). The RGC2 family, containing about 20 genes, has been classified as CNL (Christopoulou et al., 2015) and is located at a single locus in lettuce. Thus, this original study already demonstrates that genes within a single family can show widely different patterns of variation, and that domain architecture, genomic location or even sequence homology might not in themselves be determining predictors of evolutionary pattern.

Overall, allelic variation could follow different evolutionary dynamics than ortholog variation, and be determined by other evolutionary pressures and/or mechanisms. Thus, for example, P/A polymorphism between species may be governed by birth-and-death evolution, while P/A polymorphism between accessions may result from balancing selection. These results align well with previous studies which have found that P/A polymorphism based on within-species comparisons has no strong correlation to P/A polymorphism based on between-species comparisons (Guo et al., 2011) and add support to the observation that different evolutionary processes might shape NLR gene diversity over different time scales (Guo et al., 2011).

5.7 Difference Between Single and Clustered Genes

Between-species comparisons of *A. thaliana* with *A. lyrata* have shown that NLR genes present in tandem duplicate clusters are more variable than single copies (Guo et al., 2011; Chen et al., 2010). This would be expected as clustered arrangement offers more opportunity for sequence exchange. I found this pattern to be statistically significant based on interspecific comparison with *C. rubella* and *A. lyrata*, using multiple accessions. Single genes were more frequent than expected in the set of genes conserved in both *A. lyrata* and *C. rubella*, suggesting that such genes are more likely to persist through speciation.

To establish whether within-species comparisons are consistent with these findings, I looked at how clustering corresponds to the three types of variation patterns established in this study. While no significant correlation was observed overall, clustered genes tended to be found more often in the P/A and Complex categories than expected from the overall proportion of clustered genes. These results are consistent with between-species comparisons, and support the importance of clustered arrangement in facilitating NLR variation. However, I also found a number of NLR P/A genes present outside of clusters. This is also consistent with previous studies, as canonical NLR P/A gene *RPM1* is also known to be present as a single gene outside of clusters (summarized in Table 1).

5.8 *RPS5* Presence/Absence Pattern Associated with Several Growth Traits

GWAS analysis has shown that *RPS5* presence and absence in accessions has the highest association with several growth-related phenotypes of any P/A gene tested. The highest-scoring result for *RPS5* makes it the third most significant of all the P/A genes selected from the whole genome, suggesting that known NLR P/A genes such as *RPS5* are indeed unusual in having such strong effects and stand out from other P/A genes. Of the other high-scoring genes, few had clear genomic context and nucleotide diversity patterns characteristic of *RPS5* and *RPM1*. This reinforces the argument that P/A genes with variation patterns reminiscent of *RPS5* and *RPM1* might be rare in the genome, and further suggests that even out of such genes, not all may have the strong phenotypic effects characteristic of *RPS5* and *RPM1*.

These results may thus contribute to explaining how plants can afford to carry multiple P/A genes considering that their fitness costs have been shown to reach 10% in the case of *RPS5* and *RPM1* (Tian et al., 2003, Karasov et al., 2015). This would also be consistent with previous studies, which, using different sources of evidence, such as gene expression and position in the genome, proposed that P/A variation genome-wide is associated with relaxed selective constraints (Bush et al., 2014).

I found no significant or high-scoring association between the number of siliques and P/A polymorphism. Number of siliques is one of the standard fitness measures which was used in a previous study that found high fitness costs of *RPS5* presence (Karasov et al., 2015). I also found no significant or high-ranking association for *RPM1*. Some obvious explanations would be insufficient statistical power or population structure correction, which can obscure associations with traits correlated to population structure.

Surprisingly, our GWAS analysis has shown that carrying *RPS5* gene has a negative effect on several plant growth-related phenotypes (Karasov et al., 2014). A likely explanation for the difference with previous studies, which used mutants with controlled genetic backgrounds (Tian et al., 2003, Karasov et al., 2015), and the current study, which used natural variation in *A. thaliana* populations, might be attributable to the complexity of *RPS5* and *RPM1* interactions with their genomic context. However, for *RPM1*, although association was not significant, gene presence tended to have a positive effect on these high-scoring phenotypes, consistent with previous studies (Tian et al., 2003). NLR genes are known to be involved in complex networks of interactions, which might provide an explanation for these differences. It requires further investigation to say whether certain genomic backgrounds might compensate or even reverse fitness costs associated with *RPS5* and *RPM1* genes.

Another possibility is that the observed differences might in part be accounted for by the tradeoffs between different growth-related traits and fitness traits. Thus, a more complex modeling of the relevant growth and reproductive traits in terms of both, biomass and timing, as well as of the interactions and tradeoffs between the traits, might provide a more complete story needed to fully understand these differences.

In summary, I have observed unexplained complexity in the fitness effects and function of P/A genes in natural populations, thus opening these questions for further investigation. In particular, the pair of genes *RPS5* and *RPM1* potentially can have contrasting effects on growth-related traits in natural accessions, and offer a model system for elucidating these differences and interactions.

5.9 Population Structure

Previous SNP-based studies have shown both worldwide sharing of variation and an evident population structure in *A. thaliana* (Nordborg, 2005). My study has shown that clustering based on NLR conservation patterns shows weaker population structure than a clustering based on whole-genome SNP, and that isolation by distance is less evident when the patterns of conservation are compared. Thus, large-scale polymorphism in NLR genes must be attributable to other forces.

Nordborg and colleagues (2005) also described an excess of rare polymorphism in *A. thaliana*, when compared to expectations from neutral models. Thus, presence of alleles unique to the reference Col-0 accession, which I observed in this study, is not unexpected.

5.10 Immune Receptor Families: RLP, RLK and F-box Analysis

Immune receptor proteins RLP and RLK are known to function as immune receptors and to carry LRR domains, like NLR proteins. Unlike NLR proteins, however, both RLP and RLK are localized in the membrane, whereas NLR proteins are intracellular. F-box is a large family of genes whose protein products are involved in protein ubiquitination as a step in targeting proteins for degradation, and might thus be also involved in plant immune response. Previous studies have found that the most variable genes families, following NLR genes, were F-box and RLK genes (Clark et al., 2007). RLP genes were also found to be variable in between- and within-species comparisons (Guo et al., 2011). I thus expected these families to show signatures of variation.

I found that the distribution of the three categories was significantly different from the background in RLP genes, which show higher than expected number of Complex genes, and lower than expected number of Conserved genes. This enrichment shows that RLP genes have increased variability, as is also the case with NLR genes. This result is in line with previous studies (Guo et al., 2011).

RLK genes also show higher numbers of Complex genes and lower numbers of Conserved genes, but the overall pattern is not significantly different from the expected. These genes belong to a large family of above 600 genes, not all of which are involved with immune recognition, which might have diluted the signal of variability induced by pathogen pressures. Several RLK receptors with LRR domains have been associated with PAMP recognition in plant immunity (reviewed in Nürnberger and Kemmerling, 2006; Böhm et al., 2014). However, our analysis shows similar conservation values for LRR-RLK and RLK genes without LRR domains. As noted by Liu and colleagues (2017), however, protein products of LRR-RLK genes are involved in a wide variety of processes in addition to plant defence. Future work on LRR-RLK genes, further stratifying this group, might reveal additional insights.

In the PTI/ETI paradigm of immunity (see Section 1.3.1), it is considered that the protein products of RLP and RLK genes belong to the initial PTI response to pathogens, and recognize conserved pathogen-associated molecules while NLR gene products respond to the more variable effector molecules secreted by the pathogen in the second stage of immune response (Jones and Dangl, 2006; Yue et al., 2012). However, this view has also been challenged (Thomma et al., 2011). My results align with the view of similarity between NLR and RLP receptors and thus PTI and ETI. The P/A category of genes, however, shows no enrichment in either of the three gene families. This suggests that the high numbers of P/A genes observed in the NLR family are characteristic of that group.

F-box genes, like RLP genes, show greater numbers of Complex genes and fewer Conserved genes than expected from the overall proportion. This is consistent with previous results that showed F-box genes as the most variable family after NLR genes (Clark et al., 2007). This finding was based on major effect changes in the sequence that could disrupt the reading frame of the genes. This difference is also clearly detectable in the large-scale patterns of variation examined in this study, and supports classification of F-box genes as a highly variable family in plants. The lack of P/A enrichment in the F-box family, however, suggests that the pattern of variation in this gene family differs from that in the NLR family, possibly as a result of varied evolutionary pressures and/or mechanisms of allele diversification.

5.11 Interpretation and Limitations

This method is suited for the visualization of large-scale polymorphism, such as deletions or regions of high genomic variability. It is thus complementary to SNP studies and gives a high-level overview where reliable identification of SNP is not possible.

In interpreting the plots, it should be noted that information contained in the presence-of-coverage profiles is of a different kind and scope than the information contained in assembled genomic sequences, and these should not be interpreted in the same way. Consequently, the clustering of coverage profiles may differ from the clustering of DNA sequences in the following ways:

First, in heatmap coverage profiles, information from all genomes under consideration will be included in the clustering, regardless of the degree of variation relative to the reference sequence. Both, complete absence of homologous sequences as well as presence of an identical copy, can be represented – in this case, by a black and a blue line, respectively. The dataset thus obtained consists of a set of matrices with binary data (1 for presence of coverage, 0 for absence of coverage), which can be used, for example, for distance-based classification.

Second, all differences or distances highlighted in the profiles are originally defined relative to the reference. However, since the distances are position-specific, it is also possible to compare non-reference profiles to each other, reflecting whether their divergence from the reference is of a similar or of a different kind. Thus, what is being directly measured is how the complement of reference genes varies across accessions.

