THE UNIVERSITY OF CHICAGO

# NATURAL VARIATION IN SEQUENCE, DISEASE RESISTANCE, AND FITNESS IN THE *R*-GENE FAMILY OF *ARABIDOPSIS THALIANA*

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF ECOLOGY AND EVOLUTION

BY

ALICE MACQUEEN

CHICAGO, ILLINOIS

MARCH 2016

*A great man is coming to eat at my house. I do not wish to please him; I wish that he should wish to please me. I will stand here for humanity, and though I would make it kind, I would make it true.*

-Ralph Waldo Emerson

## Table of Contents

## List of Figures

## List of Tables

**Abstract**

Plant pathology developed as a field after the nineteenth century potato famine that led to more than one million deaths across Europe. One important component of plant defense is the plant resistance (*R*) gene family, which includes an *R*-gene that recognizes the causal pathogen of the potato famine. *R*-genes are a vital component of the plant's secondary defense response, and function by recognizing specific pathogen-released Avirulence genes or cellular changes effected by these genes, then inducing a strong defense response. *R*-gene evolution is thought to be driven both by selection on pathogen recognition functions and fitness costs of resistance. As a result, *R*-genes offer some of the most spectacular examples of polymorphism in plant genomes, suggesting selection may drive multiple aspects of *R*-gene variation. As my dissertation research, I investigated three facets of *R*-gene polymorphism in *Arabidopsis thaliana*.

First, I tested for evidence of adaptive variation in *R*-gene expression (MacQueen and Bergelson, in press). *R*-genes exhibit unusually extensive variation in basal expression levels when compared across accessions collected from different environments. I characterized *R*-gene expression variation to explore whether *R*-gene expression was up-regulated in environments favoring pathogen proliferation and down-regulated when risks of infection are low; down-regulation would follow if costs of *R*-gene expression negatively impact plant fitness in the absence of disease. Surprisingly, almost every change in the environment led to an increase in *R*-gene expression, a response that was distinct from the average transcriptome response and from that of other stress response genes. These changes in expression were functional in that environmental change prior to infection affected levels of specific disease resistance to isolates

of *Pseudomonas syringae*. In addition, there were strong latitudinal clines in basal *R*-gene expression and clines in *R*-gene expression plasticity with drought and high temperatures. These results suggest that variation in *R*-gene expression across environments may be shaped by natural selection to reduce fitness costs of *R*-gene expression in permissive or predictable environments.

Second, I measured fitness costs of resistance for the *R*-gene *Rps2*, which is under balancing selection for expressed resistant and susceptible clades of alleles. To do this, I conducted a large field fitness trial using 22 isogenic lines differing only in the allele of the *R*-gene *Rps2* that they carried. I found that resistant alleles of *Rps2* do not carry high fitness costs relative to susceptible alleles in the absence of disease. Instead, all alleles confer fitness benefits relative to an artificially constructed null, due to the additional role of RPS2 as a negative regulator of defense. Variation in the presence or absence of costs suggests an interplay between costs and the genetic architecture of resistance.

Third, I investigated the effects of an ancient duplication of the *R*-gene *Rpp8* on polymorphism at that locus. I examined patterns of synonymous and nonsynonymous sequence polymorphism, haplotype diversity, and linkage disequilibrium within and between homologs of *Rpp8* in 28 accessions of *A. thaliana*, and modeled the effects of duplicate distance, intergenic gene conversion rate, and number of duplicates on the patterns of polymorphism expected at this locus. The data were consistent with selection on *Rpp8* copy number and genetic architecture to generate and share an excess of novel genotypes.

In combination, the three projects above helped elucidate the forces that select on *R*-gene genomic architecture and levels of expression.

**Acknowledgements**

There are truly remarkable people at every level at the University of Chicago. First, I'm indebted to the members of staff that have helped me pass through the necessary flaming hoops during my time here. Jeff Wisniewski, Julie Steffen, Alison Anastasio, Audrey Aronowsky, Connie Homan, Bonnie Brown – thank you for knowing where the hoops are, and how best to deal with them. Second, I'd like to thank the faculty, and in particular, members of my committee: Marty Kreitman, Dick Hudson, Manyuan Long, Ilya Ruvinsky, and Jack Gilbert, but also Jerry Coyne, Molly Prezworski, and Joe Thornton. A large part of why I decided to go to the University of Chicago was the rigor and clarity of the scientific discussions I had on my interview day; that commitment never wavered in these faculty, and I'm indebted to them for their commitment to furthering my education in the sciences. Third, I am indebted to the following undergraduates, and in particular to Celia Eckhart, for their help carrying out inadvisably large fitness and gene expression experiments: Celia Eckhart, Bailey Zweifel, Alex Colbourne, and Meaghan Lyons.

I also have to thank my inner scientific circle, the Bergelson lab. Most of the things I have learned, and any success in Ecology and Evolution that I have garnered, is due directly to help, advice, smart questions and discussion, and coffee runs with members of the Bergelson lab. Though I believe it's true that "A couple of months in the laboratory can frequently save a couple of hours in the library,", it's also true that a couple of weeks of computer work can frequently save five minutes of conversation with a member of the Bergelson lab. Many thanks to Ben Brachi, Talia Karasov, Tim Morton, and Matt Perisin, as well as past members: Madlen Vetter, Wayan Muliyati, Laura Merwin, Luke Barrett, Chris Meyer, Angela Hancock, and Matt

Horton. Also, thanks to Dacheng Tian, and Xiaoqin Suen, both of whom I have yet to meet, but whose research I have picked up and to whom I am indebted for keeping clear, decipherable notes. And of course, I am deeply grateful for the time and effort Joy has spent mentoring me. From the very first, Joy has taught me essential writing skills and experimental design, given me clear and cogent advice to help handle scientific hurdles and close on projects, and tacitly assured me that I do have the abilities to go on and succeed as a scientist.

Equally important are the people who have kept me happy and sane while in the crucible of graduate school. I am thankful for the outlet of the Chicago Botanic Gardens – many thanks to Kristie Webber, Mary Plunkett, and Rob Hevey for involving me in the volunteer work there. I've learned more than you probably realize from all of you, about both how to manage and encourage volunteers and volunteer work, as well as how to interact with the public and get them excited about plants! Then, of course, graduate school would not have been tolerable at all without other graduate student friends. There are more people than I can easily list here, but many thanks to Beth Shimmyo, Laura Merwin, Carrie Albertin, Judit Pungor, and Kristen Voorhies for your close friendship, and thanks also to Wynn Meyer, Crystal Love, Dave Kennedy, Nate Upham, Will Tybursky, Emma Grieg, Paul Grabowski, Liz Scordato, Ben Krinsky, Daniel Matute, Kacy Gordon, Ben Winger, Colin Kyle, Chris Schell, Aaron Olsen, Ben Rubin, Justin Lemberg, Josh Symonds, Justin Heckman, Tim Sosa, Rob Arthur, Sarah Jackrel, and Kat Beilsmith. You guys are the best! In particular, my running buddies: the 'Darwin Posse' of Tom Stewart, Talia Karasov, and Mike Werner, and Mo Siddiq, have made graduate school so much better by bonding with me by sharing other painful, unpleasant running experiences. The people I've run with here have been some of my closest and truest friends in Chicago.

Finally, I am so grateful to my family for keeping me grounded and for supporting me throughout this process. Jessie, I was so lucky to have a sister with me in Chicago, even though I know it was your fifth choice!! We've had so much fun here, and I can't wait to see where your brilliant brain and commitment to making the world a better place will take you next. Laurie, when I started at Chicago, you were still a kid. Now, you have become a wonderful young woman – a 'small giant'. It's been wonderful watching your brain and heart grow while I've been here. You don't see it, perhaps, but I see it, and I'm so proud of you. M, I'm so grateful for your support and visits. Thank you for modelling a no-nonsense, self-reliant attitude for me. Every girl should be lucky enough to have a mother like you. D, thank you for showing me how to take on a life's work, something bigger than myself. I try to live up to my parents' examples every day. I love you all so much!

# Chapter 1

## 1.1 Introduction

Plant pathology developed as a field after the nineteenth century potato famine that led to more than one million deaths across Europe (Vanhaute *et al.,* 2006). One important component of plant defense is the plant resistance (*R*) gene family, which includes an *R*-gene that recognizes the causal pathogen of the potato famine (Ballvora *et al.,* 2002). *R*-genes are a vital component of the plant's secondary defense response, and function by recognizing specific pathogen-released Avirulence genes (Avr) or cellular changes effected by these genes, then inducing a strong defense response. *R*-gene evolution is thought to be driven both by selection on pathogen recognition functions and fitness costs of resistance (Tian *et al.,* 2003).

### *1.1.1 Study System: Arabidopsis thaliana*

*Arabidopsis thaliana* is a small annual plant species with a native range that encompasses much of the temperate zone of Europe and Central Asia, with invasive populations on Pacific islands and in North America. It is an opportunist species that favors frequently disturbed sites such as the edges of agricultural fields and roadsides, or marginal sites where competitive plant species are few. Natural populations of *A. thaliana* typically grow as winter annuals that germinate in fall, overwinter as rosettes, and bolt, flower, and set seed in early spring before the summer heat. Plants from natural populations primarily reproduce via self-fertilization: rates of outcrossing in the field are estimated to be between 1 and 3% (Platt *et al.,* 2010). This leads to naturally low levels of heterozygosity in natural populations, and inbred accessions derived from these natural populations are likely to be genetically similar to the populations from which they

were derived. A huge database of accessions of *A. thaliana* exist with which to ask questions about natural variation and local adaptation of *R*-genes.

Natural populations of *A. thaliana* follow a genetic pattern of isolation-by-distance (Sharbel *et al*., 2000; Platt *et al*., 2010); thus, a maximally diverse set of accessions of this species is a set of accessions derived from populations distributed widely around the world. Natural populations of *A. thaliana* are present in Sweden in a north-south cline that is thought to follow a natural cline in pathogen pressure, with more pressure in southern Sweden (Brachi, unpublished data). Unfortunately, it is difficult to say much about the distributions of the pathogens worldwide, except that pathogens tend to be present everywhere at sporadic intervals in time, and exhibit little population structure.

*A. thaliana* has a host of genetic resources to advance sequence-based work on natural variation in *R*-genes. A collection of thousands of accessions derived from natural populations with positional and environmental information is available (Hancock *et al*., 2011; Horton et al, 2012). Over 1001 genomes of these *A. thaliana* accessions have been sequenced, with genome sequences for over 500 accessions having already been released (Weigel and Mott 2009; Cao *et al*., 2011; Long *et al*., 2013; Schmitz *et al*., 2013). RNA-seq data is available for 144 accessions from Sweden grown in two growth conditions, as well as 19 additional diverse accessions (Gan *et al*., 2011; Dubin *et al*., 2015).

### *1.1.2 The Plant Immune System*

The plant immune system is divided into two levels of response to infection. The first line of defense is triggered by microbe-associated molecular patterns (MAMPs) that are highly conserved across microbial species, such as flagellin (Zipfel *et al.,* 2004). Recognition of these

MAMPs is known as MAMP-triggered immunity, MTI. However, many pathogens release a suite of effector molecules into plant cells that suppress MTI (Chisholm *et al.,* 2006). Plants, in turn, have a second, specific line of defense triggered by recognition of a subset of these effectors known as effector-triggered immunity (ETI). Effectors which trigger ETI are renamed Avirulence proteins. ETI typically results in a hypersensitive response (HR) by the plant that causes programmed cell death in the area of infection and chemical changes to limit pathogen spread through the plant (Greenberg 1997). HR can result in visible leaf collapse when large numbers of Avirulent bacteria infect a leaf. MTI is triggered by many microbial species, but ETI only occurs when plants encode and express a resistance gene (*R*-gene) that can recognize, indirectly or directly, one specific effector molecule.

*R*-genes are characterized by two amino acid domains: a nucleotide binding site (NBS) and a leucine-rich repeat region (LRR). A role in disease resistance has been documented in only 18 *R*-genes in *A. thaliana*; however, in the Columbia accession of *A. thaliana*, 149 candidate *R*-genes have these two amino acid domains, and two-thirds of R-genes exist in clusters with at least one *R*-gene of known function (Meyers *et al*., 2003). The NBS domain binds and hydrolyzes ATP; upon ATP hydrolysis, this domain is thought to act as a molecular switch to signal downstream defense pathways (McHale *et al.,* 2006). The LRR largely determines pathogen recognition specificity. There are 58 additional genes lacking LRR domains in *A. thaliana* that may function in ETI and *R*-gene mediated disease resistance (Meyers *et al.,* 2002); however, the LRR region has been implicated in Avirulence molecule recognition whenever a recognition mechanism has been determined (Chin *et al.,* 2001, Dodds *et al.,* 2001, Ellis *et al.,* 1999, Hwang and Williamson 2003, Wulff *et al.,* 2001). *R*-genes can be further subdivided into

classes by the presence of additional domains, including a coiled-coil (CC) domain and a TIR (Toll/Interleukin-1 Receptor) domain, both thought to be involved in protein-protein interactions (McHale *et al.,* 2006).

### *1.1.3 Natural variation in R-genes*

The 149 plant disease resistance (*R*-) genes in *Arabidopsis thaliana* provide some of the most extreme and atypical examples of polymorphism in the genome. *A. thaliana* accessions vary extensively in the basal expression of particular *R*-genes. In a survey of the expression of 45 gene families measured in 19 accessions of *A. thaliana*, the two *R*-gene subfamilies were in the top three families of differentially expressed genes (Gan *et al.,* 2011). Indeed, the extent of differential expression for *R*-genes is impressive, with up to 400-fold differences between accessions, the highest for any gene. *R*-gene coding sequences can be hypervariable, particularly for nonsynonymous changes, with $K_a$:$K_s$ ratios of up to 8 (Bergelson *et al*., 2001), and with synonymous nucleotide diversity four times higher than the genome average (Chapter 4). At least one third of *R*-loci show presence-absence polymorphisms (Guo *et al*., 2011), and two thirds of *R*-genes are clustered in the genome in groups of 2 to 12 genes (Bergelson *et al*., 2001). *R*-gene presence-absence polymorphisms and clustering of tandem duplicates have also been reported in other plant species (Zhou *et al*., 2004; Cheng *et al*., 2012; Kang *et al*., 2012; Perazzolli *et al*., 2014; Schmutz *et al*., 2014). *R*-gene presence varies substantially even between closely related species: in currently sequenced vascular and non-vascular plants, their predicted numbers range from 3-616, comprising up to 1.8% of the genes encoded in the genome (Kim *et al*., 2012).

Several types of selection may operate on different *R*-genes to drive these different patterns of polymorphism. 20% of *R*-genes are under positive selection in *A. thaliana,* as

indicated by a $K_a$:$K_s$ ratio of more than 2, while still other *R*-genes show evidence of diversifying selection (Bergelson *et al.,* 2001), and 30% of singleton *R*-loci show signatures of balancing selection (Bakker *et al*., 2006). Positive selection is the classic signature of an arms race scenario where pathogens impose selection to continually alter plant resistance specificity. Diversifying selection has been hypothesized to increase allelic diversity of *R*-gene coding regions to allow their continued recognition of fast-evolving pathogens. Copy number variation and intergenic exchange are also likely to be evolutionary responses to the need for new resistance types. Balancing selection on *R*-genes has been hypothesized to represent tradeoffs between the benefits and costs of resistance.

Four lines of evidence indicate that an additional factor in *R*-gene evolution is a cost of resistance, the cost of carrying a defense response gene in the absence of the cognate pathogen. A fitness cost for an annual plant such as *Arabidopsis thaliana* is any phenotypic change resulting in lower seed set or seed quality. There is direct evidence for two *R*-gene costs of resistance for *Rpm1* and *Rps5*, where field trials have indicated that the reduction in seed set for resistant plants in the absence of pathogens is as high as 5-10% (Tian *et al.,* 2003, Karasov *et al.,* 2014). These *R*-genes and others with long-maintained presence/absence polymorphisms or other signatures of balancing selection imply that both positive and negative selective forces have been acting on *R*-genes (Grant *et al.,* 1998, Tian *et al.,* 2002, Mauricio *et al.,* 2003, Caicedo *et al.,* 1998).

Second, overexpression or constitutive activation of single *R*-genes can also have fitness effects: overexpression of the *R*-gene ADR1-L1 using the constitutive 35S promoter resulted in both enhanced resistance to *P. syringae* infection as well as dwarfed rosette leaves and

significantly shorter primary inflorescence stems when plants were grown at 22°C (Kato *et al.,* 2011). Mutations in the *R*-gene *At1g61180* causing its constitutive activation resulted in extremely delayed and reduced flowering, with a one-fold reduction in flower formation in heterozygotes and no flower formation in individuals homozygous for the mutation (Igari *et al.,* 2008). A knockdown of the R-regulator RIN4 that led to constitutive activation of the *R*-genes RPS2 and RPM1 also gave morphological defects similar to the activation of *At1g61180* (Igari *et al.,* 2008). Reduction in silencing DNA methylation by a decrease-in-DNA-methylation mutant (DDM1) at the *R*-gene *At4g16890* led to dwarfed plants with a constitutively activated salicylic acid defense pathway (Stokes *et al.,* 2001). Even a two-fold increase in expression of a single *R*-gene can cause dwarfism and reduced seed set (Xiao *et al.,* 2002).

A more subtle line of evidence for a cost of resistance is that when whole genome duplications occur, new *R*-gene duplicates are eliminated from the genome at a significantly faster rate than other genes, leaving only paralogous *R*-genes (Nobuta *et al.,* 2005). This implies that a two-fold upregulation of many *R*-genes might be costly. At the most basic level, the dearth of *R*-genes in *A. thaliana* and many plant species relative to the number of potential pathogens implies that it is too costly to defend against every pathogen, despite the fact that defense against the correct few can be vital. In many cases, plants are under conflicting selection pressures when it comes to *R*-gene defense.

A number of major open questions remain regarding the potential adaptive nature of *R*-gene polymorphism. First, for surveillance genes which need to be expressed at low levels in order to perform their function in recognition, the high variation in *R*-gene expression between populations of *A. thaliana* that emerged from recent transcriptome work was surprising. This

expression variation could have resulted from spillover from diversifying selection on *R*-gene coding sequences, or it could itself be serving some adaptive function. Second, levels of costs of resistance found to date for *R*-genes under balancing selection for alternative functions are surprisingly high; additional mechanisms must be in place for some *R*-genes to reduce the severity of these costs. Third, the evolution and population genetics of *R*-gene duplicates and clusters remain poorly understood, as these sequences are typically inaccessible to next generation sequencing methodologies.

### 1.2 Chapter Summaries

In Chapter 2, I hypothesize that variation in *R*-gene expression is due to selection to mediate tradeoffs between the costs and benefits of disease resistance. Costs of resistance may stem from the cost of mounting a defense response or the potential for self-inflicted damage due to mis-activation of the defense response. However, as *R*-gene mediated defense is an induced response, with perfect regulation, the only cost should be the production of small amounts of a recognition protein. It is known that *R*-gene expression induction, or increases in *R*-gene expression, increase both general and specific disease resistance, but also can lead to high levels of self-inflicted cell death, associated with high fitness costs or lethality. It is thus difficult to imagine what perfect regulation of expression would look like in this fast evolving system with high tradeoffs associated with increased levels of expression. In Chapter 2, I explore *R*-gene expression as a function of environment as a proxy for regulation of *R* protein levels. I find that *R*-gene expression increases after many environmental perturbations, instead of tracking probabilities of infection in different environments. This may reflect the increased likelihood of

microbiome invasion when the microbial community is disturbed, as is predicted by theories of community ecology.

In Chapter 3, I hypothesize that the genetic architecture of two *R*-genes with documented fitness costs of resistance were exaggerated relative to the typical level of this cost. Fitness costs of 5-10% have been documented for two *R*-genes, *Rpm1* and *Rps5*. These large costs are surprising, and raise the question of how ~140 genes with such serious costs could segregate within populations. Both *Rpm1* and *Rps5* are segregating for long-lived presence/absence polymorphism, where the susceptible allele is completely missing from the genome. We expect large fitness costs to accrue to loci with presence-absence polymorphism, but there is no such expectation when the susceptible allele is also functional, and potentially segregating for some alternative or additional function. In Chapter 3, I determine the fitness of resistant and susceptible alleles of *Rps2*, an *R*-gene under balancing selection for resistance to *avrRpt2* that is present in every accession sequenced to date. I show that resistant *Rps2* alleles do not carry a large fitness cost of resistance relative to susceptible alleles, as predicted by our hypothesis. Instead, we find a large fitness benefit of any allele of *Rps2* relative to an artificial knockout, due perhaps to the role of all alleles of *Rps2* in negative regulation of the defense response. Field fitness decreased as *Rps2* expression increased, consistent with previously reported costs of *Rps2* overexpression. Our results suggest that defense loci segregating for functional alternatives have evolved to limit the manifestation of costs of resistance, in contrast to complex loci.

In Chapter 4, I hypothesize that diversifying selection for *R*-gene function might play a prominent role in the dynamics of the duplication and diversification process of *R*-genes. It has long been recognized that loci segregating for two functional alleles could eliminate the resulting

segregation load if the locus is duplicated and the alternative functional alleles are fixed at each of the sister loci. A key advantage of balancing selection as the driving force for duplication is that it provides an immediate selective advantage following duplication (Innan and Kondrashov 2009). However, there is an additional problem with this mechanism for a selfer like *A. thaliana*, and that is generating the optimal heterozygote in which the gene duplication event would be favored. I propose that when interlocus gene conversion acts on *R*-genes, it results in higher rates of exchange of novel and potentially adaptive SNPs in heterozygotes and between gene duplicates. In Chapter 4, I examine the evolution of the complex locus *Rpp8*, which segregates within *A. thaliana* for at least three distinct recognition specificities. I show that all sequenced *Rpp8* copies are undergoing gene conversion with a locus 2 Mb away of unknown function, *At5g48620*. I examine the population genetic consequences of these gene conversions and find that the triplication of *Rpp8* increases synonymous nucleotide diversity, and gene conversion between *Rpp8* and *At5g48620* greatly increases the haplotypic diversity at this locus. This data is consistent with diversifying selection on *Rpp8* copy number and genetic architecture to generate and share an excess of novel genotypes.

Thus, I study selection mitigating fitness costs of resistance and diversifying selection as the two driving selective forces acting on *R*-gene evolution. These forces may explain the action of evolution at both simple and complex *R*-loci and on the regulation of expression of *R*-genes. Through these three lines of inquiry I describe natural variation in expression, genetic architecture, and the nucleotide sequence of *R*-genes. In two of these cases, I attribute variation to fitness costs associated with *R*-genes. Understanding *R*-gene variation and its effects on phenotypes and fitness of *Arabidopsis thaliana* may shed further light on the complex and

unusual evolution of disease resistance genes in plants. An understanding of the molecular genetics of plant disease resistance may guide us in better engineering resistance in crop plants: an important part of this understanding revolves around the unique features of *R*-genes that distinguish them from more typical genomic loci.

# Chapter 2

## Modulation of *R*-gene Expression across Environments

### 2.1 Abstract

Some environments are more conducive to pathogen growth than others and, as a consequence, plants might be expected to invest more in resistance when pathogen growth is favored. Resistance (*R*-) genes in *Arabidopsis thaliana* have unusually extensive variation in basal expression when comparing the same *R*-gene among accessions collected from different environments. *R*-gene expression variation was characterized to explore whether *R*-gene expression is up-regulated in environments favoring pathogen proliferation and down-regulated when risks of infection are low; down-regulation would follow if costs of *R*-gene expression negatively impact plant fitness in the absence of disease. qRT-PCR was used to quantify the expression of 13 *R*-gene loci in plants grown in eight environmental conditions for each of 12 *A. thaliana* accessions, and large effects of the environment on *R*-gene expression were found. Surprisingly, almost every change in the environment – be it a change in biotic or abiotic conditions – led to an increase in *R*-gene expression, a response that was distinct from the average transcriptome response and from that of other stress response genes. These changes in expression were functional in that environmental change prior to infection affected levels of specific disease resistance to isolates of *Pseudomonas syringae*. In addition, there were strong latitudinal clines in basal *R*-gene expression and clines in *R*-gene expression plasticity with drought and high temperatures. These results suggest that variation in *R*-gene expression across environments may be shaped by natural selection to reduce fitness costs of *R*-gene expression in permissive or predictable environments.

## 2.2 Introduction

Plants in natural environments are frequently exposed to multiple stresses simultaneously. Plants exposed to combinations of stresses respond in ways that are distinct from, and in some cases entirely unpredictable based on gene expression in response to each stress alone (Suzuki *et al*., 2014). Expression of defense-related genes may vary with the abiotic environment due to associated changes in the probability and severity of disease. For example, high humidity and low temperatures promote more severe fungal outbreaks (Holub *et al.* 1994), while outbreaks of viral diseases are associated with cold snaps (Szittya *et al.,* 2003). Plants grown in environments prone to infection may thus invest more in the expression of defense response genes.

Resistance genes (*R*-genes) are a vital component of the plant immune system that function by producing proteins that recognize, directly or indirectly, specific effectors secreted by a pathogen. This recognition event, termed the gene-for-gene interaction, is highly specific and leads to the induction of a robust defense response. Despite the intuition that surveillance genes should be consistently expressed at low levels to allow pathogen detection, *Arabidopsis thaliana* genotypes vary extensively in the basal expression of most *R*-genes. In a survey of expression of 45 gene families measured in 19 accessions of *A. thaliana*, the two *R*-gene subfamilies were in the top three families of differentially expressed genes (Gan *et al.,* 2011). Indeed, the extent of differential expression for *R*-genes can be impressive, with up to 350-fold differences between accessions, the highest for any gene.

Natural variation in gene expression has been previously associated with important phenotypic variation (Baxter *et al*., 2010; Li *et al*., 2011; Studer *et al*., 2011; Chan *et al*., 2011).

However, it is surprising that *R*-gene expression is so variable between populations, given the large, pathogen-dependent tradeoffs in fitness associated with differences in *R*-gene expression. Plants with specific *R*-genes are up to 10% less fit than plants lacking that *R*-gene when grown in the absence of targeted pathogens (Tian *et al.*, 2003; Karasov *et al.* 2014). Fitness costs are further evident in the dwarfing that results when plant mutants constitutively overexpress *R*-genes; of 12 *R*-genes that have been overexpressed or constitutively activated in *A. thaliana*, two are lethal and nine carry significant growth defects (Mindrinos *et al*. 1994; Shah *et al.,* 1999; Shirano *et al.,* 2002; Stokes *et al*., 2002; Grant *et al.,* 2003; Xiao *et al.,* 2002; Igari *et al.,* 2008; Aboul-Soud *et al.,* 2009; Palma *et al.,* 2010; Kato *et al.,* 2011; Boccara *et al.,* 2014; Kato *et al.,* 2014). An additional study increased expression without mis-expression by introducing multiple copies of two *R*-genes under control of their native promoters. They found that increased expression decreased plant size (Xiao *et al*. 2002), presumably as a consequence of increased metabolic burden. Even a two-fold increase in expression of a single *R*-gene caused dwarfism and reduced seed set (Xiao *et al.,* 2002); basal *R*-gene expression varies even more than this between populations of *A. thaliana*.

The benefits of recognizing secreted effectors counteract these costs and potentially select for this expression variation. Fitness benefits of up to 30% have been demonstrated for plants capable of recognizing a particular *Pseudomonas syringae* isolate relative to plants that cannot (Gao *et al.,* 2009). A link between *R*-gene expression variation and fitness benefits of resistance depends on whether increases in basal *R*-gene expression provide more effective defense. This question has been addressed in two ways. First, seven studies have used mutants to ask whether basal *R*-gene overexpression impacts defense against pathogens in *A. thaliana*. Six

of these studies found that increased expression led to a significant decrease in pathogen load (Shah *et al.,* 1999; Stokes *et al*., 2002; Grant *et al.,* 2003; Aboul-Soud *et al.,* 2009; Kato *et al.,* 2011; Boccara *et al.,* 2014; but see Shirano *et al.,* 2002). In a more natural context, three studies have tested the effect of *R*-gene expression in *A. thaliana* on defense against specific pathogens; two of these studies found that increased expression led to a significant decrease in pathogen load (Van Poecke *et al.,* 2007; Zhang and Gassmann 2007; Boccara *et al.,* 2014). Thus, most studies support a clear link between *R*-gene expression and resistance to pathogens.

The net effect of *R*-gene expression on plant fitness clearly depends on the distribution of pathogens through time and space. Additionally, if the costs and benefits of *R*-gene expression vary as a function of the abiotic environment, selection may favor different levels of *R*-gene expression in different environments. Indeed, a genome wide association study (GWAS) supports an interaction between *R*-genes, climate, and pathogen load (Hancock *et al.,* 2011). In this study, SNPs associated with a suite of five climate variables overlapped with four QTLs associated with proliferation of the bacterial pathogen *P. syringae* DC3000. These climate variables (minimum temperature in the coldest month, number of consecutive cold days, relative humidity in spring, temperature seasonality, and precipitation in the driest month) may select for higher *R*-gene expression levels as they likely increase the probability that *P. syringae* infects its hosts (Underwood *et al.,* 2007; Ferrante *et al.,* 2012). Results from the Hancock *et al.* (2011) study also identified a SNP within the coding region of an *R*-gene (At5g22690) that was significantly correlated with both precipitation in the driest month and relative humidity in the spring, implying that the distribution of *R*-gene alleles may be shaped by climate.

The pathogens recognized by most *R*-genes are unknown, making environments favoring their proliferation necessarily unknown. To explore if and how plants modulate their *R*-gene expression in response to risk of infection, effects of short-term environment and historical climate on *R*-gene expression were determined. These environments were selected to represent environmental conditions likely to impact pathogen proliferation. If optimized, *R*-gene expression levels should track relative risks, increasing when the risk of infection is high, in cold and wet environments, and decreasing when it is low, in hot and dry environments. Alternatively, environmental perturbations themselves may provide sufficient reason to enhance levels of defense. Since plants contain diverse microbial communities (Bodenhausen *et al.,* 2013, Turner *et al.,* 2013, Horton *et al*., 2014), and since plant pathogens carrying *R*-gene-recognized Avirulence genes include multiple taxonomic domains, the outcome of pathogen infection may be best understood as a community assembly problem, where the pathogens are invasive species (Costello *et al.,* 2012). Pathogens are then likely to cause disease in conditions that help species invade and proliferate: that is, disturbed conditions to which the community is not well adapted (Shea & Chesson, 2002).

Here, diverse accessions of *A. thaliana* were grown in a set of eight environmental conditions to explore the relationships between abiotic perturbations, historical climate, and *R*-gene expression patterns under controlled experimental conditions. The expression of 13 *R*-genes in each of 12 *A. thaliana* accessions from disparate historical climates was determined. First, the average effect of current environment on *R*-gene expression and the effect of environment on the resistance response in one of these environments were measured. *R*-genes were upregulated on average in response to all abiotic perturbations, with extensive variation in *R*-gene expression

between accessions. A metastudy of the effect of 15 publicly available biotic and abiotic treatments on *R*-gene expression in *A. thaliana* demonstrated that the *R*-gene response to a variety of biotic perturbations differed from the responses both non-*R*-genes and other stress response genes. Functional resistance increased after an environmental perturbation followed by infection with *P. syringae*. Having found substantial variation in *R*-gene expression across environmental treatments, the geographical distribution of *R*-gene expression was examined. In particular, the expression of the 12 *A. thaliana* accessions was examined for the presence of environment-of-origin clines in both *R*-gene expression and plasticity in *R*-gene expression. RNA-seq data from 14 worldwide *A. thaliana* accessions and 144 Swedish *A. thaliana* accessions was examined to confirm these clines. In combination, these two sets of analyses allowed the investigation of the effect of environment on expression and the likelihood that variation in *R*-gene expression is shaped by the distribution of pathogens or other aspects of adaptation to the local environment.

## 2.3 Methods

### *2.3.1 Expression Study: Experimental Strategy and Overview*

The experimental design included 12 *Arabidopsis thaliana* accessions (Table 2.1) selected from a broad geographical distribution across Europe and a latitudinal cline in the Midwest USA. For each of these accessions, qPCR was used to measure expression of 13 randomly selected *R*-genes in control conditions, and after seven abiotic perturbations. Relative expression levels for these seven perturbations were considered relative to one unmanipulated control. Each perturbation was motivated by the temperature, precipitation and humidity climate variables associated with pathogen growth and recognition identified in Hancock *et al.* (2011).

Since transcriptional responses to perturbations of different lengths can differ, two treatment durations were applied. In particular, perturbations included three hours of heat shock and cold shock, as well as weeklong changes to short day, high temperatures, low water, high temperature and low water, and high humidity. Due to space constraints, a single perturbation, the weeklong change in day length, was manipulated for the Midwestern cline (Table 2.1) as this was the variable most likely to correlate with latitude. *R*-genes were chosen to represent the entire set and included singletons, genes from clusters, and genes in each of the major *R*-gene subfamilies (Table 2.2). Comparisons among the 12 accessions allowed the exploration of how historical climate impacted patterns of basal *R*-gene expression, how *R*-gene expression was modulated by weather (abiotic perturbations), and how historical climate shaped these *R*-gene responses.

**Table 2.1**. Accessions used for the expression study with details on their location of origin, treatments used, and the reasoning behind the accession's inclusion.

| Name | Ecotype ID | Location of Origin | Treatments | Reasoning |
|---|---|---|---|---|
| Col-0 | 8279 | Midwest USA | 1-8 | Clark *et al*. (2007) diverse set |
| Cvi-0 | 8281 | Canary Islands | 1-8 | Clark *et al*. (2007) diverse set |
| Est-1 | 8291 | Estonia | 1-8 | Clark *et al*. (2007) diverse set |
| Fei-0 | 8294 | Portugal | 1-8 | Clark *et al*. (2007) diverse set |
| Kin-0 | 8316 | MI, USA | 1-2 | Latitudinal cline |
| Kno-11 | 6810 | IN, USA | 1-2 | Latitudinal cline |
| Ler-1 | 8324 | Germany | 1-8 | Clark *et al*. (2007) diverse set |
| Rmx-A02 | 8370 | MI, USA | 1-2 | Latitudinal cline |
| RRS-7 | 8373 | MI, USA | 1-8 | Clark *et al*. (2007) diverse set and in latitudinal cline |
| SLSP-30 | 2274 | MI, USA | 1-2 | Latitudinal cline |
| Tsu-1 | 8394 | Italy | 1-2 | Clark *et al*. (2007) diverse set |
| Van-0 | 8400 | Western Canada | 1-8 | Clark *et al*. (2007) diverse set |

**Table 2.2.** *R*-genes and primers used in qRT-PCR study, including the GeneBank transcript number predicted to be recognized by the primers and the gene region (start-end) recognized by the primers. 13 *R*-genes were chosen to represent the larger set and included both singletons (seven) and genes from tandem arrays (six); genes from the TIR-NB-LRR subfamily (six), CC-NB-LRR subfamily (two), and NB-LRR subfamily (five); genes with overexpression mutants with known fitness consequences (two) and genes with known functions (three).

| Gene; GeneBank; Gene Name | Subfamily; Cluster Type | Phenotype | Gene Region Amplified | Primers Used |
|---|---|---|---|---|
| AT1G27180; NM102480.3; | TIR-NB-LRR; Tandem array | | 4743-4861 | AGCGTGTGTGCTTCTGTAAGTTTTT TTGTGGGAAACAAGAGTCTCCTCT |
| AT1G50180; NM103903.1; | CC-NB-LRR; Singleton | | 1262-1405 | TGCCTCCCCATGTAAAGCAATGT CCACTGTCGTGCCAGCCTCT |
| AT1G56510; NM104527.3; ADR2 | TIR-NB-LRR; Tandem array | Fitness cost of expression; resists *Albugo candida.* | 1381-1471 | CACACAGAGGCTGGCACGACA CCCTCCTCCCAACCATAACCATGC |
| AT2G14080; NM126980.2; | TIR-NB-LRR; Singleton | | 2695-3079 | TCTTGAGAGGATCACAGTGCTGGA CCAGTTTTCCGCATCCGCTGAGT |
| AT3G07040; NM111584.2; RPM1 | NB-LRR; Singleton | Fitness cost of presence; resists avrB and avrRpm1 | 2253-2372 | TGACCTGATCGCAACTGCAAGCA AGCTGAGATCCACGCAAACCCA |
| AT4G12020; NM001125496.1; MEKK4 | NB-LRR; Mixed array | | 4733-4856 | TGGAAACGGAAGAGACGGGAGC TCGGAGGCATAAATCGGCGACG |
| AT4G26090; NM118742.2; RPS2 | NB-LRR; Singleton | Resists avrRpt2 | 1854-1953 | CCCGAAACTGACAACACTGATGCT CAAGTCCAAGACTCTGAGAACAGGC |

**Table 2.2.** (continued)

| Gene; GeneBank; Gene Name | Subfamily; Cluster Type | Phenotype | Gene Region Amplified | Primers Used |
|---|---|---|---|---|
| AT5G17880 NM121794.2 CSA1 | TIR-NB-LRR Tandem array | | 2469-2541 | CTTGGGGAAACATGAGCCGC GCATAAACGACGCACGGAGA |
| AT5G04720; NM120554.1; ADR1-L2 | NB-LRR; Singleton | | 2389-2496 | TGGGAAAAGGTCCAGAAGGCGG GCTCGGAGGGAGAGAATCACGAA |
| AT5G05400; NM_120622.1; | CC-NB-LRR; Singleton | | 1660-1778 | GTCGCATGTGCCGATCCTTA AACCCATCTGGAAGGCTCGTT |
| AT5G22690 NM122175.2 | TIR-NB-LRR Singleton | | 2887-2988 | CAGGAGTCGGTGTACGCTTCC GTTCAGTCTCAGGGCATCCACA |
| AT5G40060 NM123369.3 | NB-LRR Mixed array | | 2711-2800 | CAAGGCTGACTTTTCAGACTGTGGG ATATTATCTGCTGTCGCCGCCGC |
| AT5G44870 NM123855.1 LAZ5 | TIR-NB-LRR Tandem array | | 2428-2573 | GCCGGGGATCTGACGAGACTT ACGTCCGTTGGAACGCTCTC |

*2.3.1.1 Plant materials and growth conditions*

Seeds from each accession were sown in 50:50 Farfad C2:Metromix 200 in 36-well trays and stratified at 4°C for seven days in a cold-room. Seedlings were then germinated in controlled-environment chambers at 20°C with 70% humidity, on a 16 hour light/8 hour dark cycle and a daily watering schedule. Plants were thinned and randomized into treatment flats by day seven of growth. On day seven, flats were transferred into one of eight treatment conditions. (1) Control plants remained in long day conditions in a 16 hour light/8 hour dark chamber. (2) Short day plants experienced a 12 hour light/12 hour dark cycle. (3) High temperature plants

were exposed to 28°C with 30% humidity on a 16 hour light/8 hour dark cycle. (4) High water plants were watered heavily each day in trays lacking drainage and with lids on to retain humidity, while (5) low water plants were watered every three days with drainage and without lids. (6) High temperature, low water plants were transferred into the 28°C chamber as in (3) and watered every three days. (7) Cold and (8) heat shock treated plants were treated as (1) control plants until three hours before harvest, when cold shock plants were placed on ice in a 4°C chamber and heat shock plants were placed in a 37°C incubator. To standardize the growth stage measured for different accessions, individual plants were harvested at the eight leaf stage: the above-ground tissue was flash frozen in liquid nitrogen, ground using a mortar and pestle, and stored at -80°C.

#### *2.3.1.2 RNA extraction and qRT-PCR*

To provide sufficient tissue for RNA extraction, two plants from each accession in each treatment were pooled to create one biological replicate. Three biological replicates per treatment by accession combination were processed in randomized blocks for the RNA extraction, reverse transcription, and qRT-PCR steps. RNA was extracted using a SIGMA Spectrum Plant Total RNA Kit protocol with the on-column DNAseI digestion procedure. RNA was tested for quality via nanodrop and gel electrophoresis. 1 μg RNA per 20 μL reaction was reverse transcribed per biological replicate using random hexamers (1mM, IDT), dNTPs (2mM, Takara), and 200 U buffered M-MuLV Reverse Transcriptase. Primers for qRT-PCR (Table 2.2) were designed to have an $R^2$ of 0.99 or higher and an efficiency of 1.9 - 2.1 on a three-fold dilution series. Primers were specific to a particular *R*-gene within *A. thaliana* based on Primer-BLAST when *A. thaliana*, human and fungal genome sequences were included. Additionally, primers did not non-

specifically amplify DNA from the 10 most common bacterial endophytes found in *A. thaliana* leaves. Three reference primers, for PP2A, Helicase and bHLH, were used to normalize RNA levels between samples because these genes are stably expressed in a suite of abiotic conditions (Czechowski *et al.,* 2005).

#### *2.3.1.3 Real-time qRT-PCR conditions*

20µL qRT-PCR reactions were performed in 384-well plates with an ABI 7900HT sequence detection system (Applied Biosystems) using SYBR Green for dsDNA synthesis detection, and ROX red to control for differences in fluorescence between samples. Three technical replicates were run per biological sample with the following thermal profile: 95°C for 60 s, 40 cycles of 95°C for 30 s, 51°C for 30s, and 68°C for 40s, followed by 95°C for 15s. Dissociation curves of the amplicons were determined by heating from 60°C to 95°C with a 2% ramp rate.

#### *2.3.1.4 qRT-PCR data cleanup and analysis*

Data cleanup and analysis proceeded as described by (Rieu & Powers, 2009; Vandesompele *et al.,* 2002). In addition, to reduce non-specific amplification, wells with amplicon melting temperatures below those corresponding to the product of interest were removed from the analysis. The dissociation curves for each reaction were plotted and those with irregular features were removed. After data cleanup, efficiencies (E) and Ct values were calculated from the clipped qRT-PCR output using the qpcR software in R and subtracting baseline data from cycles one to eight of the PCR reaction (Spiess *et al.,* 2008). Relative quantities (RQ) of the amplicon in the starting sample were calculated according to the formula: $RQ = 1/(E^{Ct})$. RQs were normalized by dividing the RQs by the geometric mean of the RQs for

the three reference genes. Log base two of the normalized RQ values were used as the dependent variables in a linear model (LM) in R using the nlme and lme4 packages. To infer the behavior of all *R*-genes, 13 genes were selected to represent the variation in all *R*-genes, and all models treated *R*-gene and all interaction terms including *R*-gene as random effects. LM fits and residuals were robust to the exclusion of the two genes with known common insertion-deletion polymorphisms (data not shown).

To determine which variables had large effects on *R*-gene expression (accession, treatment, or accession:treatment) a LM was constructed relating *R*-gene expression to treatment, accession, and their interaction, with *R*-gene and interaction terms that included *R*-gene treated as random effects. Variance components were estimated by REML, and 95% confidence intervals for the proportions of variance explained by the random effects were estimated by performing 1000 parametric bootstraps. To extend the model to all *R*-genes, models treated *R*-gene and interaction terms including *R*-gene as random effects. To correlate *R*-gene expression with short term environment, a LM was constructed relating *R*-gene expression to treatment, with *R*-gene, accession, and interaction terms that included *R*-gene treated as random effects, and omitting an interaction term between accession and treatment. All 12 accessions were considered in analyses involving treatment alone.

### *2.3.2 Expression Metastudy*

One limitation of the qPCR study was that it could not test whether the *R*-gene response to each perturbation was unusual compared to the response of the remainder of the transcriptome. Thus, to place the qRT-PCR results into a genomic context, a metastudy of publicly available data from the EMBL-EBI Gene Expression Atlas was conducted. Expression

data was collected for all available *R*-genes, and for two additional gene sets of equivalent number, generating a set of 433 genes (Appendix A.1). First, a control gene set was chosen to mimic the positions and distances among *R*-genes to control for co-regulation of clustered genes. Second, a stress response gene set was chosen randomly from genes with a GO annotation of response to stress (GO:0006950). Most experiments included in the metastudy considered only a single accession (typically Col-0).

Expression data was collected for the 15 perturbations described below (Table 2.3). The *R*-gene expression response was compared to the control gene expression response to test if the *R*-gene expression response is unusual; if so, then *R*-gene expression may be shaped by costs and benefits that the rest of the transcriptome does not experience. The *R*-gene expression response was also compared to the stress gene response to determine if the *R*-gene expression response was unusual compared to stress response genes. Eight conditions involved perturbations with pathogen or pathogen-associated molecular patterns and were expected to upregulate *R*-genes. Ozone exposure was expected to upregulate *R*-genes, as ozone wounds plant tissue and might allow easier infection (Sandermann, 2000). Exposure to salicylic acid (SA) was expected to upregulate *R*-genes, as SA induces the secondary immune response and can lead to systemic acquired resistance to pathogens. Four additional perturbation treatments were considered: two changes in temperature (from 20°C to 4°C or 37°C), drought stress, and 30-120 minutes of intense light after low light. *R*-genes should be upregulated in changing environmental conditions if changing environments increased the chance of microbiome invasion by pathogens. On the other hand, if *R*-gene expression tracked the risk of infection then *R*-genes should be

**Table 2.3.** Treatments considered in the metastudy and the studies from which they were derived. For the perturbation treatments, gene expression under the perturbation was compared to that of independent, unperturbed plants.

| Condition | Array Express Repository ID | # *R*-Genes | # Control Genes | # Stress Response Genes | Sampled Tissue | Experimental Setup (when not stated, plants are on soil) | Reference |
|---|---|---|---|---|---|---|---|
| ***B. graminis*** | E-GEOD-12856 | 16 | 26 | 35 | Rosettes | At growth stage 3.90, harvested 12 hpi | Jensen *et al*., 2008 |
| ***E. cichoracearum*** | E-GEOD-431 | 48 | 24 | 61 | leaves | Infected at 21 d, harvested 3 dpi | Nishimura *et al*., 2003 |
| ***Hyaloperonospora parasitica*** | E-GEOD-18329 | 60 | 54 | 95 | rosettes | Infected at 7 d, harvested 4 dpi | Bhattari *et al*., 2010 |
| ***Pseudomonas syringae* pv. DC3000 with *avrRpm1*** | E-GEOD-6176 | 14 | 10 | 32 | leaf | At growth stage 3.90, harvested 4 hpi | None associated |
| ***Pseudomonas syringae* with *AvrRpt2*** | E-GEOD-58954 | 61 | 57 | 89 | Mature leaf | Harvested 6 h post infection | None associated |
| ***Pseudomonas syringae* with *AvrRpt2*** | E-GEOD-6556 | 41 | 33 | 90 | leaf | Infected at 35 d, harvested 2 dpi | Zhang *et al*., 2007 |
| ***Pseudomonas syringae* pv. tomato with *HopZ1a*** | E-GEOD-21920 | 52 | 33 | 87 | Rosette leaf | No information | None associated |
| ***Pseudomonas syringae* pv. tomato** | E-GEOD-21920 | 26 | 12 | 53 | Rosette leaf | No information | None associated |
| **SA** | E-TABM-51 | 44 | 22 | 70 | Rosette leaves | At 42-49 d, 4, 28, or 52 h SA | Van Leeuwen *et al*., 2007 |
| **SA** | E-GEOD-22942 | 25 | 17 | 87 | Rosette leaves | At 4d, 40 min SA; harvested 10 h post treatment | None associated |

**Table 2.3** (continued)

| Condition | Array Express Repository ID | # *R*-Genes | # Control Genes | # Stress Response Genes | Sampled Tissue | Experimental Setup (when not stated, plants are on soil) | Reference |
|---|---|---|---|---|---|---|---|
| **flg22** | E-NASC-76 | 8 | 18 | 24 | seedlings on agar | At 10 d on agar; flg22 treatment for 1 and 3 h | Denoux *et al.,* 2008 |
| **4°C** | E-MEXP-1345 | 42 | 36 | 78 | leaf discs | At 45 d, 22-26 h 4°C exposure | Bieniawska *et al.,* 2008 |
| **4°C** | E-TABM-52 | 40 | 30 | 79 | bolting rosette | At bolting, to 4°C for 14 d | Hannah *et al.,* 2006 |
| **37°C** | E-GEOD-11758 | 52 | 39 | 89 | Rosette leaves | At 35 d, 1 h heat shock | Sugio *et al*., 2009 |
| **Drought** | E-GEOD-40061 | 52 | 38 | 106 | Rosette leaves | At 21 d, 10 d drought stress | Pandey *et al*., 2013 |
| **Light** | E-MTAB-392 | 72 | 53 | 140 | leaf | At 24 d, 30 – 120 min excess light (1300 µmol photons/$m^2$s) | None associated |
| **Ozone** | E-MEXP-342 | 17 | 11 | 35 | Whole plant | At 10d, 500ppb ozone | Van Aken *et al*., 2009 |

upregulated in cold and wet environments and downregulated in hot and dry environments (Holub *et al.* 1994; Szittya *et al.,* 2003).

For each gene, expression was scored as upregulated, downregulated or non-differentially expressed relative to the controls for each experimental condition. Benjamini-Hochberg-corrected chi-squared tests of numbers of *R*-genes upregulated, downregulated, or non-differentially expressed were conducted to assess if *R*-genes responded significantly differently to environmental perturbations than and control or stress response genes.

### *2.3.3 Resistance Determination*

A second limitation of the qPCR study was that it did not link an expression response upon treatment with a phenotypic response, such as a change in resistance. To determine if short-term environmental changes could influence or "prime" disease resistance, Col-0 plants were grown as in the qRT-PCR study, transferred just prior to the 21-day stage into one of three treatments, then immediately post-treatment, at the 21-day stage, infected with one of two pathovars of *P. syringae,* at an OD600 of 0.002 by blunt-end syringe inoculation (Korves and Bergelson 2003). Mock infected plants were inoculated with the buffer 10 mM MgSO4. After inoculation, plants were kept at high humidity under plastic domes for 3 days to facilitate infection. Plant resistance was measured as the log of the number of colony forming units (CFU) in the infected leaf three days post infection (dpi). CFU was measured by plating dilutions of ground leaf punches on Luria-Bertani plates with blinded colony counting. 20-26 replicates per combination of pathovar and treatment were used.

Control plants and cold shocked plants were treated as in the qRT-PCR study. The positive control, BTH treated plants, were treated as control plants until 24 hours before harvest, then sprayed with 100 μM BTH, a SA analog that both induces *R*-gene expression and increases pathogen resistance (Wang *et al*., 2006). Two pairs of DC3000 strains were used: one pair with and without the RPS2-recognized Avirulence gene *avrRpt2* (Guttman and Greenberg 2001), and one pair with and without the RPS5-recognized Avirulence gene *avrPphB* (Karasov *et al*., 2014). The DC3000 strains contained an empty version of the different vectors used to introduce the two Avirulence genes.

### *2.3.4 Gene Expression Clines: Experimental Strategy and Overview*

Clinality in a trait, or a gradient in that trait that follows an environmental gradient, is often the first line of evidence for local adaptation in that trait (Clausen *et al*., 1940). Drawing on data from the qPCR experiment described above, clinality in *R*-gene basal expression and *R*-gene expression plasticity was explored in a worldwide set of six accessions (hereafter the "Clark worldwide set") argued to capture most of the species' genetic diversity without strong effects of population structure (Clark *et al.,* 2007). One limitation of this analysis was that it could not test whether clinality in *R*-gene expression was unusual compared to the clinality in the remainder of the transcriptome. To test this, the Gan *et al*., (2011) set of accessions (hereafter the "Gan worldwide set") was reanalyzed to provide a neutral expectation for expression clinality across the transcriptome for a worldwide set of accessions. A second limitation of this analysis was that the worldwide set was not distributed along a traditional cline. Having found weak evidence for clinality in *R*-gene expression in the worldwide dataset, two additional sets of accessions collected from latitudinal clines at spatial scales of hundreds of kilometers were examined for further evidence of *R*-gene expression clinality. One of these latitudinal clines consisted of the described qPCR data for five accessions in a Midwestern US latitudinal cline (hereafter the "Midwestern set"). The second latitudinal cline consisted of RNA-seq data of 144 accessions (hereafter the "Swedish set") from two Swedish populations grown in a common growth chamber condition (Dubin *et al*., 2015). To test if *R*-gene expression clinality was unusual, the Swedish set was used to provide a neutral expectation for expression clinality across the transcriptome.

#### *2.3.4.1 Clark Worldwide Set Analyses*

Natural variation in gene expression was first explored in the worldwide set to examine *R*-gene expression patterns across the entire species range. To correlate *R*-gene expression with historical climates from which accessions derived, four climate variables from Hancock *et al*. (2011) were considered: latitude, minimum temperature in the coldest month, precipitation in the driest month, and temperature seasonality (maximum temperature in the warmest month – minimum temperature in the coldest month). With the exception of latitude, these climate variables were highly correlated: more temperate climates had both warmer winters and were drier. Two of 12 accessions were eliminated from climate analysis: Col-0 and Tsu-1, accessions with unknown locations of origin (Anastasio *et al*., 2011).

To explore the association between basal expression and climate, LMs were first constructed relating *R*-gene expression in short day and in long day to accession, with *R*-gene treated as a random effect. Then, the model estimate of *R*-gene expression for each accession was regressed against each of four climate variables and model slopes were determined. Basal expression in short day conditions is more biologically relevant than long day conditions for *A. thaliana* collected at the eight-leaf stage, as this day length corresponds to spring growth conditions. *p*-values of the model slopes were determined to identify significant effects of climate on basal *R*-gene expression. For each treatment, model slope p-values were Bonferroni corrected for the number of climate variables compared.

To explore the interaction between accession and treatment, LMs were constructed relating *R*-gene expression and accession, in each of the seven treatments separately, with *R*-gene treated as a random effect. Average *R*-gene expression plasticity, or the fold-change in

expression upon perturbation, was then determined by calculating the differences between the model estimates of average *R*-gene expression levels after each treatment and "control" levels for plants that remained in long day. This method corrected for primer bias between accessions. To explore how historic climate influenced plasticity in *R*-gene expression upon environmental perturbation, LMs of average *R*-gene plasticity against four climate variables were constructed and model slopes were determined. For each treatment, *P*-values were Bonferroni corrected for the number of climate variables compared.

#### *2.3.4.2 Gan Worldwide Set Analysis*

To test if the temperature seasonality cline in *R*-gene expression was unusual, RNA-seq data from 14 of 19 accessions in the Gan *et al.,* (2011) study with climate data in Hancock *et al*., (2011) were used, and Col-0, Hi-0, Po-0, Oy-0, and Tsu-0 were excluded. To explore the relationship between temperature seasonality and average *R*-gene expression, a LM relating the normally distributed, average expression of all *R*-genes against temperature seasonality was constructed. Second, a null distribution of t-values of 1000 LMs of the average expression of 150 randomly sampled genes against temperature seasonality was constructed. A "stress response gene" distribution was also created by creating sets of 150 genes with a GO annotation of response to stress (GO:0006950). The t-value from the LM of *R*-gene expression was compared to these two null distributions to determine if the clinality of *R*-gene expression associated with temperature seasonality was likely to have occurred by chance, or likely to respond similarly to other stress response genes.

#### *2.3.4.3 Midwestern Set Analysis*

The Midwestern set was used to explore *R*-gene expression clinality along a latitudinal cline. In the Midwestern US, *A. thaliana* populations exhibit local isolation by distance that extends hundreds of kilometers (Platt *et al.,* 2009), and natural variation in *R*-genes has previously been found to be segregating in the Midwestern US (Karasov *et al*., 2014). The interaction between basal expression and climate in the Midwestern set was explored as described for the Clark worldwide set.

#### *2.3.2.4 Swedish Set Analyses*

The Swedish set was used to explore *R*-gene expression clinality in an independent latitudinal cline. This dataset included RNA-seq data from 144 accessions with climate data in Hancock *et al*., (2011) from two Swedish populations grown in a common growth chamber condition, and is described further in Dubin *et al*., (2015). Previous work has demonstrated that these Swedish accessions should be treated as two distinct populations (Huber *et al.,* 2014). Thus, three analyses were conducted. First, a LM of average *R*-gene expression against latitude, as in the qPCR data, was constructed. Second, sets of 150 randomly chosen genes were averaged, then t-values of the coefficients of 10000 LMs of these sets against latitude were determined to create a null distribution of how an expression dataset such as that for the *R*-genes would be expected to vary with latitude at random. The t-value from the LM of average *R*-gene expression was compared to this null distribution to determine if the latitudinal correlation of *R*-gene expression was likely to have occurred by chance. To control for population structure or effects of genome size that might differ between northern and southern Sweden (Long *et al.,* 2013), the expression of each *R*-gene and each control gene was tested within each population to

determine whether expression was significantly higher in the northern or southern Swedish population. To ensure that the gene sets were similarly variable, control genes were chosen to mimic the distribution of coefficients of variation seen for *R*-gene expression for all accessions, which are significantly higher than those of the transcriptome as a whole (C.V. = 0.23, 0.13; Kolmogorov-Smirnov $p$ = 7.36e-12).

## 2.4 Results

### *2.4.1 Current and historic environment interact to influence R-gene expression levels.*

The effects of the eight abiotic treatments (seven perturbations plus control) and 12 accessions on *R*-gene expression were determined using a LM of normalized expression data (Tables 2.4,2.5). The random effect of gene explained 55% of the variation within, and 28% of the variation was explained by an interaction between gene and accession (Table 2.5). Model comparisons using AIC, log likelihood, and $\chi^2$ values indicated that the two-way interaction between accession and treatment was highly significant. This interaction had a stronger effect on *R*-gene expression than the effect of accession or treatment alone (mean squares = 22.9, 12.9, 16.9), and abiotic treatment had a stronger effect on *R*-gene expression than the effect of accession alone (mean squares = 16.9, 12.9). The effects of treatment, accession, and their interaction in this dataset were thus all further explored; the effects of accession and its interaction with treatment were explored as functions of historic climate.

The coefficients of a LM of *R*-gene expression as a function of short term environmental perturbation, or treatment, were determined. *R*-genes were significantly upregulated in all treatments (Figure 2.1). The strongest response, an average 2.5-fold upregulation of *R*-genes,

occurred after three hours of cold shock (Figure 2.1). A more modest response (1.6-fold) was seen for the weeklong temperature and water treatments, and for the three-hour heat shock.

**Table 2.4**. Fixed effects for a mixed effect linear model fitted to the $\log_2$ of the normalized relative quantities of RNA, with treatment, accession, and their interaction modeled as fixed effects, and both *R*-gene and two-way interaction including *R*-gene modeled as random intercepts.

| **Fixed Effects** | **df** | **Sum Sq.** | **Mean Sq.** | ***F*** |
|---|---|---|---|---|
| Accession | 11 | 144.92 | 13.18 | 3.93 |
| Treatment | 7 | 213.0 | 30.43 | 9.08 |
| Accession*Treatment | 46 | 1056.8 | 22.98 | 6.86 |

**Table 2.5**. Random effects for the mixed effect linear model described in Table 2.

| **Random Effects** | **Name** | **Variance** | **Variance Component (Confidence Interval)** |
|---|---|---|---|
| Gene*Accession | (Intercept) | 5.72 | 0.282 (0.165, 0.443) |
| Gene | (Intercept) | 11.21 | 0.553 (0.319, 0.746) |
| Residual | | 3.35 | 0.165 (0.0978, 0.248) |

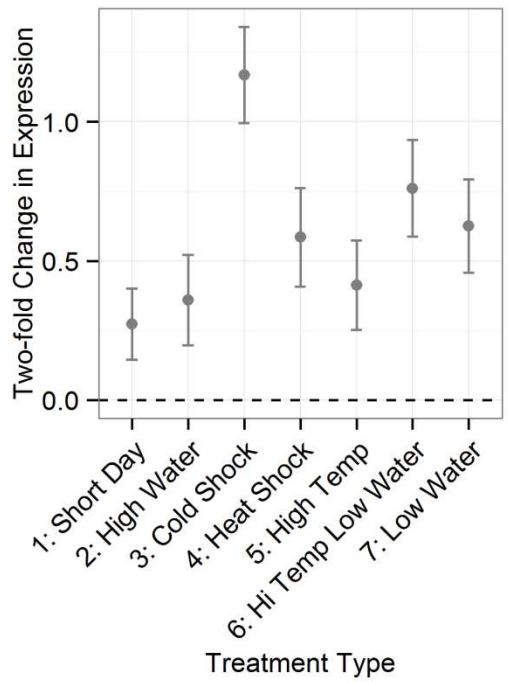



**Figure 2.1**. Coefficients of a linear model of *R*-gene expression against treatment for 13 *R*-genes in 12 accessions in *Arabidopsis thaliana*. *R*-gene and all interactions with *R*-gene were included as random effects in the model. Y-axis corresponds to a $\log_2$ fold-change in expression for each treatment relative to the control treatment.

