On the evolutionary genetics of disease resistance in the *Arabidopsis thaliana* wild pathosystem

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Alba González Hernando

aus Burgos, Spanien

Tübingen

2020

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

Tag der mündlichen Qualifikation:	27.10.2020
Stellvertretender Dekan:	Prof. Dr. József Fortágh
1. Berichterstatter:	Prof. Dr. Detlef Weigel
2. Berichterstatter:	Prof. Dr. Thorsten Nürnberger

“*Quién pudiera conocer*

todo lo que ve la luz

Los universos de ayer

los mañanas del azul.”

Silvio Rodríguez

Table of contents

Abstract

Evolution is a universal process that involves two steps. First, new mutations give rise to genetic variation in populations. Second, evolutionary forces are acting upon them, causing allele frequency changes. These frequency changes are reciprocal in interacting species and lead to coevolution, which is responsible for generating most of the biological diversity on Earth. Natural plant pathosystems are in constant coevolution and therefore exhibit extensive genetic diversity. One of the main aims of population genetic studies is to understand the genetic processes that lead to evolutionary change, including mutation, gene flow, genetic drift, matting systems, and natural selection. However, the relative importance of these processes and how diversity influences host-pathogen interactions in natural populations is mostly unknown.

In this thesis, I aimed to reveal the genetic and evolutionary mechanisms governing plant-pathogen interactions in wild populations. I adopted *Arabidopsis thaliana* - *Hyaloperonospora arabidopsidis* as a model pathosystem to interrogate the relative importance of different genetic processes that affect natural populations. I focused on North American populations of *A. thaliana*, which are outside the native range and can therefore be considered as colonizing and possibly invasive populations. Hpa is a suitable pathogen for coevolutionary studies because it is a specialist obligate biotroph of *A. thaliana*, which means it is in tight coevolution with its host.

In the first chapter, I investigated the genetic paradox of invasion, which tries to explain how colonizing species can adapt to new environments. In particular, I looked at how colonizing populations that undergo a diversity bottleneck can withstand pathogen pressures. With this goal in mind, I revealed the extent and distribution of host genetic variation using next-generation sequencing and population genomic tools. Then, I surveyed these populations for Hpa disease resistance, revealing disease resistance loci. As a result, I found that the colonizing lineage is largely susceptible to the Hpa isolates tested, but it benefits from

outcrossing with other haplogroups. From this study, I concluded that standing genetic variation and gene flow are essential in determining the phenotypic outcome of infection in invasive host populations.

In the second chapter, I moved from looking at host metapopulations and the overall distribution of genetic diversity to investigating host and pathogen genes known to be coevolving. Specifically, I examined host resistance genes and pathogen effectors that follow the gene-for-gene model of interaction. In the first part of this study, I developed and benchmarked a new target enrichment method to simultaneously capture host resistance genes and pathogen effectors in wild infected samples. Secondly, I designed a target enrichment bait set containing the most up to date collection of host resistance genes and pathogen effectors. Combining this new target enrichment technique with traditional shotgun metagenomic sequencing, I was able to assess the distribution and relative abundance of several *A. thaliana* pathogens. Lastly, I explored the presence/absence variation landscape of host resistance genes and pathogen effectors. This analysis led to the discovery of genes with genetic signatures typical from trench warfare and arms race.

Overall, this thesis presents a multiscale approach to study coevolution in natural populations of *A. thaliana* and its specialist pathogen Hpa. Moreover, it provides a new technique to study R-gene and effector coevolution and furthers our understanding of genetic processes involved in shaping plant-pathogen interactions.

Zusammenfassung

Evolution ist ein universeller Prozess, der aus zwei Schritten besteht. Zuerst erhöhen Mutationen die genetische Vielfalt von Populationen. Diesen entgegen wirken evolutionäre Kräfte, die bestimmen, ob neue Mutationen sich durchsetzen können oder ob sie wieder verschwinden. Ein wichtiges Ziel von populationsgenetischen Studien ist die Erforschung der Prozesse, die zu evolutionären Veränderungen führen: Mutation, Drift, Genfluss, Fortpflanzungssysteme sowie natürliche Selektion.

Veränderungen in der Häufigkeit von alternativen genetischen Varianten, oder Allelen, hängen bei interagierenden Spezies von wechselseitigen Interaktionen ab und führen zu Koevolution, welche maßgeblich für die Biodiversität auf der Erde verantwortlich ist. Natürliche pflanzliche Pathosysteme befinden sich in ständiger Koevolution und weisen dementsprechend eine hohe genetische Diversität sowohl auf der Seite des Wirts als auch des Pathogens auf. Hierbei ist sowohl die relative Bedeutung der eingangs erwähnten Prozesse als auch die Frage, wie Diversität die Wirt-Pathogen-Interaktionen in natürlichen Populationen beeinflusst, größtenteils ungeklärt.

In dieser Dissertation habe ich mich dem Ziel gewidmet, die genetischen und evolutionären Mechanismen, welche die Pflanzen-Pathogen-Interaktionen in wilden Populationen bestimmen, zu erforschen. Ich habe *Arabidopsis thaliana* - *Hyaloperonospora arabidopsidis* als Pathosystemmodell gewählt, um die relative Bedeutung verschiedener genetischer Prozesse, die natürliche Populationen beeinflussen, zu studieren. Hierbei beschäftigte ich mich vor allem mit nordamerikanischen *A. thaliana* Wirtspopulationen, die außerhalb ihres nativen Habitats als invasiv gelten. Auf der Pathogenseite ist die *A. thaliana* spezifische Mikrobe *H. arabidopsidis*, welche in enger Verbindung mit dem Wirt ist, für koevolutionäre Studien prädestiniert.

Im ersten Kapitel habe ich das invasive genetische Paradox erforscht, welches beschreibt, wie sich invasive Arten an neue Umgebungen anpassen

können. Insbesondere habe ich mich die Frage untersucht, wie sich einen genetischen Flaschenhals durchlaufende kolonisierende Arten dem Selektionsdruck von Pathogenen widersetzen können. Mit Hilfe von modernen Sequenzierungsverfahrenn und populationsgenomischen Ansätzen habe ich Ausmaß und Verteilung von genetischer Wirtvariation beschrieben. In einem weiteren Schritt habe ich diese Populationen auf Hpa-Befallsresistenz untersucht und deren Genloci ausgemacht. Ich bin zu dem Ergebnis gekommen, dass die *A. thaliana* Genotypen, die ursprünglich nach Nordamerika eingeschleppt wurden, zu einem Großteil anfällig für die getesteten Isolate sind, wobei sie von Auskreuzungen mit anderen, später eingeschleppten Linien profitieren. Aus dieser Studie konnte ich schließen, dass für die Bestimmung des phänotypischen Ergebnisses einer Infektion von einer invasiven Wirtpopulation bestehende genetische Variationen und Genfluss unerlässliche Faktoren sind.

Nachdem ich Wirtmetapopulationen und die allgemeine Verteilung genetischer Diversität untersucht hatte, ging ich im zweiten Kapitel der Erforschung von koevolvierenden Wirt- und Pathogengenen nach. Hierbei habe insbesondere Wirtsresistenz- und Pathogeneffektorgene, die dem Gen-für-Gen Modell folgend miteinander interagieren, untersucht. Im ersten Teil dieser Studie habe ich eine neue Anreicherungsmethode zur gleichzeitigen Erfassung von Wirtsresistenz- und Pathogeneffektorgenen in Proben, die in der Natur aufgesammelt worden waren, entwickelt und bewertet. Im zweiten Teil dieser Studie, habe ich ein Anreicherungsverfahren optimiert, welches die aktuellste Sammlung von Wirtsresistenz- und Pathogeneffektorgenen beinhaltet. Durch die Kombination dieser neuen Methode mit herkömmlicher metagenomischer Sequenzierung nach der Schrotschussmethode konnte ich die Verteilung und relative Häufigkeit von zahlreichen *A. thaliana* Pathogenen auswerten. Schließlich habe ich die Gegenwart/Abwesenheit Variationen von Wirtsresistenz- und Pathogeneffektorgenen untersucht. Dies führte zur Entdeckung von Genen mit Anzeichen dafür, die typisch für Wirts-Pathogen-Interaktionen sind, die entweder dem Grabenkrieg- oder Wettrüstungsmodell folgen.

In ihrer Gesamtheit beschreibt diese Dissertation einen breiten Ansatz zur Erforschung von Koevolution in natürlichen *A. thaliana* Populationen. Weiterhin wird ein neues Verfahren zur Erforschung der Koevolution der Wirtsreservoirs an Resistenzgenen- und des Pathogenreservoirs an Effektorgenen vorgestellt, welches unser Verständnis genetischer Prozesse in der Gestaltung von Pflanz-Pathogen-Interaktionen fördert.

General introduction

Population genetics to study natural variation

Evolutionary genetics is the field that studies the genetic basis of evolution combining principles of Darwinian evolution with Mendelian genetics [1,2], which is commonly known as the modern synthesis or neo-Darwinian theory [3]. Darwinian evolution predicts species changes in traits due to natural selection to become better suited to their environment. At the same time, Mendel's laws tell us about how these traits get inherited over generations. Within the conceptual framework of evolutionary genetics, evolution is seen as population changes in allele frequencies over time. The statistics and theory explaining expected changes in allele frequencies within populations were developed by Wright, Haldane, and Fisher in the '30s, which are considered the founders of population genetics [4–6]. The conceptual and empirical work from Huxley, Dobzhansky, and Muller provided the first empirical evidence of this field. In its beginnings, population genetics relied heavily on models because of the lack of tools to measure genetic variation in many loci [7]. The first attempts to characterize natural genetic variation in populations were made using allozymes, which migrated differently on a gel depending on its amino acid sequence variation. Results studying fruit fly and human populations concluded that there is extensive variation in natural populations in terms of polymorphic loci and heterozygosity [8,9]. Although the study of mutations in proteins with changing gel mobility did not account for the entire spectrum of possible modifications, it became apparent that there is a need to survey changes in the DNA *per se*. It was not until the late 70's when studies of DNA variation came out using restriction mapping [10]. The first genome-wide survey of natural variation was done using restriction enzyme analysis of genomic DNA of *D. melanogaster* [11]. Later, when it became possible to survey large numbers of loci throughout the genome with the aid of automated DNA sequencing, Single Nucleotide Polymorphisms (SNPs) were revealed as the most common variants on the genome. They are mostly found in non-coding regions or are synonymous variants [12]. The breakthrough of molecular population genomics

happened in the 21st century with the advent of high-throughput sequencing methods, which finally gave researchers access to complete genomes and the first catalogs of genome-wide variations from model organisms were published [13,14]. There are fundamental questions of evolution that can be addressed studying the natural genome-wide diversity of a species. First is the relative importance of the different evolutionary forces in creating and maintaining wild populations' genetic diversity [15]. Four fundamental evolutionary forces lead to evolution in populations: genetic drift, gene flow, mutation, and natural selection. Drift happens by random changes in allele frequencies that can lead to fixation, meaning the alternative allele's loss. Gene flow is the process leading to genetic exchange between populations. Mutations are the source of novel variants from which natural selection can select. Finally, natural selection changes the distribution of allele frequencies from adaptive traits in different ways, favoring intermediate frequencies of alleles (balancing selection), maintaining rare alleles (diversifying selection), or purging variation from adaptive loci (purifying selection). Understanding the distribution of allele frequencies and the sources of adaptive variation is fundamental to learn how wild populations evolve and adapt to their environments.

1.1 Colonizations to study evolution in action

Scientists have observed and described species colonizing new habitats since the times of Darwin. They often lead to species rapidly spreading over the new regions, thus called biological "invasions." Naturalists' main interest was trying to understand the ecological relevance of species invasions, which led to the birth in the '50s of the field of Invasion Biology, thanks to Elton's work on "*The Ecology of Invasions by Plant and Animals*" [16]. Although many scientists already suspected the critical role that genetics might have during colonization, it was not until 1965 that the foundations of invasion genetics were established. That year, one of the fathers of plant evolutionary biology, George Ledyard Stebbins, and the invasion ecologist and geneticist, Herbert George Baker, published together "*The Genetics of Colonizing Species"* as a first attempt to summarize and discuss the genetic aspects of invasions [17]. It became apparent that during colonization, species suffer a diversity bottleneck, so the main question was how diversity influences the adaptive potential

and ultimate success of colonizing species. Also, how gene flow resulting from multiple reintroductions, admixture, and interspecies hybridization can help alleviate the diversity loss and lead to evolutionary rescue [18]. Due to rapid globalization and human-related factors, the rate at which species get transferred to non-native environments increases and poses risks to native biodiversity resulting in significant environmental and economic consequences [19]. Therefore, there is a growing interest in the study of invasions, not only for resource managers and conservationists but also for researchers that aim to comprehend genetic causes and consequences of invasions [18]. The recent development of genetic and genomic techniques has unleashed the potential of combining evolutionary genetics and ecology with studying adaptation and evolution during biological invasions [20]. Learning what makes colonizers successful and how they adapt and evolve in the new environment relies on identifying the temporal and spatial aspects of colonization and the extent and kind of diversity that gets introduced [18]. During invasions, populations must genetically adapt to new environmental conditions and therefore represent an outstanding model to study evolution in action and to make inferences of natural populations fate during environmental changes [21].

1.2 Adaptation after invasions

One of the fascinating aspects of species' colonizations is the ability of species to become invasive even when facing diversity bottlenecks. We know that population diversity reductions are a common feature in invasive populations and can have deleterious consequences such as lack of adaptive potential and inbreeding depression [22]. Although many invasive species thrive in their introduced range, posing a genetic paradox of invasion. Consequently, scientists have been trying to understand how much and what kind of diversity is needed for a population to adapt to a new environment [23]. The genetic paradox of invasion can be solved by invasive species adapting from two distinct genetic variation sources: native standing variation or new mutations [24](**Figure 1A**). Adaptation from standing variation can happen when the bottleneck is incomplete. Alternatively, adaptation can be facilitated when enough native diversity is introduced or when subsequent reintroductions bring more native diversity on the introduced range, even creating a

new combination of alleles through admixture and introgression. On the other hand, even if the introduced diversity is low after the bottleneck, new adaptive mutations can arise. These two genetic contexts have contrasting characteristics. Adaptation from standing variation implies that the introduced variation in adaptive traits is enough for selection to act on. It often leads to faster evolution since the adaptive alleles can rise to higher frequency quicker in the populations, with their probability to become fixed increasing and the likelihood of traits variance increasing, thought to lead to evolutionary rescue [23] (**Figure 1B and C**). One of the classic examples of adaptation from standing variation is the ability to digest lactose in humans where the African ancestral derived allele has been positively selected [25]. On the other hand, studies of mutation rates and their role during invasions suggest that new adaptive mutations can rise quickly enough in populations [26,27].

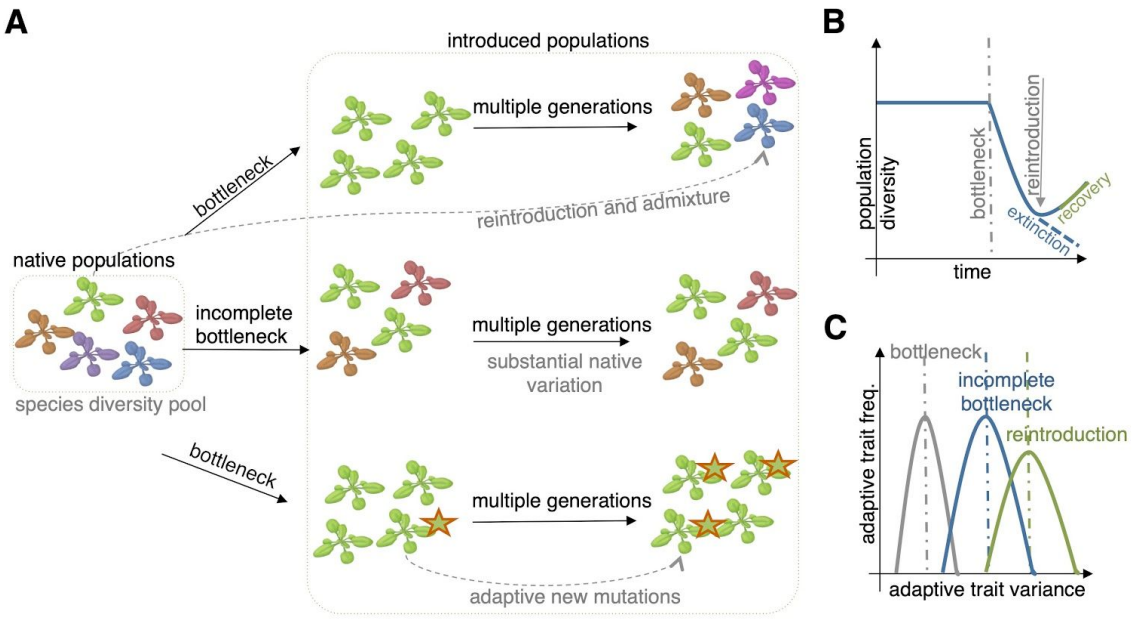


Figure 1. Adaptation after colonizations.

(**A**) The sources of adaptive variation in introduced populations can come from new mutations or standing variation. Reintroductions of native diversity after bottlenecks can result in evolutionary rescue by bringing new adaptive variation to these populations. Incomplete bottlenecks are expected to have enough source variation to adapt in situ. Substantial bottlenecks of diversity are expected to reduce adaptive potential, but new advantageous mutations can arise and sweep in the population, allowing adaptation. (**B**) Model of diversity changes after a bottleneck. The reintroduction of diversity

can help the populations to recover. (**C**) Adaptive traits controlled by large-effect loci are expected to be especially vulnerable to allele frequencies and variance changes. Subsequent introductions can bring new alleles increasing the trait adaptive variance and potentially increase the adaptive trait frequency in those populations.

1.3 The role of admixture and introgression

Intraspecies hybridization and admixture are two concepts that have been used largely interchangeably in the literature. They refer to the sexual reproduction between divergent individuals of the same species, which might experience some reproductive isolation or come from genetically discrete populations with a unique evolutionary history [19]. Introgression occurs through repeated backcrossing of the hybrids with one of the parental genotypes. As a result, they have genomic regions from one genetic background transferred to another. As mentioned earlier, one of the proposed solutions to the genetic paradox of invasion is the reintroduction of potentially adaptive diversity in colonizing populations. Admixture is a fairly common phenomenon in outcrossing species, where admixture between the colonizing lineage/lineages and newly introduced ones will likely happen [28]. One of the documented genetic benefits of admixture is hybrid vigor, heterosis, and the emergence of novel genotypes [29]. Although there are reports of the beneficial role of admixture, there have also been instances where multiple reintroductions and admixture create a geographic mosaic of maladaptation [22].

Furthermore, evidence is accumulating for the adaptive benefit of introgressions. One well-known example is the introgression of DNA from Neanderthal origin into modern-day humans. Several studies linked it to adaptive traits such as altitude adaptation, defense against pathogens, and metabolism, among others [30]. Adaptive introgression is common in plants [31] and has been shown for many species, including sunflowers, *A. thaliana,* and monkeyflowers [32–34]. On the other hand, admixture's adverse effects are evident with crosses between highly divergent sources where negative epistatic interactions lead to hybrid incompatibility [28]. These studies showcase the importance of gene flow in the forms of admixture and introgression in plant adaptation. Although we still have much to learn about the effects of intraspecies gene flow and its adaptive potential during invasions.

1.4 Plant disease resistance in the context of invasions

Adaptive traits in plants controlled by large-effect loci are known to be under frequency-dependent selection, such as self-incompatibility and responses to biotic threats [28]. These traits have essential fitness consequences and can increase in diversity thanks to multiple reintroductions and admixture. Disease resistance is a Mendelian trait controlled by large-effect loci and is thus determined by few genes, in opposition to polygenic traits controlled by many low effect loci. We have seen that adaptation from *de novo* mutations is most relevant for adaptive traits controlled by many genes because of the higher probability of mutation in one of those genes. Contrarily, the adaptation from standing variation is essential for mendelian adaptive traits such as disease resistance. This is because, during bottlenecks, the probability of losing an adaptive allele is higher than the likelihood of substantially reduced variance in quantitative traits. Although there are plenty of studies that show admixture benefits between introduced and native diversity in plants, not many have looked at admixture and introgression of disease resistance during colonization and how it affects populations' ability to fight pathogens [35]. Recent work on the African staple plant *Cassava* identified introgressed disease resistance QTL regions from the wild relative [36]. Others have shown admixture proportions predicting quantitative disease resistance in *Medicago* [37]. Positive heterosis from intraspecific admixture in the common ragweed was observed for simulated herbivory [38].

Another critical aspect to consider when looking at disease resistance in invasions is the enemy release hypothesis. It argues that the lack of native pathogens can help plants increase their fitness [39]. For instance, studies have shown that introduced plant communities have lower fungal and viral infections than their native range [40]. On the other hand, the opposing theory is that the introduced range species' interactions have negative fitness effects, therefore limiting their invasiveness [41]. Models predict the enemy release hypothesis when novel disease resistance genes arrive in wild plant populations, resulting in significant population growth and expansion [42]. All in all, the relevance of admixture and introgression for plant adaptation to pathogens should not be underestimated but expected.

1.5 *Arabidopsis thaliana* as a model to study adaptation during invasions

Finding the underlying genetic mechanisms of adaptive traits is one of the main purposes of adaptation studies. Therefore, we need species whose genomes are well characterized and phenotypes easily obtained. *Arabidopsis thaliana* (*A. thaliana*) is an annual herbaceous plant that has been increasingly used for adaptation studies because of the available genetic resources that opened the door of researchers to find the underlying genetic mechanisms behind adaptation [14,43–46]. Also, the demographic history of *A. thaliana* is understood in detail, helping researchers ask evolutionary questions [47,48]. It has gone through range expansion in the past in the Old World, and recently colonized other continents such as North America [49]. Its worldwide distribution emphasizes this species colonizing potential and poses questions of adaptation during range expansion and colonization. Previous studies have looked at the genetic sources of adaptive variation during these events. Admixture turns out to be a common phenomenon among accessions in the native range [14]. Plus, introgression between the so-called relict and non-relict populations has proved beneficial for adaptation and has been suggested as one of the potential causes of range expansion after the last glacial maximum [50]. Simultaneously, the predicted hybrid incompatibilities among divergent lineages appeared in the case of self-incompatibility locus and NLR mediated incompatibilities [51,52]. In the introduced range, the source of adaptive variation has been very recently evaluated. Work on climate adaptation revealed that native standing variation is behind climate pre-adaptation on the introduced range when the new region's climate matches the one of the native range [53]. In opposition, new mutations seem to have played an essential role in adaptation of the HPG1 colonizing lineage in North America [54]. These are examples of the distinct sources of adaptive variation during colonization. Still, they have focused on polygenic traits, which we know are expected to be less affected by genetic bottlenecks. It remains to be seen how *A. thaliana* can adapt during invasions regarding Mendelian traits such as disease resistance, and to which extent the loss of adaptive alleles can affect its colonizing and adaptation success.

1.6 Population genomics tools to study adaptation

The availability of complete high-quality genomes from *A. thaliana,* together with the decreasing genotyping costs, makes identifying population diversity and structure much more easily accessible [55]. Moreover, we can study the history of a particular locus and link it with specific variants and phenotypes [43]. Two of the most used genomic tools to map the phenotypic variation's underlying genetics are the quantitative trait loci (QTL) and genome-wide association (GWA) mapping. QTL mapping has been one of the first tools used to study variation in *A. thaliana* and allowed for identifying broad genetic regions in the other of Megabases. GWA studies are nowadays possible because we can genotype large amounts of individuals with high-density markers, helping narrow down the putative causal loci at the SNP and gene level [56]. Both QTL and GWA mapping have been proven successful in identifying adaptive loci and their natural diversity. This success was notable for the identification of disease resistance loci using GWAS [57–60] and QTL [61–63]. Despite this success, QTL and GWA mapping have significant drawbacks. Both of these methods are susceptible to allelic effect and frequency on the populations. GWA has low power to detect rare variants and high documented false positives due to population structure. On the other hand, QTL mapping has more sensitivity to rare alleles, but it requires a more substantial allelic effect for being effective. A promising strategy to overcome these pitfalls is the combined use of both mapping techniques [43], which already worked for mapping disease resistance to oomycete pathogens [64]. GWA, QTL, and admixture mapping are also adopted tools to study adaptive introgression in plants [65]. They can be used to find associations between introgressed genomic regions and an adaptive phenotype. Admixture mapping between two distinct genetic groups can help pinpoint the adaptive trait and interrogate its introgressed origin. It seems intuitive then to use mapping tools to examine disease resistance in admixed populations.

Evolutionary genetics of plant-pathogen interactions

Pathogens and hosts represent selective biotic agents because they reduce their fitness reciprocally. Pathogens infect hosts depleting their nutrients, and hosts fight pathogens, reducing their ability to reproduce. This is the basis of coevolution in antagonistic interactions. The dynamics of this interaction at the genetic level means that changes in allele frequencies from host and parasites will influence each other. This genotype by genotype interaction fuels adaptive evolution and is the basis of host-parasite coevolution [66]. Many factors will determine the infection outcome, ultimately leading to disease and thus affecting host fitness. These factors are summarised in the disease triangle, which emphasizes the role of pathogen and host diversity and environment as main disease drivers [67]. The disease triangle has been updated in the genomics era highlighting the pathogen genome by host genome interaction. We need to measure both host and pathogen genomic diversity, how it is geographically distributed, and how diversity maps to each other to learn the coevolutionary processes governing this interaction in the wild.

1.7 The geographic mosaic of coevolution

Plants and pathogens interact in the context of metapopulations. These populations are often formed by genetically distinct individuals and experience different community structures and selective environments [68]. This means that natural selection will act differently in each population and can therefore affect coevolution in various ways. For several decades now, researchers have demonstrated the geographic aspect of coevolution, leading to the Geographic Mosaic Theory of Coevolution (GMTC) [69]. This theory postulates that coevolution will happen at different temporal and geographical scales, depending on each population's diverse evolutionary forces. Much of the coevolution knowledge gained is thanks to the *Plantago lanceolata-Podosphaera plantaginis* pathosystem. From it, we have learned that wild pathosystems are highly structured in terms of host, pathogen, and phenotype diversity. The finding of different local adaptation schemes confirmed the expectations from the GMTC [70–74]. Moreover, coevolutionary

selection generates and maintains the biological diversity of wild communities [75–77]. A proposed new framework for studying pathosystem emphasizes geography as a fundamental aspect of coevolution [78].

1.8 The gene-for-gene model of molecular coevolution

Following the multiscale approach of coevolution, one can go from studying metapopulations to individual genes. While the geographic mosaic of coevolution tells us about metapopulation dynamics, the gene-for-gene model explains how plants and pathogens coevolve at the molecular level. Pieces of evidence gathered using the flax - flax rust pathosystem gave birth to this model in 1956 and continue to do so nowadays [79,80]. It showed that a pair of matching genes in the host and parasite would determine the outcome of infection. The pathogen avirulence protein will interact with the host resistance gene, sometimes directly or, as shown recently, mediated by a guardee or decoy protein [81]. This interaction is based on what is known today as the Effector Triggered Immunity (ETI), as opposed to the Pathogen Associated Molecular Pattern immunity (PTI). There are two postulated, not mutually exclusive hypotheses about how both types of resistance genes can evolve. Long-term stable polymorphisms at the host and parasite coevolving loci are characteristic of the trench warfare hypothesis [82] and are expected to promote molecular signatures of balancing selection, while the recurrent allele fixation in arms races should generate selective sweeps and transient polymorphisms [83]. The best example of arms race dynamics, which exemplifies the other broad hypothesis, is what we usually observed in crop pathosystems when the typical boom and bust cycles lead to fixation of alleles, in opposition to negative frequency-dependent selection imposed by ecological and epidemiological factors leading to trench warfare dynamics in wild pathosystems [84,85].

1.9 *Arabidopsis thaliana* and *Hyaloperonospora arabidopsidis* as a model pathosystem

The plant *Arabidopsis thaliana (A. thaliana)* has been widely employed as a model system for studying plant-pathogen interactions [86–88]. Its importance as a model system relies on the knowledge about its genetics and immune response. The *A. thaliana* immune response consists of two major steps; broad pathogen recognition, which triggers a general plant defense response, PTI, and a more targeted defense, where specific effectors from the pathogen, encoded by avirulence genes (Avr-genes), are recognized by plant resistance genes (R-genes), eliciting a more robust immune response, known as ETI [89–91]. An essential aspect of *A. thaliana* is its pervasive natural variation displayed across its native range [14,15,50,92–94], including remarkable natural variation in disease resistance, both at the phenotypic and genetic level [61,95–97].

The oomycete *Hyaloperonospora arabidopsidis* (Hpa) is a downy mildew pathogen of *A. thaliana*. As an obligate biotroph, it requires the host to remain alive. Thus, Hpa is a *bona fide* pathogen that has been under tight coevolution with its host. Hpa belongs to the phylum oomycetes, which includes many devastating crop pathogens such as *Phytophthora infestans,* the famous causal agent of the Irish potato famine. The development of Hpa as a model pathogen is based on its ability to successfully colonize wild *A. thaliana*, representing its primary oomycete pathogen [88]. The colonization of plant tissue starts when an asexual spore lands on the leaf's surface and the growing hyphae invade the epidermal cell layer until it reaches the mesophyll (**Figure 2, left**). Once there, it forms the feeding structure or "haustoria," which is also involved in the secretion of effector proteins and ultimately leads to plant immune response suppression, resulting in a compatible interaction with the pathogen or susceptibility. When the plant can recognize Hpa effectors, the recognition triggers a defense response, leading to an incompatible interaction or resistance (**Figure 2, right**). The pathogen cell cycle is complete when the hyphae produce the reproductive structures that contain asexual spores.

Eventually, if two hyphae contact each other, they can give rise to sexual spores, which in turn can overwinter, constituting a spore bank on the soil ready to infect newly growing seedlings [87].

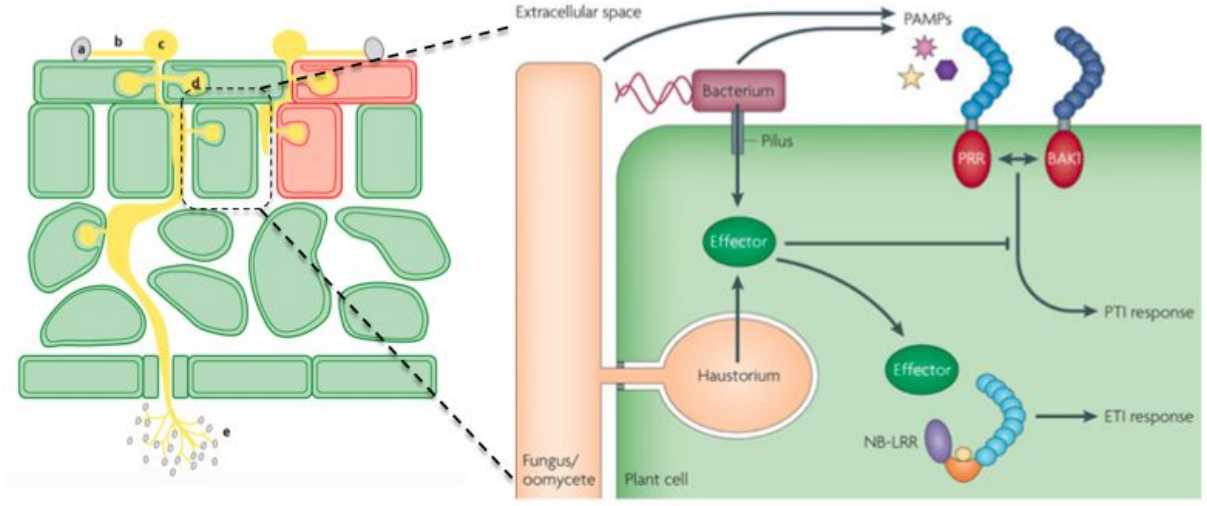


Figure 2. The life cycle of Hpa infection and plant immune response mechanism.

(Left) Schematic representation of the Hpa life cycle. There are two primary phenotypic outcomes from the host side from the interaction with Hpa; susceptible or resistant. **(Right)** Hpa - *A. thaliana* interaction at the molecular level. Pathogen PAMPs and effectors are recognized by host receptor proteins and R-genes, respectively, triggering plant defense response. (Figures modified from [87,89]).

Previous studies characterized the phenotypes of different accessions in response to various wild isolates. The results corroborated the existence of natural genetic variation in disease resistance to Hpa [98–101]. The genes that recognize *Hpa* isolates are called *Resistance to Peronospora Parasitica* (RPP). They confer resistance to specific *Hpa* isolates, while Hpa isolates also differ in ATR genes composition [58,102–104]. The different pairs of RPP and ATR genes follow the gene-for-gene model of interaction. Both R genes and Hpa effectors are highly polymorphic and show signatures of balancing selection, providing evidence of coevolution that follows a trench warfare dynamics [59,105–108].

1.10 Capturing NLR and effector diversity in environmental samples

Due to their fundamental role in molecular coevolution and immunity, there has been growing interest in identifying new resistance genes, pathogen effectors, and their diversity. Plant resistance genes encode intracellular immune receptors that belong to the nucleotide-binding leucine-rich repeat protein family. They are also commonly referred to in the literature as NB-LRR or NLRs. On the other side, oomycetes have effectors with protein motifs such as the RXLR and Crinklers motifs [109]. These conserved motifs enable the search for and annotation of NLRs and effectors in plant and pathogen genomes [110,111]. Genomic surveys have revealed the extent of NLR and effector diversity in many plant and pathogen species and led to the successful cloning of many of them [112,113]. Although the repetitive nature of these genes coupled with their complex genomic features, such as duplications, inversions, and orthologs, made the cloning and identification of these genes very challenging [109,114]. To date, perhaps the most successful technique for assessing NLR and effector diversity is target enrichment sequencing (TES).

TES is a technique that focuses on targeted amplification and sequencing of specific genes and regions of the genome. It utilizes probes complementary to regions of interest to target and enrich these molecules in an NGS library before sequencing. This technique allows for a fast, selective, high coverage, flexible, and cost-effective sequencing of desired targets [115]. These features from TES constitute an advantage over traditional whole-genome sequencing techniques. When sequencing a complex sample, the proportion of reads mapping to the target species is expected to be low, the coverage of genes insufficient, and a low quality reference genome can hamper appropriate diversity discovery by mapping. TES has been leveraged for population genetic studies when the samples exhibit high complexity and diversity of organisms, such as environmental samples. Moreover, complex and diverse gene clusters' characterization also benefits from this technique, allowing proper diversity discovery and mapping such clusters. Recent efforts seek to identify new NLRs in wild relatives of current crop cultivars for resistance breeding using

resistance gene enrichment sequencing (RenSeq) [116–119] and diagnostic resistance gene enrichment sequencing (dRenSeq) [120].

In population genetic applications, TES can be used for diverse purposes. The fact that probes can hybridize to their target regions with up to 80% sequence identity allows for the capture of diversity crucial for population genetic studies (SNPs, CNVs, INDELS) and estimates of population diversity (nucleotide diversity, F_{st}, observed heterozygosity, among others). The discovery of microbiome composition and pathogen diversity from infected plant samples in wild populations has been proven successful with the development of pathogen enrichment sequencing (PenSeq) [121,122]. PenSeq is a method used to conduct targeted population genetic studies of pathogens achieving the high depth of a substantial number of genes, mostly focused on pathogenicity determinants [123]. PenSeq has been benchmarked with oomycete and microbial population genetic studies, making it suitable for uncovering diversity in wild pathosystems. Alongside, RenSeq addresses the diversity of the host immune repertoire, allowing for the discovery of NLR diversity in a variety of plant genomes [124]. Although PenSeq has been used to identify microbial species in wild plant populations [121], either capturing housekeeping genes or ITS, few studies used PenSeq to look at the diversity of targeted microbiome pathogenicity-related genes, such as effectors, toxins, Avr genes, and type III secretion systems. Besides, the fact that PenSeq can assess the relative abundance and characterization of microbes in a plant sample opens the door for association studies between microbiome diversity and composition with the host and oomycete diversity in the same environmental sample. Complementary, RenSeq has recently been used to unveil the set of plant host immune genes, including the entire diversity in NLR genes from *A. thaliana*, resulting in the characterization of its complete pan-NLRome [125].

There is an increasing interest in measuring the degree of host infection by looking at shotgun metagenomic reads rather than estimating infection visually scoring phenotypes. This has the advantage of quantitatively measuring infections of different pathogen species within a single individual at the same time. For example,

it has been shown to be useful in relating microbial and Hpa load with disease in *A. thaliana* [126]. The correlation between pathogen load and infection state opens the door for a high potential application of PenSeq to estimate pathogen load in a given sample, which already showed promising results [121]. Although there are new methods that allow for microbial profiling of environmental samples, such as the host-associated-microbes PCR (hamPCR) [127], those are not yet able to reliably assess the diversity of these microbes at many loci, a must-have when looking at broad scale coevolution and community dynamics.

Thesis scope

To understand plant-pathogen interactions in natural populations it has become increasingly clear that there is a need to combine ecology and evolution with molecular analysis. Thus, in this thesis, I combined principles from population genomics, genetic mapping, and molecular coevolution with the aim of understanding plant-pathogen interactions in natural populations at different scales. I combined state-of-the-art sequencing technologies and large-scale phenotyping experiments to reveal the genetic and phenotypic diversity of the *A. thaliana* - Hpa pathosystem.

In Chapter 1, the main aim was to determine the role of standing variation versus introduced diversity in shaping disease resistance of the host during a recent colonization event. For this purpose, I performed a large-scale investigation of Hpa disease resistance in N. American populations of *A. thaliana* and compared it to populations in the native range. I combined disease resistance surveys, genotyping-by-sequencing, and genetic mapping techniques to obtain a comprehensive picture of Hpa disease resistance distribution and its underlying genetics. Moreover, the genetic basis of Hpa disease resistance is interrogated at multiple scales, ranging from host haplogroups to individual genes. I found that the N. American colonizing lineage is susceptible to the Hpa isolates tested and benefits from outcrossing and admixture with later-introduced haplogroups. To conclude, the main findings of this chapter highlight the relative importance of standing genetic

variation, outcrossing, and the new influx of diversity in plant adaptation to pathogen pressure during biological invasions.

In Chapter 2, the primary objective was to elucidate the coevolutionary mechanisms governing the gene-for-gene model of interaction between *A. thaliana* NLR genes and pathogen effectors. For this purpose, I first developed a new approach to capture both NLRs and effectors on the same sample based on the combination of pathogen-enrichment sequencing with R-gene enrichment sequencing. Second, I created a large and up-to-date collection of NLR and effector genes to use as gene targets for enrichment. Combining target enrichment and shotgun sequencing, I investigated the distribution and prevalence of pathogens and the presence/absence variation of NLR and effectors from wild North American populations of *A. thaliana* infected with Hpa. Despite the complexity of wild samples, I successfully enriched for the desired target organisms and genes. Looking at the presence/absence variation distribution, I found allelic frequencies typical from the trench warfare and arms race coevolutionary dynamics. Taken together, the results from this chapter shed light on the dynamics governing natural plant-pathosystems and open the door for simultaneous population genetic studies of host and pathogen disease-related genes.

To conclude, I present an overall summary of the major findings of this thesis and how they advance our knowledge in the field of evolutionary genetics of natural plant-pathosystems. Based on my results, I then reflect on how the lessons learned from this work can help future studies, and I propose new research directions.

CHAPTER 1

Invasion genomics of *A. thaliana* disease resistance during a recent colonization

Declaration of Contributions

In the present chapter, I conceived the study with input from Detlef Weigel, Gautam Shirsekar, and Fernando Rabanal. Gautam Shirsekar, Rebecca Schwab, and Jane Devos collected the North American host samples and pathogen biological material. Gautam Shirsekar and I sequenced the host samples, and he analyzed the sequencing data, imputed missing markers, and produced the genotype file used in this study. I conducted resistance screenings with the help of Anna Stepanova. I designed the experiments and analyzed all the data unless otherwise stated. Fernando Rabanal performed the linkage mapping analysis. I managed the project, generated the figures, and wrote the manuscript. Fernando Rabanal and Detlef Weigel reviewed and edited the manuscript.