Third, absence-of-coverage signal includes a range of genotypes: from ones where the sequences are present, but with more differences than allowed by the chosen edit distance, to being entirely absent. The presence of coverage, in contrast, indicates that sequences identical to the reference up to the edit distance are present in the resequenced genome, albeit not necessarily in the same position as in the reference (see below).

Fourth, this approach allows visualizing whether sequences similar to the reference are present in the genome as a whole, without requiring that they be contiguous. This means that reads aligning to a position in the reference gene might not come from an identical position in the genome of interest, but from elsewhere in the genome where a copy of that sequence exists. The effect of such cross-mapping can be investigated in advance and controlled for by adjusting the read length to the extent of repetitiveness in the reference dataset. However, such ability to detect homologous sequences regardless of whether synteny has been maintained can be of advantage in *A. thaliana* NLR gene clusters in which rearrangements are common, synteny rarely conserved (Chae et al., 2014; Bomblies et al., 2007) and relationships among genes cannot be easily determined. This offers an advantage compared to the methods that rely on primer-based resequencing or alignments of large genomic segments.

5.12 Conclusion

In this study, I addressed the challenge of characterizing whole-genome NLR patterns of variation. My proposed approach avoided many of the difficulties associated with currently widespread analysis pipelines. I identify three categories of NLR genes based on within-species variation patterns, and provide lists of NLR genes for the P/A and Conserved categories.

I found high levels of variability in F-box and RLP genes, albeit with patterns of variability different from NLR genes. My GWAS analysis identified known NLR P/A gene *RPS5* as the third highest-scoring gene among genome-wide P/A candidates. Furthermore, of other high scoring candidates, most did not show patterns of variation reminiscent of the canonical P/A genes under balancing selection, based on combined information from genomic context profiles and nucleotide diversity plots. This reinforces the need to distinguish between genes in which a long-standing P/A polymorphism is maintained and genes that are merely absent in one or more accessions.

In addition, known P/A genes *RPM1* and *RPS5* were among the most highly conserved NLR P/A genes in *A. lyrata*, based on between-species comparison, and very few P/A genes genome-wide shared nucleotide diverisity and genomic context patterns characteristic of these genes. These results cumulatively suggest that genes with polymorphism patterns characteristic of *RPM1* and *RPS5* are rare in the *A. thaliana* genome.

Two further NLR P/A genes, *RPS4* and *RPS6*, were among the most conserved in both *A. lyrata* and *C. rubella*, suggesting their importance. Another NLR P/A gene, *ADR1-L3,* was also among the most highly conserved in *C. rubella* and its genomic context and nucleotide diversity profile shared features characteristic of canonical P/A genes under balancing selection. Furthermore, it displayed a P/A like pattern of variation in both *A. lyrata* and *C. rubella*. All these lines of evidence suggest it as a candidate for having maintained P/A polymorphism across species and genera. In addition, we identified a total of 21 NLR P/A genes that appeared to be under P/A polymorphism in *A. lyrata* and 9 genes that appeared to be under P/A polymorphism in *C. rubella*, suggesting that maintenance of P/A polymorphism across species and genera might not be unusual.

Future studies can address whether these P/A gene candidates are subject to balancing selection and whether they have the same high fitness costs characteristic of other known P/A genes. My GWAS study highlighted *RPS5* as a promising candidate for further fitness effect studies in natural populations. I found that NLR genes with both TNL and CNL domain architectures were represented in the list of highly conserved genes, and in all NLR gene categories, without significant enrichment. However, clustered genes were found to be more variable than single genes both based on within- and between-species comparisons, and this effect was statistically significant in between-species comparison with *C. rubella*.

Of gene families known to be variable, I found NLR and F-box genes to be the most variable gene families based on between-species comparisons, followed by RLP genes. RLK genes were the least variable. However, while NLR genes were more variable than F-box genes based on comparison with *A. lyrata*, comparison with the more distantly related *C. rubella* revealed much less difference between the two, suggesting that variability might change with evolutionary distance.

Finally, mapping *A. lyrata* and *C. rubella* reads to their cognate reference genomes revealed that NLR genes in these species have similar variability levels to *A. thaliana*, however, the distribution of variation differs between the species.

While *A. thaliana* seems to have an increased proportion of absence alleles, *A. lyrata* seems to be enriched for NLR genes with small to moderate levels of divergence. My approach can be applied to any genomic region in individuals of the same species or closely related species.

6 References

1001 Genomes Consortium (2016) 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell 166(2): 481-491.

Aarts, N., Metz, M., Holub, E., Staskawicz, B. J., Daniels, M. J., Parker, J. E. (1998) Different requirements for EDS1 and NDR1 by disease resistance genes define at least two R gene-mediated signaling pathways in Arabidopsis. PNAS 95(17): 10306–10311.

Ågren, J. A., Wang, W., Koenig, D., Neuffer, B., Weigel, D., Wright, S. I. (2014) Mating system shifts and transposable element evolution in the plant genus Capsella. BMC Genomics, 15(1): 602.

Alcázar, R., García, A. V., Parker, J. E., and Reymond, M. (2009) Incremental steps toward incompatibility revealed by Arabidopsis epistatic interactions modulating salicylic acid pathway activation. PNAS 106(1): 334–339.

Alcázar, R., García, A. V., Kronholm, I., de Meaux, J., Koornneef, M., Parker, J. E., and Reymond, M. (2010) Natural variation at Strubbelig Receptor Kinase 3 drives immune-triggered incompatibilities between Arabidopsis thaliana accessions. Nature Genetics 42(12): 1135–1139.

Alcázar, R., von Reth, M., Bautor, J., Chae, E., Weigel, D., Koornneef, M., Parker, J. E. (2014) Analysis of a plant complex resistance gene locus underlying immune-related hybrid incompatibility and its occurrence in nature. PLOS Genetics 10(12).

Ameline-Torregrosa, C., Wang, B. B., O'Bleness, M. S., Deshpande, S., Zhu, H., Roe, B., Young, N. D., Cannon, S. B. (2008) Identification and characterization of nucleotide- binding site-leucine-rich repeat genes in the model plant Medicago truncatula. Plant Physiology 146(1): 5–21.

Andolfo, G., Jupe, F., Witek, K., Etherington, G. J., Ercolano, M. R., Jones, J. D. G. (2014) Defining the full tomato NB-LRR resistance gene repertoire using genomic and cDNA RenSeq. BMC Plant Biology 14: 120.

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408(6814): 796–815.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., Sherlock, G. (2000) Gene ontology: tool for the unification of biology. Nature Genetics 25(1): 25-29.

Ausubel, F. M. (2005) Are innate immune signaling pathways in plants and animals conserved? Nature Immunology 6: 973–979.

Baggs, E., Dagdas, G., Krasileva, K. V. (2017) NLR diversity, helpers and integrated domains: making sense of the NLR IDentity. Current Opinion in Plant Biology 38: 59-67.

Bailey, P. C., Schudoma, C., Jackson, W., Baggs, E., Dagdas, G., Haerty, W., Moscou, M., Krasileva, K. V. (2018) Dominant integration locus drives continuous diversification of plant immune receptors with exogenous domain fusions. Genome Biology 19(1): 23.

Bakker, E. G., Toomajian, C., Kreitman, M., Bergelson, J. (2006) A genome-wide survey of R gene polymorphisms in Arabidopsis. Plant Cell 18(8): 1803-1818.

Bakker, E. G., Traw, M. B., Toomajian, C., Kreitman, M., Bergelson, J. (2008) Low levels of polymorphism in genes that control the activation of defense response in Arabidopsis thaliana. Genetics 178(4): 2031-2043.

Basantani, M. K., Gupta, D., Mehrotra, R., Mehrotra, S., Vaish, S., Singh, A. (2017) An update on bioinformatics resources for plant genomics research. Current Plant Biology 11-12: 33-40.

Baumgarten, A., Cannon, S., Spangler, R., May, G. (2003) Genome-level evolution of resistance genes in Arabidopsis thaliana. Genetics 165: 309-319.

Beilstein, M. A., Nagalingum, N. S., Clements, M. D., Manchester, S. R., Mathews, S. (2010). Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. PNAS 107(43): 18724–18728.

Bent, A. F. 1996. Plant disease resistance genes: Function meets structure. Plant Cell 8: 1757–1771.

Bergelson, J., Kreitman, M., Stahl, E., Tian, D. (2001) Evolutionary dynamics of plant R-genes. Science 292: 2281–2285.

Bi, D., Johnson, K. C. M., Zhu, Z., Huang, Y., Chen, F., Zhang, Y., Li, X. (2011). Mutations in an Atypical TIR-NB-LRR-LIM Resistance Protein Confer Autoimmunity. Frontiers in Plant Science 2: 71.

Białas, A., Zess, E. K., De la Concepcion, J. C., Franceschetti, M., Pennington, H.G., Yoshida, K., Upson, J. L., Chanclud, E., Wu, C. H., Langner, T., Maqbool, A., Varden, F. A., Derevnina, L., Belhaj, K., Fujisaki, K., Saitoh, H., Terauchi, R., Banfield, M. J., Kamoun, S. (2018) Lessons in Effector and NLR Biology of Plant-Microbe Systems. Molecular Plant-Microbe Interactions 31(1): 34-45.

Bittner-Eddy, P. D., Crute, L. R., Holub, E. B., Beynon, J. L. (2000) RPP13 is a simple locus in Arabidopsis thaliana for alleles that specify downy mildew resistance to different avirulence determinants in Peronospora parasitica. Plant Journal 21: 177–188.