### *2.4.2 R-gene expression responses are distinct from non-R-gene expression responses.*

The responses of *R*-genes and non-*R*-genes were compared using 15 published conditions that involved environmental perturbation followed by expression measurements hours to days later (Table 2.6). Significant differences in the response of *R*- and control gene sets were observed for infection with two pathogens, *Hyaloperonospora parasitica* and powdery mildew, and for treatment with salicylic acid (Table 2.6). Upregulation of more *R*-genes relative to control genes drove the difference in expression (2x2 chi-squared tests). *R*-genes were upregulated on average in 10 of 15 perturbations (Table 2.6). For 14 of 15 perturbations, there was more consistency in the differentially expressed *R*-genes towards upregulation or downregulation, while control genes were more evenly split between these groups.

To determine if the *R*-gene response was typical for stress response genes, the same comparisons were made to stress response genes. Significant differences in the response of *R*-gene and stress response sets were observed for infection with two pathogens, powdery mildew and *Pseudomonas syringae* pv. *maculicola* with *AvrRpt2*, treatment with salicylic acid, and cold treatment (Table 2.6). Again, the expression response of *R*-genes to treatment was more consistent than the response of stress-related genes, with stress response genes more evenly split in terms of the number of genes up- and down-regulated in each treatment.

### *2.4.3 Environment prior to infection can prime disease resistance.*

To explore the effects of environmental perturbation on pathogen resistance, plants exposed to cold shock, the perturbation with the strongest *R*-gene response, were infected with two pairs of *P. syringae* pathovars with and without the Avirulence genes *AvrPphB2* and *AvrRpt2*, cognate Avirulence genes of the *R*-genes *Rps5* and *Rps2.* The positive control, a SA

**Table 2.6.** The number of *R-,* control and stress response genes differentially expressed after fifteen environmental perturbations. Bolded *p*-values are significant correlations, and (*****) represent significant correlations after correction for multiple testing.

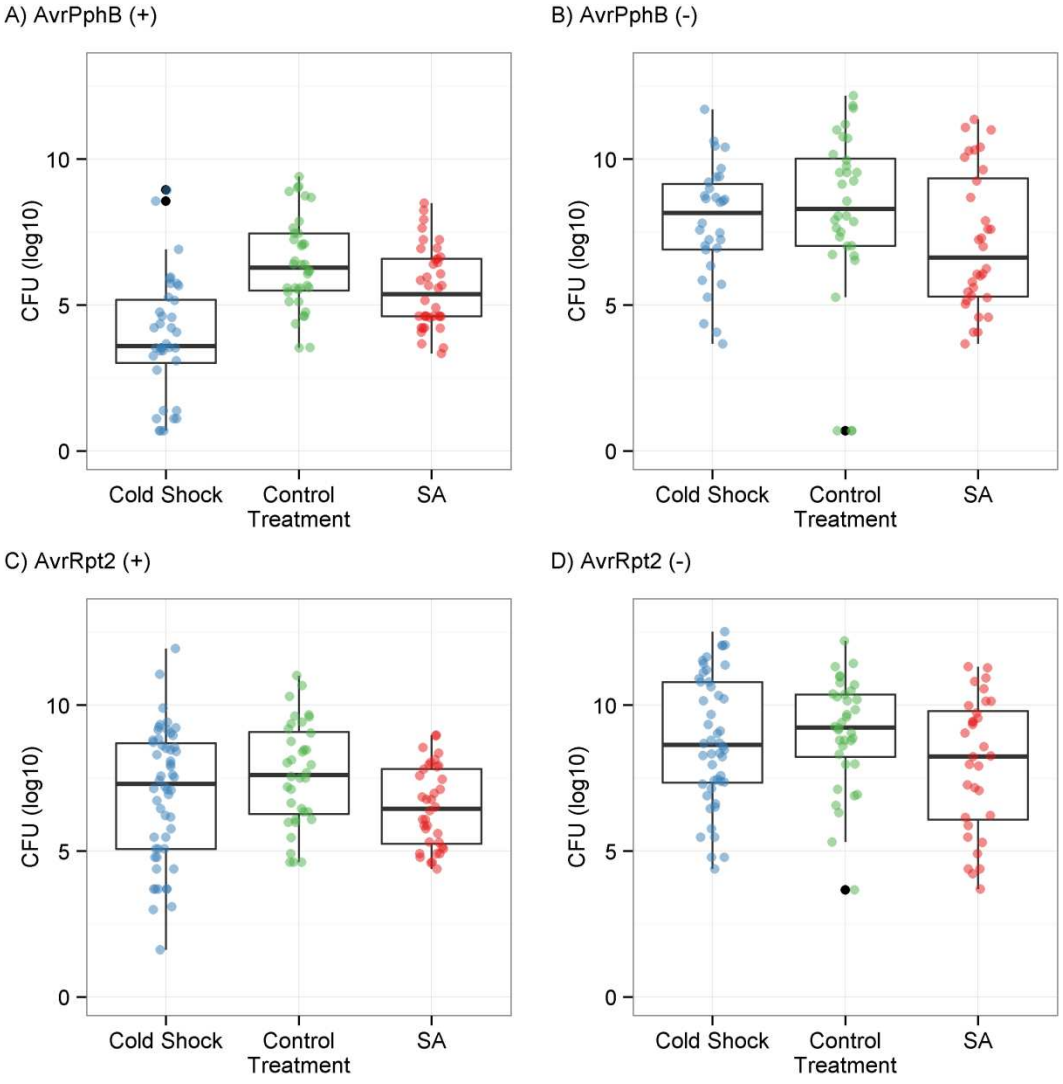
| Treatment and Gene Set | Up-regulated | No change | Down-regulated | Sets compared | $\chi^2$ value | *p*-value |
|---|---|---|---|---|---|---|
| *Blumeria graminis* vs uninfected | | | | | | |
| *R*-genes | 6 | 5 | 5 | *R* vs Control | 2.97 | 0.23 |
| Control | 5 | 15 | 6 | *R* vs Stress | 1.48 | 0.48 |
| Stress response | 11 | 7 | 17 | | | |
| *Hyaloperonospora parasitica arabidopsis Noco2* vs uninfected | | | | | | |
| *R*-genes | 33 | 25 | 2 | *R* vs Control | 18.7 | **8.7E-05*** |
| Control | 12 | 28 | 14 | *R* vs Stress | 8.65 | **0.013** |
| Stress response | 52 | 26 | 17 | | | |
| Powdery mildew (*Erisyphe cichoracearum*) vs uninfected | | | | | | |
| *R*-genes | 35 | 11 | 2 | *R* vs Control | 18.6 | **9.2E-05*** |
| Control | 5 | 18 | 1 | *R* vs Stress | 14.7 | **6.3E-04*** |
| Stress response | 22 | 31 | 8 | | | |
| *Pseudomonas syringae* pv. DC3000 with *avrRpm1*vs uninfected | | | | | | |
| *R*-genes | 8 | 3 | 3 | *R* vs Control | 2.84 | 0.24 |
| Control | 6 | 4 | 0 | *R* vs Stress | 3.28 | 0.19 |
| Stress response | 10 | 15 | 7 | | | |
| *Pseudomonas syringae* pv. *Maculicola* with *AvrRpt2* vs mock inoculated | | | | | | |
| *R*-genes | 36 | 16 | 9 | *R* vs Control | 4.73 | 0.094 |
| Control | 24 | 16 | 17 | *R* vs Stress | 12.1 | **2.3E-03*** |
| Stress response | 45 | 10 | 34 | | | |
| *Pseudomonas syringae* pv. tomato with *HopZ1a* vs mock inoculated | | | | | | |
| *R*-genes | 26 | 15 | 11 | *R* vs Control | 0.49 | 0.78 |
| Control | 16 | 8 | 9 | *R* vs Stress | 1.72 | 0.42 |
| Stress response | 40 | 20 | 27 | | | |
| *Pseudomonas syringae* pv. tomato vs mock inoculated | | | | | | |
| *R*-genes | 2 | 10 | 14 | *R* vs Control | 6.37 | **0.041** |
| Control | 4 | 6 | 2 | *R* vs Stress | 10.1 | **6.3E-03** |
| Stress response | 6 | 37 | 10 | | | |
| flg22 peptide vs water | | | | | | |
| *R*-genes | 4 | 4 | 0 | *R* vs Control | 3.05 | 0.22 |
| Control | 8 | 5 | 5 | *R* vs Stress | 7.35 | **0.03** |
| Stress response | 5 | 6 | 13 | | | |
| Salicylic acid vs Silwet, less than 24 hours after treatment | | | | | | |
| *R*-genes | 47 | 18 | 4 | *R* vs Control | 11.8 | **2.8E-03*** |
| Control | 14 | 17 | 8 | *R* vs Stress | 39.5 | **2.6E-09*** |
| Stress response | 33 | 76 | 31 | | | |

**Table 2.6.** (continued)

| Treatment and Gene Set | Up-regulated | No change | Down-regulated | Sets compared | $\chi^2$ value | *p*-value |
|---|---|---|---|---|---|---|
| Salicylic acid vs Silwet, more than 24 hours after treatment | | | | | | |
| *R*-genes | 0 | 18 | 21 | *R* vs Control | 1.24 | 0.54 |
| Control | 0 | 6 | 3 | *R* vs Stress | 9.53 | **0.01** |
| Stress response | 3 | 32 | 11 | | | |
| 0.5 or 2 hour excess light vs low light | | | | | | |
| *R*-genes | 5 | 33 | 34 | *R* vs Control | 9.22 | **0.010** |
| Control | 11 | 29 | 13 | *R* vs Stress | 10.6 | **5.1E-03** |
| Stress response | 35 | 57 | 48 | | | |
| 37°C vs 20°C | | | | | | |
| *R*-genes | 7 | 23 | 22 | *R* vs Control | 3.09 | 0.21 |
| Control | 11 | 15 | 13 | *R* vs Stress | 1.19 | 0.55 |
| Stress response | 18 | 39 | 32 | | | |
| 4°C vs 20°C | | | | | | |
| *R*-genes | 13 | 45 | 24 | *R* vs Control | 10.3 | **5.8E-03** |
| Control | 15 | 35 | 16 | *R* vs Stress | 24.8 | **4.0E-06*** |
| Stress response | 56 | 65 | 36 | | | |
| Drought vs untreated | | | | | | |
| *R*-genes | 18 | 30 | 4 | *R* vs Control | 3.50 | 0.17 |
| Control | 8 | 23 | 7 | *R* vs Stress | 8.77 | **0.012** |
| Stress response | 30 | 36 | 25 | | | |
| Ozone vs control | | | | | | |
| *R*-genes | 8 | 6 | 3 | *R* vs Control | 2.43 | 0.30 |
| Control | 2 | 6 | 3 | *R* vs Stress | 0.69 | 0.71 |
| Stress response | 17 | 9 | 9 | | | |

analog, increased resistance to all four pathogen strains (Figure 2.2). Cold shock did not lead to an increase in resistance to strains lacking Avirulence genes nor did it alter specific pathogen resistance to *P. syringae* pv. *avrRpt2* (Figure 2.2). In contrast, cold shock led to a significant increase in resistance to *P. syringae* pv. *avrPphB2* (Figure 2.2). The increase in resistance to *P. syringae* pv. *avrPphB2* was striking, a 340-fold decrease in CFU compared to an average nine-fold decrease in CFU with SA treatment.

**Figure 2.2.** Bacterial titers measured 3 days post infection for *Arabidopsis thaliana* plants exposed to three treatments then infected by two pairs of *Pseudomonas syringae* strains. Data points for barplots are included as jittered points. (A) Infection with *P. syringae* pv. *AvrPphB*; (B) Infection with *P. syringae* (-) *AvrPphB*; (C) Infection with *P. syringae* pv. *AvrRpt2*; (D) Infection with *P. syringae* (-) *AvrRpt2*.



### *2.4.4 Basal R-gene expression and R-gene expression plasticity vary clinally at a worldwide scale.*

To explore the effect of historic environment on *R*-gene expression in the Clark worldwide set, basal *R*-gene expression in short day was regressed against four climate variables. Basal *R*-gene expression in short day was correlated with temperature seasonality ($R^2 = 0.73$, $p = 0.031$), although this correlation was not significant after a Bonferroni correction. Here, higher basal expression of *R*-genes was associated with lower temperature seasonality, or a more temperate winter climate (Figure 2.3A). In the Gan worldwide dataset, average *R*-gene expression also varied significantly with temperature seasonality ($R^2 = 0.185$, $t=-2.349$, $p=0.022$; Figure 3B). However, null distributions of averaged expression of sets of randomly sampled genes and stress response genes both overlapped this value , with 4.8% of the null distribution and 16% of the stress response gene distribution as or more extreme than the relationship between *R*-genes and temperature seasonality (Figure 2.3C).

To explore the interaction between accession and treatment in the Clark worldwide set, the relationships between historical climate variables and *R*-gene expression plasticity in seven environmental perturbations in the qPCR set were examined. There were no significant clines in *R*-gene expression when plants were perturbed with short day, high water, cold shock, or heat shock. Expression changes after mild drought were significantly correlated with minimum temperature in the coldest month and temperature seasonality (Table 2.7). In particular, accessions from temperate climates, with milder winters, tended to downregulate *R*-genes significantly more than accessions from more continental climates, with colder winters (Figure 2.4D, E). Expression changes in response to high temperature, mild drought, or both were

correlated with precipitation in the driest month (Table 2.7). Accessions from dry climates downregulate *R*-genes significantly more than accessions from wetter climates in increased heat and aridity (Figure 2.4B, C, E).

**Figure 2.3**. Clinal variation in basal *R*-gene expression in *Arabidopsis thaliana*. Dashed lines are regression lines. (A;D) Y-axes corresponds to the $\log_2$ basal expression for each accession in short day conditions. (A) Expression variation of 13 *R*-genes in accessions from the Clark worldwide set against temperature seasonality at the accession's location of origin. (B) Average expression of 150 *R*-genes in accessions from the Gan worldwide set against temperature seasonality at the accession's location of origin. (C) The null distributions of t-values of correlations with temperature seasonality for 1000 sets of 150 randomly sampled gene expression profiles from the Gan worldwide set. Null distributions were drawn from all genes (blue) and stress response annotated genes (green). The dashed vertical line is the t-value of the correlation from (B). (D) Expression variation of 13 *R*-genes in accessions from the Midwestern set against latitude of origin. (E) Average expression of 150 *R*-genes in accessions from the Swedish set against latitude of origin. (F) The null distributions of t-values of correlations with temperature seasonality for 1000 sets of 150 randomly sampled gene expression profiles for the Swedish set. Null distributions were drawn from all genes (blue) and stress response annotated genes (green). The dashed vertical line is the t-value of the correlation from (E).

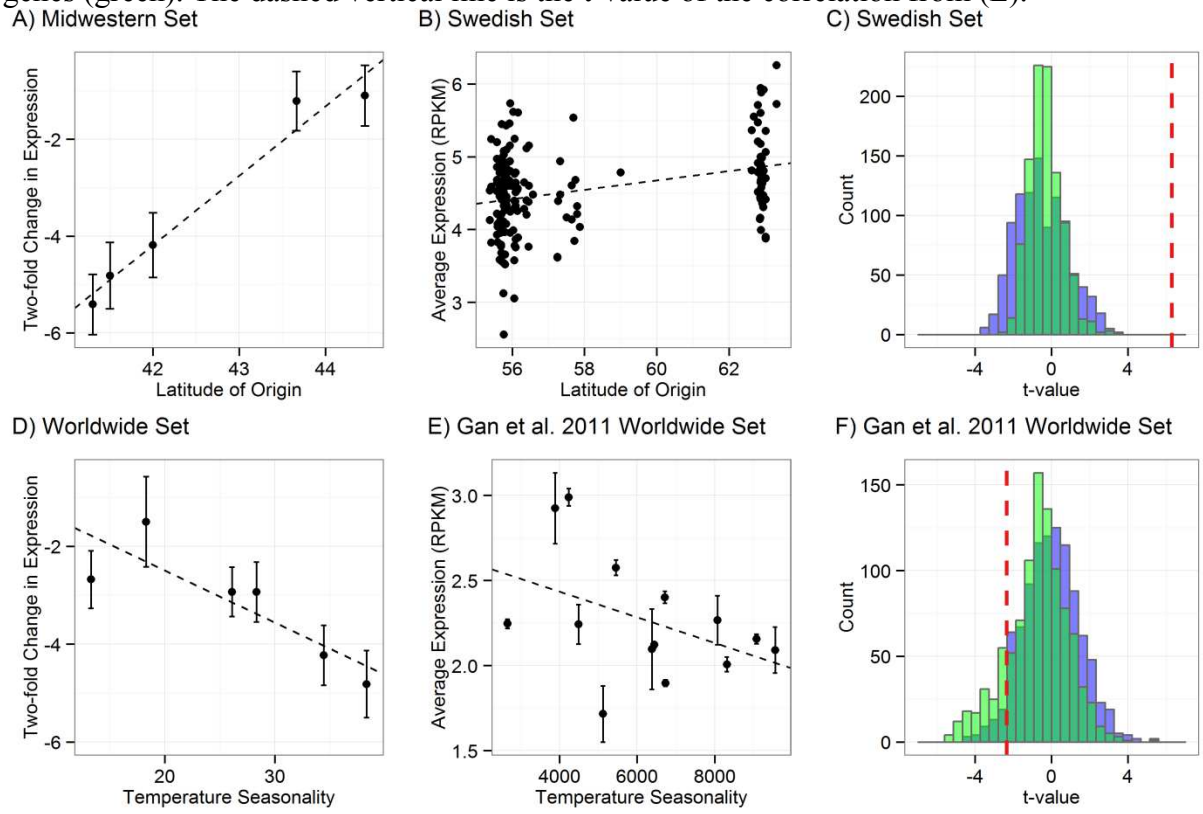
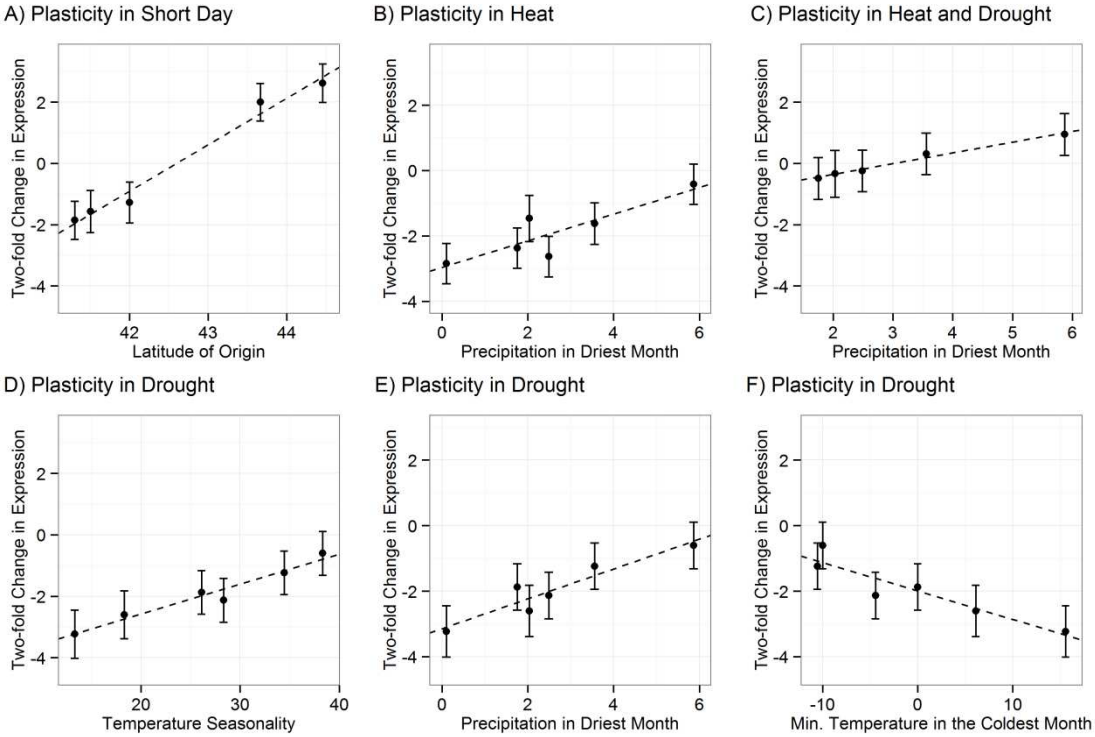


**Table 2.7**. Correlations between average *R*-gene expression and four climate variables. Significant correlations of both basal *R*-gene expression and *R*-gene expression plasticity after environmental perturbation are shown. Bolded *p*-values are significant correlations, and (*****) represent significant correlations after corrections for multiple testing to four climate variables.

| Accessions | Climate Variable | Treatment | Slope | $R^2$ | Adj. $R^2$ | *F* | *p*-value |
|---|---|---|---|---|---|---|---|
| *Basal Expression* | | | | | | | |
| Midwest | Latitude | Short Day | 1.435 | 0.964 | 0.953 | 81.2 | **0.003*** |
| Worldwide | Temperature Seasonality | Short Day | -0.107 | 0.728 | 0.661 | 10.75 | **0.031** |
| *Expression Plasticity* | | *Fold change in response to:* | | | | | |
| Midwest | Latitude | Short Day | 1.516 | 0.981 | 0.975 | 157.7 | **0.001*** |
| Worldwide | Temperature Seasonality | Low Water | 0.097 | 0.954 | 0.943 | 83.5 | **<0.001*** |
| Worldwide | Min. Temp. in the Coldest Month | Low Water | -0.087 | 0.864 | 0.830 | 25.4 | **0.007*** |
| Worldwide | Precipitation in the Driest Month | Low Water | 0.456 | 0.887 | 0.858 | 31.2 | **0.005*** |
| Worldwide | Precipitation in the Driest Month | High Temp & Low Water | 0.350 | 0.983 | 0.977 | 169.3 | **<0.001*** |
| Worldwide | Precipitation in the Driest Month | High Temp. | 0.408 | 0.764 | 0.705 | 12.9 | **0.023** |

**Figure 2.4**. Clinal variation in resistance (*R*-) gene expression plasticity in *Arabidopsis thaliana.* X-axes correspond to historical climate variables. Y-axes correspond to a $\log_2$ fold-change in expression for each treatment relative to the control treatment. Dashed lines are the regression lines. (A) Plasticity in *R*-gene expression upon change to short day correlates with latitude of origin for accessions in the Midwestern set. (B, C) Plasticity in *R*-gene expression after heat, and heat and drought stress correlates with precipitation differences at the accession's location of origin in the Clark worldwide set. (D, E, F) Plasticity in *R*-gene expression after drought stress correlates with precipitation differences and temperature differences at the accession's location of origin in the Clark worldwide set.



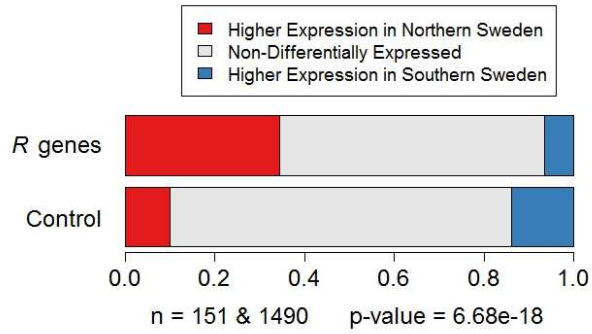

### *2.4.5 Basal R-gene expression increases with latitude in two latitudinal clines.*

In the Midwestern set, there was a significant correlation between basal *R*-gene expression in short days and latitude of origin ($R^2 = 0.96$, $p = 0.003$), with accessions from higher latitudes exhibiting higher basal expression levels (Figure 2.3D). Other climate variables were uncorrelated with basal *R*-gene expression in short day conditions. In the Midwestern set, the responsiveness of *R*-gene to a change to short day was also significantly correlated with

latitude of origin ($R^2$ = 0.97, p = 0.002; Figure 2.4A). Other climate variables were uncorrelated with changes in expression upon a transition to short day.

In the Swedish set, a LM indicated that average *R*-gene expression was significantly higher in the northern than the southern Swedish population ($R^2$ = 0.116, t=6.29, p = 2.9e-07; Figure 2.3E). The t-value of the coefficient of this relationship was higher than the entire null distribution of sets of randomly sampled genes and stress response genes (Figure 2.3F). Higher gene expression in northern Sweden was not seen on a gene-by-gene basis: control genes had roughly equal proportions of genes with higher gene expression in northern and higher gene expression in southern Sweden, while the vast majority of differentially expressed *R*-genes were expressed at higher levels in the northern population (Figure 2.5).

**Figure 2.5.** Clinal variation in basal expression with latitude or population in the Swedish RNA-seq dataset. The proportion of *R*-genes and control genes that are expressed significantly higher in the northern than the southern Swedish population, non-differentially expressed between populations, or expressed significantly higher in the southern than the northern Swedish population.



## 2.5 Discussion

*R*-genes were upregulated in response to all abiotic treatments in this expression study, with the largest upregulation occurring after cold shock (Figure 2.1). Neither the duration of the short-term treatment nor type of short-term treatment affected the direction of the average *R*-gene response. The metastudy reveals that not only are *R*-genes typically upregulated in response to perturbations but, for many perturbation responses, the *R*-gene expression response is more consistent and biased towards upregulation than the response of control or stress response genes (Table 2.6). The bias towards upregulation is not consistent with the idea that plants track the weather, increasing or decreasing *R*-gene expression in response to how favorable the environment is for pathogen growth. Instead, this suggests that changes in the environment *per se* increases risk of infection.

It is possible that the *R*-gene expression responses are adaptive because species are more likely to successfully invade a community after disruption (Shea & Chesson 2002). For bacterial communities in particular, both biotic and abiotic environmental factors have been shown to alter the plant microbiome (Kadivar & Stapleton, 2003; Suda *et al.,* 2009; Yutthammo *et al.,* 2010; Ikeda *et al.,* 2011), and changes in abiotic conditions modulate the risk of infection by both latent and novel pathogens (Szittya *et al.,* 2003; Yi *et al.,* 2004, Monteiro *et al.,* 2012). While details of the relationship between microbiome structure and plant resistance are yet to be understood, many microbial species take part: bacteria from Proteobacteria, Firmicutes, and Actinobacteria are known to contribute to the suppression of fungal and bacterial pathogens (Innerebner *et al.,* 2011; Mendes *et al.,* 2011) and axenic plants are known to be especially susceptible to infection by pathogens (Innerebner *et al.,* 2011). The significant increase in disease resistance after cold

shock demonstrates that the environment can prime disease resistance (Figure 2.2). Thus, the consistent upregulation of *R*-genes in response to environmental change could be a preparatory response to conditions that alter the probability of invasion by pathogens.

Clinal interactions between *R*-gene expression plasticity and historic environment support fine-scale genetic tinkering to adapt *R*-gene expression levels to the environment (Table 2.7). The significant interaction between accession and abiotic treatment suggests that selection has shaped the plasticity of *R*-gene expression as a function of the genotype collected from different historical climates (Table 2.4). The Clark worldwide set has an eight-fold difference in basal *R*-gene expression across the range of temperature seasonality in datasets (Figure 2.3A). This expression difference was recapitulated for all *R*-genes in the Gan worldwide set (Figure 2.3B). However, both null distributions showed more extreme clinality than *R*-genes (Figure 2.3C); thus, this expression difference could have arisen through random drift in *R*-gene expression levels between populations. Alternatively, permissive growth chamber conditions may fail to mimic the environment in which *R*-gene expression is under selection. In other words, phenotypic consequences of *R*-gene expression in different environments could affect *R*-gene expression levels in the conditions tested here. In this vein, a four-fold difference in average *R*-gene expression was observed in response to drought stress across the climatic range of temperatures and aridity in the Clark worldwide set (Figure 2.4D,E,F). Four of these plasticity clines were significant after a Bonferroni correction for multiple climate variables, while the basal expression cline was not significant after a Bonferroni correction (Table 2.7). Though basal *R*-gene expression levels were higher in accessions from more temperate climates, accessions from these climates downregulated *R*-genes in drought to a much greater degree than accessions

from colder, wetter climates (Figure 2.3, 2.4). Similarly, accessions from dry climates downregulated *R*-genes in response to increased heat and aridity more than accessions from wet climates (Figure 2.4B, C). The consistency of the accession-specific expression responses to heat, drought, and heat plus drought strengthens support for a relationship between *R*-gene expression plasticity and aridity of the historic climate (Table 2.7).

The observed expression responses are inconsistent with the idea that plants track the weather, increasing or decreasing *R*-gene expression in direct response to how favorable the environment is for pathogen growth. Instead, they point to the possibility that changes in the environment *per se* act to promote infection by pathogenic species. It is known that rare environmental disruptions are more likely to promote invasion than common environmental disruptions (Sher & Hyatt, 1999). Thus, identical weather events may differentially affect accessions depending on their historical climates. For example, wetter climates typically experience fewer droughts, and this rare disruption may select for relative upregulation of *R*-genes in plants from wetter climates. Thus, *R*-gene upregulation may be a preparatory response to an increased likelihood of microbiome invasion. These interactions between weather and climate generally support the idea that *R*-gene expression changes are a mechanism to defend against changes to the microbial community inside the plant. *R*-genes may be induced to prevent plant microbiome invasion when an unusual perturbation occurs.

Historical climate appears to particularly influence basal *R*-gene expression across latitudinal clines at medium spatial scales. The Midwestern set shows a strong trend in basal *R*-gene expression, with the northernmost accession expressing *R*-genes 20-fold more strongly than the southernmost accession (Figure 2.3D). In the Swedish set, latitude explained a small, but

significant, fraction of the variation in average *R*-gene expression. Moreover, this distribution of *R*-gene expression values is extremely unlikely to have arisen by chance or by drift in gene expression between populations (Figure 2.3F); thus, *R*-gene expression is likely under differential selection between the two Swedish populations. Gene by gene, *R*-gene expression was consistently higher in the northern Swedish population than the southern Swedish population: 26 of 29 *R*-genes that were significantly differentially expressed between populations after a Bonferroni correction were expressed at a higher level in northern Sweden (Figure 2.5). These clinal patterns are consistent with the resource allocation hypothesis, which suggests that defense should increase as plant tissue becomes more difficult to acquire (Coley *et al.,* 1985), such as in locations with shorter day lengths. Of course, many other factors, such as pathogen distributions and abundance, may select on *R*-gene expression patterns and explain the residual variation in *R*-gene expression. Unfortunately, little is known about the distribution of *A. thaliana* pathogens through time and space.

In conclusion, inducible defenses are a cost-saving strategy, and are theorized to occur only in environments where they confer a fitness benefit to the individual (Cipollini 2008). *R*-genes initiate one such inducible defense, and their upregulation can benefit plant fitness through increased pathogen resistance but can be costly in the absence of pathogens. *R*-gene family expression is distinct from an average transcriptome response after biotic perturbations and across a latitudinal cline, supporting the idea that *R*-gene expression is shaped by these additional selective forces. *R*-genes are induced when the biotic or abiotic environment is perturbed, especially when that perturbation is unusual given the historical climate. Basal *R*-gene expression increases at higher latitudes, consistent with the resource allocation hypothesis. Fine-

scale genetic tinkering for differential expression responses to climate and weather perturbations may help explain the atypically high natural variation in *R*-gene expression.

# Chapter 3

## The long-term maintenance of *Rps2* alleles in *Arabidopsis thaliana*

### 3.1 Abstract

The mounting evidence that *R*-genes incur large fitness costs raises a question: how can there be a 5-10% fitness reduction for all 149 *R*-genes in *Arabidopsis thaliana*? The two *R*-genes tested to date segregate for insertion-deletion polymorphisms where the susceptible alleles carry no costs. Since costs of resistance are measured as the differential fitness of isolines carrying resistant and susceptible alleles, insertion-deletion polymorphisms necessarily exhibit augmented costs, which may be masked in *R*-genes with other architectures. *Rps2* segregates for two expressed clades of alleles, one resistant and one susceptible. Here we show that plants with resistant *Rps2* are not less fit than those with susceptible *Rps2* alleles in the absence of disease. Instead, all expressed alleles provide a fitness benefit relative to an artificial deletion, due to the role of RPS2 as a negative regulator of defense. Variation in costs of resistance suggests an interaction between costs and *R*-gene architecture.

### 3.2 Introduction

Understanding how plants maximize fitness in response to intermittent pathogen presence is of central importance in plant pathology. Pathogen resistance may involve two distinct fitness costs. The first, a cost of surveillance, accrues from carrying *R*-genes that allow a resistance response upon attack, and the second, a cost of defense, accrues from activation of the resistance response during attack (Korves and Bergelson 2004). A cost of surveillance is measured in the absence of disease and has been measured for two *R*-genes, *Rpm1* and *Rps5*, in *Arabidopsis thaliana* (Tian *et al.*, 2003; Karasov *et al.*, 2014). *Rpm1* and *Rps5* exist in nature as long-lived

insertion-deletion polymorphisms (indels) for resistance (R) and susceptibility (S) (Grant *et al*., 1995; Stahl *et al*., 1999; Tian *et al*., 2002); in both cases, resistant isolines suffer a 5-10% fitness cost relative to null isolines (Tian *et al.*, 2003; Karasov *et al.*, 2014). However, costs of this magnitude would correspond to an impossibly high genetic load if seen for all, or many, of the ~149 *R*-genes in *A. thaliana*. We propose that indels represent an unusual architecture that may exacerbate costs of surveillance. Null alleles of indels cannot carry any burden of mis-expression or mis-activation that would reduce relative costs of surveillance. Furthermore, null alleles do not have the potential to evolve pleiotropic or alternative functions, another means of reducing costs of surveillance. *R*-genes are found with a great diversity of genetic architectures, including single loci with many, functional alleles and arrays of tandem duplicated *R*-genes (Meyers *et al*., 2003). Here, we explore the possibility that *R*-genes with alternative genetic architectures have substantially smaller costs than those measured for the indels *Rpm1* and *Rps5*. We posit that the large costs associated with these two *R*-genes are precisely the reason susceptible alleles are deleted.

*Rps2* exists as an ancient balanced polymorphism with two long-lived clades of alleles, one resistant and one susceptible to *Pseudomonas syringae* pv. *avrRpt2*, maintained at intermediate frequencies in local populations (Kunkel *et al*., 1993; Yu *et al*., 1993). *Rps2* is also present in every accession sequenced to date; that is, no sequenced accession has missing data or deletions called for *Rps2*. To measure the surveillance cost of alleles of *Rps2,* we assayed fitness in the absence of disease for a transgenically-created allelic series of *Rps2*. It is essential to control for the insertion site of the *R*-gene alleles when testing for costs of resistance (Bergelson and Purrington 1996). This is done relatively easily with a Cre-lox system when the presence or

absence of a single allele is involved (as done for *Rpm1* (Tian *et al.*, 2003) and *Rps5* (Karasov *et al.*, 2014)). Here, we extended this basic strategy to precisely integrate five alleles of *Rps2* into the same genomic location using Col-0 in which *Rps2* has been knocked out as a genetic background (Yu *et al*., 1993). We additionally verified the robustness of our results by using three genomic locations. Our results revealed that resistant *Rps2* alleles do not carry a large fitness cost of surveillance relative to susceptible alleles. Instead, we find a fitness benefit of all *Rps2* alleles relative to the artificial knockout, due to the role of RPS2 in negative regulation of the defense response. Field fitness decreased as *Rps2* expression increased, consistent with previously reported costs of *Rps2* overexpression (Mindrinos *et al*., 1994; Tao *et al*., 2000). Our results suggest that defense loci segregating for functional alternatives, rather than presence-absence polymorphisms, limit the manifestation of costs of surveillance.

## 3.3 Methods

### 3.3.1 Cre-lox Insertion of RPS2

We introduced five intact alleles of *Rps2* (Appendix B.1) into the same genomic location using a Cre-lox system into an *rps2-101c* mutant of Col-0, a plant line with a stop codon in RPS5 at amino acid 235 that is a presumed null mutation (Yu *et al.* 1993). We repeated this process for each of three genomic locations (Table 3.1). These lines were created by Xiaoqin Suen between 2007 and 2009, before I took over the project. Three alleles from the resistant clade, Col-0, Ct-0, and Ler-0 (hereafter *R* clade alleles or $Rps2^{R}$ lines), were characterized as resistant in their native genetic background (Ana *et al*., 1999; Rodney *et al*., 2003). One allele from the *R* clade, Ws-0 (hereafter $Rps2^{pR}$), was characterized as partially resistant in its native genetic background (Ana *et al*., 1999). One allele from the susceptible clade, Wu-0 (hereafter S clade allele or $Rps2^{S}$ lines)

was characterized as susceptible in its native genetic background (Kunkel *et al*., 1993). We induced excision of *Rps2* from each set of lox sites to obtain empty vector insertions in the isogenic RPS2 null background (hereafter $Rps2^{KO}$). The $Rps2^{R}$ and $Rps2^{pR}$ lines exhibited elevated HR and resistance compared to the $Rps2^{S}$ and $Rps2^{KO}$ lines upon infection with *P. syringae* pv. *avrRpt2* (Figure 3.1, 3.2).

To create lox-RPS2-lox lines, we cloned a 5.3 kb region containing *Rps2*, its natural promoter and its full terminator, from five donor accessions of *A. thaliana* between two lox sites in the vector pBS246 (Gibco-BRL catalogue no. 10349-019; Life Technologies). We then inserted the cloned *Rps2* into the binary vector pBin19, which included the selectable marker nptII outside the lox sites. This construct was introduced by vacuum infiltration into the rps2-101c mutant of Col-0.

**Table 3.1.** Genomic locations for the three insertion sites for alleles of *Rps2*.

| Insertion Site | Chromosome | Location | Annotation |
|---|---|---|---|
| 1 | 5 | 23041468 | Intergenic (At5g56910 – At5g56920) |
| 2 | 3 | 4289059 | Intergenic (At3g13270 – At3g13275) |
| 3 | 2 | 13898787 | Intergenic (At2g32750 – At2g32760) |

**Figure 3.1.** Hypersensitive Response (HR) for all 17 isogenic lines. HR was measured as differential tissue collapse 19 hours post inoculation with *P. syringae* pv. *avrRpt2* relative to *P. syringae* without *avrRpt2*. Dotted line shows the line of no differential response between the pair of pathogens.

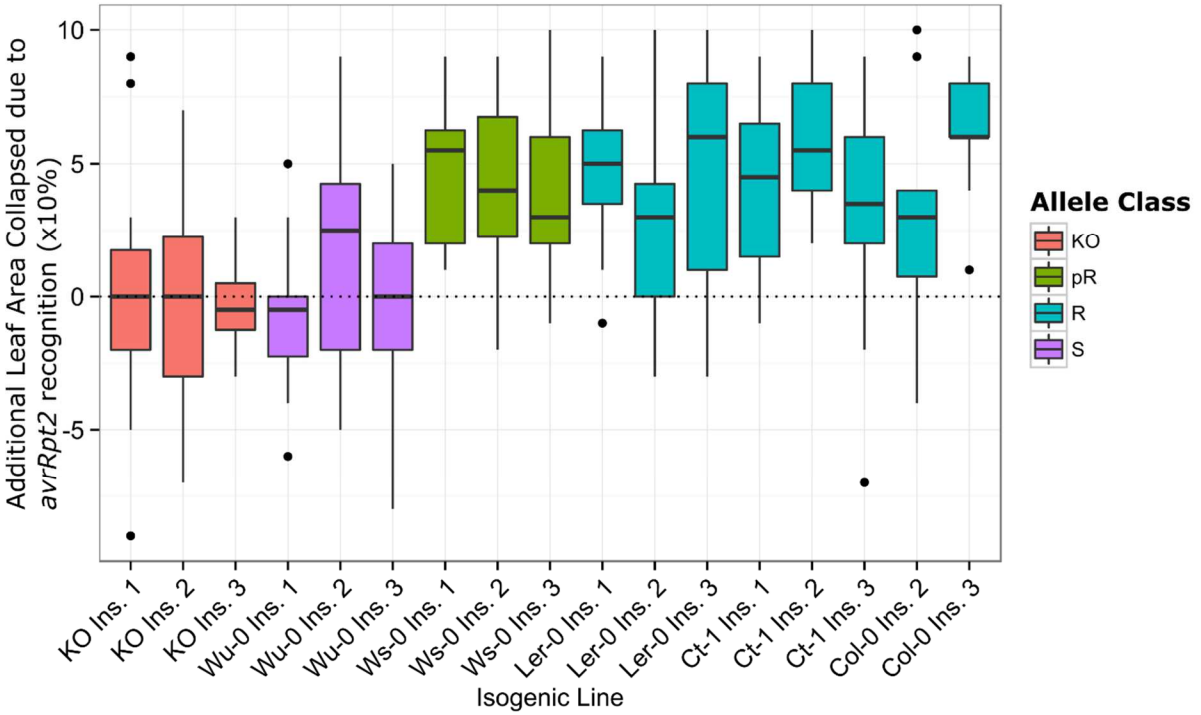
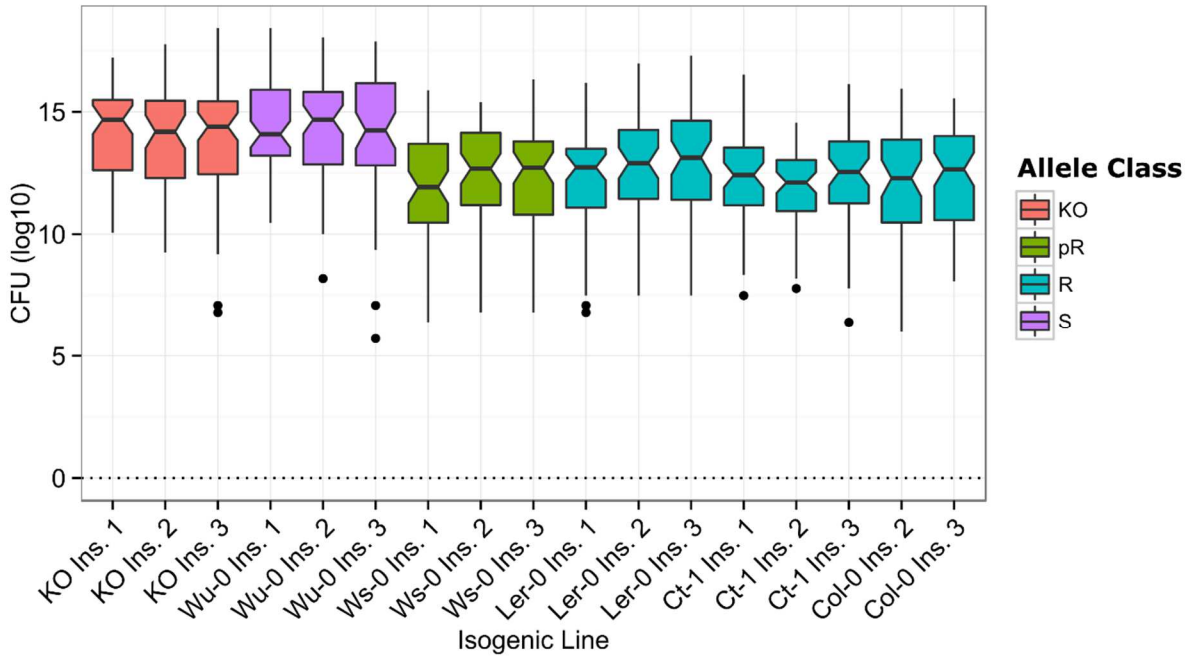



**Figure 3.2.** Resistance to *P. syringae* pv. *avrRpt2* three days post infection for all 17 isogenic lines.



To create Cre lines, we cloned a 3.43-kb region containing Cre with the 35S promoter and Nos30 terminator, as well as the selectable markers Basta and GUS, into the binary vector pCAMBIA3301. We introduced this vector into an rps2-101c Col-0 plant by agrobacterium-mediated transformation. The lox-RPS2-lox lines were emasculated and then pollinated by heterozygotes for five Cre insertions. In the F2 generation, the presence of Cre was negatively selected with Basta (AgroEvo USA) diluted 1:1,100.

The remaining plants were screened by PCR to identify individuals with each insertion site that had lost *Rps2*. After confirmation that these lines were not resistant to GUS or Basta, each pair was self-pollinated and the homozygosity of lines with *Rps2* (hereafter $Rps2^{+}$) or $Rps2^{KO}$ was confirmed by PCR. Inserts were sequenced in each *Rps2* line to confirm their integrity, and in $Rps2^{KO}$ lines to confirm clean excision. Southern blots and anchor PCR tests were consistent with the presence of only a single insert for each transformed line.

### 3.3.2 Field Fitness Experiment

I conducted the remainder of the experiments and analyses described in the Methods and Results. Seedlings of each of 17 RPS2 lines were germinated in 98-cell trays containing 50:50 Metromix 200: Farfad C2 in the University of Chicago greenhouse. Plants with the Col-0 allele at insertion site one did not germinate. Seedlings were thinned on day 7 of growth and flat location was randomly cycled daily in the greenhouse to standardize growth conditions. On day 14, 100 seedlings per line were transplanted to a tilled field site in Downers Grove, Illinois, in a randomized block design in which each block contained a plant from each *Rps2* line. Plants were set out in 15 rows of 9 blocks, spaced by 0.25m within rows and by 1m between rows. Plants were irrigated for one week to reduce transplantation shock, and then sustained only by natural

rainfall. The field was hand weeded once and plants received no other protection from competition or pests. 55% of plants died; the majority of these died in the first week presumably due to transplantation shock. Plant survival was evenly distributed among the allele types. Seven fitness proxies were measured: dry weight, undehisced seed set, seed size, silique number, average seeds per silique, number of aerial and basal branches. Total seed set was estimated by multiplying silique number by the average number of seeds per silique.

### 3.3.3 Sampling for Pathogen Presence

96 plant samples representing all 17 lines were destructively harvested from the field on days 30 and 40 of growth. The leaf endophytic community of these lines was sampled by plating aboveground leaf tissue onto KB plates and observing growth after 24 and 48 hours. No effect of plant line on leaf endophytic bacterial density was seen (data not shown), nor were visible signs of disease observed on plants in the field. The presence or absence of bacteria, *P. syringae*, and the *P. syringae* effector *avrRpt2* was determined by PCR. *P. syringae* was present in the majority of the plant samples, while *avrRpt2* was not seen in any sample. *P. syringae* pv. *avrRpt2* was thus either absent or at extremely low frequency in the field.

### 3.3.4 Growth Chamber Fitness Experiment

1600 seedlings of 7 *Rps2* lines from the P insertion site were germinated in 36-cell trays containing 25:25:50 Metromix 200:Farfad C2:Turfase in the University of Chicago growth chambers. Seedlings were thinned and accessions were randomized within flats on day 7 of growth. After day 14, plants were watered every other day to mimic stressful growth conditions in the field. After six weeks of growth, I stopped watering and allowed the plants to dry for two

weeks before processing. Two fitness proxies were measured: dry weight and undehisced seed set.

### 3.3.5 Sterile Plant Fitness Experiments

Two sets of lines were grown in sterile conditions. First, isogenic *Rps2* lines with insertion site two were grown to measure fitness of plants with alleles of *Rps2* relative to the *Rps2* knockout. Second, two isogenic *Rps2* lines from one insertion site, the Col-0 allele and the empty vector, Col-0, and an amiRNA knockdown of RPS2 in a Col-0 background were grown to measure fitness of our isogenic lines relative to the wild type and a distinct *Rps2* mutant. Seeds were surface sterilized with 70% and 100% ethanol, then dried under flame. One seed per well was germinated on MS media with agar in 12-cell sterile plates, and lines were evenly distributed between plates. At 21 days, plants were removed from agar and both the total plant mass and aboveground rosette mass were weighed.

### 3.3.6 Fitness Analysis

Two sets of nested linear models were used to generate confidence intervals for each of seven fitness proxies for the field data, and for the two fitness proxies for the growth chamber data. The field linear models included an effect of the date the plant was collected from the field. The growth chamber linear models included a block effect. The first set of models nested allele into one of five allele classes. The second set of models nested *Rps2* expression level into one of five allele classes.

### 3.3.7 Quantitative real-time PCR

Three to five replicates were flash frozen in liquid nitrogen on day 21 or 23 of growth between the fifth and the seventh hour of the light cycle. RNA was extracted from these tissues

using the protocol from Oñate-Sánchez and Vicente-Carbajosa (2008). Expression of all *Rps2* isogenic lines was measured with qPCR using primers for *Rps2* and normalizing between samples using three reference genes (Czechowski *et al*., 2005). A subset of isogenic lines was used to compare *Rps2* expression in natural accessions to expression in the allelic series. 20 µL qPCR reactions were performed in optical 96-well plates on an ABI 7900HT sequence detection system (Applied Biosystems). SYBR Green was used for dsDNA synthesis detection and ROX red was used to control for between sample fluorescence differences. Three technical replicates were run per biological sample. The following thermal profile was used for all qPCR reactions: 95 °C for 1 min, 40 cycles of 95 °C for 30 s, 51 °C for 30s, and 68 °C for 40s, followed by 95 °C for 15s. Dissociation curves of the amplicons were then determined by heating from 60 °C to 95 °C with a 2% ramp rate.

### 3.3.8 Whole Transcriptome Profiling

Plants with the Col-0 allele at insertion site two (R) and three (High R), Wu-0 allele at insertion site two (S), and empty vector at insertion site two (KO) were grown in sterile growth media as in the sterile plant fitness experiments. Plants were flash frozen and RNA was extracted as for qPCR. Two biological replicates per line were sequenced at the University of Chicago Genomics core using 50bp single-end RNA-seq indexing eight samples per lane. Sequences were aligned to the Col-0 reference sequence with bowtie2 (Langmead and Salzberg 2012), differential expression was analyzed with DESeq (Anders and Huber 2010), and enrichment of differentially expressed genes between contrasts was determined with Amigo (The Gene Ontology Consortium 2015). Specific contrasts made included R vs S; R, High R, and S vs KO; and High R vs R.

## 3.4 Results

### *3.4.1 No cost of surveillance for R alleles relative to S alleles of Rps2*

To test for surveillance costs of *Rps2* alleles, we measured the fitness of eight resistant (*Rps2*$^{R}$) and three partially resistant (*Rps2*$^{pR}$) lines relative to three susceptible (*Rps2*$^{S}$) lines in a field experiment performed in the absence of *avrRpt2* (Figure 3.3, Table 3.1). After correction for multiple testing, there were no significant differences in the fitness of isogenic plants carrying *Rps2*$^{pR}$ and *Rps2*$^{S}$ alleles at the same insertion site (Table 3.2-3.4)., and two fitness proxies supported higher fitness for *Rps2*$^{R}$ alleles than Rps2$^{S}$ alleles (Table 3.3). At the 5% level, three field fitness proxies supported *Rps2*$^{R}$ alleles as more fit than *Rps2*$^{S}$ alleles, and four supported *Rps2*$^{S}$ alleles as more fit than *Rps2*$^{R}$ alleles (Table 3.2-3.4), while one field fitness proxy supported *Rps2*$^{pR}$ alleles as more fit than *Rps2*$^{S}$ alleles, and one supported *Rps2*$^{S}$ alleles as more fit than *Rps2*$^{pR}$ alleles (Table 3.2, 3.4) Additionally, there was considerable variation in the fitness associated with particular susceptible or resistance alleles across insertion sites (Figure 3.4a-c).

**Figure 3.3**. Natural variation in *Rps2* captured by the transgenic allelic series. a) The two clades of *Rps2* allele inferred using the Cao *et al*., (2009) 80 genomes data for *Rps2* and Sanger sequencing of the five alleles used in this study. b) Amino acid variation in the alleles used in this study. Red bar indicates the Leucine-rich repeat region of *Rps2*; *Rps2*$^{R}$ and *Rps2*$^{pR}$ lines are resistant to *Pseudomonas syringae* pv. *avrRpt2*.