Author	Author position	Scientific ideas	Data generation	Analysis & interpretation	Paper writing
Alba González Hernando	1st	55%	65%	60%	75%
Gautam Shirsekar	2nd	10%	20%	10%	—
Fernando A. Rabanal	3rd	10%	5%	10%	5%
Anna Stepanova	4th	—	5%	—	—
Rebecca Schwab	5th	—	3%	—	—
Jane Devos	6th	—	2%	—	—
Detlef Weigel	7th	25%		20%	20%
Manuscript Title	“Invasion genomics of *A. thaliana* disease resistance during a recent colonization”				
Status in the publication process	To be submitted for publication				

Abstract

When humans accidentally introduce species to a new geographic territory, they typically start with limited genetic diversity. Yet, it is not unusual that some species, despite a narrow genetic basis, quickly adapt to their new home, a phenomenon known as the genetic paradox of invasion. *Arabidopsis thaliana*, a forb that has its origins in Africa and Eurasia, has been introduced in historic times to North America. Because of its many genomic and genetic resources, it presents an excellent case study for determining whether adaptation to a new environment is more likely to result from new mutations or by remixing standing variation. Here, we investigated how North American *A. thaliana* populations resist the specialist oomycete pathogen *Hyaloperonospora arabidopsidis* (Hpa). The most prevalent North American *A. thaliana* lineage, HPG1, is susceptible to two Hpa strains isolated in North America. Still, it has benefitted from outcrosses to other introduced *A. thaliana* lineages, with different resistance genes from various sources providing Hpa resistance in admixed populations. At different scales, our work highlights the adaptive value of outcrosses and genetic recombination in introduced populations.

Introduction

Increased plant invasion rates because of rapid globalization pose risks to native biodiversity and result in major environmental and economic impacts [19]. Therefore, there is a growing interest in understanding the genetic causes and consequences of invasions [18]. During invasions, populations must adapt to new environmental conditions and therefore represent an outstanding model to study evolution in action and to make inferences on the fates of natural populations during environmental changes [21]. Reduced genetic diversity levels in introduced populations have deleterious consequences, such as lack of adaptive potential and inbreeding depression [22]. Nonetheless, many invasive species thrive in their introduced range, posing a genetic paradox of invasion. Two sources of adaptive variation can solve this paradox, the appearance of de-novo mutations and the introduction of native diversity from subsequent colonizations [23,28,128].

Gene flow between colonizing lineages and newly-introduced ones is common and beneficial [28]. Hybrid vigor, heterosis, and the emergence of novel genotypes are known benefits of admixture [29]. Conversely, admixture with later colonizers can create a geographic mosaic of maladaptation [22]. Adaptive introgression helps ameliorate diversity losses; examples in plants include sunflowers, *A. thaliana*, and monkeyflowers [32–34]. On the other hand, admixture can be detrimental among highly diverged lineages, resulting in negative epistasis and hybrid incompatibility [28]. These studies showcase the importance of gene flow in the forms of admixture and introgression in plant adaptation.

Traits controlled by large-effect loci benefit the most from introgression since they can be quickly introduced in the form of new haplotypes of hundreds of kilobases [129]. This is the case for disease resistance (R) genes in plants located in clusters with multiple genes on defined genomic regions. The relevance of admixture, in this case, relies on what new variation gets introduced, rather than how much. Despite the many documented benefits of admixture in invasions, the effect of admixture and introgression in disease resistance has been largely overlooked [35]. Introgression of R loci from wild relatives has been found in the African staple plant Cassava [36]. Moreover, admixture proportions could predict quantitative disease resistance in Medicago [37]. Finally, intraspecies hybridization caused positive heterosis in the common ragweed for simulated herbivory [38]. Thanks to genomic advances, we can identify the genetic source of adaptive variants, unleashing the potential of combining evolutionary genetics and ecology to study adaptation and evolution during biological invasions [65].

The worldwide distribution of *A. thaliana* demonstrates the colonization potential of the species. Together with its extensive genetic variation, it makes it a perfect model plant to study adaptation during range expansion and colonization. Admixture is common among accessions in the native range [14]. Introgression between evolutionary relicts and non-relicts has proved beneficial for adaptation and one of the potential causes of range expansion after the last glacial age [50]. Likewise,

the predicted hybrid incompatibilities among divergent lineages appeared in the case of self-incompatibility locus and NLR mediated disease resistance [51,52].

After its exotic introduction into North America in the early 17th century, a dominant single colonizing lineage (HPG1) has prevailed in these populations due to a founder bottleneck, also known as the founder effect [49,54,130]. The genetic sources of polygenic adaptation in N. America come both from standing variation, in the case of climate pre-adaptation [53], and new mutations in the HPG1 lineage [54]. Monogenic adaptation has not been studied, despite being, as a rule, affected by diversity bottlenecks. The presence of *Hyaloperonospora arabidopsidis* (Hpa), a specialist oomycete pathogen, has been documented in these N. American populations [131]. Specialist pathogens co-evolve with their hosts, promoting the maintenance of host diversity through generations in wild populations, with allele frequency changes indicative of balancing selection [132]. This, together with the pervasive phenotypic and genetic diversity of this wild pathosystem [58,64,104,133], makes it a perfect model to study disease resistance during this recent colonization of N. America.

In this study, we combine genomics tools with resistance screening to investigate the sources and extent of host diversity in the exotic N.American range versus the Eurasian native range, and its effect on Hpa disease resistance phenotypes.

Results

Complex genetics underlies *Hpa* resistance in Europe and N. America

In specific accessions of *A. thaliana*, resistance to different races of the obligate biotroph Hpa is usually governed by single dominant R genes encoding NLR proteins. Major R gene loci conferring Hpa resistance are *RESISTANCE TO PERONOSPORA PARASITICA 1 (RPP1), RPP2, RPP4/RPP5, RPP7, RPP8,* and *RPP13* [86,100,134–137], and several of these are clustered at four genomic loci on chromosomes I, III, IV, and V that have sometimes been called Major Recognition Complexes (MRCs) [138,139]. However, despite the simple genetics of Hpa resistance in experimental crosses, genome-wide association (GWA) studies in diverse

populations have met with limited success in identifying resistance genes for specific Hpa races; while resistance genes and other genes involved in disease resistance were usually among the top hits, they did not stand out, in contrast to resistance genes recognizing bacterial effectors [58,59,64].

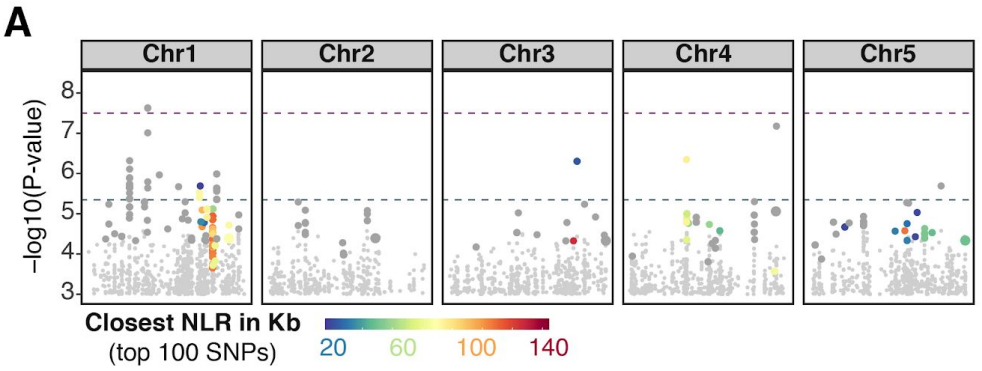
We wanted to learn whether the reduced genetic diversity of *A. thaliana* in N. America [49,54] was reflected in a different genetic architecture of resistance to Hpa. To interrogate the N. American population, we collected seeds and leaves of natural accessions from N. American populations during two consecutive years (2014, 2015); we used a collection of 480 accessions from the introduced range in our study of Hpa disease resistance. For comparison, we investigated 405 *A. thaliana* accessions from the native range in Eurasia, drawn from the 1001 Genomes collection[14]. We used published whole-genome information for the Eurasian dataset (EUR) [14]. For the N. American dataset (US), we genotyped accessions with RAD-sequencing and imputed missing markers using a subset of whole-genome sequenced accessions from N. America and Eurasia.

To ensure that our findings do not reflect a peculiarity of a single unusual Hpa isolate, we studied two N. American Hpa isolates, both of which can infect individuals that belong to the HPG1 haplogroup, the dominant lineage of *A. thaliana* in N. America [49,54]. This resulted in four *A. thaliana* populations x Hpa isolates GWA experiments. For each genotype, we infected five to ten plants at the seedling stage two independent times. We scored spore formation and disease symptoms after seven days post-infection on a quantitative scale. Resistance was present at intermediate frequencies in both populations, and the distribution of disease and resistance in both populations was 57% and 29% in the Eurasian and N. American dataset, respectively.

Our GWA results using these phenotypes were in agreement with similar previous efforts to map Hpa resistance. There were no dominant GWA peaks, and documented Hpa resistance genes did not stand out (**Fig 1A, Table S9**). Nevertheless, for the Eurasian dataset and the Hpa isolate 15IN55, we found two disease resistance-related genes as the top two GWA hits (**Fig 1B**). One was *MYB3*

(AT1G22640), which encodes a transcription factor repressing phenylpropanoids' biosynthesis, which has a well-established role in plant defense [140]. The other was a significant SNP upstream of *PRR1* (AT1G32120), which encodes a pinoresinol reductase involved in lignan biosynthesis, an important cell wall component involved in growth and defense [141]. While GWA hits' overall landscape was distinct in the N. American dataset and Hpa 15IN55, the top hits again included genes with links to disease resistance. The top GWA hit was significant and a missense variant in *LRK10L2* (AT1G66930), which encodes a receptor-like kinase, often involved in disease resistance and symbiotic interactions [142]. The second highest association tagged *GOX3* (AT4G18360), a gene that modulates ROS signal transduction during non-host and R-mediated resistance [143]. The best hits at known Hpa resistance loci of the NLR type was a SNP downstream of *RPP7* (AT1G58602), with rank eleventh among all GWA hits in Eurasia.

The picture for resistance to the Hpa isolate 14OH04 was similar, without any clear major GWA peaks. In Eurasia, the top association was next to a gene encoding a hypothetical protein (AT1G36580) of unknown function. The second-best is in a TIR-NLR gene (AT5G41550), which is part of a larger NLR supercluster that includes the *DM1* and *SSI4* genes, which can cause autoimmunity [144,145]. The third major hit was in *AMP1* (AT3G54720), which is involved in various developmental processes, including ones that can affect the severity of symptoms after bacterial infection [146,147]. In N. America, the top SNP on *4CL3* (AT1G65060), which encodes an isoform of 4-coumarate:CoA ligase (4CL), involved in phenylpropanoid biosynthesis; as mentioned above, these secondary metabolites are well-known for their involvement in plant defense [140]. The second top hit was near a gene (AT4G13580) encoding a dirigent-like protein; these proteins modulate cell walls in response to biotic and abiotic stress [148]. The third top hit was downstream of two genes (AT5G45510 and AT5G45520) encoding proteins with leucine-rich repeat like those found in NLR proteins.



B

Dataset	Isolate	Chr	Position	P-value	Allele	MAF	Distance (bp)	Gene	Consequence
Eurasia	15IN55	1*	11550957	2.35E-08	G/A	0.27	2941	PRR1	upstream gene variant
Eurasia	15IN55	1	8004344	4.81E-07	A/T	0.50	1720	MYB3	upstream gene variant
N. America	15IN55	1*	24972072	2.32E-06	G/A	0.29	0	LRK10L-2.7	missense variant
N. America	15IN55	4	10146279	4.61E-05	A/C	0.07	0	GOX3	synonymous variant
Eurasia	14OH04	1	13803850	1.09E-06	G/A	0.03	12609	h. protein	intergenic variant
Eurasia	14OH04	5	16618424	5.28E-06	C/T	0.09	0	TIR-NBS-LRR	missense variant
N. America	14OH04	1	24167859	7.60E-06	T/G	0.06	0	4CL3	intron variant
N. America	14OH04	4	7883050	1.23E-05	T/C	0.03	17324	dir-like protein	intergenic variant

* above bonferroni treshold

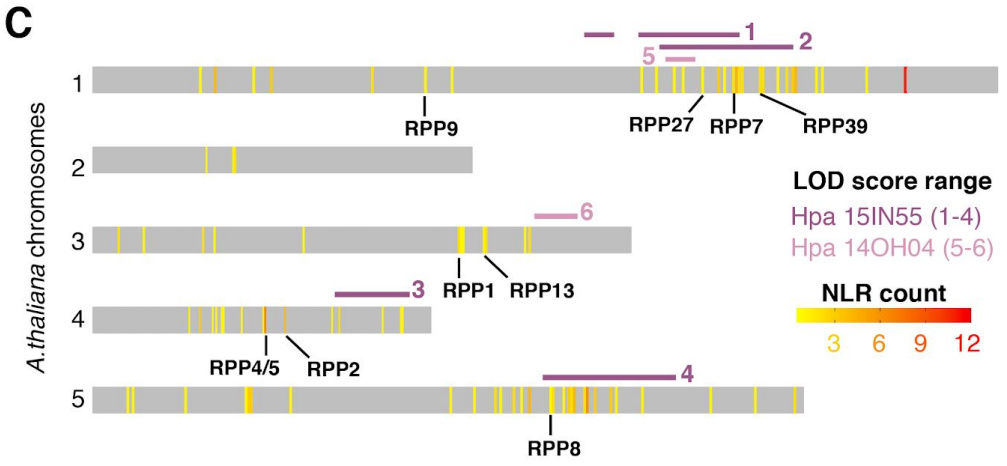


Figure 1. Hpa disease-resistance associated loci and their relation to NB-LRR clusters.

(**A**) Manhattan plot of four GWA for Hpa disease resistance for each Hpa isolate and dataset combination color-coded by the SNP genetic distance to the closest known NLR for the top 100 hits. Dashed lines represent Bonferroni thresholds for the N. America (blue) and Eurasia (purple) datasets. (**B**) List of top two GWA hits for each dataset x Hpa isolate combination (**C**) *A. thaliana* chromosomes displaying NLR cluster hotspots labeled for the known RPP genes. Top LOD scores ranges are shown on top for each isolate due to the QTL mapping of Hpa disease resistance.

That NLR genes were best among the top GWA hits but did not stand out in the GWA results suggested that *A. thaliana* accessions can resist the two Hpa isolates with different resistance alleles at the same loci, or with alleles at other

resistance genes altogether. To ascertain that the genetics of resistance was simple in experimental crosses, we carried out QTL mapping of 15IN55 resistance in two crosses and 14OH04 resistance in a single cross. The segregation ratio with the best fit for all three crosses was 1:15 (susceptible: resistant for 15IN55 and resistant: susceptible for 14OH04), suggesting the involvement of two genes with an epistatic effect, showcasing a more complex genetic basis than the classical gene-for-gene model of disease resistance (**Table S10**). Contrary to the GWA results, we found a limited number of QTL regions in each cross, with one region appearing in all three crosses and non-overlapping secondary regions (**Fig 1C**). At least two of the QTL overlapped with or were near Major Recognition Complexes MRC I and MRC V [64].

RPP4/RPP5 cluster mediates 14OH04 Hpa disease resistance in the N. American MISJCJT population

Even though GWA mapping did not produce very encouraging results, neither in the Eurasian nor the less diverse N. American populations, we suspected that the genetic basis of resistance might be more straightforward in a local population, as has been suggested before [60]. We decided to test the concept of "local" GWA using the N. American population MISJCJT, consisting of 77 individuals from Michigan that differ in about 60,000 SNPs and segregate for resistance Hpa isolate 14OH04. There was a single very clear GWA hit on chromosome 4 (**Fig 2A**). The best hit was a non-synonymous variant in AT4G17140 with an almost perfect association with resistance (**Fig 2B and C**). This gene encodes a pleckstrin homology (PH) domain, also found in the *ENHANCED DISEASE RESISTANCE 2 (EDR2)* gene [149]. However, even better candidates for the causal locus are genes in the nearby *RPP4/RPP5* cluster of NLR genes, since at least two members, *RPP4* and *RPP5*, from different *A. thaliana* accessions, provide resistance to different Hpa races [135,150–152].

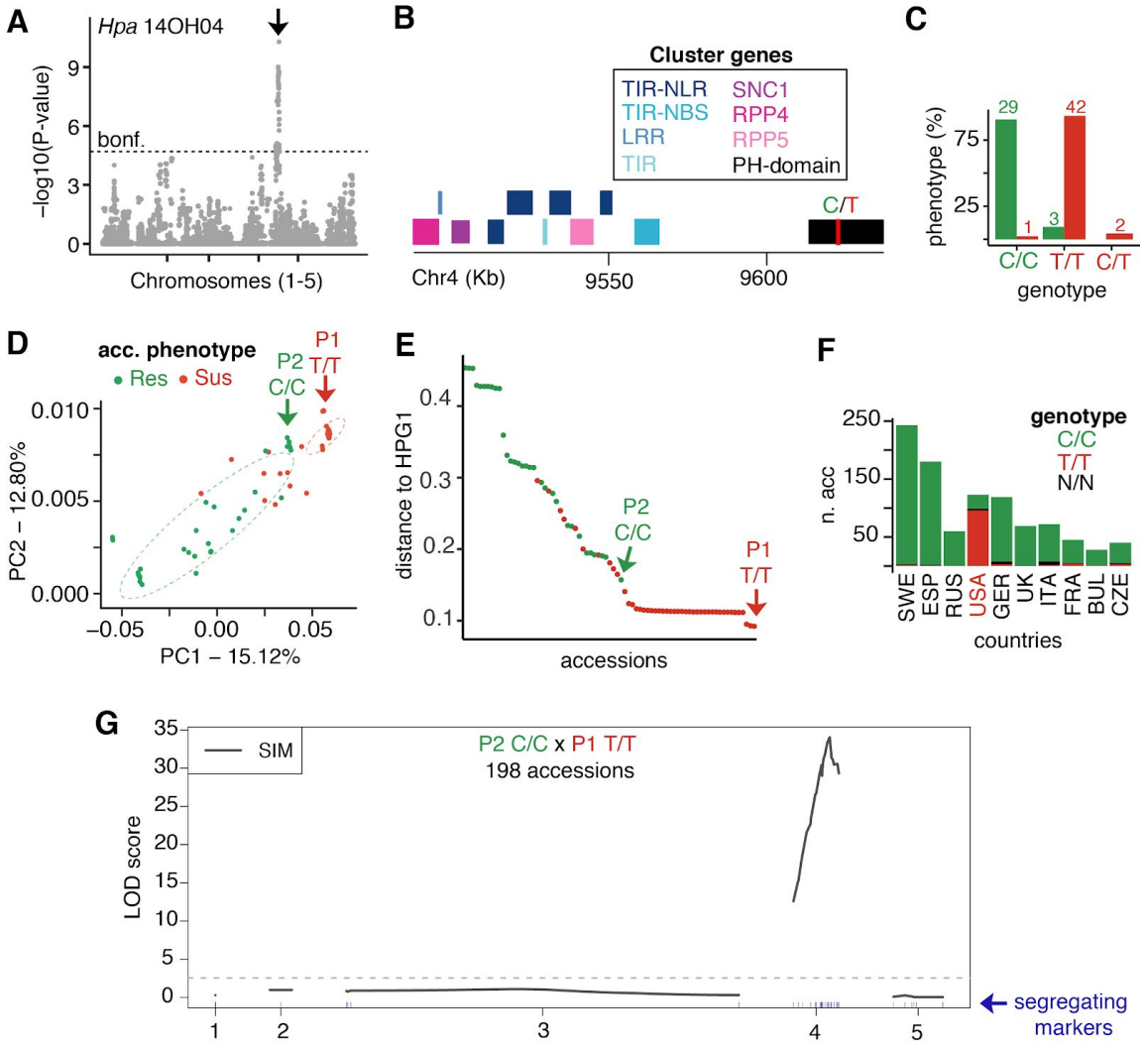


Figure 2. *RPP4/RPP5* cluster mediates 14OH04 Hpa disease resistance in the N. American MISJCJT population.

(**A**) GWA study of 77 accessions of the MISJCJT population showing the highest associated loci the *RPP4/RPP5* cluster (arrow). (**B**) Close-up of the *RPP4/RPP5* cluster with top GWA hit SNP (C/T). (**C**) Phenotypic distribution of resistance among individuals carrying the alternative alleles. (**D**) PCA of host kinship matrix color-coded by Hpa resistance phenotype. Ellipses indicate 0.7 normal confidence level. Parents for the QTL mapping cross marked with an arrow. (**E**) Distribution of genome-wide distance to HPG1 in the MISJCJT population. QTL parents marked with an arrow. (**F**) In the 1001 Genomes data set, the susceptible marker T is almost exclusively found in N. American accessions. (**G**) QTL mapping results confirming the *RPP4/RPP5* cluster as the primary locus for resistance to Hpa 14OH04.

A PCA of genome-wide SNPs revealed that accessions with the susceptible T allele at AT4G17140 and accessions with the resistant C allele at AT4G17140 only partially overlapped (**Fig 2D**). The susceptible T allele is found in the haplogroup 1 (HPG1) genome, representing the most common genetic lineage among N.

American *A. thaliana* individuals [49,54]. When we ordered MISJCJT individuals according to their genome-wide genetic distance from HPG1, we found that genetic distance was correlated with the likelihood of being resistant (**Fig 2E**). To validate the GWA candidate locus, we performed QTL mapping using a susceptible individual and a resistant individual close to the canonical HPG1 genotype, reasoning that the resistance gene might have been introgressed from another lineage into the HPG1 background (P1 and P2; **Fig 2E**). Indeed, the two parents differed minimally on all chromosomes, but chromosome 4 and QTL mapping confirmed linkage of resistance to Hpa isolate 14OH04 to the *RPP4/RPP5* region (**Fig 2G**).

We then more closely examined the distribution of both alleles of the resistance-associated marker SNP within the 1001G dataset. While the susceptible T allele is the predominant allele among N. America accessions (78%), the resistance-associated C allele is nearly fixed in every other group/country (98%) (**Fig 2F**). These results are in support of the hypothesis that the resistance carried by P2 might have been introgressed into the HPG1 background, rather than being the result of a de-novo mutation.

Introduced populations show substantial diversity and admixture levels

Since we had found evidence for introgression from another lineage likely having been important for introducing Hpa disease resistance into the susceptible HPG1 background, the dominant lineage in N. America [49,54], we wanted to learn more about the relationship between admixture and Hpa disease resistance in our material.

To reveal population diversity and structure in the EUR and US populations that we had phenotyped for Hpa disease resistance, we computed a kinship matrix and performed Principal Component Analysis (PCA). In addition, we identified distinct ancestry groups using the allele-frequency based ADMIXTURE software [153]. We used the previously identified 11 different ancestral haplogroups for the EUR dataset [154], and we estimated a total of 10 haplogroups for the US collection (**Fig 3**). It is important to emphasize that the number of estimated haplogroups does not

directly reflect the diversity in the EUR and US populations, which is much lower in the latter, as evidenced by the difference in nucleotide diversity (π) along the genome (**Fig 3D**). Of the 405 EUR accessions, all but five could be classified into an admixture-based haplogroup (**Fig 3C**). All 480 N. American accessions were classified into a haplogroup, with nearly half of the accessions belonging to the K4/US haplogroup (**Fig 3H**).

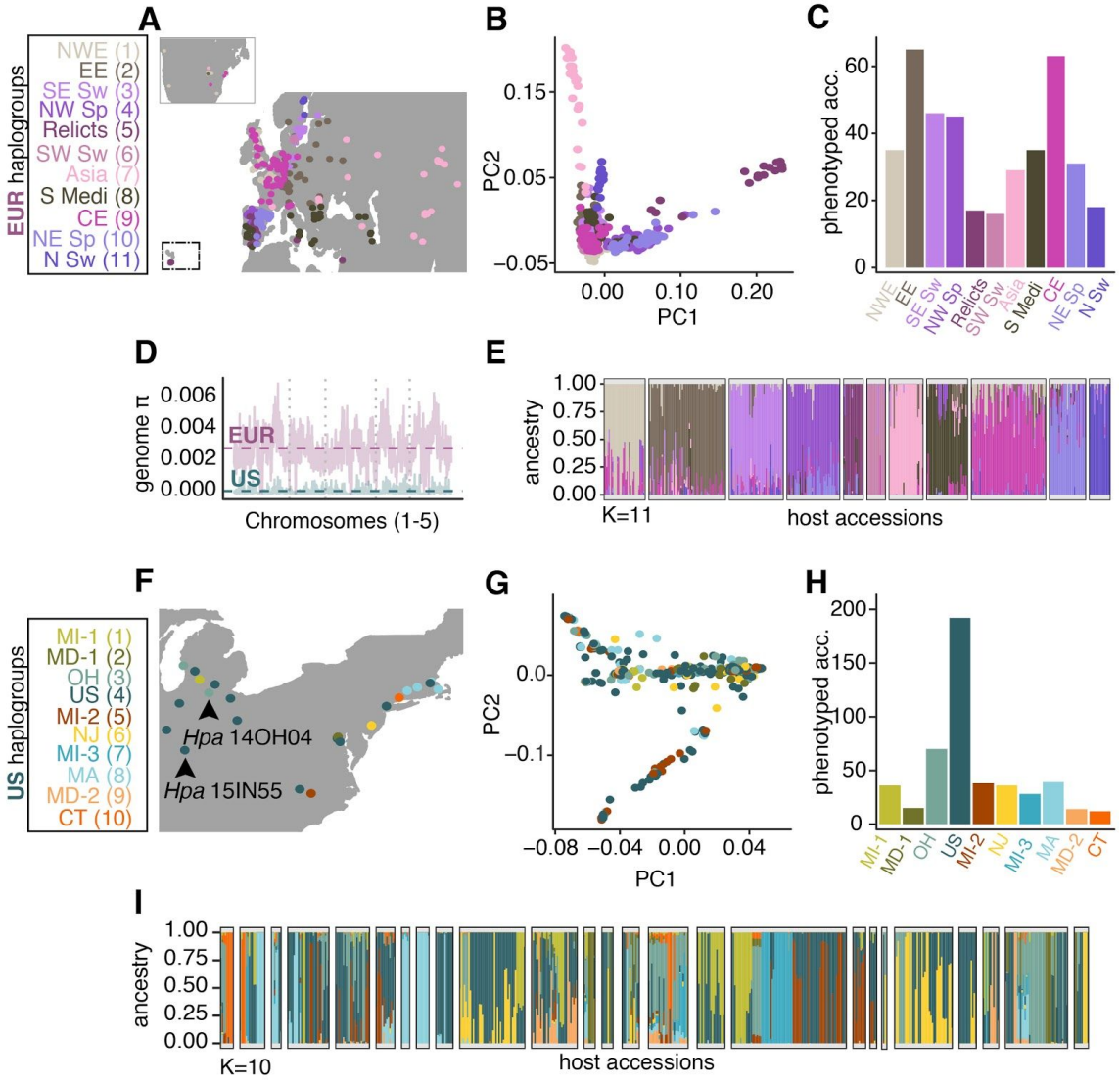


Figure 3. Introduced populations show substantial diversity and admixture levels.

(**A, F**) Geographical distribution of *A. thaliana* accessions phenotyped for Hpa disease resistance. Colored points are accessions for the Eurasian dataset, and populations represented by multiple accessions for the American dataset. Arrowheads mark the population origin of Hpa isolates used in this study, and colors indicate haplogroups. (**B, G**) PCA of accession kinship matrix color-coded by haplogroup. (**C, H**) The number of accessions phenotyped for Hpa disease resistance by haplogroup. (**D**) Genome-wide nucleotide diversity (π) of both datasets, average π for each dataset displayed by horizontal hyphenated lines. (**E, I**) ADMIXTURE ancestry proportions for each accession are

indicated by vertical bars, boxed-grouped by population in the US, and haplogroup in Eurasia and color-coded by haplogroup.

Geographic distribution of Hpa disease resistance reflects host HPG1 ancestry

Although *A. thaliana* has been introduced to N. America only in historical times [54], there is today substantial genetic diversity, indicating multiple introductions over the past 400 years or so. Besides, it has been previously observed that while accessions with identical genome-wide genotypes are much more prevalent in N. America than Eurasia, there is also an excess of near-identical, but genetically distinct pairs of accessions, suggesting that admixture is common in N. America [49,54].

If N. America were initially dominated by a single lineage, e.g., HPG1, any new outcross would have been likely to an HPG1 member, thereby continuously diluting the genetic contribution of newly arrived lines under neutral selection. Besides, we had already noticed in the MISJCJT population the presence of multiple individuals that were genome-wide closely related to HPG1. Therefore we were particularly interested in determining the genetic relatedness of non-HPG1 accessions in the US collection to the canonical HPG1 lineage. We calculated pairwise genetic distances of each accession to the available HPG1 reference genome [155]. HPG1-relatedness is geographically structured in both N. America and Eurasia (**Fig 4A and B**). Within Europe, we find that the accessions most similar to HPG1 come from the UK (**Fig 4C**), in agreement with a previous proposal that British and Western Eurasian populations are the source of HPG1 introduction [14]. The HPG1 reference genome originates from a Michigan accession; we, therefore, hypothesized that N. American populations within this state should have the most HPG1 ancestry. Two Michigan populations are on average among the most similar ones, with populations from New Jersey and Maryland having, on average, an even lower genetic distance to HPG1 (**Fig 4C**). This finding strengthens the previous evidence of an HPG1 entry point in N. America through the East coast, and later migration to the West [54].

Because HPG1 is susceptible to the two N. American Hpa tested, and because HPG1 ancestry is unevenly distributed geographically, we were curious whether resistance was also unevenly distributed geographically. Previous studies have investigated the link between the geographical origin of the accessions and distribution of Hpa disease resistance but found no correlation [64,104]. Accessions from the EUR and US collections were classified as resistant if and only if incompatible interactions for both Hpa isolates tested occurred. When we looked at the continental distribution of phenotypes (**Fig 4D**), both resistance and susceptibility were more unevenly distributed in N. America than in Eurasia (**Fig 4D and E, Fig S3**).

Of particular interest was an apparent overlap between regions with high HPG1 relatedness and susceptible geographic areas, particularly in N. America (**Fig 4A and D**). Indeed, in the US collection, we found that HPG1-like accessions are most often susceptible, supporting the hypothesis that the geographic pattern of Hpa disease resistance phenotypes might be determined to a considerable extent by each accession's genetic similarity to HPG1. As expected, such a trend was not evident in the EUR collection (**Fig 4F**). We know from previous work that resistance to Hpa is more common than susceptibility, with some variation depending on the isolate tested ([64,104]. We found a similar proportion of resistant accessions in our EUR dataset, as seen for other European Hpa races (57% vs. 60%) [64]. In the US dataset, susceptibility is more prevalent (71%) (**Fig 4F**). The difference in the proportion of resistant accessions between EUR and US was statistically significant (**Table S5**). The importance of sympatric host-pathogen interactions and local adaptation has been previously suggested for British Hpa isolates. These accessions were more likely to be susceptible to one of the local Hpa strains (Emco5) [64]. This is consistent with what we report, with N. American accessions being more susceptible to US Hpa isolates. In terms of Hpa pathogenicity, the reported numbers of infected accessions range from 17% to 46% [64]. This number is slightly higher with 30% pathogenicity for the EUR datasets and reaching up to 60% pathogenicity for the US datasets. The difference in the ability to infect *A. thaliana* accessions differs minimally between our two Hpa isolates, only ~2.5%.

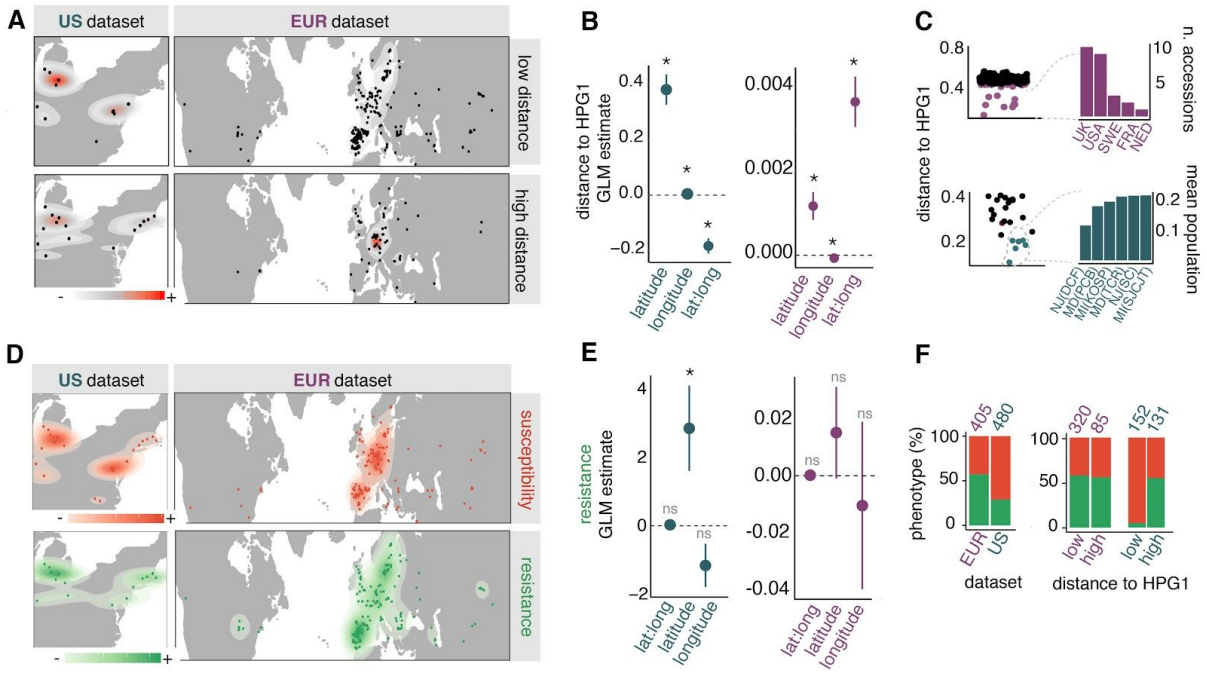


Figure 4. Geographic distribution of Hpa disease resistance is mainly driven by host HPG1 ancestry.

(**A**) The distribution of HPG1 genetic distance for each dataset. (**B**) HPG1 genetic distance GLM estimates for each geographical predictor variable. (**C**) Distribution of HPG1 genetic distances for each dataset. Countries and populations with the lowest genetic distance to HPG1 are zoomed on the left. (**D**) Geographic hotspots of Hpa disease phenotypes, colors depict the density gradient, in red for susceptibility and in green for a full resistance. (**E**) Hpa full resistance phenotype logistic GLM estimate of predictor geographical variables. Each model's estimate's statistical significance is denoted as a star (*) when associated p-value < 0.05. (**F**) Phenotype proportions for each dataset and each HPG1 genetic distance category.

Host haplotype and admixture differences in Hpa disease resistance

Having discovered a suggesting correlation between HPG1 ancestry and Hpa disease resistance, we looked at the phenotypes' distribution by haplotype and admixture levels. Based on the fact that HPG1-like accessions are susceptible, we asked if this holds at the haplogroup level, expecting haplogroups genetically closer to HPG1 also to be more susceptible. Therefore, we calculated the distribution of phenotypic proportions for each haplogroup (**Fig 5A**) and their average genetic distance to HPG1 (**Fig 5B**). In the US collection, MI-2, NJ, and usHPG1 were the most susceptible haplogroups in addition to being the most HPG1-like (**Fig 5A and**

B). Conversely, the Massachusetts (MA) haplogroup had the highest resistance among US haplogroups and was genetically the second most dissimilar to HPG1 (**Fig 5A and B**). In the EUR dataset, the situation is different, and resistance prevails in all haplogroups except in the relicts, which have only 35% resistance.

Since accessions were classified into haplogroups based on dominant ancestry proportion, we could be overlooking the putative role of admixture and diverse ancestries within an accession in Hpa disease resistance. To address this concern, we used each haplogroup's ancestry proportions to ask if admixture drives the observed phenotypes (**Fig 5C**). We only found two instances in which admixture levels were significantly different between resistant and susceptible accessions. Accessions with more MA haplogroup ancestry were, on average, the most resistant ones. This case could indicate adaptive introgression of disease resistance, where carrying genomic regions with MA ancestry conferred a selective advantage. On the other hand, EUR accessions with more Eastern Europe (EE) ancestry were, on average, more susceptible, indicating that susceptibility does not follow a simple rule in Eurasia. However, correlations between Hpa disease resistance and the host's overall genome-wide relatedness have not been detected before [104]. This is also held for our EUR dataset (**Fig S4**). In contrast, both PCA axes significantly separated susceptible and resistant accessions in N. America (**Fig S4**, **Table S4**). We also looked at the link between overall population diversity and disease resistance, counting the number of distinct haplogroups and the genome-wide nucleotide diversity in each local US population (**Fig 5D**). We found that the number of different haplogroups (but not overall genome-wide diversity) in a given population correlated significantly with the number of susceptible accessions ($r=0.56$, $p<0.05$), but not resistant accessions ($r=0.22$, $p=0.3$). From these results, it is clear that what predicts resistance is not the amount of diversity that has been introduced, but rather the admixture group composition and the amount of remaining HPG1 ancestry.

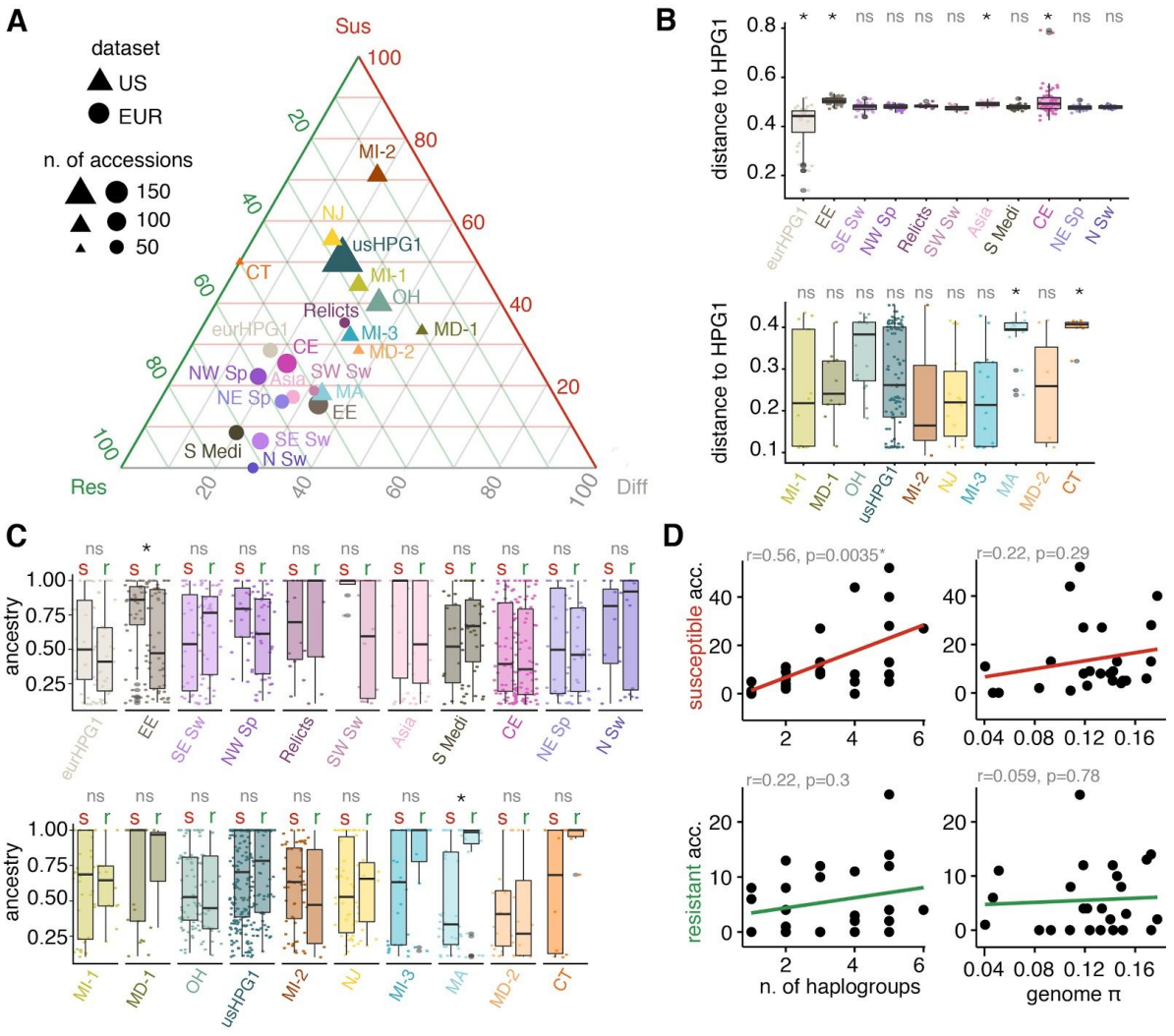


Figure 5. Host haplotype and admixture differences in Hpa disease resistance

(**A**) Phenotype proportions for each dataset and haplogroup. The three phenotypes used as triangle corners are complete susceptibility (Sus), complete resistance (Res), and differential responses to the two Hpa isolates tested (Diff). (**B**) Genetic distance to HPG1 for each haplogroup in the EUR (top) and US (bottom) datasets. Pairwise t-test of each haplogroup means against the base-mean (denoted with "*" when associated p-value < 0.01, and "ns" when non-significant p-value). (**C**) Ancestry proportions for each haplogroup grouped by Hpa disease phenotype for the EUR (top) and US (bottom) datasets. Statistical significance of the Wilcoxon test comparing pairwise means within each haplogroup is denoted with "*" when associated p-value < 0.05 and "ns" when non-significant p-value. (**D**) Correlation between the number of susceptible and resistant accessions per population with their mean genome-wide nucleotide diversity (π) and the number of distinct haplogroups. Pearson correlation coefficients (r) and their associated p-value are shown above each scatters plot, "*" denotes p-values < 0.05. Regression lines in colors.

