SEQUENCING-BASED GENE DISCOVERY AND GENE REGULATORY VARIATION EXPLORATION IN PEDIGREED POPULATIONS

by

Robert Ebow McEwan

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Horticulture

West Lafayette, Indiana

August 2022

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Brian P. Dilkes, Chair

Department of Biochemistry

Dr. Joshua R. Widhalm

Department of Horticulture and Landscape Architecture

Dr. Jody Banks

Department of Botany and Plant Pathology

Dr. Kranthi Varala

Department of Horticulture and Landscape Architecture

Approved by:

Dr. Linda S. Prokopy

Dedicated to my family, for being there through thick and thin: My mom Felicia Mary Ghansah, my sisters Patricia and Nana Afariwa McEwan, my loving wife Nana Ama McEwan, and my adorable daughter Natalie Aba Nyarkoa McEwan.

ACKNOWLEDGMENTS

Foremost thanks to my major advisor Dr. Brian Dilkes for his invaluable support, motivation, and patience.

I would also like to offer my sincere thanks to my other committee members, Dr. Jody Banks, Dr. Josh Widhalm, and Dr. Kranthi Varala. Without your support this work would not have been completed.

I also owe a debt of gratitude to Dr. Steve Rounsley, my former manager and mentor, for encouraging me to pursue this degree and introducing me to Dr. Brian Dilkes.

Last, but not the least, I thank successive leadership at Corteva Agriscience, especially my current manager Dr. John A. Crow, for giving me the financial support and cheering me over the finish line.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Forward genetics discovery of the molecular basis of induced mutants has fundamentally contributed to our understanding of basic biological processes such as metabolism, cell dynamics, growth, and development. Advances in Next-Generation Sequencing (NGS) technologies enabled rapid genome sequencing but also come with limitations such as sequencing errors, dependence on reference genome accuracy, and alignment errors. By incorporating pedigree information to help correct for some errors I optimized variant calling and filtering strategies to respond to experimental design. This led to the identification of multiple causative alleles, the detection of pedigree errors, and an ability to explore the mutational spectrum of multiple mutagens in Arabidopsis. Similar to the problems in forward genetic discovery of mutant alleles, variation in genomes complicates the analysis of gene expression affected by natural variation. The plant hypersensitive response (HR) is a highly localized and rapid form of programmed cell death that plants use to contain biotrophic pathogens. Substantial natural variation exists in the mechanisms that trigger and control HR, yet a complete understanding of the molecular mechanisms modulating HR is lacking. I explored the gene expression consequences of the plant HR in maize using a semi-dominant mutant encoding a constitutively active HR-inducing Nucleotide Binding Site Leucine Rich Repeat protein, *Rp1-D21*, derived from the receptor responsible for perceiving certain strains of the common rust *Puccinia sorghi*. Differentially expressed genes (DEG) in response to *Rp1-D21* were identified in different genetic backgrounds and hybrids that exhibit divergent enhancing (NC350) or suppressing (H95, B73) effects on the visual manifestations of HR. To enable this analysis, I created anonymized reference genomes for each comparison, so that the reference genome induced less bias in the mapping steps. Comprehensive identification of DEG corroborated the visual phenotypes and provided the identities of genes influential in plant hypersensitive response for further studies. The locations of expression quantitative trait loci (eQTL) that determined the differential response of NC350 and B73 were identified using 198 F_1 families generated by crossing B73 x NC350 RIL population and *Rp1-D21*/+ in H95. This identified 3514 eQTL controlling the variability in differential expression between mutant versus wild-type. *Trans*-eQTL were dramatically arranged in the genome and identified 17 hotspots with more than 200 genes influenced by each locus. A single locus significantly affected expression

variation in 5700 genes, 5396 (94.7%) of which were DGE. An allele specific expression analysis of NC350 x H95 and B73 x H95 F_1 hybrids with and without *Rp1-D21* identified *cis*-eQTL and ASE at a subset of these genes. Bias in the confirmation of eQTL by ASE was still present despite the anonymized reference genomes indicating that additional efforts to improve signal processing in these experiments is needed.