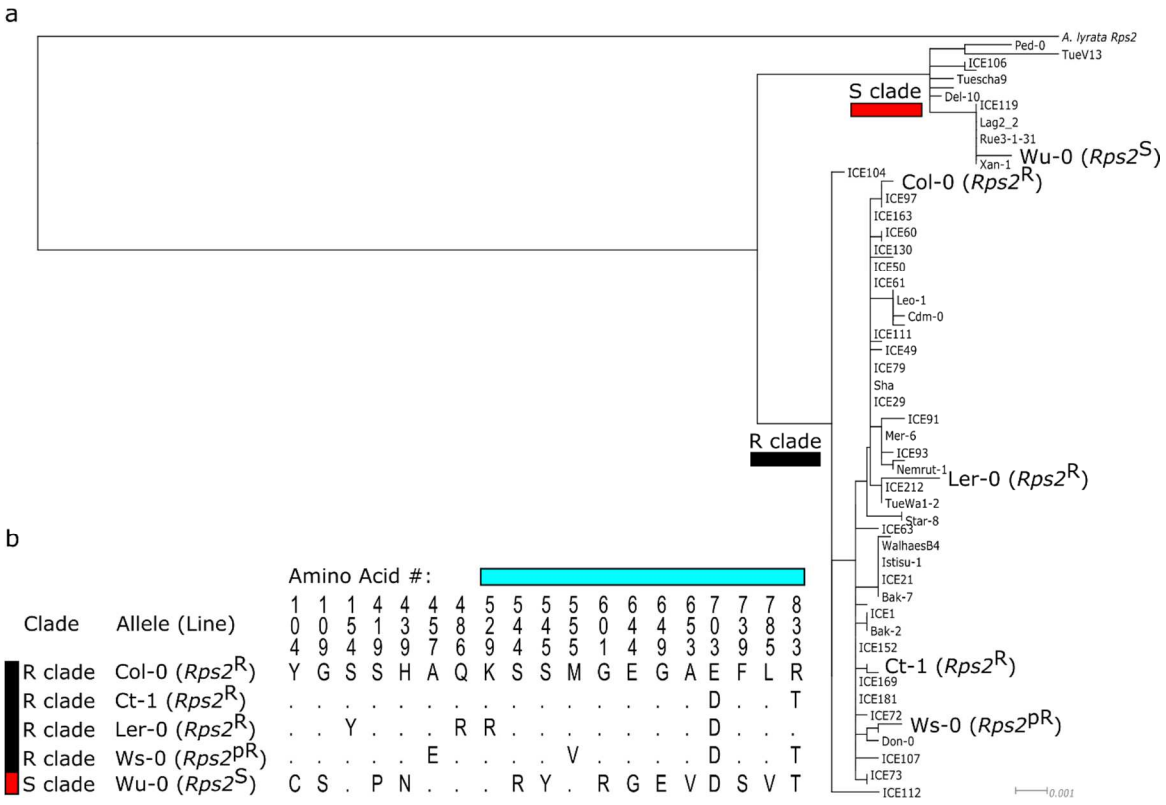


**Table 3.2.** Model coefficients for seven plant fitness proxies for insertion site one, with fitness fitted to resistance class, with allele nested into resistance class, and with date the plant was collected included. Class fitness is relative to the susceptible class of allele. Bolded values are significant after a Bonferroni correction, and stars indicate significance at the levels indicated at the base of the table.

| **Insertion Site 1** | **Number of Aerial Branches** | **Number of Basal Branches** | **Total Number of Siliques** | **Average Seeds per Silique** | **Total Seed Estimate** | **Total Collected Seed** | **Weight (mg)** |
|---|---|---|---|---|---|---|---|
| (Intercept) | -31.85 | -95.25$^{**}$ | -4387.8$^{**}$ | -710.86$^{**}$ | -198668$^{**}$ | -124576$^{**}$ | -4767.98$^{*}$ |
| | (37.40) | (36.30) | (1512.2) | (241.85) | (68910) | (46232) | (2044.04) |
| Class:pR | -0.67$^{*}$ | -0.13 | -17.51 | -1.89 | -866.34 | -215.41 | -23.74 |
| | (0.31) | (0.30) | (12.54) | (2.00) | (571.23) | (315.09) | (16.94) |
| Class:R | -1.08$^{*}$ | 0.22 | -37.12$^{*}$ | -1.52 | -1695.46$^{*}$ | -744.97 | -47.72$^{*}$ |
| | (0.42) | (0.41) | (17.11) | (2.74) | (779.58) | (440.91) | (23.12) |
| Date Collected | 0.06 | 0.16$^{**}$ | 7.22$^{**}$ | 1.19$^{**}$ | 325.39$^{**}$ | 203.30$^{**}$ | 7.86$^{*}$ |
| | (0.06) | (0.06) | (2.43) | (0.39) | (110.87) | (74.43) | (3.29) |
| Class:R (Ler-0) | 0.48 | -0.13 | 20.88 | 1.19 | 949.85 | 544.39 | 24.67 |
| | (0.44) | (0.42) | (17.67) | (2.83) | (805.24) | (458.57) | (23.89) |
| $R^2$ | 0.09 | 0.06 | 0.12 | 0.08 | 0.12 | 0.11 | 0.09 |
| Adj. $R^2$ | 0.06 | 0.03 | 0.09 | 0.05 | 0.09 | 0.07 | 0.06 |
| Num. obs. | 126 | 126 | 126 | 126 | 126 | 105 | 126 |
| RMSE | 1.31 | 1.27 | 52.96 | 8.47 | 2413.57 | 1243.81 | 71.59 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Statistical models

**Table 3.3.** Model coefficients for seven plant fitness proxies for insertion site two, with fitness fitted to resistance class, with allele nested into resistance class, and with date the plant was collected included. Class fitness is relative to the susceptible class of allele. Bolded values are significant after a Bonferroni correction, and stars indicate significance at the levels indicated at the base of the table.

| **Insertion Site 2** | **Number of Aerial Branches** | **Number of Basal Branches** | **Total Number of Siliques** | **Average Seeds per Silique** | **Total Seed Estimate** | **Total Collected Seed** | **Weight (mg)** |
|---|---|---|---|---|---|---|---|
| (Intercept) | 55.08$^{*}$ | -124.95$^{***}$ | -4155.4$^{**}$ | -229.1 | -177902$^{**}$ | -42224.6 | -4257.7$^{*}$ |
| | (23.76) | (27.70) | (1374.4) | (170.9) | (62543) | (35532.4) | (1822.2) |
| Class:pR | 0.26 | -0.04 | -2.37 | -0.43 | -99.30 | 23.81 | -3.94 |
| | (0.26) | (0.31) | (15.16) | (1.88) | (689.90) | (338.41) | (20.10) |
| Class:R | 0.14 | 0.57 | **43.12**$^{**}$ | -1.40 | 1178.53 | 459.47 | **53.24**$^{**}$ |
| | (0.25) | (0.29) | (14.37) | (1.79) | (653.78) | (323.56) | (19.05) |
| Date Collected | -0.08$^{*}$ | 0.20$^{***}$ | 6.81$^{**}$ | 0.42 | 290.65$^{**}$ | 70.08 | 6.99$^{*}$ |
| | (0.04) | (0.04) | (2.21) | (0.27) | (100.61) | (57.20) | (2.93) |
| Class:R (Ct-1) | 0.15 | 0.31 | -25.38$^{*}$ | 1.83 | -565.80 | -115.37 | -27.03 |
| | (0.22) | (0.26) | (12.81) | (1.59) | (582.76) | (298.86) | (16.98) |
| Class:R (Ler-0) | 0.05 | 0.06 | -12.03 | 4.89$^{**}$ | 240.95 | 290.83 | -11.10 |
| | (0.22) | (0.25) | (12.56) | (1.56) | (571.59) | (288.79) | (16.65) |
| $R^2$ | 0.03 | 0.17 | 0.12 | 0.06 | 0.08 | 0.06 | 0.10 |
| Adj. $R^2$ | 0.01 | 0.15 | 0.10 | 0.04 | 0.06 | 0.03 | 0.08 |
| Num. obs. | 211 | 210 | 211 | 211 | 211 | 170 | 211 |
| RMSE | 1.07 | 1.25 | 61.98 | 7.71 | 2820.49 | 1271.45 | 82.18 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Statistical models

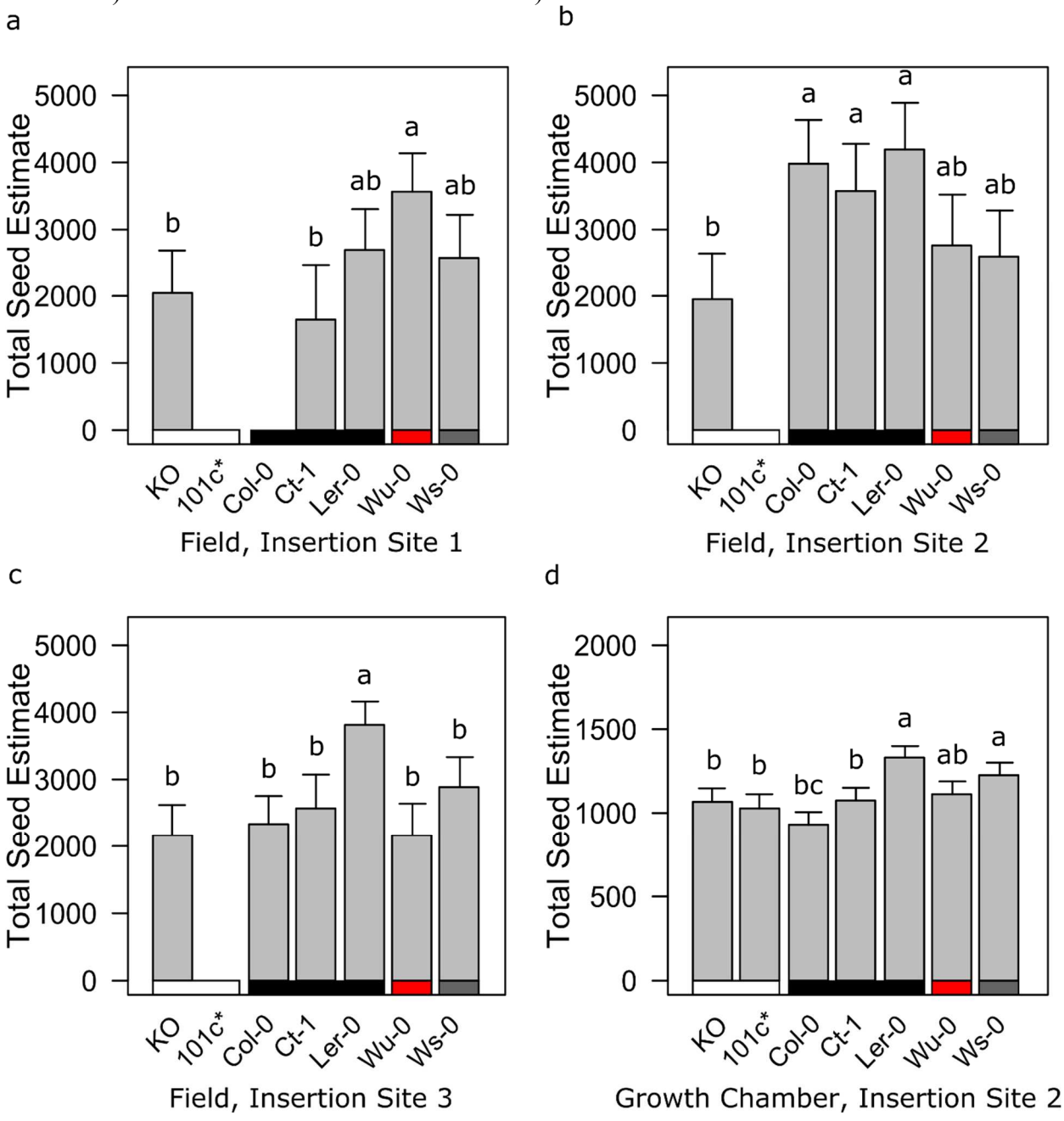
**Table 3.4.** Model coefficients for seven plant fitness proxies for insertion site three, with fitness fitted to resistance class, with allele nested into resistance class, and with date the plant was collected included. Class fitness is relative to the susceptible class of allele. Bolded values are significant after a Bonferroni correction, and stars indicate significance at the levels indicated at the base of the table.

| **Insertion Site 3** | **Number of Aerial Branches** | **Number of Basal Branches** | **Total Number of Siliques** | **Average Seeds per Silique** | **Total Seed Estimate** | **Total Collected Seed** | **Weight (mg)** |
|---|---|---|---|---|---|---|---|
| (Intercept) | -7.82 | -119.0$^{***}$ | -2959$^{**}$ | -99.5 | -114252$^{*}$ | -9340.9 | -4862$^{**}$ |
| | (22.83) | (27.32) | (1049.98) | (181.60) | (46472) | (27439.2) | (1803) |
| Class:pR | 0.05 | 0.62 | 12.82 | 5.12$^{*}$ | 812.03 | 545.08 | 30.71 |
| | (0.28) | (0.33) | (12.70) | (2.20) | (562.19) | (301.61) | (21.81) |
| Class:R | -0.43 | 0.31 | 0.82 | 4.37$^{*}$ | 268.16 | 201.36 | 6.37 |
| | (0.26) | (0.32) | (12.18) | (2.11) | (539.11) | (281.19) | (20.92) |
| Date Collected | 0.02 | 0.19$^{***}$ | 4.88$^{**}$ | 0.21 | 187.43$^{*}$ | 16.68 | 7.95$^{**}$ |
| | (0.04) | (0.04) | (1.69) | (0.29) | (74.82) | (44.19) | (2.90) |
| Class:R (Ct-1) | 0.74$^{**}$ | -0.29 | 6.10 | -2.15 | 127.56 | 8.21 | 5.15 |
| | (0.28) | (0.33) | (12.75) | (2.21) | (564.31) | (281.29) | (21.90) |
| Class:R (Ler-0) | 0.63$^{**}$ | 0.95$^{***}$ | 30.41$^{**}$ | 1.58 | 1234.92$^{**}$ | 612.01$^{**}$ | 49.75$^{**}$ |
| | (0.21) | (0.25) | (9.52) | (1.65) | (421.46) | (212.58) | (16.35) |
| $R^2$ | 0.07 | 0.29 | 0.16 | 0.07 | 0.14 | 0.10 | 0.15 |
| Adj. $R^2$ | 0.04 | 0.27 | 0.13 | 0.05 | 0.12 | 0.07 | 0.12 |
| Num. obs. | 174 | 174 | 174 | 174 | 174 | 148 | 174 |
| RMSE | 0.96 | 1.15 | 44.14 | 7.63 | 1953.69 | 953.75 | 75.81 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Statistical models

**Figure 3.4.** Significant fitness variation among lines in the allelic series in the absence of pathogen. KO is the *Rps2*-101c* mutant with an empty lox site at the insertion site for that line, or the $Rps2^{KO}$ line; 101c* is the *Rps2*-101c* mutant. Lines with *Rps2* inserted at three genomic locations, or insertion sites, were tested. Black bar is under $Rps2^{R}$ lines, grey bar is under $Rps2^{pR}$ lines, white bars are under $Rps2^{S}$ and $Rps2^{KO}$ lines. Within insertion site, alleles are grouped using Tukey's post-hoc test. a) Field fitness for insertion site one. b) Field fitness for insertion site two. c) Field fitness for insertion site three. d) Growth chamber fitness for insertion site two.



Expression levels can alter the penetrance of phenotypes (Raj *et al*., 2010), and overexpression of *Rps2* can lead to nonspecific activation of HR, or even lethality if expression levels are high enough (Mindrinos *et al*., 1994; Tao *et al*., 2000). Isoline expression measured in growth chambers varied within each genomic insertion site (Figure 3.5), and in many cases *Rps2* expression was two- to four-fold higher in the isolines than in the accessions from which the alleles derived (Figure 3.6). To investigate the relationship between fitness variation of $Rps2^R$ and $Rps2^S$ lines and *Rps2* expression, we modeled lifetime seed production in the field as a linear function of average *Rps2* expression at 23 days for each allele nested within allele class (Figure 3.7; Table 3.5). We also included as a covariate the date at which each plant was collected from the field. For both the $Rps2^R$ and $Rps2^S$ class, basal expression of *Rps2* was negatively correlated with lifetime fitness (Figure 3.7; F=9.91; df=10,789; p=0.004, 0.029). These results indicate that *Rps2* overexpression is costly in the absence of pathogens for both *R* and *S* clade alleles. Adding *Rps2* expression into the model did not reveal any differences in fitness proxies masked by differences in isoline expression between $Rps2^R$, $Rps2^{pR}$, and $Rps2^S$ lines (Table 3.5).

**Figure 3.5.** Average *Rps2* expression level at 23 days for all isogenic lines. Dotted line shows the average expression of *Rps2* detected for the knockout lines. Error bars are standard error about the mean.

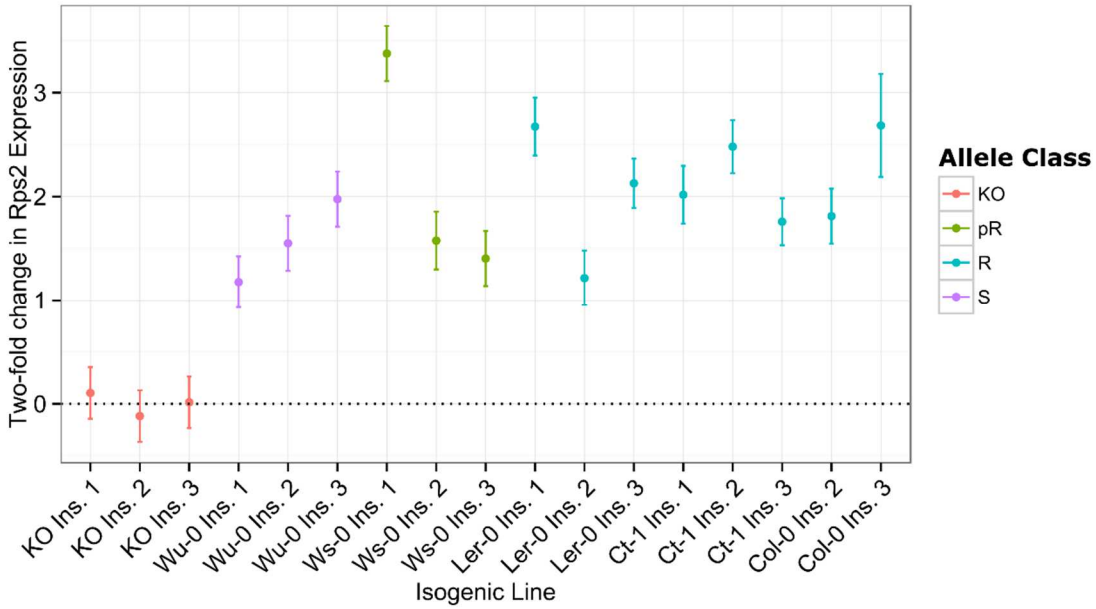



**Figure 3.6.** Average *Rps2* expression level at 21 days for a subset of isogenic lines and the native accessions from which each allele was derived. Expression of each allele in the isogenic *rps2*-101c mutant background should be compared to the expression of the accession from which that allele was derived. Dotted line shows the average expression of *Rps2* in samples where no reverse transcriptase was added; lines that overlap this expression would have *Rps2* expression that is non-detectable by qPCR. Error bars are standard error about the mean.

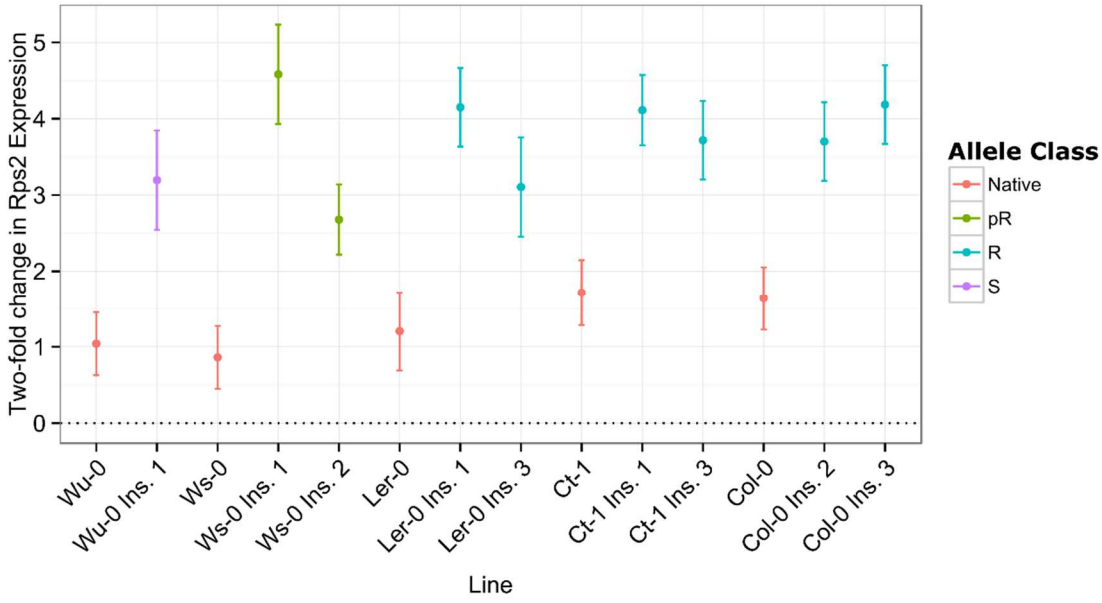


**Figure 3.7.** Average *Rps2* expression at three weeks is negatively correlated with fitness in the field. Black and red dashed lines are the regression lines for the *R* and *S* clades, respectively, for the relationship between fitness and expression nested in allelic class. Black points are $Rps2^{R}$ lines from the resistant clade of *Rps2*, and red points are $Rps2^{S}$ lines from the susceptible clade of *Rps2*. Native accessions express *Rps2* at levels to the left of the vertical dotted line; the average fitness of the *Rps2* knockouts is plotted at the horizontal dotted line.

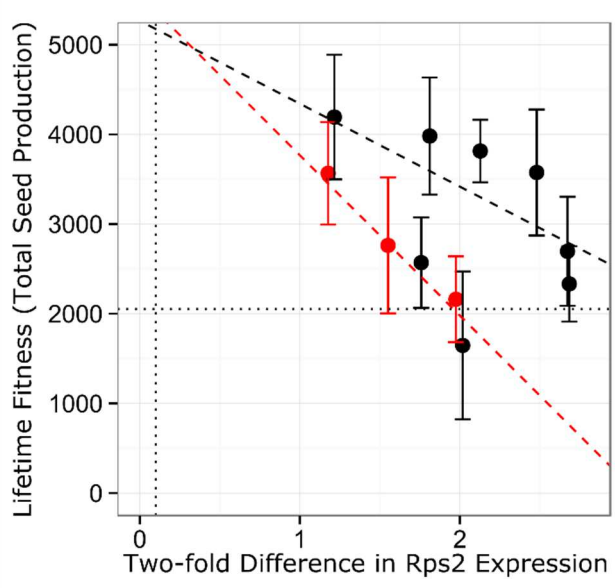


**Table 3.5.** Model coefficients for seven plant fitness proxies for all field data with fitness fitted to resistance class and with the expression of each isogenic line nested into resistance class. The date the plant was collected was also included in the model. Class fitness is relative to the susceptible class of allele. Bolded values are significant after a Bonferroni correction, and stars indicate significance at the levels indicated at the base of the table.

| | **Number of Aerial Branches** | **Number of Basal Branches** | **Total Number of Siliques** | **Average Seeds per Silique** | **Total Seed Estimate** | **Total Collected Seed** | **Weight (mg)** |
|---|---|---|---|---|---|---|---|
| (Intercept) | 3.64 | -125.4*** | -4107*** | -298.1** | -172331*** | -52997** | -5092*** |
| | (15.48) | (17.20) | (759) | (110.2) | (34085) | (20414) | (1084) |
| Class:pR | 0.09 | -1.19 | -57.06 | -2.76 | -2485.28 | -1211.51 | -78.72 |
| | (0.60) | (0.67) | (29.57) | (4.29) | (1327.75) | (690.31) | (42.23) |
| Class:R | -0.03 | -0.57 | 4.52 | -2.75 | -290.23 | -283.90 | -2.27 |
| | (0.60) | (0.67) | (29.43) | (4.27) | (1321.67) | (690.70) | (42.04) |
| Date Collected | -0.00 | 0.21*** | 6.83*** | 0.54** | 286.08*** | 89.65** | 8.47*** |
| | (0.02) | (0.03) | (1.22) | (0.18) | (54.76) | (32.83) | (1.74) |
| Class:S*_Rps2_ expression | -0.28 | -0.76 | -33.93 | -4.08 | -1661.90* | -853.05* | -51.47* |
| | (0.35) | (0.39) | (17.29) | (2.51) | (776.67) | (406.10) | (24.70) |
| Class:pR*_Rps2_ expression | -0.31* | 0.05 | 0.21 | -1.36 | -98.66 | -12.55 | -1.38 |
| | (0.13) | (0.15) | (6.54) | (0.95) | (293.84) | (149.17) | (9.35) |
| Class:R*_Rps2_ expression | -0.24 | -0.03 | -20.78** | -0.89 | -851.42** | -369.73* | -26.72** |
| | (0.14) | (0.15) | (6.66) | (0.97) | (299.09) | (156.34) | (9.51) |
| $R^2$ | 0.02 | 0.14 | 0.11 | 0.04 | 0.09 | 0.05 | 0.09 |
| Adj. $R^2$ | 0.01 | 0.13 | 0.10 | 0.02 | 0.08 | 0.04 | 0.08 |
| Num. obs. | 511 | 510 | 511 | 511 | 511 | 423 | 511 |
| RMSE | 1.12 | 1.24 | 54.86 | 7.97 | 2463.89 | 1172.30 | 78.36 |

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

Statistical models

### *3.4.2 An artificial* **Rps2** *knockout is significantly less fit in the absence of pathogen*

If $Rps2^R$ and $Rps2^S$ have equivalent fitness in the absence of pathogens, then the benefit of $Rps2^R$ resistance should have driven the $Rps2^R$ clade to fixation. However, both clades have been maintained for millions of years in *A. thaliana* (Mauricio *et al*., 2003). We hypothesized that $Rps2^S$, and perhaps $Rps2^R$, alleles must have another beneficial function to permit maintenance of the $Rps2^S$ clade. To explore this possibility, we compared the fitness of all lines with an expressed allele of *Rps2* (collectively, $Rps2^+$) to three artificial knockout lines ($Rps2^{KO}$), using the same field experiment in the absence of *avrRpt2*. In 19 out of 21 comparisons of fitness proxies contrasting *Rps2* lines, $Rps2^+$ isolines demonstrated higher performance than $Rps2^{KO}$ isolines, although after correction for multiple testing, only five of these instances were significant (Table 3.6-3.8). More importantly, $Rps2^{KO}$ individuals suffered up to a 54% decrease in seed set relative to $Rps2^+$ isolines (Figure 3.4; p=0.001), although these differences were only significant for the five lines with the lowest *Rps2* expression (Figure 3.7). These results suggest that any allele of *Rps2* is beneficial in the absence of known *Rps2*-mediated pathogen.

We considered two hypotheses to explain the observed benefit of *Rps2* alleles in the absence of *P. syringae* pv. *avrRpt2*. First, the presence of a different, and undetected, pathogen recognized by *Rps2* in the field could have compromised our ability to measure a cost of surveillance. Alternatively, a pleiotropic function for *Rps2* in the absence of disease may have contributed to this benefit. To discriminate between these hypotheses, we first repeated our fitness experiment for isolines from one insertion site in a growth chamber that mimicked the stressful environmental conditions of our field environment, but was known to be free of RPS2-recognized pathogens. As in the field experiment, after correction for multiple testing, there were

no significant differences in fitness proxies between *Rps2*$^{R}$ or *Rps2*$^{pR}$ lines relative to *Rps2*$^{S}$ lines (Table 3.9; F=9.20, df=9,1009, p>0.003) and *Rps2*$^{+}$ lines set significantly more seed than *Rps2*$^{KO}$ lines (Table 3.10; p=0.001). Thus, the growth chamber results recapitulated the results seen in the field (Figure 3.4d).

To further investigate if the fitness difference between *Rps2*$^{+}$ and *Rps2*$^{KO}$ lines was due to an interaction with an unknown microbe, we grew our isolines from one insertion site in sterile conditions on agar. Again, *Rps2*$^{+}$ plants had a higher weight than *Rps2*$^{KO}$ plants at 21 days ($p = 0.0024$). To test if the fitness difference between *Rps2*$^{+}$ and *Rps2*$^{KO}$ plants was due to some effect of our isogenic lines, we grew Col-0, a Col-0 isogenic line with the Col-0 allele at insertion site two, a Col-0 *Rps2*$^{KO}$ isogenic line with insertion site two, and an independently created Col-0 amiRNA knockdown of *Rps2* in sterile conditions on agar. There were not significant differences in weight between the two *Rps2* knockout lines, nor between the two lines with a Col-0 allele of *Rps2* ($p > 0.45$). In contrast, *Rps2*$^{+}$ plants had significantly higher weight than plants without *Rps2* ($p < 0.0005$). In combination, these results exclude the possibility that *Rps2* presence carried a fitness benefit due to recognition of pathogens, and instead suggest a pleiotropic function of *Rps2* in stressful abiotic environments in the absence of pathogen.

**Table 3.6.** Model coefficients for seven plant fitness proxies for insertion site one, with fitness fitted to resistance class, with fitness fitted to allele presence or absence and with allele nested within, and with date the plant was collected included. The fitness of *Rps2*$^+$ lines is measured relative to the knockout. Bolded values are significant after a Bonferroni correction, and stars indicate significance at the levels indicated at the base of the table.

| **Insertion Site 1** | **Number of Aerial Branches** | **Number of Basal Branches** | **Total Number of Siliques** | **Average Seeds per Silique** | **Total Seed Estimate** | **Total Collected Seed** | **Weight (mg)** |
|---|---|---|---|---|---|---|---|
| (Intercept) | -44.20 | -87.78$^{**}$ | -4038$^{**}$ | -635$^{**}$ | -179439$^{**}$ | -96578$^{*}$ | -4494$^{*}$ |
| | (33.42) | (30.91) | (1298) | (220) | (59047) | (37807) | (1733) |
| *Rps2*$^+$ | -0.08 | **0.93**$^{**}$ | 17.03 | 1.34 | 632.72 | 394.59 | 28.10 |
| | (0.33) | (0.31) | (12.93) | (2.19) | (588.15) | (324.64) | (17.26) |
| Date Collected | 0.08 | 0.14$^{**}$ | 6.60$^{**}$ | 1.07$^{**}$ | 292.21$^{**}$ | 157.23$^{*}$ | 7.33$^{**}$ |
| | (0.05) | (0.05) | (2.09) | (0.35) | (95.07) | (60.88) | (2.79) |
| *Rps2*$^+$ (Ct-1) | -0.48 | 0.13 | -21.07 | -1.23 | -959.99 | -555.58 | -24.83 |
| | (0.43) | (0.40) | (16.76) | (2.84) | (762.25) | (434.12) | (22.37) |
| *Rps2*$^+$ (Wu-0) | 0.59$^{*}$ | -0.08 | 16.48 | 0.38 | 758.25 | 222.52 | 23.25 |
| | (0.28) | (0.26) | (11.01) | (1.87) | (500.79) | (288.79) | (14.69) |
| *Rps2*$^+$ (Ws-0) | -0.07 | -0.21 | -1.26 | -1.56 | -120.60 | -7.59 | -0.68 |
| | (0.33) | (0.30) | (12.66) | (2.14) | (575.79) | (324.21) | (16.89) |
| $R^2$ | 0.08 | 0.12 | 0.13 | 0.07 | 0.13 | 0.10 | 0.13 |
| Adj. $R^2$ | 0.05 | 0.09 | 0.10 | 0.04 | 0.10 | 0.07 | 0.10 |
| Num. obs. | 152 | 152 | 152 | 152 | 152 | 130 | 152 |
| RMSE | 1.29 | 1.20 | 50.24 | 8.51 | 2285.10 | 1177.73 | 67.05 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Statistical models

**Table 3.7.** Model coefficients for seven plant fitness proxies for insertion site two, with fitness fitted to resistance class, with fitness fitted to allele presence or absence and with allele nested within, and with date the plant was collected included. The fitness of *Rps2*$^+$ lines is measured relative to the knockout. Bolded values are significant after a Bonferroni correction, and stars indicate significance at the levels indicated at the base of the table.

| **Insertion Site 2** | **Number of Aerial Branches** | **Number of Basal Branches** | **Total Number of Siliques** | **Average Seeds per Silique** | **Total Seed Estimate** | **Total Collected Seed** | **Weight (mg)** |
|---|---|---|---|---|---|---|---|
| (Intercept) | 48.11$^*$ | -126.0$^{***}$ | -4296.0$^{**}$ | -252.87 | -184747$^{**}$ | -45354 | -4466.5$^*$ |
| | (23.42) | (26.9) | (1313.7) | (167.66) | (59330) | (33318) | (1730.6) |
| *Rps2*$^+$ | 0.32 | 0.68$^*$ | **59.95**$^{***}$ | 0.90 | **2060.32**$^{**}$ | **899.73**$^{**}$ | **77.28**$^{***}$ |
| | (0.26) | (0.30) | (14.48) | (1.85) | (653.93) | (310.29) | (19.07) |
| Date Collected | -0.07 | 0.20$^{***}$ | 7.01$^{**}$ | 0.45 | 300.24$^{**}$ | 74.41 | 7.29$^{**}$ |
| | (0.04) | (0.04) | (2.11) | (0.27) | (95.41) | (53.58) | (2.78) |
| *Rps2*$^+$ (Ct-1) | 0.15 | 0.31 | -25.49$^*$ | 1.81 | -571.06 | -117.81 | -27.20 |
| | (0.22) | (0.25) | (12.38) | (1.58) | (559.30) | (284.95) | (16.31) |
| *Rps2*$^+$ (Ler-0) | 0.05 | 0.06 | -12.01 | 4.89$^{**}$ | 241.90 | 290.52 | -11.08 |
| | (0.22) | (0.25) | (12.15) | (1.55) | (548.64) | (275.40) | (16.00) |
| *Rps2*$^+$ (Wu-0) | -0.14 | -0.57$^*$ | -43.09$^{**}$ | 1.41 | -1177.18 | -459.67 | -53.20$^{**}$ |
| | (0.25) | (0.28) | (13.89) | (1.77) | (627.52) | (308.56) | (18.30) |
| *Rps2*$^+$ (Ws-0) | 0.13 | -0.61$^*$ | -45.41$^{***}$ | 0.98 | -1274.29$^*$ | -434.10 | -57.1$^{***}$ |
| | (0.21) | (0.25) | (11.92) | (1.52) | (538.49) | (267.18) | (15.7) |
| $R^2$ | 0.03 | 0.18 | 0.15 | 0.06 | 0.11 | 0.09 | 0.13 |
| Adj. $R^2$ | 0.01 | 0.15 | 0.13 | 0.04 | 0.09 | 0.06 | 0.11 |
| Num. obs. | 235 | 234 | 235 | 235 | 235 | 193 | 235 |
| RMSE | 1.07 | 1.23 | 59.94 | 7.65 | 2707.24 | 1212.51 | 78.97 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Statistical models

**Table 3.8.** Model coefficients for seven plant fitness proxies for insertion site two, with fitness fitted to resistance class, with fitness fitted to allele presence or absence and with allele nested within, and with date the plant was collected included. The fitness of *Rps2*$^+$ lines is measured relative to the knockout. Bolded values are significant after a Bonferroni correction, and stars indicate significance at the levels indicated at the base of the table.

| **Insertion Site 3** | **Number of Aerial Branches** | **Number of Basal Branches** | **Total Number of Siliques** | **Average Seeds per Silique** | **Total Seed Estimate** | **Total Collected Seed** | **Weight (mg)** |
|---|---|---|---|---|---|---|---|
| (Intercept) | 1.87 | -104.6$^{***}$ | -3154$^{***}$ | -106.03 | -119826$^{**}$ | -5418 | -4233$^{**}$ |
| | (20.54) | (22.1) | (905) | (149.96) | (39157) | (21902) | (1481) |
| *Rps2*$^+$ | -0.22 | 0.37 | 1.37 | 3.09 | 230.61 | 209.55 | 8.01 |
| | (0.22) | (0.24) | (9.84) | (1.63) | (425.50) | (207.77) | (16.09) |
| Date Collected | 0.00 | 0.17$^{***}$ | 5.19$^{***}$ | 0.22 | 196.48$^{**}$ | 10.34 | 6.93$^{**}$ |
| | (0.03) | (0.04) | (1.46) | (0.24) | (63.07) | (35.29) | (2.38) |
| *Rps2*$^+$ (Ct-1) | 0.75$^{**}$ | -0.28 | 5.92 | -2.16 | 122.30 | 10.04 | 5.74 |
| | (0.29) | (0.31) | (12.65) | (2.10) | (547.03) | (266.13) | (20.69) |
| *Rps2*$^+$ (Ler-0) | 0.65$^{**}$ | 0.98$^{***}$ | 30.00$^{**}$ | 1.56 | 1223.03$^{**}$ | 616.07$^{**}$ | 51.08$^{**}$ |
| | (0.21) | (0.23) | (9.39) | (1.56) | (406.11) | (200.66) | (15.36) |
| *Rps2*$^+$ (Wu-0) | 0.44 | -0.30 | -0.98 | -4.38$^{*}$ | -272.82 | -199.89 | -5.85 |
| | (0.27) | (0.29) | (12.08) | (2.00) | (522.66) | (266.06) | (19.77) |
| *Rps2*$^+$ (Ws-0) | 0.48 | 0.31 | 11.98 | 0.75 | 543.39 | 340.94 | 24.39 |
| | (0.26) | (0.28) | (11.37) | (1.88) | (491.83) | (247.22) | (18.60) |
| $R^2$ | 0.06 | 0.32 | 0.16 | 0.09 | 0.16 | 0.13 | 0.17 |
| Adj. $R^2$ | 0.03 | 0.30 | 0.14 | 0.07 | 0.14 | 0.10 | 0.14 |
| Num. obs. | 224 | 224 | 224 | 224 | 224 | 194 | 224 |
| RMSE | 0.99 | 1.07 | 43.82 | 7.26 | 1895.24 | 902.60 | 71.67 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Statistical models

**Table 3.9.** Model coefficients for two plant fitness proxies for insertion site two grown in the growth chamber, with fitness fitted to resistance class, with allele nested into resistance class, and with date the plant was collected included. Class fitness is relative to the susceptible class of allele. Bolded values are significant after a Bonferroni correction, and stars indicate significance at the levels indicated at the base of the table.

| | **Total Collected Seed** | **Weight (mg)** |
|---|---|---|
| (Intercept) | 687052.5*** | 34812.3*** |
| | (111999.5) | (8058.1) |
| Class:pR | 168.70* | 0.51 |
| | (82.24) | (5.91) |
| Class:R | -199.76* | 0.48 |
| | (83.23) | (5.98) |
| Date Processed | -0.03*** | -0.00*** |
| | (0.01) | (0.00) |
| Class:R (Ct-1) | 204.68* | -7.50 |
| | (83.19) | (5.98) |
| Class:R (Ler-0) | 293.98*** | 10.10 |
| | (79.48) | (5.70) |
| $R^2$ | 0.10 | 0.06 |
| Adj. $R^2$ | 0.10 | 0.06 |
| Num. obs. | 648 | 655 |
| RMSE | 641.72 | 46.30 |

***p < 0.001, **p < 0.01, *p < 0.05

Statistical models

**Table 3.10.** Model coefficients for two plant fitness proxies for insertion site two grown in the growth chamber, with fitness fitted to allele presence or absence and with allele nested within, and with date the plant was collected included. The fitness of *Rps2*$^{+}$ lines is measured relative to the knockout. Bolded values are significant after a Bonferroni correction, and stars indicate significance at the levels indicated at the base of the table.

| | **Total Collected Seed** | **Weight (mg)** |
|---|---|---|
| (Intercept) | 713562.58$^{***}$ | 26524.46$^{***}$ |
| | (87707.66) | (6311.26) |
| *Rps2*$^{+}$ | 324.43$^{***}$ | 10.34 |
| | (77.82) | (5.60) |
| Date Processed | -0.04$^{***}$ | -0.00$^{***}$ |
| | (0.00) | (0.00) |
| *Rps2*$^{KO}$ (101c*) | 107.64 | -10.30 |
| | (87.49) | (6.22) |
| *Rps2*$^{+}$ (Col-0) | -369.98$^{***}$ | 0.48 |
| | (79.72) | (5.74) |
| *Rps2*$^{+}$ (Ct-1) | -163.54$^{*}$ | -7.56 |
| | (80.25) | (5.78) |
| *Rps2*$^{+}$ (Ler-0) | -80.94 | 12.09$^{*}$ |
| | (77.17) | (5.55) |
| *Rps2*$^{+}$ (Wu-0) | -168.87$^{*}$ | -0.42 |
| | (80.42) | (5.79) |
| $R^2$ | 0.11 | 0.07 |
| Adj. $R^2$ | 0.10 | 0.06 |
| Num. obs. | 878 | 891 |
| RMSE | 627.59 | 45.36 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Statistical models

### 3.4.3 Rps2-*associated changes in defense response gene expression in the absence of pathogen*

To investigate novel functions of *Rps2* in the absence of pathogen, we determined the expression profile of two $Rps2^R$, one $Rps2^S$, and one $Rps2^{KO}$ line grown in sterile conditions. We first contrasted the expression profiles of an $Rps2^R$ and $Rps2^S$ line that shared the same insertion site and exhibited similar expression levels. The $Rps2^R$ line had 14 genes that were upregulated and two genes that were downregulated relative to the $Rps2^S$ line (Figure 3.8a). These genes were enriched for GO annotations of response to stress, particularly for response to water stress (Table 3.11; *p* value =1.24E-05). Thus, in the absence of pathogen, the difference between alleles with similar expression levels from different clades was slight (Figure 3.8b). We then contrasted all $Rps2^+$ lines with the $Rps2^{KO}$ line. $Rps2^+$ lines upregulated 538 genes relative to $Rps2^{KO}$ lines, and downregulated 312 genes. Genes upregulated in $Rps2^+$ plants were enriched for GO annotations involving photosynthesis and light response (Figure 3.9a; Table 3.12; *p* values = 2.47E-03, 3.91E-05), while genes downregulated in $Rps2^+$ plants were enriched for stimulus response, stress response, biotic stimulus response, and defense response annotations (Figure 3.9b,c; Appendix B.2; *p* values = 1.18E-16, 3.57E-16, 7.41E-11, 9.01E-11). We found substantial overlap between these genes and genes differentially expressed between single $Rps2^+$ lines and the $Rps2^{KO}$ line (Figure 3.9d). The genes found in each single-line comparison were consistently enriched for the same genes and GO annotations (Figure 3.9d; Table 3.13-3.14; Appendix B.3-B.6). Thus, removing *Rps2* from the system predominantly increases expression of genes that are induced in response to stress and pathogens. In the absence of pathogen, both clades of *Rps2* appear to function in the negative regulation of induced response genes.

**Figure 3.8.** The presence of *Rps2* alters the expression of more genes than the clade of *Rps2* allele. a) Heatmap and dendrogram of differentially expressed genes between an *S* and *R* line of *Rps2*. Genes are in rows, and biological replicates are in columns, with both dendrograms grouped by similarity of expression in the gene set displayed. KO is the *rps2*-101c* mutant with an empty lox site at the insertion site two, *R* clade and *S* clade are resistant and susceptible *Rps2* lines with similar levels of expression from insertion site two, and high *R* is the same allele of *Rps2* in the *R* clade comparison, from insertion site three, which has a higher level of *Rps2* expression. b) The overlap of differentially expressed genes significant in the *R* clade and *S* clade contrast, the *R* clade and KO contrast, and the *S* clade and KO contrast.

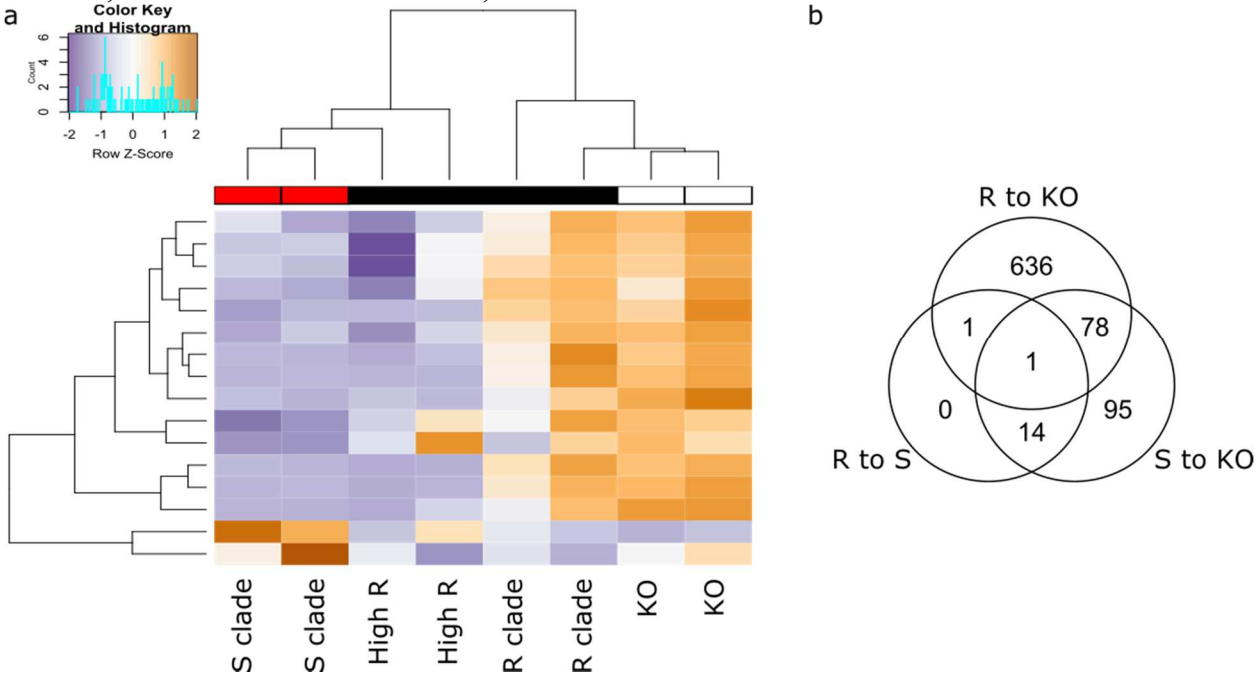


**Table 3.11.** Gene ontology enrichment of the 16 genes differentially expressed between *R* and *S* lines of *Rps2*. *R* clade and *S* clade are resistant and susceptible *Rps2* lines with similar levels of expression from insertion site two. GO annotations are from experimentally verified datasets and differentially expressed genes were compared to the entire annotated *Arabidopsis thaliana* genome. *p* values are Bonferroni corrected for multiple testing.

| GO annotation | # in Genome | # DE in *R* vs *S* | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| response to acid chemical | 547 | 7 | 0.29 | + | > 5 | 5.68E-06 |
| response to water deprivation | 148 | 5 | 0.08 | + | > 5 | 1.24E-05 |
| response to water | 149 | 5 | 0.08 | + | > 5 | 1.29E-05 |
| response to oxygen-containing compound | 734 | 7 | 0.39 | + | > 5 | 4.26E-05 |
| response to stress | 1234 | 8 | 0.65 | + | > 5 | 6.03E-05 |
| response to abscisic acid | 237 | 5 | 0.12 | + | > 5 | 1.28E-04 |
| response to alcohol | 269 | 5 | 0.14 | + | > 5 | 2.38E-04 |
| response to lipid | 317 | 5 | 0.17 | + | > 5 | 5.35E-04 |
| response to chemical | 1232 | 7 | 0.65 | + | > 5 | 1.42E-03 |
| response to inorganic substance | 489 | 5 | 0.26 | + | > 5 | 4.45E-03 |
| response to stimulus | 2216 | 8 | 1.16 | + | > 5 | 5.30E-03 |
| response to abiotic stimulus | 968 | 6 | 0.51 | + | > 5 | 6.56E-03 |
| response to hormone | 637 | 5 | 0.33 | + | > 5 | 1.60E-02 |
| response to osmotic stress | 310 | 4 | 0.16 | + | > 5 | 2.05E-02 |
| response to endogenous stimulus | 721 | 5 | 0.38 | + | > 5 | 2.90E-02 |

**Figure 3.9.** *Rps2* knockout lines differentially express stress response, defense response, and growth related genes relative to all lines with an allele of *Rps2*. (a-c) Heatmaps and dendrograms of gene sets as described below. Genes are in rows, and biological replicates are in columns, with both dendrograms grouped by similarity of expression in the gene set displayed. KO is the *Rps2*-101c* mutant with an empty lox site at the insertion site two, *R* clade and *S* clade are resistant and susceptible *Rps2* lines with similar levels of expression from insertion site two, and high *R* is the same allele of *Rps2* as in the *R* clade comparison, from insertion site three, which has a higher level of *Rps2* expression. a) Differentially expressed genes with GO annotations related to photosynthesis or response to light stimulus. b) Differentially expressed genes with GO annotations of response to stimulus or response to stress. c) Differentially expressed genes with GO annotations of defense response or response to biotic stimulus. d) The overlap of differentially expressed genes for three contrasts of lines with an *Rps2* allele with the knockout. Orange values are upregulated and blue are downregulated relative to the knockout.

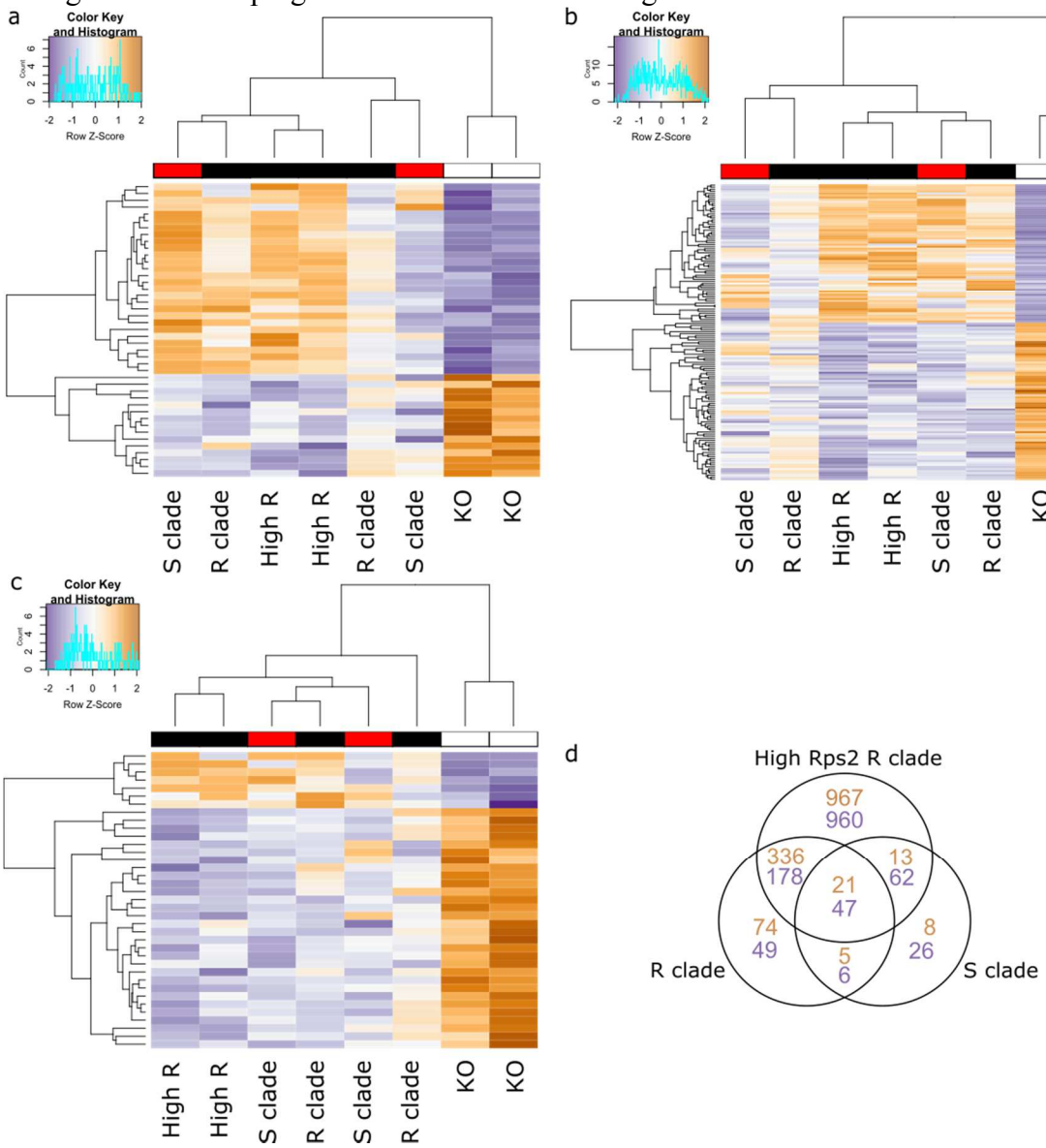


**Table 3.12.** Gene ontology enrichments of the 522 annotated genes upregulated in *Rps2*$^{+}$ lines compared to *Rps2*$^{KO}$ lines. *Rps2*$^{KO}$ is the *Rps2*-101c* mutant with an empty lox site at the insertion site two, and *Rps2*$^{+}$ included three lines: an *R* clade and *S* clade line, with resistant and susceptible *Rps2* lines with similar levels of expression from insertion site two, and a high *R* line, with the same allele of *Rps2* as in the *R* clade comparison, from insertion site three, which has a higher level of *Rps2* expression. GO annotations are from experimentally verified datasets and differentially expressed genes were compared to the entire annotated *Arabidopsis thaliana* genome. *p* values are Bonferroni corrected for multiple testing.

| **GO annotation** | **# in Genome** | **# DE in *Rps2*$^{+}$ vs *Rps2*$^{KO}$** | **# Expected** | **Over/ Under Expected** | **Fold Enrichment** | ***p* value** |
|---|---|---|---|---|---|---|
| response to light stimulus | 316 | 24 | 6.18 | + | 3.88 | 3.91E-05 |
| response to radiation | 326 | 24 | 6.38 | + | 3.76 | 6.91E-05 |
| photosynthesis | 41 | 8 | 0.8 | + | > 5 | 2.47E-03 |
| response to blue light | 36 | 7 | 0.7 | + | > 5 | 1.10E-02 |
| photosynthesis, light reaction | 25 | 6 | 0.49 | + | > 5 | 1.51E-02 |
| response to red or far red light | 95 | 10 | 1.86 | + | > 5 | 2.95E-02 |

**Table 3.13.** Gene ontology enrichments of the 415 annotated genes upregulated in *R* compared with KO lines. KO is the *Rps2*-101c* mutant with an empty lox site at the insertion site two, and *R* is a resistant *Rps2* line with a low level of *Rps2* expression. GO annotations are from experimentally verified datasets and the upregulated gene set was compared to the entire annotated *Arabidopsis thaliana* genome. *p* values are Bonferroni corrected for multiple testing.

| GO annotation | # in Genome | # DE in *R* vs KO | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| response to light stimulus | 316 | 30 | 4.91 | + | > 5 | 1.06E-11 |
| response to radiation | 326 | 30 | 5.07 | + | > 5 | 2.34E-11 |
| response to abiotic stimulus | 968 | 44 | 15.05 | + | 2.92 | 4.61E-07 |
| response to blue light | 36 | 9 | 0.56 | + | > 5 | 1.03E-05 |
| response to stimulus | 2216 | 70 | 34.46 | + | 2.03 | 1.60E-05 |
| cellular response to light stimulus | 50 | 9 | 0.78 | + | > 5 | 1.64E-04 |
| response to red light | 37 | 8 | 0.58 | + | > 5 | 2.08E-04 |
| cellular response to radiation | 52 | 9 | 0.81 | + | > 5 | 2.27E-04 |
| photosynthesis | 41 | 8 | 0.64 | + | > 5 | 4.49E-04 |
| response to red or far red light | 95 | 11 | 1.48 | + | > 5 | 5.32E-04 |
| biological process | 4190 | 104 | 65.16 | + | 1.6 | 7.34E-04 |
| response to far red light | 30 | 7 | 0.47 | + | > 5 | 7.51E-04 |
| photosynthesis, light reaction | 25 | 6 | 0.39 | + | > 5 | 4.12E-03 |
| cellular response to abiotic stimulus | 86 | 9 | 1.34 | + | > 5 | 1.32E-02 |
| chloroplast organization | 68 | 8 | 1.06 | + | > 5 | 1.79E-02 |
| generation of precursor metabolites and energy | 49 | 7 | 0.76 | + | > 5 | 1.81E-02 |

**Table 3.14.** Gene ontology enrichments of the 101 annotated genes upregulated in *S* compared with KO lines. KO is the *Rps2*-101c* mutant with an empty lox site at the insertion site two, and *S* is a susceptible *Rps2* line with a low level of *Rps2* expression. GO annotations are from experimentally verified datasets and the upregulated gene set was compared to the entire annotated *Arabidopsis thaliana* genome. *p* values are Bonferroni corrected for multiple testing.

| GO annotation | # in Genome | # DE in S vs KO | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| response to stimulus | 2216 | 22 | 8.39 | + | 2.62 | 2.96E-02 |
| response to light stimulus | 316 | 8 | 1.2 | + | > 5 | 3.63E-02 |
| response to radiation | 326 | 8 | 1.23 | + | > 5 | 4.51E-02 |

### 3.4.4 Rps2 *overexpression alters distinct defense gene expression from* Rps2 *knockouts*

Though both fitness and function of *R* and *S* clade alleles were indistinguishable when basal *Rps2* expression was controlled, this was no longer true when *Rps2* expression varied. In particular, fitness was significantly inversely correlated with the level of *Rps2* expression, suggesting a metabolic cost associated with overexpression. Indeed, lines with eight-fold higher than normal level of expression suffered a very high fitness reduction, on par with the cost of losing *Rps2* function entirely (Figure 3.7). This allelic series thus helps to delimit the range of *Rps2* expression tolerable to plants in natural conditions in the absence of pathogen. Expression data of natural accessions suggest that the $Rps2^{+}$ lines with lower basal *Rps2* expression and higher fitness had *Rps2* expression and *Rps2*-mediated phenotypes most similar to native accessions in the absence of pathogen (Figure 3.6). However, *Rps2* expression is induced during the typical course of infection (Axtell and Staskawicz 2003). The *Rps2* overexpression lines may thus elucidate plant cost of attack responses in the absence of other pathogen-mediated changes to the plant transcriptome.

To determine downstream effects of *Rps2* overexpression, we contrasted the expression profiles of two Col-0 *Rps2*$^{R}$ lines with high and low average *Rps2* expression in sterile conditions. Expression of 36 genes was upregulated in the high *Rps2* expression line relative to the low *Rps2* expression line and expression of 189 genes was downregulated (Figure 3.10a). Upregulated genes in the high *Rps2* expression line were enriched for genes involved in response to stimulus; particularly, for thalianol metabolic processes (Figure 3.10a; Table 3.15; *p* value = 1.05E-05). Downregulated genes in the high *Rps2* expression line were enriched for response to stimuli and stress; particularly, for response to biotic stimulus and for the defense response (Figure 3.10a, b; Appendix B.7; *p* values = 3.38E-10, 1.75E-08, 1.67E-03, 4.03E-04). Though these GO enrichment categories were the same as the contrasts between *Rps2*$^{+}$ and *Rps2*$^{KO}$ lines, the majority of differentially expressed genes in these two contrasts were different (Figure 3.10c). In addition, there was a distinct gene set differentially expressed in only the expression level contrast which was enriched for GO categories of response to stress and the defense response (Table 3.16; *p* values = 1.04E-02; 1.28E-02). *Rps2* overexpression thus downregulated an additional, different set of defense response genes than those induced in *Rps2* null mutants (Figure 3.10b, c).

**Figure 3.10.** *Rps2* expression level affects both stress response and defense response genes. (a,b) Heatmaps and dendrograms of gene sets as described below. Genes are in rows, and biological replicates are in columns, with both dendrograms grouped by similarity of expression in the gene set displayed. KO is the *Rps2*-101c* mutant with an empty lox site at the insertion site two, and R clade and S clade are resistant and susceptible *Rps2* lines with similar levels of expression from insertion site two. a) Differentially expressed genes with GO annotations of response to stimulus or response to stress. b) Differentially expressed genes with GO annotations of defense response or response to biotic stimulus. c) The proportional overlap of genes differentially expressed between low and high *Rps2* expression lines with genes differentially expressed between the low expression line and the knockout. Orange values are upregulated and blue are downregulated relative to the knockout.

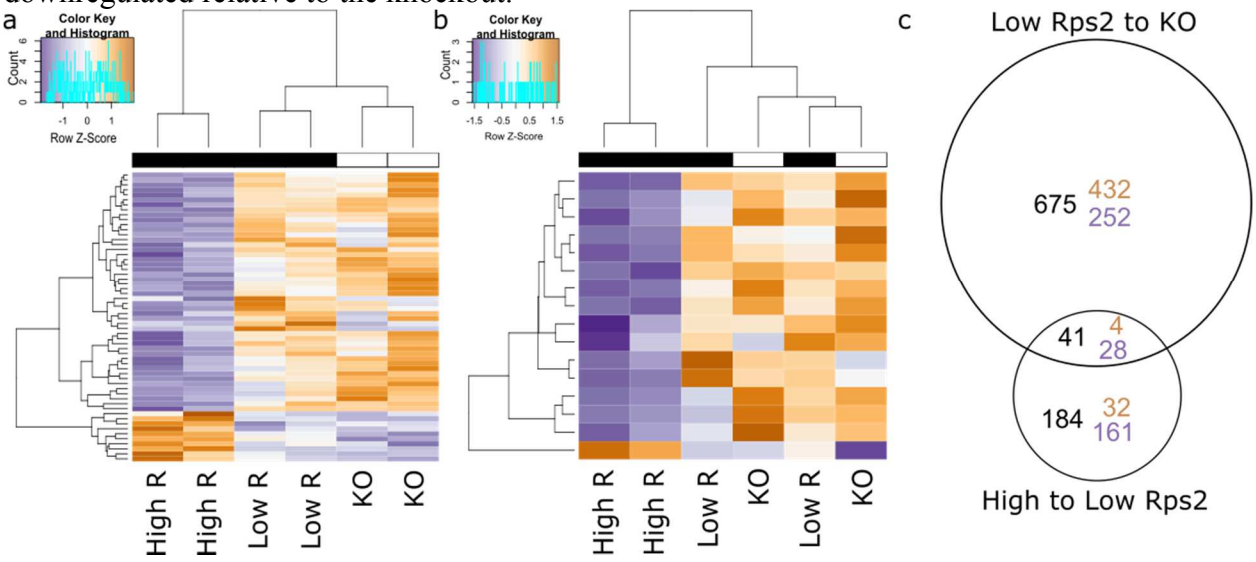


**Table 3.15.** Gene ontology enrichments of the 34 annotated genes upregulated in High R compared with Low R lines. High R is an *Rps2* line with a high level of *Rps2* expression, and Low R is an *Rps2* line with a low level of *Rps2* expression. GO annotations are from experimentally verified datasets and the upregulated gene set was compared to the entire annotated *Arabidopsis thaliana* genome. *p* values are Bonferroni corrected for multiple testing.

| GO annotation | # in Genome | # DE in High R vs Low R | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| thalianol metabolic process | 3 | 3 | 0 | + | > 5 | 1.05E-05 |
| tricyclic triterpenoid metabolic process | 4 | 3 | 0.01 | + | > 5 | 2.48E-05 |
| triterpenoid metabolic process | 10 | 3 | 0.01 | + | > 5 | 3.85E-04 |

**Table 3.16.** Gene ontology enrichments of the 70 annotated genes differentially expressed in the High R compared with Low R contrast, but not the Low R to KO contrast. High R is a *Rps2* line with a high level of *Rps2* expression, Low R is a *Rps2* line with a low level of *Rps2* expression, and KO is the *Rps2*-101c* mutant with an empty lox site at the insertion site two. GO annotations are from experimentally verified datasets and the upregulated gene set was compared to the entire annotated *Arabidopsis thaliana* genome. *p* values are Bonferroni corrected for multiple testing.

| GO annotation | # in Genome | # DE in High R vs Low R | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| response to stress | 3023 | 22 | 7.93 | + | 2.77 | 1.04E-02 |
| defense response | 1314 | 14 | 3.45 | + | 4.06 | 1.28E-02 |
| defense response to other organism | 882 | 11 | 2.31 | + | 4.75 | 3.38E-02 |

## 3.5 Discussion

While a natural null *S* allele precludes the possibility of the *S* allele being selectively favored for an alternative function, there is no such expectation when there are multiple functional alleles segregating at a locus. The fitness benefit of up to 40% of *Rps2* alleles relative

to *Rps2* knockouts in stressful conditions in the growth chamber and in the field provides a clear explanation for why *Rps2* has been found in every accession to date (Figure 3.4). $Rps2^{KO}$ lines fail to negatively regulate a number of genes involved in induced stress responses (Figure 3.9; Table 3.13; Appendix B.2); as induced genes are typically costly, the higher expression of induced genes in the knockout is a plausible mechanism for the fitness cost in $Rps2^{KO}$ lines. We expect selection against deletions when the fitness benefit of allele presence is high. The available work on the stoichiometric nature of RPS2-mediated resistance strongly suggests that the fitness advantage of RPS2 could be due to the important role of RPS2 in the NDR1-RIN4-RPS2 complex (Axtell and Staskawicz 2001, Axtell and Staskawicz 2002). Mechanistically, it is unknown whether S alleles of *Rps2* associate with RIN4 and NDR1 in the cell and how these components interact. For R alleles of RPS2, activation of disease resistance signaling mediated by RPS2 requires NDR1 (Axtell and Staskawicz 2002). RPS2 and NDR1 both associate with RIN4; RPS2 appears to detect a change in the levels of RIN4 relative to stable conditions in the cell. RIN4 knockouts are lethal unless RPS2 is also knocked out; conversely, addition of RIN4 into the system causes a time lag in the activation of RPS2-mediated defense responses (Day *et al*., 2006). Overexpression of NDR1 causes stunted growth and white lesions and hyper-activates RPS2-mediated resistance (Coppinger *et al*., 2004), probably by titrating away RIN4 from RPS2 (Day *et al*., 2006).