Evidence for adaptive introgression of disease resistance of the *RPP4/RPP5* cluster

It is becoming increasingly clear that gene flow between divergent taxa can generate new phenotypic diversity, allow for adaptation to novel environments, and contribute to speciation [129]. Gene flow through introgression events can introduce large blocks of novel variation into a population, potentially transferring an adaptive trait from the donor to the recipient population. Introgression might be particularly helpful for disease resistance traits since NLR-type resistance genes are found in clusters. A single introgression of an entire resistance gene cluster could introduce multiple, linked resistances [156,157].

In the MISJCJT population, we had mapped resistance to the Hpa 14OH04 isolate to the *RPP4/RPP5* cluster in two parents whose major genetic differences were restricted to chromosome 4, suggesting that the resistant *RPP4/RPP5* allele had likely been introgressed into the HPG1 genomic background. Therefore, we wanted to know whether this reflected a more general pattern and looked at the *RPP4/RPP5* cluster's diversity in the US and EUR accessions. We extracted 2,012 variants from 1,615 accessions across a genomic region of 148 kb in the Col-0 reference genome, including the *RPP4/RPP5* cluster and the top associated SNP. We removed accessions identical in this region, which left 1,160 accessions to build a Maximum Likelihood (ML) phylogenetic tree with RAxML.

After collapsing clades with similar average branch length, we identified eight distinct clades that contain accessions from the US dataset and a few single-lineage branches (long purple arrows) (**Fig 6A**). There were two main clades for the *RPP4/RPP5* cluster in the MISJCJT population, clade 1, including the susceptible parent (P1), and clade 2, including the resistant parent (P2). All HPG1-related accessions from the US and EUR datasets belonged to clade 1, confirming a unique susceptible haplogroup's hypothesis in this lineage. To quantitatively measure the N. American *RPP4/RPP5* cluster's similarity, we calculated pairwise percent identity (PIM) of the concatenated SNP arrays. We clustered them according to the ML tree

clades (**Fig 6B**). Members of clade 1 were very similar and had the lowest PIM with clade 2, highlighting the *RPP4/RPP5* clades' divergence.

Our objective was to track both haplogroups' geographical origin (clade 1 and clade 2) and that of the individual parental haplogroups (P1 and P2). To do so, we classified each accession into one of the ML tree clades and looked at their geographical distribution (**Fig 6C**). Clade 1 was the most common in the US collection of accessions, in line with what was previously seen with individual SNPs (**Fig S5B**). Clade 2 was most common in Central, North, and Western European accessions and rare in the US dataset (**Fig S5B**). This observation supports the idea that the resistant allele came from a non-HPG1 Eurasian accession.

To reveal potential source accessions and geographic origin of the distinct *RPP4/RPP5* haplogroups in the parental lines, we selected the top ten accessions with the highest sequence percent identity with each parent (P1 and P2) and looked at their locations (**Fig S5A**). The sequences most similar to the susceptible parent P1 were in N. American HPG1-like accessions. In contrast, the sequences most similar to the resistant parent P2 came from accessions belonging mostly to clade 2 and continental Europe. Taken together, these findings provide evidence of introgression of the resistant allele from a European source into the N. American HPG1 background. It remains to be seen whether these alleles have been maintained in the population because of their selective advantages, or by random processes such as genetic drift.

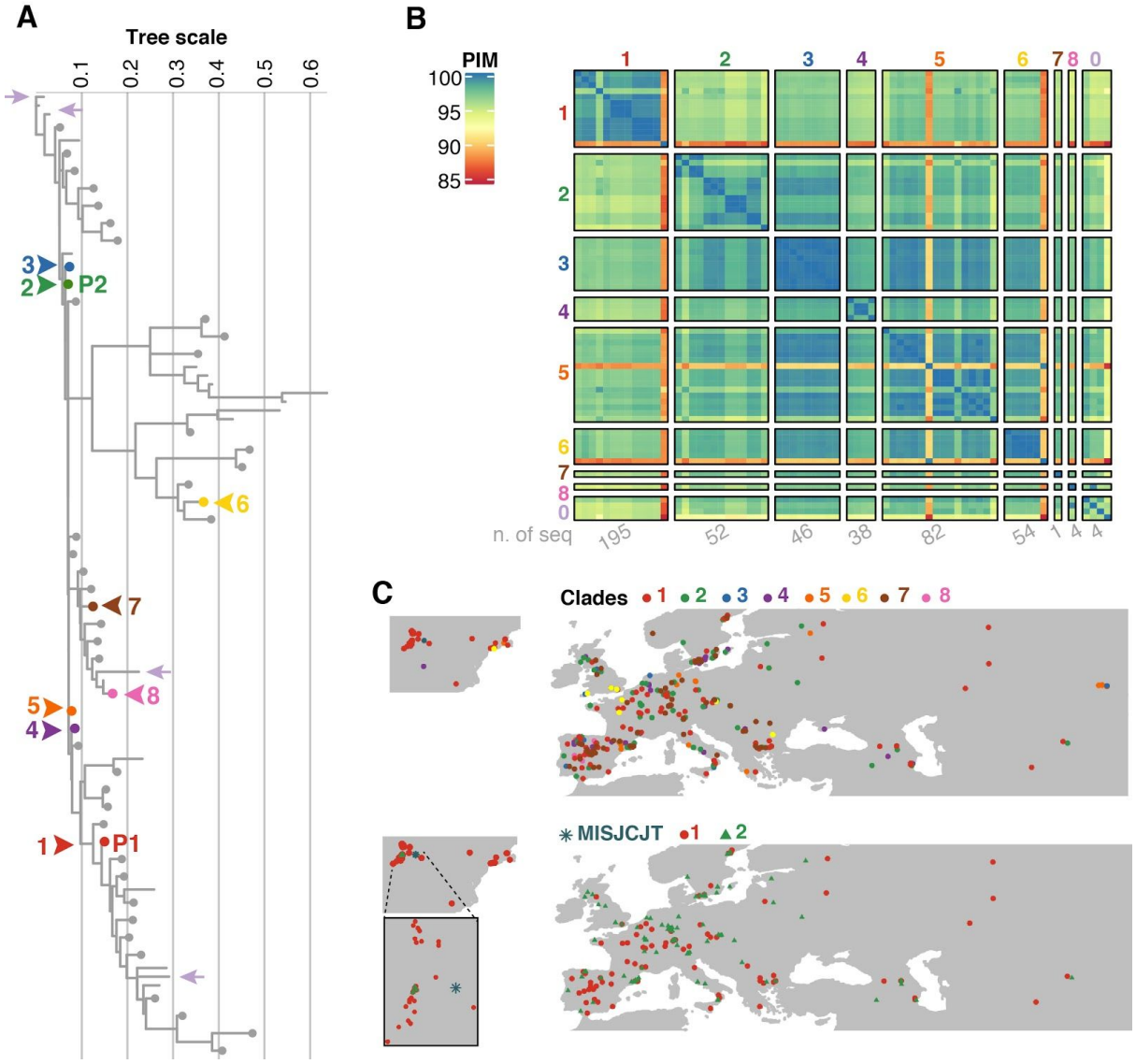


Figure 6. Origin of disease resistance of the *RPP4/RPP5* cluster.

(**A**) ML tree of the *RPP4/RPP5* cluster containing Eurasian and N.American accessions. Clades containing N. American accessions are color-coded, with the few that didn't cluster in any clade indicated by purple arrows. Parents used for QTL mapping are denoted as P1 and P2. (**B**) Pairwise Percent Identity Matrix (PIM) of concatenated SNPs clustered by ML tree clades by axes. Only non-redundant sequences are displayed, with the total number of sequences that belong to each clade noted below in gray. (**C**) Geographical distribution of Eurasian accessions color-coded by ML tree clades on the left. On the bottom, it is displaying accessions from the MISJCJT showing only the Eurasian accessions that cluster in the same clades (Clade 1 and 2) and zoom on the MISJCJT region.

Discussion

Hpa disease resistance screenings have been done in a small subset of *A. thaliana* accessions, mostly from the Nordborg collection, whose accessions come predominantly from Sweden, meaning that most of the diversity space remains unexplored [57,58,64,104]. We hoped to reveal new disease resistance loci by screening a representation of the entire host genetic diversity and major admixture groups. Moreover, we hoped to leverage the limited host diversity in the introduced range to fine-map Hpa resistant loci. However, we encountered one of GWA studies' common limitations: the link between phenotype and population structure. Correcting for population structure reduced our power to detect statistically significant associations. Another common limitation in detecting associated variants is allelic heterogeneity, which means that the same phenotype can be achieved by different combinations of genes [158]. This could have been the case for our metapopulations, considering that we grouped many different genetic clusters. If the resistant alleles were found in low frequencies, this could have also hampered our results. Despite these caveats, most of the loci that we identified are on NLRs or defense-related genes. We were also able to partially reproduce published Hpa disease resistance QTL regions containing NLR clusters. The fact that there is minor overlap between the identified QTL regions and GWA associated loci could simply mean that loci governing resistance at the continent scale are different from those of the crossed accessions since the main resistance loci from GWA do not segregate in the parental lines chosen for QTL mapping.

Founder effects decrease the allelic richness of invasive populations, affecting oligogenic traits because, for these traits, heterozygosity and allelic diversity are important [159]. Gene flow from subsequent colonizers can help alleviate this diversity loss and help plants endure new environments. Theory predicts that natural selection should favor the introgression of alleles under balancing selection [22]. Resistance genes in *A. thaliana* display signatures of balancing selection in wild populations and could be subjected to introgression events [59,106,160]. One way to test the adaptive introgression hypothesis is to find associations between introgressed

genomic regions and an adaptive phenotype. We followed the standard approach of first identifying the genomic basis of an adaptive phenotype using the concept of local GWA and QTL mapping and then looking at the distribution and source of the resistant haplotype. We showed that disease resistance in one of the Michigan populations is mediated by the *RPP4/RPP5* cluster and two segregating haplotypes. Our findings on the geographical distribution of the most similar haplotypes in Europe revealed that the resistant haplotype was brought to N. American from a non-HPG1 accession. The fact that the resistant parent genetic background was HPG1 but carried the Eurasian-origin resistant haplotype provides evidence for the adaptive introgression of disease resistance hypothesis.

Comparisons of genetic diversity in the introduced and native ranges of a species are necessary to find the invader(s) sources, the extent of introduced diversity, and potential reintroduction events [161]. By extensively sampling within the introduced range and comparing it to the native one we could identify contrasting degrees of population structure and diversity. The fact that we found very low levels of nucleotide diversity in the N. American range confirms that these populations went through founder bottlenecks [54]. However, we found the distribution and extent of N. American genetic variation to be higher than what previously reported [49], with complex admixture scenarios and the presence of other major haplogroups. This finding supports the previous hypothesis of *A. thaliana* outcrossing rate being high enough to mix haplotypes in its introduced range in N. America, contributing to the increase of genetic diversity [49].

One of the standing questions in wild plant-pathogen associations is what generates variation in disease resistance. Disease resistance has a genetic basis, but the relative importance of adaptive versus non-adaptive processes in driving resistance is still not understood [73]. Moreover, variation in specific traits over a broad geographical area driven by arms races is already predicted under the mosaic theory of coevolution. We managed to determine the genetic drivers of phenotypic diversity at different geographical scales. Our results exemplify how the distribution of disease resistance phenotypes across space can vary significantly and showcase

the importance of non-adaptive processes, in this case, the founder effect and bottleneck, in driving geographical patterns of resistance.

A popular explanation of genetic rescue after a founder event is reintroducing adaptive variation from native source populations [29]. It is also acknowledged that low genetic diversity in wild populations increases the risk of epidemics and can be counteracted by increasing host population diversity [162]. However, increasing overall genetic diversity in N. American populations did not seem to affect the disease outcome. The reverse is true: increasing the number of admixture groups present in N. American populations correlates with Hpa susceptibility. One explanation could be that most of these haplogroups are mainly susceptible; therefore, additional admixture does not bring resistant variants to the populations. When looking at the N. American dataset as a whole, we observed the beneficial effect of admixture for one haplogroup in Massachusetts. The results show that admixture does seem to impact the phenotypic outcome, but just in specific cases.

To conclude, our findings reveal significant differences in the distribution and genetic causes of Hpa disease resistance in the native versus the introduced range of *A. thaliana*. We confirmed the founder's effect and show that the N. American colonizing lineage HPG1 is susceptible to the Hpa isolates tested. It is also the main driver of the observed distribution of phenotypic differences in the introduced range. We observed substantial gene flow in the forms of admixture and introgression from native populations into introduced ones. In particular, we observed differences in disease resistance at the haplogroup level and revealed contrasting admixture effects among haplogroups. There is evidence of introgression from a native-source *RPP4/5* resistant haplotype into the HPG1 genomic background. Together, these findings underline the importance of founder events and gene flow during biological invasions in disease resistance.

Materials and Methods

Host accessions dataset

We sampled twenty-five wild populations of *A. thaliana* in two N. American regions (Mid West and East Coast) and selected 480 accessions. Seeds were collected for propagation, and those seed-derived F1 accessions were used for the Hpa disease resistance screenings in the laboratory. Leaf material was used for sequencing and genotyping. We used available seeds from the Eurasian Arabidopsis Stock Centre (NASC) and the Arabidopsis Biological Resource Center (ABRC) belonging to the 1001 Genomes collection. We selected 405 accessions from thirty-five different countries. Only those accessions contained within the 762 high-quality genomes list were selected [154]. The full list with accessions metadata can be found in **Table S1**.

Pathogen isolates revival and propagation

Hpa infected leaves from wild populations of *A. thaliana* were collected in Eppendorf tubes and kept frozen at -80 °C. Hpa isolates were revived by placing infected *A. thaliana* leaves at 4°C for 30 minutes, then leaves were washed with sterile water to liberate spores. Hpa spore suspension was drop-inoculated onto eds1-1 (Ws-0) plants and kept in a percival for 7-10 days at 15 °C and 60% RH. Hpa spore propagation was done weekly by obtaining spore-containing water suspension from Hpa sporulating eds1-1 leaves, as previously stated [163]. 10-20 days old eds1-1 seedlings were spray inoculated with an airbrush kit and placed in a covered container with a lid sprayed with ddH_2O to maintain optimal Hpa growth humidity levels (60%RH).

Disease resistance assay

We used two Hpa isolates from wild infected *A. thaliana* samples from N. American populations for resistance screenings (**Fig 3F, arrows**): Hpa isolate 15IN55 was collected in the year 2015 from Indiana, INRCT population, plant number 55. Hpa Isolate 14OH04 was collected in the year 2014 from Ohio,

OHMLNP population, plant number 04. For the host growth, *A. thaliana* seeds were surface sterilized with an ethanol wash (75% EtOH 3 min, 90% EtOH 1 min) and stratified at 4°C for five days prior sowing to ensure synchronized germination. Seeds were sown out on trays containing 60 pots (Meyer, QuickPot QPD 60/5,5; pot dimensions: 47x55 mm) and grown at short-day photoperiod (8 h light, 16 h dark) under 50 µmol m-2 s-1 light fluence rate and 23°C in growth chambers. The experiment was replicated twice, using trays with 5 to 10 seedlings per accession in each pot. Accessions' pots position was randomized within and between trays. Hpa disease resistance screenings were done by adjusting Hpa spore concentration to 510^4 spores/mL of water and spray-inoculation of the spore suspension to 14 days-old seedlings. Col-0 accession showed incompatible interaction with both Hpa isolates. It, therefore, was used as a negative control for contamination, and the eds1-1 (Ws-0) mutant was used as a positive control of Hpa infection and screening conditions; both controls' phenotypes were validated microscopically to verify the lack of cryptic infection using Trypan Blue staining as described previously [164] (**Fig S2A**). Hpa disease resistance phenotyping was done seven days post-infection (dpi) by visually scoring sporulation on leaves. Accessions were classified according to the outcome of infection and the ability to recognize the Hpa isolates as a qualitative binary phenotype, either resistant or susceptible [165](**Fig S2B**).

Supplementary Methods

Genotyping

The host genetic variant file from the 1001 Genomes dataset was downloaded from the online catalog of A. thaliana genetic variation (1001 Genomes: https://1001genomes.org/data/GMI-MPI/releases/v3.1/). Those of the accessions belonging to the N. American dataset were genotyped using RAD-sequencing by first mapping to the TAIR10 reference genome using BWA-mem and then using the GATK haplotype caller v3.8 for calling SNPs [166]. The data were filtered for 10% missing data per SNP, no singletons, no transposable elements, and 2kb up and downstream of transposable elements. Imputation was done using BEAGLE [167] and a panel of WGS accessions from both the N. American and Eurasian dataset. The

variant file was filtered to keep only biallelic SNPs, for maximum missingness and minimum allele frequency for both datasets using VCFtools v.0.1.15 filters --min-alleles 2 --max-alleles 2 --max-missing 0.9 and --maf 0.03, respectively [168].

Genome-wide nucleotide diversity

We calculated genome-wide nucleotide diversity (π), the average pairwise difference between all pairs of accessions independently for both datasets with VCFtools v.0.1.15 --window-pi using a non-overlapping window's size of 100 Kb [168].

Principal Component Analysis

To obtain insights into the populations' genetic distance and internal structure within each dataset, we computed a PCA for both datasets independently based on the variance-standardized relationship matrix using PLINK. v.1.90b4.1 option --pca [169]. We chose to plot the first two principal component axes (PC1 and PC2) since they account for most of the variation in the data (**Fig S1A**).

Admixture

For the N. American dataset, we ran ADMIXTURE software v1.3.0 [153] for K = 1 to 16 in twelve independent replicates (12 different seed values). We assessed the most likely value of K by including a 5-fold cross-validation procedure. The software ADMIXTURE consistently estimated an optimal K = 10 (**Fig S1B**), with minimal decrease in CV error for K-11 (the difference explained between two consecutive Ks was less than 3.5%). We then re-run ADMIXTURE using K=10 and 2000 bootstraps to better estimate accessions admixture proportions. For the Eurasian dataset, ADMIXTURE values and classification were taken from a previous study [154], where K=11 was estimated as the optimal number of ancestral populations. Admixture proportions were used to classify each accession into an admixture group, using the proportion of the most abundant K group as criteria.

Genetic distances

We mapped the KBS-Mac-74 accession, which corresponds to the HPG1 reference genome (Michigan Kellogg Biological Station) [155], to TAIR10 using minimap2 and -cx asm5 flags, variants were called using paftools call command [170], and the following variant VCF file was filtered to keep only biallelic SNPs using VCFtools v.0.1.15 filters --min-alleles 2 --max-alleles 2. We use this VCF file to calculate pairwise genetic distances (1 - Identity-by-State) of each individual to the HPG1 reference genome using the PLINK v.1.90b4.1 --mdist flag [169]. We also classified each accession qualitatively based on their distance to the HPG1 in two categories for the Eurasian dataset (low <=0.50, high 0.5-1) and three categories for the American dataset (low <=0.20, medium 0.2-0.4, and high 0.4-1) and use the high and low categories for analysis.

Phenotype analysis

We used a 2D kernel density estimation, which is a nonparametric technique for probability density functions, to estimate the geographical density of Hpa disease resistance phenotypes. Therefore, it can predict phenotype probability density where no accessions are present in the geographical space. The interpretation of a 2D kernel density estimation plot is the average trend of what would be the scatter plot of accession phenotypes on the geographic map. For the N. American populations, we selected the dominant phenotype within that population for representation and analysis. For calculating the 2d kernel density estimation and visual representation, we used the stat_density_2d function implemented in the ggplot2 R package [171]. Default parameters estimated the contour bandwidth. Kernel density estimates for the distribution of phenotypes geographically and across HPG1 genomic distances were computed using the function geom_density from the ggplot2 R package [171]. Phenotype proportions were calculated based on the total number of resistant and susceptible accessions per dataset and haplogroup. We used a two-sample test for equality of proportions, prop.test() from the stats package in R, for testing the null hypothesis of equality of proportions for resistant accessions in both datasets and set the alternative hypothesis to “less,” meaning we accept the alternative

hypothesis that the probability of resistance is less in the American datasets compared to the Eurasian one (**Table S5**). We run a Wilcoxon test in R for testing the likelihood of resistance within a specific haplogroup (**Table S7**). To show each host admixture group's phenotype ratios, we used a ternary plot, which graphically depicts the ratios of three variables (**Table S6**); in this case, we used three Hpa disease resistance phenotypes; resistant, susceptible, and ability to differentiate between Hpa isolates. For visualization of the ternary plot, we used the ggtern R package and ggtern function [172]. We wanted to investigate the impact of population diversity in the phenotype; therefore, we calculated Pearson pseudo-R-squared two different measures of population diversity, the number of distinct haplogroups per population, and the genome-wide nucleotide diversity, with the number of susceptible and resistant accessions on those populations (**Table S8**). Genome-wide nucleotide diversity was calculated, as explained in the above "Genome-wide nucleotide diversity" methods section.

GLMs and correlations

We use generalized linear models to infer the effect of different predictor variables using R glm() function. For binary data, such as the full Hpa disease resistance phenotype, we use the glm binomial "logit" function against geographical predictor variables (latitude and longitude). For continuous data, such as the genetic distance to HPG1, we used the glm gaussian distribution. Model coefficients, p-values, and estimates with their corresponding standard variation were extracted with the summary() function and the jtools R package export_summ() function [173] (**Table S4**). The p-value shown in the graphs shows if there is a significant relationship described by the model for those predictor variables. The pseudo-R-squared values indicate how well the model explains the data.

Genome-Wide Association

To identify variants associated with Hpa disease resistance, we run a Genome-Wide Association analysis using the EasyGWAS web platform [174], selecting all the SNPs and using the EMMAX algorithm [175] that corrects for accessions relatedness computing a kinship matrix [176]. GWA analysis was done for each dataset

of accessions separately, one for the Eurasian accessions belonging to the 1001 Genomes dataset (1001 Genomes Data, using 2.327.646 SNPs after filtering) and another for our collection of American accessions (American imputed genomes, using 77.221 SNPs after filtering). The phenotype is used as a binary trait (resistant (0) vs. susceptible (1)). Both GWAS were run using the same parameters: Minimum Allele Frequency (MAF) of 3% (optimal QQ plots fit and AIC/BIC values), additive SNP encoding, no phenotype transformation, and TAIR10 gene annotation. The first two principal components of the SNP covariance matrix were used to correct simple forms of population structure in the US dataset because it displayed a more robust population structure than the EUR dataset. The GWA on the MISJCJT population for mapping resistance loci to the Hpa isolate 14OH04 was conducted using the same described above parameters but without MAF filtering, using 77 samples and 62.159 SNPs after filtering.

NLRs distance

We obtained a list of the *A. thaliana* NLR genes [52] with their genomic coordinates and calculated the genomic distance of each SNP to the closest NLR using bedtools closest -d flag [177]. Moreover, we counted the number of NLRs in the genome using a bin size of 100 Kb towards identifying and visualizing NLR cluster regions.

QTL mapping

Accession choice criteria for Hpa disease resistance QTL mapping was done by taking screened accessions as parents with opposing phenotypes (Resistant vs. Susceptible) for segregating the resistance loci. We selected six different lines from the Eurasian dataset that corresponded with three crosses for the F2 mapping offspring screened against Hpa isolates 14OH04 and 15IN55.

To map the resistance loci in the American population and further investigate Hpa disease resistance's admixture role, we selected accessions from the admixed population MISJCJT. Then we generated a backcross between a susceptible HPG1 individual with a resistant haplotype individual. We screened two F2 populations

derived from two different crosses of these same parental lines against the Hpa isolate 140H04 and merged them for analysis. Hpa disease resistance screening and phenotyping were done as stated in the disease resistance assay method section with the only modification of using a tray containing 240 pots (Meyer, QuickPot QPD 240/6; pot dimensions: 22x22x60 mm). For looking at phenotype segregation ratios we performed a Chi-square test comparing observed and expected frequencies of the number of resistant and susceptible F2-screened accessions (**Table S10**). DNA extraction was performed by collecting plant tissue in 2 mL screw cap micro tubes filled with garnet rocks (up to 0.5 ml) and deep-frozen in liquid nitrogen, stored in -80°C. Plants were ground with Fast-Prep-24 5G (MP Biomedicals) at speed 6 for 40s. 800 ul of prewarmed (55°C) Extraction buffer (100mM Tris pH8, 50 mM EDTA, 500 mM NaCl, 1,3% SDS, and 20 mg/mL RNaseA) was added an additional grinding was done at speed 6 for 40s. Samples were incubated for 10 min at 55°C, followed by centrifugation at 12,000 g for 1 min. 400 uL of lysate was transferred to a 96-well plate with 130 uL of KAc per sample and mixed in a plate mixer for 50 sec at 800 rpm, followed by incubation at 4°C for 5 min. Samples were centrifuged for 5 min at 6200 g. 300 uL of the supernatant was transferred to a fresh 96-well plate with 300 uL of SPRI beads and mixed for 1 min at 800 rpm. Samples were placed on a magnetic stand until the beads were bound to the side. The supernatant was removed, and two series of 80% Ethanol washes were done. DNA bound to beads was re-suspended in 50 uL of water for 5 minutes. The plate was placed on the magnetic stand, and the supernatant containing the DNA was transferred to a fresh plate.

Genomic DNA libraries were constructed using a modified version of the Nextera protocol [178], modified to include smaller volumes. Briefly, 0.25-2ng of extracted DNA was sheared with the Nextera Tn5 transposase. Sheared DNA was amplified with custom primers for 14 cycles. DNA from libraries was quantified using the fluorescent dye PicoGreen® kit on a TECAN plate reader before pooling. Three different library pools were made using ~ 50 ng per library into the pool (263,210 and 251 libraries per pool, respectively). Pools were cleaned for removing amplification primers using Sera-Mag Magnetic Speedbeads “SPRI beads” (GE) at a

1:1 ratio. To determine average libraries' size and nanomolarity, verify the lack of adapter contamination and suitable sequencing size, 0.5 ng of each of the pools was run on a High Sensitivity DNA Chip (Agilent Technologies) and measured with a 2100 Bioanalyzer (Agilent Technologies). Libraries ranged from an average size of 429 to 472 bp and 2.5 nM. Libraries were then sequenced on the Genome Center of the Max Planck Institute from Dev. Biology using Illumina HiSeq 3000 150 bp paired-end reads. Conversion to FASTQ, demultiplexing, and adaptor trimming was done with bcl2fastq2 version 2.18 from Illumina. To inspect read quality, MultiQC reports were generated with MultiQC version 1.3.dev0 [179] and included analyzed information from FastQC v0.11.5 and fastq_screen V0.5.2. The genetic map was created using only segregating SNPs among the parental lines. Simple Interval Mapping (SIM) was performed with the R package R/qtl [180]. Then, Multiple QTL mapping (MQM) was performed with a two cM step size and 100 Kb as the window size. One thousand permutations were applied to estimate genome-wide significance. For comparison discussion of Hpa disease resistance QTL loci, genomic ranges from previous Hpa QTL mappings were taken [64].

Phylogenetics

We extracted the SNPs from the RPP4/5 cluster, including the gene with the top GWA marker SNP, a genomic region of 148 Kb (Chr4:9488466-9636873). We included all 1135 accessions from the 1001G and the American dataset and filtered the VCF for missingness using vcftools max-missing 0.9 flag. In the end, the total genotyping rate was 0.97, and 2012 biallelic SNPs were used to generate a maximum likelihood (ML) phylogenetic tree using RAxML v. 8.1.3 [181] after removing identical sequences. We run the GAMMA model of rate heterogeneity and the GTR model of substitution. We inferred 20 different randomized MP trees on the SNPs alignment, selected the best-scored tree, and calculated the support values for tree's nodes and branches conducting 100 bootstraps and drawing bipartitions. To have a comprehensive view of the resulting phylogenetic tree, we collapsed the clades with an average branch length below 0.12 using iTOL [182]. Percent Identity Matrix of the

SNPs alignment was calculated using ClustalO [183]. The geographical location of each accession and its corresponding clade can be found in **Table S11**.

Supplementary Figures

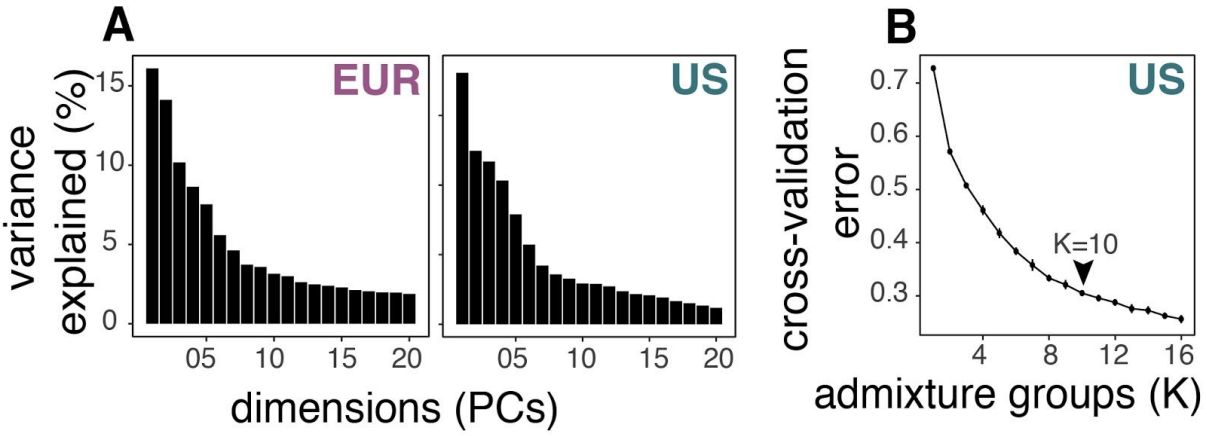


Figure S1. PCAs axis variance explained and admixture cross-validation error.
(**A**) Variance explained by each principal component (PC) dimension for each dataset. (**B**) Cross-validation error from admixture analysis for each admixture group cluster (K). Cross-validation error flattens at K10, therefore, selected as the optimal number of admixture groups for further analysis.

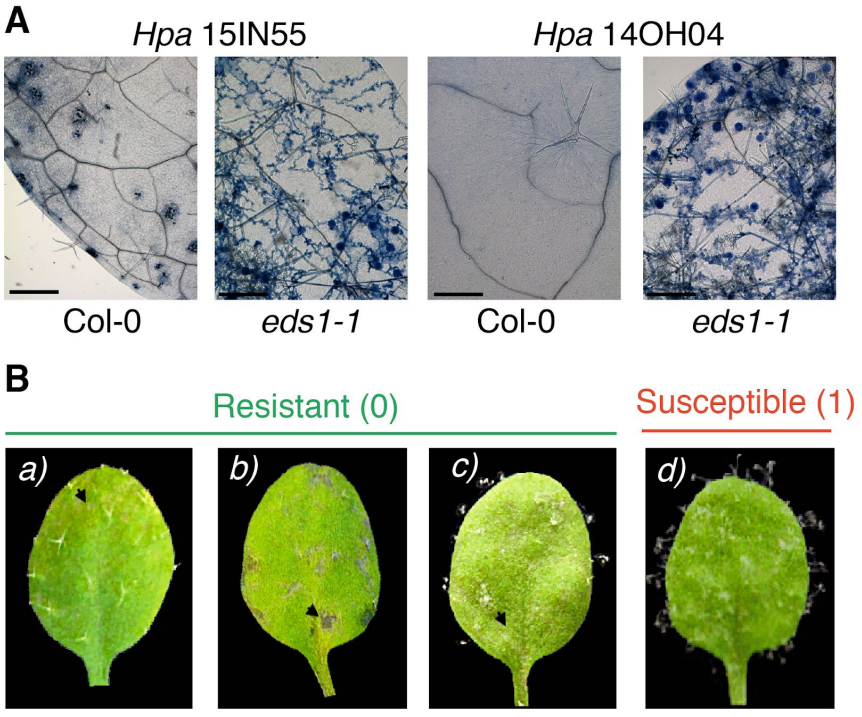


Figure S2. Infection controls and visual phenotyping criteria.
(**A**) Trypan blue staining of Hpa infected control accessions. Col-0 is used as a negative control of infection, and *eds1-1* is used as a positive control of infection and inoculation conditions, size bar 30 µm. (**B**) Plants were classified as resistant (0) if they were able to recognize the pathogen and trigger an immune response that caused three types of necrosis (black arrows, a) pitting necrosis, b) flecking necrosis, c) trailing necrosis) or susceptible (1) if the pathogen was able to successfully colonized the plant tissue and finished its reproductive cycle as evidence by profuse sporulation d).

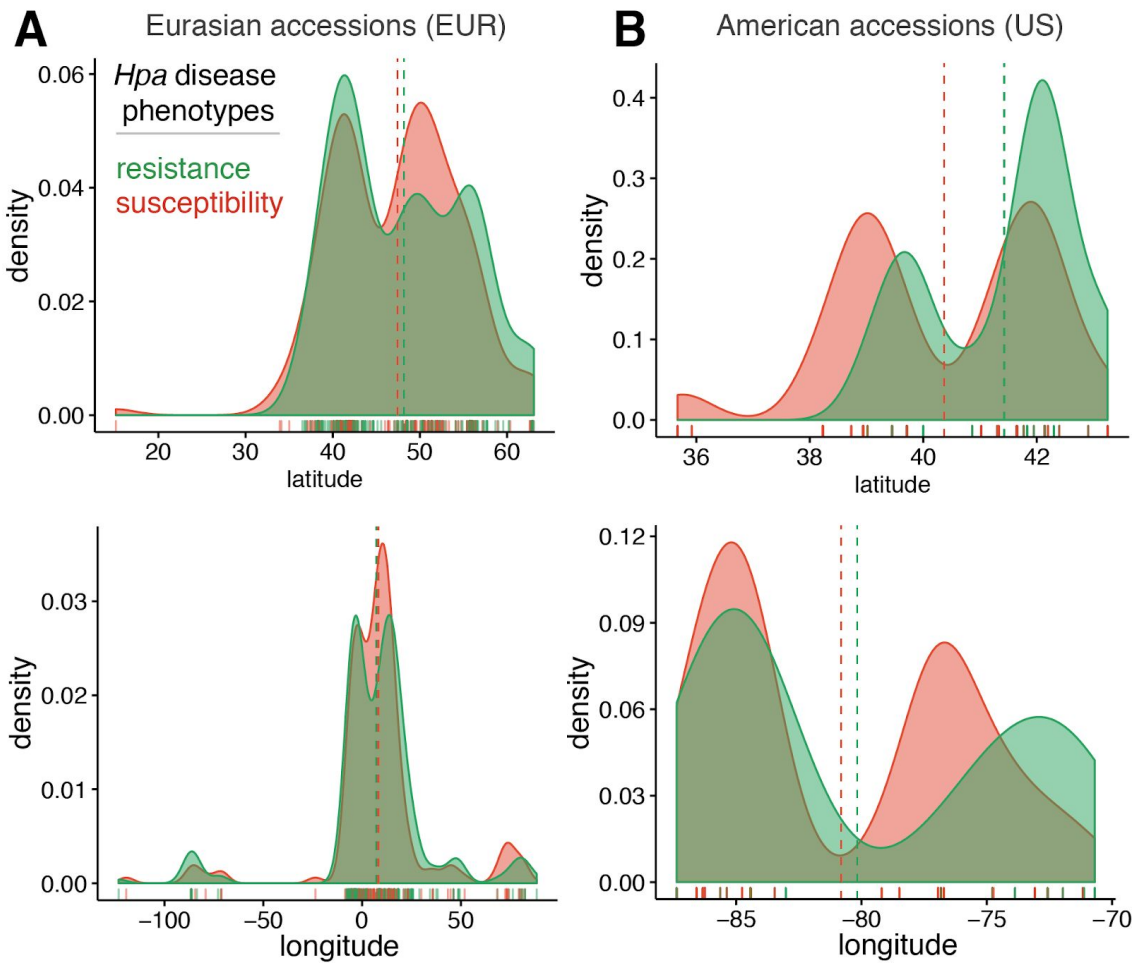


Figure S3. Hpa disease phenotype density by geographical coordinates.
(**A**) Phenotypic density of Eurasian accessions by longitude and latitude (**B**) Phenotypic density of American accessions by longitude and latitude. Hyphenated lines represent the mean density for each phenotype. Hyphenated lines are displaying the mean density of each phenotype distribution.

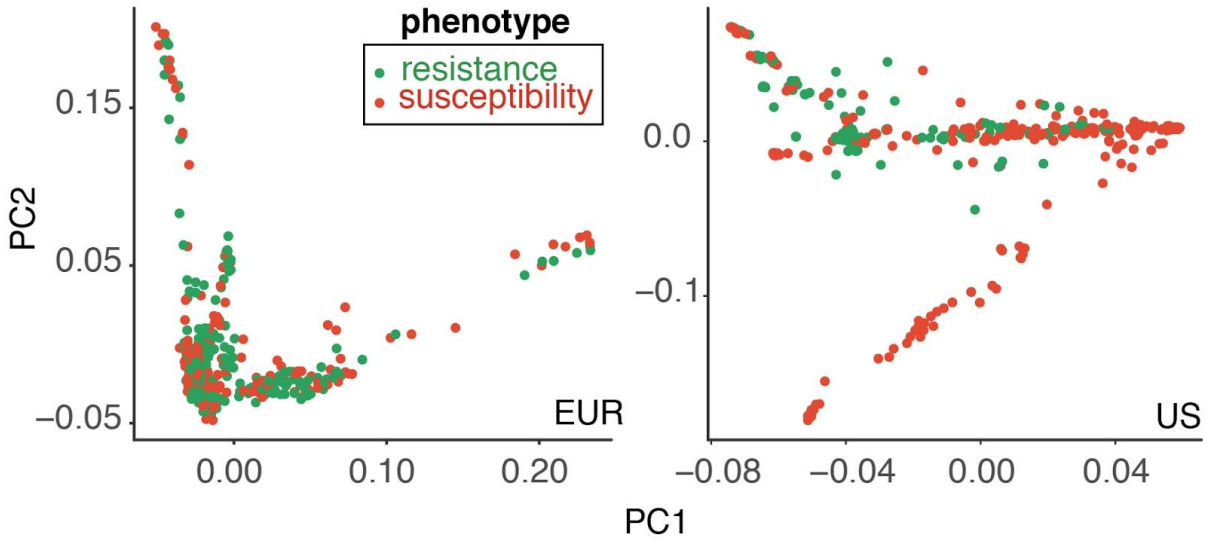


Figure S4. PCA of the *A. thaliana* kinship matrix.
Each point represents a single accession colored-coded by their *Hpa* disease resistance phenotype for each dataset. Binomial GLMs show a significant separation of Hpa disease resistance phenotype across both PC axes for the N.American dataset (**Table S4**).

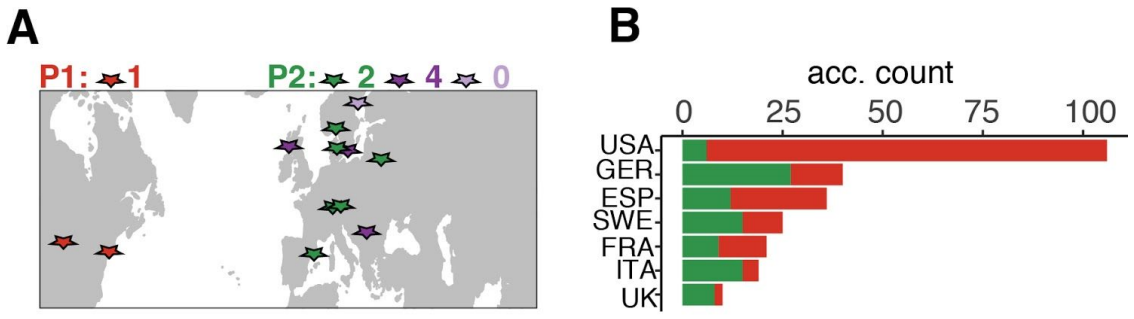


Figure S5. Complementary analyses to Figure 6.
(**A**) Top 10 Eurasian accessions with the highest PIM with the QTL parents P1 and P2 and their geographical origin. (**B**) Number of accessions in ML tree clades 1 and 2. clade 1 is the most abundant in one in the USA while clade 2 dominates in central and western Eurasian countries (UK, GER, SWE).