Bittner-Eddy, P. D., Beynon, J. L. (2001) The Arabidopsis downy mildew resistance gene RPP13-Nd, functions independently of NDR1 and EDS1 and does not require the accumulation of salicylic acid. Molecular Plant-Microbe Interactions 14: 416-421.

Böhm, H., Albert, I., Fan, L., Reinhard, A., Nürnberger, T. (2014) Immune receptor complexes at the plant cell surface. Current Opinion in Plant Biology 20: 47-54.

Bomblies, K., Lempe, J., Epple, P., Warthmann, N., Lanz, C., Dangl, J. L., Weigel, D. (2007). Autoimmune Response as a Mechanism for a Dobzhansky-Muller-Type Incompatibility Syndrome in Plants. PLoS Biology 5(9): e236.

Bomblies, K., Weigel, D. (2007) Hybrid necrosis: autoimmunity as a potential gene-flow barrier in plant species. Nature Reviews Genetics 8: 382-393.

Bomblies, K., Weigel, D. (2010) Arabidopsis and relatives as models for the study of genetic and genomic incompatibilities. Philos Trans R Soc Lond B Biol Sci 365: 1815-1823.

Borevitz, J. O., Hazen, S. P., Michael, T. P., Morris, G. P., Baxter, I. R., Hu, T. T., Chen, H., Werner, J. D., Nordborg, M., Salt, D. E., Kay, S. A, Chory, J., Weigel, D., Jones, J. D., Ecker, J. R. (2007) Genome-wide patterns of single-feature polymorphism in Arabidopsis thaliana. PNAS 104(29): 12057–12062.

Borrelli, G. M., Mazzucotelli, E., Marone, D., Crosatti, C., Michelotti, V., Valè, G., Mastrangelo, A. M. (2018) Regulation and Evolution of NLR Genes: A Close

Interconnection for Plant Immunity. International Journal of Molecular Science 19(6).

Botella, M. A., Parker, J. E., Frost, L. N., Bittner-Eddy, P. D., Beynon, J. L., Daniels, M. J., Holub, E. B., Jones, J. D. (1998) Three genes of the Arabidopsis RPP1 complex resistance locus recognize distinct Peronospora parasitica avirulence determinants. Plant Cell 10: 1847–1860.

Bouktila, D., Khalfallah, Y., Habachi-Houimli, Y., Mezghani-Khemakhem, M., Makni, M., Makni, H. (2014) Full-genome identification and characterization of NBS-encoding disease resistance genes in wheat. Molecular Genetics and Genomics 290(1): 257-271.

Brandvain, Y., Slotte, T., Hazzouri, K. M., Wright, S. I., Coop, G. (2013) Genomic identification of founding haplotypes reveals the history of the selfing species Capsella rubella. PLoS Genetics 9(9).

Bush, S. J., Castillo-Morales, A., Tovar-Corona, J. M., Chen, L., Kover, P. X., & Urrutia, A. O. (2014). Presence-absence variation in A. thaliana is primarily associated with genomic signatures consistent with relaxed selective constraints. Molecular Biology and Evolution 31(1): 59-69.

Caicedo, A. L., Schaal, B. A., Kunkel, B. N. (1999) Diversity and molecular evolution of the RPS2 resistance gene in Arabidopsis thaliana. PNAS 96(1): 302–306.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Müller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K. J., Weigel, D. (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nature Genetics 43(10): 957- 963.

Casey, L. W., Lavrencic, P., Bentham, A. R., Cesari, S., Ericsson, D. J., Croll, T., … Williams, S. J. (2016). The CC domain structure from the wheat stem rust resistance protein Sr33 challenges paradigms for dimerization in plant NLR proteins. PNAS 113(45): 12856–12861.

Cesari, S., Thilliez, G., Ribot, C., Chalvon, V., Michel, C., Jauneau, A., Rivas, S., Alaux, L., Kanzaki, H., Okuyama, Y., Morel, J. B., Fournier, E., Tharreau, D., Terauchi, R., Kroj, T. (2013) The rice resistance protein pair RGA4/RGA5 recognizes the Magnaportheoryzae effectors AVR-Pia and AVR1-CO39 by direct binding. Plant Cell 25(4): 1463–81.

Cesari, S., Bernoux, M., Moncuquet, P., Kroj, T., Dodds, P. N. (2014) A novel conserved mechanism for plant NLR protein pairs: The "integrated decoy" hypothesis. Frontiers in Plant Sciences 5: 606.

Chae, E., Bomblies, K., Kim, S. T., Karelina, D., Zaidem, M., Ossowski, S., Martín-Pizarro, C., Laitinen, R. A., Rowan, B. A., Tenenboim, H., Lechner, S., Demar, M., Habring-Müller, A., Lanz, C., Rätsch, G., Weigel, D. (2014) Species-wide genetic incompatibility analysis identifies immune genes as hot spots of deleterious epistasis. Cell 159(6): 1341-1351.

Chen, C., Chen, H., Lin, Y. S., Shen, J. B., Shan, J. X., Qi, P., Shi, M., Zhu, M. Z., Huang, X. H., Feng, Q., Han, B., Jiang, L., Gao, J. P., Lin, H. X. (2014). A two-locus interaction causes interspecific hybrid weakness in rice. Nature Communications 5: 3357.

Chen, F., Dong, W., Zhang, J., Guo, X., Chen, J., Wang, Z., Lin, Z., Tang, H., Zhang, L. (2018) The Sequenced Angiosperm Genomes and Genome Databases. Frontiers in Plant Sciences 9:418.

Chen, Q., Han, Z., Jiang, H., Tian, D., Yang, S. (2010) Strong positive selection drives rapid diversification of R-genes in Arabidopsis relatives. Journal of Molecular Evolution 70: 137–148.

Cheng, S., Melkonian, M., Smith, S., Brockington, S., Archibald, J. M., Delaux, P.-M., Li, F.-W., Melkonian, B., Mavrodiev, E. V., Sun, W., Fu, Y., Yang, H., Soltis, D. E., Graham, S. W., Soltis. P. S., Liu, X., Xu, X., Wong, G. K. (2018) 10KP: A phylodiverse genome sequencing plan. GigaScience 7(3).

Chhangawala, S., Rudy, G., Mason, C. E., Rosenfeld, J. A. (2015). The impact of read length on quantification of differentially expressed genes and splice junction detection. Genome Biology 16(1): 131.

Chini, A., Loake, G. J. (2005) Motifs specific for the ADR1 NBS-LRR protein family in Arabidopsis are conserved among NBS-LRR sequences from both dicotyledonous and monocotyledonous plants. Planta 221(4): 597-601.

Chisholm, S. T., Coaker, G., Day, B., Staskawicz, B. J. (2006) Host-microbe interactions: shaping the evolution of the plant immune response. Cell 124(4): 803–814.

Christie, N., Tobias, P. A., Naidoo, S., Külheim, C. (2016) The Eucalyptus grandis NBS-LRR gene family: Physical clustering and expression hotspots. Frontiers in Plant Science 6: 1238.

Christopoulou, M., Wo, S. R.-C., Kozik, A., McHale, L. K., Truco, M.-J., Wroblewski, T., Michelmore, R. W. (2015). Genome-Wide Architecture of Disease Resistance Genes in Lettuce. G3: Genes|Genomes|Genetics 5(12): 2655–2669.

Clark, R. M., Schweikert, G., Toomajian, C., Ossowski, S., Zeller, G., Shinn, P., Warthmann, N., Hu, T. T., Fu, G., Hinds, D. A., Chen, H., Frazer, K. A., Huson, D. H., Schölkopf, B., Nordborg, M., Rätsch, G., Ecker, J. R., Weigel, D. (2007) Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. Science 317(5836): 338-342.

Collier, S. M., Hamel, L. P., Moffett, P. (2011) Cell death mediated by the N-terminal domains of a unique and highly conserved class of NB-LRR protein. Molecular Plant-Microbe Interactions 24(8): 918-931.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. Bioinformatics 27(15): 2156–2158.

Deslandes, L., Olivier, J., Theulières, F., Hirsch, J., Feng, D. X., Bittner-Eddy, P., Beynon, J., Marco, Y. (2002) Resistance to Ralstonia solanacearum in Arabidopsis thaliana is conferred by the recessive RRS1 R gene, a member of a novel family of resistance genes. PNAS 99(4): 2404–2409.

Ding, J., Araki, H., Wang, Q., Zhang, P., Yang, S., Chen, J. Q., Tian, D. (2007) Highly asymmetric rice genomes. BMC Genomics 8: 154.

Ellis, J. G. (2016) Integrated decoys and effector traps: how to catch a plant pathogen. BMC Biology 14: 13.

Enright, A. J., Iliopoulos, I., Kyripides, N. C., Ouzounis, C. A. (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature 402(6757): 86-90.

Fei, Q., Xia, R., Meyers, B. C., (2013) Phased, secondary, small interfering RNAs in post transcriptional regulatory networks. Plant Cell 25: 2400–2415.

Flor, H. H. (1956). The complementary genic system in flax and flax rust. Advances in Genetics 8: 29–54.

Foxe, J. P., Slotte, T., Stahl, E. A., Neuffer, B., Hurka, H., Wright, S. I. (2009) Recent speciation associated with the evolution of selfing in Capsella. PNAS 106(13): 5241–5245.

Freeling, M. (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annual Review of Plant Biology 60: 433-453.

Gan, X., Stegle, O., Behr, J., Steffen, J. G., Drewe, P., Hildebrand, K. L., Lyngsoe, R., Schultheiss, S. J., Osborne, E. J., Sreedharan, V. T., Kahles, A., Bohnert, R., Jean, G., Derwent, P., Kersey, P., Belfield, E. J., Harberd, N. P., Kemen, E., Toomajian, C., Kover, P. X., Clark, R. M., Rätsch, G., Mott, R. (2011). Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature 477(7365): 419-423.