At the same time, the resistant clade of *Rps2* alleles has not swept to fixation in *A. thaliana* (Mauricio *et al*., 2003), suggesting persistent tradeoffs in the costs and benefits of resistant and susceptible alleles. Our field data indicated that, after controlling for the level of *Rps2* expression, $Rps2^{R}$ alleles were not significantly more or less fit than $Rps2^{S}$ alleles (Figure

3.7, Table 3.5). In fact, in the absence of pathogen, *Rps2*$^{R}$ and *Rps2*$^{S}$ alleles with similar levels of expression were not distinguishable: *Rps2*$^{S}$ lines suppressed the same defense response genes (Appendix B.3-B.5), and had a nearly indistinguishable expression profile from *Rps2*$^{R}$ (Figure 3.8a). Thus, we find no evidence that a cost of surveillance contributes to the maintenance of these two clades of alleles. Instead, the *S* clade may be maintained due to additional functionalities for *Rps2*$^{S}$ alleles in the presence of additional, unknown pathogens. Single *R*-loci can evolve to confer recognition of multiple Avirulence genes (Dangl and Jones, 2001): thirteen sequenced flax rust resistance alleles each have a different L locus specificity (Ellis *et al*., 1999), and alternative alleles of the *R*-gene *Rpp8* have been shown to resist a fungus and two distinct viruses (McDowell *et al*., 1998; Cooley *et al*., 2000; Takahashi *et al*., 2002).

One additional hypothesis warrants consideration. A fitness cost in the presence of pathogen in some environment may impose a fitness tradeoff sufficient to maintain a balanced polymorphism in an *R*-gene. In the presence of pathogen, *R* alleles of *Rps2* are known to suffer a net fitness cost when infected by *P. syringae* pv. *avrRpt2* in the absence of competition with conspecifics (Korves and Bergelson 2004). This fitness tradeoff in the presence of pathogen that differs by competitive environment is sufficient to maintain both clades of *Rps2* alleles for realistic parameters for competition and infection (Korves and Bergelson 2004), making a fitness tradeoff in the absence of pathogen unnecessary to explain the polymorphism at this locus.

In conclusion, our data support the idea that *R*-gene genetic architecture impacts the costs exhibited by resistant loci. Specifically, we fail to find the large surveillance costs of resistance that has been found for the indels, *Rpm1* and *Rps5,* when investigating the non-indel locus, *Rps2*. We suggest that *R*-genes with indel genetic architectures carry unusually high costs of resistance,

which may be mitigated by modulation of expression or by additional beneficial functions in *R*-genes with other genetic architectures.

# Chapter 4

## The genomic architecture of the *Rpp8* gene family drives high polymorphism at *Rpp8* and *At5g48620*

### 4.1 Abstract

Plant resistance (*R*-) genes provide some of the most extreme examples of polymorphism in the genome and are frequently under strong selection, particularly in the pathogen recognition domain encoded by the leucine-rich repeat region (LRR). Most *R*-genes exist as tandem arrays and many segregate for copy number polymorphism; studies of complex, duplicated *R*-gene loci lag far behind those of simple *R*-loci, as haplotype reconstruction of duplicated genes remains a challenge for next generation sequencing methodologies. Here we apply the Sanger method to sequence 50 *Rpp8* homologs from 28 accessions of *Arabidopsis thaliana* and 12 alleles of *A. lyrata*. With these data, we examined the evolution of a small *R*-gene family consisting of a tandem duplication with polymorphism for gene copy number, and a third homolog separated by approximately 2MB. *Rpp8* homologs exhibited unusually high levels of synonymous and nonsynonymous nucleotide diversity, 6.7-11.9 and 20-27 times that of the genomic background, and extreme levels of haplotype diversity, with 44 unique LRR haplotypes out of 50 homologs. I hypothesized that intergenic gene conversion (IGC) could be driving this increased polymorphism. I estimated IGC rates, and found higher rates between the physically distant homologs than the tandemly duplicated homologs. I demonstrated that in a selfing organism, IGC and conventional recombination between distant linked copies allows faster sharing of new mutations among *Rpp8* loci than within-copy recombination for values of IGC > $2\rho(1-s)$. I simulated variation in the recombination rate $\rho$ and the coefficient of selfing s on the minimum

IGC rate necessary to satisfy this inequality and showed that IGC rates between *Rpp8* homologs were well within the plausible range. The data were consistent with selection on *Rpp8* copy number and genetic architecture to generate and share novel genotypes. The results raise the possibility that *Rpp8* genomic architecture — one duplicated locus lying distant to a tandem duplication — may be maintained by diversifying selection on *Rpp8* haplotypic diversity.

## 4.2 Introduction

Resistance (*R*-) genes are the major component of the plant secondary immune system conferring resistance to viral, bacterial, and fungal pathogens. *R*-genes encode receptors that recognize Avirulence gene-dependent proteins, or changes to host proteins or products effected by these proteins (McDowell *et al*., 1998). Upon recognition, a defense response is activated which can lead to the death of the infected plant cell and inhibition of the spread of the pathogen (McDonald and Linde, 2002). This inducible defense response is mediated by specific gene-for-gene interactions between a plant *R*-gene and a pathogen Avirulence gene (*Avr*) (Flor, 1971).

The gene-for-gene model implies a matching *Avr* gene for every functional resistance gene (Van der Hoorn *et al*., 2002). The genome of *A. thaliana* contains 149 putative *R*-genes (Meyers *et al*., 2003), approximately 0.8% of genome content. Plant pathogens carrying *R*-gene-recognized *Avr* genes include species of bacteria, fungi, viruses, oomycetes, and parasitic nematodes (Scholthof *et al*., 2011; Dean *et al*., 2012; Mansfield *et al*., 2012; Jones *et al*., 2013; Kamoun *et al*., 2015). Each of these groups contains thousands of species and pathovars characterized as general or specific plant pathogens. In principle, ~150 recognition genes does not seem a large enough number to allow gene-for-gene recognition of the array of pathogens a plant might encounter.

Two major models have been proposed to explain this evolutionary puzzle. In the first, *R*-genes recognize conserved pathogen molecules or common pathogen targets. *R*-genes in this model should be ancient in origin; the low, stable levels of sequence polymorphism seen in ancient *R*-genes such as *Rpm1* and *Rps4* may reflect this case (Stahl *et al*., 1999; Bergelson *et al*., 2001a). In the second model, one *R*-locus may evolve to generate an allelic series that can confer recognition of multiple *Avr* genes (Dangl and Jones, 2001). The thirteen sequenced flax rust resistance alleles which each have a different L locus specificity (Ellis *et al*., 1999), and the three resistance specificities found in a sample of 24 *Rpp13* alleles (Allen *et al*., 2004), provide evidence that one *R*-locus can segregate for multiple functional alleles. Alleles of *Rpp8* confer resistance to a fungus, *Peronospora parasitica*, and two viruses, turnip crinkle virus and cucumber mosaic virus (McDowell *et al*., 1998; Cooley *et al*., 2000; Takahashi *et al*., 2002). *R*-genes in this model should be hypervariable and subject to positive selection on the solvent-exposed portions of the leucine-rich repeat (LRR) region of the protein, a binding surface often involved in protein recognition (Parniske *et al*., 1997; Bergelson *et al*., 2001a; Mondragon-Palomino *et al*., 2002).

The genomic and evolutionary mechanisms that generate high haplotypic and nucleotide diversity have not been well explored for *R*-genes (however see Allen *et al*., 2004; Rose *et al*., 2004). *R*-genes are often duplicated and found in tandem arrays, which are assumed to serve as reservoirs of variation for the generation of new *R*-gene specificities (Michelmore and Meyers 1998). These arrays could also retard the loss of conditionally advantageous alleles from the population. Interlocus gene conversion (IGC), or sequence replacement by a homologous sequence not mediated by a crossover event, is expected to increase both haplotypic and

nucleotide diversity within paralogous loci. Duplicate loci with sufficient IGC can have double the expected amount of nucleotide diversity ($\pi$) of single copy loci (Teshima and Innan 2012), owing to the doubling of the effective pool size of alleles. The conditions under which more than two copies will further increase nucleotide diversity have yet to be explored.

*Rpp8* homologs comprise a moderately simple gene family with a copy number polymorphism at the *Rpp8* locus (Figure 1; McDowell *et al*., 1998). Some ecotypes of *A. thaliana* have a tandem duplication of *Rpp8*, here designated D. We refer to the two duplicates as $D_1$ and $D_2$. A second common chromosomal haplotype carries only one copy of the gene, here designated S. $D_1$, $D_2$ and S segregate at the *Rpp8* locus. We present evidence indicating that S and $D_1$ can be considered the same gene, which we refer to as H1; we refer to the paralog $D_2$ equivalently as H2. A third paralog, *At5g48620*, henceforth called H3, is located 2.25 Mb proximal to *Rpp8* on the same chromosome (Kuang *et al*., 2008). H3 also exists as a copy number polymorphism with 0-2 copies (Kuang *et al.,* 2008). The most common haplotypes are the two-copy (*Rpp8(S), H3(S)*) or three-copy (*Rpp8($D_1$,$D_2$), H3(S)*) genotypes. There is strong evidence for IGC between members of *Rpp8* and *At5g48620* (Kuang *et al*., 2008).

Of particular interest is how mutations in the *Rpp8* gene family might move between $D_1$ or $D_2$ and S chromosomal haplotypes. By conventional intragenic crossing over or gene conversion, this can only occur in an outcrossed individual heterozygous for the two chromosomal haplotypes. Given the low outcrossing rate in *A. thaliana*, the rate of intragenic recombination may be insufficient to generate the observed levels of haplotype diversity and sharing of polymorphism. As an alternative, we explored a more circuitous route that a mutation may use to move between D and S haplotypes. This process involves three steps. First, IGC

moves the mutation to a paralogous locus, a process that occurs in chromosomal homozygotes, *i.e.,* selfed lineages. Then, a conventional crossover event in a heterozygote anywhere between the original and duplicated locus moves the mutation onto a new chromosomal background. A second IGC event then moves the mutation from the paralogous locus back to the original locus, albeit on a different chromosomal background. Though this possibility involves two additional steps compared with intralocus recombination within a heterozygote, each of these steps may be substantially more likely than the occurrence of this recombination event within a heterozygote, particularly in this highly inbred species.

Here, we described patterns of polymorphism in the *Rpp8* gene family consistent with the occurrence of IGC between all three homologs. We also explored possible mechanisms by which mutations spread across haplotypes in this gene family. To analyze the effect of IGC on sharing of mutations between copies of *Rpp8*, an analytical approximation was found for the rate of IGC required for the expected waiting time for the sharing of a mutation to be shorter through the second route than the first. To explore the effect of duplicate distance on haplotype diversity, we ignored the rarer IGC events with H2 and simulated IGC events between H1 and H3 only. This analysis showed that IGC between distant duplicates in a selfing species allows conventional crossing over to circulate mutations across chromosomal haplotypes faster than IGC alone for a large proportion of realistic IGC rates. Combined, these data suggest that *R*-gene copy number, as well as *R*-gene genomic architecture, is under selection for *R*-genes under diversifying selection to generate, share and maintain functional alleles.

## 4.3 Methods

Accurate insertion-deletion calling for regions larger than a few hundred base pairs remains a challenge for next generation sequencing technology (Cridland and Thornton 2010). Because of this, *Rpp8* homologs from 37 individuals from 28 accessions of *A. thaliana* and two individuals from one ecotype of *A. lyrata* were genotyped with PCR. A subset of 16 *A. thaliana* individuals and one *A. lyrata* individual were Sanger sequenced for the full gene and flanking regions, and the remainder were sequenced only for the most highly polymorphic 1038 bp LRR region.

### *4.3.1. Plant material*

Accessions were selected to include a world-wide set of representatives from across the geographic range of the species. Seeds were obtained from the collections of J. Bergelson and R. Mauricio, or were received from the Arabidopsis stock centers at Ohio State University and the University of Nottingham. *A. lyrata* seeds were obtained from the seed collections of D. Jacobson. Information on the accessions used in this study is provided in Table 4.1.

### *4.3.2 Genotyping*

Genotyping and sequencing were conducted by Dacheng Tian. PCR primers used to genotype and sequence *Rpp8* and *At5g48620* in *A. thaliana* and *A. lyrata* can be found in Table 4.2, and a schematic of all *Rpp8* homologs and all rounds of PCR can be found in Figure 4.1. Prior to PCR genotyping, individuals of each ecotype were grown in the greenhouse and leaf tissue was flash frozen in liquid nitrogen. DNA was extracted using a modified CTAB mini-prep protocol (Current Protocols in Molecular Biology, 1991) and DNA was purified by 9% PEG and 0.7M NaCl. For products 1kb in size or less, PCR was performed in 20 μL containing 20 ng of

**Table 4.1.** Genotypes and Phenotypes of Sampled Accessions. D represents the chromosomal haplotype carrying *Rpp8* tandem variants ($D_1$, $D_2$); S, one copy variants; and H3, the presence of *At5g48620*. * indicates an accession with the full coding sequence for at least one *Rpp8* homolog (Sequences of Ler-0, Col-0 and Di17 came from GeneBank); / designates no genotyping or phenotyping for this homolog or accession.

| Species | Accession | Origin | Genotype *Rpp8, At5g48620* | *Rpp8* Phenotype |
|---|---|---|---|---|
| *A. thaliana* | Ang-0 | Belgium | D, H3 | / |
| | Bla-2 | Spain | D, H3 | S |
| | Bur-0* | Ireland | D, H3 | R |
| | Ct-1 | Italy | D, H3 | R |
| | Cul-1 | England | D, H3 | S |
| | Cvi-0* | Cape Verdi Isl. | D, Δ | R |
| | Inv* | England | D, H3 | / |
| | Kz-1, 4, 7, 13 | Kazakhstan | D, H3 | / |
| | Ler-0* | Germany | D, H3 | R |
| | NFE3, 13 | England | D, H3 | / |
| | GR24* | N. Carolina, USA | D, H3 | R |
| | Pu-4, 5, 8, 16, 23 | Czechoslovakia | D, H3 | / |
| | Tamm-07 | Finland | D, H3 | / |
| | Wu-0* | Germany | D, H3 | S |
| | Zu-0* | Switzerland | D, H3 | S |
| | AB-27 | Indiana, USA | S, H3 | / |
| | Anh-3* | Germany | S, H3 | / |
| | Col-0* | Columbia, USA | S, H3 | S |
| | Di17* | France | S, / | R |
| | FM-15 | New York, USA | S, H3 | / |
| | HS-12 | Mass., USA | S, H3 | / |
| | Kas-1* | India | S, H3 | R |
| | Lip-0* | Poland | S, H3 | R |
| | Mt-0* | Libya | S, H3 | / |
| | NFC-5 | England | S, H3 | / |
| | Pog-0* | B. C., Canada | S, H3 | R |
| | RF-4* | Indiana, USA | S, H3 | S |
| | Tsu-0 | Japan | S, H3 | S |
| | UP-14 | Michigan, USA | S, H3 | / |
| *A. lyrata* | CE* (3 individuals) | Michigan, USA | D, H3 | / |
| | CH* (4 individuals) | Illinois, USA | / | / |

template DNA, 0.2 μM of each primer, 0.15mM dNTPs, 1 U TAQ polymerase, 1.2 μL of 25mM $MgCl_2$ and 2 μL 10x PCR buffer. Products were amplified in a MJ Research PTC-200

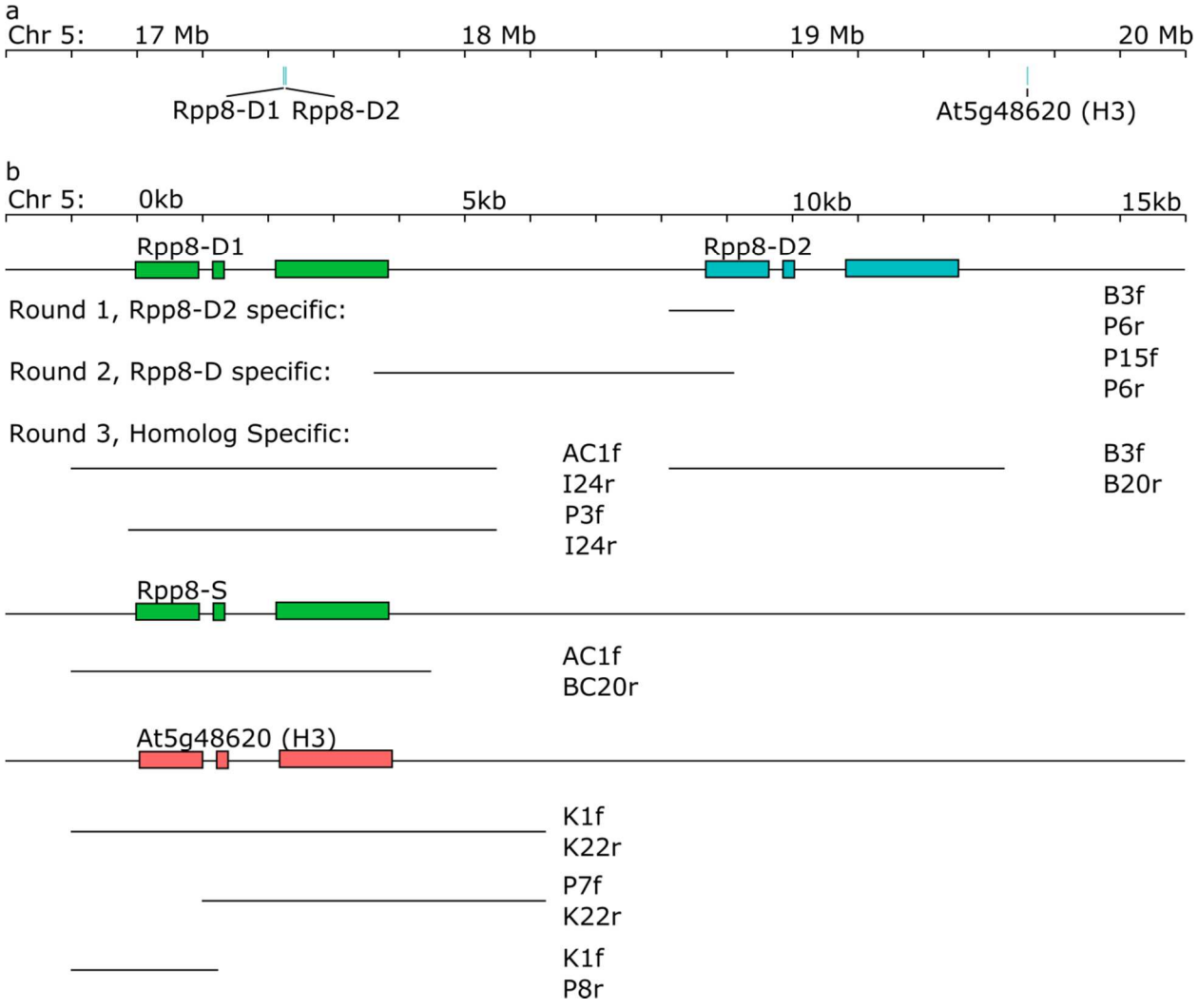
thermocycler using the following thermal profile: 94 °C for 180 s, 35 cycles of 94 °C for 30 s, 55 °C for 40s, and 72 °C for 60s, followed by 72 °C for 180s. For PCR products larger than 1kb, the Expand Long Template System (Roche) was used. The PCR reaction was identical except for containing 0.25mM dNTPs, 0.8 U enzyme and 2 μL 10x buffer 3 with 22.5 mM $MgCl_2$. The following thermal profile was used for long PCR products: 93 °C for 120 s, 10 cycles of 92 °C for 10s, 57 °C for 30s, and 68 °C for 120s, 20 cycles increasing each step at 68 for 10s, followed by 68 °C for 240s.

Three rounds of PCR were used to distinguish between the one-copy (S) and two-copy ($D_1$, $D_2$) variants of *Rpp8* (Table 4.2). In the first round, *RPP8*-D2 specific primers B3f and P6r

**Table 4.2**. Location and sequence of primers for locus of *Rpp8.* The locations and sequences of primers are based on Ler-0 sequence (accession no. AF089710) for gene D & AT5G48620, or based on Col (AF089711) in GeneBank; Sequence of Primer AC1 came from accession AB025638 in GeneBank.

| Primer | Location: from start codon of gene A B C | sequence (5' -3') |
|---|---|---|
| B3f | 6826 –447 | GGGAAGAAGATGCCTGGGAGTGA |
| AC1f | -1001 -982 | GATCAATGCAGCGAAGGTGTA |
| BC20r | 11798 4525 4570 | CACCAATCTGAACTGAAACCTAC |
| I24r | 5481 | AGTTTTAGTTTTGATGTATGTG |
| P3f | -44, 7229 -44 -44 | GTTCTTGTACTGGTTCATCGTAG |
| P5f | 452, 7715 443 452 | AGGGAGATCCGACAAACGTAT |
| P6r | 534, 7797 525 534 | TGAACATCATTCTCCACCAAA |
| P7f | 975, 8236 966 975 | CCCTAGCATGAGAAACACAAA |
| P8r | 1038, 8298 1026 1038 | CAGCATGTATCCCAACACCTT |
| P9f | 1327, 8593 1321 1327 | CTAAAAACGTATGGTAATCCA |
| P10r | 2150, 9415 2141 2150 | GATCCATCGTAAATCCCTTCT |
| P11f | 2524, 9800 2527 2530 | TTGCTCAGGGTGTTGGATCTT |
| P12r | 2641, 9917 2644 2647 | GTTCCGCATAGTAGAAGGTAG |
| P15f | 3473, 10748 3476 3479 | ACAAAGTCCAACACATTCCCG |
| P16r | 3648, 10924 3650 3655 | CTTCTTGGTCTTTCCTGCATC |
| P20r | 4441, 11687 4414 4459 | TGTTGTTACTAGAAGGCATGGTC |
| K1f | | |
| K22r | | |

**Figure 4.1.** Regions amplified with PCR to identify *Rpp8* homologs in *A. thaliana*. a) Positions of the coding sequence of *Rpp8* and *At5g48620* on chromosome 5. b) Homology of loci in the *Rpp8* gene family. Positions are shown in kilobases, relative to the start codon of the first homolog at that chromosomal location. Exons of *Rpp8* loci are shown as boxes; PCR fragments amplified at each locus are shown below the locus.




were used to generate a 971bp PCR product for D variants or no product for S variants. In the second round, a two-copy genotype was re-confirmed using primer pair P15f and P6r to generate a 5.5kb PCR product that encompassed parts of homologs $D_1$ and $D_2$ and the entire 4.3kb intergenic region. For one-copy genotypes, no product was produced. In the third round, the

presence (or absence) of the ($D_1$,$D_2$) or S genotype was confirmed with long PCR reactions using variant-specific primers, AC1f or P3f and I24r for $D_1$, B3f and B20r for $D_2$, and AC1f and BC20r for S, as given in Table 4.2 (Figure 4.2).

For half of the accessions, an 8 kb fragment of *At5g48620* was amplified using primers K1f and K22r (Figure 4.1, Table 4.2). For the remaining accessions, an overlapping two kb and six kb fragment of *At5g48620* were amplified by the primer pairs of P7f and K22r and K1f and P8r. Only one copy of *At5g48620* was sequenced for all accessions. The genotype of *A. lyrata* was determined by primer pair P15f and P6r and subsequent sequencing.

### *4.3.3 Sequencing*

To amplify homologs of the *Rpp8* family, multiple sets of primers were designed from the low polymorphic regions of the two sequenced copies of *Rpp8* in Ler-0 and one sequenced copy of *Rpp8* and *At5g48620* in Col-0 (Table S1; McDowell *et al*., 1998). To sequence all variants of *Rpp8* and *At5g48620* and their flanking regions, the long PCR products from genotyping were cut and extracted from gels to provide templates for short PCRs. The overlapping PCR products for each gene were sequenced directly using ABI cycle sequencing, Bigdye chemistry, and an ABI 377 automated sequencer. $D_1$ sequence in Ler-0 aligned with *Rpp8*-Ler sequence from McDowell *et al*., (1998), while $D_2$ sequence in Ler-0 aligned with RPH8A.

In *A. lyrata*, no primers amplified sequence from intergenic or flanking regions, which meant that individual homologs could not be distinguished with these primers. A long PCR product that spanned the $D_1$ and $D_2$ genes and contained the full intergenic sequence was cloned and partly sequenced to aid primer design. Of 16 primer sets then designed between primers in

adjacent ORFs and conserved primers in *Rpp8*, one pair gave a long PCR product, which was sequenced to obtain the 5' flanking region and the full $D_1$ coding sequence. The complete $D_2$ coding sequence and 1 kb 3' flanking regions were produced by anchored PCR.

### *4.3.4 Data Analysis*

Using the above generated data, I conducted all data analyses described herein. Homologs of *Rpp8* were aligned with Muscle and manually refined to minimize sequence mismatches. To obtain a general picture of the population genetics of this small gene family, $K_a$:$K_s$ ratios, synonymous and nonsynonymous nucleotide diversity in the coding region and framed LRR region, divergence from *A. lyrata*, and sliding window analyses of nucleotide diversity and Tajima's D were determined with DNAsp (Librado and Rozas 2009). Linkage disequilibrium (LD) values within and between loci were determined in R and plotted using the R package 'LDheatmap'. To infer sequence similarity in this extensively recombining gene family, I used maximum parsimony, a method that does not assume a shared history for the entire sequence in question, but rather gives an overall picture of similarity. Maximum parsimony trees of the coding sequence excluding the LRR and trees of the LRR alone were constructed to contrast the evolution of these portions of the *Rpp8* sequence.

To estimate IGC between the paralogous *Rpp8* loci, site frequency spectra between all possible pairs of loci were compared to theoretical expectations (Innan 2003a). A site frequency spectrum (SFS) describes the frequencies of two types of derived polymorphism segregating within a population: polymorphisms shared between paralogous loci, and polymorphisms specific to one paralog and not found in the other (Innan 2003a). To infer ancestral and derived polymorphisms for *Rpp8*, maximum parsimony trees of the coding sequence and the entire

Sanger sequenced region were constructed in Paup* using *Rpp8* alleles of *A. lyrata* as the outgroup (Swofford 2003). Site frequency spectra (SFS) of derived SNPs in all three homologs were then calculated in R. To ascertain the similarity between site frequency spectra and the theoretical expectations of Innan (2003a), 1000 SFS distributions of 10 samples were created by determining, for the 7-16 sequences available for each homolog, the expected frequencies for each SNP in a set of 10 samples, using random gamma distributions with a scale of 1 and a size parameter equivalent to each SNP's proportion times 10. Then, Kolmogorov-Smirnov tests between each resampled set and each of three theoretical distributions, where R=1, n=10, and c=0.2, 1, or 5, were conducted to obtain the distances between each resampled set and each theoretical distribution. Each resampled SFS was counted as closest to the expected distribution to which it had the minimum Kolmogorov-Smirnov distance. Thus, for each of the six pairs of SFS between the three members of the *Rpp8* gene family, there were 1000 resampled datasets binned as closest to one of three expected spectra, with gene conversion rates of 0.2, 1, and 5. A chi-squared test of these counts was conducted to determine significance. To determine the distribution of shared and specific derived polymorphisms across the sequenced region, site frequency spectra were additionally replotted as derived SNP frequencies against position across the sequenced region. This allowed us to distinguish between regions which had undergone distinct patterns of IGC.

## 4.4 Results

### *4.4.1 Variation in Rpp8 Genomic Architecture, Phenotypes, and Sequence*

Our sample of 28 *A. thaliana* ecotypes yielded 14 single copy variants (hereafter S variants) and 14 two copy variants (hereafter D variants $D_1$ and $D_2$) of *Rpp8* (Figure 4.2). Both

copy number variants (CNVs) were widely distributed throughout the species' range, though there were more D variants in Europe and more S variants in the US (Table 4.1). A CNV at *At5g48620* was also confirmed, with no copies of this gene detected in the Cvi-0 accession. The vast majority of accessions sequenced, 27/28 accessions, had at least one copy of *At5g48620* (hereafter *Rpp8* homolog three or H3). In *A. lyrata*, the D variant of *Rpp8* and one copy of H3 were detected. There was no correlation between *Rpp8* copy number and resistance to *P. parasitica* in *Rpp8*. Five of nine *Rpp8* D variants and three of six *Rpp8* S variants were resistant to *P. parasitica*. The Cvi-0 variant without *At5g48620* was also resistant to *P. parasitica*.

It was important to differentiate between the single copy and tandem duplicate variants at *Rpp8*, because the likelihood of recombination and gene conversion events in heterozygotes between S and $D_1$ or S and $D_2$ may differ based on the divergence between loci and by their probability of pairing during meiosis (Teshima and Innan 2004; Teshima and Innan 2012). I measured the fixed differences and frequencies of shared polymorphism for each *Rpp8* homolog to determine how *Rpp8* S and D sequences likely pair in meiosis. To contextualize these numbers, sequenced members of the *Rpp8* gene family had 470 total segregating sites and a range of 174-358 segregating sites within each paralog (Table 4.3). In all cases, the majority of segregating sites were found in the leucine-rich repeat region (Table 4.3). *Rpp8* S and $D_1$ alleles had no fixed differences, while S and $D_2$ alleles had 31 fixed differences (Table 4.4). S and $D_1$ alleles had a much higher number of shared derived polymorphisms than S and $D_2$ alleles (Table 4.5). Frequencies of shared polymorphisms were strongly correlated between S and $D_1$ alleles ($R^2$=0.67) but not between S and $D_2$ alleles ($R^2$=0.07; Table 4.6). I concluded that in the majority

of cases, S alleles pair with $D_1$ alleles during meiosis. I thus treated S and $D_1$ alleles as one orthologous gene, H1, in subsequent sections, and $D_2$ alleles as a paralogous gene (H2).

**Figure 4.2.** *Rpp8* homologs in *A. thaliana*. a) Positions of the coding sequence of *Rpp8* and *At5g48620* on chromosome 5. b) Homology of loci in the *Rpp8* gene family. Positions are shown in kilobases, relative to the start codon of the first homolog at that chromosomal location. Exons of *Rpp8* loci are shown as boxes; dashed line indicates deleted regions; orange line indicates region of *At5g48620* with no homology to *Rpp8*.

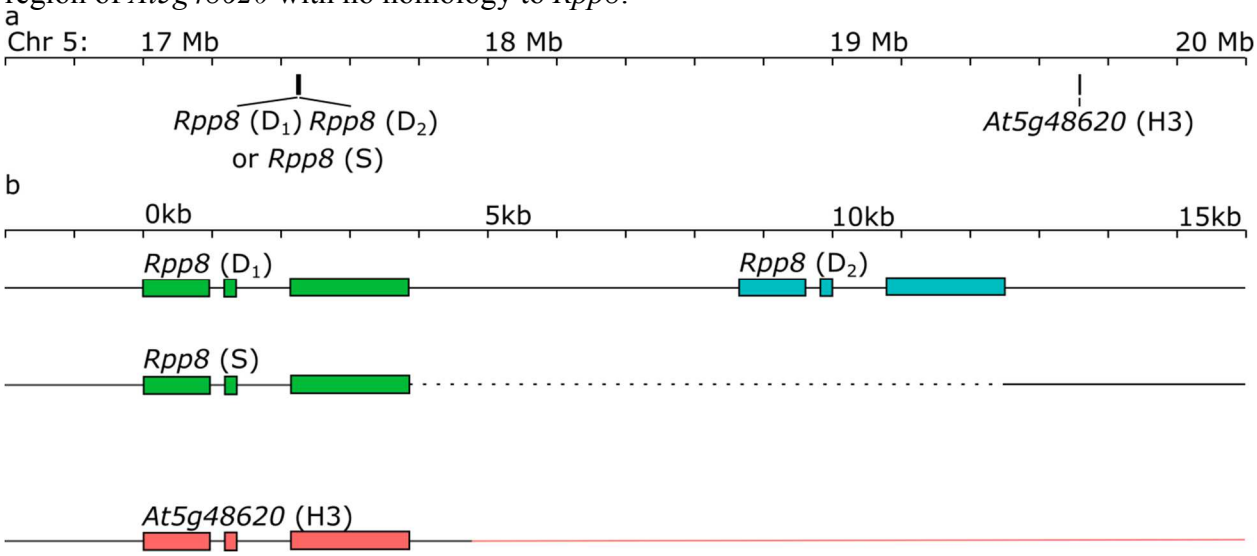



**Table 4.3.** Segregating sites, average number of different sites, and number of unique haplotypes found within the stated groups of loci of the *Rpp8* gene family. Values for both the coding sequence and leucine rich repeat (LRR) region are shown. Values were determined excluding gaps only in pairwise comparisons.

| | All coding sites (2766 bp) | | | LRR (888 bp) | | |
|---|---|---|---|---|---|---|
| | Segregating sites (S) | Average Number of Differences | Number of Unique Haplotypes | Segregating sites (S) | Average Number of Differences | Number of Unique Haplotypes |
| H1 | 358 | 108.0 | 15/15 | 254 | 61.2 | 20/21 |
| S | 276 | 109.3 | 8/8 | 153 | 54.2 | 11/11 |
| $D_1$ | 246 | 103.4 | 7/7 | 212 | 68.0 | 9/10 |
| $D_2$ | 174 | 80.5 | 7/7 | 162 | 51.8 | 17/19 |
| H3 | 302 | 110.8 | 9/9 | 179 | 63.7 | 9/10 |
| all | 470 | 118.0 | 31/31 | 327 | 60.4 | 45/50 |

**Table 4.4.** Fixed derived SNPs in the sequenced duplicated region at locus X, rows, not segregating at locus Y, columns. There are 470 segregating sites when all sequenced paralogs of the *Rpp8* gene family are included.

| X / Y | S | $D_1$ | $D_2$ | H3 |
|---|---|---|---|---|
| S | - | 0 | 27 | 0 |
| $D_1$ | 0 | - | 14 | 0 |
| $D_2$ | 4 | 0 | - | 3 |
| H3 | 4 | 0 | 27 | - |

**Table 4.5.** Number of shared derived polymorphisms in the sequenced duplicated region between locus X, rows, and Y, columns. There are 470 segregating sites when all sequenced paralogs of the *Rpp8* gene family are included.

| X / Y | S | $D_1$ | $D_2$ | H3 |
|---|---|---|---|---|
| S | - | 316 | 183 | 326 |
| $D_1$ | - | - | 161 | 307 |
| $D_2$ | - | - | - | 162 |
| H3 | - | - | - | - |

**Table 4.6.** $R^2$ values for linear models of shared polymorphism SNP frequencies in the sequenced duplicated region for comparisons between locus X, rows, and Y, columns. ** represents correlations significant at the <0.01 level; *** and bolded values represent correlations significant at the <0.001 level.

| X / Y | $D_1$ | $D_2$ | H3 |
|---|---|---|---|
| S | $R^2$ = **0.673***** | $R^2$ = **0.0699***** | $R^2$ = **0.506***** |
| $D_1$ | | $R^2$ = 0.0458** | $R^2$ = **0.473***** |
| $D_2$ | | | $R^2$ = 0.04313** |

Interestingly, there were no fixed differences between $D_1$ and H3 loci, and only four fixed differences between S and H3 loci (Table 4.4). In addition, there were a similar number of shared polymorphisms between S, $D_1$, and H3 alleles (Table 4.5). The frequencies of shared polymorphisms were also strongly correlated between S, $D_1$, and H3 alleles (Table 4.6). Thus there was sufficient sequence exchange between the distant duplicates H1 and H3, presumably through ectopic gene conversion, to maintain homogenization of alleles; this exchange could not be mediated by ectopic crossing over, as this would lead to the loss of sequence between *Rpp8* and *At5g48620* (~300 genes), which has never been observed.

### *4.4.2 Rates of IGC Between Rpp8 Homologs*

*Rpp8* and *At5g48620*, or H3, undergo IGC (Kuang *et al.,* 2008); however, with previously used methodology, it was difficult to detect all gene conversion events in order to estimate an IGC rate. Identification of gene conversion tracts is not feasible when the gene conversion rate is high (Mansai and Innan 2010), in part because only extremely long regions can be significant by a randomization test such as the one GENECONV employs when the gene conversion rate is high. However, higher rates of IGC increase the power to detect gene conversion for two alternative methods: the analysis of shared polymorphism using a SFS, and the analysis of the phylogeny for the entire region (Mansai and Innan 2010). In addition, previous work did not consider IGC between both D variants of *Rpp8*. For these reasons, we obtained evidence for IGC with three methods: LD, SFS, and phylogenetic inference, and estimated the rates of IGC using SFS.

IGC in a region reduces LD to an extent that depends on the rate of IGC (Hartasanchez *et al*., 2014). Recombination hotspots within a region undergoing IGC are also theorized to reduce LD, but only in the hotspot region (Hartasanchez *et al*., 2014). The pattern of LD did not vary substantially within the sequenced region for any homolog, and was thus not consistent with predictions of a recombination hotspot (Figure 4.3). However, $r^2$ values were significantly less than the genome-wide expectation for the entire length of the sequenced region of H1 (Figure 4.3c). The average $r^2$ between SNPs 1 kb apart in H1 was 0.078 (+/- 0.13), significantly less than the $r^2$ of 0.52 for this distance measured genome wide in *A. thaliana* (Kim *et al*., 2007). H2 also had reduced LD between SNPs 1kb apart, 0.232 (+/- 0.25), though LD at distances greater than 1kb was not significantly different than that of the genome on average (Figure 4.3c). The average

$r^2$ between SNPs 1kb apart in H3 was 0.167 (+/- 0.20). $r^2$ values of SNPs 2kb apart or less were significantly less than genome-wide expectations in H3 (Figure 4.3c). After the duplication breakpoint 3' of H3, $r^2$ values became more typical (Figure 4.3a, b), and $r^2$ values were not atypical between $D_1$ and $D_2$ variants of *Rpp8*, nor between *Rpp8* and *At5g48620*. Qualitatively, mild reductions in LD are theorized to result from IGC rates of ~1, and strong reductions in LD are due to IGC rates of >>1 (Hartasanchez *et al*., 2014). The observed pattern of LD was consistent with a low rate of IGC between the other two homologs and H2, and a high level of IGC between H1 and H3, IGC >> 1.

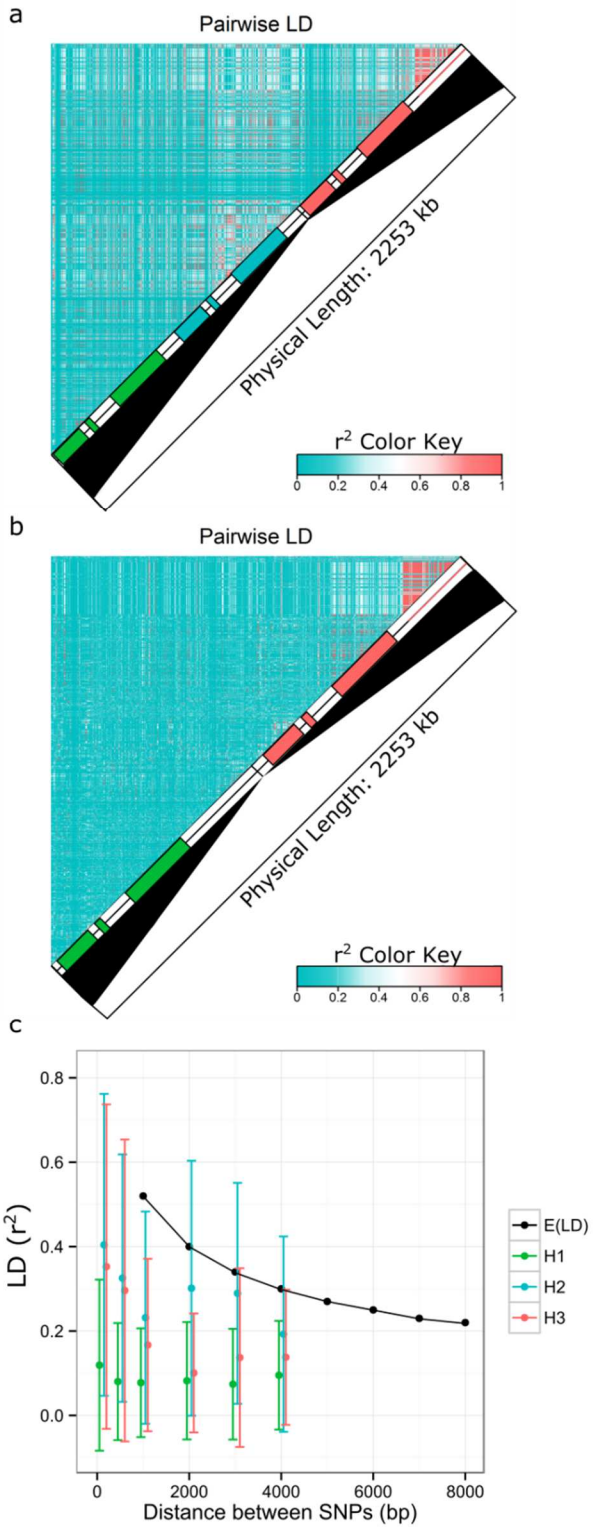



**Figure 4.3.** LD within and between the three members of the *Rpp8* gene family. Green, blue, and orange boxes represent positions of exons of H1, H2, and H3, respectively; red line indicates the 3' region of At5g48620 with no homology with other members of the *Rpp8* gene family. a) LD within and between $D_1$ and $D_2$ variants and H3. b) LD within and between H1 and H3 variants. c) Expectations for LD decay from Kim *et al*., (2007) versus observed decay in each member of the *Rpp8* gene family.

To compare rates of IGC for *Rpp8* homologs to theoretical expectations, we constructed site frequency spectra between each pair of *Rpp8* homologs, which resulted in two site frequency spectra per pair. These SFS showed the frequencies of SNPs in the first, "acceptor" homolog which were shared with the second, "donor" homolog or specific to the acceptor homolog (Figure 4.4). We contrasted these spectra with expected spectra for duplicates undergoing gene conversion at three rates: c=0.2, 1, and 5 (Innan 2003a). Both SFS with H1 as the donor homolog were most similar to the expected SFS where c=1 (95.5% and 98.7% of trials; $p < 2.2e-16$). Both SFS with H2 as the donor homolog were most similar to the expected SFS where c=0.2 (85.5% and 97.7% of trials; $p<2.2e-16$). The SFS with H3 as the donor and H2 as the acceptor were most similar to the expected SFS where c=1 (99.9% of trials; $p<2.2e-16$), while the SFS with H3 as the donor and H1 as the acceptor were most similar to the expected SFS where c=5 (88.1% of trials; $p<2.2e-16$). The data was consistent with a rate of IGC of 1 or higher between H1 and H3, and with a lower rate, of 0.2 to 1, between H2 and the other homologs.

To further explore the site frequency spectra for different regions of the *Rpp8* homolog sequence, we plotted frequencies of derived shared and private SNPs against position in the gene for all combinations of *Rpp8* homolog. The majority of the polymorphism shared between all three homologs was found in the third exon of *Rpp8*, which contains the LRR (Figure 4.5a-c). We also observed a large number of SNPs shared between two homologs that were not found in the third (Figure 4.5d-i). Differences in the numbers of shared polymorphisms may indicate the relative frequency of IGC events between homologs, just as a site frequency spectrum indicates the relative frequency of IGC events between two homologs. The majority of SNPs shared with two homologs but not with the third homolog were found in H1 and H3 but not H2 (Figure 4.5f-

**Figure 4.4.** Site Frequency Spectra between *Rpp8* paralogs compared to the most similar SFS expectation from Innan (2003a). Green, blue, and orange represent spectra of frequencies in H1, H2, and H3, respectively, while grey represents the expected spectra. a-b) Site Frequency Spectrum between H1 and H2 showing frequencies of derived alleles in H1 and H2, respectively. c-d) Site Frequency Spectrum between H1 and H3 showing frequencies of derived alleles in H1 and H3, respectively. e-f) Site Frequency Spectrum between H2 and H3 showing frequencies of derived alleles in H2 and H3 respectively.

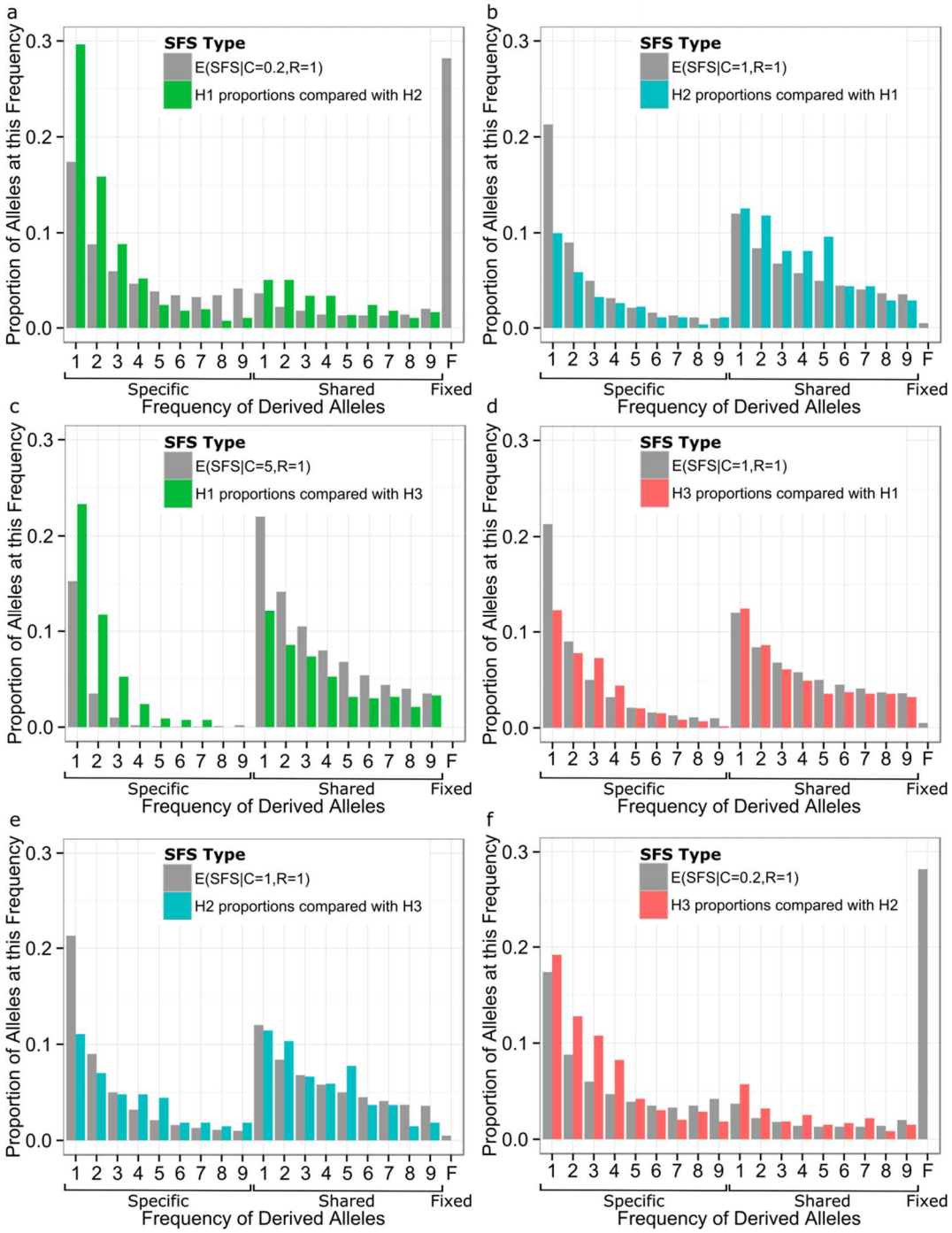
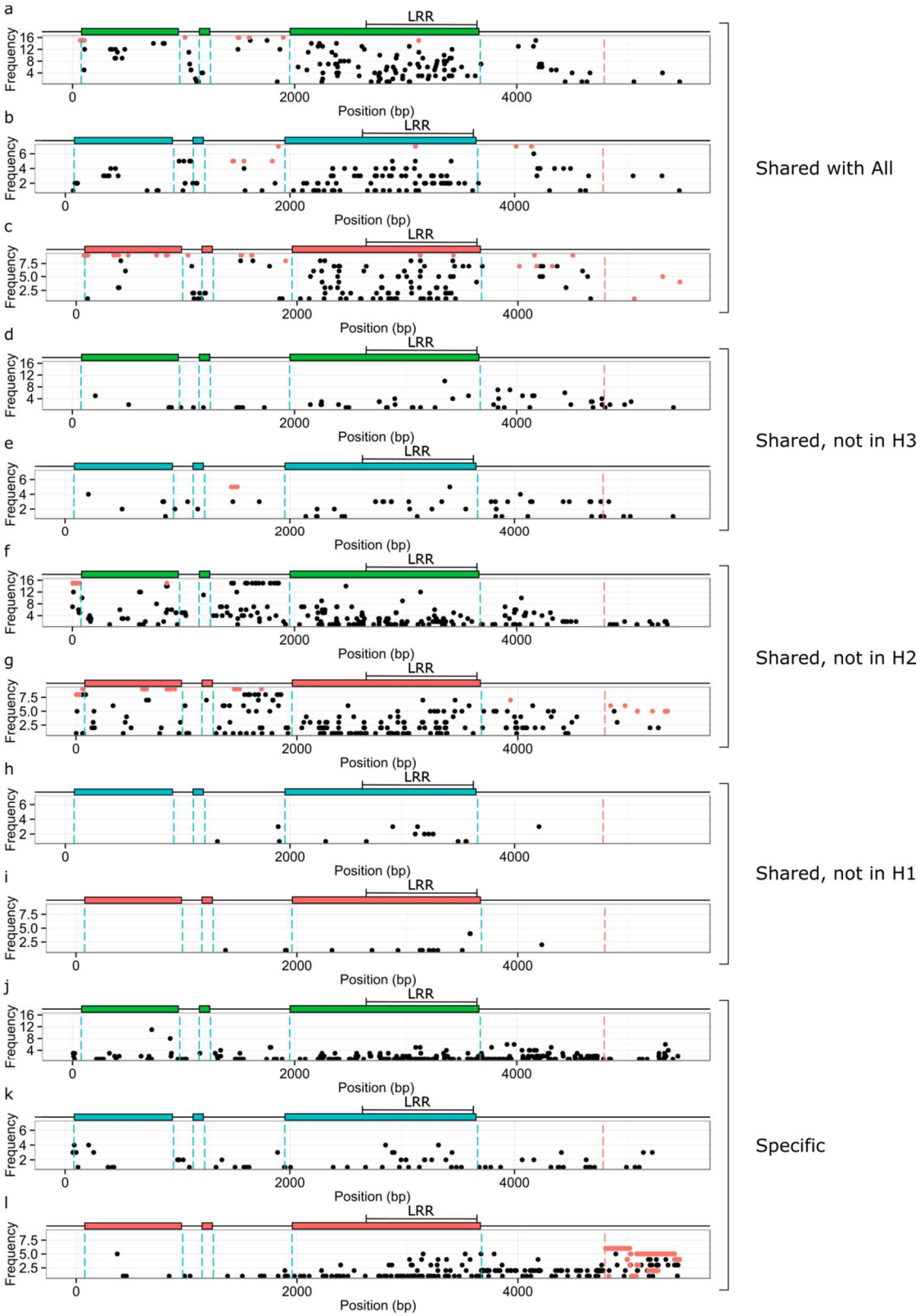


**Figure 4.5**. (see following page for caption)



*(from previous page)* **Figure 4.5.** Polymorphism Frequencies by Site, out of 470 segregating sites within the *Rpp8* gene family. Plots show proportions of derived SNPs shared with other homologs or specific to that homolog against the position of the SNP on the sequence. Green, blue, and orange boxes represent positions of exons of H1, H2, and H3, respectively. Dotted blue lines show the intron and exon boundaries. Dotted red line shows the downstream duplication boundary for AT5G48620. Red dots represent fixed derived alleles at that site. a-c) Shared polymorphisms found in all three members of the *Rpp8* gene family; frequencies of that SNP in H1, H2, and H3, respectively. d-i) Polymorphisms shared between two *Rpp8* paralogs that were not observed in the third. d-e) Polymorphisms shared between H1 and H2; frequencies in H1 and H2, respectively. f-g) Polymorphisms shared between H1 and H3; frequencies in H1 and H2, respectively. h-i) Polymorphisms shared between H2 and H3; frequencies in H1 and H2, respectively. j-l) Polymorphisms specific to each of the three *Rpp8* homologs; SNP frequencies in H1, H2, and H3, respectively.

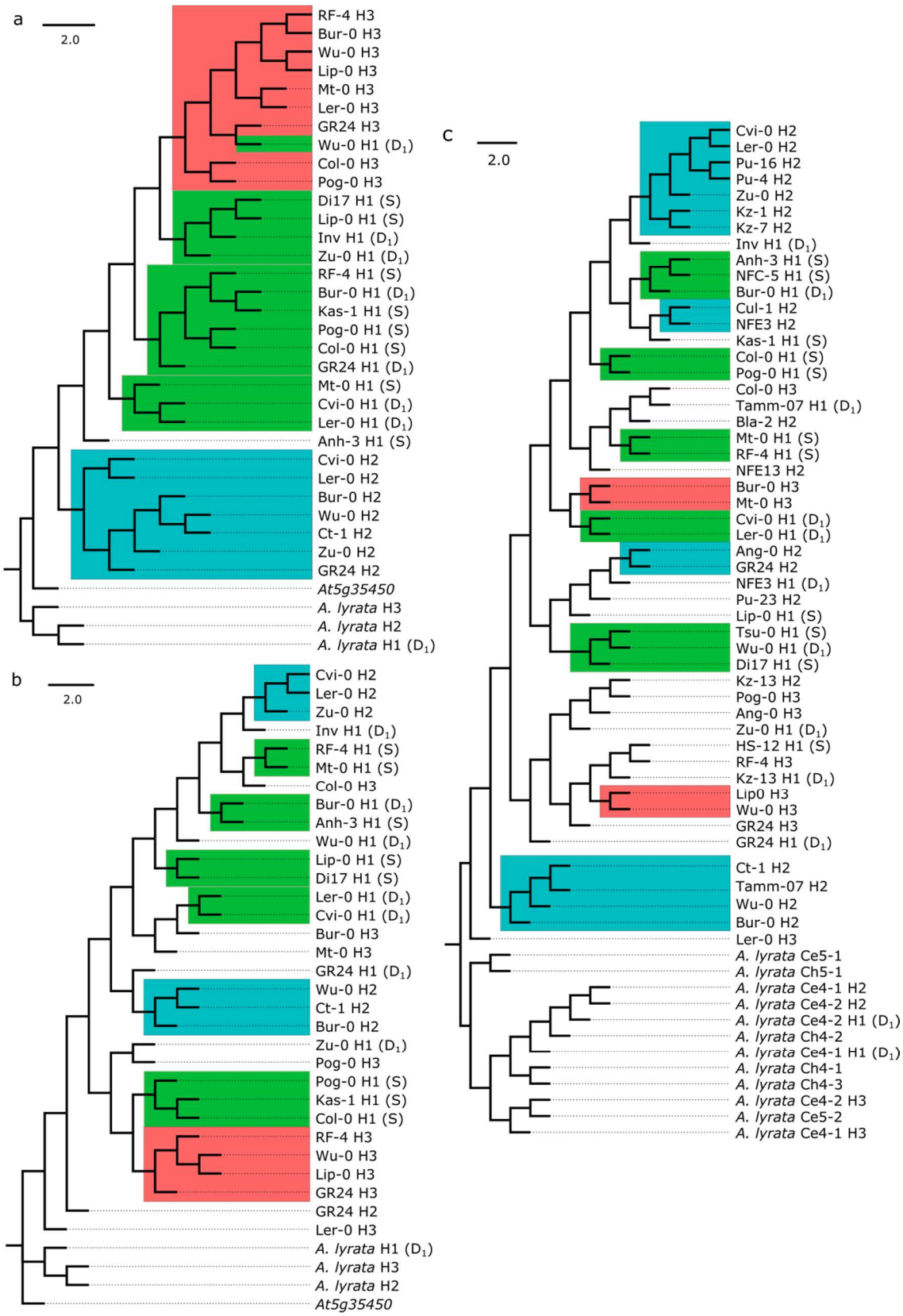
g). Very few SNPs were shared between H2 and H3 but not with H1, and these SNPs were found at low frequencies (Figure 4.5h-i). These data indicated that the most common route of IGC is between H1 and H3. A lower level of IGC was indicated between H1 and H2, and the lowest level of IGC was indicated to be between H2 and H3. Finally, private derived SNPs may indicate regions where IGC does not occur. No private polymorphisms were fixed in the duplicated region of *Rpp8* (Figure 4.5j-l). The majority of private polymorphisms were found at low frequencies, with all but 11 of these SNPs occurring at a frequency of 50% or below in the sampled sequences. These data were consistent with the occurrence of IGC across the entire duplicated region of *Rpp8*, as private polymorphisms at low frequencies have a low likelihood of being included in an IGC event.

We hypothesized that the LRR of *Rpp8* and *At5g48620* might be under diversifying selection to maintain novel haplotypes, while the remainder of the gene might be evolving more neutrally. There was enormous allelic diversity at *Rpp8*, particularly in the LRR region. All 31 fully sequenced alleles contained different haplotypes in the LRR region. To explore the potential for diversifying selection in the LRR, we constructed maximum parsimony trees of the

hypothesized LRR codons and the non-LRR codons from McDowell *et al*. (1998) (Figure 4.6). Comparing these trees revealed strikingly different evolutionary forces acting on different *Rpp8* homologs. The non-LRR portion of the sequence supported interlocus exchange between $D_1$ and S at *Rpp8*, but not sequence exchange between *Rpp8* and *At5g48620* or sequence exchange between $D_1$ and $D_2$ or S and D2 (Figure 4.6a). The $D_2$ alleles formed a monophyletic outgroup to all of the remaining *Rpp8* homologs (Figure 4.6a). Within the $D_1$ and S clade, the *At5g48620* alleles formed a nearly monophyletic clade, with the exception of one S allele. (Figure 4.6a) The non-LRR phylogeny supported recombination between S and $D_1$ alleles at *Rpp8*, not between S and $D_2$ alleles. This phylogeny was consistent with a duplication event from S or $D_1$ to create H3. As H3 shares flanking sequence with $D_1$, it was likely derived from this homolog before the speciation of *A. thaliana* and *A. lyrata*.

In contrast, the LRR portion of these sequences supported an entirely different phylogenetic history (Figure 4.6b). Here, S, $D_1$, H2, and H3 alleles were all distributed paraphyletically on the tree. The largest clade was of four H3 loci. Most H3 alleles had H1 alleles as their closest relative on the tree (Figure 4.6b). H2 alleles clustered into two smaller clades that had $D_1$ alleles as their closest outgroups (Figure 4.6b). This tree supported frequent gene conversion events between H3 and H1 alleles. It also supported gene conversion events between $D_1$ and H2 alleles.