Supplementary Tables

Complete tables can be found in the digital version of the thesis.

S1. Accessions metadata, geographical variables, admixture proportions, HPG1 genetic distance, and PCA eigenvalues

S2. Admixture cross-validation error values

S3. Genome-wide nucleotide diversity (π)

S4. Generalized linear models (GLM)

S5. Proportion test for equality of proportions for Hpa disease resistance phenotypes.

S6. Phenotypic proportions for each haplogroup

S7. Wilcoxon test for admixture proportions and Hpa disease resistance phenotypes.

S8. Measures of N.American average population diversity and number of distinct haplogroups with their respective Hpa disease resistance phenotypes.

S9. GWA and NLR distance

S10. F2 mapping population segregation ratios and Chi-square test

S11. RPP4/5 clades for each accession and geographical distribution

Table S1. Accession information

Accessions information, geographical variables, Hpa disease resistance phenotypes

accessions information			geography		*Hpa* disease resistance phenotype		
IID	Dataset	Population	Latitude(º)	Longitude(º)	Hpa14OH04	Hpa15IN55	Full resistance
108	eur	FRA	48.5167	-4.06667	0	0	1
159	eur	FRA	47.35	3.93333	0	0	1
265	eur	FRA	44.65	-1.16667	0	0	1
630	eur	USA	40.7777	-72.9069	0	0	1
763	eur	KGZ	42.3	74.3667	0	0	1
765	eur	KGZ	42.1833	73.4	1	0	0
766	eur	KGZ	42.5833	73.6333	0	1	0
768	eur	KGZ	42.8	76.35	0	0	1
772	eur	TJK	37.35	72.4667	0	1	0
801	eur	USA	37.9169	-84.4639	1	1	0
870	eur	USA	41.8266	-86.4366	0	0	1
915	eur	USA	41.8972	-71.4378	0	0	1
932	eur	USA	42.3634	-71.1445	1	1	0
992	eur	SWE	55.3833	14.05	0	0	1
1002	eur	SWE	55.3833	14.05	1	0	0
1062	eur	SWE	55.7167	14.1333	1	0	0
1063	eur	SWE	55.7167	14.1333	0	1	0
1317	eur	SWE	59.5667	16.8667	0	0	1
1552	eur	SWE	63.0833	18.3667	0	0	1
1890	eur	USA	43.5139	-86.1859	0	0	1
2016	eur	USA	43.5356	-86.1788	0	0	1
2171	eur	USA	42.148	-86.431	0	1	0
2317	eur	USA	42.03	-86.514	0	0	1
4779	eur	UK	50.4	-4.9	1	1	0
4807	eur	UK	50.4	-4.9	0	0	1
5151	eur	UK	52.2	-1.7	0	0	1
5165	eur	UK	51.3	0.4	0	0	1
5577	eur	UK	54.7	-3.4	1	1	0
5644	eur	UK	54.4	-2.9	0	0	1
5741	eur	UK	56.6	-4.1	0	0	1
5768	eur	UK	54.1	-1.5	1	1	0
5772	eur	UK	54.1	-2.3	0	0	1
5779	eur	UK	51	-3.1	0	0	1
5784	eur	UK	56.4	-5.2	1	0	0
5800	eur	UK	57.4	-5.5	0	0	1
5811	eur	UK	52.9	-3.1	0	0	1
5831	eur	SWE	56.3333	15.9667	0	0	1
5865	eur	SWE	55.76	14.12	0	0	1
5950	eur	CZE	49.4112	16.2815	0	0	1
6009	eur	SWE	62.877	18.177	0	0	1
6011	eur	SWE	62.877	18.177	0	0	1
6013	eur	SWE	62.877	18.177	0	0	1
6024	eur	SWE	55.7509	13.3712	0	1	0
6025	eur	SWE	62.6437	17.7339	0	1	0
6038	eur	SWE	56.1	13.74	0	0	1
6040	eur	SWE	55.66	13.4	0	0	1
6042	eur	SWE	56.09	13.9	0	0	1
6073	eur	SWE	56.1481	15.8155	0	0	1
6074	eur	SWE	56.4573	16.1408	1	0	0
6094	eur	SWE	55.6494	13.2147	1	0	0
6104	eur	SWE	55.7	13.2	1	0	0
6108	eur	SWE	55.7989	13.1206	0	0	1
6114	eur	SWE	55.8097	13.1342	0	0	1
6118	eur	SWE	55.7	13.2	0	0	1
6122	eur	SWE	55.8364	13.3075	0	0	1

Table S2. ADMIXTURE Cross validation error values

The software ADMIXTURE consistently estimated an optimal K = 10
with very little decrease in CV error for K=11 (the difference explained between two consecutive Ks was less than 3.5%.)

Admixture cluster # (K)	mean	sd	porcent3.5	difference	diff_tested	
1	0.728	0.000	0.025	NA	NA	
2	0.571	0.000	0.020	0.157	no	
3	0.508	0.003	0.018	0.064	no	
4	0.461	0.008	0.016	0.047	no	
5	0.418	0.008	0.015	0.043	no	
6	0.384	0.006	0.013	0.034	no	
7	0.358	0.009	0.013	0.026	no	
8	0.333	0.005	0.012	0.024	no	
9	0.321	0.007	0.011	0.013	no	
10	0.305	0.004	0.011	0.016	no	selected
11	0.296	0.005	0.010	0.009	yes	
12	0.288	0.005	0.010	0.008	yes	
13	0.276	0.007	0.010	0.012	no	
14	0.273	0.006	0.010	0.003	yes	
15	0.263	0.005	0.009	0.010	no	
16	0.256	0.006	0.009	0.006	yes	

Table S3. Genome-wide nucleotide diversity (π)

EUR Dataset					US Dataset				
CHROM	BIN_START	BIN_END	N_VARIANTS	PI	CHROM	BIN_START	BIN_END	N_VARIANTS	PI
1	1	100000	1079	0.0019066	1	1	100000	55	0.000082
1	100001	200000	1278	0.0027073	1	100001	200000	56	0.000070
1	200001	300000	1570	0.0029627	1	200001	300000	27	0.000062
1	300001	400000	856	0.0017363	1	300001	400000	73	0.000181
1	400001	500000	1737	0.0031582	1	400001	500000	61	0.000176
1	500001	600000	1689	0.0032704	1	500001	600000	15	0.000046
1	600001	700000	1246	0.0021028	1	600001	700000	57	0.000138
1	700001	800000	957	0.0016513	1	700001	800000	66	0.000172
1	800001	900000	749	0.0013426	1	800001	900000	18	0.000039
1	900001	1000000	928	0.0017532	1	900001	1000000	21	0.000055
1	1000001	1100000	1040	0.0018472	1	1000001	1100000	51	0.000094
1	1100001	1200000	1503	0.0030001	1	1100001	1200000	8	0.000021
1	1200001	1300000	1223	0.0023494	1	1200001	1300000	78	0.000241
1	1300001	1400000	613	0.0012584	1	1300001	1400000	42	0.000052
1	1400001	1500000	1068	0.0020792	1	1400001	1500000	124	0.000280
1	1500001	1600000	1538	0.0031529	1	1500001	1600000	2	0.000002
1	1600001	1700000	913	0.0015878	1	1600001	1700000	15	0.000068
1	1700001	1800000	1361	0.0024159	1	1700001	1800000	50	0.000091
1	1800001	1900000	1104	0.0017054	1	1800001	1900000	1	0.000001
1	1900001	2000000	1029	0.0016309	1	1900001	2000000	39	0.000080
1	2000001	2100000	1121	0.0018469	1	2000001	2100000	67	0.000192
1	2100001	2200000	854	0.0016522	1	2100001	2200000	19	0.000057
1	2200001	2300000	1153	0.0024091	1	2300001	2400000	107	0.000236
1	2300001	2400000	1427	0.0027832	1	2400001	2500000	132	0.000225
1	2400001	2500000	1149	0.0018408	1	2500001	2600000	73	0.000182
1	2500001	2600000	1124	0.0021006	1	2600001	2700000	31	0.000070
1	2600001	2700000	843	0.0017665	1	2700001	2800000	286	0.000620
1	2700001	2800000	1475	0.002948	1	2800001	2900000	62	0.000058
1	2800001	2900000	1031	0.0021007	1	2900001	3000000	159	0.000285
1	2900001	3000000	1235	0.0023773	1	3000001	3100000	3	0.000002
1	3000001	3100000	882	0.0016521	1	3100001	3200000	254	0.000822
1	3100001	3200000	1190	0.0026081	1	3200001	3300000	601	0.001033
1	3200001	3300000	2223	0.0050909	1	3300001	3400000	144	0.000358
1	3300001	3400000	1999	0.0036555	1	3400001	3500000	18	0.000063
1	3400001	3500000	974	0.0017204	1	3500001	3600000	59	0.000119
1	3500001	3600000	1311	0.001886	1	3600001	3700000	43	0.000079
1	3600001	3700000	948	0.0017277	1	3700001	3800000	13	0.000065
1	3700001	3800000	1509	0.0026872	1	3800001	3900000	49	0.000172
1	3800001	3900000	1906	0.0035258	1	3900001	4000000	48	0.000043
1	3900001	4000000	1717	0.0025839	1	4000001	4100000	224	0.000683
1	4000001	4100000	1914	0.0037624	1	4100001	4200000	155	0.000406
1	4100001	4200000	1806	0.0032064	1	4200001	4300000	100	0.000216
1	4200001	4300000	1483	0.0030108	1	4300001	4400000	143	0.000244
1	4300001	4400000	2156	0.0039123	1	4400001	4500000	10	0.000041
1	4400001	4500000	1024	0.0019025	1	4700001	4800000	41	0.000095
1	4500001	4600000	1684	0.002937	1	4800001	4900000	11	0.000024
1	4600001	4700000	1220	0.0025489	1	4900001	5000000	162	0.000379
1	4700001	4800000	778	0.0013055	1	5000001	5100000	13	0.000054
1	4800001	4900000	1494	0.003303	1	5100001	5200000	248	0.000366
1	4900001	5000000	1116	0.0023601	1	5200001	5300000	17	0.000043
1	5000001	5100000	1534	0.002821	1	5300001	5400000	108	0.000167
1	5100001	5200000	1993	0.0043685	1	5400001	5500000	9	0.000013
1	5200001	5300000	661	0.0014677	1	5500001	5600000	281	0.001234
1	5300001	5400000	2013	0.0026655	1	5600001	5700000	150	0.000383
1	5400001	5500000	1080	0.0022337	1	5700001	5800000	117	0.000361
1	5500001	5600000	1429	0.0024799	1	5800001	5900000	88	0.000210
1	5600001	5700000	1505	0.0027429	1	5900001	6000000	3	0.000006
1	5700001	5800000	1349	0.002483	1	6000001	6100000	46	0.000065
1	5800001	5900000	1680	0.0032021	1	6300001	6400000	74	0.000137
1	5900001	6000000	1724	0.0026207	1	6400001	6500000	36	0.000057

Table S4. Generalized linear models (GLMs)

*** p < 0.001; ** p < 0.01; * p < 0.05.

1. Gaussian GLMs predictors for Hpg1 genomic distance

(A) Geography

Model = glm(hpg1_genomic_distance ~ latitude * longitude, data = dataset)

Model summary

	US	EUR
(Intercept)	0.29 ***	0.49 ***
latitude	0.03 ***	0.00 *
longitude	0.01***	0.01 ***
latitude:longitude	0.04 ***	-0.01 ***
	-0.01	0
N	480	405
AIC	-723.22	-1423.67
BIC	-702.36	-1403.65
Pseudo R2	-0.04	-0.01

Model estimates

estimate	std.err	lowerci	upperci	geo	p-value(z)	dataset
0.37920	0.05441	0.32479	0.43361	lat	***	us
-0.18260	0.02750	-0.21010	-0.15510	long	***	us
0.00452	0.00067	0.00384	0.00519	lat:long	***	us
0.00117	0.00033	0.00084	0.00150	lat	***	eur
0.00366	0.00060	0.00306	0.00426	long	***	eur
-0.00007	0.00001	-0.00008	-0.00005	lat:long	***	eur

2. Logistic GLMs predictors for full *Hpa* disease resistance

(A) Geography

glm(full_res ~ latitude * longitude, data = dataset, family = binomial(link="logit"))

Model summary

	US	EUR
(Intercept)	-0.97 ***	0.29 **
	-0.12	-0.11
latitude	0.85 ***	0.12
	-0.13	-0.11
longitude	0.21	-0.05
	-0.12	-0.13
latitude:longitude	0.28	0.04
	-0.14	-0.14
N	480	405
AIC	532.58	559.19
BIC	549.28	575.2
Pseudo R2	0.15	0.01

Model estimates

estimate	std.err	lowerci	upperci	predictor	p-value(z)	dataset
2.85327	1.24725	1.60602	4.10052	latitude	***	us
-1.15574	0.63349	-1.78923	-0.52225	longitude	0.0681	us
0.02936	0.01532	0.01404	0.04468	lat:long	0.0554	us
0.0151408	0.0161267	-0.00099	0.0312675	latitude	0.348	eur
-0.0104806	0.0294801	-0.03996	0.0189995	longitude	0.722	eur
0.0001833	0.0006664	-0.00048	0.0008497	lat:long	0.783	eur

(B) Principal Components of Host Kinship Matrix

Model = glm(full_res ~ PC1 + PC2, data =dataset, family = binomial(link="logit"))

Model summary

	US	EUR
(Intercept)	-1.19 ***	0.29 **
	-0.13	-0.1
PC1	-1.18 ***	-0.06
	-0.14	-0.1
PC2	0.64 ***	-0.04
	-0.15	-0.1
N	480	405
AIC	440.65	558.31
BIC	453.17	570.32
Pseudo R2	0.37	0

Model estimates

estimate	std.err	lowerci	upperci	predictor	p-value(z)	dataset
-25.7624	2.9627	-28.7251	-22.7997	PC1	***	us
13.9356	3.289	10.6466	17.2246	PC2	***	us
-1.2099	2.0069	-3.2168	0.797	PC1	0.5466	eur
-0.781	2.0136	-2.7946	-3.5756	PC2	0.6981	eur

Table S5. Proportion test for equality of phenotypic proportions

	res	total
US	139	480
EUR	232	405

```
resistant_ind <- prop.test(x = c(139, 232), n = c(480, 405),  alternative = "less")
# Printing the results
resistant_ind
    2-sample test for equality of
    proportions with continuity
    correction

data:  c(139, 232) out of c(480, 405)
X-squared = 71.228, df = 1, p-value < 2.20E-16
alternative hypothesis: less
95 percent confidence interval:
 -1.0000000 -0.2281195
sample estimates:
   prop 1    prop 2
0.2895833 0.5728395
```

Table S6. Ternary plot phenotypic proportions

	phenotypes%				
kgroup	diff	sus	res	size	dataset
1	17.142857	28.571429	54.285714	35	eur
2	33.846154	15.384615	50.769231	65	eur
3	26.086957	6.5217391	67.391304	46	eur
4	17.777778	22.222222	60	45	eur
5	29.411765	35.294118	35.294118	17	eur
6	31.25	18.75	50	16	eur
7	27.586207	17.241379	55.172414	29	eur
8	20	8.5714286	71.428571	35	eur
9	22.222222	25.396825	52.380952	63	eur
10	25.806452	16.129032	58.064516	31	eur
11	27.777778	0	72.222222	18	eur
1	27.777778	44.444444	27.777778	36	us
2	46.666667	33.333333	20	15	us
3	34.285714	40	25.714286	70	us
4	21.354167	50.520833	28.125	192	us
5	18.421053	71.052632	10.526316	38	us
6	16.666667	55.555556	27.777778	36	us
7	32.142857	32.142857	35.714286	28	us
8	33.333333	17.948718	48.717949	39	us
9	35.714286	28.571429	35.714286	14	us
10	1	50	50	12	us

Table S7. Wilcoxon test for admixture proportions and Hpa disease phenotype

EUR

kgroup	.y.	group1	group2	p	p.adj	p.format	p.signif	method
K1	prop	0	1	0.3981	1	0.398	ns	Wilcoxon
K2	prop	0	1	0.0408	0.45	0.041	*	Wilcoxon
K3	prop	0	1	0.3766	1	0.377	ns	Wilcoxon
K4	prop	0	1	0.0517	0.52	0.052	ns	Wilcoxon
K5	prop	0	1	0.6933	1	0.693	ns	Wilcoxon
K6	prop	0	1	0.0573	0.52	0.057	ns	Wilcoxon
K7	prop	0	1	0.3431	1	0.343	ns	Wilcoxon
K8	prop	0	1	0.1865	1	0.187	ns	Wilcoxon
K9	prop	0	1	0.6937	1	0.694	ns	Wilcoxon
K10	prop	0	1	0.8332	1	0.833	ns	Wilcoxon
K11	prop	0	1	0.6557	1	0.656	ns	Wilcoxon

US

kgroup	.y.	group1	group2	p	p.adj	p.format	p.signif	method
K1	prop	0	1	0.9272	1	0.9272	ns	Wilcoxon
K2	prop	0	1	0.8491	1	0.8491	ns	Wilcoxon
K3	prop	0	1	0.3178	1	0.3178	ns	Wilcoxon
K4	prop	0	1	0.2200	1	0.22	ns	Wilcoxon
K5	prop	0	1	0.4401	1	0.4401	ns	Wilcoxon
K6	prop	0	1	0.7361	1	0.7361	ns	Wilcoxon
K7	prop	0	1	0.0798	0.72	0.0798	ns	Wilcoxon
K8	prop	0	1	0.0085	0.085	0.0085	**	Wilcoxon
K9	prop	0	1	0.4985	1	0.4985	ns	Wilcoxon
K10	prop	0	1	0.1827	1	0.1827	ns	Wilcoxon

Table S8. Measures of N.American average population diversity and number of distinct haplogroups.

population	n_ind	n_sus	n_res	n_admix_groups	average_PI*1000
CTDERBY	9	9	0	3	0.1238942
CTPNR	17	5	12	5	0.1420491
CTRP	7	3	4	2	0.1218742
INRCT	28	28	0	5	0.1727818
INTHRR	23	13	10	3	0.1455254
INWIER	13	13	0	3	0.09279855
MAAA	6	0	6	1	0.04668964
MAMSSF	9	1	8	1	0.1085371
MAUR	12	11	1	2	0.04043171
MDPCB	44	44	0	4	0.108061
MDPGF	31	27	4	6	0.13339
MDSR	8	5	3	4	0.1527912
MDTCR	8	8	0	3	0.1324927
MIKOSP	12	8	4	5	0.1188129
MIMSP	27	13	14	5	0.1725809
MIMSUK	19	6	13	2	0.1690745
MISJCJT	77	52	25	5	0.1160059
NCARS	9	9	0	2	0.1423423
NCTBF	5	5	0	1	0.1501961
NJDCF	2	2	0	2	0.08354163
NJSC	39	27	12	3	0.1186249
NYBG	12	4	8	2	0.1487964
OHLAOBT	11	0	11	4	0.05139768
OHMLNP	42	40	2	5	0.1776123
OHPR	10	8	2	4	0.1401942

Table S9. GWA SNPs with distance to closest NLR

Only SNPs with –log10(P-value) > 2 and the distances for top 400 SNPs are displayed

chromosom	position	–log10(P-va	dataset	isolate	NLR distance kb
1	22613740	4.353	eur	14OH04	0
1	22613742	4.353	eur	14OH04	0
5	16618424	5.277	eur	14OH04	0
1	21743065	5.688	eur	15IN55	3.288
5	5956525	4.666	eur	15IN55	4.906
5	18109686	4.426	eur	14OH04	4.979
5	18423877	5.034	us	14OH04	5.169
3	19113386	6.303	eur	15IN55	8.421
1	22538477	4.781	eur	15IN55	12.854
5	16673628	4.755	eur	14OH04	14.997
5	16670519	4.333	eur	14OH04	18.106
5	16670457	4.333	eur	14OH04	18.168
5	14591003	4.565	eur	14OH04	18.373
1	21881411	4.796	eur	14OH04	20.872
4	10689152	4.579	eur	15IN55	31.87
4	10689184	4.579	eur	15IN55	31.902
5	21008764	4.532	eur	14OH04	33.848
5	21008790	4.532	eur	14OH04	33.874
5	26679074	4.333	eur	14OH04	35.689
5	19678171	4.387	eur	14OH04	38.304
5	19678158	4.505	eur	14OH04	38.317
5	19678136	4.505	eur	14OH04	38.339
5	19677899	4.505	eur	14OH04	38.576
5	19677889	4.505	eur	14OH04	38.586
5	19677821	4.505	eur	14OH04	38.654
5	19677763	4.399	eur	14OH04	38.712
5	19677603	4.399	eur	14OH04	38.872
5	19677257	4.505	eur	14OH04	39.218
5	19677227	4.505	eur	14OH04	39.248
5	19677128	4.505	eur	14OH04	39.347
5	19677072	4.399	eur	14OH04	39.403
5	19676994	4.63	eur	14OH04	39.481
5	19676928	4.523	eur	14OH04	39.547
4	6943604	4.762	eur	15IN55	44.474
4	6943605	4.762	eur	15IN55	44.475
4	9443481	4.732	eur	15IN55	44.984
1	24246120	5.119	us	14OH04	46.357
4	6754465	5.001	eur	15IN55	56.661
4	6750402	4.99	eur	15IN55	60.724

Table S10. F2 QTL mapping *Hpa* disease resistance segregation

			Chi-square test with Yates' correction for continuity					
Hpa isolate	Parents	Phenotype	Obs. pheno		Seg. ratio	Predicted interaction	X^2	P-value
			R	S				
14OH04	9405♀ x 8247♂	R x S	14	112	2 genes, 1: 15	dominant epistasis	3.9	0.0484
15IN55	9941♀ x 9971♂	R x S	98	7	2 genes, 1: 15	dominant epistasis	0	0.8339
15IN55	1063 ♀ x 6989 ♂	S x R	92	11	2 genes, 1: 15	dominant epistasis	3.6	0.0583
	MISJCJT population							
14OH04	29MI2014♀ x 28MI2015♂	R x S	160	79	2 genes, 2:1	dominant epistasis	0	0.891

Parents	QTL ranges
9405♀ x 8247♂	5, 6
9941♀ x 9971♂	1, 3
1063 ♀ x 6989 ♂	2, 4

Table S11. RPP4/5 Clades for each accession

IID	Clade	country	latitude	longitude	CS_number
9503	P8	UK	55.8877	-3.21072	CS76640
9544	P8	ESP	39.4	-5.33	CS76894
9539	P8	ESP	40.29	-6.67	CS76793
9880	P8	ESP	42.72	-3.44	CS77175
6943	P7	UK	51.4083	-0.6383	CS77126
7320	P7	FRA	49.4424	1.09849	CS76591
9606	P5	MAR	31.48	-7.45	CS76649
9550	P5	ESP	43.05	-5.37	CS76946
7063	P5	ESP	29.2144	-13.4811	CS76740
6008	P5	CZE	49.1	16.2	CS76824
9569	P5	ESP	42.87	-6.45	CS77166
7343	P5	GER	52.5339	13.181	CS76603
7319	P5	ITA	42	12.1	CS76590
9655	P5	ITA	38.92	16.47	CS77071
9733	P5	SVK	48.47	18.94	CS76697
15593	P5	AUT	48.331467	14.715867	CS78941
7250	P5	GER	51.9183	10.1138	CS76549
7236	P5	LTU	NA	NA	CS76543
8334	P5	SWE	55.71	13.2	CS77056
7013	P5	GER	52.4584	13.287	CS76445
7223	P5	GER	50.3833	8.0666	CS76541
8419	P5	LTU	54.6833	25.3167	CS78855
7161	P5	GER	53.5	10.5	CS76491
7424	P5	CZE	49.2	16.6166	CS76519
9901	P5	ESP	42.27	-2.98	CS78824
9666	P5	ITA	46.36	11.28	CS78909
9979	P5	ITA	46.36	11.23	CS76352
9669	P5	ITA	46.37	11.28	CS77086
9973	P5	ITA	46.36	11.28	CS76354
9667	P5	ITA	46.36	11.28	CS78910
9727	P5	GRC	37.63	21.62	CS77144
9594	P5	ESP	42.04	1.01	CS78837
9540	P5	ESP	41.81	2.34	CS76838
9557	P5	ESP	42.46	0.7	CS77102
9567	P5	ESP	42.34	1.3	CS77159
9899	P5	ESP	42.54	0.84	CS77342
9876	P5	ESP	41.34	0.99	CS77158
8420	P5	GER	50.0667	8.5333	CS76525
403	P5	CZE	49.3667	16.2667	CS78873
7477	P5	USA	41.7302	-71.2825	CS78853
9641	P5	RUS	51.9	80.06	CS77203
9640	P5	RUS	51.87	80.06	CS77202
9427	P5	SWE	62.8815	18.4055	CS77122
5860	P5	SWE	62.6814	18.0165	CS77913
6069	P5	SWE	62.9513	18.2763	CS77137

CHAPTER 2

Using target enrichment sequencing for population genetics of NLR and pathogenicity genes in wild *A. thaliana* samples

Declaration of Contributions

In the present chapter, I conceived the study with input from Detlef Weigel, Oliver Deutsch, Talia Karasov, and Sophien Kamoun. Gautam Shirsekar collected the North American visually infected plants used as biological material. Talia Karasov and Michael Giolai designed the *Pseudomonas* probes. Anna-Lena Van de Weyer provided the NLR isoform file from which the NLR probes were designed. Joe Win provided and adapted the secretome annotation pipeline used for Hpa effector discovery. Talia Karasov and Manuela Neumann performed the control infections for *Pseudomonas.* I performed the control infections for Hpa. I generated all the sequencing libraries, and Oliver Deutsch analyzed the sequencing data and assembled the Hpa genomes. Oliver Deutsch and I designed the Hpa probes. Oliver Deutsch filtered and analyzed the sequencing files for the shotgun and target enrichment sequencing. I generated the genotype files and analyzed the data. I managed the project, generated the figures, and wrote the manuscript with input from Oliver Deutsch and Talia Karasov. Detlef Weigel reviewed and edited the manuscript.

Author	Author position	Scientific ideas	Data generation	Analysis & interpretation	Paper writing
Alba González Hernando	1st	50%	40%	60%	70%
Oliver Deutsch	2nd	10%	20%	35%	10%
Gautam Shirsekar	3rd	—	25%	—	—
Talia Karasov	4th	5%	5%	5%	5%
Manuela Neumann	5th	—	2.5%	—	—
Michael Giolai	6th	—	2.5%	—	—
Anna-Lena Van de Weyer	7th	—	2.5%	—	—
Joe Win	8th	—	2.5%	—	—
Sophien Kamoun	9th	5%	—	—	—
Detlef Weigel	10th	30%	—	—	15%
Manuscript Title	“Using target enrichment sequencing to capture the diversity of NLR and pathogenicity genes in wild *A. thaliana* samples”				
Status in the publication process	To be submitted for publication				

Abstract

Plants and their pathogens constantly coevolve. Thus, they exert selection pressure on each other leading to reciprocal genetic change. At the molecular level, much of their interaction can be explained by the gene-for-gene model in which host resistance genes (R-genes) recognize pathogen effectors. Although molecular coevolution shapes the diversity of wild pathosystems, little is known about the evolution of R-genes and pathogen effectors. Therefore, describing the amount and distribution of diversity in these key molecular players is the first step towards understanding coevolutionary dynamics in natural plant populations. One of the main limitations in studying wild pathosystems is the sample complexity and the repetitive nature of R-genes and effectors, representing a minor portion of the genome. As a result, enrichment methods that reduce sample complexity have been developed, leading to the creation of R-gene enrichment sequencing (RenSeq) and pathogen-enrichment sequencing (PenSeq). Here we combine RenSeq and PenSeq to elucidate the distribution of presence/absence variation in natural populations of *Arabidopsis thaliana* (*A. thaliana*) infected with the oomycete pathogen *Hyaloperonospora arabidopsidis* (Hpa). We successfully estimated pathogen relative abundance using shotgun and target enrichment sequencing. In addition, we observed a wide range of infection levels between populations and related it to their population structure. We verified the enrichment of target organisms, described gene presence/absence variation, and recapitulated known host-pathogen compatibilities using reconstruction experiments. Also, we elucidated the distribution of presence/absence variation of the *A. thaliana* NLR-ome and Hpa effector-ome in natural populations. Finally, we observed that some NLR and effector genes at intermediate frequencies, while others were rare. Thus, we provide evidence for both arms race and trench warfare models of molecular coevolution. The possibility of capturing at large scale resistance and avirulence genes opens the door for molecular coevolutionary studies in natural plant pathosystems.

Introduction

The participants in natural plant pathosystems are engaged in a constant coevolutionary battle. Pathogens affect plant fitness by limiting growth and reducing seed production, while plants ward off pathogens, trying to evade infections. Therefore, there is reciprocal selection pressure for resistance on the plant side and for virulence of the pathogen side. Increasing interest exists in understanding plant-pathogen coevolution dynamics in natural environments by combining an ecological and population genetic approach [84]. In ecological studies, the geographic mosaic of coevolution theory expects a range of pathogen prevalence among populations and gene flow between them, maintaining polymorphisms. Complementary, population genetics investigates signatures associated with arms race dynamics (low genetic variation) and trench warfare dynamics (high level of polymorphisms at intermediate frequencies)[85].

Arabidopsis thaliana constitutes an excellent model pathosystem for coevolutionary studies because it is susceptible to a range of pathogens in its natural habitat. Obligate biotrophs, such as *Hyaloperonospora arabidopsidis* (Hpa), are considered *bona fide* pathogens of *A. thaliana* and thus in tight coevolution with its host [101]. Moreover, the interaction between *A. thaliana* and Hpa shows substantial genetic and phenotypic variation in wild populations, making it a perfect model pathosystem to study coevolutionary interactions [98,138]. Bacteria from the genus *Pseudomonas* are commonly found in wild populations, constituting the bacterial model pathogen of *A. thaliana* [184].

Both of these pathogens have been successfully used for investigating the molecular basis of the gene-for-gene model of interaction involved in Effector-triggered immunity (ETI). This model proposes a matching pair of effector and NLR alleles that ultimately determine the infection outcome. Therefore, ETI is the plant immunity layer involved in fighting host-adapted pathogens engaged in molecular coevolution. There are multiple instances of NLR and effectors under trench warfare coevolutionary dynamics, which is characterized by balancing polymorphisms. Balancing polymorphisms result from frequency-dependent

selection, heterozygote advantage, or spatio-temporal selection of alternative alleles [185].

Balancing selection of effectors and NLRs can be observed as presence/absence variation of these genes at intermediate frequencies among and within populations [186] [187]. They are the result of negative-frequency dependent selection, in which rare alleles are often the most advantageous ones, because there is little incentive for the partner to evolve mechanisms to counter the effects of these rare alleles. Moreover, having presence/absence variation in NLR genes can be evolutionary safer than mutations, avoiding the fitness cost associated with the absence of effectors and autoimmunity caused by divergent alleles[185].

A good example are two NLR genes that recognize *Pseudomonas* effector and that both show presence/absence variation in *A. thaliana* populations; these are *RPM1* and *RPS5*. Both of these genes confer fitness cost in the absence of the cognate effector, and they are found at intermediate frequencies in the global set of *A. thaliana* accessions [59,82,188–190]. In addition, fitness-growth tradeoffs and hybrid incompatibilities have been found for NLRs that recognize Hpa effectors [191] [52] . Because of the lack of Hpa population genetic studies, there is a missing link between the distribution and presence/absence variation of RPP genes and Hpa effectors.

From the pathogen side, there are numerous examples of oomycetes displaying presence/absence polymorphisms of effector genes, including *Phytophthora* effectors *Avr1d*, *Avh245*, and *PiAVR2* [192–194]. Conversely, *ATR1* is the only identified Hpa effector that showcases presence/absence variation [195]. Due to the dynamic nature of bacterial genomes, the presence/absence of effectors is common and well documented. For instance, the effector repertoire and presence/absence variation from *Pseudomonas* species are well known [196–198] .

Identifying the presence/absence variation of many genes per sample is nowadays possible thanks to target enrichment sequencing methods[115]. Population genetic studies of presence/absence variation of NLR and effector genes have been independently validated using target enrichment sequencing [121,122,124] [160]. However,

there is a lack of simultaneous analysis of presence/absence variation in host NLRs and pathogen effectors and their distribution in natural populations, which is key to finding coevolutionary signatures and ultimately understanding the coevolutionary dynamics of wild pathosystems.

In this study, we combined Pathogen-enrichment sequencing (PenSeq) and R-gene enrichment sequencing (RenSeq) to simultaneously capture the effectorome and NLRome of infected *A. thaliana* samples from wild N. American populations. This, coupled with shotgun sequencing and population genetic analysis, allowed us to interrogate the distribution of presence/absence variation of key disease resistance genes and effectors. Moreover, we gathered evidence for arms race and trench warfare dynamics for different NLRs and effectors. The combination of PenSeq and RenSeq (PRenSeq) opens the door for large-scale coevolutionary studies in wild plant pathosystems.

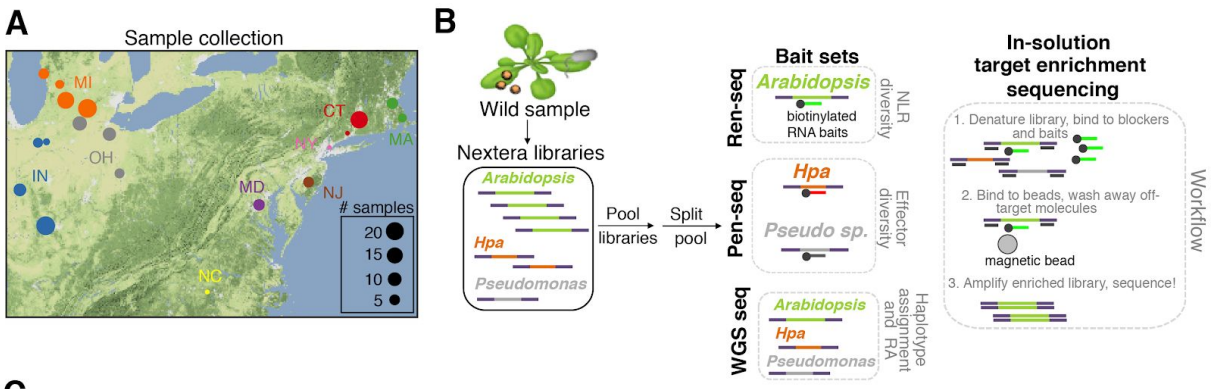
Results

Combination of target enrichment sequencing and shotgun sequencing for population genetic studies of infected *A. thaliana* samples and their pathobiomes

In order to have a representation of the N. American diversity, we collected multiple *A. thaliana* leaves that were visually infected with Hpa during three consecutive years in twenty different populations (**Fig 1A**). We included a set of control *A. thaliana* samples grown and infected in laboratory conditions to validate the capture protocol (**Figure S1**). We used *A. thaliana* accessions Col-0 and the *eds1-1* (Ws-0) mutant as controls (compatible vs. incompatible interactions). To deduce the pathobiome, we inoculated plants with five Hpa isolates (14OH04,15IN55, Emoy2, Cala2 and Waco9), two *Pseudomonas* strains (DC3000 and p13.g4), or water (mock) and collected samples after the infection onset. For population genetic analysis, each sample was enriched and sequenced following the target enrichment protocol and shotgun sequencing (**Fig 1B**). Shotgun sequenced samples were used to identify population structure and infection levels. Target

enrichment sequenced samples were used to assess infection levels, distribution, and presence/absence of targeted genes.

RNA baits were used to enrich samples for host R-genes (RenSeq) and pathogen genes from Hpa and *Pseudomonas* (PenSeq) (**Fig 1C**). In brief, the *A. thaliana* bait set was designed to target 13,167 NLR genes coming from sixty-four different accessions[160]. The Hpa bait set included as targets annotated effectors and other pathogenicity genes from previous publications [108,195,199–202]. Because the majority of effectors comes from the Emoy2 reference genome, we *de-novo* annotated additional assemblies from British Hpa isolates and de-novo assembled N. American isolates (**Table S1 and S2**). Annotated proteins were fed to a custom-built secretome prediction pipeline. Moreover, we included putative housekeeping genes as controls that should show much less variation between samples (**Table S4**). The final Hpa bait set included 2,504 genes from a total of fifteen different Hpa isolates. The *Pseudomonas* spp. bait set consisted of 372 pathogenicity genes annotated from 1,524 strains collected from *A. thaliana* in Southwest Germany [198]. Baits from capture enrichment can hybridize with target sequences that have up to 80% sequence identity. Thus, we first clustered the target sequences and then mapped the reads to genes representing sequence clusters with 80% or more sequence identity (**Table S5**). The total number of reference genes used for mapping were 1,855 for Hpa, 589 for *A.thaliana*, and 220 for *Pseudomonas*.



C

Bait set	# of strains	Target genes	Source	# of sequences for baits	# of ref. seqs for mapping	Target Mb (including Ns)	Bait length (nt)	Tiling (X)	Baits (Mb)
A. thaliana	64	NLRs	Van de Weyer et al.(2019)	13,167	589	32,3	80	3	6,3
Hpa	15	Secretome Housekeeping Evol. signatures	multiple	2,504	1,855	2,6	80	1,5	3,2
Pseudomonas sp.	1,524	pathogenicity determinants	Karasov et al.(2018)	372	220	5	70	2 (50 nt apart)	5,8

Figure 1. Sequencing strategy and bait library design.

(**A**) Geographic origin of the sample collection, color-coded by state and number of samples in each population. (**B**) Sequencing strategy from library preparation to target enrichment sequencing workflow. (**C**) Bait library design used for in-solution target enrichment sequencing.

Arabidopsis thaliana samples from wild populations have a wide spectrum of pathogen abundance

We calculated the relative abundance of *A. thaliana*, Hpa, and *Pseudomonas spp. reads* in each sample by mapping reads to a multi-reference genome, including two *A. thaliana* host accessions, the classically used reference genome TAIR10 and the HPG1 reference genome representing the majority of North American diversity KBS-Mac-74 [155], six high-quality Hpa assemblies; Cala2, Emoy2, Noks1, 15IN54, 15IN55, 14OH04, and 1,524 *Pseudomonas* assemblies [198].

Although the host phyllosphere's microbiome composition is complex, we know that *Pseudomonas* infections are common, with a single lineage dominating across Southwestern Germany populations (OTU5)[198]. Thus, we went on to identify the *Pseudomonas* strains present in our samples and their prevalence. We

determined this by mapping reads to 1,524 *Pseudomonas* assemblies and then looking at the strains that had the highest number of reads mapped in each sample (**Table S7 and Fig S2**). *P. viridiflava* was the most common bacterial taxon in our dataset, having the highest number of reads mapped in most of our samples (78%). The OTU5 pathogenic lineage was the most abundant OTU in our sample set (67%, 100/149), agreeing with previous reports [198]. The most common isolate was p7.E10 in our sample set (64%).

We calculated relative abundance (RA) as the fraction of reads that mapped to each of our target organisms (*A.thaliana* , Hpa, and *Pseudomonas*) [198]. After read mapping and quality filtering, 149 out of 150 samples had enough mapped reads to calculate relative abundances (**Fig 2A**). The RA of *A. thaliana* ranged from 9% to 96%, with an average of 41%. The RA of Hpa ranged from 2% to 90%, with an average of 42%. Since we visually observed a few samples co-infected with *Albugo* in our sample collection, we wanted to measure the extent of these co-infections. Therefore, we included in our multi-genome reference twenty-three genomes from other oomycetes (Ooo) (**Table S6**). The RA of other oomycetes ranged from 0.25% to 40%, with an average of 9%. This suggests that other oomycetes usually co-infect *A. thaliana* leaves when Hpa is present. Finally, the RA of *Pseudomonas* ranged from 0.08% to 34%, with an average of 7%. Taken altogether, the proportion of reads that mapped to pathogen genomes was, on average, higher than the proportion of reads mapping to the host. This should be expected from samples with visual levels of infection.