Gassmann, W., Hinsch, M. E., Staskawicz, B. J. (1999) The Arabidopsis RPS4 bacterial-resistance gene is a member of the TIR-NBS-LRR family of disease-resistance genes. Plant Journal 20: 265–277.

Grant, M. R., J. M. McDowell, A. G. Sharpe, de Torres Zabala, M., Lydiate, D. J., Dangl, J. L. (1998) Independent deletions of a pathogenresistance gene in Brassica and Arabidopsis. PNAS 95: 15843–15848.

The Gene Ontology Consortium. (2017) Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Research, 45: D331–D338.

Guo, Y.-L., Bechsgaard, J. S., Slotte, T., Neuffer, B., Lascoux, M., Weigel, D., Schierup, M. H. (2009). Recent speciation of Capsella rubella from Capsella grandiflora, associated with loss of self-incompatibility and an extreme bottleneck. PNAS 106(13): 5246–5251.

Guo, Y.-L., Fitz. J., Schneeberger, K., Ossowski, S., Cao, J., Weigel, D. (2011) Genome-wide comparison of nucleotide-binding site-leucine-rich repeat-encoding genes in Arabidopsis. Plant Physiology 157: 757-769.

Hämälä, T., Mattila, T. M., Savolainen, O. (2018) Local adaptation and ecological differentiation under selection, migration, and drift in Arabidopsis lyrata. Evolution [Electronically published ahead of print].

Hehl, R. (2017) From experiment-driven database analyses to database-driven experiments in Arabidopsis thaliana transcription factor research. Plant Science 262: 141-147.

Henk, A. D., Warren, R. F., Innes, R. W. (1999) A new Ac-like transposon of Arabidopsis is associated with a deletion of the RPS5 disease resistance gene. Genetics 151: 1581–1589.

Hofberger, J. A., Zhou, B., Tang, H., Jones, J. D., Schranz, M. E. (2014) A novel approach for multi-domain and multi-genefamily identification provides insights into evolutionary dynamics of disease resistance genes in core eudicot plants. BMC Genomics 15: 966.

Holub, E. B. (2001) The arms race in ancient history in Arabidopsis, the wildflower. Nature Reviews Genetics 2: 516-527.

Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J. F., Clark, R. M., Fahlgren, N., Fawcett, J. A., Grimwood, J., Gundlach, H., Haberer, G., Hollister, J. D., Ossowski, S., Ottilar, R. P., Salamov, A. A., Schneeberger, K., Spannagl, M., Wang, X., Yang, L., Nasrallah, M. E., Bergelson, J., Carrington, J. C., Gaut, B. S., Schmutz, J., Mayer, K. F., Van de Peer, Y., Grigoriev, I. V., Nordborg, M., Weigel, D., Guo, Y. L. (2011) The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nature Genetics 43(5): 476-481.

Hua, Z., Zou, C., Shiu, S. H., Vierstra, R. D. (2011) Phylogenetic comparison of F-Box (FBX) gene superfamily within the plant kingdom reveals divergent evolutionary histories indicative of genomic drift. PLoS One 6(1).

Hua, Z., Pool, J. E., Schmitz, R. J., Schultz, M. D., Shiu, S. H., Ecker, J. R., Vierstra, R. D. (2013) Epigenomic programming contributes to the genomic drift evolution of the F-Box protein superfamily in Arabidopsis. PNAS 110(42): 16927-16932.

Hulbert, S. H., Webb, C. A., Smith, S. M., Sun, Q. (2001) Resistance gene complexes: evolution and utilization. Annual Review of Phytopathology 39: 285-312.

Jacob, F., Vernaldi, S., and Maekawa, T. (2013). Evolution and conservation of plant NLR functions. Frontiers in Immunology 4: 297.

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., Tomsho, L. P., Hu, Y., Liang, H., Soltis, P. S., Soltis, D. E., Clifton, S. W., Schlarbaum, S. E., Schuster, S. C., Ma, H., Leebens-Mack, J., dePamphilis, C. W. (2011) Ancestral polyploidy in seed plants and angiosperms. Nature 473(7345): 97-100.

Johnston, J. S., Pepper, A. E., Hall, A. E., Chen, Z. J., Hodnett, G., Drabek, J., Lopez, R., Price, H. J. (2005) Evolution of genome size in Brassicaceae. Annals of Botany 95: 229–235.

Jones, J. D. and Dangl, J. L. (2006) The plant immune system. Nature. 444(7117): 323-329.

Jones, J. D., Vance, R. E., Dangl, J. L. (2016) Intracellular innate immune surveillance devices in plants and animals. Science 354(6316).

Jupe, F., Witek, K., Verweij, W., Sliwka, J., Pritchard, L., Etherington, G. J., Maclean, D., Cock, P. J., Leggett, R. M., Bryan, G. J., Cardle, L., Hein, I., Jones, J. D. G. (2013) Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. The Plant Journal 76: 530–544.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., Sabatti, C., Eskin, E. (2010) Variance component model to account for sample structure in genome-wide association studies. Nature Genetics 42: 348-354.

Kang, Y. J., Kim, K. H., Shim, S., Yoon, M. Y., Sun, S., Kim, M. Y., Van, K., Lee, S.-H. (2012) Genome-wide mapping of NBS-LRR genes and their association with disease resistance in soybean. BMC Plant Biology 12: 139.

Karasov, T. L., Kniskern, J. M., Gao, L., DeYoung, B. J., Ding, J., Dubiella, U., Lastra, R. O., Nallu, S., Roux, F., Innes, R. W., Barrett, L. G., Hudson, R. R., Bergelson, J. (2014) The long-term maintenance of a resistance polymorphism through diffuse interactions. Nature 512(7515): 436-440.

Koch, M. A., Kiefer, M. (2005) Genome evolution among cruciferous plants: a lecture from the comparison of the genetic maps of three diploid species - Capsella rubella, Arabidopsis lyrata subsp. petraea, and A. thaliana. American Journal of Botany 92(4): 761-767.

Koenig, D. and Weigel, D. (2015) Beyond the thale: comparative genomics and genetics of Arabidopsis relatives. Nature Reviews Genetics 16(5): 285-298.

Koornneef, M., Meinke, D. (2010) The development of Arabidopsis as a model plant. Plant Journal 61(6): 909-921.

Krämer, U. (2015) Planting molecular functions in an ecological context with Arabidopsis thaliana. Elife 4.

Kroj, T., Chanclud, E., Michel-Romiti, C., Grand, X.,. Morel, J. B. (2016) Integration of decoy domains derived from protein targets of pathogen effectors into plant immune receptors is widespread. New Phytologist 210(2): 618-626.

Kuang, H., Woo, S. S., Meyers, B. C., Nevo, E., Michelmore, R. W. (2004) Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce. Plant Cell 16(11): 2870-2894.

Leipe, D. D., Koonin, E. V., Aravind, L. (2004) STAND, a class of P-loop NTPases including animal and plant regulators of programmed cell death: multiple, complex domain architectures, unusual phyletic patterns, and evolution by horizontal gene transfer. Journal of Molecular Biology 343: 1–28.

Leister, D. (2004) Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance gene. Trends in Genetics 20: 116-122.

Li, F., Pignatta, D., Bendix, C., Brunkard, J. O., Cohn, M. M., Tung, J., Sun, H., Kumar, P., Baker, B. (2012) Micro RNA regulation of plant innate immune receptors. PNAS 109(5): 1790–1795.

Li, F. W., Harkess, A. (2018) A guide to sequence your favorite plant genomes. Applications in Plant Sciences 6(3).

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.

Li, H., Handsaker, B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25: 2078-2079.

Li, J., Ding, J., Zhang, W., Zhang, Y., Tang, P., Chen, J. Q., Tian, D., Yang, S. (2010) Unique evolutionary pattern of numbers of gramineous NBS–LRR genes. Molecular Genetics and Genomics 283: 427–438.

Li, X., Cheng, Y., Ma, W., Zhao, Y., Jiang, H., Zhang, M. (2010) Identification and characterization of NBS-encoding disease resistance genes in Lotus japonicus. Plant Systematics and Evolution 289: 101–110.

Li, X., Kapos, P., Zhang, Y. (2015) NLRs in plants. Current Opinion in Immunology 32: 114-121.

Liu, P. L., Du, L., Huang, Y., Gao, S. M., Yu, M. (2017) Origin and diversification of leucine-rich repeat receptor-like protein kinase (LRR-RLK) genes in plants. BMC Evolutionary Biology 17(1): 47.

Long, Q., Rabanal, F. A., Meng, D., Huber, C. D., Farlow, A., Platzer, A., Zhang, Q., Vilhjálmsson, B. J., Korte, A., Nizhynska, V., Voronin, V., Korte, P., Sedman, L., Mandáková, T., Lysak, M. A., Seren, Ü., Hellmann, I., Nordborg, M. (2013). Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. Nature Genetics 45: 884-890.

Lysak, M. A., Berr, A., Pecinka, A., Schmidt, R., McBreen, K., Schubert, I. (2006) Mechanisms of chromosome number reduction in Arabidopsis thaliana and related Brassicaceae species. PNAS 103: 5224–5229.

Lv, Q., Lan, Y., Shi, Y., Wang, H., Pan, X., Li, P., Shi, T. (2017) AtPID: a genome-scale resource for genotype-phenotype associations in Arabidopsis. Nucleic Acids Research 45(D1): D1060-D1063.