**Figure 4.6.**


a
2.0
RF-4 H3
Bur-0 H3
Wu-0 H3
Lip-0 H3
Mt-0 H3
Ler-0 H3
GR24 H3
Wu-0 H1 (D1)
Col-0 H3
Pog-0 H3
Di17 H1 (S)
Lip-0 H1 (S)
Inv H1 (D1)
Zu-0 H1 (D1)
RF-4 H1 (S)
Bur-0 H1 (D1)
Kas-1 H1 (S)
Pog-0 H1 (S)
Col-0 H1 (S)
GR24 H1 (D1)
Mt-0 H1 (S)
Cvi-0 H1 (D1)
Ler-0 H1 (D1)
Anh-3 H1 (S)
Cvi-0 H2
Ler-0 H2
Bur-0 H2
Wu-0 H2
Ct-1 H2
Zu-0 H2
GR24 H2
At5g35450
A. lyrata H3
A. lyrata H2
A. lyrata H1 (D1)
b
2.0
Cvi-0 H2
Ler-0 H2
Zu-0 H2
Inv H1 (D1)
RF-4 H1 (S)
Mt-0 H1 (S)
Col-0 H3
Bur-0 H1 (D1)
Anh-3 H1 (S)
Wu-0 H1 (D1)
Lip-0 H1 (S)
Di17 H1 (S)
Ler-0 H1 (D1)
Cvi-0 H1 (D1)
Bur-0 H3
Mt-0 H3
GR24 H1 (D1)
Wu-0 H2
Ct-1 H2
Bur-0 H2
Zu-0 H1 (D1)
Pog-0 H3
Pog-0 H1 (S)
Kas-1 H1 (S)
Col-0 H1 (S)
RF-4 H3
Wu-0 H3
Lip-0 H3
GR24 H3
GR24 H2
Ler-0 H3
A. lyrata H1 (D1)
A. lyrata H3
A. lyrata H2
At5g35450
c
2.0
Cvi-0 H2
Ler-0 H2
Pu-16 H2
Pu-4 H2
Zu-0 H2
Kz-1 H2
Kz-7 H2
Inv H1 (D1)
Anh-3 H1 (S)
NFC-5 H1 (S)
Bur-0 H1 (D1)
Cul-1 H2
NFE3 H2
Kas-1 H1 (S)
Col-0 H1 (S)
Pog-0 H1 (S)
Col-0 H3
Tamm-07 H1 (D1)
Bla-2 H2
Mt-0 H1 (S)
RF-4 H1 (S)
NFE13 H2
Bur-0 H3
Mt-0 H3
Cvi-0 H1 (D1)
Ler-0 H1 (D1)
Ang-0 H2
GR24 H2
NFE3 H1 (D1)
Pu-23 H2
Lip-0 H1 (S)
Tsu-0 H1 (S)
Wu-0 H1 (D1)
Di17 H1 (S)
Kz-13 H2
Pog-0 H3
Ang-0 H3
Zu-0 H1 (D1)
HS-12 H1 (S)
RF-4 H3
Kz-13 H1 (D1)
Lip0 H3
Wu-0 H3
GR24 H3
GR24 H1 (D1)
Ct-1 H2
Tamm-07 H2
Wu-0 H2
Bur-0 H2
Ler-0 H3
A. lyrata Ce5-1
A. lyrata Ch5-1
A. lyrata Ce4-1 H2
A. lyrata Ce4-2 H2
A. lyrata Ce4-2 H1 (D1)
A. lyrata Ch4-2
A. lyrata Ce4-1 H1 (D1)
A. lyrata Ch4-1
A. lyrata Ch4-3
A. lyrata Ce4-2 H3
A. lyrata Ce5-2
A. lyrata Ce4-1 H3

*(from previous page)* **Figure 4.6.** Phylogenies of LRR and non-LRR regions for all three *Rpp8* homologs. Clades comprised of alleles from one locus are boxed. Green, blue, and orange boxes represent H1, H2, and H3, respectively. a) Phylogeny of non-LRR region. b) Phylogeny of the framed LRR region for the same accessions as in (a). c) Phylogeny of the 888bp LRR sequence obtained for 50 alleles of the *A. thaliana Rpp8* gene family and 12 alleles of the *A. lyrata Rpp8* gene family.

To further explore the allelic diversity and the phylogeny in the LRR region, we sequenced an 888bp region containing the last 12 LRRs of the 14 LRRs of *Rpp8* for 22 additional alleles in the *Rpp8* gene family. 45 of 50 alleles had unique haplotypes in the LRR region (Table 4.3). A phylogeny of these sequences also demonstrated high levels of paraphyly in alleles from different genomic loci (Figure 4.6c). H3 alleles had H1 alleles as their closest relatives, and H2 alleles had $D_1$, S, and H3 alleles as their closest relatives. In particular, alleles from the same populations did not necessarily fall into the same clades or closely related clades. Most Nfe, Pu, and Kz alleles, which were isolated from the same populations, were paraphyletic on the tree (Figure 4.6c). This result indicated that similar levels of diversity and a huge number of distinct alleles were maintained both between, and also within populations.

### *4.4.3 IGC and signatures of selection at Rpp8*

If *Rpp8* homologs are under positive selection, we expected to see a $K_a$:$K_s$ ratio higher than that expected on average for comparisons of genes between *A. lyrata* and *A. thaliana*. The distribution of $K_a$:$K_s$ ratios between genes found in both *A. lyrata* and *A. thaliana* is known (Hu *et al.,* 2011). Across the entire coding region, $K_a$:$K_s$ ratios within homologs varied between 0.53 and 0.61 (Table 4.7), higher than 91-94% of the known $K_a$:$K_s$ distribution. This enhancement was due to an excess of $K_a$ relative to the genome on average; $K_s$ values were not within the tail of the expected distribution (Table 4.7). Within the LRR, $K_a$:$K_s$ ratios varied between 0.75 and

0.97 (Figure 4.7), higher than 96-98% of the $K_a$:$K_s$ distribution, and indicated the action of positive selection to fix nonsynonymous mutations within the LRR.

If *Rpp8* is under balancing selection, we expected to see an excess of positive values of Tajima's D in the region. However, gene duplication with IGC is theorized to result in an underdispersed distribution of Tajima's D, with more than 95% of values within the region undergoing IGC falling between -1 and 1 in the neutral case, instead of between -2 and 2 in the single-copy gene case (Innan 2003). As predicted by this work, a sliding window analyses of Tajima's D found no regions with values above or below +/- 1.5 (Figure 4.8). However, in each homolog, 23% of windows fell above or below 1. For H2, 14 of 61 300bp windows, all in coding regions, and the majority in the LRR, had Tajima's D's above one, which may indicate balancing selection in a two-copy system (Innan 2003). For H1 and H3, 14 and 15 300bp windows, again mainly in the LRR, had a Tajima's D less than negative one, which may indicate positive selection in a two-copy system.

**Table 4.7.** Within-homolog synonymous and nonsynonymous nucleotide diversity and divergence for both the entire coding region and the leucine-rich repeat region. Average values for these variables across coding regions in the genome are also shown, and 95% confidence intervals, where available, are shown in parentheses.

| | Coding Region | | | | | | LRR | |
|---|---|---|---|---|---|---|---|---|
| | $\pi_s$ | $\pi_a$ | $\pi_a/\pi_s$ | $K_s$ | $K_a$ | $K_a/K_s$ | $\pi_a/\pi_s$ | $K_a/K_s$ |
| Genome Average | 0.005 (0.004-0.006) | 0.0014 | 0.23 | 0.13 (0.02,0.24) | 0.025 (0, 0.12) | 0.19 (0,0.07) | n/a | n/a |
| H1 | 0.0429 | 0.0355 | 0.829 | 0.143 | 0.0789 | 0.527 | 1.56 | 0.746 |
| H2 | 0.0341 | 0.0279 | 0.815 | 0.130 | 0.0823 | 0.612 | 2.27 | 0.965 |
| H3 | 0.0459 | 0.0383 | 0.830 | 0.144 | 0.0814 | 0.538 | 1.06 | 0.751 |

**Figure 4.7.** Sliding window analysis of within-species polymorphism and divergence between *A. thaliana* and *A. lyrata* in the *Rpp8* region. Vertical lines indicate boundaries of coding regions of *Rpp8*, as shown in the *Rpp8* schematic above each plot. Green, blue, and orange boxes above the plots represent positions of exons of H1, H2, and H3, respectively. The leucine rich repeat region (LRR) is also indicated. Orange and blue horizontal lines indicate average levels of Pi(a):Pi(s) and $K_a$:$K_s$ within *A. thaliana and between A. thaliana and A. lyrata*; grey line is the 95% right-hand tail for $K_a$:$K_s$. A) Homolog 1 at *Rpp8*. b) Homolog 2 at *Rpp8*. c) Homolog 3 at *At5g48620*.

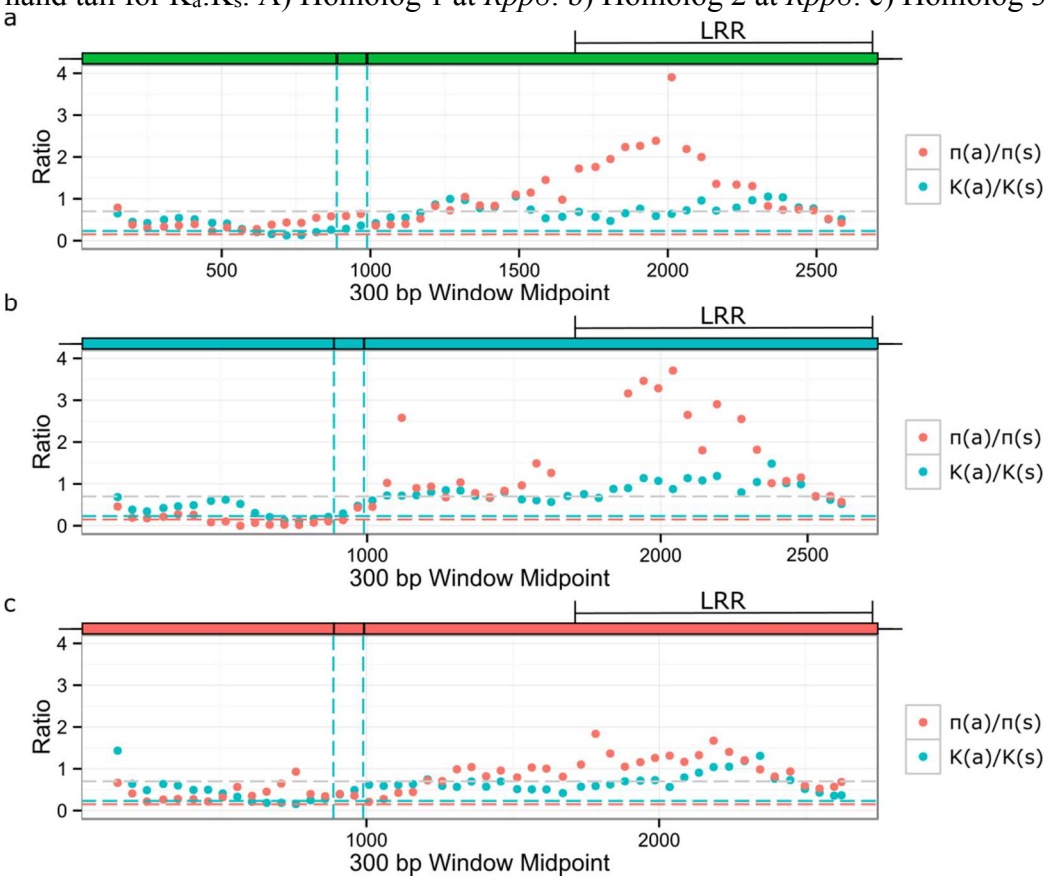
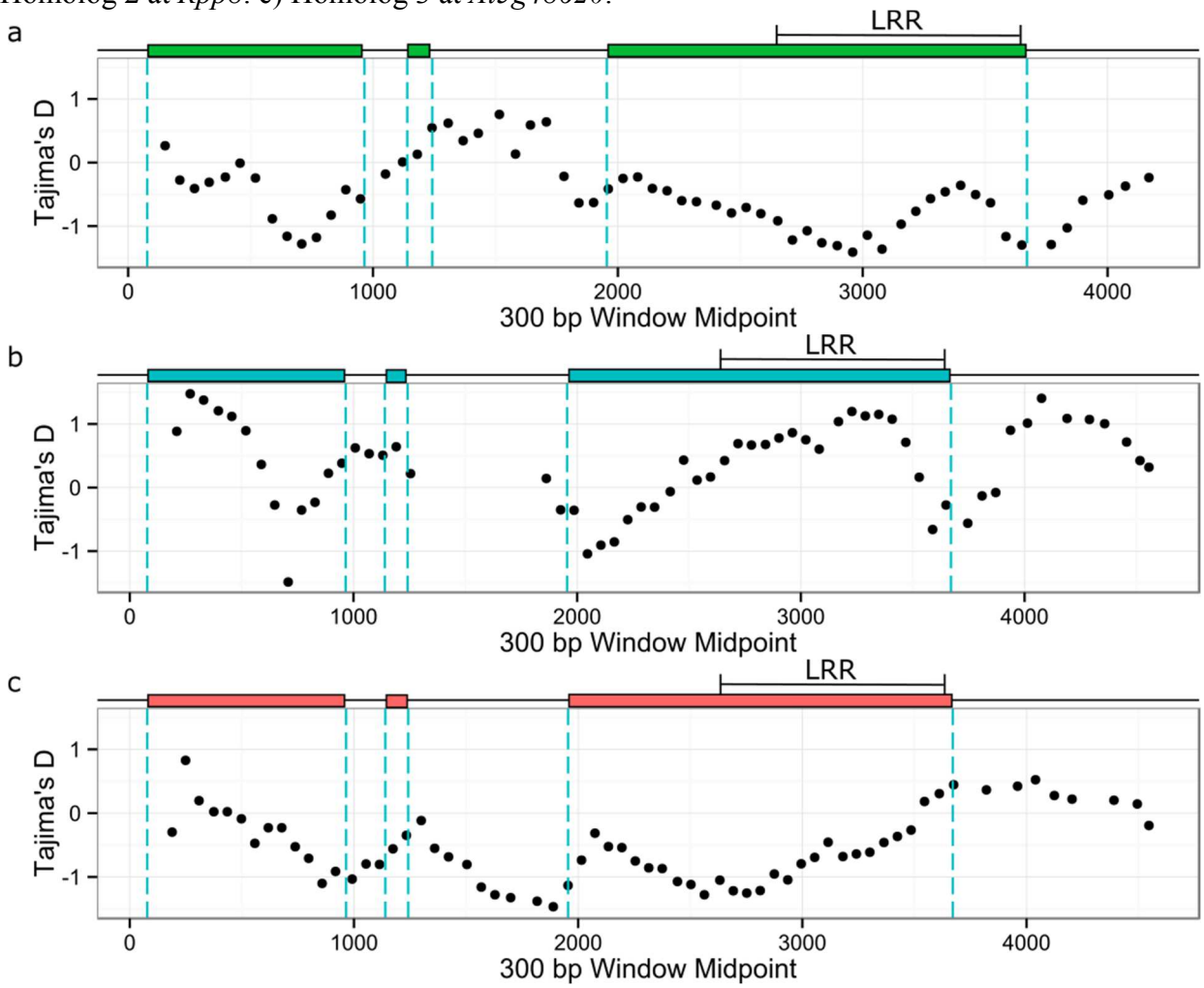


**Figure 4.8**. Sliding window analysis of Tajima's D across the sequenced regions. Vertical lines indicate boundaries of coding regions of *Rpp8*, as shown in the *Rpp8* schematic above each plot. Green, blue, and orange boxes above the plots represent positions of exons of H1, H2, and H3, respectively. The leucine rich repeat region (LRR) is also indicated. a) Homolog 1 at *Rpp8*. b) Homolog 2 at *Rpp8*. c) Homolog 3 at *At5g48620*.




### *4.4.4 IGC Generates Excess $\pi_s$ at Rpp8*

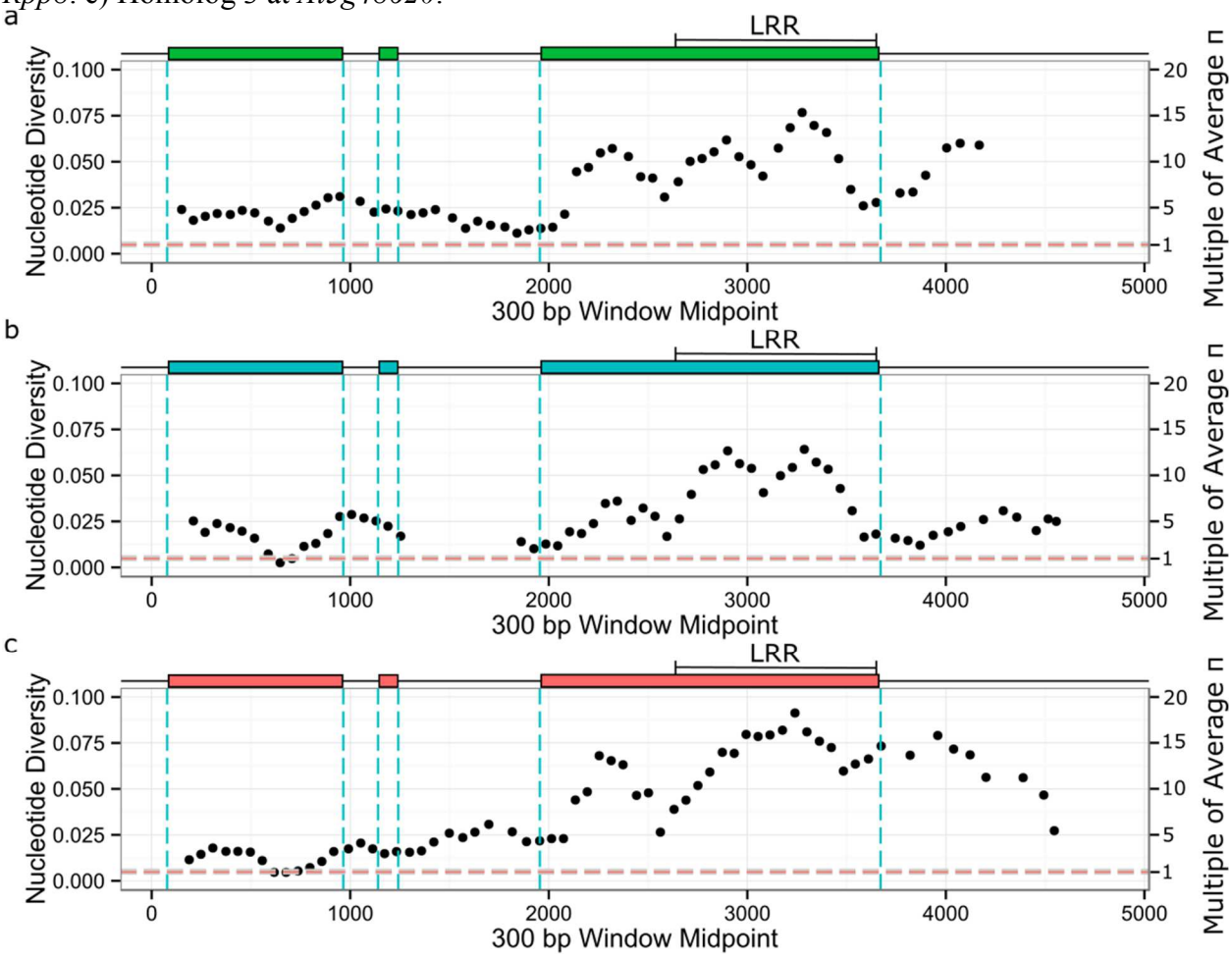
Synonymous nucleotide diversity ($\pi_s$) within and between the coding regions of *Rpp8* homologs varied between 0.0333 and 0.0597, six to twelve times the genome average of ~0.005 (Table 4.7,4.8). Nonsynonymous nucleotide diversity ($\pi_a$) within the coding region of *Rpp8* homologs was 20-27 times the genome average of ~0.0014 (Table 4.8). Ratios of $\pi_a$/ $\pi_s$ were also higher than the genome average, particularly in the LRR (Figure 4.7). Interestingly, no

comparisons of $\pi_s$ between homologs were significantly different (Table 4.8). A sliding window analysis of $\pi$ across each homolog showed that $\pi$ varied between 2-15 times the genome average for H1, 0.5-12 times the genome average for H2, and 0.9-18 times the genome average for H3 (Figure 4.9). The 700 bp intron sequence 5' of the LRR had $\pi_s$ levels 3.7, 3.4, and 4.5 times the genome average for H1-H3, respectively (Figure 4.9), while the LRR had $\pi_s$ levels 10.3, 5.6, and 13 times the genome average. Synonymous divergence between *A. thaliana* and *A. lyrata* in the *Rpp8* region was not in excess of the genome-wide expectations, indicating that a high mutation rate was not increasing $\pi$ or $\pi_s$ in these regions (Figure 4.7). Gene duplicates undergoing IGC are theorized to increase $\pi_s$ in both copies to two times the rate of single copy genes (Teshima and Innan 2012); values for *Rpp8* homologs were far in excess this expectation, even within the presumably neutrally-evolving intron.

**Table 4.8.** Synonymous nucleotide diversity ($\pi_s$) in the coding region between locus X, rows, and Y, columns, and the variance in $\pi_s$.

| X / Y | S | $D_1$ | $D_2$ | H3 |
|---|---|---|---|---|
| S | 0.0477 ± 0.0100 | 0.0449 ± 0.0079 | 0.0537 ± 0.0067 | 0.0526 ± 0.0130 |
| $D_1$ | - | 0.0392 ± 0.0110 | 0.0559 ± 0.0076 | 0.0476 ± 0.0138 |
| $D_2$ | - | - | 0.0333 ± 0.0156 | 0.0597 ± 0.0140 |
| H3 | - | - | - | 0.0451 ± 0.0195 |

**Figure 4.9.** Sliding window analysis of nucleotide diversity across the sequenced regions. Vertical lines indicate boundaries of coding regions of *Rpp8*, as shown in the *Rpp8* schematic above each plot. Green, blue, and orange boxes above the plots represent positions of exons of H1, H2, and H3, respectively. The leucine rich repeat region (LRR) is also indicated. Horizontal line indicates average level of nucleotide diversity within *A. thaliana*; line width is the confidence interval for average nucleotide diversity. a) Homolog 1 at *Rpp8*. b) Homolog 2 at *Rpp8*. c) Homolog 3 at *At5g48620*.



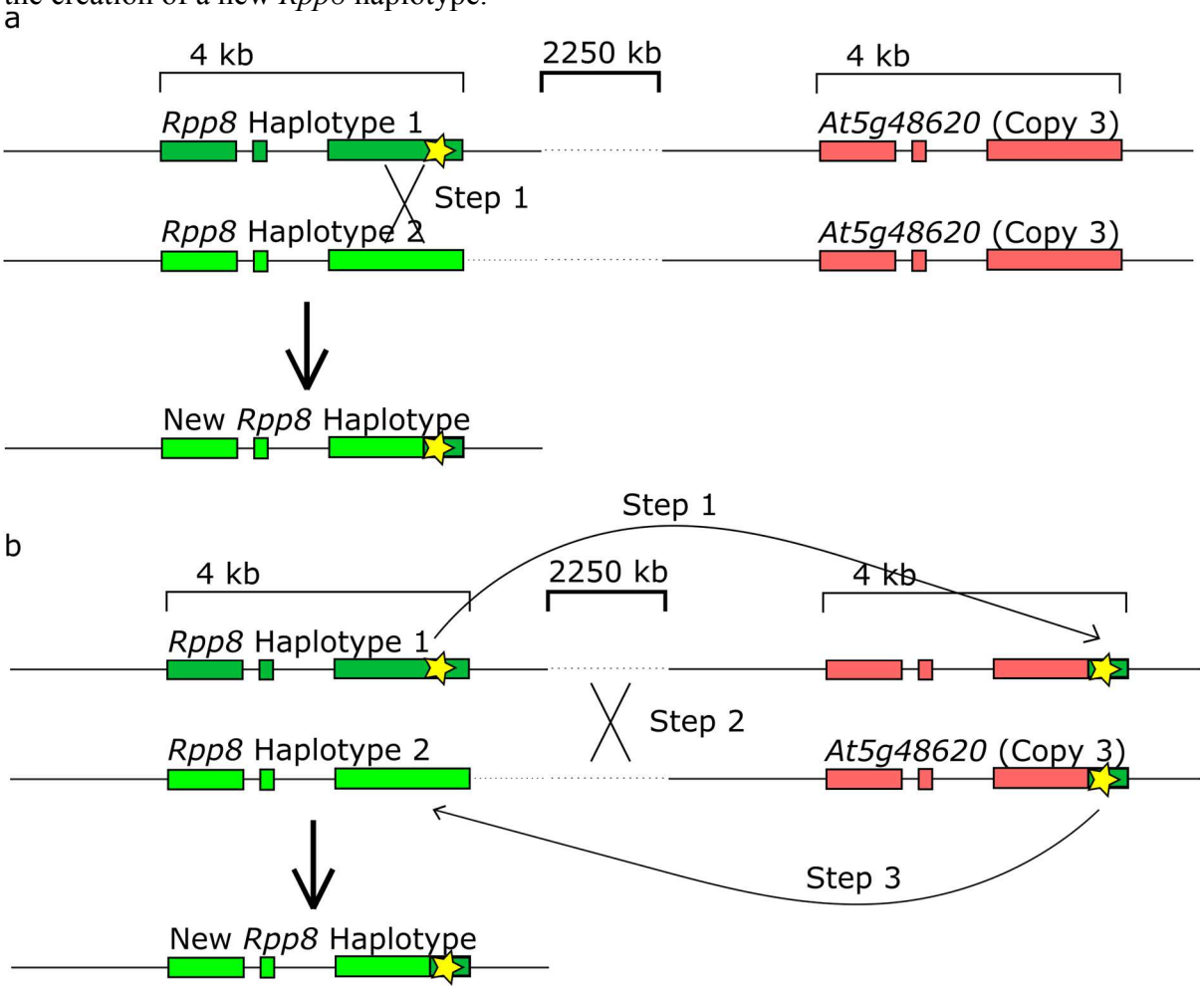

### *4.4.5 Waiting time for intergenic recombination*

The extensive sharing of SNPs between $D_1$, S, and H3, and the extraordinarily high shared nucleotide diversity at the three loci, raised the possibility that mutations were moved between $D_1$ and S via H3, as opposed to intragenic crossing over or intragenic gene conversion directly between $D_1$ and S. Here I explored the waiting times for recombination under the two

scenarios. As explained in the introduction, I envision two mechanisms by which a mutation is moved between genetic backgrounds at *Rpp8*: a rare, intragenic crossover event within *Rpp8* in a heterozygote (Figure 4.10a), or a three step process involving IGC between duplicates to move a mutation to the distant duplicate, a crossover event between duplicates, and another IGC between duplicates encompassing the same mutation to move the mutation to its original locus (Figure 4.10b). To analyze the effect of IGC on sharing of mutations between copies of *Rpp8*, an analytical solution was found for the rate of IGC required for the expected waiting time for the sharing of a mutation to be shorter through the second, IGC route than the first, rare crossover route. To explore the effect of duplicate architecture on haplotype diversity, we ignored IGC events with H2 and simulated IGC events between H1 and H3 only. Standard coalescent approximations were used for both recombination and gene conversion events. The rare crossover route involved a coalescent approximation for the waiting time to recombination event within the *Rpp8* region; the IGC route involved coalescent approximations for the waiting times to two independent gene conversion events within the *Rpp8* region and a coalescent approximation for the waiting time to a recombination event anywhere between H1 and H3.

In a coalescent framework, the waiting time to a recombination event is exponentially distributed with a rate parameter equal to (R/2), where R is the expected number of recombination events between the ends of the region being considered per 4N generations (Hudson 1983). R scales as a function of region size, $R = 4N\rho L$, where N is the population size, and $\rho$ is the recombination rate per generation between the ends of L, the length of the region being considered. We denote $L_1$ as the length of the region allowing within-gene recombination, and $L_2$ for the length of the region between the original location of the mutation and the location

**Figure 4.10.** Routes by which a new mutation could be moved onto a new genetic background at *Rpp8*. Two example chromosomes are shown. Exons of *Rpp8* loci are shown as boxes; lengths of *Rpp8* homologs and between *Rpp8* homologs in kilobases (kb) are shown above each figure. Stars indicate the new mutation. Colors indicate haplotypes of *Rpp8* homologs. Crossovers between chromosomes are indicated by X's; intergenic gene conversion events are indicated by thin arrows, and the resultant new haplotype is indicated by a thick arrow. a) A rare crossover event within *Rpp8* in an *Rpp8* heterozygote allows the creation of a new *Rpp8* haplotype. b) Two gene conversion events and a crossover event in the region between *Rpp8* and *At5g48620* allows the creation of a new *Rpp8* haplotype.




of the paralogous site in the duplicate. ρ has been approximated as 0.8 per kb in *A. thaliana* (Kim *et al*., 2007). In species where selfing is common, such as *A. thaliana*, the proportion of recombination events in heterozygotes decreases according to the function R(1-s), where s is the fraction of offspring produced by selfing (Nordborg 2000). Thus, the waiting time to a

recombination event within a gene region is exponentially distributed with a rate parameter equal to $4N\rho L_1(1\text{-}s)/2$; the waiting time to a recombination event between two gene regions is exponentially distributed with a rate parameter equal to $4N\rho L_2(1\text{-}s)/2$.

In a coalescent framework, the waiting time to a gene conversion event is exponentially distributed with a rate parameter of G/2, where G is the expected number of gene conversion events per sequence per 4N generations (Wiuf and Hein 2000). G scales as a function of region size, G = 4NcL, analogous to the parametrization adopted for the coalescent with recombination, with c here as the IGC rate between the duplicated regions (Wiuf and Hein 2000). Thus, the waiting time to exchange by IGC can be thought of as exponentially distributed with a rate parameter equal to $2*4NcL_1/2$. We simulated the proportion of cases where the waiting time to a rare crossover within *Rpp8* was much greater than the waiting time to two gene conversion events and a crossover between *Rpp8* paralogs for different values of s, $\rho$, c, $L_1$, and $L_2$. We also solved for the threshold c for which the average expected waiting time for a rare crossover event was equal to that of an exchange by IGC event. The average expected waiting time for an exponential distribution with rate parameter λ is equivalent to 1/λ. Thus, we solved the following inequality for c:

$$\frac{1}{(\frac{4N\rho L_1(1-s)}{2})} > \frac{1}{(\frac{4N\rho L_2(1-s)}{2})} + \frac{1}{(\frac{4NcL_1}{2})} + \frac{1}{(\frac{4NcL_1}{2})}$$

It is simple to show that this inequality is true for:

$$c > \frac{2\rho L_2(1-s)}{(L_2 - L_1)}$$

When $L_2$ is large compared to $L_1$, this inequality simplifies further to

$$c > 2\rho(1-s).$$

### *4.4.6 IGC Generates Novel Haplotypes at Rpp8*

For known values of s for *A. thaliana*, $L_1$ for *Rpp8*, and $L_2$ between *Rpp8* H1 and H3, we found that IGC had to occur at a rate of $0.06\rho$ or higher in order for IGC via H3 to be as fast as sharing polymorphism via a rare recombination event within *Rpp8*. This was a far lower rate than that required for an obligate outcrosser, where the rate is approximately $2\rho$. Both of these values assumed that every allele of the locus is distinct; that is, that any recombination event will necessarily move a mutation onto a new genetic background. This was not an unrealistic assumption for *Rpp8*, where there were 49 haplotypes in a sample of 51 homologs. However, the relative speed of multiple IGC events compared to a rare crossover event should also increase as the number of variant haplotypes at a locus decreases. This is because the number of within-locus crossover events that are visible, or that move a mutation onto a new genetic background, decrease as the number of haplotypes decreases. In contrast, the haplotypes at paralogous loci should typically be distinct. Thus, for genes under relatively recent diversifying selection, duplication and moderate IGC may carry an additional selective benefit in both generating and spreading variation. In particular, for selfing individuals, duplication and moderate IGC may be essential to generate diversity at a locus.

I simulated the proportion of cases where the waiting time to crossover events within Rpp8 was greater than the waiting time to exchange by IGC for different values of s,$\rho$, c, $L_1$, and $L_2$. Within-locus rates of gene conversion of 1-3 have been estimated in Arabidopsis (Kim *et al*., 2007), but estimates of IGC rates do not exist for *Rpp8* or *At5g48620*. In the absence of specific IGC estimates, I considered IGC rates of greater than three, or greater than that of within-locus gene conversion, unlikely, and thus considered a rate of three a “critical threshold” above which

IGC is unlikely to generate the majority of haplotypic variation. I note that I estimated the rate of IGC between *Rpp8* and *At5g48620* to be one or higher. For the estimated values of s, $\rho$, $L_1$, and $L_2$ for *Rpp8* and *At5g48620* in *A. thaliana*, IGC was faster than recombination on average when c > 0.048, and faster than recombination 95% of the time when c>0.8 (Figure 4.11a). The speed of IGC scaled logarithmically with the distance between duplicates. For an $L_1$ of 4kb, $L_2$ needed to be at least 114kb for a c of 3 or less to be faster than recombination 95% of the time, and greater than the length of $L_1$, equivalent to a tandem duplicate, for a c of 3 or less to be faster than recombination on average (Figure 4.11b). The speed of IGC scaled linearly with increasing s. For *Rpp8* values of $L_1$ and $L_2$, s needed to be 0.91 or greater for a c of 3 or less to be faster than recombination 95% of the time, and could be any value greater than 0 for a c of 3 or less to be faster than recombination on average (Figure 4.11c). Actual values of these parameters for the *Rpp8* system in *A. thaliana* were far in excess of the minimum requirements for IGC to be faster than recombination 95% of the time. In a predominantly selfing species, IGC between duplicates may thus be an important mechanism for sharing new mutations within a locus.

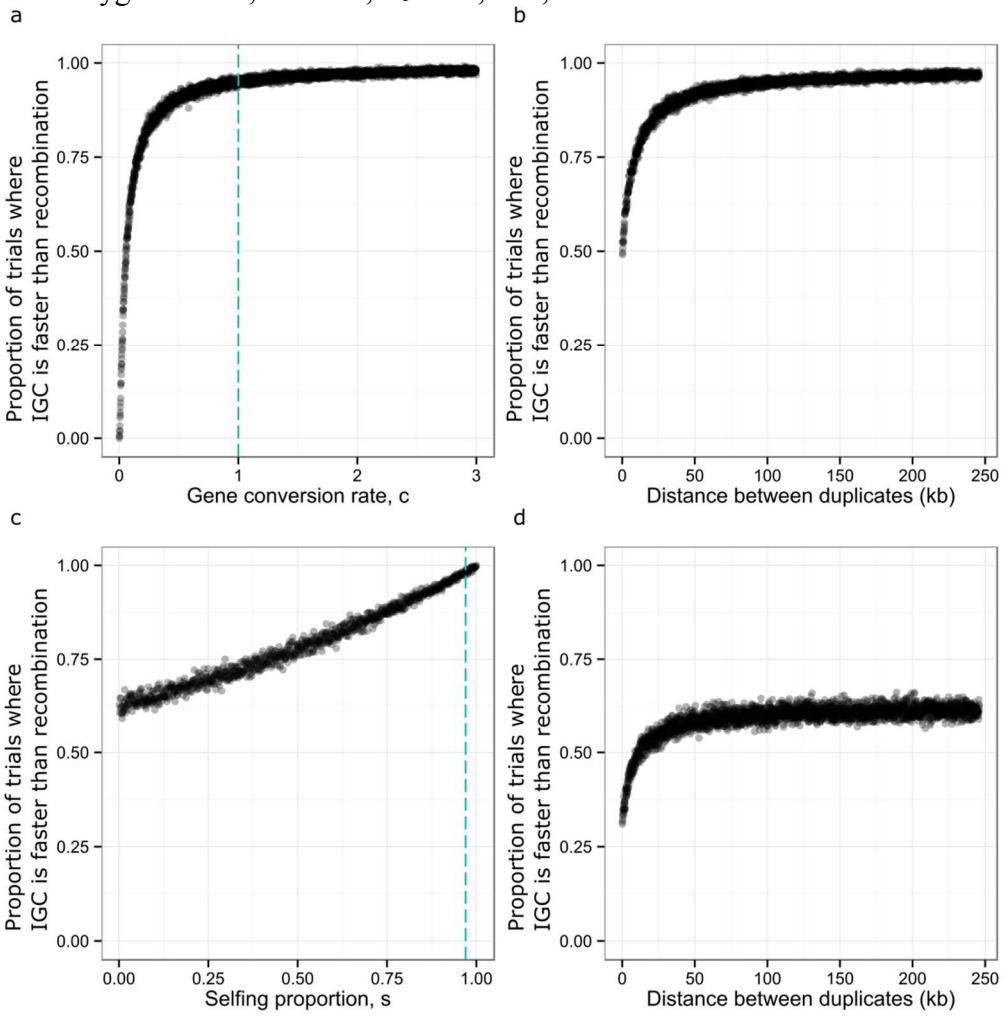
## 4.5 Discussion

The population genetics of the *Rpp8* homologs aided in the determination of how selection acts to generate an allelic series that can confer recognition for multiple Avr genes. The correlated frequencies of shared polymorphism, few or no fixed differences, reduced LD, and phylogeny of the LRR of the three homologs of *Rpp8* in the *A. thaliana* genome were consistent with IGC between all three sequenced homologs, with a higher rate of IGC between the distant duplicates H1 and H3 than between the tandem duplicates at *Rpp8*. The site frequency spectra between pairs of homologs were most consistent with expectations for a c of ~0.2-1 between H2

and the other two loci, and a c of ~1-5 between H1 and H3. Using established expectations from the coalescent for waiting times for recombination and gene conversion events in the history of the sample, we found that rates of IGC would need to be higher than c=0.8 for *Rpp8* in *A. thaliana* for IGC between H1 and H3 to be faster than a rare recombination event in H1 for moving mutations to different haplotypes. Thus, IGC between H1 and H3 was well within the range for it to be the major driver of the extreme haplotypic variation seen for loci in the *Rpp8* gene family.

We note that *R*-genes in many plant species are present in large, tandem arrays, which are frequently located on the ends of chromosomes, far from the pericentrochromatic suppression of recombination (Meyers *et al*., 1998; Meyers *et al*., 2003; Leister *et al*., 2004; Cheng *et al*., 2012; Schmutz *et al*., 2014; Perrazzolli *et al*., 2014). Additionally, IGC between *R*-gene homologs has been previously reported (Meagher *et al*., 1989; Meyers *et al*., 1998; Kuang *et al*., 2008). Ectopic recombination is often invoked as a likely mechanism to generate *R*-gene diversity in these clusters (Meyers *et al*., 1998; Parniske and Jones 1999; Meyers *et al*., 2003); however, evidence for ectopic crossover events involving gene gain or loss have thus far rarely been observed in studies of *R*-gene polymorphism (Baumgarten *et al*., 2003). *Rpp8* is known to be near a recombination hotspot (Kim *et al*., 2007). The presence of a tandem duplicate and distant duplicate for *Rpp8* allowed us to address whether non-allelic exchange primarily occurred through ectopic recombination in D variants or through IGC with *At5g48620*. In the former case, more SNPs should be shared between the tandem duplicates than between alleles at *Rpp8* and alleles at *At5g48620,* and equal levels of polymorphism shared between S, $D_1$, and $D_2$ loci; there should also be correlated frequencies of these polymorphisms in these alleles. In the latter case,

**Figure 4.11.** Simulations of the effects of IGC rate, selfing, and distance between duplicates on the speed of sharing of mutations by within-locus recombination or via IGC through a distant duplicate. a) Effect of c, gene conversion rate, on the proportion of trials where the expected waiting time is less for IGC between distant duplicates than for recombination in a heterozygote. Here, rho=0.8, $L_1$=4kb, $L_2$=2250kb, and s=0.97, all estimated from *A. thaliana* data. Blue dashed line shows the lower bound of the estimate of IGC between H1 and *At5g48620*, or H3, in *A. thaliana* from this study. B) Effect of $L_2$, the distance between duplicated genes, on the proportion of trials where the expected waiting time is less for IGC between distant duplicates than for recombination in a heterozygote. Here, rho=0.8, $L_1$=4kb, c=3, and s=0.97, all estimated from *A. thaliana* data. Actual distance between *Rpp8* and *At5g48620* is ten times larger than the right bound displayed. c) Effect of s, the selfing rate, on the proportion of trials where the expected waiting time is less for IGC between distant duplicates than for recombination in a heterozygote. Here, rho=0.8, $L_1$=4kb, $L_2$=2250kb, and c=3, all estimated from *A. thaliana* data. The blue dashed line shows the estimate of selfing in *A. thaliana*. d) Effect of $L_2$, the distance between duplicated genes, for an obligate outcrosser on the proportion of trials where the expected waiting time is less for IGC between distant duplicates than for recombination in a heterozygote. Here, rho=0.8, $L_1$=4kb, c=3, and s=0.



there should be more shared SNPs with correlated frequencies observed between *Rpp8* and *At5g48620* than between the tandem duplicates. The data supported higher sharing between the distant duplicates. We hypothesized that this difference was due to the ease of moving mutations onto new genetic backgrounds by IGC with a distant duplicate. IGC events were always faster than rare recombination events in selfing populations, no matter the distance between duplicates; they were faster for obligate outcrossers for distances > 12 kb, but only faster than recombination 60% of the time (Figure 4.11d). In particular, then, IGC could generate *R*-gene diversity in selfing species, by adding additional genomic regions of diversity which are much less likely to become homogenized than a single locus. In selfers, IGC may be more likely to generate new haplotypes when it occurs between distant homologs, more than 25kb apart, in large tandem arrays of *R*-genes, because the speed of IGC relative to recombination depends on the distance between duplicates (Figure 4.11b).

*R*-gene duplicates undergoing IGC could also raise the background level of $\pi$ expected within both loci. By extension from a two-copy model, I assume that the level of $\pi$ should increase as the number of loci undergoing IGC increases. With sufficient exchange, I assume the level of polymorphism should increase linearly with the number of loci. Thus, tandemly duplicated arrays of *R*-genes, if undergoing IGC, could likely maintain levels of $\pi$ at least as high as the product of the number of genes undergoing IGC. The effect of the number of duplicates on nucleotide diversity deserves greater attention, either analytically or by simulation. A simulation framework exists for two loci that could be extended to include additional loci undergoing IGC. Our analysis suggests that the *Rpp8* gene family could be thought of as a three locus model. The *Rpp8* triplication undergoing IGC is sufficient to explain the level of $\pi$ found in the introns of

*Rpp8* (Figure 4.9), but does not account for the entire excess within the LRR and the regions adjoining the LRR. Two selective possibilities could explain this excess. First, the LRR could have undergone a number of soft selective sweeps for functionality, in one or multiple homologs. This would allow the spread of numerous alleles to intermediate frequencies. Alternatively, negative frequency dependent selection could prevent fixation of swept alleles. In either scenario, IGC between *Rpp8* and the additional two homologs could allow fast return of $\pi$ in the swept homolog to their former levels, or allow a selected mutation to 'sweep' on multiple genetic backgrounds, in effect erasing evidence of a selective sweep.

Frequent IGC in tandem arrays of *R*-genes could help explain the impenetrability of *R*-gene arrays to next generation sequencing approaches. High IGC rates mix sequence fragments between duplicates, giving no fixed differences between long-diverged copies and, indeed, intermixing regions undergoing high IGC so that they are paraphyletic on a phylogeny and unresolvable to the genomic location of either duplicate with short read sequencing. I observed both of these signatures for homologs of *Rpp8* (Figure 4.6b; Table 4.3); despite this, all three copies of *Rpp8* were collinear on the syntenic region of chromosome 8 in *A. lyrata* (Hu *et al*., 2011), indicating that this gene family and gene family architecture are ancient. Long read sequencing methods such as Sanger sequencing or emerging technologies such as PacBio will be necessary to view and understand the patterns of polymorphism in more complex *R*-gene arrays.

Finally, we note that *At5g48620* may have no canonical function in pathogen resistance, but instead may exist to increase diversity at *Rpp8* proper. Alternatively, *At5g48620* may have similar resistance functions as *Rpp8* obtained through IGC. This may be the case for many *R*-genes present in tandem arrays. I observed patterns of polymorphism in *Rpp8* which were

consistent with IGC between all duplicates, and with faster rates of IGC between distant duplicates. These features both generated and shared an excess of novel variants compared to single copy genomic regions. We suggested these features may be common for *R*-genes under diversifying selection and present in large tandem arrays. *R*-genes in tandem arrays have thus far been inaccessible to next generation sequencing technologies; we described specific expectations for levels of nucleotide diversity and patterns of IGC between *R*-genes within a large array which should be tested as long-read next generation sequencing comes online.

# Chapter 5

## Conclusions

This dissertation work expands on our previous knowledge of natural variation of *R*-genes in *Arabidopsis thaliana*. Chapter 2 is the first work to consider neutral and adaptive explanations for the unusually high natural variation in *R*-gene expression, and supports distinct selective forces shaping *R*-gene expression relative to the transcriptome as a whole. Chapter 3 seeks to understand differences in fitness costs between *R*-loci with and without insertion-deletion polymorphisms, and the selective mechanisms driving these differences. It suggests that long-maintained insertion-deletion polymorphisms exacerbate costs of resistance, and that *R*-loci segregating for functional alternatives may have evolved either additional, beneficial functions, or reduced levels of expression, that mask or eliminate fitness costs of resistance. Chapter 4 describes the population genetics of complex *R*-gene loci, *Rpp8* and *At5g48620*, and offers a mechanism that could explain why most *R*-genes, in most sequenced species, occur in arrays of tandemly duplicated genes: interlocus gene conversion. Particularly in predominantly selfing plant species, interlocus gene conversion can be a fast mechanism for the generation and sharing of novel haplotypes, an essential requirement for fast-evolving pathogen recognition proteins. In addition to this work on *R*-gene evolution, I made a co-first author contribution to a paper looking at variation in and features of a transcriptional mark, $N^6$-methyladenosine methylation of mRNA in *A. thaliana*. I find a unique genomic pattern for this editing mark that is associated with plant-specific, chloroplast-related pathways (Appendix F). In the following sections, I address known caveats of the previous chapters and offer future directions for these datasets and

future projects to increase our understanding of components of fitness and natural variation in fast-evolving, responsive gene families.

### 5.1 Chapter 2 Caveats and Future Work

Both *R*-gene overexpression mutants and patterns of *R*-gene expression polymorphism support a link between *R*-gene expression and plant fitness. However, this relationship is only correlative. Two experimental manipulations could causally link *R*-gene expression variation and plant fitness. First, *R*-gene expression knockdowns could be created using amiRNAs. Measurements of fitness for these lines in the absence of pathogens would help determine if natural levels of *R*-gene expression are sufficient to reduce plant fitness. Knockdowns of multiple individual *R*-genes would describe the variation in any "cost of expression" for *R*-genes. However, it may be difficult to measure the potentially small fitness benefits of knockdowns of single *R*-genes. amiRNA lines offer the potential to knockdown multiple related targets in the same genetic background. I have designed an amiRNA that potentially knocks down up to 25 closely related *R*-genes. Lines like this could reveal a collective benefit of removing many *R*-genes in the absence of pathogen. Second, promoter swaps between *R*-alleles with high and low levels of expression may reveal fitness costs of *R*-alleles masked in their natural promoter context. CRISPR may soon allow sophisticated replacements and knockouts in accessions other than Columbia, and the varied and costly *R*-genes would be excellent early candidates for manipulations in *A. thaliana*.

Three changes to the *R*-gene expression datasets from the first chapter would have helped increase the power of this study and the certainty in the results: a larger accession sample for the qPCR experiment, and a better chosen sample for clinality in the RNA-seq expression dataset,

and a larger worldwide RNA-seq dataset. The qPCR dataset was limited by the cost of qPCR reagents, while the RNA-seq datasets were generated to study questions unrelated to clinality in expression. Population structure is a concern in both the Midwestern and Swedish latitudinal clines, and any future work would ideally choose three or more independent clines. As *R*-genes are ubiquitous in plant species, expression clines could be examined within multiple plant species subject to multiple, independent adaptive radiations along similar climatic gradients to give truly independent datasets. A dataset of this nature could answer an additional question of intense interest in evolutionary biology by generating novel insights into the repeatability of adaptive evolution to climate.

Within *A. thaliana*, future work with the currently generated RNA-seq datasets should include confirmation of candidates from GWA using *R*-gene expression as a phenotype. Confirmation of cis-variants and trans-candidates that affect variation in *R*-gene expression would elucidate important levels of regulation of these costly genes.

### 5.2 Chapter 3 Caveats and Future Work

Our fitness measurements of an allelic series of *Rps2* in the absence of pathogen could explain the likely reason that *Rps2* is not an insertion-deletion polymorphism: the large fitness benefit of plants with any allele of *Rps2* relative to an artificial knockout of *Rps2*. RNA-seq results indicate that RPS2 presence negatively regulates a number of components of the defense response and induced stress responses; however, the mechanism of this function remains unknown. To find an alternative function, it would be helpful to know whether S alleles still associate and interact with two other important members of the RPS2 complex, RIN4 and NDR1, or whether they instead interact with some new component of the plant defense machinery. Co-

immunoprecipitation experiments or yeast two-hybrid experiments with labeled S clade RPS2, RIN4, and NDR1 could address this question. In addition, RNA-seq results indicate that RPS2 presence negatively regulates a number of components of the defense response and induced stress responses. Thus, it is possible that RPS2 regulates NDR1 activity just as NDR1 regulates RPS2.

Our fitness measurements in the absence of pathogen did not explain why S alleles of *Rps2* have been maintained. If there is a benefit to both R and S alleles in the absence of pathogen, but a benefit to only R alleles in the presence of pathogen, then R alleles should sweep to fixation. Thus, there is likely an alternative function for S alleles in the presence of some pathogen species that our work in the absence of pathogen was unable to determine. A simple possibility is that S alleles may have an alternative resistance functionality. Single *R*-loci can evolve to confer recognition of multiple Avirulence genes (Dangl and Jones, 2001): thirteen sequenced flax rust resistance alleles each have a different L locus specificity (Ellis *et al*., 1999), and alleles of the *R*-gene *Rpp8* have been documented that resist a fungus and two distinct viruses (McDowell *et al*., 1998; Cooley *et al*., 2000; Takahashi *et al*., 2002). Alternatively, in the presence of pathogen, *R* alleles of *Rps2* are known to suffer a net cost with attack in the absence of competition. This cost in the presence of pathogen has been modeled as sufficient to maintain both clades of *Rps2* alleles without an additional resistance function for S alleles (Korves and Bergelson 2004).

If I had designed this allelic series, I would have chosen more alleles from the S clade of *Rps2* to test. I would have attempted to create an allelic series with three R alleles, two strongly resistant (sR), three S alleles, and the artificial knockout of *Rps2*. In our study, there was only

one S allele to compare against numerous R alleles. Substantial variation in fitness between R alleles was noted in this study, but this allelic series could not capture any variation in fitness between S alleles. This could skew our results if the S allele chosen happened to be particularly fit or unfit. Additionally, severely resistant, or sR lines were not included in the study design, and it is possible they could carry a cost of resistance relative to S lines, as they by definition have the highest level of resistance to *P. syringae* pv. *avrRpt2*. As another control, both the field and growth chamber experiments should have had at least Col-0 and Wu-0 planted along with the transgenic allelic series. This would have reassured us that the overexpression of *Rps2* lines was not causing a fitness cost in the transgenic series relative to natural lines; if there was a cost, it would have informed us of its magnitude. Col-0, the genetic background alleles of *Rps2* were added on to, would have been particularly helpful to have had field fitness data for to infer these costs. For the field experiment, twice as many plants should have been planted per line in the field experiment. This would have greatly increased the power to discriminate between lines within each insertion site, and would have perhaps prevented some repetitions of fitness experiments for this chapter. If I did the field experiment again, I would plant each set of lines created at a different insertion sites on a different day, thus spreading the work of planting out over three days and allowing many more plants of each insertion site to be planted per day. Instead of planting 2200 plants in one day, I would plant 1600 plants per day over three days, 200 plants per line for each of the 6 lines of the allelic series, as well as Col-0 and Wu-0.

### 5.3 Chapter 4 Caveats and Future Work

In this chapter, I articulate specific requirements for IGC relative to recombination rate to allow IGC between *Rpp8* and *At5g48620* to increase haplotype diversity at *Rpp8*. I also draw

general conclusions about the level of IGC in this system by determining the level of similarity of observed spectra to expectations of spectra with low, medium, and high gene conversion rates. Unfortunately, apart from these correlations, we still lack a precise genetic method to determine an estimate of IGC. Bayesian methods may help to increase the precision of this estimate. Alternatively, empirical estimates of rates of recombination and gene conversion in these regions may be possible and may help fill in the blanks on how this system truly behaves. This has been attempted in Drosophila (Langley *et al*., 2000). However, it would be very difficult to sequence the number of plants necessary to get empirical estimates of these quantities (>10,000).

I am currently working with collaborators to simulate the effects of a long-maintained triplication on levels of nucleotide diversity by modifying a forward-time simulator of the effects of gene duplication on nucleotide diversity to take into account effects of gene triplication. We intuit that, for n copies of a gene undergoing interlocus gene conversion, levels of nucleotide diversity should increase to at least n times the background level of the genome (Innan, Hartasanchez, personal communication). Thus far, our model results confirm our intuition that a three copy system should have three times the level of nucleotide diversity as a one copy system. We are also modeling the effect of the copy number polymorphism at *Rpp8* H2. If this copy number polymorphism is under balancing selection, it is possible that this is generating an increased level of polymorphism at *Rpp8* H1, which is tightly linked to H2. Balancing selection on *Rpp8* H2 copy number could also explain the decrease in diversity we see at H2 relative to the other *Rpp8* homologs (Table 4.8). In addition, this copy number polymorphism could explain the asymmetry in the IGC rates we see between homologs of *Rpp8* (Figure 4.4): fewer copies of H2 that could undergo IGC should be reflected in a lower estimated rate of IGC with H2, which we

see reflected in the IGC estimates (Figure 4.4). We are currently actively modeling the effects of parameter settings on polymorphism in a three copy gene family and on polymorphism in a three copy gene family with one copy under balancing selection for presence or absence.

Multiple attempts to use NGS data for analysis of sequence polymorphism at *Rpp8* were ultimately discarded from this chapter. I assembled a set of 18 accessions where I had both Sanger data (with *Rpp8* called as a one-copy or two-copy variant) and NGS data, and tested various methods to call duplicates for *Rpp8* on this set. A commonly used program to call duplications, DELLY (Rausch *et al*., 2012), only calls one of five accessions correctly. *Rpp8* region coverage aligned to a one-copy variant should be twice as high for duplicates compared with single copy genes; however, this method worked for only 20% of accessions. Independent assembly of the *Rpp8* region into contigs with either one or two copies of *Rpp8* with velvet were not able to assemble two known duplicates of *Rpp8* into two-copy regions. Distance matrices of SNP calls when accessions were aligned to a two-copy variant called six of ten accessions correctly, a result little better than chance. Finally, *Rpp8* two-copy variants have a unique 2000bp region that one-copy variants lack. Measuring coverage of this unique region when accessions were aligned to a two-copy variant called 16 of 18 accessions correctly, the highest success rate that I had. Despite this, I was concerned that NGS assembly methods would obscure the extreme haplotype diversity we found using Sanger sequencing by incorrectly assigning SNPs to homologs. Thus, I analyze only Sanger sequences in this chapter. The read lengths obtained by NGS technology needs to increase to at least a few kilobases to allow variants within *Rpp8* to be tagged by fixed mutations outside the 4kb duplicated region undergoing interlocus gene conversion, before this kind of analysis will be possible using NGS.

## 5.4 Appendix F Caveats and Future Work

Our preliminary adaptive hypothesis regarding variation in the $m^6A$ methylome in *A. thaliana* was that that $m^6A$ edits were important for energy regulation in the cell, and likely varied in plants from different historic light environments. Our experiments did not support variation in $m^6A$ editing associated with light environment at a within-species level. To better understand this variation, new adaptive hypotheses could be generated after determining the scale at which this variation occurs. Multiple plant species could be assayed to determine when a 5' editing mark evolved, if this mark is always associated with pathways involving the chloroplast, and if this editing mark is conserved within plants or has evolved more recently. Future work within *A. thaliana* could extend these analyses to additional accessions to explore environmental correlates for this variation (Hancock *et al*., 2011; Carniero *et al*., 2013; Phifer-Rixey *et al*., 2014; Lasky *et al*., 2015). At the individual level, the variation between different plant tissues for this editing mark could also be described, as well as the dynamism of this transcriptional mark across different environments and growth stages. For example, we have some indication that the 5' $m^6A$ editing mark is associated with chloroplast proteins; thus, we could explore if this mark has a circadian rhythm or is altered when photosynthesis is and is not occurring. As our dataset is the first to characterize this transcriptional mark in plants, a huge number of avenues for follow-up remain for this work.

## References


Aboul-Soud MA, Chen X, Kang J-GG, Yun B-WW, Raja MU, Malik SI & Loake GJ (2009) Activation tagging of ADR2 conveys a spreading lesion phenotype and resistance to biotrophic pathogens. *New Phytol.* **183:** 1163–1175

Allen RL, Bittner-Eddy PD, Grenville-Briggs LJ, Meitz J, Rehmany AP, Rose LE, Beynon JL (2004) Host-parasite coevolutionary conflict between *Arabidopsis* and downy mildew. *Science* **306:** 1957-1960

Anastasio AE, Platt A, Horton M, Grotewold E, Scholl R, Borevitz JO, Nordborg M & Bergelson J (2011) Source verification of mis-identified *Arabidopsis thaliana* accessions. *Plant J.* **67:** 554–566

Anders S and Huber W (2010) Differential expression analysis for sequence count data. *Genome Biology* **11**: R106

Ausubel, Frederick M (1991) *Current Protocols in Molecular Biology*. New York: Greene Pub. Associates and Wiley-Interscience, Print.

Axtell M & Bowman J (2008) Evolution of plant microRNAs and their targets. *Trends Plant Sci.* **13:** 343–349

Axtell M & Staskawicz B (2001) Mutational Analysis of the Arabidopsis RPS2 Disease Resistance Gene and the Corresponding *Pseudomonas syringae avrRpt2* Avirulence Gene. *MPMI* **14**: 2

Axtell M & Staskawicz B (2002) Initiation of RPS2-Specified Disease Resistance in Arabidopsis Is Coupled to the *AvrRpt2*-Directed Elimination of RIN4. *Cell* **112**

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

Ballvora A, Ercolano M, Weiß J, Meksem K, Bormann C, Oberhagemann P, Salamini F & Gebhardt C (2002) The R1 gene for potato resistance to late blight (Phytophthora infestans) belongs to the leucine zipper/NBS/LRR class of plant resistance genes. *Plant J* **30:** 361–371

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

Baxter I, Brazelton JN, Yu D, *et al*. **(**2010) A coastal cline in sodium accumulation in *Arabidopsis thaliana* is driven by natural variation of the sodium transporter AtHKT1;1. *PLoS Genet.* **6**:e1001193

Bergelson J & Purrington C (1996) Surveying Patterns in the Cost of Resistance in Plants. *Am. Nat.* **148**(3): 536-558

Bergelson J, Dwyer G & Emerson J (2001) Models and Data on Plant-Enemy Coevolution. *Annu. Rev. Genet.* **35:** 469-499

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

Bhattarai K, Atamian H, Kaloshian I & Eulgem T (2010) WRKY72-type transcription factors contribute to basal immunity in tomato and *Arabidopsis* as well as gene-for-gene resistance mediated by the tomato R gene Mi-1. *Plant J. Cell Mol. Biology* **63**: 229–240

Bieniawska Z, Espinoza C, Schlereth A, Sulpice R, Hincha D, Hannah M (2008) Disruption of the *Arabidopsis* circadian clock is responsible for extensive variation in the cold-responsive transcriptome. *Plant physiology* **147**: 263–79

Bisgrove S, Simonich M, Smith N, Sattler A & Innes R (1994) A Disease Resistance Gene in Arabidopsis with Specificity for Two Different Pathogen Avirulence Genes. *The Plant Cell* **6:** 927

Boccara M, Sarazin A, Thiébeauld O, Jay F, Voinnet O, Navarro L & Colot V (2014) The Arabidopsis miR472-RDR6 Silencing Pathway Modulates PAMP- and Effector-Triggered Immunity through the Post-transcriptional Control of Disease Resistance Genes. *PLoS Pathog.* **10:** e1003883

Bodenhausen N, Horton MW & Bergelson J (2013) Bacterial communities associated with the leaves and the roots of *Arabidopsis thaliana*. *PLoS ONE* **8:** e56329

Bodi Z, Zhong S, Surbhi M, Song J, Graham N, Li H, May S, Fray R (2012) Adenosine Methylation in Arabidopsis mRNA is Associated with the 3′ End and Reduced Levels Cause Developmental Defects. Frontiers in Plant Genetics and Genomics **3**: 48

Bokar JA, Rath-Shambaugh ME, Ludwiczak R, Narayan P, Rottman F (1994) Characterization and partial purification of mRNA N6-adenosine methyltransferase from HeLa cell nuclei. Internal mRNA methylation requires a multisubunit complex. The Journal of biological chemistry. 269: 17697–17704

Caicedo AL, Schaal BA & Kunkel BN (1999) Diversity and molecular evolution of the RPS2 resistance gene in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U.S.A.* **96:** 302–306

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, Wang X, Ott F, Müller J, Alonso-Blanco C, Borgwardt K, Schmid KJ & Weigel D (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* **43:** 956–963

Carneiro M, Baird S, Afonso S, Ramirez E, Tarroso P, Teotónio H, Villafuerte R, Nachman M & Ferrand N (2013) Steep clines within a highly permeable genome across a hybrid zone between two subspecies of the European rabbit. *Mol. Ecol.* **22:** 2511–2525

Chan EK, Rowe HC, Corwin JA, Joseph B, Kliebenstein DJ (2011) Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in *Arabidopsis thaliana*. *PLoS Biol.* **9**: e1001125

Cheng Y, Li X, Jiang H, Ma W, Miao W, Yamada T & Zhang M (2012) Systematic analysis and comparison of nucleotide-binding site disease resistance genes in maize. *Febs J.* **279:** 2431–43

Chin DB, Arroyo-Garcia R, Ochoa OE, Kesseli RV, Lavelle DO & Michelmore RW (2001) Recombination and spontaneous mutation at the major cluster of resistance genes in lettuce (*Lactuca sativa*). *Genetics* **157:** 831–849

Chisholm ST, Coaker G, Day B & Staskawicz BJ (2006) Host-microbe interactions: shaping the evolution of the plant immune response. *Cel* **124:** 803–814

Cipollini D (2008) Constitutive expression of methyl jasmonate-inducible responses delays reproduction and constrains fitness responses to nutrients in *Arabidopsis thaliana*. *Evolutionary Ecology* **24**

Clancy MJ, Shambaugh ME, Timpte CS, Bokar JA (2002) Induction of sporulation in Saccharomyces cerevisiae leads to the formation of N6-methyladenosine in mRNA: a potential mechanism for the activity of the IME4 gene. *Nucleic acids research*. **30**:4509–4518

Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, Warthmann N, Hu TT, Fu G, Hinds DA, Chen H, Frazer KA, Huson DH, Schölkopf B, Nordborg M, Rätsch G, Ecker JR & Weigel D (2007) Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* **317:** 338–342

Clausen, J, Keck D, & Hiesey WH (1940) Experimental studies on the nature of species. I. Effects of varied environments on western North American plants. Carnegie Institution of Washington, Washington, DC.