When oomycetes infect *A. thaliana*, these can promote the growth of other pathogens. For instance, some HaRxLs effectors enhance *Pst* bacterial growth by suppressing the PTI immune response [203]. In addition, some *Albugo* species can suppress non-host resistance [204], allowing non-compatible Hpa isolates to grow and changing the microbial composition of *A. thaliana* leaves [205]. To test the hypothesis that co-infection might result from pathogens' presence, we computed a pairwise correlation between each organism's relative abundance (**Fig 2B**). We found a negative correlation between the relative abundance of *A. thaliana* and all the other pathogens. Moreover, we observed a substantial negative correlation between Hpa

and *A. thaliana* (r= -0.74), as expected for a biotrophic pathogen colonizing the plant tissue and causing host cell death. On the other hand, we found a very high positive correlation between the relative abundances of other oomycetes and *Pseudomonas* (r= +0.76). This supports the idea of oomycetes promoting bacterial growth and leading to co-infections. On the other hand, we did not see an interaction between Hpa and other oomycetes.

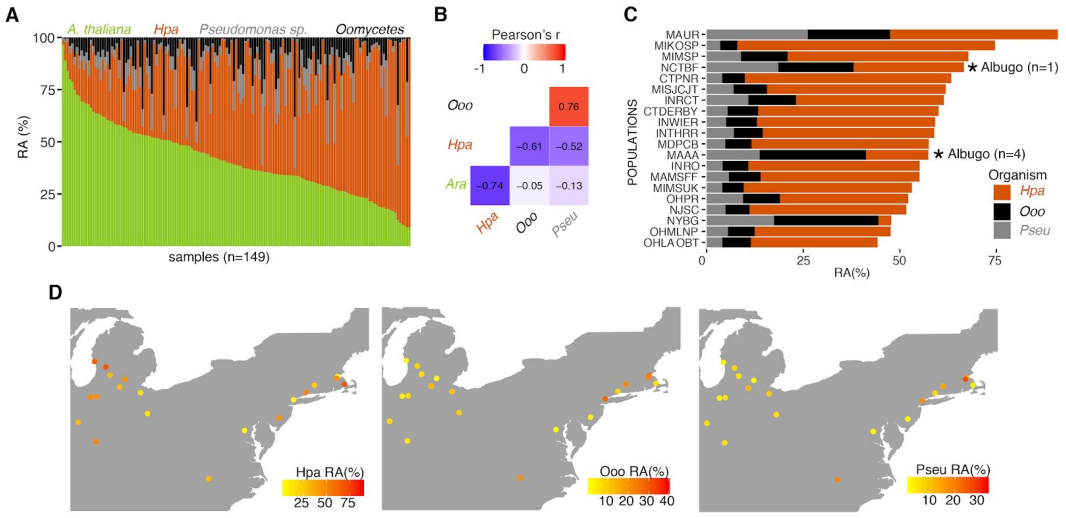


Figure 2. Distribution of relative abundances from each taxon in whole-genome sequenced samples.

(**A**) Relative Abundance (RA) as a fraction of total reads mapped to each organism per sample, color-coded by organism. (**B**) Correlation of RA for each pairwise organism combination, displaying and color-coding the Pearson's correlation coefficient. (**C**) Distribution of RA from pathogens per N. American population, populations with samples visually infected with *Albugo*, are denoted with a star "*." (**D**) Geographic distribution of RA for each pathogen taxon, color-coded by RA percentage.

Since we are observing a wide abundance range from pathogens, we were curious to see their distribution among and within the different sampled populations (**Fig 3C and D**). The highest Hpa infection levels were seen in populations from Michigan and Connecticut, ranging from 66% to 46%. The lowest Hpa infection levels were found in two East Coast populations, MAAA and NYBG, with Hpa relative abundances of 16% and 3%. In contrast, populations from the East Coast had the highest infection levels of *Pseudomonas* and of other oomycetes, ranging from 26% to 13% for *Pseudomonas* and 27% to 19% for other oomycetes. Populations with high Hpa infection levels do not overlap with those of *Pseudomonas* and other oomycetes. This lack of overlap could result from niche

competition between Hpa and other oomycetes. Niche competition has already been observed for *Albugo,* which has a competitive growth advantage over Hpa [206].

A. thaliana and Hpa haplogroups undergo admixture but display different levels of population structure

The population structure of *A. thaliana* in its native and introduced range has already been interrogated [49]. Still, nothing is known about Hpa and *Pseudomonas* population structure and how it maps to its host. Therefore, we mapped our samples to one reference genome for each organism (TAIR10 for *A. thaliana*, 14OH04 for Hpa, and p7.E10 for *Pseudomonas*). We used the Bayesian variant detector Freebayes [207] because it exploits population information to call variants confidently. After performing variant quality filtering, we obtained 81,942 SNPs for *A. thaliana*, 59,732 SNPs for Hpa, and 1,294 SNPs for *Pseudomonas*. The amount of *Pseudomonas* variants was insufficient to estimate its population structure, so we only proceeded with *A. thaliana* and Hpa. To reveal the number of putative ancestral haplogroups in our samples, we used the allele-frequency based ADMIXTURE software [153]. We estimated two ancestral haplogroups for Hpa and six haplogroups for *A. thaliana* as having the lowest cross-validation error (**Fig 3A**). We also observed different levels of admixture between these haplogroups (**Fig 3B and C**). The CTPNR population showed a very distinct population structure, having most of the haplogroup two samples from Hpa and haplogroup six samples from *A. thaliana* (**Fig 3B**). Several populations had samples with both Hpa haplogroups, but it did not correlate with specific *A. thaliana* haplogroups. To have a better picture of the samples' diversity distribution, we computed a kinship matrix-based PCA (**Fig 3C**). There was a strong host population structure with admixture between distinct haplogroups, but we did not observe strong Hpa population clusters.

Finally, we wanted to know whether some host haplogroups are more susceptible to pathogens than others and therefore have higher infection levels. Hence, we looked at the relationship between each haplogroup and the relative abundance of pathogens (**Fig 3D**). Samples with admixed Hpa haplogroups had lower Hpa infection levels than samples infected with pure Hpa haplogroups. Host

samples from haplogroup 4 had considerably higher Hpa infection levels than admixed hosts.

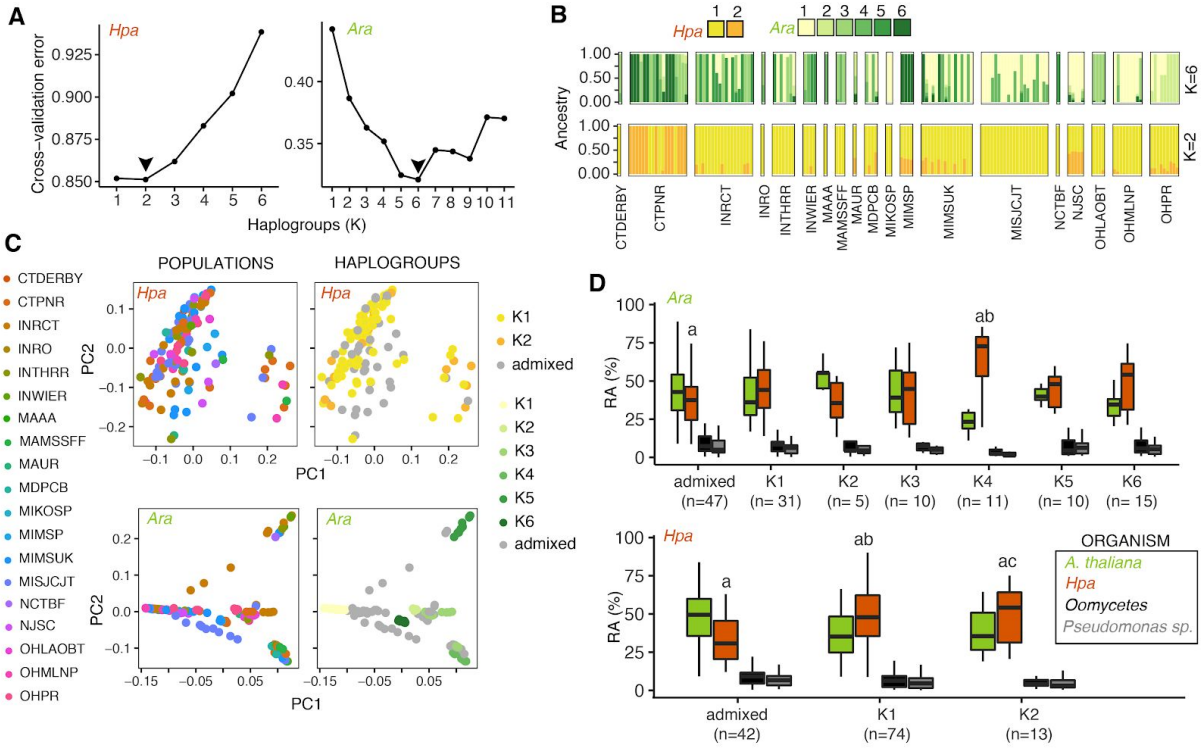


Figure 3. Population structure of *A. thaliana* and Hpa whole-genome sequenced samples.

(**A**) Cross-validation error from ADMIXTURE analysis for each tested ancestry haplogroup (K). Arrows show the selected number of optimal haplogroups (**B**). Ancestry proportions for each sample are displayed by vertical bars, boxed-grouped by population, and color-coded by haplogroup. (**C**) PCA of samples kinship matrix for *A. thaliana* and Hpa color-coded by haplogroup and population. (**D**) Relative abundance of each organism grouped by *A. thaliana* (top) or Hpa (bottom) samples' haplogroup. A Tukey HSD, pairwise comparison test, was done, groups statistically significant (p-value < 0.05) are denoted with letters.

Reconstruction experiments show successful target enrichment and recapitulate known host compatibilities

Samples collected from the wild can be considered metagenomic samples since they contain a complex mix of colonizing microbes such as bacterial communities, fungi, viruses, and oomycetes. We wanted to ensure that baits capture only the desired organisms and avoid retrieving DNA sequences that belong to any

other colonizing microbes. Thus, we carried out a sequence homology search of our bait sequences to the BLAST database and removed those with high homology with other microbes. Moreover, we looked for pairwise sequence similarity between our three target organisms and removed redundant baits among the three bait sets. Using this filtering strategy, we ensured the unique and non-overlapping capture of genes present in each bait set and organism.

For the target enrichment protocol, each sequencing library was split into three tubes, one for each bait set. Then, libraries were re-pooled for sequencing at different ratios for each organism (**Fig 1B**). For samples coming from the reconstruction experiment, we multiplexed twenty-two samples per capture reaction and per pool (control pool; 18 control samples and four German samples as library concentration input test). We obtained an average sequencing yield of ~ 4 Gb, ranging from ~1 Gb to ~12 Gb (**Fig S2**). We multiplexed 30 samples per capture reaction and pool for wild samples, achieving an overall sequencing yield that we considered optimal (**Fig S2**). Reads that were demultiplexed and QC filtered before being mapped to the target reference sequences were considered high-quality reads and used for further analyses (PRenSeq reads; Pen+Ren-seq reads). We used the PRenSeq reads to calculate the percent on-target reads by mapping them to the target reference sequences. Our definition of on-target is used in a non-standard way since we only mapped the reads back to the reference sequences and not to complete reference genomes. We later filtered the mapped reads by mapping quality score, and it, therefore, changed the final percent on-target values. Mismatches caused by variants (SNPs, INDELS, etc.) affect the mapping quality score. Considering the polymorphic nature of the target genes and the presence of only one reference gene per cluster, we decided to report both metrics.

To assess how well we captured the desired organisms, we first analyzed the control samples from the reconstruction experiment. We achieved a high fraction of on-target reads in control samples, ranging from 7% to 40% (**Fig 4**). The highest percentage on-target was achieved for Hpa infected samples, and the lowest for mock-treated samples. We observed a very low percentage of reads in each sample that map to Hpa, even if a sample had not been infected with Hpa (0.4% to

0.6%, **Fig 4; first column**). When we filtered the reads by mapping quality, these numbers decreased to an almost negligible amount, ranging from 0.02% to 0.07%, showing no substantial off-target mappings. *Pseudomonas* background mappings in non-infected samples were much lower than for Hpa (0% to 0.06%, **Fig 4; third column**) and almost non-existent after filtering for mapping quality (0% to 0.04%). I could recapitulate known compatibilities because there were clear differences in the percent of reads on-target for the inoculated organisms in compatible vs. incompatible hosts. I could also reproduce known differences in pathogenicity levels for Hpa isolates. For example, Waco9 had the second-highest percent of reads on-target, and it is known to be a highly virulent isolate [108]. Also, we know from our study that Col-0 allows for the marginal growth of the Hpa isolate 15IN55, causing cell death (**Chapter 1, Fig S2**). In agreement, Col-0 infected with 15IN55 had the highest percent of reads mapping to Hpa within the incompatible interactions category. The same observation applied for *Pseudomonas* isolate p13.g4, which is known to be less virulent than DC3000 [198]. In cases where the reaction was compatible, the fraction of on-target reads for the host (1% to 6%) was lower than when it was incompatible (6% to 9%) (**Fig 4; second column**). This is not unexpected because highly infected samples undergo cell death, and therefore the relative abundance of host DNA decreases significantly.

The percent of on-target reads in wild samples ranged from 2% to 17%. Similar to what we observed in the reconstruction experiment, Hpa had the highest percent of on-target reads (0.48% to 16%), followed by *A. thaliana* (0.7% to 8%), and finally *Pseudomonas* (0% to 4%). One explanation for why fewer Hpa reads were mapped could be that infection levels in the wild are lower than in laboratory settings. Moreover, the target genes were by design present in the Hpa isolates used in the reconstruction experiment while they might not be present in the wild isolates with a priori unknown genetics.

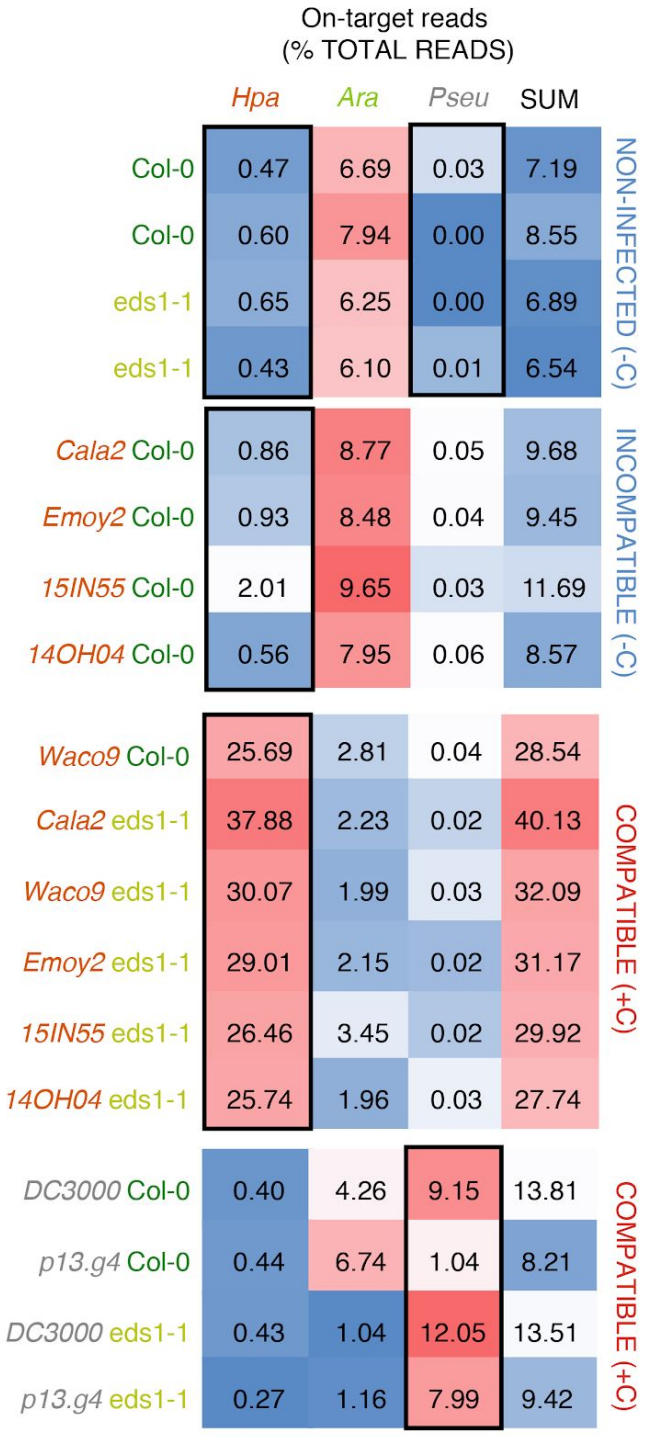


Figure 4. Percent of total PRenSeq reads mapped on-target in control samples

A. thaliana accessions used as genotype controls (Col-0 and *eds1-1*) infected with single Hpa isolates or *Pseudomonas* strains for positive infection controls (+ve C). Negative controls (-C) were non-infected plants grown alongside infected plants. When the genotype allows pathogen growth is denoted as a compatible interaction, when the given strain cannot grow on a given host genotype, the interaction is indicated as incompatible.

Estimates of pathogen infection levels are comparable between TES and shotgun sequencing

It has been previously proposed that the relative abundance of pathogen DNA can be estimated using PenSeq [121]. We aimed to compare pathogen infection levels between shotgun sequencing data, which provide a direct measurement of both microbial and plant DNA and therefore a DNA-based measure of pathogen load[127], and TES (**Fig 2 and 5**). We observed a comparable range of infection levels, although the sample distribution was skewed towards samples with high pathogen levels for Hpa levels, and low pathogen levels for *Pseudomonas* (**Fig 5A**). This could be because the number of Hpa genes used for mapping was higher (1855) than for *A. thaliana* and *Pseudomonas* genes (589 and 220). We then looked at the distribution of on-target reads by population (**Fig 5B**). The population with the highest fraction of on-target reads from Hpa was NYBG, with an average above 75%. This finding was at variance with what had been observed with shotgun sequencing, where NYBG samples showed low Hpa infection levels (**Fig 2C**). There was nevertheless a positive correlation between the fraction of on-target reads and WGS measured abundance of Hpa (**Fig 5C**).

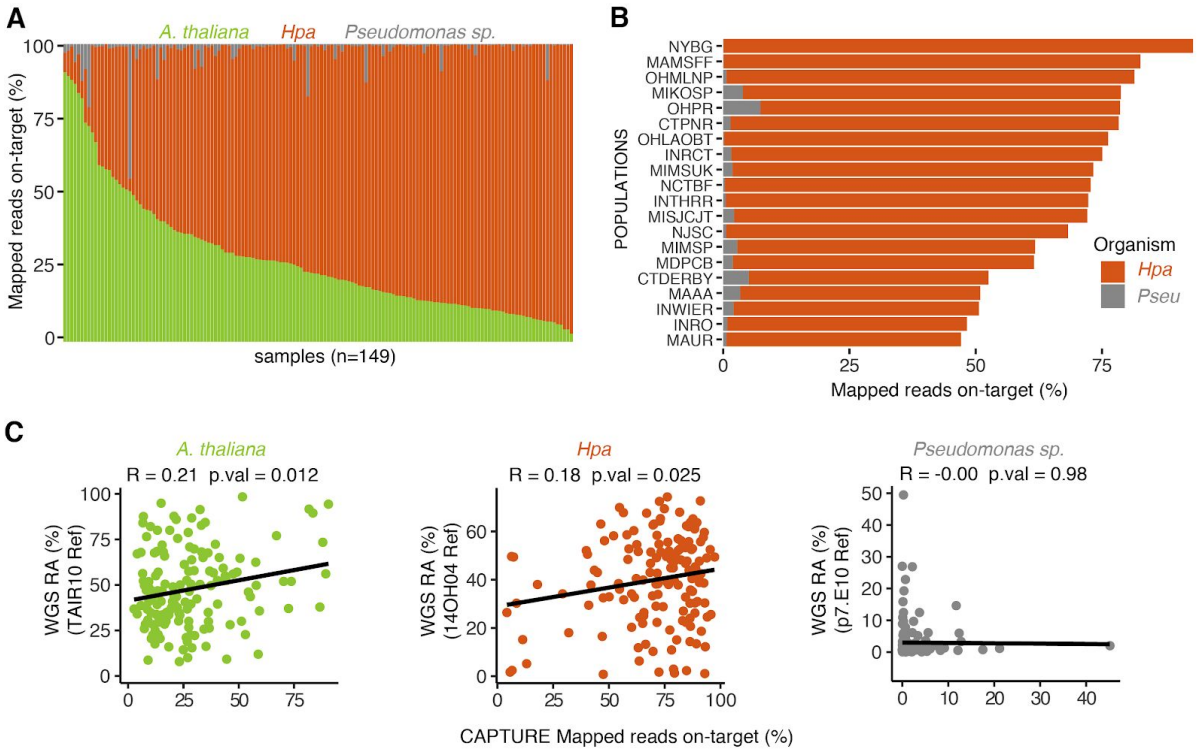


Figure 5. Distribution of PRenSeq reads mapped on-target and their relationship with relative abundances from shotgun sequenced samples.

(**A**) Percentage of PenSeq reads mapped on-target for each organism per sample, color-coded by organism. (**B**) Distribution of on-target reads from pathogens per N.American population. (**C**) Correlation between relative abundances (RA) estimated from WGS and on-target read fractions in PRenSeq for each organism. Correlation coefficients and p-values are displayed.

Reconstruction experiments recover expected presence/absence polymorphisms in key NLR and pathogenicity genes

After confirming the target species' enrichment, I determined each gene's presence in our samples based on gene coverage, mean read depth, and the number of mapped reads. These are the criteria that have been used in previous target enrichment sequencing studies [121,122].

I first focused on the P/A of genes in the reconstruction experiment (**Fig 6**). Since I used two *A. thaliana* accessions (Col-0 and Ws-0), I expected to see two distinct clusters of NLRs present for each accession. By looking at the P/A matrix, I could clearly distinguish between Col-0 and Ws-0 associated NLRs (**Fig 6B**). I saw

some variability between samples, mostly driven by infected samples having lower amounts of host DNA, which increased the likelihood of missing the capture of all genes in these samples

Regarding the pathogens, I confirmed the absence or near absence of reads mapping to the target Hpa and *Pseudomonas* genes in the negative control samples, which had not been inoculated (**Fig 6A and C**). Only one of the negative control samples had four *Pseudomonas* genes captured (**Fig 6C**). Considering that *Pseudomonas'* negative controls were grown next to *Pseudomonas* infected samples, there could have been slight contamination and *Pseudomonas* growth in this control sample. Another plausible explanation is the presence of *Pseudomonas* in the soil where the plants were grown leading to spontaneous infection. The number of *Pseudomonas* genes robustly captured for both isolates was consistent among compatible host genotypes (**Fig 6C**). We captured ~ 57% of the genes in DC3000 inoculated plants (126 in Col-0 and 127 in *eds1-1*), including 21 effectors out of 30 known to be present in DC3000[196]. Moreover, we captured ~35% of the genes in p13.g4 inoculated plants (76 in Col-0 and 71 in *eds1-1*). Most of the known DC3000 effectors were captured, but *avrE*, *hopM*, *hopD*, hopH, *hopC*, *hopY*, *hopAM*, and hopB were missing. Effectors known to be absent from DC3000, such as *avrRpm1*, *avrRps4*, *avrB*, *avrRpt2*, *hopAE*, and *hopAS*, were as expected not detected in samples from DC3000 inoculated plants. In the case of p13.g4 inoculated plants, seven hop genes were present (*hopAD1, hopG1, hopK1, hopO1, hopR1, hopV1,* and *hopX1*). The lineage p13.g4 is known to share the *avrE* effector with DC3000 but was not detectable in our p13.g4 infected samples [198].

I recapitulated known compatibilities between Hpa isolates and *A. thaliana* accessions (**Fig 6A**). I captured a substantial number of genes when there was compatibility. In contrast, only a few genes were captured in incompatible Hpa infections. That any Hpa sequences were captured at all could either be from the inoculum of pathogen spores and/or limited growth that was not apparent morphologically. Moreover, we wanted to verify the known presence/absence polymorphisms of ATR effectors in the tested Hpa isolates (**Fig S4**). The Hpa isolate

Waco9 lacks the *ATR1* effector and therefore evades recognition by the resistance gene *RPP1* in the Col-0 accession [195]. As expected, the *ATR1* effector in samples infected with Waco9 was not detected (**Fig S4**).

Another effector that has been shown to have P/A polymorphisms is *ATR39*. The inconclusive distribution of *ATR39* alleles and their associated phenotype led to the hypothesis that some isolates might be heterozygous at this locus or two different copies of the same effector [208]. I observed both alleles in all our isolates besides 15IN55, which only shows the presence of *ATR39-2* (**Fig S5**).

Finally, I aimed to test if I could differentiate between two polymorphic alleles of the same effector. Therefore, I focused on *ATR13*, which has 15 protein variants classified into two categories, those that can be recognized by RPP13-Nd and those that are not [209]. All the tested isolates carried the Bico1 *ATR13* allele, which is recognized by RPP13-Nd. In addition, we observed the Hind2 *ATR13* allele in the Waco9 isolate (**Fig S5**). This could be because Waco9 is heterozygous at this locus or because there are two copies of the same effector in the genome. Another explanation could be that reads between different *ATR13* alleles are similar enough to have cross-mappings.

We had access to the *ATR2* effector sequence from Cala2 and *ATR2L* effector from Emoy2 (DaeSung Kim, personal communication) (**Fig S5**). It is worth mentioning that all isolates carried the *ATR2/ATR2L* effector.

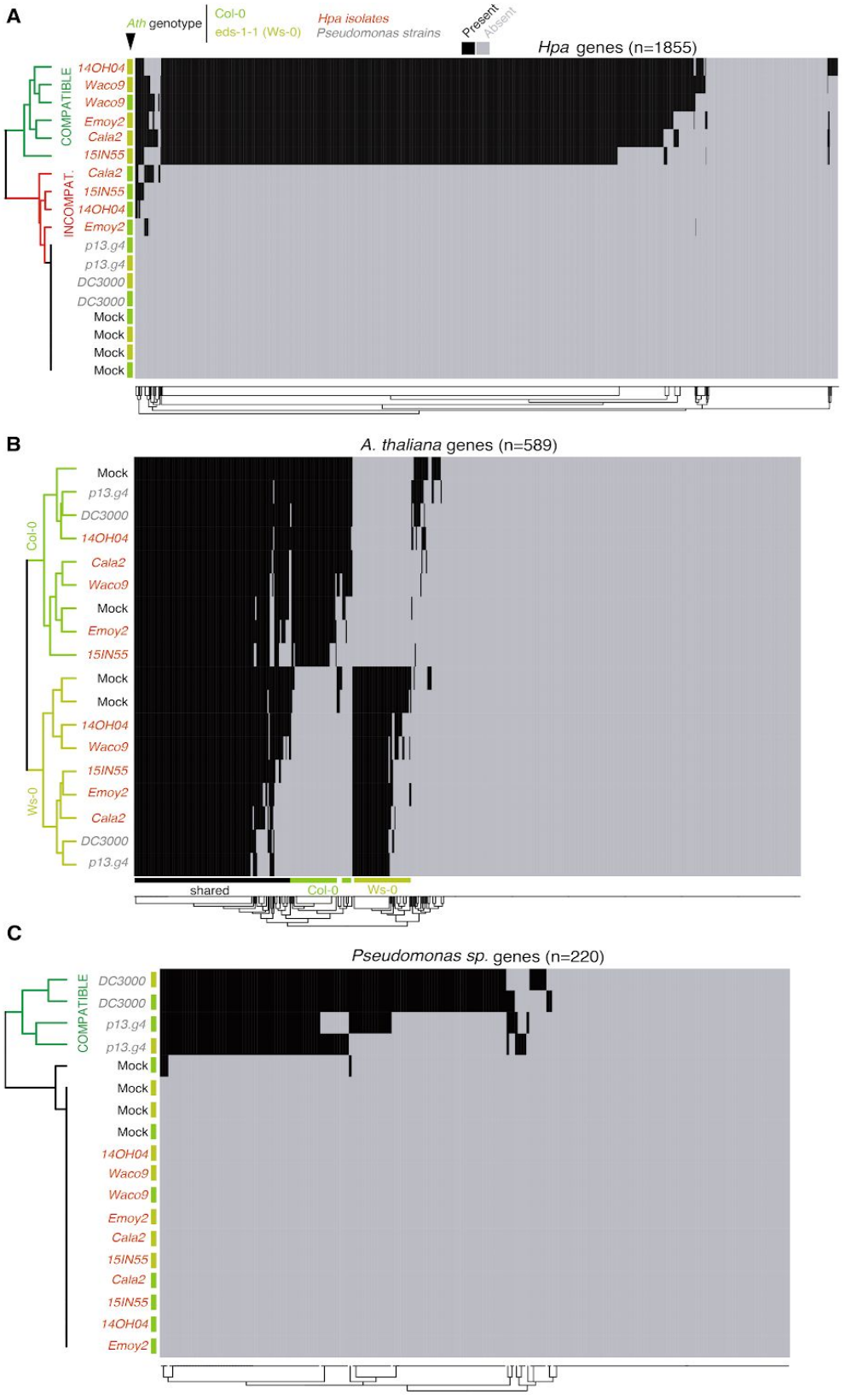


Figure 6. Presence/Absence (P/A) polymorphisms in control samples.

(**A**) Clustering of samples by P/A of a subset of Hpa pathogenicity genes, (**B**) by P/A of a subset of *A. thaliana* NLR genes, (**C**) by P/A of a subset of *Pseudomonas* pathogenicity genes.

Presence/Absence polymorphisms in the *A. thaliana* NLRome

After successfully determining the P/A of target genes in control samples, I looked at the distribution of P/A polymorphisms in our wild *A. thaliana* samples from N. America.

I first looked at the overall distribution of captured NLR genes in our sample collection (**Fig 7, Fig S5**). I could verify the presence of 67% of the representative NLR orthogroups (375/589). The sample with the highest number of present NLRs was 29CTPNR2016, with 207 present out of 589 total genes. On average, 150 NLRs were captured in each sample. Three samples had insufficient sequencing yield for determining the presence of NLR genes. Three NLRs were present in all of the 147 remaining samples (1925.T133.R1 (*LAZ5*), 7413.T164.R1 (*RPS4*), and 6924.T416.R1 (AT5G58120)).

I observed a set of NLRs present at 100% frequency in each population, representing the core NLRs (**Fig 7A**). Moreover, there is a subset of NLRs that are found at intermediate frequencies (~50%) in all populations, which could be good candidates for being under balancing selection. I also observed a set of shared genes among populations found at low frequencies (< 25%). Overall, there is a wide range of NLR frequencies within populations, although the majority of NLRs is present at high frequencies (> 75%) (**Fig 7B**). The number of NLRs found increased with the number of individuals in a sampled population, in agreement with each non-identical individual in a population carrying a different set of NLR genes.

After looking at individual populations, I moved on to investigate the distribution of individual NLRs with known function in our global sample collection (**Fig 8**). *ZAR1*, which encodes a protein that recognizes several type III secretion system (T3S) effector families from *P. syringae* and which has been reported before to be conserved among *A. thaliana* accessions [14,210], was also conserved in our sample set. *CHS3* is known to segregate at lower-altitude populations [211], and was found in ~75% of my samples. Finally, genes in the *RPW8*/*HR* cluster, which encodes NLR related proteins, showed presence/absence variation . HR2, HR3, and

HR4 were largely absent, whereas HR1 was found in one-third of the samples. Structural variation and variation in copy number of the HR cluster gene members have already been reported [212–215].

Several *RESISTANCE TO PERONOSPORA PARASITICA* (*RPP*) genes were common in my collection, including *RRP2B*, *RPP8*, *RPP9*, *RPP13*, *RPP28* and *RPP39*). Allele matching mediates recognition specificity for some of these NLR-effector pairs, explaining their maintenance in populations [107,208,216]. The canonical *RPP4* and *RPP5* genes from the *RPP4*/*RPP5* cluster could only be captured in six samples, but some of the cluster members were present in up to twenty-two samples (i.e., *SNC1*)[151]. Copy number variation and extensive sequence divergence in the *RPP4*/*RPP5* cluster has been reported and could explain these observations [186]. *RPP1* was captured in one-third of the interrogated samples, but other members of the *DM2*/*RPP1* cluster were largely missing. The *DM2*/*RPP1* cluster is known to feature extreme copy number variation, although presence/absence variation has been reported only for one gene in the cluster [52,186,217].

Finally, *WRR4*, *RML3*, *RPM1*, *RFL1*, and *ADR1* were candidates for balancing selection, being present in about a third to half of the samples, while *RPS5*, *RBA1*, and *CSA1* were rarely present [218–224].

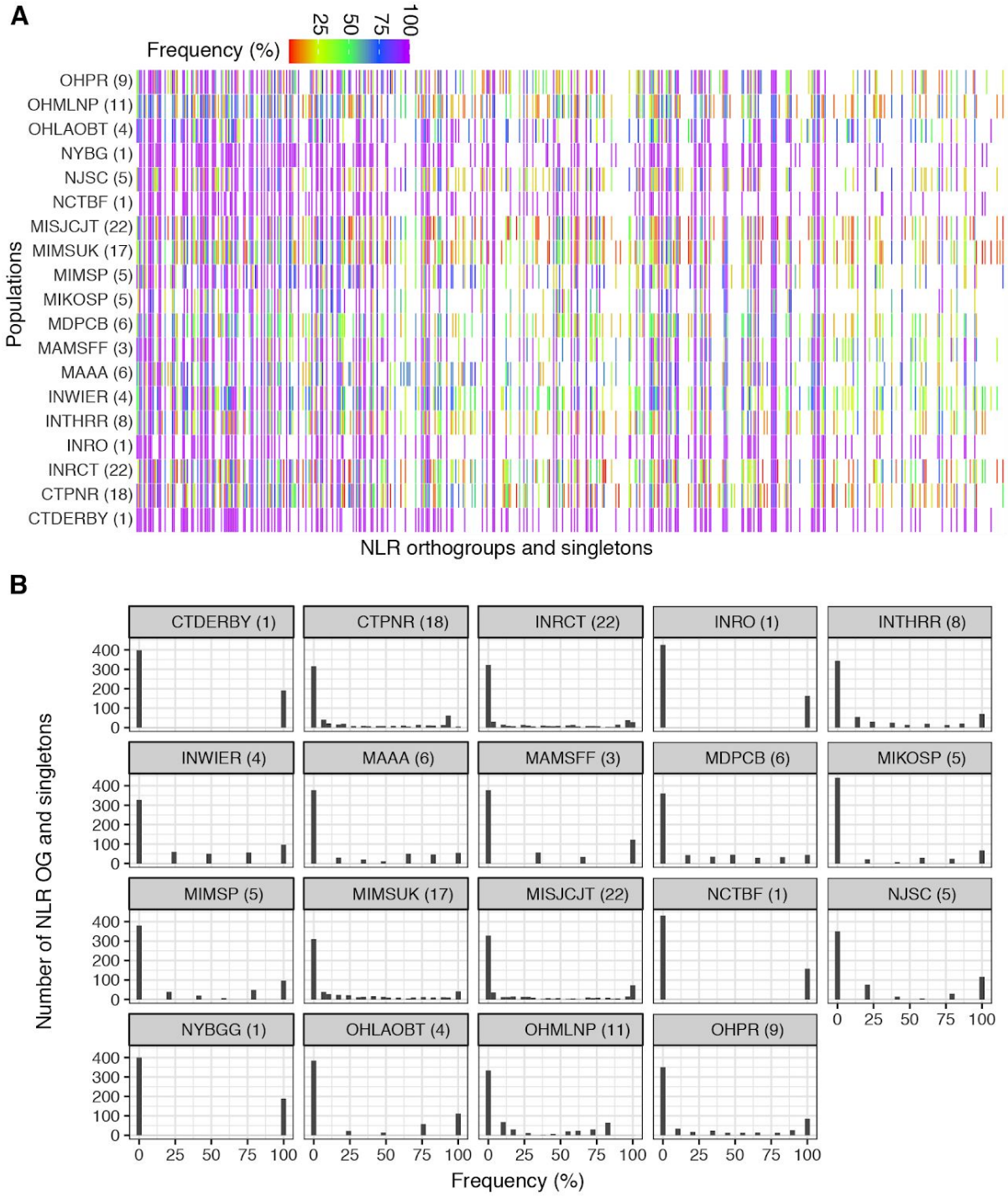


Figure 7. Frequency of NLRs and singletons in N. American populations.

(**A**) Frequency of individual NLR orthogroups and singletons within each population of *A. thaliana*. Purple represents higher frequencies whereas red represents lower frequencies. (**B**) The number of NLR orthogroups and singletons that are found within N. American populations of *A. thaliana* at defined frequencies. One bar plot is displayed for each population.

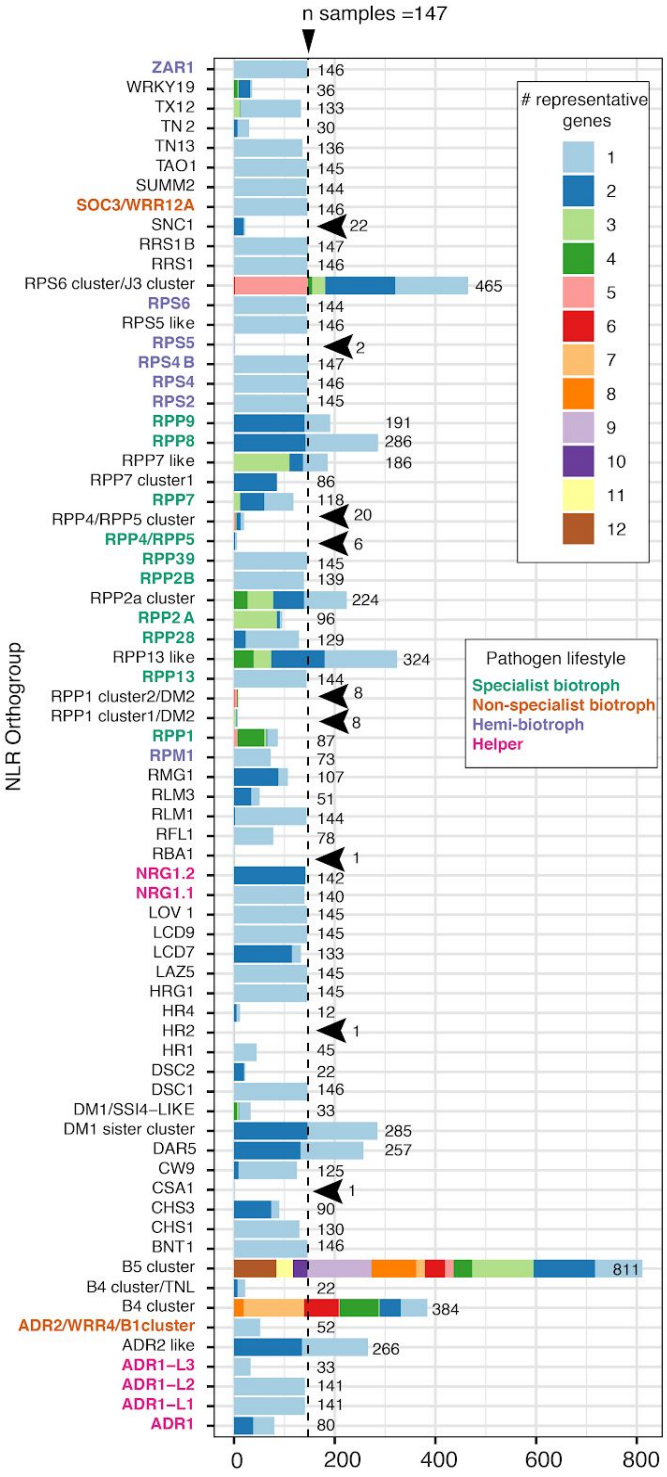


Figure 8. P/A polymorphisms of key NLR orthogroups in N. America.

P/A of NLR orthogroups in wild *A. thaliana* samples, named by representative genes with known function in each orthogroup, are shown. Names are color-coded by pathogen lifestyle, where known. Bar colors show the number of representative genes per orthogroup that attracted mappings. The numbers in black represent the number of samples where that cluster was present. Black arrowheads highlight rare NLR orthogroups.

Patterns of P/A distribution identify Hpa effectors with different frequencies within populations

After analyzing the distribution of NLR genes, I moved on to investigate Hpa pathogenicity genes. I expected to capture Hpa genes in all samples, since they were all visually infected.