Maekawa, T., Kufer, T. A., Schulze-Lefert, P. (2011) NLR functions in plant and animal immune systems: so far and yet so close. Nature Immunology 12(9): 817-826.

Manzanares, C., Barth, S., Thorogood, D., Byrne, S. L., Yates, S., Czaban, A., Asp, T., Yang, B., Studer, B. (2016) A Gene Encoding a DUF247 Domain Protein Cosegregates with the S Self-Incompatibility Locus in Perennial Ryegrass. Molecular Biology and Evolution 33(4): 870-884.

Marcotte, E. M., Pellegrini, M., Ng, H.-L., Rice, D. W., Yeates, T. O., Eisenberg, D. (1999) Detecting protein function and protein-protein interactions from genome sequences. Science 285: 751–753.

Mauricio, R., E. Stahl, T. Korves, D. Tian, M. Kreitman, Bergelson, J. (2003) Natural selection for polymorphism in the disease resistance gene RPS2 of Arabidopsis. Genetics 163: 735–746.

McDowell, J. M., Dhandaydham, M., Long, T. A., Aarts, M. G., Goff, S., Holub, E. B., Dangl, J. L. (1998). Intragenic recombination and diversifying selection contribute to the evolution of downy mildew resistance at the RPP8 locus of Arabidopsis. The Plant Cell 10(11): 1861–1874.

McDowell, J. M., Simon, S. A. (2006) Recent insights into R gene evolution. Molecular Plant Pathology 7(5): 437-448.

McHale, L., Tan, X., Koehl, P., Michelmore, R. W. (2006). Plant NBS-LRR proteins: adaptable guards. Genome Biology 7(4): 212.

Metzker, M. L. (2010) Sequencing technologies - the next generation. Nature Reviews Genetics 11(1): 31-46.

Meunier, E., Broz, P. (2017) Evolutionary Convergence and Divergence in NLR Function and Structure. Trends in Immunology 38(10): 744-757.

Meyers, B. C., Shen, K. A., Rohani, P., Gaut, B. S., Michelmore, R. W. (1998). Receptor-like genes in the major resistance locus of lettuce are subject to divergent selection. The Plant Cell, 10(11): 1833–1846.

Meyers, B. C., Dickerman, A. W., Michelmore, R. W., Sivaramakrishnan, S., Sobral, B. W., Young, N. D (1999) Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. Plant Journal 20: 317-332.

Meyes, B. C., Morgante, M., Michelmore, R. W. (2002) TIR-X and TIR-NBS proteins: two new families related to disease resistance TIR-NBS-LRR proteins encoded in Arabidopsis and other plant genomes. Plant Journal 32(1): 77-92.

Meyers, B. C., Kozik, A., Griego, A., Kuang, H., Michelmore, R. W. (2003) Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. Plant Cell 15(4): 809-834.

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., Thomas, P. D. (2017). PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Research 45: D183–D189.

Michelmore, R. W., Meyers, B. C. (1998) Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Research 8: 1113-1130.

Mondragón-Palomino M (2002) Patterns of positive selection in the complete NBS-LRR gene family of Arabidopsis thaliana. Genome Research 12(9): 1305-1315.

Mondragón-Palomino, M., Gaut, B. S. (2005). Gene conversion and the evolution of three leucine-rich repeat gene families in Arabidopsis thaliana. Molecular Biology and Evolution 22(12): 2444-2456.

Monteiro, F., Nishimura, M. T. (2018) Structural, Functional, and Genomic Diversity of Plant NLR proteins: An Evolved Resource for Rational Engineering of Plant Immunity. Annual Review of Phytpathology 56: 12.1-12.25.

Mukhtar, M. S., Carvunis, A. R., Dreze, M., Epple, P., Steinbrenner, J., Moore, J., Tasan, M., Galli, M., Hao, T., Nishimura, M. T., Pevzner, S. J., Donovan, S. E., Ghamsari, L., Santhanam, B., Romero, V., Poulin, M. M., Gebreab, F., Gutierrez, B. J., Tam, S., Monachello, D., Boxem, M., Harbort, C. J., McDonald, N., Gai, L., Chen, H., He, Y., European Union Effectoromics Consortium, Vandenhaute, J., Roth, F. P., Hill, D. E., Ecker, J. R., Vidal, M., Beynon, J., Braun, P., Dangl, J. L. (2011) Independently evolved virulence effectors converge onto hubs in a plant immune system network. Science 333(6042): 596-601.

Narusaka, M., Shirasu, K., Noutoshi, Y., Kubo, Y., Shiraishi, T., Iwabuchi, M., Narusaka, Y. (2009) RRS1 and RPS4 provide a dual Resistance-gene system against fungal and bacterial pathogens. Plant Journal 60(2): 218-226.

Nishimura, M. T., Dangl, J. L. (2010) Arabidopsis and the plant immune system. Plant Journal 61(6): 1053-1066.

Noël, L., T. L. Moores, E. A. van der Biezen, M. Parniske, M. J. Daniels, Parker, J. E., Jones, J. D. (1999) Pronounced intraspecific haplotype divergence at the RPP5 complex disease resistance locus of Arabidopsis. Plant Cell 11: 2099–2111.

Nordborg, M., Hu, T. T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R., Jakobsson, M., Kim, S., Morozov, Y., Padhukasahasram, B., Plagnol, V., Rosenberg, N. A, Shah, C., Wall, J. D., Wang, J., Zhao, K., Kalbfleisch, T., Schulz, V., Kreitman, M., Bergelson, J. (2005). The pattern of polymorphism in Arabidopsis thaliana. PLoS Biology 3(7): 1289-1299.

Nürnberger, T., Brunner, F. (2002) Innate immunity in plants and animals: emerging parallels between the recognition of general elicitors and pathogen associated molecular patterns. Current Opinion in Plant Biology 5: 318–324.

Nürnberger, T., Kemmerling, B. (2006) Receptor protein kinases - pattern recognition receptors in plant immunity. Trends in Plant Science 11(11): 519-22.

Ossowski, S., Schneeberger, K., Clark, R. M., Lanz, C., Warthmann, N., and Weigel, D. (2008). Sequencing of natural strains of Arabidopsis thaliana with short reads. Genome Research 18: 2024-2033.

Oyama, R. K., Clauss, M. J., Formanová, N., Kroymann, J., Schmid, K. J., Vogel, H., Weniger, K., Windsor, A. J., Mitchell-Olds, T. (2008) The shrunken genome of Arabidopsis thaliana. Plant Systematics and Evolution 273: 257–271.

Paetsch, M., Mayland-Quellhorst, S., Neuffer B (2006) Evolution of the self-incompatibility system in the Brassicaceae: Identification of S-locus receptor kinase (SRK) in self-incompatible Capsella grandiflora. Heredity 97(4): 283–290.

Paetsch, M., Mayland-Quellhorst, S., Hurka, H., Neuffer B (2010) Evolution of the mating system in the genus Capsella (Brassicaceae). Evolution in Action. Editor: Glaubrecht, M. Springer, pp. 77–100.

Petersen, A., Spratt, J., Tintle NL. (2013) Incorporating prior knowledge to increase the power of genome-wide association studies. Methods in Molecular Biology 1019: 519-541.

Plantegenet, S., Weber, J., Goldstein, D. R., Zeller, G., Nussbaumer, C., Thomas, J., Weigel, D., Harshman, K., Hardtke, C. S. (2009) Comprehensive analysis of Arabidopsis expression level polymorphisms with simple inheritance. Molecular Systems Biology 5: 242.

Provart, N. J., Alonso, J., Assmann, S. M., Bergmann, D., Brady, S. M., Brkljacic, J., Browse, J., Chapple, C., Colot, V., Cutler, S., Dangl, J., Ehrhardt, D., Friesner, J. D., Frommer, W. B., Grotewold, E., Meyerowitz, E., Nemhauser, J., Nordborg, M., Pikaard, C., Shanklin, J., Somerville, C., Stitt, M., Torii, K. U., Waese, J., Wagner, D., McCourt, P. (2016) 50 years of Arabidopsis research: highlights and future directions. New Phytologist 209(3): 921-944.

Rairdan, G., Moffett, P. (2007) Brothers in arms? Common and contrasting themes in pathogen perception by plant NB-LRR and animal NACHT-LRR proteins. Microbes and Infection 9(5): 677-686.

Rawat, V., Abdelsamad, A., Pietzenuk, B., Seymour, D. K., Koenig, D., Weigel, D., Pecinka, A., Schneeberger, K., (2015) Improving the Annotation of Arabidopsis lyrata Using RNA-Seq Data. PloS ONE 10(9).

Richard, M. M. S., Gratias, A., Thareau, V., Kim, K. D., Balzergue, S., Joets, J., Jackson, S., Geffroy, V. (2018) Genomic and epigenomic immunity in common bean: The unusual features of NB-LRR gene family. DNA Research 25: 161–172.

Rieseberg, L. H., Blackman, B. K. (2010) Speciation genes in plants. Annals of Botany 106(3): 439-455.

Rose, L. E., Bittner-Eddy, P. D., Langley, C. H., Charles, H., Holub, E. B., Michelmore, R. W., Beynon, J. L. (2004) Maintenance of extreme amino acid diversity at the disease resistance gene, RPP13, in Arabidopsis thaliana. Genetics 166: 1517–1527.

Schneeberger, K., Hagmann, J., Ossowski, S., Warthmann, N., Gesing, S., Kohlbacher, O., Weigel, D. (2009) Simultaneous alignment of short reads against multiple genomes. Genome Biology 10(9): 17.

Seren, Ü., Grimm, D., Fitz, J., Weigel, D., Nordborg, M., Borgwardt, K., Korte, A. (2017) AraPheno: a public database for Arabidopsis thaliana phenotypes. Nucleic Acids Research 45(D1): D1054-D1059.