Coley PD, Bryant JP & Chapin FS (1985) Resource availability and plant antiherbivore defense. *Science* **230:** 895–899

Cooley M, Pathirana S, Wu H, Kachroo P & Klessig D (2000) Members of the Arabidopsis HRT/RPP8 Family of Resistance Genes Confer Resistance to Both Viral and Oomycete Pathogens. *Plant Cell* **12:** 663676

Coppinger P, Repetti P, Day B, Dahlbeck D, Mehlert A & Staskawicz B (2004) Overexpression of the plasma membrane-localized NDR1 protein results in enhanced bacterial disease resistance in *Arabidopsis thaliana*. *The Plant Journal* **40:** 225–237

Costello EK, Stagaman K, Dethlefsen L, Bohannan BJ & Relman DA (2012) The application of ecological theory toward an understanding of the human microbiome. *Science* **336:** 1255–1262

Cridland J & Thornton K (2010) Validation of rearrangement break points identified by paired-end sequencing in natural populations of *Drosophila melanogaster*. *Genome Biol Evol* **2:** 83–101

Czechowski T, Stitt M, Altmann T, Udvardi MK & Scheible W-RR (2005) Genome-wide identification and testing of superior reference genes for transcript normalization in Arabidopsis. *Plant Physiol.* **139:** 5–17

Dangl JL & Jones JD (2001) Plant pathogens and integrated defence responses to infection. *Nature* **411:** 826–33 Available at: http://www.scholaruniverse.com/ncbi-linkout?id=11459065

Day B, Dahlbeck D & Staskawicz B (2006) NDR1 Interaction with RIN4 Mediates the Differential Activation of Multiple Disease Resistance Pathways in Arabidopsis. *The Plant Cell Online* **18:** 2782–2791

Dean R, Kan J, Pretorius Z, Hammond-Kosack K, Pietro A, Spanu P, Rudd J, Dickman, Kahmann R, Ellis J & Foster G (2012) The Top 10 fungal pathogens in molecular plant pathology. *Mol. Plant Pathology* **13:** 414–30

Denoux C, Galletti R, Mammarella N, Gopalan S, Werck D, Lorenzo GD, Ferrari S, Ausubel FM, Dewdney J (2008) Activation of Defense Response Pathways by OGs and Flg22 Elicitors in *Arabidopsis* Seedlings. *Molecular Plant* **1**

Dickey L, Petracek M, Nguyen T, Hansen E & Thompson W (1998) Light Regulation of Fed-1 mRNA Requires an Element in the 5' Untranslated Region and Correlates with Differential Polyribosome Association. *Plant Cell* **10:** 47

Dodds PN, Lawrence GJ & Ellis JG (2001) Six amino acid changes confined to the leucine-rich repeat beta-strand/beta-turn motif determine the difference between the P and P2 rust resistance specificities in flax. *Plant Cell* **13:** 163–178

Dominissini D, Moshitch-Moshkovitz S, Salmon-Divon M, Amariglio N & Rechavi G (2013) Transcriptome-wide mapping of N6-methyladenosine by m6A-seq based on immunocapturing and massively parallel sequencing. *Nat. Protoc.* **8:** 176–189

Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, Cesarkas K, Jacob-Hirsch J, Amariglio N, Kupiec M, Sorek R & Rechavi G (2012) Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nat.* **485:** 201–206

Dubin M, Zhang P, Meng D, Remigereau M-S, Osborne E, Casale F, Drewe P, Kahles A, Jean G, Vilhjálmsson B, Jagoda J, Irez S, Voronin V, Song Q, Long Q, Rätsch G, Stegle O, Clark R & Nordborg M (2015) DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. *eLife* **4**: e05255

Ellis JG, Lawrence GJ, Luck JE & Dodds PN (1999) Identification of regions in alleles of the flax rust resistance gene L that determine differences in gene-for-gene specificity. *Plant Cell* **11:** 495–506

Ferrante P, Fiorillo E, Marcelletti S, Marocchi F, Mastroleo M, Simeoni S, & Scortichini M (2012) The importance of the main colonization and penetration sites of *Pseudomonas syringae* pv. *Actinidiae* and prevailing weather conditions in the development of epidemics in yellow kiwifruit, recently observed in central Italy. *Journal of Plant Pathology* **94**: 455-461

Flor HH (1971) Current status of the gene-for-gene concept. *Annual review of phytopathology.* Available at: http://www.annualreviews.org/doi/pdf/10.1146/annurev.py.09.090171.001423

Fu Y, Dominissini D, Rechavi G, He C (2014) Gene expression regulation mediated through reversible m(6)A RNA methylation. *Nature Reviews Genetics*. **15**: 293–306

Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan VT, Kahles A, Bohnert R, Jean G, Derwent P, Kersey P, Belfield EJ, Harberd NP, Kemen E, Toomajian C, Kover PX, *et al.,* (2011) Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature* **477:** 419–423

Gao L, Roux F & Bergelson J (2009) Quantitative fitness effects of infection in a gene-for-gene system. *New Phytol.* **184:** 485–494

Gautier L, Cope L, Bolstad B & Irizarry R (2004) affy-analysis of Affymetrix GeneChip data at the probe level. *Bioinform.* **20:** 307–315

Grant J, Chini A, & Basu D (2003) Targeted activation tagging of the *Arabidopsis* NBS-LRR gene, *ADR1*, conveys resistance to virulent pathogens. *Molecular plant-microbe interactions: MPMI* **16**(8): 669-680

Grant M, Godiard L, Straube E, Ashfield T, Lewald J, Sattler A, Innes R, Dangl J (1995) Structure of the Arabidopsis RPM1 gene enabling dual specificity disease resistance. *Science* **11**;269(5225): 843-846

Grant MR, McDowell JM, Sharpe AG, de Torres Zabala M, Lydiate DJ & Dangl JL (1998) Independent deletions of a pathogen-resistance gene in Brassica and Arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.* **95:** 15843–15848

Greenberg J (1997) Programmed Cell Death in Plant-Pathogen Interactions. *Annu. Rev. Plant. Physiol. Plant. Mol. Biol.* **48:** 525–545

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

Gutiérrez R, Ewing R, Cherry J & Green P (2002) Identification of unstable transcripts in Arabidopsis by cDNA microarray analysis: Rapid decay is associated with a group of touch- and specific clock-controlled genes. *Proc. Natl. Acad. Sci.* **99:** 11513–11515

Guttman DS & Greenberg JT (2001) Functional analysis of the type III effectors *AvrRpt2* and *AvrRpm1* of *Pseudomonas syringae* with the use of a single-copy genomic integration system. *Mol. Plant Microbe Interact.***14:** 145–155

Hancock AM, Brachi B, Faure N, Horton MW, Jarymowycz LB, Sperone FG, Toomajian C, Roux F & Bergelson J (2011) Adaptation to climate across the *Arabidopsis thaliana* genome. *Science* **334:** 83–6

Hannah M, Wiese D, Freund S, Fiehn O, Heyer A, Hincha D (2006) Natural genetic variation of freezing tolerance in Arabidopsis. *Plant Physiology* **142**: 98–112

Hartasánchez D, Vallès-Codina O, Brasó-Vives M & Navarro A (2014) Interplay of interlocus gene conversion and crossover in segmental duplications under a neutral scenario. *G3 (Bethesda)* **4:** 1479–1489

Heinz S, Benner C, Spann N, Bertolino E, Lin Y, Laslo P, Cheng J, Murre C, Singh H & Glass C (2010) Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38**

Holub EB, Brose E, Tör M, Clay C, Crute IR & Beynon JL (1995) Phenotypic and genotypic variation in the interaction between *Arabidopsis thaliana* and *Albugo candida*. *Mol. Plant Microbe Interact.* **8:** 916–928

Horling F, Lamkemeyer P, König J, Finkemeier I, Kandlbinder A, Baier M & Dietz K-J (2003) Divergent Light-, Ascorbate-, and Oxidative Stress-Dependent Regulation of Expression of the Peroxiredoxin Gene Family in Arabidopsis. *Plant Physiology* **131:** 317–325

Horton M, Bodenhausen N, Beilsmith K, Meng D, Muegge B, Subramanian S, Vetter M, Vilhjálmsson B, Nordborg M, Gordon J & Bergelson J (2014) Genome-wide association study of Arabidopsis thaliana leaf microbial community. *Nat Comms* **5:**5320

Hu T, Pattyn P, Bakker E, Cao J, Cheng J-F, Clark R, Fahlgren N, Fawcett J, Grimwood J, Gundlach H, Haberer G, Hollister J, Ossowski S, Ottilar R, Salamov A, Schneeberger K, Spannagl M, Wang X, Yang L, Nasrallah M, *et al.,* (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nature Genetics* **43:** 476–481

Huang D, Sherman B, Tan Q, Collins J, Alvord G, Roayaei J, Stephens R, Baseler MW, Lane H & Lempicki R (2007) The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome biology.* **8**: R183

Huber C, Nordborg M, Hermisson J, & Hellmann I (2014) Keeping it local: Evidence for Positive Selection in *Arabidopsis thaliana*. *Mol. Biol. Evol.* **31**(11): 3026-3039

Hudson RR (1983) Properties of a Neutral Allele Model with Intragenic Recombination. *Theoretical Population Biology* **23**: 183-201

Hwang C-FF & Williamson VM (2003) Leucine-rich repeat-mediated intramolecular interactions in nematode recognition and cell death signaling by the tomato resistance protein Mi. *Plant J.* **34:** 585–589

Igari K, Endo S, Hibara K, Aida M, Sakakibara H, Kawasaki T & Tasaka M (2008) Constitutive activation of a CC-NB-LRR protein alters morphogenesis through the cytokinin pathway in Arabidopsis. *Plant J.* **55:** 14–27

Ikeda S, Anda M, Inaba S, Eda S, Sato S, Sasaki K, Tabata S, Mitsui H, Sato T, Shinano T & Minamisawa K (2011) Autoregulation of nodulation interferes with impacts of nitrogen fertilization levels on the leaf-associated bacterial community in soybeans. *Appl. Environ. Microbiol.* **77:** 1973–1980

Innan H (2003) A two-locus gene conversion model with selection and its application to the human RHCE and RHD genes. *Proc. Natl. Acad. Sci. U.S.A.* **100:** 8793–8798

Innan H (2003) The coalescent and infinite-site model of a small multigene family. *Genetics* **163:** 803–810

Innerebner G, Knief C & Vorholt JA (2011) Protection of Arabidopsis thaliana against leaf-pathogenic Pseudomonas syringae by Sphingomonas strains in a controlled model system. *Appl. Environ. Microbiol.* **77:** 3202–3210

Jensen M, Hagedorn P, Torres-Zabala M, Grant M, Rung J, Collinge D, Lyngkjaer M **(**2008) Transcriptional regulation by an NAC (NAM-ATAF1,2-CUC2) transcription factor attenuates ABA signalling for efficient basal defence towards *Blumeria graminis* f. sp. *hordei* in *Arabidopsis*. *Plant J. Cell Mol. Biology* **56**: 867–880.

Jia G, Fu Y & He C (2012) Reversible RNA adenosine methylation in biological regulation. *Trends Genetics* **29**

Jia G, Fu Y, Zhao X, Dai Q, Zheng G, Yang Y, Yi C, Lindahl T, Pan T, Yang Y-G & He C (2011) N6-Methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nat. Chem. Biology* **7:** 885–887

Jones J, Haegeman A, Danchin E, Gaur H, Helder J, Jones M, Kikuchi T, Manzanilla-López R, Palomares-Rius J, Wesemael W & Perry R (2013) Top 10 plant-parasitic nematodes in molecular plant pathology. *Mol. Plant Pathology* **14:** 946–961

Juntawong P & Bailey-Serres J (2012) Dynamic Light Regulation of Translation Status in *Arabidopsis thaliana*. *Front. Plant Sci.* **3:** 66

Kadivar H & Stapleton AE (2003) Ultraviolet radiation alters maize phyllosphere bacterial diversity. *Microb. Ecol.* **45:** 353–61

Kamoun S, Furzer O, Jones J, Judelson H, Ali G, Dalio R, Roy S, Schena L, Zambounis A, Panabières F, Cahill D, Ruocco M, Figueiredo A, Chen X, Hulvey J, Stam R, Lamour K, Gijzen M, Tyler B, Grünwald N, *et al.,* (2015) The Top 10 oomycete pathogens in molecular plant pathology. *Mol. Plant Pathology* **16:** 413–434

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

Karasov T, Kniskern J, Gao L, DeYoung B, Ding J, Dubiella U, Lastra R, Nallu S, Roux F, Innes R, Barrett L, Hudson R & Bergelson J (2014) The long-term maintenance of a resistance polymorphism through diffuse interactions. *Nature* **512:** 436–40

Kato H, Saito T, Ito H, Komeda Y & Kato A (2014) Overexpression of the TIR-X gene results in a dwarf phenotype and activation of defense-related gene expression in Arabidopsis thaliana. *J. Plant Physiol.***171:** 382–388

Kato H, Shida T, Komeda Y, Saito T & Kato A (2011) Overexpression of the Activated Disease Resistance 1-like1 (ADR1-L1) Gene Results in a Dwarf Phenotype and Activation of Defense-Related Gene Expression in Arabidopsis thaliana. *Journal of Plant Biology* **54:** 172-179

Kim J, Lim CJ, Lee B-WW, Choi J-PP, Oh S-KK, Ahmad R, Kwon S-YY, Ahn J & Hur C-GG (2012) A genome-wide comparison of NB-LRR type of resistance gene analogs (RGA) in the plant kingdom. *Mol. Cells* **33:** 385–392

Kim S, Plagnol V, Hu T, Toomajian C, Clark R, Ossowski S, Ecker J, Weigel D & Nordborg M (2007) Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nat. Genet.* **39:** 1151–1155

Kimura M & Weiss G (1964) The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance. *Genetics* **49:** 561–576

Korves T & Bergelson J (2004) A novel cost of R gene resistance in the presence of disease. *Am. Nat.***163:** 489–504

Korves TM & Bergelson J (2003) A developmental response to pathogen infection in Arabidopsis. *Plant Physiol.* **133:** 339–347

Krug R, Morgan M & Shatkin A (1976) Influenza viral mRNA contains internal N6-methyladenosine and 5'-terminal 7-methylguanosine in cap structures. *J. Virology* **20:** 45–53

Kuang H, Caldwell K, Meyers B & Michelmore R (2008) Frequent sequence exchanges between homologs of RPP8 in Arabidopsis are not necessarily associated with genomic proximity. *Plant J.* **54:** 69–80

Kunkel B, Bent A, Dahlbeck D, Innes R & Staskawicz B (1993) RPS2, an Arabidopsis Disease Resistance Locus Specifying Recognition of *Pseudomonas syringae* Strains Expressing the Avirulence Gene *avrRpt2*.*Plant Cell Online* **5:** 865–887

Lamesch P, Berardini T, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander D, Garcia-Hernandez M, Karthikeyan A, Lee C, Nelson W, Ploetz L, Singh S, Wensel A & Huala E (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40:** D1202–D1210

Langley C, Lazzaro B, Phillips W, Heikkinen E & Braverman J (2000) Linkage disequilibria and the site frequency spectra in the su(s) and su(w(a)) regions of the Drosophila melanogaster X chromosome. *Genetics* **156:** 1837–1852

Langmead B, Salzberg S (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**:357-359

Lasky J, Upadhyaya H, Ramu P, *et al*. (2015) Genome-environment associations in sorghum landraces predict adaptive traits. *Sci Adv* 1: e1400218

Leister D (2004) Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. *Trends in Genetics* Available at: http://www.sciencedirect.com/science/article/pii/S0168952504000228

Levis R, Penman S (1978) 5′-terminal structures of poly(A)+ cytoplasmic messenger RNA and of poly(A)+ and poly(A)-heterogeneous nuclear RNA of cells of the dipteran Drosophila melanogaster. *Journal of molecular biology.* **120**: 487–515

Li Y, Fan C, Xing Y, *et al.* (2011) Natural variation in GS5 plays an important role in regulating grain size and yield in rice. *Nat. Genet.* **43**: 1266-1269

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

Liu J, Yue Y, Han D, Wang X, Fu Y, Zhang L, Jia G, Yu M, Lu Z, Deng X, Dai Q, Chen W & He C (2014) A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. *Nat. Chem. Biology* **10:** 93–95

Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, Platzer A, Zhang Q, Vilhjálmsson BJ, Korte A, Nizhynska V, Voronin V, Korte P, Sedman L, Mandáková T, Lysak MA, Seren U, Hellmann I & Nordborg M (2013) Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nat. Genet.* **45**(8): 884-890

Luo, GZ, MacQueen, A, Zheng, G, Duan, H, Dore, L, Lu, Z, Liu, J, Jia, G, Bergelson, J & He, C (2014) Unique features of the $m^6A$ methylome in *Arabidopsis thaliana*. *Nat.Comm.* **5**:5630

Machnicka M, Milanowska K, Oglou O, Purta E, Kurkowska M, Olchowik A, Januszewski W, Kalinowski S, Dunin-Horkawicz S, Rother K, Helm M, Bujnicki J & Grosjean H (2013) MODOMICS: a database of RNA modification pathways—2013 update. *Nucleic Acids Res.* **41:**D262–D267

Mansai S & Innan H (2010) The Power of the Methods for Detecting Interlocus Gene Conversion. *Genetics* **184:** 517–527

Mansfield J, Genin S, Magori S, Citovsky V, Sriariyanum M, Ronald P, Dow M, Verdier V, Beer S, Machado M, Toth I, Salmond G & Foster G (2012) Top 10 plant pathogenic bacteria in molecular plant pathology. *Mol. Plant Pathology* **13:** 614–629

Mauricio R, Stahl EA, Korves T, Tian D, Kreitman M & Bergelson J (2003) Natural selection for polymorphism in the disease resistance gene *Rps2* of *Arabidopsis thaliana*. *Genetics* **163:** 735–746

McDonald B & Linde C (2002) Pathogen population genetics, evolutionary potential, and durable resistance. *Annual review of phytopathology* **40:** 349–379

McDowell JM, Dhandaydham M, Long TA, Aarts MG, Goff S, Holub EB & Dangl JL (1998) Intragenic recombination and diversifying selection contribute to the evolution of downy mildew resistance at the RPP8 locus of Arabidopsis. *Plant Cell* **10:** 1861–1874

McHale L, Tan X, Koehl P & Michelmore RW (2006) Plant NBS-LRR proteins: adaptable guards. *Genome Biol.* **7:** 212

Meagher RB, Berry-Lowe S & Rice K (1989) Molecular evolution of the small subunit of ribulose bisphosphate carboxylase: nucleotide substitution and gene conversion. *Genetics* **123:** 845–863

Mendes R, Kruijt M, de Bruijn I, Dekkers E, van der Voort M, Schneider JH, Piceno YM, DeSantis TZ, Andersen GL, Bakker PA & Raaijmakers JM (2011) Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science* **332:** 1097–1100

Meyer K, Saletore Y, Zumbo P, Elemento O, Mason C & Jaffrey S (2012) Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell* **149:** 1635–1646

Meyer KD, Jaffrey SR (2014) The dynamic epitranscriptome: N6-methyladenosine and gene expression control. *Nature reviews. Molecular cell biology*. **15**:313–326

Meyers B, Chin D, Shen K, Sivaramakrishnan S, Lavelle D, Zhang Z & Michelmore R (1998) The major resistance gene cluster in lettuce is highly duplicated and spans several megabases. *Plant Cell***10:** 1817–1832

Meyers B, Morgante M & Michelmore R (2002) TIR-X and TIR-NBS proteins: two new families related to disease resistance TIR-NBS-LRR proteins encoded in Arabidopsis and other plant genomes. *Plant J.* **32:** 77–92

Meyers BC, Kozik A, Griego A, Kuang H & Michelmore RW (2003) Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. *Plant Cell* **15:** 809–834

Michelmore RW & Meyers BC (1998) Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res.* **8:** 1113–1130

Mindrinos M, Katagiri F, Yu GL & Ausubel FM (1994) The *A. thaliana* disease resistance gene RPS2 encodes a protein containing a nucleotide-binding site and leucine-rich repeats. *Cell* **78:** 1089–1099

Monde R, Schuster G & Stern D (2000) Processing and degradation of chloroplast mRNA. *Biochimie* **82:** 573–582

Mondragón-Palomino M, Meyers B, Michelmore R & Gaut B (2002) Patterns of positive selection in the complete NBS-LRR gene family of *Arabidopsis thaliana*. *Genome Res.* **12:** 1305–1315

Monteiro R, Balsanelli E, Wassem R, Marin A, Brusamarello-Santos L, Schmidt M, Tadra-Sfeir M, Pankievicz V, Cruz L, Chubatsu L, Pedrosa F & Souza E (2012) Herbaspirillum-plant interactions: microscopical, histological and molecular aspects. *Plant and Soil* **356**

Nichols J (1979) 'Cap' structures in maize poly(A)-containing RNA.*Biochimica Et Biophysica Acta* **563:** 490–495

Nilsen TW (2014) Internal mRNA methylation finally finds functions. *Science*. **343**: 1207–1208

Nishimura MT, Stein M, Hou B-HH, Vogel JP, Edwards H, Somerville SC (2003) Loss of a callose synthase results in salicylic acid-dependent disease resistance. *Science* **301**: 969–972

Nobuta K, Ashfield T, Kim S & Innes RW (2005) Diversification of non-TIR class NB-LRR genes in relation to whole-genome duplication events in Arabidopsis. *Mol. Plant Microbe Interact.* **18:** 103–109

Nordborg M (2000) Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics* **154:** 923–929

Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R, Jakobsson M, Kim S, Morozov Y, Padhukasahasram B, Plagnol V,

Rosenberg NA, Shah C, Wall JD, Wang J, Zhao K, *et al.,* (2005) The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* **3:** e196

Oñate-Sánchez L, Vicente-Carbajosa J (2008) DNA-free RNA isolation protocols for Arabidopsis thaliana, including seeds and siliques. *BMC research notes*. **1**: 93

Palma K, Thorgrimsen S, Malinovsky FG, Fiil BK, Nielsen HB, Brodersen P, Hofius D, Petersen M & Mundy J (2010) Autoimmunity in Arabidopsis *acd11* is mediated by epigenetic regulation of an immune receptor. *PLoS Pathog.* **6:** e1001137

Pandey N, Ranjan A, Pant P, Tripathi R, Ateek F, Pandey H, Patre U, Sawant S (2013) CAMTA 1 regulates drought responses in *Arabidopsis thaliana*. *Bmc Genom*. **14**: 216

Parniske M & Jones J (1999) Recombination between diverged clusters of the tomato Cf-9 plant disease resistance gene family.*Proc. Natl. Acad. Sci.* **96:** 5850–5585

Parniske M, Hammond-Kosack KE, Golstein C, Thomas CM, Jones DA, Harrison K, Wulff BB & Jones JD (1997) Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the Cf-4/9 locus of tomato. *Cell* **91**:821-832

Perazzolli M, Malacarne G, Baldo A, Righetti L, Bailey A, Fontana P, Velasco R & Malnoy M (2014) Characterization of Resistance Gene Analogues (RGAs) in Apple (Malus × domestica Borkh.) and Their Evolutionary History of the Rosaceae Family. *PLoS ONE* **9**(2): e83844

Phifer-Rixey M, Bomhoff M & Nachman M (2014) Genome-Wide Patterns of Differentiation Among House Mouse Subspecies. *Genetics* **198:** 283–297

Ping X-L, Sun B-F, Wang L, Xiao W, Yang X, Wang W-J, Adhikari S, Shi Y, Lv Y, Chen Y-S, Zhao X, Li A, Yang Y, Dahal U, Lou X-M, Liu X, Huang J, Yuan W-P, Zhu X-F, Cheng T, *et al.,* (2014) Mammalian WTAP is a regulatory subunit of the RNA N6-methyladenosine methyltransferase. *Cell Res.* **24:** 177–189

Platt A, Horton M, Huang YS, Li Y, Anastasio AE, Mulyati NW, Agren J, Bossdorf O, Byers D, Donohue K, Dunning M, Holub EB, Hudson A, Le Corre V, Loudet O, Roux F, Warthmann N, Weigel D, Rivero L, Scholl R, *et al.,* (2010) The scale of population structure in Arabidopsis thaliana. *PLoS Genet.* **6:** e1000843

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features.*Bioinformatics*. **26**: 841–842

Raj S, Rifkin S, Andersen E, & Oudenaarden A (2010) Variability in gene expression underlies incomplete penetrance. *Nature* **463**: 913-919

Rausch T, Zichner T, Schlattl A, Stütz A, Benes V & Korbel J (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinform.* **28:** i333–i339

Rieu I & Powers SJ (2009) Real-time quantitative RT-PCR: design, calculations, and statistics. *Plant Cell* **21:** 1031–1033

Rose LE, Bittner-Eddy PD, Langley CH, Holub EB, Michelmore RW, Beynon JL (2004) The maintenance of extreme amino acid diversity at the disease resistance gene, *RPP13*, in *Arabidopsis thaliana*. *Genetics* **166:** 1517-1527

Sandermann H (2000) Ozone/biotic disease interactions: molecular biomarkers as a new experimental tool. *Environ. Pollut.* **108:** 327–332

Schibler U, Kelley D & Perry R (1977) Comparison of methylated sequences in messenger RNA and heterogeneous nuclear RNA from mouse L cells. *J. Mol. Biology* **115:** 695–714

Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, Alix A, McCosh RB, Chen H, Schork NJ & Ecker JR (2013) Patterns of population epigenomic diversity. *Nature* **495:** 193-198

Schmutz J, McClean PE, Mamidi S, *et al.,* (2014) A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.* **46:** 707–713

Scholthof K, Adkins S, Czosnek H, Palukaitis P, Jacquot E, Hohn T, Hohn R, Saunders K, Candresse T, Ahlquist P, Hemenway C & Foster G (2011) Top 10 plant viruses in molecular plant pathology. *Molecular Plant Pathology* **12:** 938–954

Schwartz S, Agarwala S, Mumbach M, Jovanovic M, Mertins P, Shishkin A, Tabach Y, Mikkelsen T, Satija R, Ruvkun G, Carr S, Lander E, Fink G & Regev A (2013) High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. *Cell* **155:** 1409–1421

Shah J, Kachroo P & Klessig DF (1999) The Arabidopsis *ssi1* mutation restores pathogenesis-related gene expression in npr1 plants and renders defensin gene expression salicylic acid dependent. *Plant Cell* **11:** 191–206

Sharbel TF, Haubold B & Mitchell-Olds T (2000) Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. *Mol. Ecol.* **9:** 2109–2118

Shea K & Chesson P (2002) Community ecology theory as a framework for biological invasions. *Trends in Ecology & Evolution* **17:** 170–176

Sher A & Hyatt L (1999) The disturbed resource-flux invasion matrix: a new framework for patterns of plant invasion. *Biological Invasions* **1**

Shirano (2002) 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**

Soll J & Schleiff E (2004) Protein import into chloroplasts. *Nat. Rev. Mol. Cell Biology* **5:** 198–208

Spiess A-NN, Feig C & Ritz C (2008) Highly accurate sigmoidal fitting of real-time PCR data by introducing a parameter for asymmetry. *BMC Bioinformatics* **9:** 221

Stahl EA, Dwyer G, Mauricio R, Kreitman M & Bergelson J (1999) Dynamics of disease resistance polymorphism at the *Rpm1* locus of Arabidopsis. *Nature* **400:** 667–71

Stark A, Brennecke J, Bushati N, Russell R & Cohen S (2005) Animal MicroRNAs Confer Robustness to Gene Expression and Have a Significant Impact on 3′UTR Evolution. *Cell* **123**

Stokes TL, Kunkel BN & Richards EJ (2002) Epigenetic variation in Arabidopsis disease resistance. *Genes Dev.* **16:** 171–82

Studer A, Zhao Q, Ross-Ibarra J, Doebley J (2011) Identification of a functional transposon insertion in the maize domestication gene tb1. *Nat. Genet.* **43**:1160-1163

Suda W, Nagasaki A & Shishido M (2009) Powdery Mildew-Infection Changes Bacterial Community Composition in the Phyllosphere. *Microbes and Environments* **24**

Sugio A, Dreos R, Aparicio F & Maule A (2009) The Cytosolic Protein Response as a Subcomponent of the Wider Heat Shock Response in Arabidopsis. *Plant Cell Online* **21**: 642–654

Suzuki N, Rivero R, Shulaev V, Blumwald E, & Mittler R (2014) Abiotic and biotic stress combinations. *New Phytologist* **203**: 32-43

Swofford, DL (2003) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Szittya G, Silhavy D, Molnár A, Havelda Z, Lovas A, Lakatos L, Bánfalvi Z & Burgyán J (2003) Low temperature inhibits RNA silencing-mediated defence by the control of siRNA generation. *EMBO J.* **22:** 633–640

Takahashi H, Miller J, Nozaki Y, Sukamto1 S, Takeda M, Shah J, Hase S, Ikegami M, Ehara Y & Dinesh-Kumar S (2002) RCY1, an *Arabidopsis thaliana* RPP8/HRT family resistance gene, conferring resistance to cucumber mosaic virus requires salicylic acid, ethylene and a novel signal transduction mechanism. *Plant J* **32:** 655–667

Tang L, Bhat S & Petracek M (2003) Light control of nuclear gene mRNA abundance and translation in tobacco. *Plant Physiology* **133:** 1979–90

Tao Y, Yuan F, Leister RT, Ausubel FM & Katagiri F (2000) Mutational analysis of the Arabidopsis nucleotide binding site-leucine-rich repeat resistance gene RPS2. *Plant Cell* **12:** 2541–2554

Teshima K & Innan H (2004) The Effect of Gene Conversion on the Divergence Between Duplicated Genes. *Genetics* **166:** 1553–1560

Teshima K & Innan H (2012) The Coalescent with Selection on Copy Number Variants. *Genetics* **190:** 1077–86

The Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. *Nucl. Acids Res.* **43**: D1049–D1056

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

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

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols.* **7**: 562–578

Turner TR, James EK & Poole PS (2013) The plant microbiome. *Genome Biol.* **14:** 209

Underwood W, Melotto M & He SY (2007) Role of plant stomata in bacterial invasion. *Cell. Microbiol.* **9:** 1621–1629

Vainonen J, Hansson M & Vener A (2005) STN8 Protein Kinase in *Arabidopsis thaliana* Is Specific in Phosphorylation of Photosystem II Core Proteins. *J. Biological Chem.* **280:** 33679–33686

Van Aken O, Zhang B, Carrie C, Uggalla V, Paynter E, Giraud E, Whelan J (2009) Defining the Mitochondrial Stress Response in *Arabidopsis thaliana*. *Mol. Plant* **2**: 1310–1324

Van der Hoorn R, de Wit P & Joosten M (2002) Balancing selection favors guarding resistance proteins. *Trends Plant Sci.* **7:** 67–71

Van Leeuwen H, Kliebenstein D, West M, Kim K, Poecke R, Katagiri F, Michelmore R, Doerge R, Clair D **(**2007) Natural variation among *Arabidopsis thaliana* accessions for transcriptome response to exogenous salicylic acid. *Plant Cell* **19**: 2099–110

Van Poecke RM, Sato M, Lenarz-Wyatt L, Weisberg S & Katagiri F (2007) Natural variation in RPS2-mediated resistance among *Arabidopsis* accessions: correlation between gene expression profiles and phenotypic responses. *Plant Cell* **19:** 4046–60

Van Valen L (1973) A New Evolutionary Law. *Evolutionary Theory* **1**: 1-30

Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A & Speleman F (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* **3:** RESEARCH0034

Vanhaute E, Paping R., & Grada C (2006) The European subsistence crisis of 1845-1850: a comparative perspective. IEHC Helsinki: **123**

Wang D, Amornsiripanitch N, & Dong X (2006) A Genomic Approach to Identify Regulatory Nodes in the Transcriptional Network of Systemic Acquired Resistance in Plants. *PLoS Pathogens* **2**(11): e123

Wang X, Lu Z, Gomez A, Hon G, Yue Y, Han D, Fu Y, Parisien M, Dai Q, Jia G, Ren B, Pan T & He C (2014) N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature* **505:** 117–120

Wang Y, Li Y, Toth J, Petroski MD, Zhang Z, Zhao JC (2014) N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nature cell biology*. **16**: 191–198

Wei CM, Gershowitz A, Moss B (1975) Methylated nucleotides block 5′ terminus of HeLa cell messenger RNA. *Cell* **4**:379–386

Weigel D & Mott R (2009) The 1001 Genomes Project for *Arabidopsis thaliana*. *Genome Biol* **10:** 107

Wiuf C & Hein J (2000) The coalescent with gene conversion. *Genetics* **155:** 451–62

Wu H-J, Ma Y-K, Chen T, Wang M & Wang X-J (2012) PsRobot: a web-based plant small RNA meta-analysis toolbox. *Nucleic Acids Res.* **40:** W22–W28

Wulff BB, Thomas CM, Smoker M, Grant M & Jones JD (2001) Domain swapping and gene shuffling identify sequences required for induction of an Avr-dependent hypersensitive response by the tomato Cf-4 and Cf-9 proteins. *Plant Cell* **13:** 255–272

Xiao S, Brown S, Patrick E, Brearley C, & Turner J (2002) Enhanced transcription of the *Arabidopsis* disease resistance genes *RPW8.1* and *RPW8.2* via a salicylic acid-dependent amplification circuit is required for hypersensitive cell death. *The Plant Cell* **15**: 33–45

Yi SY, Kim J-HH, Joung Y-HH, Lee S, Kim W-TT, Yu SH & Choi D (2004) The pepper transcription factor CaPF1 confers pathogen and freezing tolerance in *Arabidopsis*. *Plant Physiol.* **136:** 2862–2874

Yu GL, Katagiri F, Ausubel F (1993) *Arabidopsis* mutations at the RPS2 locus result in loss of resistance to *Pseudomonas syringae* strains expressing the avirulence gene avrRpt2. *MPMI* **6**(4): 434-443

Yutthammo C, Thongthammachat N, Pinphanichakarn P & Luepromchai E (2010) Diversity and activity of PAH-degrading bacteria in the phyllosphere of ornamental plants. *Microb. Ecol.* **59:** 357–368

Zhai J, Jeong D-HH, De Paoli E, Park S, Rosen BD, Li Y, González AJ, Yan Z, Kitto SL, Grusak MA, Jackson SA, Stacey G, Cook DR, Green PJ, Sherrier DJ & Meyers BC (2011) MicroRNAs as master regulators of the plant NB-LRR defense gene family via the production of phased, trans-acting siRNAs. *Genes Dev.* **25:** 2540–2553

Zhang X-CC & Gassmann W (2007) Alternative splicing and mRNA levels of the disease resistance gene RPS4 are induced during defense responses. *Plant Physiol.* **145:** 1577–1587

Zhang Z, Li Q, Li Z, Staswick P, Wang M, Zhu Y, He Z (2007) Dual regulation role of GH3.5 in salicylic acid and auxin signaling during *Arabidopsis-Pseudomonas* syringae interaction. *Plant Physiology* **145**: 450–464

Zheng G, Dahl J, Niu Y, Fedorcsak P, Huang C-M, Li C, Vågbø C, Shi Y, Wang W-L, Song S-H, Lu Z, Bosmans R, Dai Q, Hao Y-J, Yang X, Zhao W-M, Tong W-M, Wang X-J, Bogdan F, Furu K, *et al.,* (2012) ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. *Mol. Cell* **49:** 18–29

Zhong S, Li H, Bodi Z, Button J, Vespa L, Herzog M & Fray R (2008) MTA Is an *Arabidopsis* Messenger RNA Adenosine Methylase and Interacts with a Homolog of a Sex-Specific Splicing Factor. *Plant Cell Online* **20:** 1278–1288

Zhou T, Wang Y, Chen J-Q, Araki H, Jing Z, Jiang K, Shen J & Tian D (2004) Genome-wide identification of NBS genes in japonica rice reveals significant expansion of divergent non-TIR NBS-LRR genes. *Mol. Genetics Genom.* **271:** 402–415

Zipfel C, Robatzek S, Navarro L, Oakeley EJ, Jones JD, Felix G & Boller T (2004) Bacterial disease resistance in *Arabidopsis* through flagellin perception. *Nature* **428:** 764–777

## Appendix A

## Additional Data, Chapter 2

**List A.1.** Genes considered in the metastudy and the gene set they were in.

| Gene | Gene Set | Gene | Gene Set | Gene | Gene Set |
|---|---|---|---|---|---|
| AT1G01040 | Stress response gene | AT1G08170 | Control gene | AT4G26090 | *R*-gene |
| AT1G02730 | Stress response gene | AT1G08210 | Control gene | AT1G10920 | *R*-gene |
| AT1G04810 | Stress response gene | AT1G08270 | Control gene | AT1G12210 | *R*-gene |
| AT1G06390 | Stress response gene | AT1G08280 | Control gene | AT1G12220 | *R*-gene |
| AT1G08600 | Stress response gene | AT1G10940 | Control gene | AT1G12280 | *R*-gene |
| AT1G09970 | Stress response gene | AT1G13320 | Control gene | AT1G12290 | *R*-gene |
| AT1G11820 | Stress response gene | AT1G16890 | Control gene | AT1G15890 | *R*-gene |
| AT1G13260 | Stress response gene | AT1G24140 | Control gene | AT1G33560 | *R*-gene |
| AT1G14755 | Stress response gene | AT1G24150 | Control gene | AT1G50180 | *R*-gene |
| AT1G16130 | Stress response gene | AT1G33580 | Control gene | AT1G53350 | *R*-gene |
| AT1G17860 | Stress response gene | AT1G34540 | Control gene | AT1G58390 | *R*-gene |
| AT1G19380 | Stress response gene | AT1G42410 | Control gene | AT1G58410 | *R*-gene |
| AT1G21710 | Stress response gene | AT1G50350 | Control gene | AT1G59124 | *R*-gene |
| AT1G24000 | Stress response gene | AT1G52180 | Control gene | AT1G59218 | *R*-gene |
| AT1G26700 | Stress response gene | AT1G53300 | Control gene | AT1G59620 | *R*-gene |
| AT1G28680 | Stress response gene | AT1G53340 | Control gene | AT1G59780 | *R*-gene |
| AT1G31180 | Stress response gene | AT1G53340 | Control gene | AT1G61180 | *R*-gene |
| AT1G33060 | Stress response gene | AT1G53440 | Control gene | AT1G61190 | *R*-gene |
| AT1G37130 | Stress response gene | AT1G54510 | Control gene | AT1G61300 | *R*-gene |
| AT1G47600 | Stress response gene | AT1G54620 | Control gene | AT1G61310 | *R*-gene |
| AT1G50140 | Stress response gene | AT1G58050 | Control gene | AT1G62630 | *R*-gene |
| AT1G51790 | Stress response gene | AT1G58150 | Control gene | AT1G63350 | *R*-gene |
| AT1G53240 | Stress response gene | AT1G58160 | Control gene | AT3G07040 | *R*-gene |
| AT1G55010 | Stress response gene | AT1G58450 | Control gene | AT3G14460 | *R*-gene |
| AT1G56650 | Stress response gene | AT1G60830 | Control gene | AT3G14470 | *R*-gene |
| AT1G59870 | Stress response gene | AT1G61370 | Control gene | AT3G15700 | *R*-gene |
| AT1G61360 | Stress response gene | AT1G61750 | Control gene | AT3G46530 | *R*-gene |
| AT1G63522 | Stress response gene | AT1G61760 | Control gene | AT3G46710 | *R*-gene |
| AT1G64670 | Stress response gene | AT1G61770 | Control gene | AT3G46730 | *R*-gene |
| AT1G66280 | Stress response gene | AT1G61850 | Control gene | AT3G50950 | *R*-gene |
| AT1G68090 | Stress response gene | AT1G61860 | Control gene | AT4G10780 | *R*-gene |
| AT1G69828 | Stress response gene | AT1G61860 | Control gene | AT4G14610 | *R*-gene |
| AT1G71220 | Stress response gene | AT1G61870 | Control gene | AT4G19050 | *R*-gene |
| AT1G72850 | Stress response gene | AT1G61880 | Control gene | AT4G27190 | *R*-gene |

| | | | | | |
|---|---|---|---|---|---|
| AT1G73805 | Stress response gene | AT1G61970 | Control gene | AT4G27220 | *R*-gene |
| AT1G75560 | Stress response gene | AT1G61980 | Control gene | AT4G33300 | *R*-gene |
| AT1G77120 | Stress response gene | AT1G62090 | Control gene | AT5G04720 | *R*-gene |
| AT1G78600 | Stress response gene | AT1G65300 | Control gene | AT5G05400 | *R*-gene |
| AT1G80460 | Stress response gene | AT1G66550 | Control gene | AT5G35450 | *R*-gene |
| AT2G02470 | Stress response gene | AT1G67810 | Control gene | AT5G43470 | *R*-gene |
| AT2G03955 | Stress response gene | AT1G67840 | Control gene | AT5G43730 | *R*-gene |
| AT2G06530 | Stress response gene | AT2G10090 | Control gene | AT5G43740 | *R*-gene |
| AT2G15080 | Stress response gene | AT2G21030 | Control gene | AT5G45510 | *R*-gene |
| AT2G17380 | Stress response gene | AT2G21040 | Control gene | AT5G47250 | *R*-gene |
| AT2G18880 | Stress response gene | AT3G04050 | Control gene | AT5G47260 | *R*-gene |
| AT2G20465 | Stress response gene | AT3G09240 | Control gene | AT5G47280 | *R*-gene |
| AT2G21870 | Stress response gene | AT3G18160 | Control gene | AT5G48620 | *R*-gene |
| AT2G23620 | Stress response gene | AT3G18180 | Control gene | AT5G63020 | *R*-gene |
| AT2G25620 | Stress response gene | AT3G19700 | Control gene | AT5G66630 | *R*-gene |
| AT2G27145 | Stress response gene | AT3G26540 | Control gene | AT1G17600 | *R*-gene |
| AT2G29045 | Stress response gene | AT3G42530 | Control gene | AT1G27170 | *R*-gene |
| AT2G31180 | Stress response gene | AT3G44410 | Control gene | AT1G27180 | *R*-gene |
| AT2G32730 | Stress response gene | AT3G44490 | Control gene | AT1G31540 | *R*-gene |
| AT2G34600 | Stress response gene | AT3G44640 | Control gene | AT1G56510 | *R*-gene |
| AT2G36530 | Stress response gene | AT3G44680 | Control gene | AT1G56520 | *R*-gene |
| AT2G38560 | Stress response gene | AT3G48790 | Control gene | AT1G56540 | *R*-gene |
| AT2G40030 | Stress response gene | AT3G48810 | Control gene | AT1G63730 | *R*-gene |
| AT2G41370 | Stress response gene | AT3G53060 | Control gene | AT1G63740 | *R*-gene |
| AT2G43090 | Stress response gene | AT3G53590 | Control gene | AT1G63750 | *R*-gene |
| AT2G44920 | Stress response gene | AT3G53600 | Control gene | AT1G63860 | *R*-gene |
| AT2G46450 | Stress response gene | AT4G08970 | Control gene | AT1G63870 | *R*-gene |
| AT2G48150 | Stress response gene | AT4G09360 | Control gene | AT1G63880 | *R*-gene |
| AT3G02400 | Stress response gene | AT4G10200 | Control gene | AT1G64070 | *R*-gene |
| AT3G03710 | Stress response gene | AT4G10220 | Control gene | AT1G65850 | *R*-gene |
| AT3G04945 | Stress response gene | AT4G10460 | Control gene | AT1G69550 | *R*-gene |
| AT3G06050 | Stress response gene | AT4G11400 | Control gene | AT1G72840 | *R*-gene |
| AT3G08580 | Stress response gene | AT4G11470 | Control gene | AT1G72860 | *R*-gene |
| AT3G09940 | Stress response gene | AT4G16080 | Control gene | AT2G14080 | *R*-gene |
| AT3G11410 | Stress response gene | AT4G16530 | Control gene | AT2G17050 | *R*-gene |
| AT3G12810 | Stress response gene | AT4G16540 | Control gene | AT2G17060 | *R*-gene |
| AT3G14210 | Stress response gene | AT4G16550 | Control gene | AT3G04220 | *R*-gene |
| AT3G15950 | Stress response gene | AT4G16560 | Control gene | AT3G25510 | *R*-gene |
| AT3G17155 | Stress response gene | AT4G17860 | Control gene | AT3G44400 | *R*-gene |
| AT3G18910 | Stress response gene | AT4G17970 | Control gene | AT3G44480 | *R*-gene |
| AT3G20550 | Stress response gene | AT4G17980 | Control gene | AT3G44630 | *R*-gene |

| | | | | | |
|---|---|---|---|---|---|
| AT3G22780 | Stress response gene | AT4G18000 | Control gene | AT3G44670 | *R*-gene |
| AT3G23870 | Stress response gene | AT4G18020 | Control gene | AT3G51560 | *R*-gene |
| AT3G25265 | Stress response gene | AT4G18030 | Control gene | AT3G51570 | *R*-gene |
| AT3G27150 | Stress response gene | AT4G18040 | Control gene | AT4G08450 | *R*-gene |
| AT3G28910 | Stress response gene | AT4G18360 | Control gene | AT4G09360 | *R*-gene |
| AT3G44380 | Stress response gene | AT4G18600 | Control gene | AT4G09430 | *R*-gene |
| AT3G46630 | Stress response gene | AT4G26890 | Control gene | AT4G11170 | *R*-gene |
| AT3G48190 | Stress response gene | AT4G26920 | Control gene | AT4G12010 | *R*-gene |
| AT3G49810 | Stress response gene | AT4G30090 | Control gene | AT4G12020 | *R*-gene |
| AT3G51450 | Stress response gene | AT4G32220 | Control gene | AT4G14370 | *R*-gene |
| AT3G52930 | Stress response gene | AT4G32230 | Control gene | AT4G16860 | *R*-gene |
| AT3G55320 | Stress response gene | AT4G38070 | Control gene | AT4G16890 | *R*-gene |
| AT3G57130 | Stress response gene | AT4G38300 | Control gene | AT4G16900 | *R*-gene |
| AT3G59660 | Stress response gene | AT5G03720 | Control gene | AT4G16920 | *R*-gene |
| AT3G61190 | Stress response gene | AT5G04400 | Control gene | AT4G16940 | *R*-gene |
| AT3G66652 | Stress response gene | AT5G13290 | Control gene | AT4G16950 | *R*-gene |
| AT4G02070 | Stress response gene | AT5G14300 | Control gene | AT4G16960 | *R*-gene |
| AT4G03560 | Stress response gene | AT5G14400 | Control gene | AT4G19500 | *R*-gene |
| AT4G08170 | Stress response gene | AT5G14490 | Control gene | AT4G19510 | *R*-gene |
| AT4G09984 | Stress response gene | AT5G14560 | Control gene | AT4G19520 | *R*-gene |
| AT4G11290 | Stress response gene | AT5G14980 | Control gene | AT4G19530 | *R*-gene |
| AT4G12880 | Stress response gene | AT5G14990 | Control gene | AT4G36140 | *R*-gene |
| AT4G14630 | Stress response gene | AT5G15000 | Control gene | AT4G36150 | *R*-gene |
| AT4G16845 | Stress response gene | AT5G18690 | Control gene | AT5G11250 | *R*-gene |
| AT4G17750 | Stress response gene | AT5G31927 | Control gene | AT5G17680 | *R*-gene |
| AT4G19530 | Stress response gene | AT5G32450 | Control gene | AT5G17880 | *R*-gene |
| AT4G21600 | Stress response gene | AT5G36270 | Control gene | AT5G17890 | *R*-gene |
| AT4G23150 | Stress response gene | AT5G36280 | Control gene | AT5G17970 | *R*-gene |
| AT4G24220 | Stress response gene | AT5G36780 | Control gene | AT5G18350 | *R*-gene |
| AT4G25200 | Stress response gene | AT5G40470 | Control gene | AT5G18360 | *R*-gene |
| AT4G26780 | Stress response gene | AT5G40730 | Control gene | AT5G18370 | *R*-gene |
| AT4G29033 | Stress response gene | AT5G40740 | Control gene | AT5G22690 | *R*-gene |
| AT4G30340 | Stress response gene | AT5G42260 | Control gene | AT5G36930 | *R*-gene |
| AT4G31910 | Stress response gene | AT5G42510 | Control gene | AT5G38340 | *R*-gene |
| AT4G33420 | Stress response gene | AT5G42840 | Control gene | AT5G38350 | *R*-gene |
| AT4G34710 | Stress response gene | AT5G43200 | Control gene | AT5G38850 | *R*-gene |
| AT4G36430 | Stress response gene | AT5G43260 | Control gene | AT5G40060 | *R*-gene |
| AT4G38130 | Stress response gene | AT5G43360 | Control gene | AT5G40100 | *R*-gene |
| AT4G39850 | Stress response gene | AT5G43370 | Control gene | AT5G40910 | *R*-gene |
| AT5G02840 | Stress response gene | AT5G43480 | Control gene | AT5G40920 | *R*-gene |
| AT5G04930 | Stress response gene | AT5G43490 | Control gene | AT5G41540 | *R*-gene |

| | | | | | |
|---|---|---|---|---|---|
| AT5G06720 | Stress response gene | AT5G43490 | Control gene | AT5G41550 | *R*-gene |
| AT5G08120 | Stress response gene | AT5G43520 | Control gene | AT5G41740 | *R*-gene |
| AT5G09978 | Stress response gene | AT5G43520 | Control gene | AT5G41750 | *R*-gene |
| AT5G11670 | Stress response gene | AT5G43530 | Control gene | AT5G44510 | *R*-gene |
| AT5G13680 | Stress response gene | AT5G43550 | Control gene | AT5G44870 | *R*-gene |
| AT5G15180 | Stress response gene | AT5G44060 | Control gene | AT5G45050 | *R*-gene |
| AT5G16990 | Stress response gene | AT5G44140 | Control gene | AT5G45060 | *R*-gene |
| AT5G18610 | Stress response gene | AT5G44250 | Control gene | AT5G45200 | *R*-gene |
| AT5G20230 | Stress response gene | AT5G44480 | Control gene | AT5G45210 | *R*-gene |
| AT5G22500 | Stress response gene | AT5G44700 | Control gene | AT5G45230 | *R*-gene |
| AT5G24530 | Stress response gene | AT5G44760 | Control gene | AT5G45240 | *R*-gene |
| AT5G26860 | Stress response gene | AT5G44770 | Control gene | AT5G45250 | *R*-gene |
| AT5G34850 | Stress response gene | AT5G44910 | Control gene | AT5G45260 | *R*-gene |
| AT5G37510 | Stress response gene | AT5G44930 | Control gene | AT5G46260 | *R*-gene |
| AT5G39740 | Stress response gene | AT5G45490 | Control gene | AT5G46270 | *R*-gene |
| AT5G41360 | Stress response gene | AT5G45530 | Control gene | AT5G46450 | *R*-gene |
| AT5G42980 | Stress response gene | AT5G45570 | Control gene | AT5G46470 | *R*-gene |
| AT5G44080 | Stress response gene | AT5G45610 | Control gene | AT5G46490 | *R*-gene |
| AT5G45220 | Stress response gene | AT5G45740 | Control gene | AT5G46510 | *R*-gene |
| AT5G46420 | Stress response gene | AT5G45750 | Control gene | AT5G46520 | *R*-gene |
| AT5G47390 | Stress response gene | AT5G46830 | Control gene | AT5G48770 | *R*-gene |
| AT5G48905 | Stress response gene | AT5G47170 | Control gene | AT5G49140 | *R*-gene |
| AT5G50720 | Stress response gene | AT5G47510 | Control gene | AT5G51630 | *R*-gene |
| AT5G52605 | Stress response gene | AT5G50620 | Control gene | AT5G58120 | *R*-gene |
| AT5G54230 | Stress response gene | AT5G62990 | Control gene | AT1G47370 | *R*-gene |
| AT5G55920 | Stress response gene | AT5G63070 | Control gene | | |
| AT5G57970 | Stress response gene | AT5G67630 | Control gene | | |
| AT5G59520 | Stress response gene | AT1G59830 | Control gene | | |
| AT5G61560 | Stress response gene | | | | |
| AT5G63320 | Stress response gene | | | | |
| AT5G64520 | Stress response gene | | | | |
| AT5G66130 | Stress response gene | | | | |

## Appendix B

## Additional Data, Chapter 3

**Table B.1.** Natural variation in the nucleotide sequences of the five alleles of *Rps2* used to create the isogenic allelic series. *Rps2*$^{R}$ and *Rps2*$^{pR}$ lines are resistant to *Pseudomonas syringae* pv. *avrRpt2*. Minor allele is bolded.

| | Position in Coding Sequence: | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 3 | 3 | 4 | 4 | 0 | 0 | 0 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 4 | 4 |
| **Clade** | **Allele (Line)** | 1 | 2 | 2 | 6 | 1 | 4 | 9 | 3 | 4 | 5 | 1 | 1 | 5 | 7 | 7 | 3 | 4 |
| | | 1 | 5 | 6 | 1 | 7 | 2 | 2 | 3 | 5 | 5 | 1 | 5 | 9 | 0 | 4 | 8 | 0 |
| R clade | Col-0 (*Rps2*$^{R}$) | A | G | A | C | G | T | A | C | A | T | C | C | T | C | T | T | G |
| R clade | Ct-0 (*Rps2*$^{R}$) | A | G | A | C | G | T | A | C | A | T | C | C | T | C | T | T | G |
| R clade | Ler-0 (*Rps2*$^{R}$) | A | G | A | **A** | G | **C** | A | C | A | T | C | C | T | C | T | T | G |
| R clade | Ws-0 (*Rps2*$^{pR}$) | A | G | A | C | **A** | T | A | C | A | T | C | C | T | **A** | T | T | G |
| S clade | Wu-0 (*Rps2*$^{S}$) | **G** | **A** | **C** | C | G | T | **T** | **T** | **T** | **C** | **T** | **A** | **C** | C | **C** | **C** | **A** |

| | Position in Coding Sequence: | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 4 | 4 | 5 | 5 | 5 | 6 | 6 | 6 | 6 | 6 | 6 | 8 | 8 | 9 | 9 | 9 | 9 |
| **Clade** | **Allele (Line)** | 5 | 5 | 4 | 6 | 8 | 3 | 3 | 6 | 6 | 8 | 9 | 0 | 1 | 1 | 2 | 2 | 3 |
| | | 7 | 8 | 8 | 9 | 6 | 2 | 4 | 3 | 5 | 9 | 8 | 1 | 5 | 1 | 3 | 6 | 1 |
| R clade | Col-0 (*Rps2*$^{R}$) | A | G | C | C | A | C | C | A | G | **A** | G | G | A | C | C | T | A |
| R clade | Ct-0 (*Rps2*$^{R}$) | A | G | C | C | A | C | C | A | G | G | G | G | A | C | C | T | A |
| R clade | Ler-0 (*Rps2*$^{R}$) | **G** | G | C | C | **G** | C | C | A | G | G | G | G | A | C | C | T | A |
| R clade | Ws-0 (*Rps2*$^{pR}$) | A | G | C | C | A | C | C | **G** | **A** | G | G | G | A | C | C | T | A |
| S clade | Wu-0 (*Rps2*$^{S}$) | A | **A** | **T** | **T** | A | **A** | **A** | A | G | G | **T** | **A** | **G** | **T** | **G** | **A** | **G** |

| | Position in Coding Sequence: | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 9 | 9 | 1 | 2 | 2 | 2 | 2 | 3 | 4 |
| **Clade** | **Allele (Line)** | 4 | 5 | 0 | 1 | 3 | 5 | 8 | 5 | 9 |
| | | 6 | 8 | 9 | 6 | 5 | 0 | 3 | 3 | 8 |
| R clade | Col-0 (*Rps2*$^{R}$) | G | C | **A** | T | G | G | G | C | **G** |
| R clade | Ct-0 (*Rps2*$^{R}$) | G | C | T | T | G | **T** | G | C | C |
| R clade | Ler-0 (*Rps2*$^{R}$) | G | C | T | T | **A** | G | **T** | C | **G** |
| R clade | Ws-0 (*Rps2*$^{pR}$) | G | C | T | T | G | G | G | C | C |
| S clade | Wu-0 (*Rps2*$^{S}$) | **A** | **T** | T | **C** | G | **T** | G | **G** | C |

**List B.2.** Gene ontology enrichments of the 295 annotated genes downregulated in *Rps2*$^{+}$ lines compared to *Rps2*$^{KO}$ lines. *Rps2*$^{KO}$ is the *Rps2*-101c* mutant with an empty lox site at the insertion site two, and *Rps2*$^{+}$ included three lines: an *R* clade and *S* clade line, with resistant and susceptible *Rps2* lines with similar levels of expression from insertion site two, and a high *R* line, with the same allele of *Rps2* as in the *R* clade comparison, from insertion site three, which has a higher level of *Rps2* expression. GO annotations are from experimentally verified datasets and differentially expressed genes were compared to the entire annotated *Arabidopsis thaliana* genome. *p* values are Bonferroni corrected for multiple testing.