I investigated the distribution of genes that are known to elicit ETI. I focused first on genes encoding ATR proteins that have been experimentally shown to have avirulence function (**Fig 7A and C**): *ATR1*, *ATR2*, *ATR5*, *ATR13*, and *ATR39* [87]. These effectors are required for host invasion and are directly or indirectly recognized by RPP proteins; therefore, they are prime candidates for playing major roles in coevolutionary conflict. We could verify the presence of all ATR effectors in our samples. *ATR13* was captured in 31% of the samples, with all samples carrying the Bico1 allelic version that is recognized by RPP13-Nd, with only one sample having the Hind2 allele of *ATR13*. *ATR13* was relatively similarly distributed across populations (**Fig 9C**). *ATR13* displays extensive allelic variation, and it is the best example in the *A. thaliana*-Hpa pathosystem for coevolution with its cognate resistance gene [107]. The fact that we only observed this gene's presence at low frequencies within each population could indicate selective pressure for Hpa isolates to lose this effector to avoid recognition.

A prevalence of 37% was found for *ATR39*, an effector with two conserved polymorphic alleles [208]. I captured both alleles, *ATR39-2* in 12% of the samples, and *ATR39-1* in 25% of samples. Thus, I looked closely at our samples and found that all samples carrying *ATR39* were either homozygous for *ATR39-1* or heterozygous. This finding agrees with what we observed for the control Hpa isolates, in which most of them were indeed heterozygous for this locus. Since control samples were only infected with a single Hpa isolate, this argues against the wild samples having been co-infected with two different Hpa genotypes. These results are in agreement with balancing selection caused by heterozygous advantage at this locus. *ATR1* and *ATR2* were found at approx. 25% to 75% prevalence within populations, while *ATR5* was the most prevalent effector in our samples, being captured in 85% of all

samples. This effector was also highly prevalent within populations, suggesting strong positive selection for its maintenance. Interestingly, the cognate R-genes in the host, *RPP4*, and *RPP5*, were only captured in four samples, and in these samples, *ATR5* was absent. This is consistent with N. American Hpa strains maintain *ATR5* when the host lacks *RPP4/RPP5*.

The second group of genes investigated for presence/absence polymorphisms were RXLR and RXLR-like effectors (**Fig S6**). We observed clear presence/absence signatures in nine of them, suggesting potential balancing selection and making them potential targets for further investigation. For example, we found *HaRxL47* at intermediate frequencies, encoding an effector that can promote pathogen growth [203]. However, there were also other effectors that are known to promote pathogen growth in a large number of accessions, and which were found at low frequency in our samples (i.e., *HaRxL62*, *HaRxL63*, *HaRxLL464*).

I also investigated other pathogenicity genes involved in PTI, such as those encoding Nep-1 like proteins, cysteine-rich proteins, elicitins, and pectins (**Fig 9B**). NLPs are secreted proteins that can affect plant growth [225]. The most common NLP genes found in our samples were those known to reduce plant growth (*NLP2* at 67%, *NLP3* at 37%, *NLP4* at 86%, and *NLP9* at 38%), while those that had shown no significant plant growth reduction in the laboratory were only found at lower prevalences (*NLP1* at 27% and *NLP7* at 8%). Moreover, I found *ELL1* (elicitin-like 1) in 27% and *Pect1* (pectin methyl esterase 1) in 47% of samples.

NLR proteins can recognize the presence of cysteine-rich proteins (CR) in fungal pathogens, and a total of sixteen CR proteins have been identified in the oomycete Hpa [108]. I captured them at frequencies between 1.3% to 47% in our samples, with the exception of *CR9* and *CR16*, which were not . Finally, I investigated the presence of a newly proposed AVR gene, the *H. arabidopsidis cryptic1* gene (*HAC1*), and its suppressor alleles (*S-HAC1* and *s-hac1*) [200]. I could detect *s-HAC1* in 72% of the samples, while *HAC1* was only present in 47% of samples.

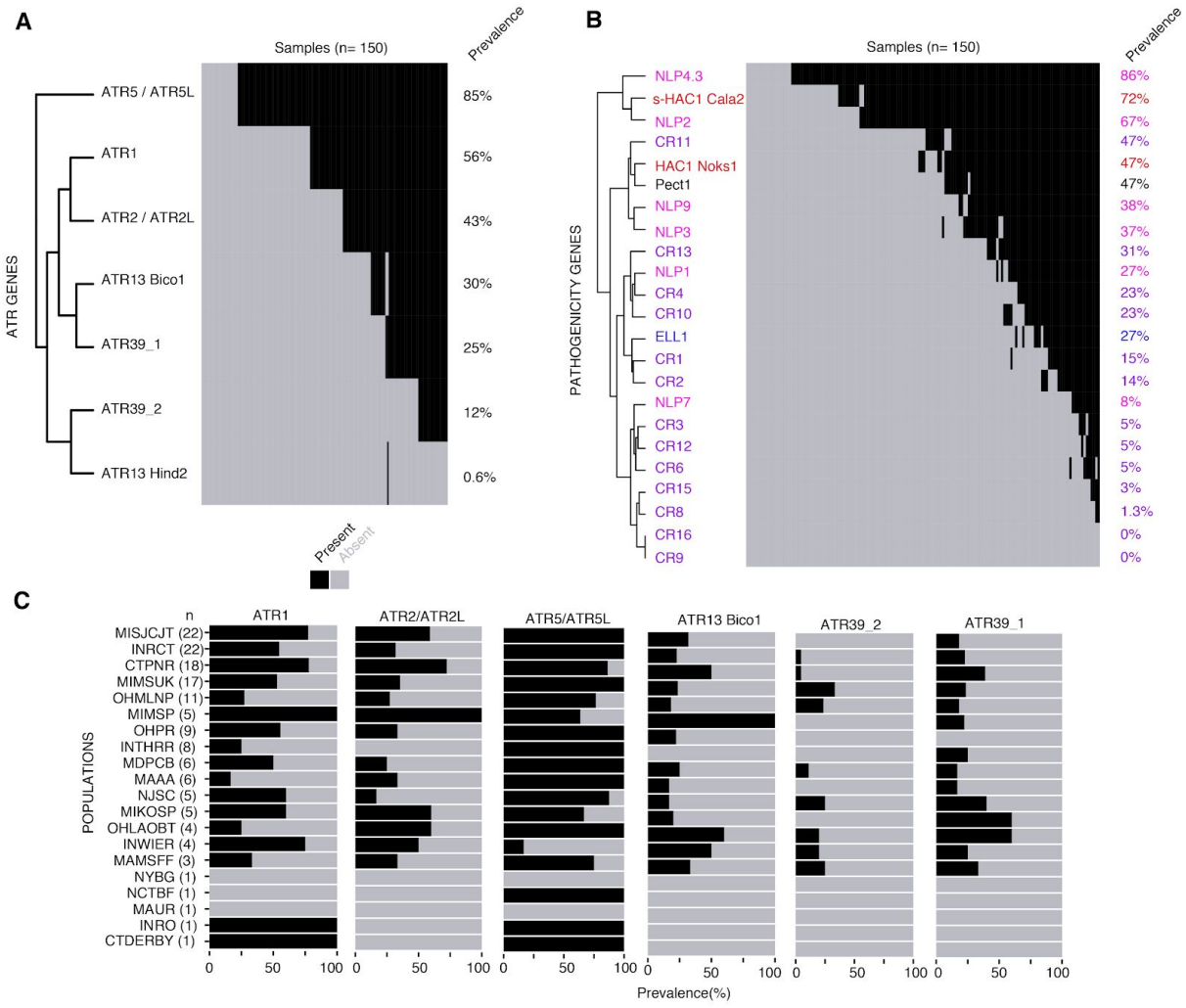


Figure 9. Presence/Absence polymorphisms (P/A) of Hpa pathogenicity genes in wild samples.

(**A**) Main *ATR* (*Arabidopsis thaliana* recognized) genes, clustered by P/A patterns across samples. (**B**) P/A of representative pathogenicity genes (cysteine-rich proteins, CR; Nep1-like proteins, NLP; *H. arabidopsidis cryptic1*, *HAC1*; Pectin1, Pec1; and Elicitin-like1, ELL1) (**C**) Prevalence of main *ATR* genes in each N. American population (*ATR13-2* not shown, since it was only present in one sample). The number of samples in each population is shown in parentheses.

Absence of *Pseudomonas* OTU5 pathogenicity genes in wild N. American samples

As a final step, we investigated the presence of Avr, Hop, and other pathogenicity genes from *Pseudomonas* spp. OTU5 in our wild samples[198] (**Fig 10**). 132 out of 220 genes used as targets for mapping were captured at least once in our samples, but none of the Avr genes. The sample with most genes captured was 53MISJCJT2015, having 56/220 genes captured. Out of the 220 genes used as

targets, 88 genes were Hop genes, and from those eleven were detected at least once, with the most common being *hopAA1*. Out of the remaining 119 pathogenicity genes, 80 (67%) could be detected at least once. The most prevalent genes were an *ABC transporter permease* gene (55 samples), a *MFS transporter phthalate permease family* gene (45 samples), and an *ABC transporter substrate-binding protein* (44 samples). Because of the small number of captured genes per sample, and given that we did not capture any genes in most samples, we did not further analyze presence/absence polymorphisms for *Pseudomonas.*

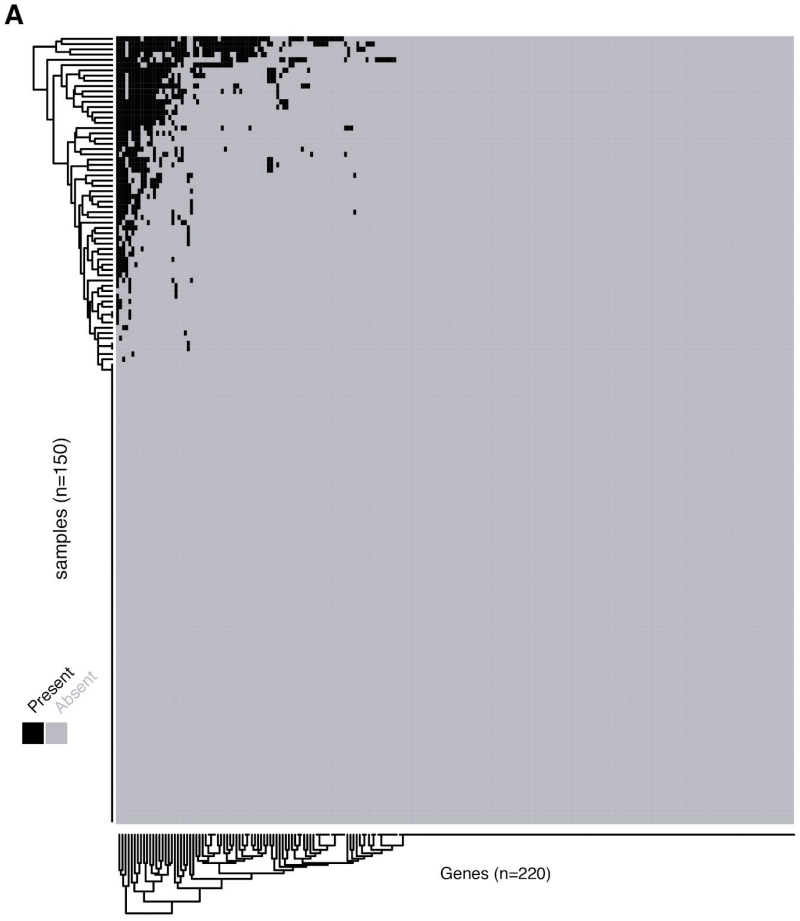


Figure 10. P/A of *Pseudomonas* spp. OTU5 pathogenicity genes in wild A. thaliana samples from N. America.

(**A**) Samples were clustered according to P/A patterns of the targeted *Pseudomonas* genes.

Discussion

Revealing the extent of host and pathogen diversity and how it maps onto each other is a standing question in coevolutionary studies[84]. By combining two previously published techniques, PenSeq and RenSeq, I aimed to shed light on the distribution of host resistance genes and pathogen effectors in wild populations of N. American *A. thaliana*. Moreover, I combined target enrichment with shotgun sequencing to uncover population structure and diversity distribution of N. American host and pathogen populations.

Using shotgun sequencing, I could assess host and pathogen relative abundances and their overall population structure. I observed a wide range of relative abundances in wild samples, emphasizing the quantitative nature of infections. The data were consistent with the promotion of bacterial growth in samples infected with oomycetes, as previously seen in other PenSeq studies[121]. Moreover, samples infected with other oomycetes and *Pseudomonas* did not overlap with those heavily infected with Hpa. These findings align with previous results and highlight the importance of niche specificity and coinfections in natural populations[206].

Prior to this study, the extent of Hpa population structure has been unknown, since only a few isolates from a limited geographical region have been sequenced [108,195,199]. My analysis revealed little population differentiation. It led us to hypothesize that gene flow between Hpa strains in populations might be common, as already evidenced by the host [49]. On the other hand, we discovered two distinct Hpa haplogroups that were often found in the same populations. It remains unclear if the observed admixture is co-infections from two distinct haplogroups or true admixed individuals infecting the same host. We have already seen that host haplogroup can have a significant impact on the Hpa disease phenotype (unpublished), equivalent to what we see in this study, where the host haplogroup 4 had, on average, higher infection levels of Hpa.

The results from the reconstruction experiment demonstrate three things; First, we can enrich a complex metagenomic sample for a given organism of interest. Second, we can reliably measure infection levels and recapitulate known host compatibilities. Third, we can accurately retrieve genes known to be present in the organism of interest and reveal the presence of known polymorphisms.

Overall, the results demonstrated here are similar to previous publications from PenSeq and RenSeq [121,122,124]. Although, the exact comparison of our combined PRenSeq approach with them is not possible due to different experimental settings, including different sequencing and mapping approaches, sample complexity, multiplexing, the nature of targeted genes, and the number of Megabases selected for enrichment.

In the benchmark PenSeq studies, the authors enriched for pathogenicity determinants in *Phytophthora infestans* and *P. capsici* mycelia grown in cultured media, therefore having a much lower sample complexity than ours. The samples were sequenced using paired-end reads of 300 bp, while I used paired-end reads of 150 bp; therefore, the original study design afforded a more precise read mapping. Additionally, both PenSeq studies with *Albugo* and *Phytophthora* mapped PenSeq reads to the same reference genome from which the baits/target genes were obtained, instead of mapping to stand-alone clustered target sequences as done here [121,122]. The most significant difference might be the bait library's size and complexity since the original studies targeted only 500 kb [122] (587 genes) and 2 Mb [121](65 genes + 400 kb contig). In contrast, 3 and 5 Mb were targeted for Hpa and *Pseudomonas*, and a total of 2504 and 372 genes. Besides these differences, comparable results were achieved.

To identify P/A polymorphisms in *P. infestans* and *P. capsici* with PenSeq, the authors of the previous studies considered a gene present when its read coverage was above 82% [122]. A similar criterion was used to determine P/A polymorphisms of *Albugo* effectors, calling a gene present with a minimum depth of ten reads and full lenght coverage of the gene [121].

RenSeq has been used for interrogating P/A variation in other plants, including tomato and potato [124]. A coverage of 20 over 500 consecutive bp was required for calling an NLR present.

In summary, PenSeq and RenSeq studies have used coverage and depth as the main criteria determining the presence/absence of genes. Hence, we combined these criteria and used gene coverage, the average number of mapped reads, and minimum depth.

Overall, the P/A analysis results in control samples confirmed the successful capture of host NLR and pathogen effector genes, recapitulating known compatibility and P/A variation in the different host and pathogen genotypes. One of the limitations observed was that certain *Pseudomonas* genes expected to be present were not detected. This might be a problem of not achieving enough sequencing depth or that the P/A criteria were too stringent.

In wild *A. thaliana* samples, two similar NLR genes were present in all samples, *RPS4* and *LAZ5*. *RPS4* is an important disease resistance gene the product of which recognizes the AvrRps4 type III effector from *P. syringae* [226]. *LAZ5* has sequence similarity to *RPS4* and can trigger cell death [227]. The fact that these NLRs were strongly conserved in N. American accessions hints towards their crucial role in fighting bacterial pathogens, and that they might be experiencing strong positive selection. The NLR singleton AT5G58120 was also retrieved in all samples; it has similarities with another TIR-NLR in the B4 cluster, but its recognition specificities are unknown. The most interesting finding was the seemingly low prevalence of *RPP4* and *RPP5* sequences in N. American populations. This goes in hand with *ATR5*, the matching Hpa effector gene being found at high frequencies in N. America. Another explanation for the lack of *RPP4/RPP5* sequences is that they were lost during the diversity bottleneck that preceded the N. American colonization, or that they have fitness costs for the plant in the absence of *ATR5* and Hpa isolates with *ATR5* having been introduced only later in N. America. Much of the observed P/A variation in NLRs has also been reported in the native range, suggesting that N. American accessions are experiencing similar coevolutionary dynamics.

Finally, we identified *ATR5* as a core effector gene in N. American Hpa strains, whereas other known effectors seem to be under balancing selection. Other groups of pathogenicity genes engaged in PTI were present at higher to intermediate frequencies, revealing their importance in helping Hpa infect its host.

Together, our findings in presence/absence polymorphisms highlight the different dynamics of plant-pathogen coevolution. We observe patterns consistent with arms race dynamics for the *RPP5-ATR5* NLR-effector gene pair, with fixation of *ATR5* and near absence of RPP4/RPP5. On the other hand, we observed patterns more consistent with trench warfare dynamics for other NLRs and Hpa effector pairs such as *RPP1-ATR1* and *ATR39*, which suggest negative-frequency dependent selection or heterozygous advantage as main evolutionary drivers. To conclude, we present PRenSeq as a combined target enrichment sequencing approach to capture host and pathogen genes in the same sample for population genetics and coevolutionary studies.

Materials and Methods

Sample collection

A. thaliana infected leaves (with Hpa, *Albugo*, or both) were collected from wild N.American populations in Midwest and East Coast regions during three consecutive years; 2014, 2015, and 2016. Leaf material was frozen and stored at -80 ˚C until DNA extraction took place. We used frozen stocks of Hpa isolates 14OHMLNP04 and 15INRCT55 from N. America. Parker's lab provided frozen stocks of British Hpa isolates Cala2, Waco9, and Emoy2. All these isolates were revived and used for controlled infections for the Hpa bait set. Control *Pseudomonas* syringae DC3000 [228] and local *Pseudomonas* strain p13.g4 [198] were used as controls for the *Pseudomonas* bait set.

Control infections

Control samples represent reconstruction experiments to verify that DNA from targeted organisms and loci were enriched and sequenced. Col-0 and *eds1-1(Ws-0)*

two weeks old seedlings were sprayed inoculated with Hpa isolates; inoculations were done by adjusting Hpa spore concentration to $5x10^4$ spores/mL of water and spray-inoculating the spore suspension on the plant. *Pseudomonas* infections were done by syringe inoculation of bacteria liquid culture at an OD=0.0002. Negative control plants from *Pseudomonas* (without inoculation) were grown alongside inoculated plants. Negative control plants from Hpa were kept separately. Samples were collected seven days post-infection (dpi) for Hpa and three dpi for *Pseudomonas* and frozen at -80 until DNA extraction was performed.

DNA extraction and sequencing of Hpa isolates

DNA was extracted from spores from three different Hpa isolates (14OH04, 15IN54, 15IN55) and was isolated using the CTAB DNA extraction method [229]. Isolates 14OH04 and 15IN55 were sequenced in 2017 using 150 bp paired-end reads with Illumina HiSeq 3000. Isolate 15IN54 was sequenced in 2015 using TruSeq PCR free 2ug of material on MiSeq 2000 with 300 bp paired-end reads.

The 14OH04 Hpa isolate was additionally sequenced using long-read technology. Hpa spores were harvested in water from 3 week old eds1-1 infected plants. The spore water solution was centrifuged, and the spores pellet frozen at -20 C until DNA was extracted. ~500 uL of spore pellets were used as a starting point for DNA extraction following the extraction protocol from [230]. We performed a DNA size selection cutoff of > 10 Kb using the BluePippin (Sage Science). The sample was sequenced in a PacBio Sequel I System from the Max Planck of Developmental Biology. We achieved a N50 read length of ~ 28 Kb after sequencing, yielding 9.9 Gb and 1231537 reads. Quality control was performed with FastQC (Andrews and Others 2010). Sequence reads were trimmed based on quality scores using Trimmomatic v0.36 [231] using a sliding window approach (SLIDINGWINDOW:10:20). Reads were trimmed when average quality dropped below a threshold and discarded if the remaining read dropped below a minimum length (MINLEN:200).

Genome Assembly of Hpa isolates

Short reads were assembled into contigs with SPAdes v3.7.1 in paired-end mode [232]. For assembling the 14OH04 Hpa isolate with long-read data, first, we removed *A. thaliana* reads by mapping to the TAIR10 reference genome using minimap2 (2.11-r797) and long-read specific parameters (-x map-pb -H). The remaining reads were assembled using the Flye assembler (version 2.4.1-gbdbd33e) [233] and an estimated genome size of 80 Mb.

For both short and long reads, BLOB tools (version 1.0) were used to identify and visualize contamination; for this purpose, reads were aligned to the contigs using bowtie2 (version 2.3.4). Contigs were aligned to the non-redundant protein reference database of the (NCBI NR, released January 23rd, 2018) using diamond (version v0.9.14.115) [234] to obtain taxonomic annotation. Only contigs unambiguously classified as Peronosporales on the NCBI taxonomy tree's order level using the Lowest Common Ancestor (LCA) approach were retained. The assembly statistics for all Hpa genomes can be found in **Table S1**.

Gene prediction for Hpa genomes

Gene prediction was performed with the webserver of AUGUSTUS v3.3 [235]. First, we used as an oomycete gene prediction training set the Hpa Emoy2 reference genome file from ENSEMBL Genomes database (file: Hyaloperonospora_arabidopsidis.HyaAraEmoy2_2.0.cdna.all.fa as cDNA file and Hpa_Emoy2_V8.3.fa as Genome file). Gene prediction was made with this training set for the rest of our Hpa isolates with the following parameters: (User set UTR prediction: false; Report genes on both strands; Alternative transcripts: few; Allowed gene structure: predict any number of (possibly partial) genes; Ignore conflicts with other strands: false). Augustus job ID and number of predicted genes and proteins can be found in **Table S2.**

Secretome prediction and annotation

We used a modified version of a pipeline previously described in [236] to identify the diverse Hpa genomes' secretome. The predicted proteins with Augustus are used as the input files for secretome annotation and prediction. In brief, the first step consists of predicting signal peptides (SP) using Neural Networks and Hidden Markov Models (HMMs). To confirm an SP, it must be predicted by both tools and pass the threshold score. Second, transmembrane domains are predicted using TMHMM. Third, proteins subcellular localization are predicted using TargetP. Finally, only proteins that have SP and do not localize on the mitochondria and chloroplast and do not contain transmembrane domains are kept. These putative secreted proteins are then used to predict effector motifs (RXLR, CRINKERS, and WY) using HMMs. An extra manual step was done to remove any protein with a hit on the curated Swissprot database, except associated pathogenicity enzymes that we identified using several keywords (**Table S3**). The redundancy of the database was removed by performing a self-blast.

Target enrichment genes

For Hpa, we included baits targeting neutrally evolving genes and genes under selection (purifying and diversifying selection) based on Tajima's D and Fu's Fs statistics considering eight different Hpa isolates analyzed with DNAsp software [237] [201] . Neutrally evolving genes were considered when Fu's Fs and Tajima's D statistics were between -0.3 and 0.3. Genes under selection were considered when Fu's Fs and TD were higher or equal to + 2.5 and Tajima's D lower or equal to -1.8. An additional twenty-four housekeeping genes were retrieved from the EumicrobeDB Emoy2 genome to compare diversity with the rest of the gene set (**Table S4**). Effector and Pathogenicity genes were retrieved by extensive literature data mining and examination of the Emoy2 reference genome; Nep1-like proteins gene sequences that match with protein file from Emoy2 were retrieved, as well as Hpa isolates Emoy2 and Waco9 cysteine-rich proteins, putative elicitins and RXLR effector gene sequences [108]. Hpa Emoy2 high-confidence and putative RXLR effector candidates gene sequences were retrieved from a previous publication [195].

Hpa race-specific effector alleles sequences were taken from GenBank database (https://www.ncbi.nlm.nih.gov/genbank/) by manual search. HaRxL96 effector published sequence was retrieved from Emoy2 reference genome files (Hpa802236)[202]. Extracellular glucanases HaEGL12-1, HaEGL12-2, and pectinase HaPect1 protein sequences were retrieved and matched with corresponding gene sequence ID (HpaG808599, HpaG814377, and HpaG809280, respectively) in the latest available gene annotated Emoy2 reference genome[199].

For *Pseudomonas*, ~2500 species with 1524 assembled genomes were annotated for effectors, hormones, hrp-hrc, phytotoxins, and core genes [198].

For capturing the *A. thaliana* NLRome, we included the 13,167 annotated NLR available genes as targets[160].

Bait design for target sequence enrichment

Target sequences for the three different species; *A. thaliana*, Hpa, and *Pseudomonas* were sent for bait design to myBaits® custom kits from Arbor Biosciences®. Biotinylated RNA baits sequences to hybridize with target sequences were designed after applying the following filtering criteria. For the Hpa bait set design, 2,504 target sequences were provided (2,558,865 bp total, including Ns), sequences were soft masked for simple repeats (0.59% masked), 80 nt baits were designed with 1.5x tiling (40,395 baits). Only baits that didn't match the *Pseudomonas* collection didn't match the *A. thaliana* genome and were ≤ 25% masked (39,945 baits) were kept. For the A. thaliana baits set design, 13,167 sequences were provided (32,297,012 bp total), 80 nt baits were designed with 3x tiling (1,133,909 baits), identical baits were collapsed (326,527 baits), only baits that matched regions of the *A. thaliana* genome that were ≤ 25% repeat masked, had ≤ 20 BLAST hits, didn't match the mitochondrial/plastid genomes, didn't match the *Pseudomonas* collection, and didn't match the oomycete collection (316,755 baits) were kept, filtered baits that were 97.3% identical over 83% of the sequence were collapsed (78,817 baits). For generating the *Pseudomonas* bait set design, 41,598 sequences were provided (120 bp each), sequences were soft masked for simple repeats (0.21% masked), two 70 nt baits per locus were designed (50 nt apart,

83,196 baits), only baits with no BLAST hits to *A. thaliana* or the Hpa Emoy2 oomycete genome, and were ≤ 25% masked were kept (82,874 baits).

DNA extraction and library preparation target enrichment

Plant tissue was collected in 2 ml screw cap m filled with garnet rocks (up to 0.5 ml) and deep-frozen in liquid nitrogen, stored at -80°C. Plants were ground with Fast-Prep-24 5G (MP Biomedicals) at speed 6 for 40s. 800 ul of prewarmed (55°C) Extraction buffer (100mM Tris pH8, 50 mM EDTA, 500 mM NaCl, 1,3% SDS, and 20 mg/mL RNaseA) was added an additional grinding was done at speed 6 for 40s. Samples were incubated for 10 min at 55°C followed by centrifugation at 12,000 g for 1 min. 400 uL of lysate was transferred to a 96-well plate with 130 uL of KAc per sample and mixed in a plate mixer for 50 sec at 800 rpm, followed by incubation at 4°C for 5 min. Samples were centrifuged for 5 min at 6200 g. 300 uL of the supernatant was transferred to a fresh 96-well plate with 300 uL of SPRI beads and mixed for 1 min at 800 rpm. Samples were placed on a magnetic stand until the beads were bound to the side. The supernatant was removed, and two series of 80% Ethanol washes were done. DNA bound to beads was resuspended in 50 uL of water for 5 minutes. The plate was placed on the magnetic stand, and the supernatant containing the DNA was transferred to a fresh plate. Genomic DNA libraries were constructed using a modified version of the Nextera protocol [178], modified to include smaller volumes. Briefly, 0.25-2ng of extracted DNA was sheared with the Nextera Tn5 transposase. Sheared DNA was amplified with custom primers for 14 cycles. DNA from libraries was quantified using the fluorescent dye PicoGreen® kit on a TECAN plate reader before pooling. For the control pool (22 samples), between 10 to 450 ng were pooled. For the US samples, 5 pools of 30 samples each were made, individual libraries were pooled at a normalized amount of 200 ng each. Pools were cleaned for removing amplification primers using Sera-Mag Magnetic Speedbeads “SPRI beads” (GE) at a 1:1 ratio. For host whole-genome sequencing, three pools of 50 libraries (~65 ng each) were prepared. Pools were cleaned for removing library amplification primers using SPRI beads (1:1).

In-solution target enrichment

myBaits® (Arbor Biosciences) Hybridization Capture for Targeted NGS Manual v4.01 April 2018 was used with small modifications. Each Nextera library pool was split into three subsamples of 20 uL as an input for the Appendix A2 of the protocol "Pre-treating libraries made with Nextera kits" and repooled afterward. Pools were cleaned for removing amplification primers using SPRI beads at a 1:1 ratio. The Hybridization mix setup was done using the following components per reaction: 9.25 uL of Hyb N, 3.5 uL of Hyb D, 0.5 uL of Hyb S, 1.25 uL of Hyb R, and 5.5 uL of Baits. The blockers mix setup was done mixing per reaction 0.5 uL of Block A (Nextera adaptors blockers), 2.5 uL of Block C, and 2.5 uL of Block O. library pools were split into three 7 uL ranging from 300 ng to 1 ug of pooled libraries DNA for each capture bait set reaction (Hpa, *Pseudomonas* and *A. thaliana*). The 7 uL of pooled libraries were mixed with 5uL of blockers mix and during reaction assembly, mixed with 18uL of the hybridization mix. Hybridization was done at 65 °C for 16 hours. Captured libraries cleanup was done using a tube-compatible MPC, and library amplification was done directly on 15 uL of on-bead enriched library suspension as a template for 11 PCR cycles using the 2X KAPA HiFi HotStart Ready Mix DNA Polymerase (KAPA Biosystems) and reamp p.5 and p.7 primers from [238]. Captured pools were cleaned for removing amplification primers using SPRI beads at a 1:1 ratio. Each initially split captured pool (Hpa, *Pseudomonas*, and *A. thaliana*) was quantified using Qubit and pooled back together at the following DNA amount (ng) proportions; Control pool (36.5 % Hpa, 40% *A. thaliana,* 22% *Pseudomonas*) and N.American samples pools (10% Hpa, 30% *A. thaliana*, 60% *Pseudomonas*).

Target enrichment sequencing and QC filtering

To determine average libraries size and nanomolarity and to verify the lack of adaptor contamination and suitable sequencing size, 0.5 ng of each of the pools (6 final captured library pools and the three whole-genome sequence pools) were run on a High Sensitivity DNA Chip (Agilent Technologies) and measured with a 2100 Bioanalyzer (Agilent Technologies). Libraries ranged from an average size of 370 to

645 bp. Libraries were then sequenced on the Genome Center of the Max Planck Institute from Dev. Biology using Illumina HiSeq 3000 150 bp paired-end reads. Conversion to FASTQ and demultiplexing was done with bcl2fastq2 version 2.18 from Illumina. To inspect read quality, MultiQC reports were generated with MultiQC version 1.3.dev0 [179] and included information from FastQC v0.11.5 and fastq_screen V0.5.2. Reads were adapter and quality trimmed using trimmomatic [231] and the following settings: ILLUMINACLIP:NexteraPE-PE.fa:2:30:10 and SLIDINGWINDOW:6:30 MINLEN:60. The same sequencing libraries used for in-solution target enrichment were used for whole-genome-sequencing using 150 bp paired-end reads with Illumina HiSeq 3000. Quality filtering was done with the same filtering criteria as described before.

Read mapping for target enrichment sequencing

Since baits can hybridize with target DNA sequences with 80% sequence identity or more, we use the clustering algorithm uclust [239] to cluster the reference sequences used for mapping the captured reads. We ran usearch -cluster_fast command on the three different reference sequences set independently (*A. thaliana*, Hpa, *Pseudomonas*) with a sequence identity cutoff of 0.8 (-id 0.8) and used the centroid sequences (-centroids) of each cluster as a reference sequence to map against (**Table S5**). Then, reads were mapped to these reference centroid sequences using bwa mem standard parameters [240] (version 0.7.17-r1188) for each reference set.

Mapped bam files were filtered to remove PCR duplicates using samtools rmdup [241]. We removed alignments with mapping quality smaller than 20 using samtools view (-q 20), supplementary mappings were removed (-F 2048) and clipped bases.

Percentage on-target reads

We calculated the percentage of total PRenSeq high-quality reads (trimmed and paired reads) that were mapped to each capture reference set (*A. thaliana,* Hpa, and *Pseudomonas*); we called this metric the percent of total reads on-target even

though we only mapped to the reference sequences. We also calculated the percent of mapped reads on-target as the proportion of total mapped reads after quality filtering that map to each reference (*A. thaliana,* Hpa, and *Pseudomonas*); this is the percent of mapped reads on-target.

Read mapping for whole-genome-sequencing

For calculating relative abundance, reads were mapped using bwa mem using standard parameters [240] (version 0.7.17-r1188) against a reference database of concatenated genomes containing *A. thaliana* assemblies, Hpa, *Pseudomonas*, and other Oomycetes. For *A. thaliana* accessions: TAIR10 and KBS-Mac-74 [155]. For Hpa isolates Cala2, Emoy2, Noks1, 15IN54, 15IN55, 14OH04. For other oomycetes, we used a list of 24 genomes (**Table S6**). For *Pseudomonas* strains, we used 1524 assemblies [198].

For genotyping and calculating the percentage of total mapped reads to each organism, we mapped the reads against TAIR10 for *A. thaliana*, 14OH04 for Hpa, and p7.E10 for *Pseudomonas*. We selected the *Pseudomonas* reference genome by identifying for each sample the *Pseudomonas* strain that gets most of the reads mapped to in all the samples, meaning the most abundant strain in our sample collection. After mapping, three P. viridiflava strains were the topmost abundant ones. The strain p7.E10 (AthOTU5) was the top hit in 96 samples and chosen as reference (p11.H11 in 16 samples and p9.C4 in 9 samples) (**Table S7, Figure S3**).

Genotyping

For detecting variants, we used freebayes, a haplotype-based variant detector [207]. We ran freebayes in the multi-sample VCF (population mode), and we only called variants that had a minimum coverage of 5 (--min-coverage 5) and a minimum of two observations (-C 2). We first filtered out variants with quality less or equal to 20 (--minQ20), keeping only SNPs (--remove-indels). Then, we look at the proportion of missing data per individual (--missing-indv) and filtered those with equal or less 0.9 of missing data in the individuals mapped to *A. thaliana* TAIR10 and Hpa 14OH04 reference files. We then look at the average depth per site to

calculate the depth maximum limit to filter out SNPs above that depth, set as the average mean depth * 2 for each file (--max-meanDP)(**Table S8**). We only kept SNPs with a Minor Allele Count (MAC) of 3 for *A. thaliana* and Hpa (--mac 3) and MAC 2 for *Pseudomonas* (--mac 2). We only kept biallelic SNPs (--min-alleles 2 --max-alleles 2). Thus, we ended up with 129 individuals for each dataset and a total number of SNPs for; *A. thaliana* TAIR10 81942 SNPs, Hpa 14OH04 59732, and *Pseudomonas* 1294.

Presence/Absence of genes

We determined the presence/absence of a gene in each sample based on gene coverage, mean read depth, and the number of mapped reads. These were calculated using samtools coverage [241] on the quality-filtered mapped reads bam files. We considered a gene present when it matched the following criteria: coverage above or equal to 85%, depth above or equal to 2, number of reads above or equal to the half the average number of reads for all samples (Hpa; 200 reads, Ara; 200 reads, Pseudo; 70 reads. We then created a presence/absence binary matrix for each gene x sample combination.

Principal Component Analysis

To obtain insights into the samples' genetic distance and internal structure within each dataset, we computed a PCA for each dataset independently (Ara and Hpa) based on the variance-standardized relationship matrix using PLINK v.1.90b4.1 option --pca [169]. We chose to plot the first two principal component axes (PC1 and PC2) since they account for most n in the data variation.

Admixture Analysis

We ran ADMIXTURE software v1.3.0 [153] for Hpa and *A. thaliana* VCF files. For *A. thaliana,* we selected K = 1 to 11 to assess the most likely value of K, including a 5-fold cross-validation procedure. The software ADMIXTURE estimated an optimal K = 6, having the lowest error. We then re-run ADMIXTURE using K=6 and 2000 bootstraps to better estimate accessions admixture proportions. For Hpa,

we selected K = 1 to 6 to assess the most likely value of K, including a 5-fold cross-validation procedure. The software ADMIXTURE estimated an optimal K = 2, having the lowest error. Admixture proportions were used to classify each sample organism into an admixture group. We considered a pure individual when admixture >=99% ancestry and admixed when it carries >=1% of another ancestry group.

Supplementary Figures

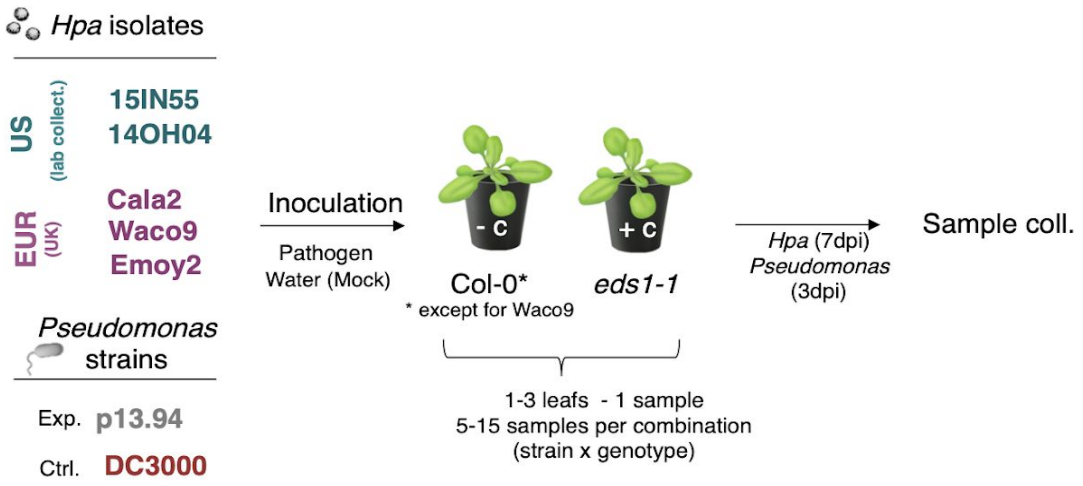


Figure S1. Overview of the experimental set up for the reconstruction experiment (control samples).

In brief, five Hpa isolates and two *Pseudomonas* isolates were inoculated independently on two different host genotypes. Samples were collected seven days post-inoculation for Hpa and three days post-inoculation for *Pseudomonas.*

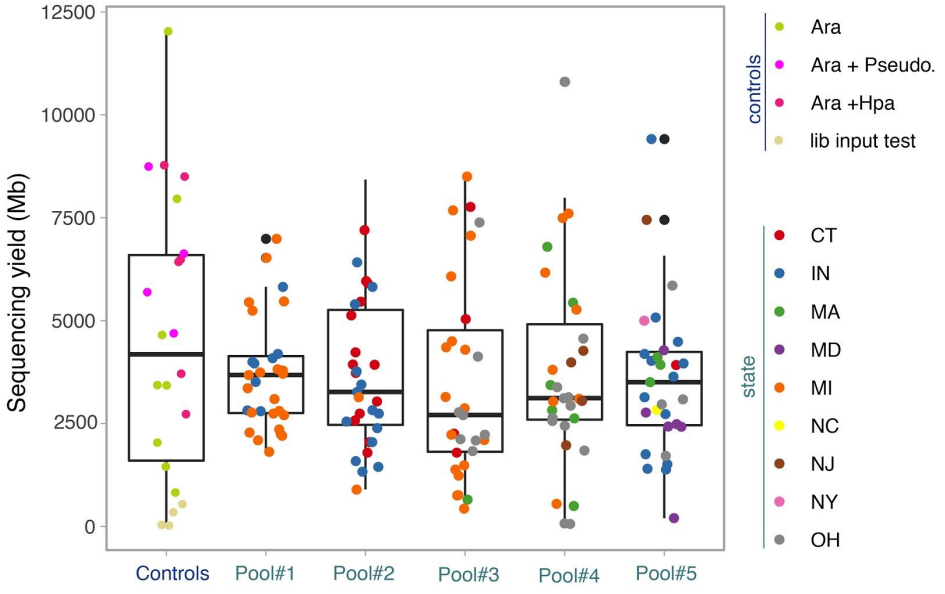


Figure S2. Sequencing yield from the target enrichment sequencing experiment.

We sequenced five pools of wild samples and one pool of control samples. Wild samples are color-coded by the provenance state of those samples in N. America. Control samples are color-coded by organisms present in those samples.