Seymour, D. K., Koenig, D., Hagmann, J., Becker, C., Weigel, D. (2014). Evolution of DNA Methylation Patterns in the Brassicaceae is Driven by Differences in Genome Organization. PLoS Genetics 10(11).

Shen, J., Araki, H., Chen, L., Chen, J. Q., Tian, D. (2006) Unique Evolutionary Mechanism in R-Genes Under Presence/Absence Polymorphism in Arabidopsis thaliana. Genetics 172: 1243-1250.

Shi, J., Zhang, M., Zhai, W., Meng, J., Gao, H., Zhang, W., Han, R., Qi, F. (2018) Genome-wide analysis of nucleotide binding site-leucine-rich repeats (NBS-LRR) disease resistance genes in Gossypium hirsutum. Physiological and Molecular Plant Pathology 104: 1-8.

Shivaprasad, P. V., Chen, H.-M., Patel, K., Bond, D. M., Santos, B. A. C. M., Baulcombe, D. C. (2012) A microRNA superfamily regulates nucleotide binding site-leucine-rich repeats and other mRNAs. Plant Cell 24: 859–874.

Shiu, S. H., and Bleecker, A. B. (2001). Receptor-like kinases from Arabidopsis form a monophyletic gene family related to animal receptor kinases. PNAS 98(19): 10763–10768.

Shiu, S.-H., Karlowski, W. M., Pan, R., Tzeng, Y.-H., Mayer, K. F. X., Li, W.-H. (2004). Comparative Analysis of the Receptor-Like Kinase Family in Arabidopsis and Rice. The Plant Cell 16(5): 1220–1234.

Slotte, T., Hazzouri, K. M., Ågren, J. A., Koenig, D., Maumus, F., Guo, Y. L., Steige, K., Platts, A. E., Escobar, J. S., Newman, L. K., Wang, W., Mandáková, T., Vello, E., Smith, L. M., Henz, S. R., Steffen, J., Takuno, S., Brandvain, Y., Coop, G., Andolfatto, P., Hu, T. T., Blanchette, M., Clark, R. M., Quesneville, H., Nordborg, M., Gaut, B. S., Lysak, M. A., Jenkins, J., Grimwood, J., Chapman, J., Prochnik, S., Shu, S., Rokhsar, D., Schmutz, J., Weigel, D., Wright, S. I. (2013) The Capsella rubella genome and the genomic consequences of rapid mating system evolution. Nature Genetics 45(7): 831–835.

Staal, J., Dixelius, C. 2007. Tracing the ancient origins of plant innate immunity. Trends in Plant Science 12: 334–342.

Stahl, E. A., Dwyer, G., Mauricio, R. Kreitman, M., Bergelson, J. (1999) Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. Nature 400(6745): 667-671.

Tan, S., Wu, S. (2012) Genome wide analysis of nucleotide-binding site disease resistance genes in Brachypodium distachyon. Comparative and Functional Genomics 2012: 418208.

Tan, S., Zhong, Y., Hou, H., Yang, S., Tian, D. (2012) Variation of presence/absence genes among Arabidopsis populations. BMC Evolutionary Biology 12:86.

Tarr, D. E., Alexander, H. M. (2009) TIR-NBS-LRR genes are rare in monocots: evidence from diverse monocot orders. BMC Research Notes 2: 197.

Thomma, B. P., Nürnberger, T., Joosten, M. H. (2011) Of PAMPs and effectors: the blurred PTI-ETI dichotomy. Plant Cell 23(1): 4-15.

Tian, D., Araki, H., Stahl, E., Bergelson, J., Kreitman, M. (2002) Signature of balancing selection in Arabidopsis. PNAS 99(17): 11525–11530.

Tian, D., Traw, M. B., Chen, J. Q., Kreitman, M., Bergelson, J. (2003) Fitness costs of R-gene-mediated resistance in Arabidopsis thaliana. Nature 423(6935): 74-7.

Togninalli, M., Seren, Ü., Meng, D., Fitz, J., Nordborg, M., Weigel, D., Borgwardt, K., Korte, A., Grimm, D. G. (2018) The AraGWAS Catalog: a curated and standardized Arabidopsis thaliana GWAS catalog. Nucleic Acids Research 46(D1): D1150-D1156.

Uehling, J., Deveau, A., Paoletti, M. (2017) Do fungi have an innate immune response? An NLR-based comparison to plant and animal immune systems. PLoS Pathog. 13(10).

Urbach, J. M., Ausubel, F. M. (2017) The NBS-LRR architectures of plant R-proteins and metazoan NLRs evolved in independent events. PNAS 114(5): 1063-1068.

Vasseur, F., Bresson, J., Wang, G., Schwab, R., & Weigel, D. (2018a). Image-based methods for phenotyping growth dynamics and fitness components in Arabidopsis thaliana. Plant Methods 14(1): 63.

Vasseur, F., Exposito-Alonso, M., Ayala-Garay, O. J., Wang, G., Enquist, B. J., Vile, D., Violle, C., Weigel, D. (2018b). Adaptive diversification of growth allometry in the plant Arabidopsis thaliana. PNAS 115(13): 3416–3421.

Vasseur, F., Exposito-Alonso, M., Ayala-Garay, O. J., Wang, G., Enquist, B. J., Vile, D., Violle, C., Weigel, D. (2018c) Data from: Adaptive diversification of growth allometry in the plant Arabidopsis thaliana. Dryad Digital Repository. https://doi.org/10.5061/dryad.343bd84

Wang, G., Ellendorff, U., Kemp, B., Mansfield, J. W., Forsyth, A., Mitchell, K., Bastas, K., Liu, C. M., Woods-Tör, A., Zipfel, C., de Wit, P. J., Jones, J. D., Tör, M., Thomma, B. P. (2008) A genome-wide functional investigation into the roles of receptor-like proteins in Arabidopsis. Plant Physiology 147: 503-517.

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., Li, M., Zheng, T., Fuentes, R. R., Zhang, F., Mansueto, L., Copetti, D., Sanciangco, M., Palis, K. C., Xu, J., Sun, C., Fu, B., Zhang, H., Gao, Y., Zhao, X., Shen, F., Cui, X., Yu, H., Li, Z., Chen, M., Detras, J., Zhou, Y., Zhang, X., Zhao, Y., Kudrna, D., Wang, C., Li, R., Jia, B., Lu, J., He, X., Dong, Z., Xu, J., Li, Y., Wang, M., Shi, J., Li, J., Zhang, D., Lee, S., Hu, W., Poliakov, A., Dubchak, I., Ulat, V. J., Borja, F. N., Mendoza, J. R., Ali, J., Li, J., Gao, Q., Niu, Y., Yue, Z., Naredo, M. E. B., Talag, J., Wang, X., Li, J., Fang, X., Yin, Y., Glaszmann, J. C., Zhang, J., Li, J., Hamilton, R. S., Wing, R. A., Ruan, J., Zhang, G., Wei, C., Alexandrov, N., McNally, K. L., Li, Z., Leung, H. (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557(7703): 43-49.

Warwick, S. I., Francis, A., Al-Shehbaz, I. A. (2006) Brassicaceae: species checklist and database on CD-Rom. Plant Systematics and Evolution 259: 249–258.

Weigel, D. (2012) Natural Variation in Arabidopsis: From Molecular Genetics to Ecological Genomics. Plant Physiology 158: 2–22.

Weigel, D., Nordborg, M., (2015) Population Genomics for Understanding Adaptation in Wild Plant Species. Annual Review of Genetics 49: 315–338.

Weßling, R., Epple, P., Altmann, S., He, Y., Yang, L., Henz, S. R., McDonald, N., Wiley, K., Bader, K. C., Gläßer, C., Mukhtar, M. S., Haigis, S., Ghamsari, L., Stephens, A. E., Ecker, J. R., Vidal, M., Jones, J. D., Mayer, K. F., Ver Loren van Themaat, E., Weigel, D., Schulze-Lefert, P., Dangl, J. L., Panstruga, R., Braun, P. (2014) Convergent targeting of a common host protein-network by pathogen effectors from three kingdoms of life. Cell Host Microbe 16(3): 364-375.

Wicker, T., Yahiaoui, N., Keller, B. (2007) Illegitimate recombination is a major evolutionary mechanism for initiating size variation in plant resistance genes. Plant Journal 51: 631–641.

Williams, S. J., Sohn, K. H., Wan, L., Bernoux, M., Sarris, P. F., Segonzac, C., Ve, T., Ma, Y., Saucet, S. B., Ericsson, D. J., Casey, L. W., Lonhienne, T., Winzor, D. J., Zhang, X., Coerdt, A., Parker, J. E., Dodds, P. N., Kobe, B., Jones, J. D. (2014) Structural basis for assembly and function of a heterodimeric plant immune receptor. Science 344(6181): 299–303.

Williams, S. J., Yin, L., Foley, G., Casey, L. W., Outram, M. A., Ericsson, D. J., Lu, J., Boden, M., Dry, I. B., Kobe, B. (2016). Structure and Function of the TIR Domain from the Grape NLR Protein RPV1. Frontiers in Plant Science 7: 1850.

Woodward, A. W., Bartel, B. (2018) Biology in Bloom: A Primer on the Arabidopsis thaliana Model System. Genetics 208(4): 1337-1349.

Wu, C.-H., Derevnina, L., Kamoun, S. (2018) Receptor networks underpin plant immunity. Science 360(6395): 1300-1301.