| GO annotation | # in Genome | # DE in *Rps2*$^{+}$ vs *Rps2*$^{KO}$ | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| response to stimulus | 2216 | 77 | 24.5 | + | 3.14 | 1.18E-16 |
| response to stress | 1234 | 56 | 13.64 | + | 4.1 | 3.57E-16 |
| biological process | 4190 | 109 | 46.32 | + | 2.35 | 7.70E-16 |
| response to external stimulus | 660 | 39 | 7.3 | + | > 5 | 3.92E-14 |
| response to inorganic substance | 489 | 31 | 5.41 | + | > 5 | 1.37E-11 |
| response to metal ion | 289 | 24 | 3.19 | + | > 5 | 5.77E-11 |
| response to other organism | 485 | 30 | 5.36 | + | > 5 | 7.04E-11 |
| response to external biotic stimulus | 485 | 30 | 5.36 | + | > 5 | 7.04E-11 |
| response to biotic stimulus | 486 | 30 | 5.37 | + | > 5 | 7.41E-11 |
| defense response to other organism | 325 | 25 | 3.59 | + | > 5 | 9.01E-11 |
| multi-organism process | 660 | 34 | 7.3 | + | 4.66 | 2.29E-10 |
| response to chemical | 1232 | 46 | 13.62 | + | 3.38 | 8.51E-10 |
| defense response | 369 | 25 | 4.08 | + | > 5 | 1.40E-09 |
| small molecule metabolic process | 414 | 26 | 4.58 | + | > 5 | 2.60E-09 |
| oxoacid metabolic process | 295 | 22 | 3.26 | + | > 5 | 5.37E-09 |
| metabolic process | 1454 | 49 | 16.07 | + | 3.05 | 5.37E-09 |
| organic acid metabolic process | 296 | 22 | 3.27 | + | > 5 | 5.73E-09 |

| GO annotation | # in Genome | # DE in *Rps2*$^{+}$ vs *Rps2*$^{KO}$ | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| cellular metabolic process | 1314 | 46 | 14.53 | + | 3.17 | 7.57E-09 |
| organonitrogen compound metabolic process | 251 | 20 | 2.77 | + | > 5 | 1.63E-08 |
| single-organism metabolic process | 880 | 36 | 9.73 | + | 3.7 | 2.94E-08 |
| response to cadmium ion | 215 | 18 | 2.38 | + | > 5 | 8.02E-08 |
| sulfur compound biosynthetic process | 43 | 10 | 0.48 | + | > 5 | 1.13E-07 |
| single-organism process | 2481 | 64 | 27.43 | + | 2.33 | 1.70E-07 |
| S-glycoside biosynthetic process | 21 | 8 | 0.23 | + | > 5 | 1.92E-07 |
| glucosinolate biosynthetic process | 21 | 8 | 0.23 | + | > 5 | 1.92E-07 |
| glycosinolate biosynthetic process | 21 | 8 | 0.23 | + | > 5 | 1.92E-07 |
| sulfur compound metabolic process | 81 | 12 | 0.9 | + | > 5 | 2.47E-07 |
| cellular process | 2355 | 61 | 26.04 | + | 2.34 | 4.37E-07 |
| single-organism cellular process | 1724 | 50 | 19.06 | + | 2.62 | 5.84E-07 |
| response to bacterium | 224 | 17 | 2.48 | + | > 5 | 1.18E-06 |
| glycosyl compound biosynthetic process | 27 | 8 | 0.3 | + | > 5 | 1.35E-06 |
| defense response to bacterium | 178 | 15 | 1.97 | + | > 5 | 2.93E-06 |
| single-organism biosynthetic proc. | 415 | 22 | 4.59 | + | 4.8 | 3.01E-06 |

| GO annotation | # in Genome | # DE in *Rps2*$^{+}$ vs *Rps2*$^{KO}$ | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| organonitrogen compound biosynthetic process | 128 | 13 | 1.42 | + | > 5 | 3.94E-06 |
| response to oxidative stress | 213 | 16 | 2.35 | + | > 5 | 4.28E-06 |
| organic acid biosynthetic process | 129 | 13 | 1.43 | + | > 5 | 4.32E-06 |
| carboxylic acid biosynthetic process | 129 | 13 | 1.43 | + | > 5 | 4.32E-06 |
| response to abiotic stimulus | 968 | 34 | 10.7 | + | 3.18 | 5.20E-06 |
| carbohydrate derivative metabolic process | 109 | 12 | 1.21 | + | > 5 | 6.63E-06 |
| S-glycoside metabolic process | 34 | 8 | 0.38 | + | > 5 | 8.01E-06 |
| glucosinolate metabolic process | 34 | 8 | 0.38 | + | > 5 | 8.01E-06 |
| glycosinolate metabolic process | 34 | 8 | 0.38 | + | > 5 | 8.01E-06 |
| glycosyl compound metabolic process | 50 | 9 | 0.55 | + | > 5 | 8.95E-06 |
| cellular biosynthetic process | 483 | 23 | 5.34 | + | 4.31 | 9.38E-06 |
| cellular response to stress | 294 | 18 | 3.25 | + | > 5 | 1.03E-05 |
| organic substance metabolic process | 1313 | 40 | 14.52 | + | 2.76 | 1.07E-05 |
| organic substance biosynthetic process | 501 | 23 | 5.54 | + | 4.15 | 1.82E-05 |
| defense response by callose deposition in cell wall | 14 | 6 | 0.15 | + | > 5 | 1.97E-05 |

| GO annotation | # in Genome | # DE in *Rps2*$^{+}$ vs *Rps2*$^{KO}$ | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| defense response by cell wall thickening | 14 | 6 | 0.15 | + | > 5 | 1.97E-05 |
| response to oxygen-containing compound | 734 | 28 | 8.11 | + | 3.45 | 2.44E-05 |
| indole-containing compound biosynthetic process | 26 | 7 | 0.29 | + | > 5 | 2.89E-05 |
| cell wall thickening | 15 | 6 | 0.17 | + | > 5 | 2.95E-05 |
| callose deposition in cell wall | 15 | 6 | 0.17 | + | > 5 | 2.95E-05 |
| defense response by callose deposition | 15 | 6 | 0.17 | + | > 5 | 2.95E-05 |
| nitrogen compound metabolic process | 564 | 24 | 6.24 | + | 3.85 | 3.61E-05 |
| carboxylic acid metabolic process | 257 | 16 | 2.84 | + | > 5 | 5.61E-05 |
| cellular nitrogen compound biosynthetic process | 136 | 12 | 1.5 | + | > 5 | 7.26E-05 |
| polysaccharide localization | 18 | 6 | 0.2 | + | > 5 | 8.56E-05 |
| callose localization | 18 | 6 | 0.2 | + | > 5 | 8.56E-05 |
| biosynthetic process | 558 | 23 | 6.17 | + | 3.73 | 1.24E-04 |
| small molecule biosynthetic process | 175 | 13 | 1.93 | + | > 5 | 1.45E-04 |
| carbohydrate derivative biosynthetic process | 70 | 9 | 0.77 | + | > 5 | 1.53E-04 |
| indole-containing compound metabolic process | 35 | 7 | 0.39 | + | > 5 | 2.13E-04 |

| GO annotation | # in Genome | # DE in *Rps2*$^{+}$ vs *Rps2*$^{KO}$ | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| aromatic amino acid family metabolic process | 22 | 6 | 0.24 | + | > 5 | 2.75E-04 |
| indolalkylamine metabolic process | 12 | 5 | 0.13 | + | > 5 | 3.66E-04 |
| tryptophan metabolic process | 12 | 5 | 0.13 | + | > 5 | 3.66E-04 |
| defense response to fungus | 130 | 11 | 1.44 | + | > 5 | 3.90E-04 |
| secondary metabolic process | 82 | 9 | 0.91 | + | > 5 | 5.65E-04 |
| cellular response to stimulus | 572 | 22 | 6.32 | + | 3.48 | 7.46E-04 |
| secondary metabolite biosynthetic process | 63 | 8 | 0.7 | + | > 5 | 8.45E-04 |
| cellular biogenic amine metabolic process | 28 | 6 | 0.31 | + | > 5 | 1.11E-03 |
| cellular amine metabolic process | 28 | 6 | 0.31 | + | > 5 | 1.11E-03 |
| response to fungus | 179 | 12 | 1.98 | + | > 5 | 1.29E-03 |
| amine metabolic process | 46 | 7 | 0.51 | + | > 5 | 1.30E-03 |
| cell wall modification | 30 | 6 | 0.33 | + | > 5 | 1.64E-03 |
| monocarboxylic acid biosynthetic process | 70 | 8 | 0.77 | + | > 5 | 1.84E-03 |
| auxin biosynthetic process | 17 | 5 | 0.19 | + | > 5 | 2.00E-03 |
| response to reactive oxygen species | 96 | 9 | 1.06 | + | > 5 | 2.04E-03 |
| indoleacetic acid biosynthetic process | 8 | 4 | 0.09 | + | > 5 | 2.87E-03 |
| response to osmotic stress | 310 | 15 | 3.43 | + | 4.38 | 3.30E-03 |

| GO annotation | # in Genome | # DE in *Rps2*$^{+}$ vs *Rps2*$^{KO}$ | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| cellular amino acid metabolic process | 77 | 8 | 0.85 | + | > 5 | 3.68E-03 |
| response to organic substance | 840 | 26 | 9.29 | + | 2.8 | 3.86E-03 |
| response to light stimulus | 316 | 15 | 3.49 | + | 4.29 | 4.15E-03 |
| indoleacetic acid metabolic process | 9 | 4 | 0.1 | + | > 5 | 4.57E-03 |
| cellular aromatic compound metabolic process | 500 | 19 | 5.53 | + | 3.44 | 5.25E-03 |
| response to salt stress | 284 | 14 | 3.14 | + | 4.46 | 5.77E-03 |
| heterocycle biosynthetic process | 140 | 10 | 1.55 | + | > 5 | 5.93E-03 |
| response to radiation | 326 | 15 | 3.6 | + | 4.16 | 6.01E-03 |
| single-organism localization | 370 | 16 | 4.09 | + | 3.91 | 6.30E-03 |
| cellular nitrogen compound metabolic process | 466 | 18 | 5.15 | + | 3.49 | 7.56E-03 |
| monocarboxylic acid metabolic process | 145 | 10 | 1.6 | + | > 5 | 8.03E-03 |
| alpha-amino acid metabolic process | 65 | 7 | 0.72 | + | > 5 | 1.22E-02 |
| organic cyclic compound metabolic process | 540 | 19 | 5.97 | + | 3.18 | 1.53E-02 |
| single-organism cellular localization | 71 | 7 | 0.78 | + | > 5 | 2.15E-02 |
| cellular hormone metabolic process | 28 | 5 | 0.31 | + | > 5 | 2.19E-02 |
| response to transition metal nanoparticle | 73 | 7 | 0.81 | + | > 5 | 2.56E-02 |
| response to extracellular stim. | 101 | 8 | 1.12 | + | > 5 | 2.57E-02 |

| GO annotation | # in Genome | # DE in *Rps2*$^{+}$ vs *Rps2*$^{KO}$ | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| auxin metabolic process | 29 | 5 | 0.32 | + | > 5 | 2.59E-02 |
| cell wall organization | 51 | 6 | 0.56 | + | > 5 | 3.26E-02 |
| cellular biogenic amine biosynthetic process | 15 | 4 | 0.17 | + | > 5 | 3.34E-02 |
| amine biosynthetic process | 15 | 4 | 0.17 | + | > 5 | 3.34E-02 |
| heterocycle metabolic process | 473 | 17 | 5.23 | + | 3.25 | 3.38E-02 |
| localization | 426 | 16 | 4.71 | + | 3.4 | 3.47E-02 |
| response to nutrient levels | 81 | 7 | 0.9 | + | > 5 | 4.92E-02 |

**List B.3.** Gene ontology enrichments of the 267 annotated genes downregulated in *R* compared with KO lines. KO is the *Rps2*-101c* mutant with an empty lox site at the insertion site two, and *R* is a resistant *Rps2* line with a low level of *Rps2* expression. GO annotations are from experimentally verified datasets and the downregulated gene set was compared to the entire annotated *Arabidopsis thaliana* genome. *p* values are Bonferroni corrected for multiple testing.

| GO annotation | # in Genome | # DE in *R* vs KO | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| response to stimulus | 2216 | 67 | 22.17 | + | 3.02 | 2.31E-13 |
| response to chemical | 1232 | 47 | 12.33 | + | 3.81 | 4.69E-12 |
| biological process | 4190 | 93 | 41.93 | + | 2.22 | 1.79E-11 |
| response to stress | 1234 | 45 | 12.35 | + | 3.64 | 9.46E-11 |
| response to oxygen-containing compound | 734 | 32 | 7.34 | + | 4.36 | 6.18E-09 |
| response to inorganic substance | 489 | 26 | 4.89 | + | > 5 | 1.03E-08 |
| response to other organism | 485 | 25 | 4.85 | + | > 5 | 5.02E-08 |

| GO annotation | # in Genome | # DE in *R* vs KO | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| response to external biotic stimulus | 485 | 25 | 4.85 | + | > 5 | 5.02E-08 |
| response to biotic stimulus | 486 | 25 | 4.86 | + | > 5 | 5.24E-08 |
| response to external stimulus | 660 | 28 | 6.6 | + | 4.24 | 2.67E-07 |
| response to wounding | 113 | 13 | 1.13 | + | > 5 | 2.68E-07 |
| response to acid chemical | 547 | 25 | 5.47 | + | 4.57 | 5.88E-07 |
| multi-organism process | 660 | 26 | 6.6 | + | 3.94 | 5.53E-06 |
| response to organic substance | 840 | 29 | 8.41 | + | 3.45 | 1.24E-05 |
| response to endogenous stimulus | 721 | 26 | 7.21 | + | 3.6 | 3.22E-05 |
| defense response to other organism | 325 | 17 | 3.25 | + | > 5 | 5.90E-05 |
| response to fungus | 179 | 13 | 1.79 | + | > 5 | 5.92E-05 |
| defense response | 369 | 18 | 3.69 | + | 4.88 | 6.65E-05 |
| oxoacid metabolic process | 295 | 16 | 2.95 | + | > 5 | 9.09E-05 |
| organic acid metabolic process | 296 | 16 | 2.96 | + | > 5 | 9.51E-05 |
| defense response to fungus | 130 | 11 | 1.3 | + | > 5 | 1.45E-04 |
| single-organism cellular process | 1724 | 41 | 17.25 | + | 2.38 | 3.25E-04 |
| single-organism process | 2481 | 52 | 24.82 | + | 2.09 | 3.35E-04 |
| small molecule metabolic process | 414 | 18 | 4.14 | + | 4.35 | 3.54E-04 |
| response to metal ion | 289 | 15 | 2.89 | + | > 5 | 4.05E-04 |
| single-organism metabolic process | 880 | 27 | 8.81 | + | 3.07 | 4.20E-04 |
| response to bacterium | 224 | 13 | 2.24 | + | > 5 | 7.35E-04 |

| GO annotation | # in Genome | # DE in *R* vs KO | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| response to hormone | 637 | 22 | 6.37 | + | 3.45 | 8.11E-04 |
| response to osmotic stress | 310 | 15 | 3.1 | + | 4.84 | 9.64E-04 |
| single-organism biosynthetic process | 415 | 17 | 4.15 | + | 4.09 | 1.70E-03 |
| cellular metabolic process | 1314 | 32 | 13.15 | + | 2.43 | 4.74E-03 |
| cellular process | 2355 | 47 | 23.56 | + | 1.99 | 5.30E-03 |
| metabolic process | 1454 | 34 | 14.55 | + | 2.34 | 5.49E-03 |
| response to salt stress | 284 | 13 | 2.84 | + | 4.57 | 9.48E-03 |
| organic acid biosynthetic process | 129 | 9 | 1.29 | + | > 5 | 9.65E-03 |
| carboxylic acid biosynthetic process | 129 | 9 | 1.29 | + | > 5 | 9.65E-03 |
| monocarboxylic acid biosynthetic process | 70 | 7 | 0.7 | + | > 5 | 1.03E-02 |
| cellular biosynthetic process | 483 | 17 | 4.83 | + | 3.52 | 1.23E-02 |
| response to oxidative stress | 213 | 11 | 2.13 | + | > 5 | 1.60E-02 |
| response to zinc ion | 29 | 5 | 0.29 | + | > 5 | 1.61E-02 |
| response to cadmium ion | 215 | 11 | 2.15 | + | > 5 | 1.74E-02 |
| carboxylic acid metabolic process | 257 | 12 | 2.57 | + | 4.67 | 1.75E-02 |
| organic substance biosynthetic process | 501 | 17 | 5.01 | + | 3.39 | 1.95E-02 |
| defense response to bacterium | 178 | 10 | 1.78 | + | > 5 | 1.95E-02 |
| monocarboxylic acid metabolic process | 145 | 9 | 1.45 | + | > 5 | 2.41E-02 |

| GO annotation | # in Genome | # DE in *R* vs KO | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| response to jasmonic acid | 115 | 8 | 1.15 | + | > 5 | 3.15E-02 |

**List B.4.** Gene ontology enrichments of the 136 annotated genes downregulated in *S* compared with KO lines. KO is the *Rps2*-101c* mutant with an empty lox site at the insertion site two, and *S* is a susceptible *Rps2* line with a low level of *Rps2* expression. GO annotations are from experimentally verified datasets and the downregulated gene set was compared to the entire annotated *Arabidopsis thaliana* genome. *p* values are Bonferroni corrected for multiple testing.

| GO annotation | # in Genome | # DE in S vs KO | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| response to stress | 1234 | 39 | 6.29 | + | > 5 | 2.29E-17 |
| response to inorganic substance | 489 | 26 | 2.49 | + | > 5 | 6.89E-16 |
| response to oxygen-containing compound | 734 | 29 | 3.74 | + | > 5 | 1.30E-14 |
| response to stimulus | 2216 | 45 | 11.29 | + | 3.98 | 2.98E-13 |
| response to chemical | 1232 | 33 | 6.28 | + | > 5 | 3.96E-12 |
| response to abiotic stimulus | 968 | 28 | 4.93 | + | > 5 | 1.12E-10 |
| biological process | 4190 | 56 | 21.36 | + | 2.62 | 1.31E-09 |
| response to osmotic stress | 310 | 16 | 1.58 | + | > 5 | 9.52E-09 |
| response to acid chemical | 547 | 20 | 2.79 | + | > 5 | 9.60E-09 |
| cellular response to reactive oxygen species | 25 | 7 | 0.13 | + | > 5 | 1.03E-07 |
| response to reactive oxygen species | 96 | 10 | 0.49 | + | > 5 | 1.26E-07 |
| response to oxidative stress | 213 | 13 | 1.09 | + | > 5 | 1.28E-07 |
| cellular response to oxidative stress | 28 | 7 | 0.14 | + | > 5 | 2.24E-07 |
| response to transition metal nanoparticle | 73 | 9 | 0.37 | + | > 5 | 2.59E-07 |
| response to salt stress | 284 | 14 | 1.45 | + | > 5 | 3.73E-07 |

| GO annotation | # in Genome | # DE in S vs KO | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| response to metal ion | 289 | 14 | 1.47 | + | > 5 | 4.66E-07 |
| response to water deprivation | 148 | 11 | 0.75 | + | > 5 | 4.86E-07 |
| response to water | 149 | 11 | 0.76 | + | > 5 | 5.21E-07 |
| cellular response to oxygen-containing compound | 176 | 11 | 0.9 | + | > 5 | 2.90E-06 |
| multi-organism process | 660 | 18 | 3.36 | + | > 5 | 1.11E-05 |
| response to organic substance | 840 | 20 | 4.28 | + | 4.67 | 1.49E-05 |
| response to iron ion | 34 | 6 | 0.17 | + | > 5 | 3.60E-05 |
| response to other organism | 485 | 15 | 2.47 | + | > 5 | 4.20E-05 |
| response to external biotic stimulus | 485 | 15 | 2.47 | + | > 5 | 4.20E-05 |
| response to biotic stimulus | 486 | 15 | 2.48 | + | > 5 | 4.31E-05 |
| response to temperature stimulus | 255 | 11 | 1.3 | + | > 5 | 1.22E-04 |
| defense response to other organism | 325 | 12 | 1.66 | + | > 5 | 1.65E-04 |
| cellular process | 2355 | 32 | 12 | + | 2.67 | 2.93E-04 |
| response to external stimulus | 660 | 16 | 3.36 | + | 4.76 | 3.81E-04 |
| cellular response to stress | 294 | 11 | 1.5 | + | > 5 | 4.94E-04 |
| cellular response to ethylene stimulus | 28 | 5 | 0.14 | + | > 5 | 5.04E-04 |
| cellular response to metal ion | 28 | 5 | 0.14 | + | > 5 | 5.04E-04 |
| defense response | 369 | 12 | 1.88 | + | > 5 | 6.27E-04 |
| phloem transport | 12 | 4 | 0.06 | + | > 5 | 6.56E-04 |
| vascular transport | 12 | 4 | 0.06 | + | > 5 | 6.56E-04 |
| cellular response to chemical stimulus | 303 | 11 | 1.54 | + | > 5 | 6.62E-04 |
| cellular response to inorganic substance | 30 | 5 | 0.15 | + | > 5 | 7.06E-04 |
| cellular response to nitric oxide | 13 | 4 | 0.07 | + | > 5 | 9.00E-04 |

| GO annotation | # in Genome | # DE in S vs KO | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| cellular response to reactive nitrogen species | 14 | 4 | 0.07 | + | > 5 | 1.21E-03 |
| response to nitric oxide | 14 | 4 | 0.07 | + | > 5 | 1.21E-03 |
| nicotianamine biosynthetic process | 4 | 3 | 0.02 | + | > 5 | 1.68E-03 |
| nicotianamine metabolic process | 4 | 3 | 0.02 | + | > 5 | 1.68E-03 |
| tricarboxylic acid biosynthetic process | 4 | 3 | 0.02 | + | > 5 | 1.68E-03 |
| cellular response to stimulus | 572 | 14 | 2.92 | + | 4.8 | 1.96E-03 |
| single-organism cellular process | 1724 | 25 | 8.79 | + | 2.85 | 2.53E-03 |
| response to nitrogen compound | 161 | 8 | 0.82 | + | > 5 | 2.57E-03 |
| single-organism process | 2481 | 31 | 12.64 | + | 2.45 | 2.81E-03 |
| response to bacterium | 224 | 9 | 1.14 | + | > 5 | 3.28E-03 |
| single-organism metabolic process | 880 | 17 | 4.49 | + | 3.79 | 3.49E-03 |
| iron ion homeostasis | 19 | 4 | 0.1 | + | > 5 | 4.01E-03 |
| cellular response to organic substance | 238 | 9 | 1.21 | + | > 5 | 5.33E-03 |
| response to endogenous stimulus | 721 | 15 | 3.67 | + | 4.08 | 5.84E-03 |
| cellular response to nitrogen compound | 47 | 5 | 0.24 | + | > 5 | 6.21E-03 |
| response to hormone | 637 | 14 | 3.25 | + | 4.31 | 6.68E-03 |
| defense response to fungus | 130 | 7 | 0.66 | + | > 5 | 6.78E-03 |
| cellular response to iron ion | 23 | 4 | 0.12 | + | > 5 | 8.48E-03 |
| reactive oxygen species metabolic process | 24 | 4 | 0.12 | + | > 5 | 1.00E-02 |
| transition metal ion transport | 24 | 4 | 0.12 | + | > 5 | 1.00E-02 |

| GO annotation | # in Genome | # DE in S vs KO | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| cellular metabolic process | 1314 | 20 | 6.7 | + | 2.99 | 1.53E-02 |
| single-organism transport | 351 | 10 | 1.79 | + | > 5 | 1.81E-02 |
| cellular response to light intensity | 9 | 3 | 0.05 | + | > 5 | 1.88E-02 |
| tricarboxylic acid metabolic process | 9 | 3 | 0.05 | + | > 5 | 1.88E-02 |
| metabolic process | 1454 | 21 | 7.41 | + | 2.83 | 1.98E-02 |
| response to zinc ion | 29 | 4 | 0.15 | + | > 5 | 2.09E-02 |
| cellular response to UV | 10 | 3 | 0.05 | + | > 5 | 2.57E-02 |
| single-organism localization | 370 | 10 | 1.89 | + | > 5 | 2.82E-02 |
| oxoacid metabolic process | 295 | 9 | 1.5 | + | > 5 | 2.89E-02 |
| organic acid metabolic process | 296 | 9 | 1.51 | + | > 5 | 2.96E-02 |
| transport | 382 | 10 | 1.95 | + | > 5 | 3.68E-02 |
| cellular response to acid chemical | 118 | 6 | 0.6 | + | > 5 | 4.44E-02 |
| establishment of localization | 392 | 10 | 2 | + | > 5 | 4.57E-02 |
| response to cold | 177 | 7 | 0.9 | + | > 5 | 4.82E-02 |
| transition metal ion homeostasis | 36 | 4 | 0.18 | + | > 5 | 4.84E-02 |
| defense response to bacterium | 178 | 7 | 0.91 | + | > 5 | 5.00E-02 |

**List B.5.** Gene ontology enrichments of the 590 genes upregulated in High R line compared with a KO line. KO is the *Rps2*-101c* mutant with an empty lox site at the insertion site two, and High R is a resistant *Rps2* line with a high level of *Rps2* expression. GO annotations are from experimentally verified datasets and the upregulated gene set was compared to the entire annotated *Arabidopsis thaliana* genome. *p* values are Bonferroni corrected for multiple testing.

| GO annotation | # in Genome | # DE in High R vs KO | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| myo-inositol hexakisphosphate biosynthetic process | 43 | 10 | 0.95 | + | > 5 | 1.21E-04 |
| myo-inositol hexakisphosphate metabolic process | 43 | 10 | 0.95 | + | > 5 | 1.21E-04 |
| inositol phosphate biosynthetic process | 43 | 10 | 0.95 | + | > 5 | 1.21E-04 |
| polyol biosynthetic process | 45 | 10 | 0.99 | + | > 5 | 1.83E-04 |
| inositol phosphate metabolic process | 60 | 11 | 1.33 | + | > 5 | 2.84E-04 |
| response to blue light | 74 | 12 | 1.64 | + | > 5 | 2.87E-04 |
| response to red light | 61 | 11 | 1.35 | + | > 5 | 3.34E-04 |
| protein metabolic process | 3151 | 35 | 69.67 | - | 0.5 | 1.93E-03 |
| polyol metabolic process | 76 | 11 | 1.68 | + | > 5 | 2.79E-03 |
| cellular developmental process | 1040 | 48 | 23 | + | 2.09 | 3.70E-03 |
| cellular protein metabolic process | 2795 | 30 | 61.8 | - | 0.49 | 3.87E-03 |
| cellular response to light stimulus | 69 | 10 | 1.53 | + | > 5 | 8.21E-03 |
| cell differentiation | 884 | 42 | 19.55 | + | 2.15 | 8.22E-03 |
| response to gibberellin | 124 | 13 | 2.74 | + | 4.74 | 1.07E-02 |
| cellular response to radiation | 72 | 10 | 1.59 | + | > 5 | 1.18E-02 |
| response to hormone | 1218 | 52 | 26.93 | + | 1.93 | 1.22E-02 |

| GO annotation | # in Genome | # DE in High R vs KO | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| alcohol biosynthetic process | 244 | 18 | 5.39 | + | 3.34 | 2.37E-02 |
| response to endogenous stimulus | 1383 | 56 | 30.58 | + | 1.83 | 2.42E-02 |
| root morphogenesis | 319 | 21 | 7.05 | + | 2.98 | 2.52E-02 |
| trichoblast differentiation | 246 | 18 | 5.44 | + | 3.31 | 2.63E-02 |
| organ development | 1254 | 52 | 27.73 | + | 1.88 | 2.67E-02 |
| system development | 1876 | 70 | 41.48 | + | 1.69 | 2.87E-02 |
| response to red or far red light | 275 | 19 | 6.08 | + | 3.12 | 3.35E-02 |
| multicellular organismal development | 2471 | 86 | 54.64 | + | 1.57 | 3.78E-02 |
| root development | 437 | 25 | 9.66 | + | 2.59 | 4.10E-02 |
| root system development | 437 | 25 | 9.66 | + | 2.59 | 4.10E-02 |
| cellular response to abiotic stimulus | 142 | 13 | 3.14 | + | 4.14 | 4.35E-02 |
| root epidermal cell differentiation | 256 | 18 | 5.66 | + | 3.18 | 4.41E-02 |
| transcription, DNA-templated | 1348 | 54 | 29.81 | + | 1.81 | 4.63E-02 |

**List B.6.** Gene ontology enrichments of the 581 genes downregulated in High R line compared with a KO line. KO is the *Rps2*-101c* mutant with an empty lox site at the insertion site two, and High R is a resistant *Rps2* line with a high level of *Rps2* expression. GO annotations are from experimentally verified datasets and the downregulated gene set was compared to the entire annotated *Arabidopsis thaliana* genome. *p* values are Bonferroni corrected for multiple testing.

| GO annotation | # in Genome | # DE in High R vs KO | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| response to stimulus | 2216 | 133 | 48.25 | + | 2.76 | 9.66E-24 |
| response to stress | 1234 | 92 | 26.87 | + | 3.42 | 1.78E-21 |
| biological process | 4190 | 182 | 91.23 | + | 1.99 | 6.30E-18 |
| response to chemical | 1232 | 77 | 26.82 | + | 2.87 | 2.83E-13 |
| response to external stimulus | 660 | 52 | 14.37 | + | 3.62 | 5.29E-12 |
| response to oxygen-containing compound | 734 | 55 | 15.98 | + | 3.44 | 6.33E-12 |
| response to abiotic stimulus | 968 | 63 | 21.08 | + | 2.99 | 3.33E-11 |
| response to other organism | 485 | 42 | 10.56 | + | 3.98 | 1.25E-10 |
| response to external biotic stimulus | 485 | 42 | 10.56 | + | 3.98 | 1.25E-10 |
| response to biotic stimulus | 486 | 42 | 10.58 | + | 3.97 | 1.34E-10 |
| response to acid chemical | 547 | 44 | 11.91 | + | 3.69 | 3.76E-10 |
| response to osmotic stress | 310 | 32 | 6.75 | + | 4.74 | 1.37E-09 |
| response to organic substance | 840 | 53 | 18.29 | + | 2.9 | 1.33E-08 |
| defense response to other organism | 325 | 31 | 7.08 | + | 4.38 | 2.20E-08 |
| multi-organism process | 660 | 45 | 14.37 | + | 3.13 | 4.57E-08 |
| defense response | 369 | 31 | 8.03 | + | 3.86 | 4.66E-07 |
| response to salt stress | 284 | 27 | 6.18 | + | 4.37 | 4.78E-07 |
| cellular process | 2355 | 98 | 51.28 | + | 1.91 | 6.47E-07 |
| response to endogenous stimulus | 721 | 43 | 15.7 | + | 2.74 | 6.57E-06 |

| GO annotation | # in Genome | # DE in High R vs KO | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| single-organism cellular process | 1724 | 76 | 37.54 | + | 2.02 | 7.17E-06 |
| single-organism process | 2481 | 98 | 54.02 | + | 1.81 | 9.18E-06 |
| response to inorganic substance | 489 | 33 | 10.65 | + | 3.1 | 2.51E-05 |
| response to hormone | 637 | 38 | 13.87 | + | 2.74 | 5.18E-05 |
| cellular response to stimulus | 572 | 35 | 12.45 | + | 2.81 | 9.86E-05 |
| response to fungus | 179 | 18 | 3.9 | + | 4.62 | 1.79E-04 |
| response to bacterium | 224 | 20 | 4.88 | + | 4.1 | 2.40E-04 |
| cellular response to stress | 294 | 23 | 6.4 | + | 3.59 | 2.98E-04 |
| defense response to fungus | 130 | 15 | 2.83 | + | > 5 | 3.59E-04 |
| single-organism localization | 370 | 26 | 8.06 | + | 3.23 | 3.79E-04 |
| defense response to bacterium | 178 | 17 | 3.88 | + | 4.39 | 7.96E-04 |
| induced systemic resistance | 22 | 7 | 0.48 | + | > 5 | 9.04E-04 |
| oxoacid metabolic process | 295 | 22 | 6.42 | + | 3.43 | 1.18E-03 |
| organic acid metabolic process | 296 | 22 | 6.44 | + | 3.41 | 1.25E-03 |
| localization | 426 | 27 | 9.28 | + | 2.91 | 1.58E-03 |
| cellular amino acid metabolic process | 77 | 11 | 1.68 | + | > 5 | 1.84E-03 |
| regulation of biological quality | 259 | 20 | 5.64 | + | 3.55 | 2.18E-03 |
| cellular response to iron ion starvation | 4 | 4 | 0.09 | + | > 5 | 2.73E-03 |
| cell wall organization or biogenesis | 100 | 12 | 2.18 | + | > 5 | 3.68E-03 |
| response to lipid | 317 | 22 | 6.9 | + | 3.19 | 3.71E-03 |
| response to carbohydrate | 56 | 9 | 1.22 | + | > 5 | 6.50E-03 |

| GO annotation | # in Genome | # DE in High R vs KO | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| innate immune response | 145 | 14 | 3.16 | + | 4.43 | 6.75E-03 |
| cell wall modification | 30 | 7 | 0.65 | + | > 5 | 6.83E-03 |
| immune response | 148 | 14 | 3.22 | + | 4.34 | 8.48E-03 |
| immune system process | 169 | 15 | 3.68 | + | 4.08 | 8.51E-03 |
| response to wounding | 113 | 12 | 2.46 | + | 4.88 | 1.24E-02 |
| cellular biosynthetic process | 483 | 27 | 10.52 | + | 2.57 | 1.50E-02 |
| single-organism transport | 351 | 22 | 7.64 | + | 2.88 | 1.77E-02 |
| ion transport | 159 | 14 | 3.46 | + | 4.04 | 1.86E-02 |
| organonitrogen compound metabolic process | 251 | 18 | 5.47 | + | 3.29 | 1.89E-02 |
| alpha-amino acid metabolic process | 65 | 9 | 1.42 | + | > 5 | 2.09E-02 |
| transport | 382 | 23 | 8.32 | + | 2.77 | 2.12E-02 |
| reactive oxygen species metabolic process | 24 | 6 | 0.52 | + | > 5 | 2.19E-02 |
| hormone biosynthetic process | 36 | 7 | 0.78 | + | > 5 | 2.19E-02 |
| transition metal ion homeostasis | 36 | 7 | 0.78 | + | > 5 | 2.19E-02 |
| cell wall organization | 51 | 8 | 1.11 | + | > 5 | 2.56E-02 |
| small molecule metabolic process | 414 | 24 | 9.01 | + | 2.66 | 2.56E-02 |
| carboxylic acid metabolic process | 257 | 18 | 5.6 | + | 3.22 | 2.56E-02 |
| single-organism biosynthetic process | 415 | 24 | 9.04 | + | 2.66 | 2.66E-02 |
| establishment of localization | 392 | 23 | 8.54 | + | 2.69 | 3.15E-02 |

| GO annotation | # in Genome | # DE in High R vs KO | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| indole-containing compound biosynthetic process | 26 | 6 | 0.57 | + | > 5 | 3.41E-02 |
| single-organism metabolic process | 880 | 39 | 19.16 | + | 2.04 | 3.96E-02 |
| organic acid biosynthetic process | 129 | 12 | 2.81 | + | 4.27 | 4.44E-02 |
| carboxylic acid biosynthetic process | 129 | 12 | 2.81 | + | 4.27 | 4.44E-02 |

**List B.7.** Gene ontology enrichments of the 34 annotated genes downregulated in High R compared with Low R lines. High R is an *Rps2* line with a high level of *Rps2* expression, and Low R is an *Rps2* line with a low level of *Rps2* expression. GO annotations are from experimentally verified datasets and the downregulated gene set was compared to the entire annotated *Arabidopsis thaliana* genome. *p* values are Bonferroni corrected for multiple testing.

| GO annotation | # in Genome | # DE in High R vs Low R | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| response to stimulus | 2216 | 48 | 14.95 | + | 3.21 | 3.38E-10 |
| response to abiotic stimulus | 968 | 31 | 6.53 | + | 4.75 | 8.68E-10 |
| response to stress | 1234 | 33 | 8.32 | + | 3.96 | 1.75E-08 |
| response to osmotic stress | 310 | 16 | 2.09 | + | > 5 | 6.58E-07 |
| response to salt stress | 284 | 15 | 1.92 | + | > 5 | 1.71E-06 |
| response to fungus | 179 | 12 | 1.21 | + | > 5 | 6.01E-06 |
| response to chemical | 1232 | 29 | 8.31 | + | 3.49 | 6.28E-06 |
| biological process | 4190 | 57 | 28.26 | + | 2.02 | 9.15E-05 |
| defense response to fungus | 130 | 9 | 0.88 | + | > 5 | 4.03E-04 |
| response to hypoxia | 41 | 6 | 0.28 | + | > 5 | 5.61E-04 |

| GO annotation | # in Genome | # DE in High R vs Low R | # Expected | Over/ Under Expected | Fold Enrichment | *p* value |
|---|---|---|---|---|---|---|
| response to decreased oxygen levels | 47 | 6 | 0.32 | + | > 5 | 1.23E-03 |
| response to oxygen levels | 47 | 6 | 0.32 | + | > 5 | 1.23E-03 |
| response to acid chemical | 547 | 16 | 3.69 | + | 4.34 | 1.46E-03 |
| response to other organism | 485 | 15 | 3.27 | + | 4.58 | 1.63E-03 |
| response to external biotic stimulus | 485 | 15 | 3.27 | + | 4.58 | 1.63E-03 |
| response to biotic stimulus | 486 | 15 | 3.28 | + | 4.58 | 1.67E-03 |
| response to inorganic substance | 489 | 15 | 3.3 | + | 4.55 | 1.80E-03 |
| response to cadmium ion | 215 | 10 | 1.45 | + | > 5 | 3.13E-03 |
| defense response to other organism | 325 | 12 | 2.19 | + | > 5 | 3.31E-03 |
| response to cold | 177 | 9 | 1.19 | + | > 5 | 4.94E-03 |
| response to organic substance | 840 | 19 | 5.67 | + | 3.35 | 6.08E-03 |
| response to metal ion | 289 | 11 | 1.95 | + | > 5 | 6.55E-03 |
| defense response | 369 | 12 | 2.49 | + | 4.82 | 1.18E-02 |
| response to temperature stimulus | 255 | 10 | 1.72 | + | > 5 | 1.37E-02 |
| response to oxygen-containing compound | 734 | 17 | 4.95 | + | 3.43 | 1.46E-02 |
| multi-organism process | 660 | 16 | 4.45 | + | 3.59 | 1.53E-02 |
| response to external stimulus | 660 | 16 | 4.45 | + | 3.59 | 1.53E-02 |

# Appendix C

## RNA Extraction and High Throughput RNA extraction protocols

The following protocols are directly from and modified from the protocol of Oñate-Sánchez and Vicente-Carbajosa (2008). They allow easy, cheap, high-yield extraction of RNA from *Arabidopsis thaliana* in microgram to milligram quantities.

### C.1 Sample preparation and general considerations

You will need: RNAseZap (ThermoFisher AM9780), RNAse free 1.5mL Eppendorf tubes and autoclaved pestles (for microgram quantities of RNA) or RNAse free 15mL or 50mL conicals and RNAse free mortar and pestles (for milligram quantities of RNA), liquid nitrogen, latex gloves, RNAse free pipette tips and pipettors, DNAseI and 10x DNAseI buffer, the seven RNA extraction solutions made up into RNAse free glassware, and a RNAse free workstation.

When working with RNA, it is essential that an RNAse free workstation be maintained. RNA is an incredibly transient molecule; it begins to be degraded by RNAses - which are present within every living cell - as soon as the living tissue dies when the tissue is at any temperature above -80ºC. RNAseZap removes ambient RNAses (which, in contrast to RNA, are some of the most stable enzymes on the planet, and notably are not destroyed by autoclaving) by suspending them in the RNAseZap spray, which you then rinse or remove from whatever you are making RNAse free. To decontaminate glassware and Eppendorf racks, spray a layer of RNAseZap into the glassware to coat, then rinse 2-3 times with deionized water. To decontaminate pipettors or your workstation, spray a layer of RNAseZap and then wipe with KimWipes. Wear gloves sprayed with RNAseZap at all times when using RNAseZap and when working with RNA (even

if you have been careful to remove RNAses during RNA extraction, your skin is a fantastic source of new RNAses). If this is your first time working with RNA, I'd advise being as paranoid about RNAse decontamination and keeping your samples cold/frozen as possible your first run through; then, you can relax some steps slightly if you get high yields and low sample degradation. CONSTANT VIGILANCE!

In the following protocols, samples are first frozen in liquid nitrogen, and then ground to a powder with a pestle without allowing the sample to thaw above -80ºC. Frozen tissue samples can be maintained in the -80ºC freezer for up to 6 months before RNA extraction without sample degradation. You must grind tissue to a powder before freezing, as any thawing of the tissue before RNA extraction will lead to unacceptably high degradation. All centrifugations are at the maximum speed that the Eppendorf/conical you are using can handle in a table top microfuge (Eppendorfs) or larger centrifuge (conicals). Typically, the faster the speed of centrifugation, the greater the yield of RNA, but any speed over about 10000 rpm is typically sufficient for good yields. Centrifugation steps are at room temperature unless otherwise indicated.

Once extracted, RNA can be maintained in the -20ºC for approximately two months, or at -80ºC for approximately 6 months before sample degradation. RNA can only really be thawed for use once without unacceptably high degradation; samples should be thawed on ice.

### C.2 RNA extraction reagents

1. Cell Lysis Solution: 2% SDS, 68mM sodium citrate, 132 mM citric acid, 1mM EDTA.
    a. NB: A master mix of EDTA is typically floating around the Bergelson lab; if not, I'd recommend making a 1M EDTA solution as EDTA takes a long time to go into solution.

2. Protein-DNA Precipitation Solution: 4M NaCl, 16mM sodium citrate, 32 mM citric acid
3. Isopropanol - regular grade is fine.
4. 70% Ethanol - regular grade ethanol is ok. Make an RNAse free stock using DEPC treated water.
5. DEPC Water
    a. NB: DEPC treatment removes RNAses.
6. DNAseI and 10x DNAse Buffer
    a. NB: DNAseI is incredibly unstable. Keep at -20ºC and never vortex, shake, drop, or centrifuge the DNAseI.
7. 7.5 M ammonium acetate or sodium acetate, $NH_4Ac$ or Na Ac.
    a. NB: Acetates should be kept refrigerated and appear to have expiration dates.
8. 100% Ethanol - regular grade ethanol is ok.

### C.3 Vegetative tissue protocol (microgram quantities)

Microgram quantities of RNA are sufficient for the majority of applications.

1. Harvest 50-100mg tissue (50mg is approximately the size of a thumb nail) into 1.5mL Eppendorfs.

2. Add 300µL of cell lysis solution to ground tissue and homogenize quickly by vortexing 2s and inverting and flicking the tube gently. Leave tubes at room temperature up to 5 minutes.

a. CRITICAL: The less time spent at room temperature, the better. Less than 5 minutes is critical for ensuring acceptable yields. If you are processing many samples, batch process samples until the 4ºC incubation step to ensure samples stay at room temperature for 5 minutes or less – samples can stay at 4ºC for up to half an hour.

3. Add 100µL of protein-DNA precipitation solution to the cell lysate, mix by inverting the tubes gently, and incubate at 4ºC for at least 10 minutes. Spin at 4ºC for 10 minutes.

4. Transfer the supernatant to a new Eppendorf tube (NB: The less precipitate transferred, the better the eventual RNA quality – leave some supernatant if transferring the supernatant will transfer precipitate as well). Add 300µL isopropanol and mix the sample by inverting the tube gently. Spin 4 minutes and carefully pour off the supernatant. Wash the pellet with 10-50µL 70% ethanol, aspirate off the ethanol, and air dry the RNA. Re-suspend the RNA in 25 µL DEPC water.

5. Add 3µL 10x DNAse buffer and 2 units of DNAseI. Incubate 30 minutes at 37ºC.

6. Add 70µL DEPC-water to the ~30µL of RNA, 50µL of ammonium acetate, and 400µL 100% ethanol and mix well. Spin 20 minutes at 4ºC and carefully pour off the supernatant. Wash the pellet with 10-50µL 70% ethanol, aspirate off the ethanol, and air dry the RNA. Re-suspend the RNA in 15 µL DEPC water.

7. RNA is now ready for quantification and downstream uses. A nanodrop can be used to get a rough estimate of RNA yield. To check RNA quality, a bioanalyzer can be used (best practice but expensive) or an aliquot can be run on a gel. Look for two dark bands at ~800 and ~1100bp – those are the ribosomal RNAs. Generally the larger rRNA band should be twice as bright as the smaller band. Equal brightnesses are ok, but a very dim or nonexistent larger band and a bright smudge 0-200bp in size are signals your RNA is degraded.

## C.4 Vegetative tissue protocol (milligram quantities)

For larger harvests, several plants can be harvested at once into a large ceramic mortar and pestle (freeze mortar by pouring liquid nitrogen into the mortar). Use the volume adjustments that follow (Table C.5).

**Table C.5.** Volume adjustments for different quantities of RNA.

| | 1.5mL Eppendorf | 15mL Conical | 50mL Conical |
|---|---|---|---|
| Harvest Tissue | 50-100mg | 0.6g | 2.1g |
| Add **Cell Lysis Solution** and vortex/mix | 300uL | 6mL | 21mL |
| Add **Protein-DNA precipitation solution** and mix | 100uL | 2mL | 7mL |
| Incubate at 4ºC for at least 10 min, spin at 4ºC for 10 min, Transfer s/n to new tube | | | |
| Add **Isopropanol**, mix, spin 4 min | 300uL | 6mL | 21mL |
| Wash pellet and air dry RNA; Resuspend in **DEPC water** | 25µL water | 0.5mL | 1.75mL |
| Add 10x **DNAse buffer** and **DNAseI**; incubate 30 min at 37ºC | 3µL and 2 units | 60µL and 40 units | 140µL and 95 units |
| Add **DEPC water**; **ammonium acetate**; **100% ethanol** and mix. Spin 20min at 4ºC | 70uL; 50uL; 400uL | 1.4mL; 1mL; 8mL | 4.9mL; 3.5mL; 28mL |
| Wash pellet and air dry RNA; Resuspend in **DEPC water** | 15 µL | 0.4mL | 1.4mL |

Additionally, for the second resuspension and for storage of RNA from larger extractions, it's best to use RNAse free Eppendorfs. Good luck!

# Appendix D

## qPCR Workflow: Primer Design, Protocol, Analysis

The following protocols are for relative quantification of reverse transcribed RNA and/or DNA with quantitative PCR.

### D.1 Primer design protocol for qPCR of cDNA

1. Go to TAIR (www.arabidopsis.org) and enter the name of the gene. You need to find the **GeneBank Accession** (i.e. NM_101094.2 for RPS5) for Primer Blast. So click on the protein coding gene model of your choice (usually the most likely one) and in Nucleotide Sequence section, there should be a GeneBank Accession.

2. Go to Primer Blast (http://www.ncbi.nlm.nih.gov/tools/primer-blast/) and enter your GeneBank Accession into the "Enter accession, gi, or FASTA sequence" box.

-Fill in one of the four boxes specifying the range. Usually I do the forward primer, from ~600bp from the end of the sequence, 3000 (bp) for example. You want primers near the 3' end of the sequence to increase the likelihood that you're only counting full mRNA sequences. Increment this number downward by ~500bp per round of primer selection.

-Change the PCR product size, minimum should be 90 and max 150.

-Increase the # of primers to return to 10 or so.

-Change the primer melting temperature to 55 – 60, optimum is 57.

-I don't usually change the Exon/Intron Selection because R-genes often don't have introns.

-Add more organisms under the Primer Pair Specificity Checking Parameters – at the least add Arabidopsis (3702), fungi (4751), and if you can, bacteria (2), as well as Homo sapiens, which is the default. I don't change any of the other defaults in this grey box.

3. Once you have primers generated, test them using the IDT Oligoanalyzer (http://www.idtdna.com/analyzer/applications/oligoanalyzer/ ) using the following guidelines, especially those that are bolded:

**– SYBR green binds to any dsDNA, therefore primer-dimers and non-specific products needs to be excluded**

– amplicon length 90 – 150bp (high levels of fluorescence signal obtained without compromising PCR efficiencies)

– primer length 18 – 30 nt (mine usually 22 +/- 2); Tm of for/rev primer < 2 K differences; Tm 55-60C

**– GC content : 30-80 % (ideally 40-60 %)**

**– avoid mismatches between primers and target, especially towards the 3' end of the primer**

**– avoid runs of identical nucleotides, especially of 3 or more Gs or Cs at the 3' end**

**– avoid 3' end T (higher probability of mismatches)**

– avoid complementarities within primers **(especially of 2 or more bases at the 3' ends)** to avoid hairpins

4. Ideally your primers will not have homodimers or heterodimers with affinities greater than -6 kcal/mol, although I've taken up to -10 on the 5' end when selection is limited. If hairpins exist, they should be short and not encompass the most 3' base. On Primer-BLAST, your primers should ideally not pick up any other Arabidopsis gene, especially *R*-genes, but where there are P/A polymorphisms for *R*-genes you can test that your primer is specific by seeing if it picks up anything in an *R*-gene absent plant.

5. Order the primers on Buysite from IDT, using their Custom DNA Oligo Service. Or: use this somewhat hidden page at lifescience (primers are synthesized by Operon) which gives you $3.75 per primer if you order more than 25 ($5 per primer if you order less than 25). The price is calculated per oligo and not per base!! It's up to 40mer, 25mM, standard desalted, within a day, shipped dry. https://www.lifetechnologies.com/order/custom-oligo/enterSequences

6. Reconstitute the primers to make a 100mM stock solution (typically, add a volume of water in microliters equivalent to the nanograms of primer).

7. Make 2uM primer stock for testing primers on the qPCR machine.

8. Test primers for their efficiency on the qPCR machine. You will need 3 technical reps for each of 6 3-fold dilutions for your DNA of interest and 3 reps of a water (-) control. If you want to test for non-specific amplification of bacterial endophytes, you also need 3 reps of a mix of 10 common bacterial DNAs. Primers need to be between 90% and 110% efficiency and amplify the dilution curve linearly, and not pick up a strong signal (Ct at least 35+) from the bacterial DNA, or water samples. Only primers that pass these hurdles should be used.

## D.2 Example qPCR protocol: 384 well qPCR for 16 primers (13 tests and 3 controls) and 8 cDNAs

1. Prep DNA for plate: Take reactions from cDNAs _____and add 215µL water to get enough DNA for 2 384 well plates. Mark wells with a red dot once diluted.
2. Prep NoRT controls: Add 78µL sH20 to each NoRT reaction and mark wells with a red dot.
3. Make NoRT (-) for plate by taking 3µL of each NoRT and adding to NoRT Eppendorf.
4. **Add 9µL of NoRT mix to 16 specific wells.** Then add 9 µL DNA mix to each of the remaining 47 wells of the plate.
5. Add 125 µL of 20x SYBR solution to each of 3 1mL tubes of 2.5x RealMasterMix SYBR ROX. Keep SYBR covered in foil/shielded from light when possible.
6. Prep 16 sub-master mixes, all with 25x reaction (24 wells, 1 extra):
    a. F primer: 25 µL (add primers first; SYBR is light sensitive)
    b. R primer: 25 µL
    c. SYBR Mix: 225 µL (2x 114uL, don't let extra SYBR drip)
7. Add 10.5 µL sub-master Mix to DNA in each of 24 wells of the plate.

**Figure D.3.** Example Plate Layout. A – P on bottom of plate; 24-1 on right side of plate. A-P columns contain eight 2x repeated DNA samples, numbered rows contain sixteen 3x repeated primers.

No RT wells:

| | |
|---|---|
| A22 | I22 |
| B19 | J19 |
| C16 | K16 |
| D13 | L13 |
| E10 | M10 |
| F7 | N7 |
| G4 | O4 |
| H1 | P1 |

A→

| | | |
|---|---|---|
| 24 | 22690_3 | CSA1_4 |
| | Ref_2_dal_PP2A | ADR2_1 |
| 21 | Ref_3_dev_Helicase | 14080_2 |
| | RPS5_8? | ADR1_L2_1 |
| 18 | | |
| | Ref_4_abio_bHLH | 50180_2 |
| 15 | | |
| | 40060_1 | RPM1_2 |
| 12 | | |
| | MEKK4_2 | 27180_1 |
| 9 | | |
| | RPS2_3 | LAZ5_2 |
| 6 | | |

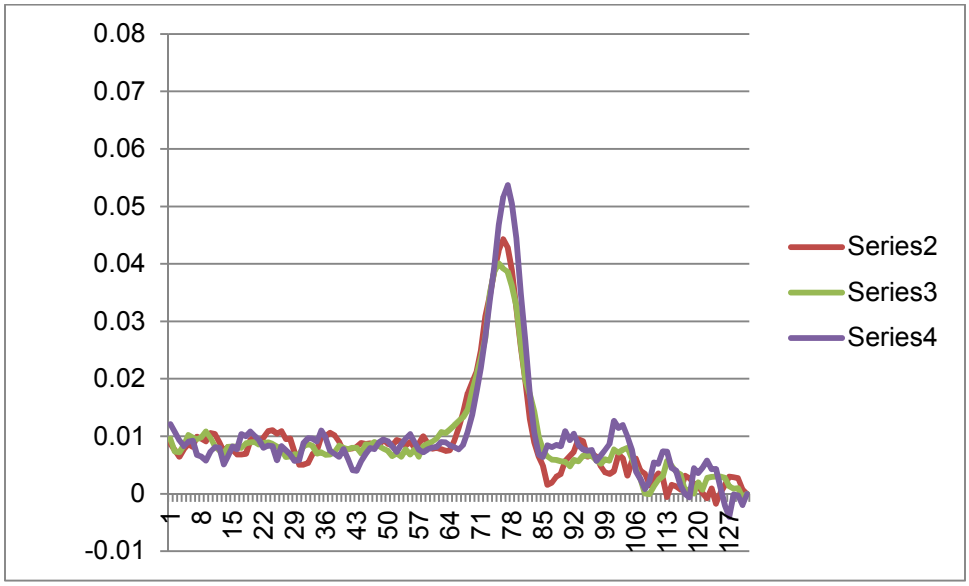
A … … … … H I … … … P

## D.4 qPCR analysis steps

1. Export the Dissociation, Results, and Clipped data files from the SDS program.
2. You basically don't want to do the rest of the analysis until you've done the qPCRs for the whole dataset. Then you can do the whole dataset at one time and save a lot of hassle.
3. Look at the Results file first. This will give general output for the run for each well of the 96 well plate, organized alphabetically by whatever primer names you named the wells for before you started the run. You're mostly looking at columns A through H for this analysis.

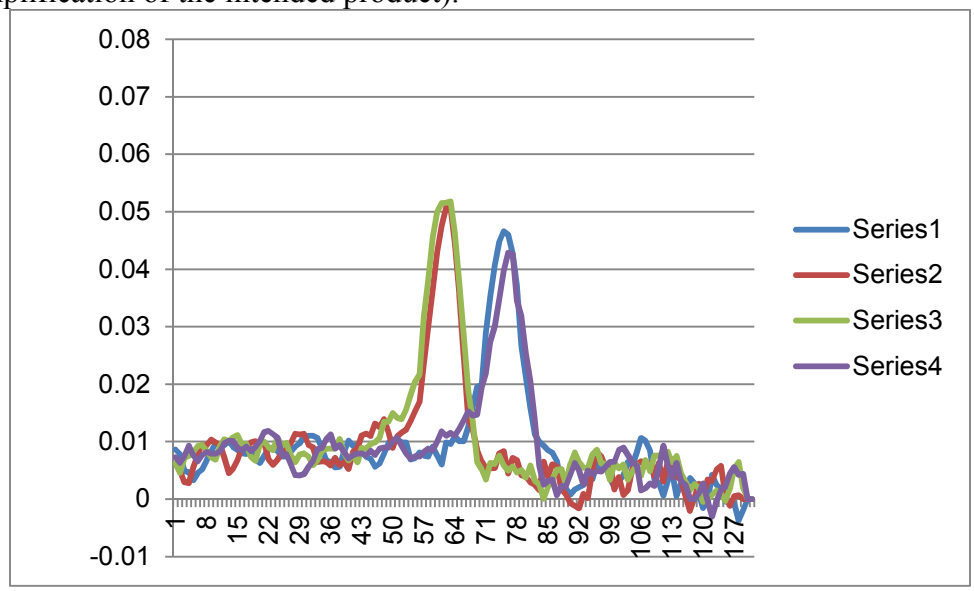
4. Screen the Results file for wells with abnormal melting temperatures. What is abnormal varies by primer set (and you should QC your primers to find out what the normal range is!) but typically Tm's less than 70 should be cause for concern and possibly removal of those wells as they usually indicate your product is a primer dimer, not what you were actually trying to amplify.
5. Based on your plate layout, you need to now add the sample name and number (in separate columns for later use) to each well. It helps to sort the file by well so you can copy and paste the samples in the correct order for each primer set. Well 1 is A1 on the plate, well 2 A2, and so on to well 96: H12.
6. Now take the well # through sample # columns and paste them into the bottom half of the "Dissociation" file (you'll probably have to add a couple of columns. Make sure the well numbers are matched up for the two documents!!!) The dissociation file measures the temperatures at which the DNA you amplified splits into two strands, or melts.
7. Now you want to look at the dissociation graphs for each well. You should sort the samples by primer type so you can compare several curves for the same primer to one another to look for outliers. Usually I compare 3 or 4 or so at a time and flag outliers by coloring in the name of that sample in yellow (if it's questionable) or red (if it's really bad). Here's an example of a good set of dissociation curves:

**Figure D.5.** A good set of dissociation curves from three technical replicates of qPCR.



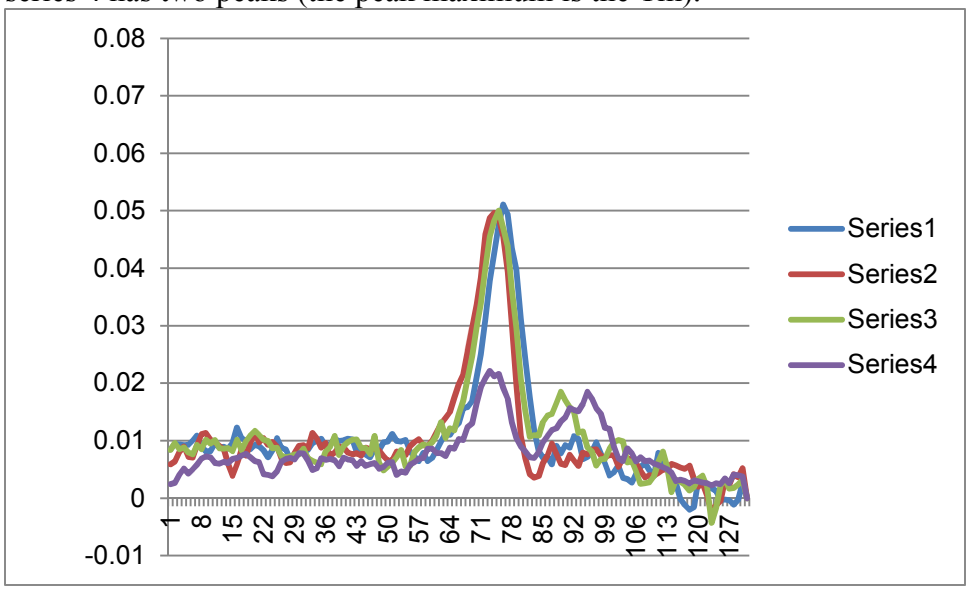

8. You basically want wells amplified by the same primer set to be the same thing (thus only one peak) and unique (no multiple peaks). Here's an example I would flag two of:

**Figure D.6.** Dissociation curves showing two wells, series 2 and 3, with possible primer dimers (no amplification of the intended product).