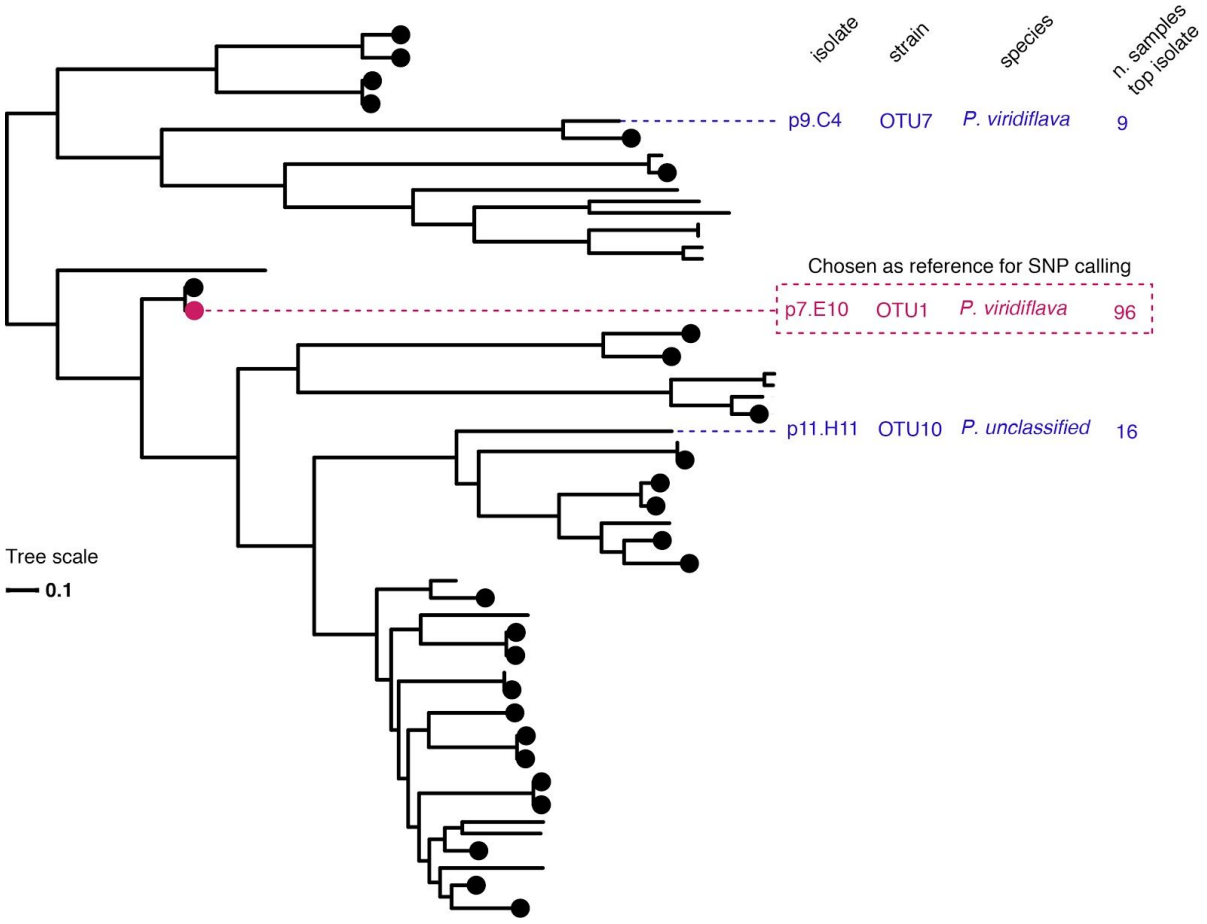


Figure S3. Phylogenetic tree of *Pseudomonas* strains used as mapping references for whole-genome sequenced samples.

Phylogenetic tree built with core genome SNPs of 1524 *Pseudomonas spp*. strains from http://panx.weigelworld.org/. The three most abundant *Pseudomonas spp.* isolates (higher number of mapped reads) from whole-genome sequenced samples are highlighted on the tree. We selected the strain p7.E10 as the reference genome to map to for variant calling.

	ATR1	ATR2/ATR2L	ATR5/ATR5L	*Bico1_ATR13	Hind2_ATR13	ATR39_1 (Emoy2)	ATR39_2 (Cala2)
Mock eds1-1	0	0	0	0	0	0	0
Mock Col-0	0	0	0	0	0	0	0
Mock eds1-1	0	0	0	0	0	0	0
Mock Col-0	0	0	0	0	0	0	0
DC3000 Col-0	0	0	0	0	0	0	0
DC3000 eds1-1	0	0	0	0	0	0	0
p13.g4 eds1-1	0	0	0	0	0	0	0
p13.g4 Col-0	0	0	0	0	0	0	0
Emoy2 eds1-1	1	1	1	1	0	1	1
14OH04 eds1-1	1	1	1	1	0	1	1
Waco9 Col-0	0	1	1	1	1	1	1
Waco9 eds1-1	0	1	1	1	1	1	1
15IN55 eds1-1	1	1	1	1	0	0	1
Cala2 eds1-1	1	1	1	1	0	1	1
Cala2 Col-0	0	0	0	0	0	0	0
Emoy2 Col-0	0	0	0	0	0	0	0
15IN55 Col-0	0	0	0	0	0	0	0
14OH04 Col-0	0	0	0	0	0	0	0

*RPP13-Nd recognized

Figure S4. Presence/Absence matrix of main ATR effectors from Hpa in control samples.

Presence (1) / Absence (0) matrix of Hpa ATR effectors for control samples (first column; inoculated pathogen and host accession). Blue cells are negative controls where no Hpa is inoculated. Green cells are compatible host genotypes with the inoculated Hpa isolate, whereas red cells denote incompatible interactions. Yellow cells highlight the absence of the ATR gene.

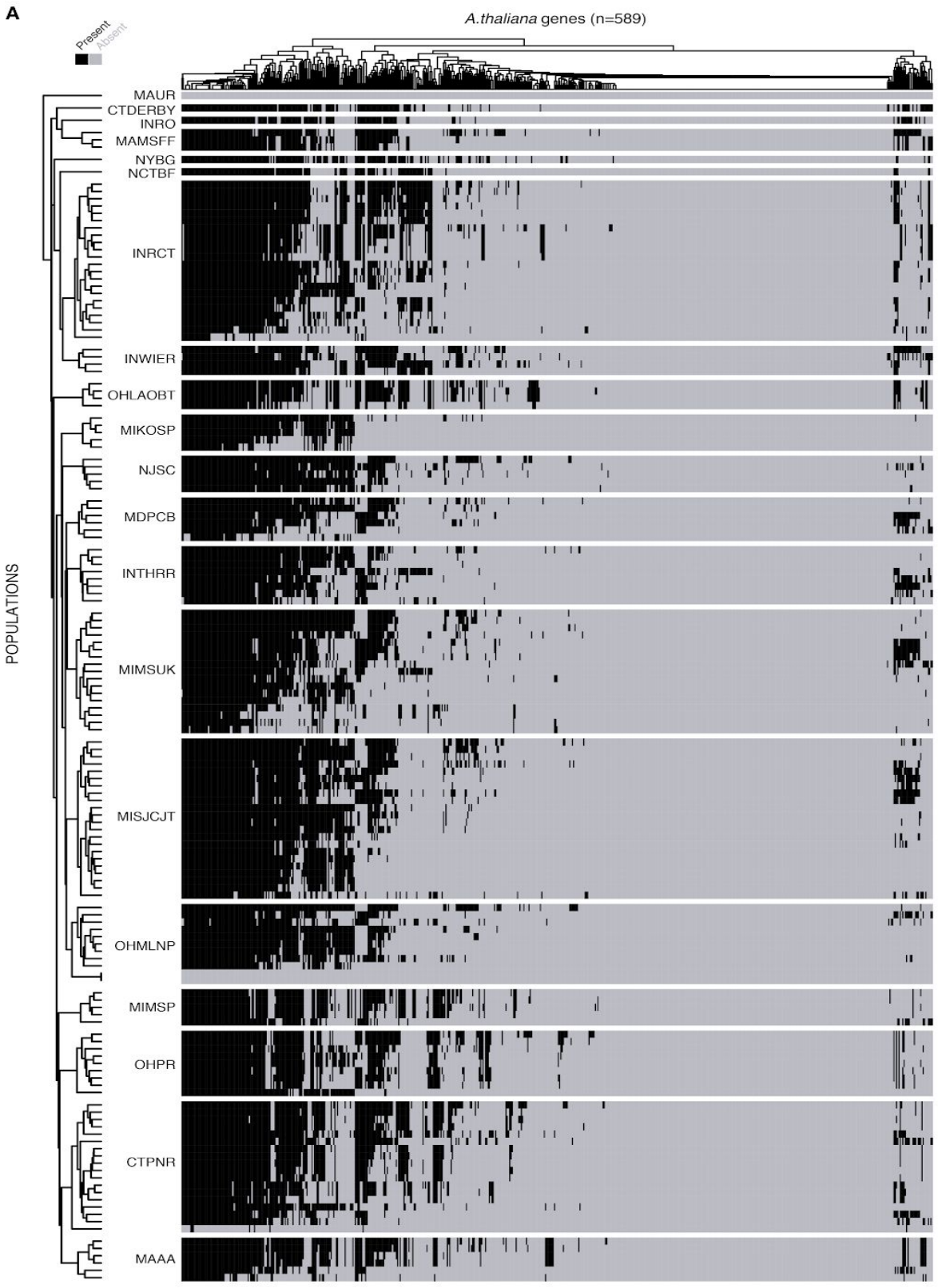


Figure S5. Presence/Absence polymorphisms of *A. thaliana* NLR genes in N. American populations.

P/A of NLR clusters representative genes in N. American populations.

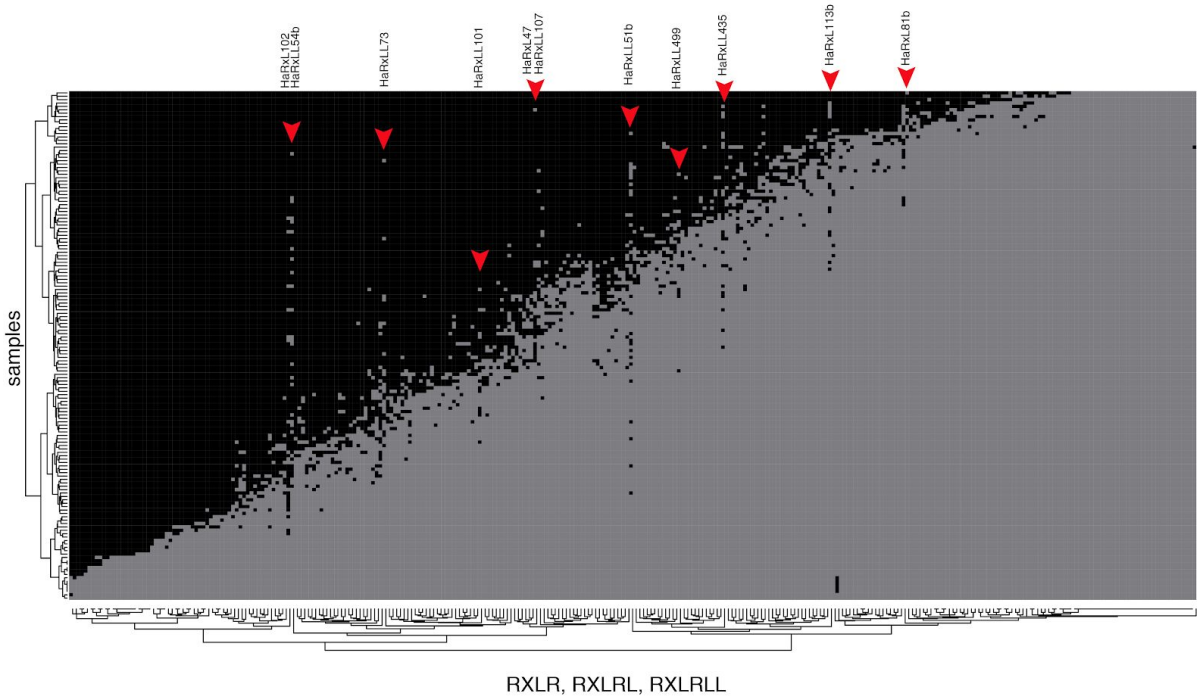


Figure S6. Presence/Absence polymorphisms of Hpa RXLR genes

P/A matrix of RXLR and RXLR-like genes from Hpa. Red crosses indicate genes under putative presence/absence variation.

Supplementary Tables

Sequencing platform	Hpa Isolate	Number of contigs	Total length	Max length	N50	N90
PacBio Sequel	14OH04	996	55696136	492031	83599	27980
Illumina 3000 2x150 bp	14OH04	5787	54695430	187515	34496	5843
Illumina 2000 2 x 300 bp	15IN54	3205	59531833	771396	94702	15495
Illumina 3000 2x150 bp	15IN55	5573	55253210	183023	36685	5972

Table S1. De-novo assembly statistics of N. American Hpa isolates.

Hpa isolate	Augustus Job ID	Genes	Proteins
Emoy2	preddkWW3v7x	13241	15417
Noks1	predPSqanH5Y	18454	20844
Cala2	predqbteqjn7	15183	17435
14OH04	predE_x2YLhd	10798	12698
15IN55	predqVqnEnBL	11985	14057
15IN54	predzvaAYMNT	11697	13701
Training set	trainaCiiA9JW		

Table S2. De-novo annotation statistics from AUGUSTUS.

Keyword search to keep
cutinase
glycosil hydrolase
polygalacturonase
cellulase
endoglucanase
exoglucanase
protease
elicitin
Beta-galactosidase
Catalase-peroxidase
cellobiohydrolase
Probable endo-beta-1,4-glucanase
pectin lyase
Endo-beta-1,4-glucanase
Superoxide dismutase
endopolygalacturonase
Endo-1,4-beta-xylanase
Carbonic anhydrase
beta-glucosidase
exo-1,4-beta-xylosidase
virulence

Table S3. Keyword search from Swissprot annotation of Hpa secretome to keep for bait design.

HK Category	Subcategory	Source	Gene ID
Gene Expression(GeneExp)	Transcription factor(TF)	EumicrobeDB	800403
Gene Expression(GeneExp)	Transcription factor(TF)	EumicrobeDB	802647
Gene Expression(GeneExp)	Transcription factor(TF)	EumicrobeDB	813941
Gene Expression(GeneExp)	Translation factor(TLF)	EumicrobeDB	807596
Gene Expression(GeneExp)	Translation factor(TLF)	EumicrobeDB	805099
Gene Expression(GeneExp)	Translation factor(TLF)	EumicrobeDB	812409
Gene Expression(GeneExp)	Ribosomal Protein(RB)	EumicrobeDB	810725
Gene Expression(GeneExp)	Ribosomal Protein(RB)	EumicrobeDB	802908
Gene Expression(GeneExp)	Ribosomal Protein(RB)	EumicrobeDB	806632
Metabolism(MT)	Proteasome(Prot)	EumicrobeDB	812818
Metabolism(MT)	Proteasome(Prot)	EumicrobeDB	814368
Metabolism(MT)	Proteasome(Prot)	EumicrobeDB	803605
Metabolism(MT)	ATPase(ATP)	EumicrobeDB	800391
Metabolism(MT)	ATPase(ATP)	EumicrobeDB	803739
Metabolism(MT)	ATPase(ATP)	EumicrobeDB	806272
Metabolism(MT)	Cytochrome(Cytc)	EumicrobeDB	814543
Metabolism(MT)	Cytochrome(Cytc)	EumicrobeDB	812595
Metabolism(MT)	Cytochrome(Cytc)	EumicrobeDB	812904
Structural(ST)	Cytoskeletal(Cytk)	EumicrobeDB	806432
Structural(ST)	Cytoskeletal(Cytk)	EumicrobeDB	810539
Structural(ST)	Organelle(Org)	EumicrobeDB	802202
Structural(ST)	Organelle(Org)	EumicrobeDB	803423
Structural(ST)	Mitochondrion(Mit)	EumicrobeDB	811601
Structural(ST)	Mitochondrion(Mit)	EumicrobeDB	801580

Table S4. List of selected putative housekeeping genes for Hpa bait capture.

	Sequences	Clusters	Max Size	Avg size	Min size	Singletons	% Sequences	% Clusters
Hpa	2504	1855	14	1.3	1	1508	60.2	81.3
Pseudo	372	220	15	2	1	151	40.6	68.6
Ara	9509	589	221	22.4	1	134	1.4	22.8

Table S5. Summary of *uclust* target sequences clustering.

Oomycete genome
Albugo candida AcNc2
Aphanomyces invadans NJM9701
Phytophthora agathidicida NZFS 3772
Phytophthora kernoviae CBS 122049
Phytophthora multivora NZFS 3378
Phytophthora parasitica INRA-310
Phytophthora pinifolia CBS 122922
Phytophthora pluvialis LC9-1
Phytophthora ramorum
Phytophthora sojae P6497
Pilasporangium apinafurcum JCM 30514
Plasmopara viticola INRA-PV221
Pseudoperonospora cubensis MSU-1
Pythium aphanidermatum CBS 132490
Pythium arrhenomanes CBS 324.62
Pythium insidiosum Pi-S
Pythium irregulare CBS 250.28
Pythium iwayamai CBS 132417
Pythium oligandrum Po37
Pythium ultimum var. ultimum DAOM BR144
Pythium vexans CBS 119.80
Saprolegnia diclina VS20
Sclerospora graminicola UoM-SG-Pathotype1

Table S6. List of oomycete genomes (Other oomycetes category) used for relative abundance calculation (shotgun sequencing reads).

Strain	Number of samples where top mapped strain	OTU #	Taxonomic classification
p7.E10	96	OTU1	*Pseudomonas viridiflava*
p11.H11	16	OTU10	Pseudomonas unclassified
p9.C4	9	OTU7	*Pseudomonas viridiflava*
p23.A5	7	OTU6	*Pseudomonas viridiflava*
p7.A9	3	OTU3	*Pseudomonas veronii*
p4.B4	3	OTU3	*Pseudomonas veronii*
p4.D11	3	OTU2	Pseudomonas unclassified
p2.B6	3	OTU9	Pseudomonas unclassified
p8.H7	2	OTU1	*Pseudomonas viridiflava*
p6.D4	1	OTU1	*Pseudomonas viridiflava*
p11.H6	1	OTU1	*Pseudomonas viridiflava*
p11.C6	1	OTU6	*Pseudomonas viridiflava*
p8.G2	1	OTU3	*Pseudomonas veronii*
p2.C11	1	OTU3	*Pseudomonas veronii*
p6.B5	1	OTU2	Pseudomonas unclassified
p2.G2	1	OTU2	Pseudomonas unclassified

Table S7. List of *Pseudomonas* strains and the number of samples in which they were the top mapped strain.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	MAX DP
ARA TAIR10	3.000	5.991	7.424	19.071	10.454	3492.730	38.142
ARA KBS	2.915	6.518	8.509	18.618	13.421	1990.030	37.236
HPA 14OH04	3.537	9.534	11.419	15.136	14.276	1483.800	30.272
HPA EMOY	1.42	10.11	12.38	29.83	17.47	1528.55	59.66
PSEU p7.E10	2.982	4.604	5.101	5.921	5.677	868.860	11.842

Table S8. Variant mean depth for the different reference genomes used to map shotgun sequencing data for variant calling.

Concluding remarks

In the last chapter of this thesis, I present a summary of my main research findings and how they advance our knowledge of the role of genetic diversity in shaping natural plant-pathosystems interaction and coevolution. Moreover, I propose the next logical steps for future research work. In the end, I suggest a research framework for the study of plant-pathosystems evolutionary genetics bringing together insights made at multiple scales.

1. Better be mixed; disease resistance during biological invasions

When species are introduced to new habitats, they commonly suffer a genetic bottleneck, which reduces their diversity. Although diversity is the source of adaptation, we observed species thriving in newly colonized environments. There are multiple proposed hypotheses on how species solve this genetic paradox of invasion and manage to adapt.

In the first chapter of my thesis, I used populations of *A. thaliana* that colonized North America in historic times to investigate the genetic sources of disease resistance after a diversity bottleneck. By revealing the extent and distribution of introduced diversity, I found evidence of gene flow from native sources. These newly introduced individuals, through outcrossing, got mixed and introgressed with the colonizing lineage. These different genetic groups have different levels of Hpa disease resistance, being the colonizing lineage, HPG1, the most susceptible one to the isolates tested. This finding supports the expectation that traits controlled by major effect loci, such as disease resistance to specialist pathogens, would be the most impacted by genetic bottlenecks. Finally, the introgression and mapping analysis of a local N. American population provide evidence towards the progression of disease resistance loci from native sources. Collectively, these findings reveal the importance of outcrossing and gene flow from native sources for the adaptation to pathogen pressure in colonizing populations (**Figure 1**). Future research should further develop and confirm these initial findings

by measuring the adaptive value of introduced variation and the exact fitness effect of pathogen infections in natural settings.

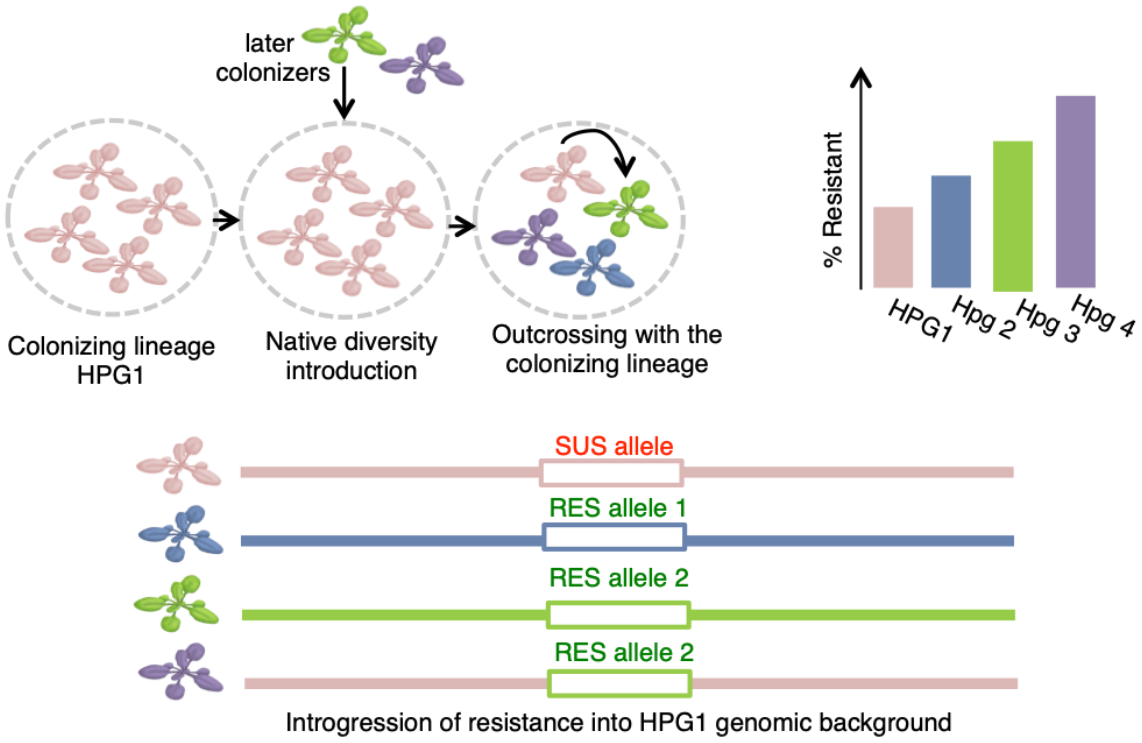


Figure 1. The effect of gene flow in disease resistance in colonizing populations.

The genetic paradox of invasion can be solved by introducing native diversity into the initially established populations. Once new diversity is brought into these populations through outcrossing, new genotypes are generated. These new haplogroups exhibit different resistance levels, having, in principle, a competitive advantage compared to the colonizing lineage. Thanks to outcrossing, the colonizing lineage imports new variation that can be potentially adaptive and kept in its genomic background through introgression. Disease resistance alleles as large-effect loci are good candidates to be favored by introgression.

2. Global versus local; fine-mapping of Hpa disease resistance

The study of natural variation can help us estimate wild populations' adaptive potential and achieve durable resistance in crops. Therefore, recent efforts have been made towards the identification of variants underlying natural variation in disease resistance. In this context, mapping disease resistance genes through experimental crosses in *A. thaliana* has been an ongoing effort. In particular, the study of natural variation underlying disease resistance to the biotrophic pathogen

Hpa has led to significant findings on several fronts. For example, we have learned that resistance to specific Hpa isolates is mediated by single dominant resistant genes in the host. Thus, the genetic interaction between *A. thaliana* and Hpa often follows Flor's gene-for-gene model of interaction[79]. Under the assumption that simple genetics governs the Hpa disease resistance, the second aim of the first chapter was to identify and compare natural variation in disease resistance in two metapopulations of *A. thaliana* representing the introduced and native range of the species. Instead of using only experimental crosses, I aimed to leverage the power of Genome-Wide Association (GWA) mapping to map resistance loci. By running GWA analysis on a global set of individuals from different locations and genetic groups, I could not detect significant associations. Which led us to hypothesize that the genetics of Hpa disease resistance might be complex. On the other hand, when experimentally crossing accessions or performing GWA mapping on a single population, I could fine-map genetic regions known to be involved in Hpa disease resistance. These findings add to a growing corpus of research showing that the use of GWA mapping on spatially and genetically structured populations might not be the best strategy to identify disease resistance loci [56,60]. Instead, future studies should try to mitigate the effect of allelic heterogeneity and population structure in global GWA. Our work suggests that local GWA is a promising approach for mapping ecologically relevant disease resistance variation by challenging a wild population with sympatric pathogen lines.

3. Finding the needle in the haystack; target enrichment sequencing as a promising population genetics tool

To understand coevolution between host resistance genes and pathogen effector genes, we need to interrogate diversity and distribution of both in the same populations. Unfortunately, the simultaneous study of host resistance genes and pathogen effectors in natural populations has previously encountered multiple challenges. In particular, there are technical challenges associated with sample composition and complexity, the structural characteristics of the targeted genes, and downstream mapping analysis. The second chapter's aim was first to develop a new approach to successfully enrich disease resistance genes and pathogen effectors

within the same sample. Second, to demonstrate that target enrichment sequencing (TES) can be used to interrogate large-scale presence and absence variation of disease resistance genes and pathogen effectors in the *A. thaliana* - Hpa pathosystem. The results from chapter two are broadly consistent with previous studies applying TES for the study of oomycete pathogens. I was able to capture the desired species and target genes successfully. Moreover, from the study of presence/absence variation, one can infer evolutionary signatures associated with distinct allelic frequencies within and among populations. One promising application of combining PenSeq and RenSeq can be the characterization of pathogens' core genome versus accessory genome, revealing which effectors are conserved and which ones are dynamic. The concept of core vs. accessory genome has been already investigated in fungal pathogens, but it is still an underexplored field for oomycetes [242,243]. This concept is especially relevant for pathogens that cause disease in economically important crops because it can help predict the successful deployment of resistant crop varieties. Target enrichment sequencing can also be leveraged to identify new allelic variants of known effectors and, therefore, predict key protein regions and amino acids involved in their physical interaction with resistance genes[244]. Because target enrichment sequencing reduces sequencing costs, it can be used to scale up interaction studies between not just one pathogen with its host but also the entire microbiome effectorome present on multiple populations or a species.

4. Towards a systems-biology approach to investigate plant pathosystems

The field of evolutionary genetics of natural plant-pathosystems is shifting towards an integrative view of evolutionary processes that govern interactions at multiple levels. Although insightful, the conclusions drawn from this study have come from investigating genotype-by-genotype interactions. This is still a reductionist approach considering the complex nature of plant pathosystems in terms of the diverse interacting factors, including time, space, and environment.

Environmental factors are the understudied side of the disease triangle, and their effect as disease drivers should not be underestimated [245]. In crop pathosystems, the environment plays a crucial role in determining disease incidence, and it is the cause of significant yield loss. On the other hand, studies investigating the environmental aspect of plant-pathogen interactions in natural populations are sparse. Still, we know from molecular studies in plants that temperature modulates disease resistance. The rise of global temperatures associated with climate change is expected to increase plant disease severity and incidence [246]. By investigating the effect of temperature in wild plant-pathogen interactions, we might gain valuable insights that help alleviate the effects of climate change in agricultural settings. Moreover, airborne plant pathogens rely on rain and optimal humidity conditions to grow and propagate. It will be critical then to also look at how shifts in rain regimes and overall humidity levels could impact pathogens' infectivity. Thus, future research should explore in more detail the link between environmental variables and disease incidence in natural populations, for instance, by investigating natural populations along environmental clines and simulating environmental effects in common garden experiments. Overall, one of the next aims for future studies should be integrating the three sides of the disease triangle model of interaction, investigating the effects of host and pathogen genetic diversity and the ecological and environmental aspects of their interaction.

By definition, evolution is change over time. Thus, including temporal aspects in the study of plant-pathogen coevolution is vital. Although we can infer from molecular signatures the effect of past evolutionary forces acting upon populations and genomes, we are often left alone with a temporal snapshot, which might not reflect diverse evolutionary processes' exact contribution. The Geographic Mosaic Theory of Coevolution (GMTC) postulates that coevolution will occur at different temporal scales, leading to coevolutionary hot and cold spots [68]. Therefore, there is an imperative need to look at populations over time to infer the effect of natural selection and other evolutionary forces. Recent attempts to incorporate the temporal aspect to the study of adaptation in *A. thaliana* populations include evolution

experiments such as the GrENE Project [247], the use of herbarium genomics [54], and the recurrent visit and sampling of natural populations such as Pathodopsis [248].

Individuals are found within populations and these populations within a range of environments. Thus, natural populations have an uneven distribution of individuals and exhibit different population structure and genetic diversity levels. The geographic aspect of coevolution is also emphasized in the GMTC and led to the development of the metapopulation concept [249]. We have seen that disease prevalence and local adaptation schemes can change significantly in spatially structured populations. Overall, my results highlight the effect of host population structure and evolutionary history in determining disease resistance. This finding would not have been possible without looking at multiple populations. It is a question of future research to investigate how both hosts and pathogens metapopulations interact and differ from one another.

As a result of the insights from my work, I propose the evolutionary disease triangle as a framework for future studies of plant pathosystems (**Figure 2**). It combines static principles from the traditional disease triangle, the interaction between host, pathogen, and environment, with populations' spatial and temporal aspects. Overall, it provides a systems-level approach in which we integrate knowledge from individual genes, genomes, and haplogroups from pathogens and hosts with local and global environmental factors. All these factors are integrated at different temporal and geographic layers ranging from metapopulations to individuals.

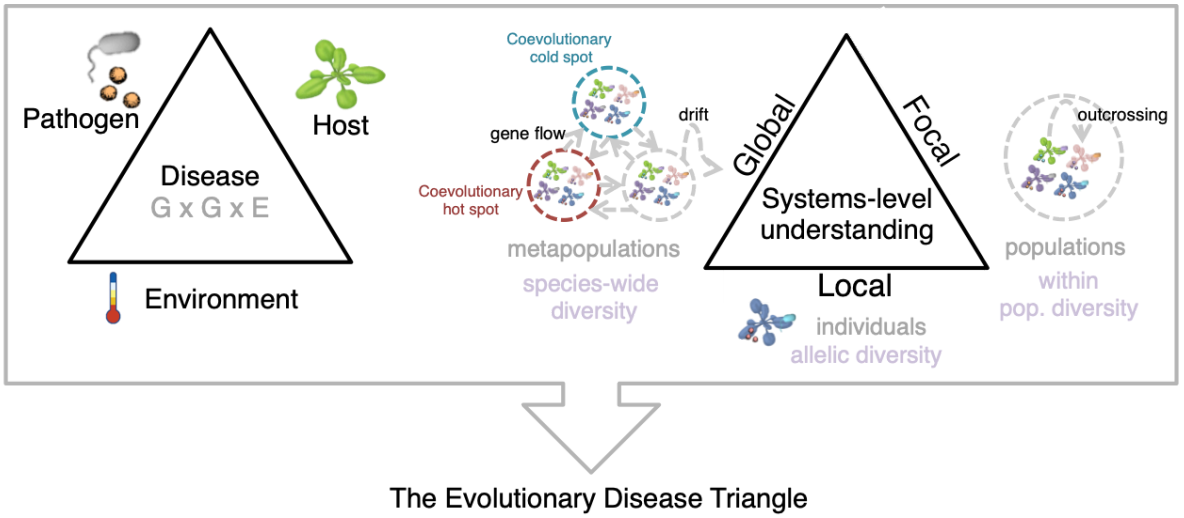


Figure 2. The evolutionary disease triangle as a proposed framework for the study of plant-pathogen evolutionary dynamics.

On the left, the traditional disease triangle model explains that the disease outcome depends on the interaction of three main factors; the pathogen, the host, and the environment. **On the right**, the disease triangle model in the context of population genetics and the geographic mosaic theory of coevolution. The evolutionary disease triangle is the proposed framework for future studies where the interactions between pathogens, hosts, and the environment should be investigated at different scales, from metapopulations to single genes, combining principles of epidemiology, molecular genetics, and evolutionary genetics.

List of abbreviations

ATR - *Arabidopsis thaliana* recognized

DM - Dangerous-mix

ETI - Effector Triggered Immunity

GWA - Genome-wide association

HMM - Hidden Markov Model

HR - Hypersensitive response

Hpa - Hyaloperonospora arabidopsidis

Hpg1 - Haplogroup 1

INDEL - Insertions and Deletions

MRC - Major Recognition Complexes

NGS - Next generation sequencing

NLP - Nep1-Like Proteins

NLR - CC-NB-LRR resistance gene

OTU - Operational Taxonomic Unit

P/A - Presence / Absence

PAMP - Pathogen-associated Molecular Pattern

PCR - Polymerase chain reaction

PRenSeq - Pen-seq and Ren-seq combined

PTI - PAMP-triggered immunity

Pen-Seq - Pathogen enrichment sequencing

Pseudomonas - *Pseudomonas* species (general)

QTL - Quantitative trait loci

R-gene - NB-LRR resistance gene

RA - Relative Abundance

RAD-sequencing - Restriction site Associated sequencing

ROS - Reactive oxygen species

RPP - Resistance to Peronospora parasitica

Ren-Seq - R-gene enrichment sequencing

SNP - Single Nucleotide Polymorphism

TES - Target enrichment sequencing

WGS - Whole genome sequencing

References

1. Darwin, C. The Origin of Species. (1959) doi:10.9783/9780812200515.
2. Mendel, G. Experiments in plant hybridization (translation). *The Origins of Genetics: A Mendel Source Book* (1866).
3. Huxley, J. S. Evolution, the Modern Synthesis. *Synthesis* **2**, (1942).
4. Wright, S. Evolution in Mendelian Populations. *Genetics* **16**, 97–159 (1931).
5. Haldane, J. B. S. The Time of Action of Genes, and Its Bearing on some Evolutionary Problems. *Am. Nat.* **66**, 5–24 (1932).
6. Fisher, R. A. The genetical theory of natural selection. (1930) doi:10.5962/bhl.title.27468.
7. Lewontin, R. C. & Others. *The genetic basis of evolutionary change*. vol. 560 (Columbia University Press New York, 1974).
8. Lewontin, R. C. & Hubby, J. L. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. *Genetics* **54**, 595–609 (1966).
9. Harris, H. C. Genetics of Man Enzyme polymorphisms in man. *Proceedings of the Royal Society of London. Series B. Biological Sciences* **164**, 298–310 (1966).
10. Nei, M. & Tajima, F. Maximum likelihood estimation of the number of nucleotide substitutions from restriction sites data. *Genetics* **105**, 207–217 (1983).
11. Langley, C. H., Montgomery, E. & Quattlebaum, W. F. Restriction map variation in the Adh region of Drosophila. *Proc. Natl. Acad. Sci. U. S. A.* **79**, 5631–5635 (1982).
12. Kreitman, M. Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster. *Nature* **304**, 412–417 (1983).
13. Consortium, T. 1000 G. P. & The 1000 Genomes Project Consortium. An integrated map

of genetic variation from 1,092 human genomes. *Nature* vol. 491 56–65 (2012).

14. 1001 Genomes Consortium. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **166**, 481–491 (2016).
15. Weigel, D. Natural variation in Arabidopsis: from molecular genetics to ecological genomics. *Plant Physiol.* **158**, 2–22 (2012).
16. Elton, C. S. *The ecology of invasions by plants and animals*. (Methuen, 1958).
17. Baker, H. G. & Stebbins, G. L. *The Genetics of Colonizing Species: Proceedings of the First International Union of Biological Sciences Symposia on General Biology*. (Academic Press, 1965).
18. Barrett, S. C. H. Foundations of invasion genetics: the Baker and Stebbins legacy. *Mol. Ecol.* **24**, 1927–1941 (2015).
19. Ward, S. M., Gaskin, J. F. & Wilson, L. M. Ecological Genetics of Plant Invasion: What Do We Know? *ipsm* **1**, 98–109 (2008).
20. Colautti, R. I. *et al.* Invasion genetics of the Eurasian spiny waterflea: evidence for bottlenecks and gene flow using microsatellites. *Mol. Ecol.* **14**, 1869–1879 (2005).
21. Colautti, R. I., Alexander, J. M., Dlugosch, K. M., Keller, S. R. & Sultan, S. E. Invasions and extinctions through the looking glass of evolutionary ecology. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372**, (2017).
22. Dlugosch, K. M. & Parker, I. M. Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. *Mol. Ecol.* **17**, 431–449 (2008).
23. Estoup, A. *et al.* Is There a Genetic Paradox of Biological Invasion? *Annu. Rev. Ecol. Evol. Syst.* **47**, 51–72 (2016).
24. Barrett, R. D. H. & Schluter, D. Adaptation from standing genetic variation. *Trends Ecol. Evol.* **23**, 38–44 (2008).
25. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74**, 1111–1120 (2004).

26. Ossowski, S. *et al.* The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *Science* **327**, 92–94 (2010).
27. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* vol. 290 1151–1155 (2000).
28. Dlugosch, K. M., Anderson, S. R., Braasch, J., Cang, F. A. & Gillette, H. D. The devil is in the details: genetic variation in introduced populations and its contributions to invasion. *Mol. Ecol.* **24**, 2095–2111 (2015).
29. Rius, M. & Darling, J. A. How important is intraspecific genetic admixture to the success of colonising populations? *Trends Ecol. Evol.* **29**, 233–242 (2014).
30. Racimo, F., Sankararaman, S., Nielsen, R. & Huerta-Sánchez, E. Evidence for archaic adaptive introgression in humans. *Nat. Rev. Genet.* **16**, 359–371 (2015).
31. Schmickl, R., Marburger, S., Bray, S. & Yant, L. Hybrids and horizontal transfer: introgression allows adaptive allele discovery. *J. Exp. Bot.* **68**, 5453–5470 (2017).
32. Whitney, K. D. *et al.* Quantitative trait locus mapping identifies candidate alleles involved in adaptive introgression and range expansion in a wild sunflower. *Mol. Ecol.* **24**, 2194–2211 (2015).
33. Arnold, B. J. *et al.* Borrowed alleles and convergence in serpentine adaptation. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 8320–8325 (2016).
34. Stankowski, S. & Streisfeld, M. A. Introgressive hybridization facilitates adaptive divergence in a recent radiation of monkeyflowers. *Proc. Biol. Sci.* **282**, (2015).
35. Barker, B. S. *et al.* Potential limits to the benefits of admixture during biological invasion. *Mol. Ecol.* **28**, 100–113 (2019).
36. Nzuki, I. *et al.* QTL Mapping for Pest and Disease Resistance in Cassava and Coincidence of Some QTL with Introgression Regions Derived from Manihot glaziovii. *Front. Plant Sci.* **8**, 1168 (2017).
37. Gentzbittel, L. *et al.* WhoGEM: an admixture-based prediction machine accurately

predicts quantitative functional traits in plants. *Genome Biol.* **20**, 106 (2019).