Xu, G., Ma, H., Nei, M., Kong, H. Evolution of F-box genes in plants: different modes of sequence divergence and their relationships with functional diversification. PNAS 2009, 106(3): 835-840.

Xu, S., Clark, T., Zheng, H., Vang, S., Li, R., Wong, G. K., Wang, J., Zheng, X. (2008) Gene conversion in the rice genome. BMC Genomics 9: 93.

Xu, X. M., Møller, S. G. (2011) The value of Arabidopsis research in understanding human disease states. Current Opinion in Biotechnology 22(2): 300-307.

Yamamoto, E., Takashi, T., Morinaka, Y., Lin, S., Wu, J., Matsumoto, T., Kitano, H., Matsuoka, M., and Ashikari, M. (2010). Gain of deleterious function causes an autoimmune response and Bateson-Dobzhansky-Muller incompatibility in rice. Molecular Genetics and Genomics 283: 305–315.

Yang, H., Shi, Y., Liu, J., Guo, L., Zhang, X., Yang, S. (2010) A mutant CHS3 protein with TIR-NB-LRR-LIM domains modulates growth, cell death and freezing tolerance in a temperature-dependent manner in Arabidopsis. Plant Journal 63(2): 283-296.

Yang, S., Feng, Z., Zhang, X., Jiang, K., Jin, X., Hang, Y., Chen, J. Q., Tian, D. (2006) Genome-wide investigation on the genetic variations of rice disease resistance genes. Plant Molecular Biology 62: 181–193.

Yang, S., Zhang, X., Yue, J-X., Tian, D., Chen, J-Q. (2008) Recent duplications dominate NBS- encoding gene expansion in two woody species. Molecular Genetics and Genomics 280: 187–198.

Yang, X., Wang, J. (2016) Genome-Wide analysis of NBS-LRR genes in sorghum genome revealed several events contributing to NBS-LRR gene evolution in grass species. Evolutionary Bioinformatics 12: 9–21.

Yogeeswaran, K., Frary, A., York, T. L., Amenta, A., Lesser, A. H., Nasrallah, J. B., Tanksley, S. D., Nasrallah, M. E. (2005) Comparative genome analyses of Arabidopsis spp.: inferring chromosomal rearrangement events in the evolutionary history of A. thaliana. Genome Research 15: 505–515.

Yue, J. X., Meyers, B. C., Chen, J. Q., Tian, D., Yang, S. (2012) Tracing the origin and evolutionary history of plant nucleotide-binding site-leucine-rich repeat (NBSLRR) genes. New Phytologist 193: 1049-1063.

Zeller, G., Clark, R. M., Schneeberger, K., Bohlen, A., Weigel, D., Ratsch, G. (2008) Detecting polymorphic regions in Arabidopsis thaliana with resequencing microarrays. Genome Research 18: 918–929.

Zhang, X., Dodds, P. N., Bernoux, M. (2017) What Do We Know About NOD-Like Receptors in Plant Immunity? Annual Review of Phytopathology 55: 205-229.

Zhang, Y. M., Shao, Z. Q., Wang, Q., Hang, Y. Y., Xue, J. Y., Wang, B., Chen, J. Q. (2016) Uncovering the dynamic evolution of nucleotide-binding site-leucine-rich repeat (NBS-LRR) genes in Brassicaceae. Journal of Integrative Plant Biology 58(2): 165-177.

Zhao, Y., Weng, Q., Song, J., Ma, H., Juan, J., Dong, Z., Liu, Y. (2016) Bioinformatics analysis of NBS-LRR encoding resistance genes in Setaria italica. Biochemical Genetics 54: 232–248.

7 Appendix

7.1 Supplementary Table 1

Conservation statistics of 163 NLR genes in 80 accessions of *A. thaliana*. Average column represents average over 80 accessions of *A. thaliana*, while Minimum and Maximum represent the reange of conservation values in the 80 accessions. Conservation values were calculated as the fraction of total CDS length with non-zero read coverage for any given gene and accession combination.

Gene ID	Category	Average	Minimum	Maximum
AT3G15700	Conserved	1.00	1.00	1.00
AT1G17615	Conserved	1.00	0.99	1.00
AT1G17600	Conserved	1.00	0.99	1.00
AT5G22690	Conserved	1.00	0.99	1.00
AT1G12290	Conserved	1.00	0.99	1.00
AT1G17610	Conserved	1.00	0.95	1.00
AT5G04720	Conserved	1.00	0.98	1.00
AT1G12280	Conserved	1.00	0.98	1.00
AT3G04220	Conserved	1.00	0.99	1.00
AT5G18370	Conserved	1.00	0.99	1.00
AT1G52660	Conserved	1.00	0.95	1.00
AT4G26090	Conserved	1.00	0.97	1.00
AT3G14460	Conserved	1.00	0.85	1.00
AT5G66900	Conserved	1.00	0.96	1.00
AT1G72950	Conserved	1.00	0.96	1.00
AT5G40100	Conserved	1.00	0.93	1.00
AT5G40090	Conserved	0.99	0.95	1.00
AT4G33300	Conserved	0.99	0.98	1.00
AT2G17060	Conserved	0.99	0.79	1.00
AT3G50950	Conserved	0.99	0.97	1.00
AT3G14470	Conserved	0.99	0.88	1.00
AT5G46450	Conserved	0.99	0.97	1.00

AT1G63740	Conserved	0.99	0.50	1.00
AT1G50180	Conserved	0.99	0.96	1.00
AT5G17680	Conserved	0.99	0.78	1.00
AT1G27170	Conserved	0.99	0.77	1.00
AT1G63730	Conserved	0.99	0.63	1.00
AT1G53350	Conserved	0.99	0.88	1.00
AT5G46470	P/A	0.99	0.36	1.00
AT1G10920	P/A	0.99	0.15	1.00
AT1G65850	Conserved	0.99	0.91	1.00
AT1G63750	P/A	0.99	0.31	1.00
AT5G38850	Conserved	0.99	0.93	1.00
AT5G18360	Conserved	0.99	0.79	1.00
AT3G46710	Conserved	0.99	0.94	1.00
AT2G17050	Conserved	0.99	0.78	1.00
AT1G59620	Conserved	0.98	0.56	1.00
AT5G45210	Conserved	0.98	0.94	1.00
AT5G45200	Conserved	0.98	0.95	1.00
AT4G08450	Conserved	0.98	0.92	1.00
AT5G66910	Conserved	0.98	0.95	1.00
AT5G47250	Conserved	0.98	0.64	1.00
AT1G51480	Conserved	0.98	0.89	1.00
AT1G33560	P/A	0.97	0.00	1.00
AT5G63020	Conserved	0.97	0.91	1.00
AT5G11250	P/A	0.97	0.23	1.00
AT3G25510	Conserved	0.97	0.92	1.00
AT1G72890	Conserved	0.97	0.91	1.00
AT5G45050	Conserved	0.97	0.92	1.00
AT5G41550	Conserved	0.96	0.57	1.00
AT5G45250	P/A	0.96	0.01	1.00
AT5G44510	Conserved	0.96	0.79	1.00
AT5G46260	Conserved	0.96	0.91	1.00
AT3G04210	Complex	0.96	0.53	1.00
AT5G38340	P/A	0.96	0.16	1.00
AT5G45060	Conserved	0.95	0.90	1.00
AT2G16870	Complex	0.95	0.73	1.00
AT1G56540	Complex	0.95	0.14	1.00
AT5G45260	P/A	0.95	0.01	1.00

AT1G72900	Complex	0.95	0.68	1.00
AT5G46270	P/A	0.94	0.31	1.00
AT1G12210	Conserved	0.94	0.82	1.00
AT4G14370	Conserved	0.94	0.87	1.00
AT5G58120	Complex	0.93	0.78	1.00
AT1G64070	P/A	0.92	0.15	1.00
AT5G43470	Conserved	0.92	0.80	1.00
AT1G31540	Conserved	0.92	0.79	1.00
AT4G12010	Complex	0.92	0.67	1.00
AT5G41540	P/A	0.92	0.07	1.00
AT1G72940	Complex	0.92	0.70	1.00
AT5G17970	P/A	0.90	0.00	1.00
AT1G58390	Complex	0.90	0.48	1.00
AT3G44480	Complex	0.89	0.39	1.00
AT5G48620	Conserved	0.88	0.77	0.95
AT3G44630	Complex	0.88	0.48	1.00
AT1G57650	P/A	0.88	0.07	1.00
AT4G11170	Complex	0.87	0.56	1.00
AT1G72920	Complex	0.87	0.51	1.00
AT1G61190	Complex	0.86	0.70	1.00
AT3G44400	Complex	0.85	0.32	1.00
AT1G72910	Complex	0.85	0.10	1.00
AT1G72930	Complex	0.85	0.12	1.00
AT3G51570	P/A	0.85	0.00	1.00
AT3G51560	P/A	0.85	0.00	1.00
AT3G44670	Complex	0.83	0.36	1.00
AT4G16990	P/A	0.82	0.00	1.00
AT4G16960	Complex	0.82	0.18	1.00
AT3G07040	P/A	0.82	0.00	1.00
AT1G62630	Complex	0.81	0.49	0.99
AT4G12020	Complex	0.81	0.18	1.00
AT1G61310	Complex	0.81	0.61	1.00
AT5G44870	P/A	0.81	0.01	1.00
AT4G16920	Complex	0.80	0.37	1.00
AT1G61180	Complex	0.79	0.62	1.00
AT4G16950	Complex	0.79	0.28	1.00
AT1G59780	Complex	0.79	0.41	1.00