9. Series 2 and 3 have different peaks than 1 and 4, and are left-shifted, probably indicating primer dimers. Essentially, the "amount" of RNA called in the end is related to the area under the curve, so non-specific amplification (the wrong peak, or more than one peak) can really mess up your estimates for a particular RNA.

**Figure D.7.** qPCR Dissociation Curves with Multiple peaks (multiple products amplified). Here, series 4 has two peaks (the peak maximum is the Tm).




10. Wells flagged as problems are the big reason why you do technical replicates. Now what you need to do is compare the technical replicates, specifically the Tm and Ct that the Results file calls. If a problem well is similar enough to the other technical replicate in Ct (<1 different, usually) I'll keep both wells. If I've called both replicates as problems, I'll usually remove that primer set from further analysis. This is a ****really important step**** of qPCR that is a pain to do. If you don't do it you can get incredibly misleading qPCR results, basically because you are not measuring the expression of what you think you are measuring, the expression of your gene of interest, you are measuring something random and unknown.

11. Now you've flagged wells to drop from my analysis, you can do the analysis on the remainder of the wells. (NB: I'll usually do the analysis on the whole dataset anyway and delete the problem data at a slightly later point; otherwise, it's a pain to do the rest of the analysis…)
12. Concatenate the primer used and the sample and make a file from the Clipped data to run through my program in R. Making the file format acceptable to R is another trick. You need to transpose the first 40 rows of the clipped data to columns. The first column needs to be numbered 1-40 with the first row in that column empty. The first row should be the concatenated names for each of the 96 wells of your data.
13. Now export this file as a csv and open run the Ct and Efficiency R code, D.7 in this document. All you _should_ have to do is add a pcrimport2() line with your file name. Section D.8 here contains the relevant lines of that R program, so you should be able to run it and then just add a write.csv line to get your results.
14. The file it spits out should have 3 columns: the well number, Eff, and Ct. Paste these into a new sheet in your Results file, making sure the well number match!!!
15. Then, calculate some new columns: Block, AveEff, RQ, Norm, NRQ, and log2NRQ.
16. AveEff is the efficiency averaged across all wells for that particular primer set.
17. RQ is the relative quantity of your DNA of interest, calculated with the formula 1/E^Ct. E is AveEff and Ct is "Ctml1". An explanation of this formula can be found at: http://pathmicro.med.sc.edu/pcr/realtime-home.htm

18. Norm is the normalization factor so you can compare RQs between different samples. It is calculated as the geometric mean of the RQs of your three (at least 3) reference primers. So there is a different normalization factor for each sample.
19. NRQ is the normalized relative quantity of your DNA of interest, just the RQ for each well divided by the normalization factor for each sample.
20. Log2NRQ is the log base 2 of the NRQ. This quantity is normally distributed so you can do a lot more statistical analyses with it than with NRQ: build linear models, do averages, compare one quantity to another, test if one set is significantly different from another with a t-test, and so on. Good luck!

### D.8 Ct and efficiency R code

```
library(qpcR)
pcrimport2(file=Formatted_qPCR_Data.csv', sep=',', header=T, dec='.', quote="",
colClasses='numeric') -> allgenes
## Important note if you get errors in the following: FCT can be found if you use R x64 2.13.0
## Then if it still messes up, there's probably a well which was flagged and should have been
deleted.
mlallg2 <- modlist(allgenes,1,2:97, model=l4, backsub=1:8)
effml2 <- sapply(mlallg2, function(x) efficiency(x)$eff)
# barplot(as.numeric(effml2)) # optional to view output
Ctml1 <- sapply(mlallg2, function(x) efficiency(x)$cpD2)
# barplot(as.numeric(Ctml1)) # optional to view output
exportmatrix <- cbind(effml2, Ctml1)
```

# Appendix E

## High Throughput Infection Protocol

The following protocol is modified from one developed by Madlen Vetter. With it, a single researcher can easily infect with *Pseudomonas syringae*, harvest, and count colonies for up to 150 plants per week, or for 72-108 plants per round. If that doesn't sound impressive, you've probably yet to try infections in *A. thaliana*…

### E.1 Timing and general considerations

Due to the incredible variability in the number of colony forming units (CFU) resulting from infection of *A. thaliana*, 20 to 25 biological replicates with two leaves infected and measured per replicate are recommended. The inoculant that you will infect with needs to be started the night before infection. Plants are typically infected at the 21 day stage; the following protocol assumes you are infecting three 36-cell flats of *A. thaliana*. Jiffy pots are an incredible help in standardizing growth stage and thus CFU and are highly recommended for plantings for infection. Plants are typically harvested three days post infection. Plates can be counted about 24 hours after they are plated, and can be stored at 4ºC once colonies have reached the correct size for counting to prolong the amount of time you have for counting. If you're not sure the dilution factor that works best for your pathogen, you might do a small trial infection with a 25-fold dilution series of four dilutions to determine the best titer and dilution factor to do a larger experiment with. Counting colonies is the rate limiting factor in this protocol; if you have reliable help to count colonies, you could likely infect even more plants per round. Depending on how many rounds of plants your experiment requires, the following schedule works well and allows you to do two rounds of infection, harvesting, and counting of colonies per week.

**Table E.2.** Suggested schedule to do two rounds of infection for CFU per week. This schedule overlaps weeks, so an infection done on Thursday would be harvested on Monday and counted on Tuesday.

| Round 1 | Round 2 | Day |
|---|---|---|
| | Harvest (all day) | M |
| | Count (all day) | T |
| Harvest (all day) | Start Inoculant (5 min, pm) | W |
| Count (all day) | Infect (3 h, pm) | R |
| Start Inoculant (5 min, pm) | | F |
| Infect (3 h, pm) | | Sa |

## E.2 Materials and reagents

KB broth, KB plates (2-3 per 96 well plate) with appropriate selective agent

Sterile 10 mM MgSO4

Sterile mortar and pestles

Sterile deep well 96-well dilution plates (1 per 16 plants or 32 leaves infected)

Forceps, hole punch, silica beads, syringes

*P. syringae* strain of interest

21 day old *A. thaliana* plants (grown in Jiffy pots to standardize if possible)

Genogrinder

## E.3 Methods

Assuming a Thursday infection:

1. Monday: Remove bacteria from -80°C stock and plate on Petri dish (KB with antibiotic)
2. Monday: Incubate at 28°C for 48h.
3. Wednesday: Start O/N cultures of bacterial strains in 5 ml KB with antibiotic. Start growing around 4:30 pm, shaker, 28°C.

4. Thursday: Dilute O/N cultures 1:10 (500 µl in 5 ml KB + antibiotic). Grow with shaking at 28°C for 3-4 hours.
5. Thursday: Prepare inoculation solution:

• Centrifuge 1.5 ml bacterial culture for 5 min at 5000 rpm

• Remove supernatant, resuspend in 1 ml 10 mM $MgSO_4$

• Measure OD600 of 1:10 dilution in clean cuvette (first blank with 900 µl $MgSO_4$, then add 100 µl culture to cuvette)

• Goal: Infiltrate with 10^6 bacteria

• OD600 = 0.2 corresponds to 2.5*10^8 bacteria / ml

• First dilute bacterial culture to have a DO of about 0.2, then dilute 250x to get 10^6 bacteria (200 µl in 50 ml $MgSO_4$ with 25 mg/l streptomycin

• Syringe infiltrate two leaves per plant. Choose leaves by marking the base of each leaf with a pink/red sharpie. Choose leaves of similar/identical growth stages to standardize CFU. Include mock inoculated controls infiltrated with 10mM MgSO4 alone to control for contamination/endophytic bacteria.

• Let dry for ~ 1h before placing in high humidity

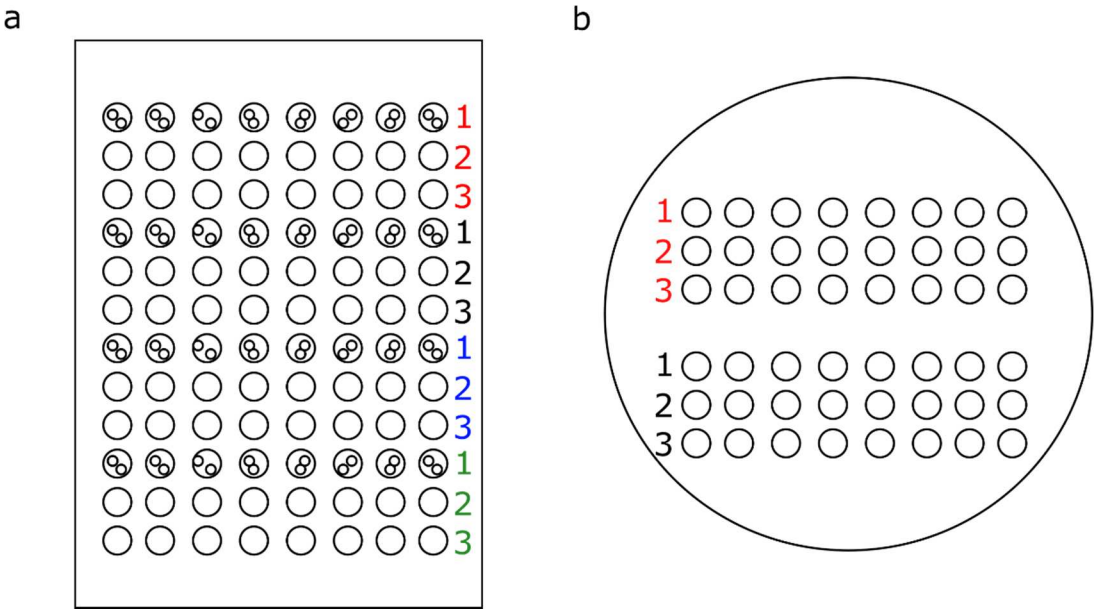
6. Thursday: To grow plants post infection:

• If growing in green-house, water plants after infection, then cover with domes with no holes to raise humidity.

• If growing in growth chamber, set humidity to 95-100% for the first 24h, for the remaining days, set humidity to 85%. *optional* cover with domes without holes, freshly washed with 70% ethanol, during the first 24h.

7. Harvesting the following Monday: pre-label harvesting sheet with identities of plants/leaves that will be placed in each well of the 96 well plate. If doing three dilutions per sample, place two silica beads in every well of every third row of a deep-well 96 well plate (Figure E.4a). Use 1/3 of wells for grinding and 2/3 for the dilution series. Add 200 µl 10 mM MgSO4 to wells with beads for grinding.

**Figure E.4.** Example plates for a three dilution series. A) 96-well plate. Put beads (small circles) in every third well and harvest plant leaf punches into these wells. Two dilutions from the 96-well plate (red and black numbers) can be plated per KB plate. B) KB plate with two sets of eight dilutions plated.




8. Pluck marked leaf with tweezers dipped in 70% ethanol. Remove leaf disc with a hole punch dipped in 70% ethanol. To surface sterilize leaf disc (optional): immerse leaf disc in 70% ethanol for 10-15 sec. Dry disc completely, place in 200 µL 10 mM MgSO4 in Eppendorf tube or 96-well plate. Place each leaf disc in a separate well with silica beads in the 96-well plate.

9. When all leaf discs are collected, place lids on 96-well plates and release bacteria by genogrinding for 1.5-3min at 1750rpm. Then spin down using the plate centrifuge going briefly (~1s) to maximum speed.

10. Now bacteria are released, continue plating/remove lids in the hood only. Allow KB plates to dry, open, in the hood for 30 minutes prior to plating for best plating conditions.

11. For a 40-fold dilution, fill the remaining 2/3 wells with 180µL 10mM MgSO4. Serial dilute 20µL from the first well into subsequent wells using a multipipettor.

12. Just before plating, add 800µL of 10mM MgSO4 to all wells of the 96-well plate. This should mix the samples sufficiently. Using a multipipettor, plate 5µL of each dilution series on the KB plate, pipetting from lowest to highest dilution on the plate. There should be enough room on each KB plate to plate 2 dilution series (Figure E.4b). Then, allow the plate to dry with the lid off.

13. Allow plates to grow at 28ºC for 12-16 hours or room temperature for 24 hours. Colonies can now be counted under a dissection scope. Use a sharpie to mark counted colonies on each plate. Determine the empirical dilution factor for each experiment by counting two dilutions on the plate whenever possible and averaging between the 1st and 2nd or 2nd and 3rd dilutions. Good luck!

# Appendix F

## Other Work: Unique Features of the m$^6$A methylome in *Arabidopsis thaliana*

For the following paper, I helped to carry out the experiments, including plant growth and harvesting, RNA extraction, and RT-qPCR follow-up. I developed a methodology to extract large quantities (>10mg) of RNA from *A. thaliana*, and carried out all data analysis except for GO categorization, miRNA analysis and microarray analysis. Full text follows; supplementary data and supplementary figures referred to in the appendix can be found at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4248235/#SD1. Full citation is as follows:

Luo GZ, **MacQueen A**, Zheng G, **(Luo, MacQueen, and Zheng are co-first authors)**, Duan H, Dore L, Lu Z, Liu J, Jia G, Bergelson J, He C (2014) Unique features of the m$^6$A methylome in *Arabidopsis thaliana*. *Nat.Comm*. **5**:5630. DOI: 10.1038/ncomms6630

## F.1 Abstract

Recent discoveries of reversible *N*$^6$-methyladenosine (m$^6$A) methylation on messenger RNA (mRNA) and mapping of m$^6$A methylomes in mammals and yeast have revealed potential regulatory functions of this RNA modification. In plants, defects in m$^6$A methyltransferase cause an embryo-lethal phenotype, suggesting a critical role of m$^6$A in plant development. Here, we profile m$^6$A transcriptome-wide in two accessions of *Arabidopsis thaliana* and reveal that m$^6$A is a highly conserved modification of mRNA in plants. Distinct from mammals, m$^6$A in *A. thaliana* is enriched not only around the stop codon and within 3′ untranslated regions (3′ UTRs), but also around the start codon. Gene ontology analysis indicates that the unique distribution pattern of m$^6$A in *A. thaliana* is associated with plant-specific pathways involving the chloroplast. We also discover a positive correlation between m$^6$A deposition and the mRNA abundance, suggesting a regulatory role of m$^6$A in plant gene expression.

## F.2 Introduction

$N^6$-methyladenosine ($m^6A$) is the most prevalent internal messenger RNA (mRNA) modification (Fu *et al.,*2014; Machnicka *et al*., 2013; Meyer and Jaffrey 2014) in eukaryotes including mammals (Wei *et al.,*1975), plants (Nichols 1979; Zhong *et al.,*2008), *Drosophila* (Levis and Penman1978) and yeast (Clancy *et al.,*2002), as well as viruses with a nuclear phase (Krug *et al.,*1976). This modification is installed by $N^6$-adenosine methyltransferase. A 70kD SAM (S-adenosylmethionine)-binding subunit methyltransferase like 3 (*METTL3*, also called *MT-A70*) was identified as one component of a methyltransferase complex in mammalian cells (Bokar *et al.,* 1994). Recent studies have characterized this complex, which consists of *METTL3*, methyltransferase like 14 (*METTL14*) and Wilms tumor 1 associated protein (*WTAP*) (Liu *et al.,* 2014; Wang *et al.,* 2014; Ping *et al.,* 2014). *METTL14* and *METTL3* are two active methyltransferases that form a heterodimer to catalyze $m^6A$ RNA methylation, while WTAP interacts with this complex and substantially affects the mRNA methylation inside cells but not *in vitro* (Liu *et al.,* 2014). Knockdowns of these methyltransferases affect mouse embryonic stem cell differentiation (Wang *et al.,* 2014). Since 2011, two $m^6A$ RNA demethylases of *FTO* and *ALKBH5* have been discovered; these demethylases are involved in mammalian development, RNA metabolism and fertility (Jia *et al.,* 2011; Zheng *et al.,* 2013). These findings reveal the first examples of reversible RNA modification and indicate regulatory functions of reversible $m^6A$ methylation on mRNA and certain non-coding RNAs that contain $m^6A$ (Jia *et al.,* 2013). Subsequent profiling of $m^6A$ distributions in mammalian transcriptomes (Dominissini *et al.,* 2012; Meyer *et al.,* 2012) and the recent mapping of the yeast $m^6A$ methylome in the meiotic state (Schwartz *et al.,* 2013) further confirm the dynamic nature of

m$^6$A modification. These studies revealed that m$^6$A is enriched around the stop codon and at 3′ UTRs, as well as in long internal exons and at the transcription start site (TSS) (Dominissini *et al.,* 2012; Meyer *et al.,* 2012; Schwartz *et al.,* 2013). Cellular proteins have also been found to preferentially bind m$^6$A-containing RNA (Dominissini *et al.,* 2012; Wang *et al.,* 2014). The human YTH domain family 2 (*YTHDF2*) has recently been characterized to specifically recognize m$^6$A-methylated mRNA and accelerate the decay of the bound mRNA (Wang *et al.,* 2014).

In *Arabidopsis thaliana*, the m$^6$A content in mRNA varies across tissues with a high ratio of m$^6$A/A found in flower buds (Zhong *et al.,* 2008). This variation correlates with the expression levels of the plant methyltransferase *MTA* (the plant homolog of human *METTL3*, encoded by *At4g10760*) (Zhong *et al.,* 2008). Previous studies have also shown that m$^6$A predominantly locates at the 3′ end of transcripts in a region 100-150 bp before the poly(A) tail in *A. thaliana* mRNA (Bodi *et al*., 2012). Inactivation of *MTA* prevents the progression of the developing embryo from passing the globular stage; an embryo-lethal phenotype with seed arrestment has been observed (Zhong *et al.,* 2008). Reduced expression of *MTA* in *A. thaliana* leads to decreased m$^6$A level in mRNA and abnormal growth with reduced apical dominance, abnormal organ definition, and increased trichome branching (Bodi *et al*., 2012). These data demonstrate that m$^6$A in mRNA plays functional roles in plant development.

In order to further investigate the functions of m$^6$A and to facilitate future studies of m$^6$A in plants, we report here transcriptome-wide m$^6$A profiling in two accessions of *A. thaliana*, Can-0 and Hen-16. These accessions are wild-collected natural lines from the two extremes of the natural range of photosynthetically active radiation (PAR) in the spring (Hancock *et al.,*

2011). We show that $m^6A$ is a highly conserved RNA modification in mRNA across these two accessions. Intriguingly, $m^6A$ in *A. thaliana* is enriched not only around the stop codon and within 3′ UTRs, as in yeast and mammalian systems, but also around the start codon, a property distinct from other known $m^6A$ methylomes (Dominissini *et al.,* 2012; Meyer *et al.,* 2012). A positive correlation between $m^6A$ deposition and mRNA levels indicates a regulatory role of $m^6A$ in plant gene expression.

## F.3 Methods

### *F.3.1 Plant Material*

Seeds from the Can-0 and Hen-16 accessions of *Arabidopsis thaliana* were sown in 50:50 Metromix 200:Farfad C2 soil in 48-cell flats and stratified for 5 days at 4 °C to synchronize germination. Seedlings were germinated in controlled-environment growth chambers at the University of Chicago greenhouses on a 16-hour light, 8-hour dark cycle. Plants were thinned between days 5-7 of growth. Above-ground tissue was harvested between the fifth and seventh hour of the light cycle on day 21. The tissue was flash frozen in liquid nitrogen, ground using a mortar and pestle, and stored at −80 °C.

### *F.3.2 High-throughput $m^6A$ sequencing*

To obtain sufficient (10 mg) total RNA for immunoprecipitation of $m^6A$-containing mRNA, approximately 200 plants from each accession were harvested and pooled in 0.6 and 2.1 g quantities. RNA was extracted using a protocol modified from a published procedure (Onate-Sanchez & Vicente-Carbajosa 2008); reactions took place in 15 ml or 50 ml conicals, and reagents were scaled up linearly with respect to the increased tissue mass, with 20 or 70 times the amount of each reagent, respectively. RNA was tested for quality via Nanodrop and gel

electrophoresis. Polyadenylated RNA was extracted using FastTrack MAG Maxi mRNA isolation kit (Invitrogen). RNA was randomly fragmented to ~200 nt by RNA Fragmentation Reagents (Ambion). Fragmented RNA was incubated for 2 h at 4 °C with $m^6A$ antibody (Synaptic Systems Cat. No. 202003, diluted to 0.5 μg $\mu l^{-1}$) in IP buffer (50 mM Tris-HCl, 750 mM NaCl and 0.5% Igepal CA-630) supplemented with BSA (0.5 μg $\mu l^{-1}$). The mixture was then incubated with protein-A beads and eluted with elution buffer (1× IP buffer and 6.7 mM $m^6A$). Eluted RNA was precipitated by 75% ethanol. The eluted RNA was treated with RNasin (Ambion Cat No. AM2694) according to the manufacturer's instructions. TruSeq Stranded mRNA Sample Prep Kit (Illumina) was used to construct the library from immunoprecipitated RNA and input RNA according to a published protocol (Trapnell *et al.,* 2012). Sequencing was done on an Illumina HiSeq machine with 2x100 cycles Solexa paired-end sequencing.

### *F.3.3 RT-qPCR validation for $m^6A$ enriched genes*

Eleven genes enriched in the $m^6A$ IP and six genes not differentially expressed between the IP and non-IP samples were tested by RT-qPCR. Data cleanup and analysis proceeded as described by previous protocol (Rieu and Powers 2009). The dissociation curves for each reaction were plotted and those with irregular features were removed. RNA passed through a beads only column, which should not bind any RNA, was treated as the input control for the IP step. Ct values from qPCR on the flow-through from the $m^6A$ IP and the $m^6A$ IP were expressed as the percent input of the beads only sample. Primer sequences are listed in Supplementary Data 9. Detailed information for plotting the qPCR figure can be found in an online manual at: http://www.lifetechnologies.com/us/en/home/life-science/epigenetics-noncoding-rna-research/chromatin-remodeling/chromatin-immunoprecipitation-chip/chip-analysis.html

### *F.3.4 Data Analysis*

Sequence data were analyzed according to the procedure described by Meyer *et al.,* (2012). Briefly, Tophat (Trapnell *et al.,* 2012) with Bowtie (Langmead and Salzberg 2012) was run in order to align the input and IP sequenced samples to the Columbia reference genome and annotation file (Tair10) (Lamesch *et al.,* 2012) The BEDTools tool (Quinlan and Hall 2010) was used to divide the aligned accepted hits into 1 bp intervals. The read depth was then averaged for each 25 nt discrete, non-overlapping genomic window using an ad hoc R program (Supplementary Data 10). To identify 25 nt windows enriched for $m^6A$, the number of reads that mapped to each window for the IP and input sample, and the total reads for each, were compared using Fisher's exact tests and corrected for multiple testing using Benjamini-Hochberg to reduce FDR to 0.05. To determine which of these windows cluster to form distinct peaks, we concatenated adjacent significant windows together and filtered out peaks less than 100 nt in length. Significant peaks with FDR<0.05 in Can-0, Hen-16, or both were annotated using an ad hoc R script (Supplementary Data 10). IntersectBED with the Columbia reference genome and annotation file was used to further annotate this set of significant peaks (Quinlan and Hall 2010). Peaks that shared more than 50% overlapping length were defined as recurrent peaks. For a peak to be classified as strain-specific, it should not overlap (1 nucleotide) any peak in any two replicates of the other strain. Sequence motifs were identified by using Homer (Heinz *et al.,* 2010). Gene expression was calculated by Cufflinks using the input sequencing reads (Trapnell *et al.,* 2012). Cuffdiff was used to find the differentially expressed genes (DE genes) between Can-0 and Hen-16 (Trapnell *et al.,* 2012). Gene function analysis (GO enrichment) was performed with the DAVID tool (Huang *et al.,* 2007). Plant miRNA targets were predicted by

psRobot (Wu *et al.,* 2012). Microarray data were downloaded from GEO (accession ID: GSE349243) and RMA method (Gautier *et al.,* 2004) was introduced to calculate gene expression and differentially expressed genes between *mta* mutant and wide type.

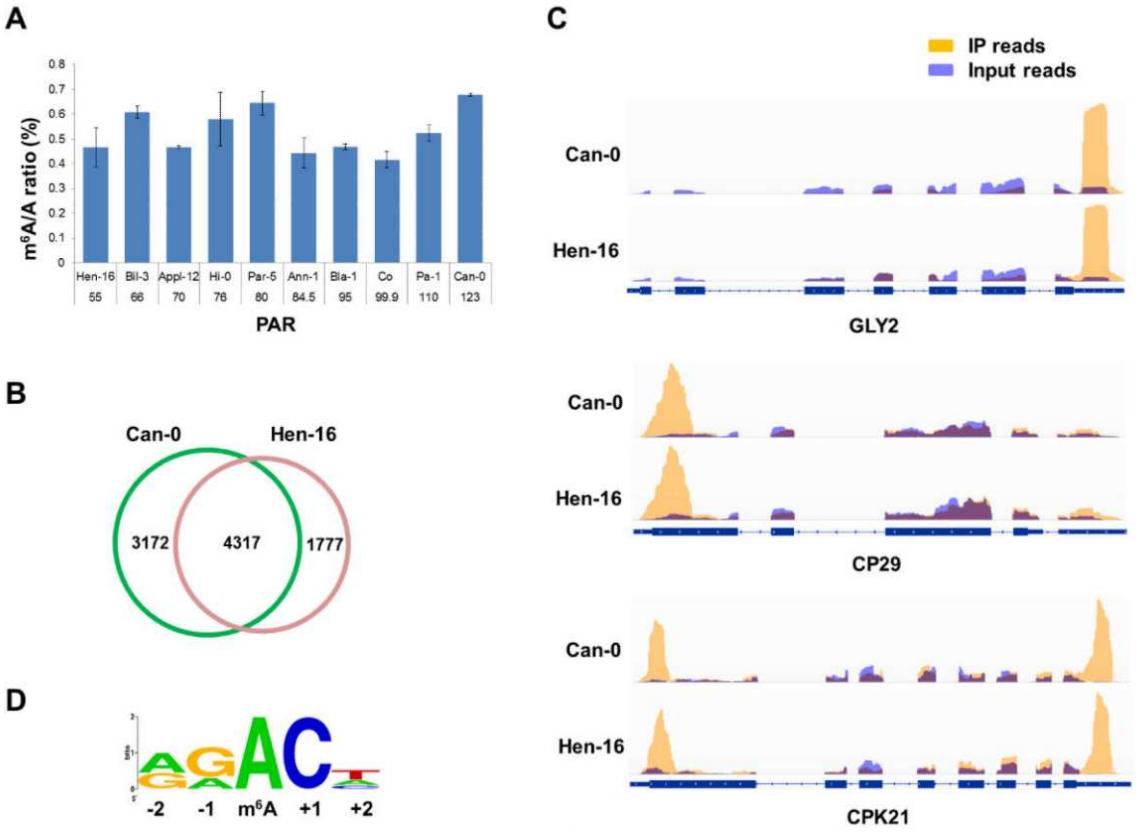
## F.4 Results

### *F.4.1 m$^6$A is abundant and conserved in* A. thaliana *mRNA*

m$^6$A is known to be a relatively abundant internal modification in *A. thaliana* mRNA[6]. We selected ten geographically diverse accessions of *A. thaliana* to grow in a common laboratory environment in order to measure the m$^6$A/A ratio of purified mRNA (Supplementary Table 1). These wild-collected natural lines were collected from sites that vary widely in PAR values (Hancock *et al.,* 2011). We observed that the ratio of m$^6$A/A in total mRNA from these ten accessions varied within the range of 0.45-0.65% (Figure F.1a), although not directly related to PAR, suggesting that the m$^6$A methylation level in mRNA is relatively stable but potentially affected by complex environmental factors.

To obtain the transcriptome-wide m$^6$A map of whole *A. thaliana* plants, we interrogated two accessions (Can-0 and Hen-16) using the m$^6$A-targeted antibody coupled with high-throughput sequencing (Dominissini *et al.,* 2012; Meyer *et al.,* 2012). Can-0 was originally collected from the Canary Islands where PAR in spring is 123.74, the highest seen for 1,191 accessions of *A. thaliana* (Hancock *et al.,* 2011). Hen-16 was obtained from northern Sweden where spring PAR is 55.29 (Hancock *et al.,* 2011), the lowest end of the range. The m$^6$A level in mRNA isolated from Can-0 is higher than that from Hen-16 as measured by LC-MS/MS (Figure F.1a).

**Figure F.1.** Overview of m6A methylome in *A. thaliana.* a) The m6A/A ratio of mRNA isolated from each *A. thaliana* strain. Error bars are calculated as the standard deviation from three replicates. PAR values are displayed below the strain names. b) Numbers of strain-specific and common m6A peaks. c) Examples of m6A peaks conserved between Can-0 and Hen-16. Orange color represents IP reads while blue color represents input reads. The purple color comes from mixing orange with blue. d) The RRACH conserved sequence motif for m6A-containing peak regions.




More than 70% of the m$^6$A peaks of Can-0 and Hen-16 were consistently detected in two biological replicates for each accession. We used these recurrent peaks as high-confidence m$^6$A sites for further analysis. In total, we identified 7,489 m$^6$A peaks representing the transcripts of 6,289 genes in Can-0, and 6,094 m$^6$A peaks representing transcripts of 5,416 genes in Hen-16 (Supplementary Data 1). Among them, 4,317 m$^6$A peaks were detected within both Can-0 and Hen-16 (P<1e-5, Chi-squared test), indicating that m$^6$A is highly conserved across *A.*

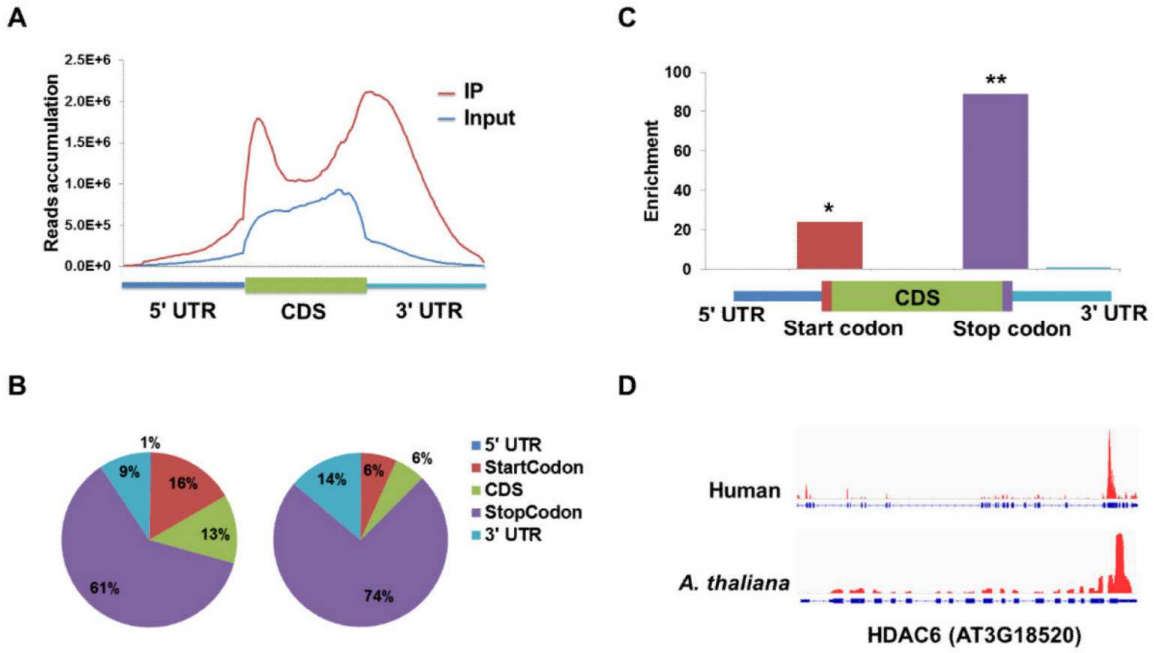
*thaliana* accessions (Figure F.1b-c, Supplementary Data 1). We validated 11 m$^6$A peak containing genes by RT-qPCR and all of them showed significant enrichment in IP-pulldown samples (Supplementary Fig. 1). Based on these results, we estimated that the *A. thaliana* transcriptome contains 0.5-0.7 m$^6$A peaks per 1,000 nucleotides, or 0.7-1.0 m$^6$A peaks per actively expressed transcript (Supplementary Table 2, Supplementary Fig. 2). These levels are comparable to those obtained in mammals (Dominissini *et al.,* 2012). The relative abundance of m$^6$A peaks in mRNAs from Can-0 and Hen-16 is consistent with total m$^6$A levels measured in mRNAs isolated from these two accessions (Figure F.1a-b).

To determine if the m$^6$A peaks that we identified contained the m$^6$A consensus sequence of RRACH (where R represents purine, A is m$^6$A, and H is a non-guanine base) (Wei *et al.,* 1976; Schibler *et al.,* 1977), we analyzed the top 1,000 most significant peaks (Supplementary Fig. 3). We found at least one such motif in 934 peaks (Figure F.1d). The same sequence motif appears to be necessary for m$^6$A methylation in plant mRNA as has been observed in mammals and yeast mRNA (Supplementary Fig. 4) (Dominissini *et al.,* 2012; Meyer *et al.,* 2012; Schwartz *et al.,* 2013).

### *F.4.2 m6A distribution exhibits a distinct topology in* A. thaliana

We next analyzed the distribution of m$^6$A in the whole transcriptome for both strains of *A. thaliana*. We determined the distribution of m$^6$A reads along transcripts in the m$^6$A-IP and non-IP (input) samples, respectively. Intriguingly, we found that reads from m$^6$A-IP are highly enriched around the start codon, stop codon and within 3′ UTRs in both strains (Figure F.2a and Supplementary Fig. 5). The prevalence of m$^6$A-IP reads around the start codon has not been observed in mammals or yeast.

**Figure F.2.** Distribution pattern of m6A peaks along transcripts. a) Accumulation of m6A-IP reads along transcripts. Each transcript is divided into 3 parts: 5′ UTRs, CDs and 3′ UTRs. b) The m6A peak distribution within different gene contexts. Left panel: total genes with m6A peaks; right panel: genes conserved in human and Arabidopsis. c) The m6A peak distribution along a metagene. Enrichment scores are calculated as. n: number of peaks belonging to each category; N: number of total peaks; p:proportion of each category within the genome by length. *P<2.2e-16, **P<1e-30. P-values are determined by Chi-squared test. d) An example of homologous genes with m6A peaks conserved in human and *A. thaliana*.




To further confirm the preferential locations of m$^6$A on transcripts, we investigated the metagene profiles of m$^6$A peaks. Consistent with the distribution of reads, m$^6$A peaks are abundant near the stop codon (61%) and start codon (16%), followed by the coding regions (CDs, 13%) and then 3′ UTRs (9%) (Figure F.2b). After segment normalization by the total length of each gene portion, we observed that m$^6$A is exclusively enriched around the start codon and stop codon (Figure F.2c). We mapped the number of m$^6$A peaks around the start codon, the stop codon and upstream of transcription termination site (TTS) (Supplementary Fig. 6a-c). The m$^6$A peaks are enriched at three locations: the region 60 bp downstream and 50 bp upstream of

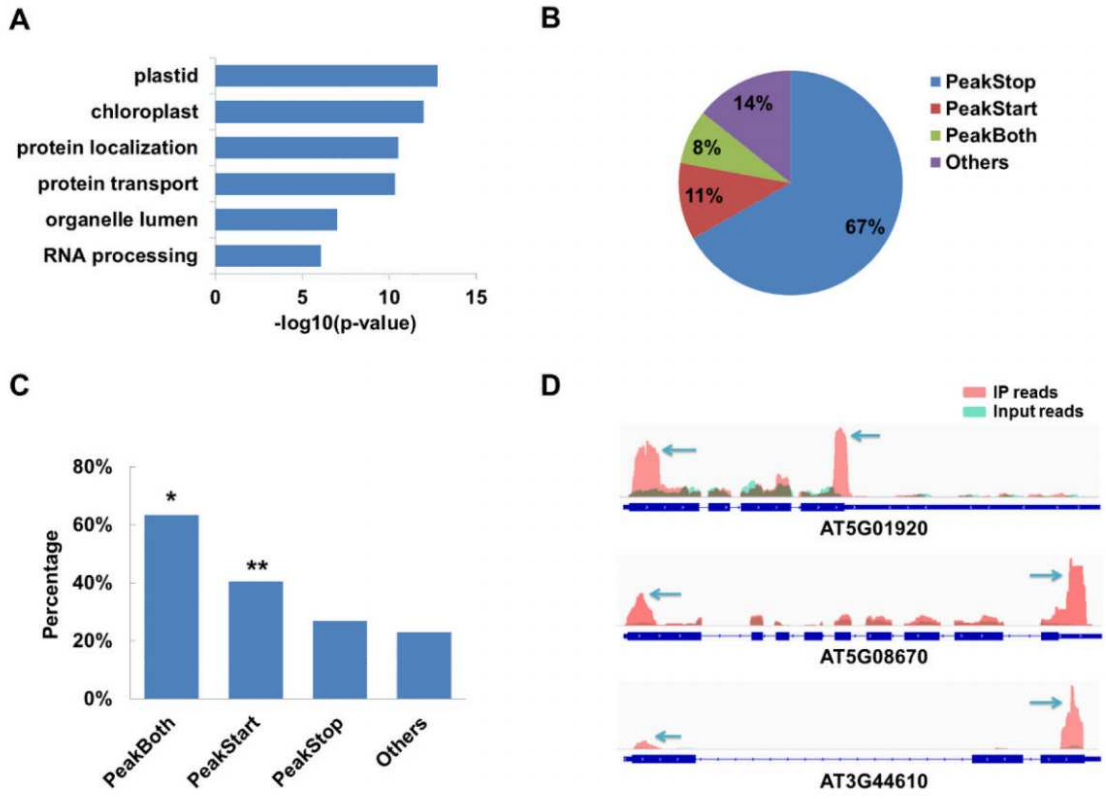
the start codon, the region 80 bp downstream and 100 bp upstream of the stop codon, and the region 80-220 bp before the poly(A) tail; the last location of enrichment is consistent with the previous finding that m$^6$A preferentially locates at the 3′ end of transcripts in a region 100-150 bp before the poly(A) tail in *Arabidopsis* (Bodi *et al.,* 2012). We focused on the m$^6$A peaks near the start codon and identified several new sequence motifs by comparing peak regions with scrambled background sequences, suggesting the possibility that distinct sequence motifs may exist for m$^6$A methylation in plant mRNA (Supplementary Fig. 7).

Using strict criteria of peptide similarity, we identified 1,684 m$^6$A-containing transcripts/genes in *A. thaliana* that are conserved in humans. Through comparison with the published data (Dominissini *et al.,* 2012), we found that 813 of these transcripts (48%) are also methylated in humans (Supplementary Data 2). Interestingly, almost all of the conserved methylated sites are located at the stop codon (74%) and 3′ UTRs (14%) of transcripts (Figure F.2b, F.2d), thus highlighting that m$^6$A peaks around the start codon are specific to *A. thaliana*. We depicted the m$^6$A peak distribution along transcripts in available datasets of human and mouse, confirming the observation that m$^6$A is exclusively enriched at the stop codon and 3′ UTRs in these systems (Supplementary Fig. 8). Clearly, the enrichment of m$^6$A near the start codon in *A. thaliana* is distinct. This feature of the m$^6$A distribution may play a unique role in plant-specific pathways.

### *F.4.3. m6A-containing mRNAs in important biological pathways*

The presence of m$^6$A is critical for normal plant development (Zhong *et al.,* 2008, Bodi *et al.,* 2012). In order to uncover further functional insights about m$^6$A in *A. thaliana*, we selected genes containing m$^6$A in both Can-0 and Hen-16, and identified the enriched gene ontology

**Figure F.3.** Functional annotation of genes with m$^6$A. a) GO enrichment analysis of all the genes with m6A peaks. GO categories are maintained by Gene Ontology Consortium. P-values are calculated using the DAVID tool. b) Percentages of subgroups of genes divided by the position pattern of m6A peaks. c) Percentages of genes characterized as chloroplast-related for each subgroup. *P<6.0e-25, **P<5.2e-18. P-values are calculated using the DAVID tool. d) Examples of chloroplast genes with m6A peaks at both the start and stop codon. The m6A-IP peaks are indicated by arrows.




(GO) terms using the DAVID tool. We found that these genes are highly enriched in chloroplast/plastid and protein transport/localization categories (Figure F.3a).

We next sought to determine if the unique m$^6$A position patterns are related to plant-specific GO categories. We classified genes into four subgroups according to the distribution of m$^6$A peaks: PeakStart (m$^6$A peaks around start codon), PeakStop (m$^6$A peaks around stop codon), PeakBoth (m$^6$A peaks around both start and stop codons) and others (Figure

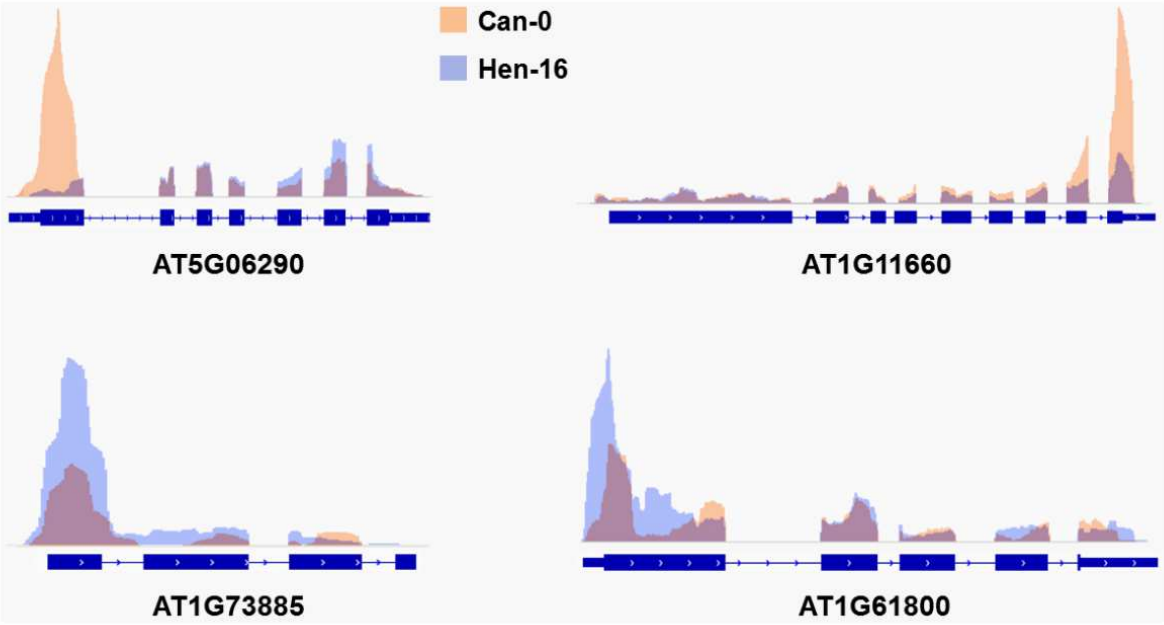
F.3b and Supplementary Data 1). Then we performed GO enrichment analysis for each subgroup. All subgroups of m$^6$A-containing genes exhibit high enrichment of chloroplast-related cellular components, indicating that a large proportion of these genes produce proteins localized in chloroplast though encoded by nuclear genome (Supplementary Fig. 9). Strikingly, more than 60% of genes belonging to the PeakBoth subgroup could be attributed to chloroplast components (Figure F.3c). Similarly, about 40% of the genes belonging to the PeakStart subgroup are associated with chloroplast. Photosynthesis is one of the most important functions in plants; 10%-20% of nuclear genes encode proteins that are imported to chloroplast (Soll *et al.,* 2004). Apparently, genes with m$^6$A enrichment around the start codon are highly enriched in chloroplast components.

The significant enrichment of chloroplast-related GO categories among transcripts with m$^6$A peaks around the start codon prompted us to examine photosynthesis-related genes in PeakStart and PeakBoth subgroups. Indeed, we identified dozens of well-studied photosynthesis-related genes that carry m$^6$A peaks (Supplementary Data 3). For instance, AT5G01920 (or STN8) is an important chloroplast thylakoid protein kinase that is specific to the phosphorylation of N-terminal threonine residues in D1, D2 and CP43 proteins and Thr-4 in PsbH of the photosystem II (Vainonen *et al.,* 2005). In our m$^6$A-IP data, the transcript of STN8 contains two clear m$^6$A peaks around the start and stop codons (Figure F.3d). The large fraction of m$^6$A-containing genes associated with chloroplast suggests a relationship between m$^6$A mRNA methylation and photosynthesis, one of the defining processes of plants.

### *F.4.4 Strain-specific m6A marking of mRNAs*

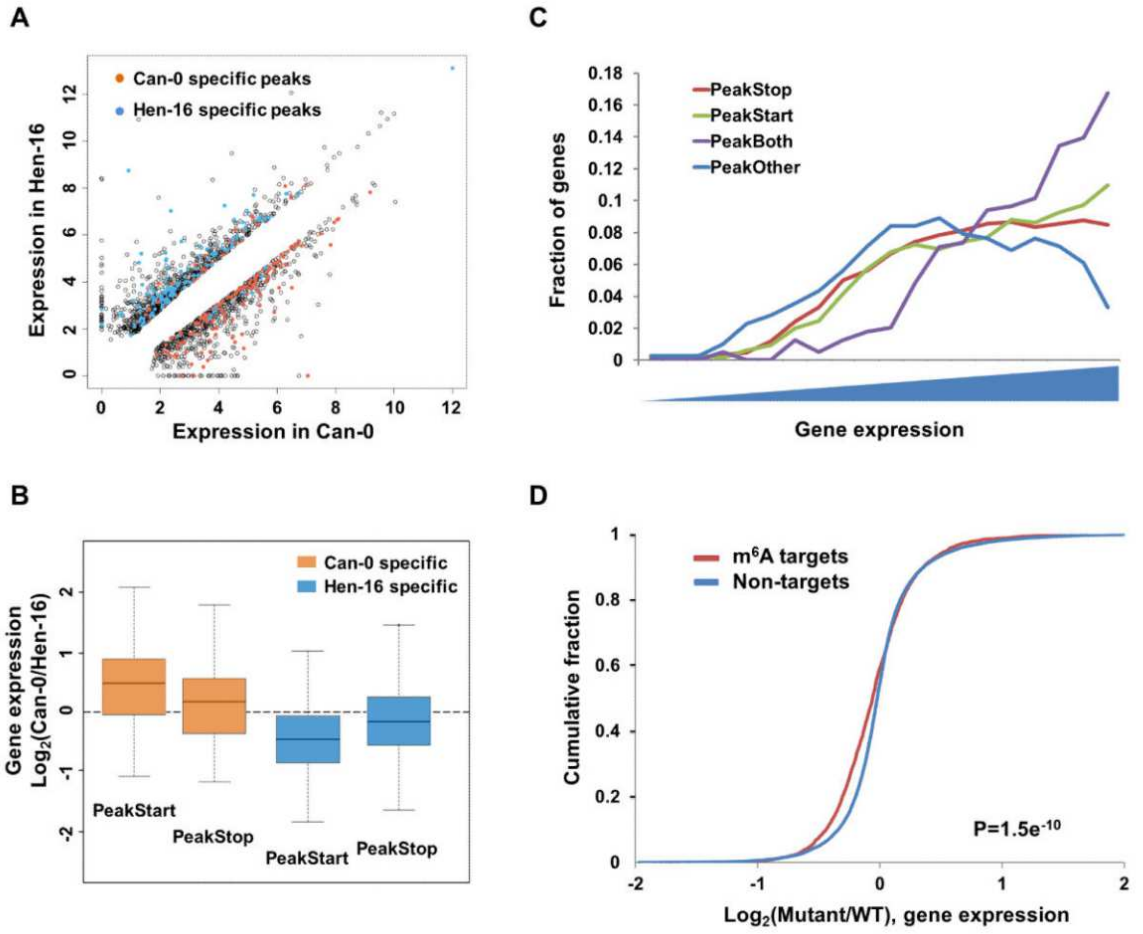
Although most m$^6$A peaks are shared between the two *A. thaliana* strains, we could detect a proportion of strain-specific peaks, even using strict criteria (Materials and Methods). In total, we identified 1,319 Can-0-specific peaks and 546 Hen-16-specific peaks (Supplementary Data 4). Furthermore, some of the common m$^6$A peaks in these two strains showed altered intensity (Figure F.4). GO analysis indicated that the genes with dynamic m$^6$A peaks are enriched in several categories of fundamental biological functions including mRNA metabolic process, response to stimulus, and regulation of translational elongation (Supplementary Data 5). For instance, *AT5G06290* (or *2-Cys Prx B*) has been shown to be sensitive to light intensities and oxidative stress (Horling *et al.,* 2003). In our m$^6$A-IP data, we observed significant m$^6$A peaks around the start codon of *AT5G06290* in Can-0 but not in Hen-16 (Figure F.4).

**Figure F.4.** Dynamic m6A peaks in two Arabidopsis strains. Different colors illustrate the accumulation of m6A-IP reads from two accessions.



As an initial exploration into the functional implications of $m^6A$ methylation differences across genomes, we asked whether $m^6A$ methylation could underlie expression differences. Using the RNA-Seq data, we calculated gene expression to assign differentially expressed genes (DE genes) of the two strains. Indeed, we found 801 more highly expressed genes in Can-0 (Can-high), and 1,011 more highly expressed genes in Hen-16 (Hen-high) (Figure F.5a and Supplementary Data 6), perhaps reflecting, at least in part, the substantial differences in PAR in the native habitat of these accessions. Within the Can-high list, we detected many more genes that contained $m^6A$ peaks in Can-0 than in Hen-16 (195/81, $P<0.01$, Fisher's exact test). Correspondingly, in the list of Hen-high genes, we detected more genes containing $m^6A$ peaks in Hen-16 than in Can-0 (221/140, $P<0.01$, Fisher's exact test) (Figure F.5a). This observation suggests that each strain possesses its own characteristic $m^6A$ methylation sites that appear to be associated with gene activation. This hypothesis seems to contradict the recent discovery that a main function of $m^6A$ is to mediate mRNA degradation in mammalian cells (Liu *et al.,* 2014; Wang Y *et al.,* 2014; Ping *et al.,* 2014; Wang X *et al.,* 2014). However, when we checked positions of the $m^6A$ peaks involved in up-regulated genes, we found that a large portion of these peaks are located at the 5′ end of the corresponding DE genes (Supplementary Data 6). To confirm this observation, we divided strain specific $m^6A$-containing genes into two groups and examined how their expression levels correlate with the locations of $m^6A$ peaks. Our analysis showed that $m^6A$ peaks at the 5′ end of transcripts correlate with higher expression levels of each strain. Genes in the PeakStart category possesses higher overall expression levels and correlate well with strain-specific $m^6A$ peaks (Figure F.5b, note that the gene expression ratios of Can-

**Figure F.5.** Relationship between m6A peaks and mRNA level. a) Differentially expressed mRNAs in Can-0 and Hen-16. Genes with Can-0-specific m6A peaks are highlighted in orange, and genes with Hen-0-specific m6A peaks are highlighted in blue. b) The ratio of mRNA expression levels in two samples containing strain-specific m6A peaks. Genes are divided to two categories (PeakStart and PeakStop) according to the peak positions. c) Fraction of genes belonging to each subgroup defined by the m6A distribution pattern. Genes are sorted by expression levels. d) Cumulative distribution of mRNA expression changes between the *mta* mutant and WT for m6A modified genes (red) and non-target genes (blue). P-values are calculated by two-sided Mann-Whitney test.




0/Hen-16 are shown). We further examined the fraction of each subgroup of genes based on their expression levels. Genes with both 5′ and 3′ m$^6$A peaks (PeakBoth) are enriched in the high-expression fraction, while genes with m$^6$A peaks at other locations (PeakOther) tend to be lower-expressed (Figure F.5c). Thus, m$^6$A enrichment at the 5′ end of the plant transcripts correlates

with higher expression level in *A. thaliana* in general, perhaps by stabilizing the transcripts via reader proteins or interaction with translation machineries.

In a recent study, Bodi and colleagues generated an $m^6A$ reduction mutant of *A. thaliana* and profiled gene expression (Bodi *et al.,* 2012). They identified 883 up-regulated and 654 down-regulated genes in the mutant plant (Bodi *et al.,* 2012). By comparing the differentially expressed genes with our list of $m^6A$ peaks, we found that 116 among 883 up-regulated transcripts are modified by $m^6A$ whereas 142 among 654 down-regulated transcripts contain $m^6A$ (Supplementary Data 7). These data suggest that transcripts carrying $m^6A$ tend to be down-regulated in the $m^6A$ reduction mutant ($P<0.01$, Chi-squared test). We extended this analysis to the whole transcriptome. Indeed, among genes with $m^6A$ modification, more are down-regulated than up-regulated in the $m^6A$ reduction mutant plant. This trend did not exist for genes without $m^6A$ modification (Figure F.5d). Together, their data supported our hypothesis that $m^6A$ tends to positively correlate with gene expression in a large fraction of transcripts in *A. thaliana*.

Mammalian microRNA (miRNA)-binding sites are mostly found within 3′ UTRs (Stark *et al.,* 2005); however, in plants they occur typically in the coding regions of target transcripts (Axtell *et al.,* 2008). Unlike mammals, miRNAs in plants recognize the target sites by near-perfect sequence complementarity. By comparing the transcripts containing $m^6A$ with the public miRNA targets database (TAIR) (Lamesch *et al.,* 2012), we found that only 5% of known miRNA-targeted transcripts contain $m^6A$. Furthermore, we employed a *de novo* miRNA target prediction method (Wu *et al.,* 2012) to inspect $m^6A$ peak regions. Among the top 1,000 most significant $m^6A$ peak regions, only 11 regions could be potentially targeted by miRNA

(Supplementary Data 8). Taken together, our results indicate that m$^6$A is unlikely to directly impact miRNA binding sites in *A. thaliana* although such an association has been proposed in animals (Meyer *et al.,* 2012). However, we could not exclude the possibilities that m$^6$A may affect miRNA maturation, or impact miRNA targeting sites through reader proteins of m$^6$A or structural changes induced by methylation.

## F.5 Discussion

The discovery of m$^6$A demethylases and the mapping of the m$^6$A methylomes in mammalian systems indicate that m$^6$A methylation of mRNA is a reversible and dynamic process with regulatory functions (Jia *et al.,* 2011; Zheng *et al.,* 2013; Jia *et al.,* 2013; Dominissini *et al.,* 2012; Meyer *et al.,* 2012; Nilsen 2014; Fu *et al.,* 2014). The importance of m$^6$A in post-transcriptional regulation of gene expression is further reinforced by the discovery and characterization of mammalian reader proteins that recognize m$^6$A modifications of mRNA and subsequently affect the stability of the target transcripts (Wang *et al.,* 2014). Previous studies have shown that m$^6$A plays a critical role in plant development (Zhong *et al.,* 2008; Bodi *et al.,* 2012). Here, we report the transcriptome-wide m$^6$A distributions in two accessions of *A. thaliana*, Can-0 and Hen-16. We found that m$^6$A is highly conserved across *A. thaliana* accessions. The methylation sites in a portion of transcripts are also conserved in corresponding transcripts of humans, indicating a fundamental functional role of m$^6$A in eukaryotes. Nevertheless, there are differences between the two *A. thaliana* accessions with a higher total m$^6$A level in Can-0 than Hen-16, and with ~1,400 more m$^6$A peaks identified in Can-0. The m$^6$A distribution could be influenced by differences in the accession's climates of origin and genetic

backgrounds. The modification sites could be conserved but the modification fraction at each site could vary depending on environmental factors.

Importantly, we discovered features of the m$^6$A distribution in *A. thaliana* mRNA that are distinct from those of mammals: i) m$^6$A in both accessions is enriched around the stop codon and at 3′ UTRs, as has been found in mammals, but also around the start codon; ii) the m$^6$A methylation around the start codon is heavily associated with the chloroplast, a photosynthesis organelle in plants. In addition, m$^6$A is less likely to be directly associated with microRNA recognition sites in *A. thaliana*. These distinct differences, namely the start codon enrichment of m$^6$A and its association with chloroplast, strongly suggest additional, plant-specific functions of this mRNA methylation. Previous studies indicated that the currently available m$^6$A antibody could also recognize *N*$^6$-,2′-O-dimethyladenosine (m$^6$Am) (Dominissini *et al.,* 2012; Schwartz *et al.,* 2013), which might lead to enrichment peaks at TSS. However, in our results, the start codon enrichment is clearly distinct from the m$^6$Am peak observed previously (Dominissini *et al.,* 2012) (Figure F.1). More accurate, base-resolution methods are highly desirable to determine the exact sites and modification fractions in the future.

Noticeably, the distinct enrichment of m$^6$A around the start codon correlates with the overall up-regulation of mRNA expression level. This relationship contradicts observations in mammalian systems, in which m$^6$A methylation around the stop codon and at 3′ UTRs is negatively correlated with gene expression (Liu *et al.,* 2014; Wang X *et al.,* 2014; Wang Y *et al.,* 2014). Published microarray data using a different accession of *Arabidopsis* (Col-0) also supports our findings with high correlations with our RNA-seq data (Supplementary Fig. 10) (Bodi *et al.,* 2012). Although our analysis demonstrates significant overlap of m$^6$A sites between

the two ecotypes (Can-0 and Hen-16), more precise analysis would benefit from future studies on coincident accessions. Our results suggest that m$^6$A reader proteins may exist in *A. thaliana* to recognize m$^6$A at the 5′ end and subsequently affect the stability of the target mRNA. This function could directly impact translation through the methylation itself or through the reader protein(s). It should be noted that several RNA binding proteins have been pulled down as potential "m$^6$A readers" with a few biochemically confirmed to specifically recognize methylated mRNAs (Dominissini *et al.,* 2012; Wang *et al.,* 2014). In particular, *YTHDF2* has been shown to bind and accelerate the decay of m$^6$A-modified mRNA (Wang *et al.,* 2014). However, the functions of other m$^6$A readers are still unknown; some of them could promote translation of a specific set of transcripts. Our discovery suggests the versatile roles of m$^6$A in plants beyond mediating mRNA decay.

It has been demonstrated that m$^6$A methyltransferase in *A. thaliana* is critical for normal plant development (Zhong *et al.,* 2008; Bodi *et al.,* 2012). We found that transcripts with m$^6$A are highly enriched in chloroplast/plastid and protein transport/localization categories, indicating a mechanism by which m$^6$A affects plant-specific metabolism. Past studies have demonstrated that light and circadian cycles impact the stability and translation of specific plant transcripts (Dickey *et al.,* 1998; Gutierrez *et al.,* 2002; Tang *et al.,* 2003; Juntawong *et al.,* 2012). However, the mechanisms of post-transcriptional gene regulation in response to light availability have not been clearly understood. The stability and translation of chloroplast mRNAs are known to be regulated by RNA-binding proteins that reside at 5′ UTRs and 3′ UTRs (Monde *et al.,* 2000).

Intriguingly, our data show that more than 60% of mRNAs containing $m^6A$ at both the start and stop codons encode proteins that could be components of the chloroplast (Figure F.3c). Given that 10%-20% of nuclear-encoded genes are associated with the chloroplast and photosynthesis in plant (Soll *et al.,* 2004), this percentage represents a significant enrichment. It is possible that the dynamic $m^6A$ methylation at 5′ and 3′ ends modulate the RNA affinities of RNA-binding proteins, thereby controlling mRNA transport and localized expression. Technologies that can provide more accurate measurements of the $m^6A$ sites are required in the future to gain deeper insights (e.g., splicing), the first $m^6A$ transcriptome-wide map of a plant species *A. thaliana* presented here provides a starting roadmap for uncovering $m^6A$ functions that may affect/control plant metabolism in the future.