38. Hahn, M. A. & Rieseberg, L. H. Genetic admixture and heterosis may enhance the invasiveness of common ragweed. *Evol. Appl.* **10**, 241–250 (2017).
39. Liu, H. & Stiling, P. Testing the enemy release hypothesis: a review and meta-analysis. *Biol. Invasions* **8**, 1535–1545 (2006).
40. Mitchell, C. E. & Power, A. G. Release of invasive plants from fungal and viral pathogens. *Nature* **421**, 625–627 (2003).
41. Maron, J. L. & Vilà, M. When Do Herbivores Affect Plant Invasion? Evidence for the Natural Enemies and Biotic Resistance Hypotheses. *Oikos* **95**, 361–373 (2001).
42. Godfree, R. C., Thrall, P. H. & Young, A. G. Enemy release after introduction of disease-resistant genotypes into plant-pathogen systems. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 2756–2760 (2007).
43. Weigel, D. & Nordborg, M. Population Genomics for Understanding Adaptation in Wild Plant Species. *Annu. Rev. Genet.* **49**, 315–338 (2015).
44. Cao, J. *et al.* Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat. Genet.* **43**, 956–963 (2011).
45. Fournier-Level, A. *et al.* A map of local adaptation in Arabidopsis thaliana. *Science* **334**, 86–89 (2011).
46. Hancock, A. M. *et al.* Adaptation to climate across the Arabidopsis thaliana genome. *Science* **334**, 83–86 (2011).
47. Krämer, U. Planting molecular functions in an ecological context with Arabidopsis thaliana. *Elife* **4**, (2015).
48. Durvasula, A. *et al.* African genomes illuminate the early history and transition to selfing in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 5213–5218 (2017).
49. Platt, A. *et al.* The scale of population structure in Arabidopsis thaliana. *PLoS Genet.* **6**, (2010).

50. Lee, C.-R. *et al.* On the post-glacial spread of human commensal Arabidopsis thaliana. *Nat. Commun.* **8**, 14458 (2017).
51. Shimizu, K. K. *et al.* Darwinian selection on a selfing locus. *Science* **306**, 2081–2084 (2004).
52. Chae, E. *et al.* Species-wide genetic incompatibility analysis identifies immune genes as hot spots of deleterious epistasis. *Cell* **159**, 1341–1351 (2014).
53. Hamilton, J. A., Okada, M., Korves, T. & Schmitt, J. The role of climate adaptation in colonization success in Arabidopsis thaliana. *Mol. Ecol.* **24**, 2253–2263 (2015).
54. Exposito-Alonso, M. *et al.* The rate and potential relevance of new mutations in a colonizing plant lineage. *PLoS Genet.* **14**, e1007155 (2018).
55. A Catalog of Arabidopsis thaliana Genetic Variation. https://1001genomes.org/.
56. Burghardt, L. T., Young, N. D. & Tiffin, P. A Guide to Genome-Wide Association Mapping in Plants. *Curr Protoc Plant Biol* **2**, 22–38 (2017).
57. Aranzana, M. J. *et al.* Genome-wide association mapping in Arabidopsis thaliana identifies previously known genes responsible for variation in flowering time and pathogen resistance. *PLoS Genetics* vol. preprint e60 (2005).
58. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* **465**, 627–631 (2010).
59. Karasov, T. L. *et al.* The long-term maintenance of a resistance polymorphism through diffuse interactions. *Nature* **512**, 436–440 (2014).
60. Bartoli, C. & Roux, F. Genome-Wide Association Studies In Plant Pathosystems: Toward an Ecological Genomics Approach. *Front. Plant Sci.* **8**, 763 (2017).
61. Iakovidis, M. *et al.* Effector-Triggered Immune Response in Arabidopsis thaliana Is a Quantitative Trait. *Genetics* **204**, 337–353 (2016).
62. Wilson, I. W., Schiff, C. L., Hughes, D. E. & Somerville, S. C. Quantitative trait loci analysis of powdery mildew disease resistance in the Arabidopsis thaliana accession

kashmir-1. *Genetics* **158**, 1301–1309 (2001).

63. Rajarammohan, S. *et al.* Genetic Architecture of Resistance to Alternaria brassicae in Arabidopsis thaliana: QTL Mapping Reveals Two Major Resistance-Conferring Loci. *Front. Plant Sci.* **8**, 260 (2017).
64. Nemri, A. *et al.* Genome-wide survey of Arabidopsis natural variation in downy mildew resistance using combined association and linkage mapping. *Proceedings of the National Academy of Sciences* **107**, 10302–10307 (2010).
65. Suarez-Gonzalez, A., Lexer, C. & Cronk, Q. C. B. Adaptive introgression: a plant perspective. *Biol. Lett.* **14**, (2018).
66. Wade, M. J. 11 - Selection in Metapopulations: The Coevolution of Phenotype and Context. in *Ecology, Genetics and Evolution of Metapopulations* (eds. Hanski, I. & Gaggiotti, O. E.) 259–273 (Academic Press, 2004).
67. Stevens, R. B. Pages 357-429 in: Plant Pathology, an Advanced Treatise, Vol. 3. JG Horsfall and AE Dimond, eds. (1960).
68. Gomulkiewicz, R., Thompson, J. N., Holt, R. D., Nuismer, S. L. & Hochberg, M. E. Hot Spots, Cold Spots, and the Geographic Mosaic Theory of Coevolution. *Am. Nat.* **156**, 156–174 (2000).
69. Thompson, J. N. The Geographic Mosaic of Coevolution. *University of Chicago Press* https://press.uchicago.edu/ucp/books/book/chicago/G/bo3533766.html (2005).
70. Laine, A.-L. Resistance variation within and among host populations in a plant–pathogen metapopulation: implications for regional pathogen dynamics. *J. Ecol.* **92**, 990–1000 (2004).
71. Laine, A.-L. Spatial scale of local adaptation in a plant-pathogen metapopulation. *J. Evol. Biol.* **18**, 930–938 (2005).
72. Tack, A. J. M., Thrall, P. H., Barrett, L. G., Burdon, J. J. & Laine, A.-L. Variation in infectivity and aggressiveness in space and time in wild host–pathogen systems: causes

and consequences. *J. Evol. Biol.* **25**, 1918–1936 (2012).

73. Laine, A.-L., Burdon, J. J., Dodds, P. N. & Thrall, P. H. Spatial variation in disease resistance: from molecules to metapopulations. *J. Ecol.* **99**, 96–112 (2011).
74. Tack, A. J. M. & Laine, A.-L. Ecological and evolutionary implications of spatial heterogeneity during the off-season for a wild plant pathogen. *New Phytol.* **202**, 297–308 (2014).
75. Laine, A.-L. Role of coevolution in generating biological diversity: spatially divergent selection trajectories. *J. Exp. Bot.* **60**, 2957–2970 (2009).
76. Thrall, P. H. *et al.* Rapid genetic change underpins antagonistic coevolution in a natural host-pathogen metapopulation. *Ecol. Lett.* **15**, 425–435 (2012).
77. Nemri, A., Barrett, L. G., Laine, A.-L., Burdon, J. J. & Thrall, P. H. Population processes at multiple spatial scales maintain diversity and adaptation in the Linum marginale--Melampsora lini association. *PLoS One* **7**, e41366 (2012).
78. Borer, E. T., Laine, A.-L. & Seabloom, E. W. A Multiscale Approach to Plant Disease Using the Metacommunity Concept. *Annu. Rev. Phytopathol.* **54**, 397–418 (2016).
79. Flor, H. H. The Complementary Genic Systems in Flax and Flax Rust. in *Advances in Genetics* (ed. Demerec, M.) vol. 8 29–54 (Academic Press, 1956).
80. Ravensdale, M., Nemri, A., Thrall, P. H., Ellis, J. G. & Dodds, P. N. Co-evolutionary interactions between host resistance and pathogen effector genes in flax rust disease. *Mol. Plant Pathol.* **12**, 93–102 (2011).
81. Bernoux, M., Moncuquet, P., Kroj, T., Dodds, P. N. & Others. A novel conserved mechanism for plant NLR protein pairs: the ‘integrated decoy hypothesis’. *Front. Plant Sci.* **5**, 606 (2014).
82. Stahl, E. A., Dwyer, G., Mauricio, R., Kreitman, M. & Bergelson, J. Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. *Nature* **400**, 667–671 (1999).

83. Bergelson, J., Dwyer, G. & Emerson, J. J. Models and data on plant-enemy coevolution. *Annu. Rev. Genet.* **35**, 469–499 (2001).
84. Brown, J. K. M. & Tellier, A. Plant-parasite coevolution: bridging the gap between genetics and ecology. *Annu. Rev. Phytopathol.* **49**, 345–367 (2011).
85. Tellier, A., Moreno-Gámez, S. & Stephan, W. Speed of adaptation and genomic footprints of host-parasite coevolution under arms race and trench warfare dynamics. *Evolution* **68**, 2211–2224 (2014).
86. Slusarenko, A. J. & Schlaich, N. L. Downy mildew of Arabidopsis thaliana caused by Hyaloperonospora parasitica (formerly Peronospora parasitica). *Mol. Plant Pathol.* **4**, 159–170 (2003).
87. Coates, M. E. & Beynon, J. L. Hyaloperonospora Arabidopsidis as a pathogen model. *Annu. Rev. Phytopathol.* **48**, 329–345 (2010).
88. McDowell, J. M. Hyaloperonospora arabidopsidis: A Model Pathogen of Arabidopsis. in *Genomics of Plant-Associated Fungi and Oomycetes: Dicot Pathogens* (eds. Dean, R. A., Lichens-Park, A. & Kole, C.) 209–234 (Springer Berlin Heidelberg, 2014).
89. Dodds, P. N. & Rathjen, J. P. Plant immunity: towards an integrated view of plant–pathogen interactions. *Nat. Rev. Genet.* **11**, 539–548 (2010).
90. Jones, J. D. G. & Dangl, J. L. The plant immune system. *Nature* **444**, 323–329 (2006).
91. Dangl, J. L. emergence of Arabidopsis thaliana as a model for plant-pathogen interactions. *Adv. Plant Pathol.* **10**, (1993).
92. Koornneef, M., Alonso-Blanco, C. & Vreugdenhil, D. Naturally occurring genetic variation in Arabidopsis thaliana. *Annu. Rev. Plant Biol.* **55**, 141–172 (2004).
93. Shindo, C., Bernasconi, G. & Hardtke, C. S. Natural genetic variation in Arabidopsis: tools, traits and prospects for evolutionary ecology. *Ann. Bot.* **99**, 1043–1054 (2007).
94. Kawakatsu, T. *et al.* Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. *Cell* **166**, 492–505 (2016).

95. van Leeuwen, H. *et al.* Natural variation among Arabidopsis thaliana accessions for transcriptome response to exogenous salicylic acid. *Plant Cell* **19**, 2099–2110 (2007).
96. Salvaudon, L., Héraudet, V. & Shykoff, J. A. Genotype-specific interactions and the trade-off between host and parasite fitness. *BMC Evol. Biol.* **7**, 189 (2007).
97. Ahmad, S., Gordon-Weeks, R., Pickett, J. & Ton, J. Natural variation in priming of basal resistance: from evolutionary origin to agricultural exploitation. *Mol. Plant Pathol.* **11**, 817–827 (2010).
98. Holub, E. B., Beynon, J. L. & Crute, I.R. (Department of Plant Pathology and Weed Science, Horticulture Research International-East Malling, West Malling, Kent ME19 6BJ (UK)). Phenotypic and genotypic characterization of interactions between isolates of Peronospora parasitica and accessions of Arabidopsis thaliana. *Molecular Plant-Microbe Interactions (United Kingdom)* **7**, (1994).
99. Parker, J. E. *et al.* Characterization of eds1, a mutation in Arabidopsis suppressing resistance to Peronospora parasitica specified by several different RPP genes. *Plant Cell* **8**, 2033–2046 (1996).
100. Botella, M. A. *et al.* Three genes of the Arabidopsis RPP1 complex resistance locus recognize distinct Peronospora parasitica avirulence determinants. *Plant Cell* **10**, 1847–1860 (1998).
101. Holub, E. B. The arms race is ancient history in Arabidopsis, the wildflower. *Nat. Rev. Genet.* **2**, 516–527 (2001).
102. Van Damme, M. *et al.* Identification of arabidopsis loci required for susceptibility to the downy mildew pathogen Hyaloperonospora parasitica. *Mol. Plant. Microbe. Interact.* **18**, 583–592 (2005).
103. Holub, E. B. Natural variation in innate immunity of a pioneer species. *Curr. Opin. Plant Biol.* **10**, 415–424 (2007).
104. Krasileva, K. V. *et al.* Global analysis of Arabidopsis/downy mildew interactions reveals

prevalence of incomplete resistance and rapid evolution of pathogen recognition. *PLoS One* **6**, e28765 (2011).

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

106. Bakker, E. G., Toomajian, C., Kreitman, M. & Bergelson, J. A Genome-Wide Survey of R Gene Polymorphisms in Arabidopsis. *Plant Cell* **18**, 1803–1818 (2006).

107. Hall, S. A. *et al.* Maintenance of genetic variation in plants and pathogens involves complex networks of gene-for-gene interactions. *Mol. Plant Pathol.* **10**, 449–457 (2009).

108. Cabral, A. *et al.* Identification of Hyaloperonospora arabidopsidis transcript sequences expressed during infection reveals isolate-specific effectors. *PLoS One* **6**, e19328 (2011).

109. Dong, S., Raffaele, S. & Kamoun, S. The two-speed genomes of filamentous pathogens: waltz with plants. *Curr. Opin. Genet. Dev.* **35**, 57–65 (2015).

110. Steuernagel, B., Jupe, F., Witek, K., Jones, J. D. G. & Wulff, B. B. H. NLR-parser: rapid annotation of plant NLR complements. *Bioinformatics* **31**, 1665–1667 (2015).

111. Tabima, J. F. & Grünwald, N. J. effectR: An Expandable R Package to Predict Candidate RxLR and CRN Effectors in Oomycetes Using Motif Searches. *Mol. Plant. Microbe. Interact.* **32**, 1067–1076 (2019).

112. Baggs, E., Dagdas, G. & Krasileva, K. V. NLR diversity, helpers and integrated domains: making sense of the NLR IDentity. *Curr. Opin. Plant Biol.* **38**, 59–67 (2017).

113. Białas, A. *et al.* Lessons in Effector and NLR Biology of Plant-Microbe Systems. *Mol. Plant. Microbe. Interact.* **31**, 34–45 (2018).

114. Weigel, D. All in the Family: The First Whole-Genome Survey of NLR Genes. *Plant Cell* **31**, 1212–1213 (2019).

115. Mamanova, L. *et al.* Target-enrichment strategies for next-generation sequencing. *Nat. Methods* **7**, 111–118 (2010).

116. Arora, S. *et al.* Resistance gene cloning from a wild crop relative by sequence capture and association genetics. *Nat. Biotechnol.* **37**, 139–143 (2019).

117. Witek, K. *et al.* Accelerated cloning of a potato late blight-resistance gene using RenSeq and SMRT sequencing. *Nat. Biotechnol.* **34**, 656–660 (2016).

118. Giolai, M. *et al.* Targeted capture and sequencing of gene-sized DNA molecules. *Biotechniques* **61**, 315–322 (2016).

119. Giolai, M. *et al.* Comparative analysis of targeted long read sequencing approaches for characterization of a plant's immune receptor repertoire. *BMC Genomics* **18**, 564 (2017).

120. Armstrong, M. R. *et al.* Tracking disease resistance deployment in potato breeding by enrichment sequencing. *Plant Biotechnol. J.* **17**, 540–549 (2019).

121. Jouet, A. *et al.* Albugo candida race diversity, ploidy and host-associated microbes revealed using DNA sequence capture on diseased plants in the field. *New Phytol.* (2018) doi:10.1111/nph.15417.

122. Thilliez, G. J. A. *et al.* Pathogen enrichment sequencing (PenSeq) enables population genomic studies in oomycetes. *New Phytol.* (2018) doi:10.1111/nph.15441.

123. Kale, S. D. PenSeq: coverage you can count on. *New Phytol.* **221**, 1177–1179 (2019).

124. Jupe, F. *et al.* 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. *Plant J.* **76**, 530–544 (2013).

125. Van de Weyer, A.-L. *et al.* The Arabidopsis thaliana pan-NLRome. *bioRxiv* 537001 (2019) doi:10.1101/537001.

126. Karasov, T. L., Neumann, M. & Duque-Jaramillo, A. The relationship between microbial population size and disease in the Arabidopsis thaliana phyllosphere. *BioRxiv* (2020).

127. Lundberg, D. S., Ayutthaya, P. P. N., Strauss, A. & Shirsekar, G. Measuring both microbial load and diversity with a single amplicon sequencing library. *bioRxiv* (2020).

128. Colautti, R. I. & Lau, J. A. Contemporary evolution during invasion: evidence for

differentiation, natural selection, and local adaptation. *Mol. Ecol.* **24**, 1999–2017 (2015).

129. Goulet, B. E., Roda, F. & Hopkins, R. Hybridization in Plants: Old Ideas, New Techniques. *Plant Physiol.* **173**, 65–78 (2017).

130. Hagmann, J. *et al.* Century-scale Methylome Stability in a Recently Diverged Arabidopsis thaliana Lineage. *PLoS Genet.* **11**, (2015).

131. Heidel, A. J., Clarke, J. D., Antonovics, J. & Dong, X. Fitness costs of mutations affecting the systemic acquired resistance pathway in Arabidopsis thaliana. *Genetics* **168**, 2197–2206 (2004).

132. Karasov, T. L., Horton, M. W. & Bergelson, J. Genomic variability as a driver of plant-pathogen coevolution? *Curr. Opin. Plant Biol.* **18**, 24–30 (2014).

133. Nordborg, M. *et al.* The pattern of polymorphism in Arabidopsis thaliana. *PLoS Biol.* **3**, e196 (2005).

134. Sinapidou, E. *et al.* Two TIR:NB:LRR genes are required to specify resistance to Peronospora parasitica isolate Cala2 in Arabidopsis. *Plant J.* **38**, 898–909 (2004).

135. Parker, J. E. *et al.* The Arabidopsis downy mildew resistance gene RPP5 shares similarity to the toll and interleukin-1 receptors with N and L6. *Plant Cell* **9**, 879–894 (1997).

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

137. McDowell, J. M. *et al.* Intragenic recombination and diversifying selection contribute to the evolution of downy mildew resistance at the RPP8 locus of Arabidopsis. *Plant Cell* **10**, 1861–1874 (1998).

138. Holub, E. B. & Beynon, J. L. Symbiology of Mouse-Ear Cress (Arabidopsis Thaliana) and Oomycetes. in *Advances in Botanical Research* (eds. Andrews, J. H., Tommerup, I. C. &

Callow, J. A.) vol. 24 227–273 (Academic Press, 1997).

139. Speulman, E., Bouchez, D., Holub, E. B. & Beynon, J. L. Disease resistance gene homologs correlate with disease resistance loci ofArabidopsis thaliana. *The Plant Journal* vol. 14 467–474 (1998).

140. Dixon, R. A. *et al.* The phenylpropanoid pathway and plant defence-a genomics perspective. *Molecular Plant Pathology* vol. 3 371–390 (2002).

141. Xie, M. *et al.* Regulation of Lignin Biosynthesis and Its Role in Growth-Defense Tradeoffs. *Front. Plant Sci.* **9**, 1427 (2018).

142. Kim, S. *et al.* Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. *Nat. Genet.* **46**, 270–278 (2014).

143. Rojas, C. M. *et al.* Glycolate oxidase modulates reactive oxygen species-mediated signal transduction during nonhost resistance in Nicotiana benthamiana and Arabidopsis. *Plant Cell* **24**, 336–352 (2012).

144. Bomblies, K. *et al.* Autoimmune response as a mechanism for a Dobzhansky-Muller-type incompatibility syndrome in plants. *PLoS Biol.* **5**, e236 (2007).

145. Shirano, Y., Kachroo, P., Shah, J. & Klessig, D. F. A Gain-of-Function Mutation in an Arabidopsis Toll Interleukin1 Receptor–Nucleotide Binding Site–Leucine-Rich Repeat Type R Gene Triggers Defense Responses and Results in Enhanced Disease Resistance. *Plant Cell* **14**, 3149–3162 (2002).

146. Helliwell, C. A. *et al.* The Arabidopsis AMP1 gene encodes a putative glutamate carboxypeptidase. *Plant Cell* **13**, 2115–2125 (2001).

147. Lee, M. W. *et al.* ALTERED MERISTEM PROGRAM1 has conflicting effects on the tolerance to heat shock and symptom development after Pseudomonas syringae infection. *Biochem. Biophys. Res. Commun.* **480**, 296–301 (2016).

148. Paniagua, C. *et al.* Dirigent proteins in plants: modulating cell wall metabolism during abiotic and biotic stress exposure. *J. Exp. Bot.* **68**, 3287–3301 (2017).

149. Tang, D., Ade, J., Frye, C. A. & Innes, R. W. Regulation of plant defense responses in Arabidopsis by EDR2, a PH and START domain-containing protein. *Plant J.* (2005).

150. Reignault, P. *et al.* Four Arabidopsis RPP loci controlling resistance to the Noco2 isolate of Peronospora parasitica map to regions known to contain other RPP recognition specificities. *MPMI-Molecular Plant Microbe Interactions* **9**, 464–473 (1996).

151. Noël, L. *et al.* Pronounced intraspecific haplotype divergence at the RPP5 complex disease resistance locus of Arabidopsis. *Plant Cell* **11**, 2099–2111 (1999).

152. Van Der Biezen, E. A., Freddie, C. T., Kahn, K. & Parker\P. Arabidopsis RPP4 is a member of the RPP5 multigene family of TIR-NB-LRR genes and confers downy mildew resistance through multiple signalling components. *Plant J.* **29**, 439–451 (2002).

153. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

154. Exposito-Alonso, M. *et al.* Genomic basis and evolutionary potential for extreme drought adaptation in Arabidopsis thaliana. *Nat Ecol Evol* **2**, 352–358 (2018).

155. Michael, T. P. *et al.* High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nat. Commun.* **9**, 541 (2018).

156. Monteiro, F. & Nishimura, M. T. Structural, Functional, and Genomic Diversity of Plant NLR Proteins: An Evolved Resource for Rational Engineering of Plant Immunity. *Annu. Rev. Phytopathol.* **56**, 243–267 (2018).

157. van Wersch, S. & Li, X. Stronger When Together: Clustering of Plant NLR Disease resistance Genes. *Trends Plant Sci.* **24**, 688–699 (2019).

158. Bergelson, J. & Roux, F. Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana. *Nat. Rev. Genet.* **11**, 867–879 (2010).

159. Parker, I. M. & Gilbert, G. S. The Evolutionary Ecology of Novel Plant-Pathogen Interactions. *Annu. Rev. Ecol. Evol. Syst.* **35**, 675–700 (2004).

160. Van de Weyer, A.-L. *et al.* A Species-Wide Inventory of NLR Genes and Alleles in

Arabidopsis thaliana. *Cell* **178**, 1260–1272.e14 (2019).

161. Sakai, A. K. *et al.* The population biology of invasive species. *Annu. Rev. Ecol. Syst.* **32**, 305–332 (2001).

162. King, K. C. & Lively, C. M. Does genetic diversity limit disease spread in natural host populations? *Heredity* **109**, 199–203 (2012).

163. McDowell, J. M., Hoff, T., Anderson, R. G. & Deegan, D. Propagation, Storage, and Assays with Hyaloperonospora arabidopsidis: A Model Oomycete Pathogen of Arabidopsis. in *Plant Immunity: Methods and Protocols* (ed. McDowell, J. M.) 137–151 (Humana Press, 2011).

164. Holt, B. F., 3rd, Belkhadir, Y. & Dangl, J. L. Antagonistic control of disease resistance protein stability in the plant immune system. *Science* **309**, 929–932 (2005).

165. Holub, E. B. Natural history of Arabidopsis thaliana and oomycete symbioses. in *European Journal of Plant Pathology* (2008). doi:10.1007/s10658-008-9286-1.

166. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. 201178 (2017) doi:10.1101/201178.

167. Ayres, D. L. *et al.* BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Syst. Biol.* **61**, 170–173 (2012).

168. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

169. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

170. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

171. Wickham, H. ggplot2 - Elegant Graphics for Data Analysis (2nd Edition). *Journal of Statistical Software, Book Reviews* **77**, 1–3 (2017).

172. Hamilton, N. & Others. ggtern: An Extension to ‘ggplot2’, for the Creation of Ternary

Diagrams. *R package version* **2**, (2016).

173. Long, J. A. jtools: Analysis and Presentation of Social Scientific Data. R package version 2.0.0. (2019).

174. Grimm, D. G. *et al.* easyGWAS: A Cloud-Based Platform for Comparing the Results of Genome-Wide Association Studies. *Plant Cell* **29**, 5–19 (2017).

175. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).

176. Hayes, B. J., Visscher, P. M. & Goddard, M. E. Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res.* **91**, 47–60 (2009).

177. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

178. Caruccio, N. Preparation of Next-Generation Sequencing Libraries Using Nextera™ Technology: Simultaneous DNA Fragmentation and Adaptor Tagging by In Vitro Transposition. in *High-Throughput Next Generation Sequencing: Methods and Applications* (eds. Kwon, Y. M. & Ricke, S. C.) 241–255 (Humana Press, 2011).

179. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

180. Broman, K. W., Wu, H., Sen, S. & Churchill, G. A. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889–890 (2003).

181. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

182. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–5 (2016).

183. Sievers, F. & Higgins, D. G. Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol. Biol.* **1079**, 105–116 (2014).

184. Jakob, K. *et al.* Pseudomonas viridiflava and P. syringae—Natural Pathogens of

Arabidopsis thaliana. *Mol. Plant. Microbe. Interact.* **15**, 1195–1203 (2002).

185. Borrelli, G. M. *et al.* Regulation and Evolution of NLR Genes: A Close Interconnection for Plant Immunity. *Int. J. Mol. Sci.* **19**, (2018).

186. Lee, R. R. Q. & Chae, E. Variation Patterns of NLR Clusters in Arabidopsis thaliana Genomes. *Plant Communications* **1**, 100089 (2020).

187. Shen, J., Araki, H., Chen, L., Chen, J.-Q. & Tian, D. Unique evolutionary mechanism in R-genes under the presence/absence polymorphism in Arabidopsis thaliana. *Genetics* **172**, 1243–1250 (2006).

188. Tian, D., Traw, M. B., Chen, J. Q., Kreitman, M. & Bergelson, J. Fitness costs of R-gene-mediated resistance in Arabidopsis thaliana. *Nature* **423**, 74–77 (2003).

189. Grant, M. R. *et al.* Independent deletions of a pathogen-resistance gene in Brassica and Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 15843–15848 (1998).

190. Tian, D., Araki, H., Stahl, E., Bergelson, J. & Kreitman, M. Signature of balancing selection in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 11525–11530 (2002).

191. Alcázar, R. *et al.* Analysis of a plant complex resistance gene locus underlying immune-related hybrid incompatibility and its occurrence in nature. *PLoS Genet.* **10**, e1004848 (2014).

192. Yin, W. *et al.* The Phytophthora sojae Avr1d gene encodes an RxLR-dEER effector with presence and absence polymorphisms among pathogen strains. *Mol. Plant. Microbe. Interact.* **26**, 958–968 (2013).

193. Ye, W. *et al.* Digital gene expression profiling of the Phytophthora sojae transcriptome. *Mol. Plant. Microbe. Interact.* **24**, 1530–1539 (2011).

194. Gilroy, E. M. *et al.* Presence/absence, differential expression and sequence polymorphisms between PiAVR2 and PiAVR2-like in Phytophthora infestans determine virulence on R2 plants. *New Phytol.* **191**, 763–776 (2011).

195. Asai, S. *et al.* Expression profiling during arabidopsis/downy mildew interaction reveals a

highly-expressed effector that attenuates responses to salicylic acid. *PLoS Pathog.* **10**, e1004443 (2014).

196. Baltrus, D. A. *et al.* Dynamic evolution of pathogenicity revealed by sequencing and comparative genomics of 19 Pseudomonas syringae isolates. *PLoS Pathog.* **7**, e1002132 (2011).

197. Lindeberg, M., Cunnac, S. & Collmer, A. Pseudomonas syringae type III effector repertoires: last words in endless arguments. *Trends Microbiol.* **20**, 199–208 (2012).

198. Karasov, T. L. *et al.* Arabidopsis thaliana and Pseudomonas Pathogens Exhibit Stable Associations over Evolutionary Timescales. *Cell Host Microbe* **24**, 168–179.e4 (2018).

199. Baxter, L. *et al.* Signatures of Adaptation to Obligate Biotrophy in the Hyaloperonospora arabidopsidis Genome. *Science* **330**, 1549–1551 (2010).

200. Woods-Tör, A. *et al.* A Suppressor/Avirulence Gene Combination in Hyaloperonospora arabidopsidis Determines Race Specificity in Arabidopsis thaliana. *Front. Plant Sci.* **9**, 265 (2018).

201. Ishaque, N. An Investigation into the Signatures of Evolution in Pathogen Effector Genes. (University of East Anglia, 2012).

202. Anderson, R. G. *et al.* Homologous RXLR effectors from Hyaloperonospora arabidopsidis and Phytophthora sojae suppress immunity in distantly related plants. *Plant J.* **72**, 882–893 (2012).

203. Fabro, G. *et al.* Multiple candidate effectors from the oomycete pathogen Hyaloperonospora arabidopsidis suppress host plant immunity. *PLoS Pathog.* **7**, e1002348 (2011).

204. Cooper, A. J. *et al.* Basic compatibility of Albugo candida in Arabidopsis thaliana and Brassica juncea causes broad-spectrum suppression of innate immunity. *Mol. Plant. Microbe. Interact.* **21**, 745–756 (2008).

205. Agler, M. T. *et al.* Microbial Hub Taxa Link Host and Abiotic Factors to Plant Microbiome

Variation. *PLoS Biol.* **14**, e1002352 (2016).

206. Ruhe, J. *et al.* Obligate biotroph pathogens of the genus Albugo are better adapted to active host defense compared to niche competitors. *Front. Plant Sci.* **7**, 820 (2016).

207. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv [q-bio.GN]* (2012).

208. Goritschnig, S., Krasileva, K. V., Dahlbeck, D. & Staskawicz, B. J. Computational prediction and molecular characterization of an oomycete effector and the cognate Arabidopsis resistance gene. *PLoS Genet.* **8**, e1002502 (2012).

209. Allen, R. L. *et al.* Natural variation reveals key amino acids in a downy mildew effector that alters recognition specificity by an Arabidopsis resistance gene. *Mol. Plant Pathol.* **9**, 511–523 (2008).

210. Laflamme, B. *et al.* The pan-genome effector-triggered immunity landscape of a host-pathogen interaction. *Science* **367**, 763–768 (2020).

211. Günther, T., Lampei, C., Barilar, I. & Schmid, K. J. Genomic and phenotypic differentiation of Arabidopsis thaliana along altitudinal gradients in the North Italian Alps. *Mol. Ecol.* **25**, 3574–3592 (2016).

212. Barragan, C. A. *et al.* RPW8/HR repeats control NLR activation in Arabidopsis thaliana. *PLoS Genet.* **15**, e1008313 (2019).

213. Orgil, U., Araki, H., Tangchaiburana, S., Berkey, R. & Xiao, S. Intraspecific genetic variations, fitness cost and benefit of RPW8, a disease resistance locus in Arabidopsis thaliana. *Genetics* **176**, 2317–2333 (2007).

214. Xiao, S. *et al.* Origin and maintenance of a broad-spectrum disease resistance locus in Arabidopsis. *Mol. Biol. Evol.* **21**, 1661–1672 (2004).

215. Xiao, S. *et al.* Broad-spectrum mildew resistance in Arabidopsis thaliana mediated by RPW8. *Science* **291**, 118–120 (2001).

216. Rose, L. E. *et al.* The maintenance of extreme amino acid diversity at the disease

resistance gene, RPP13, in Arabidopsis thaliana. *Genetics* **166**, 1517–1527 (2004).

217. Jiao, W.-B. & Schneeberger, K. Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat. Commun.* **11**, 989 (2020).

218. Borhan, M. H. *et al.* WRR4 encodes a TIR-NB-LRR protein that confers broad-spectrum white rust resistance in Arabidopsis thaliana to four physiological races of Albugo candida. *Mol. Plant. Microbe. Interact.* **21**, 757–768 (2008).

219. Staal, J., Kaliff, M., Dewaele, E. & Persson, M. RLM3, a TIR domain encoding gene involved in broad-range immunity of Arabidopsis to necrotrophic fungal pathogens. *The Plant* (2008).

220. Grant, J. J., Chini, A., Basu, D. & Loake, G. J. Targeted activation tagging of the Arabidopsis NBS-LRR gene, ADR1, conveys resistance to virulent pathogens. *Mol. Plant. Microbe. Interact.* **16**, 669–680 (2003).

221. Debener, T., Lehnackers, H., Arnold, M. & Dangl, J. L. Identification and molecular mapping of a single Arabidopsis thaliana locus determining resistance to a phytopathogenic Pseudomonas syringae isolate. *Plant J.* **1**, 289–302 (1991).

222. Simonich, M. T. & Innes, R. W. A disease resistance gene in Arabidopsis with specificity for the avrPph3 gene of Pseudomonas syringae pv. phaseolicola. *Mol. Plant. Microbe. Interact.* **8**, 637–640 (1995).

223. Nishimura, M. T. *et al.* TIR-only protein RBA1 recognizes a pathogen effector to regulate cell death in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E2053–E2062 (2017).

224. Faigón-Soverna, A. *et al.* A constitutive shade-avoidance mutant implicates TIR-NBS-LRR proteins in Arabidopsis photomorphogenic development. *Plant Cell* **18**, 2919–2928 (2006).

225. Oome, S. *et al.* Nep1-like proteins from three kingdoms of life act as a microbe-associated molecular pattern in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* **111**,

16955–16960 (2014).

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

227. Palma, K. *et al.* Autoimmunity in Arabidopsis acd11 is mediated by epigenetic regulation of an immune receptor. *PLoS Pathog.* **6**, e1001137 (2010).

228. Buell, C. R. *et al.* The complete genome sequence of the Arabidopsis and tomato pathogen Pseudomonas syringae pv. tomato DC3000. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 10181–10186 (2003).

229. Porebski, S., Bailey, L. G. & Baum, B. R. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Mol. Biol. Rep.* **15**, 8–15 (1997).

230. McDonald, M. High quality DNA from Fungi for long read sequencing e.g. PacBio, Nanopore MinION. https://www.protocols.io/view/high-quality-dna-from-fungi-for-long-read-sequenci-k6qczdw (2017) doi:10.17504/protocols.io.k6qczdw.

231. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

232. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

233. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).

234. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

235. Hoff, K. J. & Stanke, M. Predicting Genes in Single Genomes with AUGUSTUS. *Curr. Protoc. Bioinformatics* e57 (2018).

236. Saunders, D. G. O. *et al.* Using hierarchical clustering of secreted protein families to classify and rank candidate effectors of rust fungi. *PLoS One* **7**, e29847 (2012).

237. Librado, P. & Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451–1452 (2009).

238. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010**, db.prot5448 (2010).

239. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).

240. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).

241. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

242. Sánchez-Vallet, A. *et al.* The Genome Biology of Effector Gene Evolution in Filamentous Plant Pathogens. *Annu. Rev. Phytopathol.* (2018) doi:10.1146/annurev-phyto-080516-035303.

243. Plissonneau, C. *et al.* Using Population and Comparative Genomics to Understand the Genetic Basis of Effector-Driven Fungal Pathogen Evolution. *Front. Plant Sci.* **8**, 119 (2017).

244. Leonelli, L. *et al.* Structural elucidation and functional characterization of the Hyaloperonospora arabidopsidis effector protein ATR13. *PLoS Pathog.* **7**, e1002428 (2011).

245. Scholthof, K.-B. G. The disease triangle: pathogens, the environment and society. *Nat. Rev. Microbiol.* **5**, 152–156 (2007).

246. Cohen, S. P. & Leach, J. E. High temperature-induced plant disease susceptibility: more than the sum of its parts. *Curr. Opin. Plant Biol.* **56**, 235–241 (2020).

247. Genomics of rapid Evolution in Novel Environments. https://grenenet.wordpress.com/.

248. The Pathodopsis Team. Pathodopsis - the co-evolution of A. thaliana and its pathogens. http://pathodopsis.org/.

249. Thrall, P. H. & Burdon, J. J. Host-Pathogen Dynamics in a Metapopulation Context: The Ecological and Evolutionary Consequences of Being Spatial. *J. Ecol.* **85**, 743–753 (1997).

Acknowledgements

One of the most enriching parts of my life comes to an end, and I would like to thank everyone that has made this journey possible in one way or another. First and foremost, I am incredibly grateful to Detlef Weigel for believing in me, for his continuous support and making me grow both scientifically and personally. Your life and scientific wisdom inspire me every day. I keep these two with me; "*There are very few bad people in this world*" and "*You can always learn something positive from anyone*."

I am thankful to the mentors that have accompanied my scientific journey; From the University of Málaga, I want to thank Dr. Josefa Ruiz Sánchez, Dr. Manuel Marí Beffa, and, in particular, Dr. Francisco Cánovas for teaching me all about plant genetics. I wish to express my gratitude to Dr. Leslie Sieburth from the University of Utah, whose generosity and support led me to Tübingen. I would also like to thank my Thesis Advisory Committee members; Dr. Karl Schmidt, Dr. Thorsten Nurnberger, and Dr. Felicity Jones, who reviewed my progress, providing valuable feedback. I am thankful to the Sainsbury Laboratory past and present members for being wonderful hosts; Dr. Jonathan Jones, Dr. Sophien Kamoun, Dr. Ksenia Krasileva, and Dr. Joe Win.

I have had the pleasure of overlapping with wonderful colleagues who supported me along my scientific journey and shared valuable moments. I am thankful to the Cánovas Lab and Sieburth Lab members, particularly to Jorge, Belén, Malia, Dave, and Reed, for being in the front line with me. I am grateful to the entire Weigel Lab for research support and intellectual discussions. In particular, I want to thank those who made me stand up stronger and be more resilient. I want to particularly mention Alejandra, Fernando, Talia, and Oliver for always supporting me and making me a better scientist.

I wish to express my gratitude to the Ph.D. community at the Max Planck Institute for Developmental Biology, who taught me how to think outside of the box, and to the past and present Ph.D. representatives with whom I shared the passion

for making our institute a better working place. I would also like to thank the genome center and the IT support team that made my research possible.

I am thankful to Dr. Dagmar Sigurdardottir, the International Ph.D. Program Coordinator, for her inspiring dedication to the students' community and helping me many times. To the colleagues from the other side of the hill, I want to thank Farid, Mayank, Kyrylo, and Louis for making the ZMBP a place where I always felt welcomed and for cheering the Ph.D. students community.

Because the Ph.D. is also a personal journey, I want to thank all my daring friends who supported me behind the scenes.

I am thankful to mis "supernenas," Marina and María. I learned from you more than I could ever have imagined. Thanks for being there through thick and thin. I am thankful to my "german" friends, Elena and Ilona, for showing me the beauty of the German language and making me feel at home. *Du wirst immer in meinem Herzen sein.* I especially want to thank Maite for being the most unconditional friend. I would also like to thank Elena for our failed attempts to get fit together and wonderful moments around Tübingen. Finally, I want to thank destiny for putting another Alba in my life; you have been such a remarkable life coincidence.

Quiero hacer una mención especial a mis niñas de Marbella; Tere, Davinia, Isa, Mj y Marina. Gracias por vuestra amistad, por todo lo que hemos vivido juntas y por crecer a mi lado. Gracias por hacer de Marbella mi hogar y por siempre recibirme con los brazos abiertos a pesar del tiempo y la distancia. Os quiero.

And last but not least, the people that I consider my family.

Manfred, Monika, and Andi, you welcomed me in your home with the arms and the heart open; I could not have asked for more. Moritz, you are the love of my life. I'm genuinely excited to be your life companion and share endless love and support. Thank you for everything.

Familia, quiero recordaros una vez más mi amor incondicional y daros las gracias por todo. A mis abuelos que ya no están pero me dieron tanto. A mi abuelo

Josele, por sus gracias, risas y por su ingenio. A mi abuela Josefa, por darme una educación y la mejor madre que se pueda pedir. A mi abuelo Nisio, por compartir aficiones juntos y hacerme la más feliz pescando cangrejos. A mi abuela Esperanza, la virtuosa, que siempre tenía amor del bueno para todos. A “Los González”, que a pesar de las dificultades se siguen manteniendo unidos. A mi tia Mariajo, por ser una segunda madre y descubrirme el mundo a través de los viajes. A mi otra mitad, Jimena. Tan pequeña y tan grande a la vez. Siempre serás mi inspiración y siempre estaré ahí, siempre. A mi padre, por saber escuchar y por enseñarme que de sueños también se vive. Gracias a ti ahora soy una idealista y romántica empedernida, que remedio. A mi madre, por aguantarme y quererme, por hacerme mejor persona y enseñarme a amar a la vida. Por hacerme sensible y luchadora, por mantener viva mi curiosidad, y sobre todo, por su amor incondicional.

i stand
on the sacrifices
of a million women before me
thinking
what can i do
to make this mountain taller
so that women after me
can see farther

legacy - rupi kaur