AT4G36150	Complex	0.78	0.16	1.00
AT5G35450	Complex	0.78	0.25	1.00
AT4G16860	Complex	0.77	0.39	0.94
AT3G46530	Complex	0.76	0.60	1.00
AT4G09430	P/A	0.75	0.00	1.00
AT5G47260	Complex	0.75	0.52	1.00
AT4G09420	P/A	0.75	0.00	1.00
AT5G40910	Complex	0.75	0.16	1.00
AT2G14080	Complex	0.75	0.66	1.00
AT5G43730	Complex	0.74	0.37	1.00
AT1G66090	Complex	0.74	0.50	1.00
AT5G41750	Complex	0.74	0.31	1.00
AT4G09360	P/A	0.74	0.00	1.00
AT1G63360	Complex	0.74	0.32	1.00
AT1G61300	Complex	0.73	0.56	1.00
AT4G16890	Complex	0.72	0.34	1.00
AT5G38350	Complex	0.72	0.51	1.00
AT5G46490	Complex	0.71	0.47	1.00
AT1G27180	Complex	0.71	0.34	1.00
AT1G56510	P/A	0.71	0.05	1.00
AT1G56520	Complex	0.71	0.03	1.00
AT1G59124	Complex	0.70	0.27	0.91
AT1G63880	Complex	0.70	0.28	1.00
AT4G16900	Complex	0.68	0.13	1.00
AT5G49140	P/A	0.67	0.00	1.00
AT1G63870	P/A	0.66	0.03	1.00
AT5G51630	Complex	0.66	0.00	1.00
AT5G41740	Complex	0.65	0.27	0.98
AT5G18350	P/A	0.65	0.02	1.00
AT4G16940	Complex	0.65	0.13	1.00
AT1G63860	P/A	0.64	0.03	1.00
AT1G58807	Complex	0.64	0.24	0.92
AT1G69550	Complex	0.63	0.39	0.97
AT5G43740	Complex	0.62	0.34	1.00
AT1G58410	P/A	0.60	0.00	1.00
AT1G12220	P/A	0.60	0.00	1.00
AT1G63350	Complex	0.60	0.03	1.00

AT4G19530	P/A	0.57	0.31	1.00
AT1G58848	Complex	0.57	0.15	0.91
AT1G59218	Complex	0.57	0.19	0.87
AT3G46730	P/A	0.56	0.12	1.00
AT4G36140	P/A	0.55	0.06	1.00
AT1G58602	Complex	0.53	0.13	0.98
AT1G58400	P/A	0.52	0.05	0.99
AT5G46520	Complex	0.49	0.17	1.00
AT5G17880	P/A	0.47	0.00	1.00
AT5G17890	P/A	0.47	0.00	1.00
AT4G10780	P/A	0.38	0.00	1.00
AT4G19520	P/A	0.37	0.00	1.00
AT5G05400	P/A	0.37	0.00	1.00
AT5G46510	Complex	0.36	0.04	1.00
AT4G27220	P/A	0.35	0.00	1.00
AT4G19510	P/A	0.34	0.11	1.00
AT4G27190	P/A	0.34	0.00	1.00
AT5G47280	P/A	0.34	0.00	1.00
AT5G48780	P/A	0.32	0.00	1.00
AT5G48770	P/A	0.32	0.00	1.00
AT4G19500	P/A	0.30	0.07	1.00
AT1G72860	P/A	0.26	0.00	1.00
AT1G72870	P/A	0.25	0.00	1.00
AT1G72840	P/A	0.25	0.00	1.00
AT1G72850	P/A	0.24	0.00	1.00
AT5G45220	Complex	0.20	0.10	1.00
AT5G36930	P/A	0.18	0.02	1.00
AT1G15890	P/A	0.17	0.01	1.00
AT5G45240	P/A	0.10	0.00	1.00
AT5G45230	P/A	0.10	0.00	1.00

7.2 Supplementary Table 2

P/A genes identified as sharing characteristics of *RPM1* and *RPS5* based on visual inspection of genomic context and nucleotide diversity plots. Multiple genes located within the same deletion are shown as a single entry.

Gene ID
AT1G02250
AT1G12220
AT1G16120/30
AT1G27610
AT1G33530
AT1G50520/30
AT1G64260
AT1G71390
AT2G19230
AT3G07040
AT3G16750
AT3G21080
AT3G47580
AT3G47920
AT3G51560/70
AT4G14905
AT4G21260
AT4G23290
AT4G23590
AT4G31710
AT5G01140
AT5G02930
AT5G05400
AT5G11290
AT5G17960/70
AT5G27100
AT5G47280
AT5G48320
AT5G48770/80
AT5G49140

7.3 Supplementary Methods

7.3.1 Illumina Read Mapping

For mapping reads, version 0.7.15-r1140 of BWA-backtrack algorithm (Li and Durbin, 2009) was used with default parameters, with the exception of the following: maximum edit distance in the seed (*k* in bwa aln command) was set to one and maximumu number of alignments to output (*n* in bwa samse command) was set to 10000. Maximal number of mismatches allowed was one, with zero gaps and a total distance of one. Paired end information was discarded.

The output mapped files were processed with samtools mpileup command version 1.9 (Li et al., 2009), with the following parameters: output all positions including unused sequences (*aa*); maximum per-file depth of 10000 (*d*); base quality threshold of zero (*Q*).

7.3.2 *k* means Clustering

Clustering was carried out using *k* means algorithm with one dimentional data. The three cluster centers were initiated at 0, 0.5 and 1.

7.3.3 Gene Onthology Analysis

Analysis Type:	PANTHER Overrepresentation Test (Released 20181010)
Annotation Version and	GO Ontology database Released 2018-09-06
Reference List:	Arabidopsis thaliana (all genes in database)
Annotation Data Set:	GO biological process complete

	GO molecular function complete
	GO cellular component complete
Test Type:	Fisher's Exact with FDR multiple test correction

7.3.4 Genome-Wide Association Study Analysis

Genome-Wide Association Study Analysis (GWAS) was carried out using EMMAX algorithm (Kang et al., 2010) through the EPACTS v3.2.6 pipeline. Presence and Absence calls were encoded as SNP. To correct for population structure, kinship matrix was created based on whole-genome SNP of the 1001 genomes project (1001 Genomes Sequencing Consortiun, 2016), with Version 3 of the vcf file. Kinship matrix was created with the following parameters: minimum Minor Allele Frequency (--min-maf) of 0.01 and a minimum call rate (--min-callrate) of 0.95. Subsequently Emmax was run with minimum MAF (--min-maf) of 0.05.

7.3.5 Comparison with *A. lyrata* and *C. rubella*

C. rubella reads were downloaded from SRA project PRJEB6689. Run identifiers: ERR636124 ERR636127 ERR636130 ERR636144 ERR636147 ERR636155 ERR636157 ERR636158 ERR636160 ERR636161 ERR636163 ERR636164 ERR636165 ERR636166 ERR636167 ERR636169 ERR636170 ERR636171 ERR636172 ERR636173 ERR636174 ERR636162.

C. rubella accession with the run identifier ERR636144 was excluded from the analysis when comparing *C. rubella* reads to their own reference due to being an obvious outlier corresponding to the reference or closely related to the reference, with NLR average non-zero coverage fraction of 0.99, compared to the remaining 21 accessions, which were all in the range of 0.74 to 0.85.

A. lyrata reads were downloaded from SRA project PRJNA459481. Run identifiers:
SRR7119548 SRR7119547 SRR7119546 SRR7119545 SRR7119544
SRR7119543 SRR7119542 SRR7119541 SRR7119540 SRR7119539
SRR7119538 SRR7119537 SRR7119536 SRR7119535 SRR7119534
SRR7119533 SRR7119532 SRR7119531 SRR7119530 SRR7119529
SRR7119528 SRR7119527 SRR7119526 SRR7119525 SRR7119524
SRR7119523.

8 Abbreviations

ADR1	*ACTIVATED DISEASE RESISTANCE 1*
CC	Coiled-Coil
CDS	Coding Sequence
CHS3	*CHILLING SENSITIVE 3*
DSC1	*DOMINANT SUPRESSOR OF CAMTA3 NUMBER 1*
GWAS	Genome-Wide Association Study
LRR	Leucine-Rich Repeat
MAF	Minor Allele Frequency
NACHT	NAIP, CIITA, HET-E and TP1
NB-ARC	Nucleotide-Binding Adaptor Shared with APAF-1, Plant Resistance Proteins, and CED-4
NLR	Nucleotide-Binding Domain Leucine-Rich Repeat Nucleotide Oligomerization Domain (NOD)-Like Receptors
PPR	Pentatricopeptide Repeat
RGC2	*RESISTANCE GENE CANDIDATE 2*
P/A	Presence/Absence
PV	Pathovar
RLK	Receptor Like Kinase
RPM1	*RESISTANCE TO PSEUDOMONAS SYRINGAE PV. MACULICOLA 1*

RPP1	*RESISTANCE TO PERONOSPORA PARASITICA 1*
RPS5	*RESISTANCE TO PSEUDOMONAS SYRINGAE 5*
RPW8	*RESISTANCE TO POWDERY MILDEW 8*
RRS1	*RESISTANCE TO RALSTONIA SOLANACEARUM 1*
SNP	Single Nucleotide Polymorphism
SOC3	*SUPPRESSOR OF CHS1-2 3*
STAND	Signal Transduction ATPases with Numerous Domains
SUMM2	*SUPPRESSOR OF MKK1 MKK2 2*
TAIR	The Arabidopsis Information Resource
TIR	Toll/Interleukin 1 Receptor
TLR	Toll-like receptors
WRKY	tryptophan (W), arginine (R), lysine (K), tyrosine (Y) motif-containing domain
Y2H	Yeast two-hybrid