CHAPTER 1. INTRODUCTION

1.1 Signal processing of sequence information

Forward genetics screens are the gold standard for understanding how genes control basic biological processes such as growth and development. The basic principles underlying forward genetics approaches are remarkably consistent across many biological systems (Candela & Hake, 2008; Forsburg, 2001; Jorgensen & Mango, 2002; Kile & Hilton, 2005; Page & Grossniklaus, 2002; Patton & Zon, 2001; Shuman & Silhavy, 2003; St Johnston, 2002). Briefly, mutations are induced in a population of individuals, typically in an accession or variety with a well-characterized genetic background, using physical or chemical mutagens. A screening of this mutant collection then identifies individuals with alterations in phenotypes of interest. Identifying the causative mutation behind the phenotype of interest entails mapping it to a chromosomal region through genetic linkage analysis and combing through candidate mutations within the mapped area. Confirming a mutated locus within a candidate gene as the causal variant behind the phenotype is a multistep procedure involving first genotyping wild type as well as mutants at the putative locus followed by complementation tests to demonstrate that a transgenic wild-type allele of the candidate gene can rescue the phenotype (Schneeberger, 2014).

Advances in Next-Generation Sequencing (NGS) technologies have spurred a rapid evolution in the methods of mapping and cloning mutations in most model organisms (Doitsidou et al., 2010; Minevich et al., 2012; Moresco et al., 2013; Obholzer et al., 2012; Schneeberger, 2014; Schneeberger et al., 2009). Consequently, the process of identifying causative mutations has been greatly simplified, and the length of time and effort required has also been significantly reduced (Doitsidou et al., 2016). Along with these advancements have also come new limitations and pitfalls. Error rates inherently associated with commercially available NGS platforms range from 0.1% to as high as 30% (Goodwin et al., 2016). Often this error affects a predictable subset of sites in the genome. For example, the sequence-by-ligation approach used by the SOLiD sequencing system, which enables a class-leading accuracy of about 99.99%, produces substitution errors and an under-representation of coverage at AT and GC-rich regions (Harismendy et al., 2009). Similarly, the sequence-by-synthesis approach of cyclic reversible termination utilized by the Illumina platforms affords it an overall accuracy of more than 99.5% but it too is limited by under-

representation in AT-rich and GC-rich regions (Harismendy et al., 2009; Minoche et al., 2011). Sequences produced through nucleotide addition approaches, as used by the Ion Torrent System and others, are riddled with insertion and deletion (indel) errors (Goodwin et al., 2016), and inaccuracies in homopolymer regions over 6-8 bp (Loman et al., 2012). Error rates on long-read sequencing instruments are even higher. For the PacBio platform — the most popular long-read sequencing technology — the rates of errors, mostly indels, have been as high as 15% (Carneiro et al., 2012). Moreover, the Oxford Nanopore technology has difficulties sequencing homopolymer regions and error rates for 1D reads that can go as high 30% (Jain et al., 2015)

Additional errors arise from sequence data analysis and are independent of the sequencing technology used to generate the data. These tend to only affect specific positions within the genome, rendering them prone to errors during variant calling. These sources of errors include misalignment of reads harboring indels and other structural variants or reads from genomic locations that have greatly diverged from the reference (R. Li et al., 2009). Others are reference assembly errors and poor performance of mapping tools (A. Y. Cheng et al., 2014), mapping to duplicated regions in the reference, and mapping reads from genomic regions that are poorly assembled or otherwise missing from the reference genome (Teo et al., 2012). While errors associated with sample processing and sequencing can usually be corrected through additional and costly deep sequencing (Addo-Quaye et al., 2017), errors from sequence processing, which collectively can result in erroneous inference, must be controlled to improve mapping and cloning procedures. The absence of efficient mitigation strategies and tools to exclude these errors often results in the production of false positives more than the actual changes induced by mutagenesis. We hypothesize that the rate of detecting false-positive variants can be limited through a combination of improved experimental design and optimized variant calling and filtering strategies. Reducing the false-positive rate will lead to a reduction in the number of time-consuming and costly validation experiments and render the process of identifying genes from mutants robust and rapid.

1.2 Signal Processing in Differentially Expressed Genes (DEG)

For over a decade now RNA-sequencing (RNA-seq), defined as massively parallel sequencing of RNA-derived cDNA libraries, has replaced microarrays as the tool to study differential expression of genes between conditions. RNA-seq simultaneously provides sequence

information on the entire transcriptome and if provided with an annotated reference sequence of all the genes can quantify transcript abundance (Stevenson et al., 2013). Specifically, sequence reads are mapped to the part of the genome that it is most similar to and through based on the positions of the genome annotated as corresponding to each gene the number of sequence reads associated with a gene is counted and used as a representation for transcript abundance (Mortazavi et al., 2008). The use of a single linear reference however has its limitations. Reads containing non-reference alleles, especially reads from hypervariable sections in the genome (Brandt et al., 2015), can be incorrectly mapped and confound downstream results. (Slabaugh et al., 2019) reported that the choice of reference not only impacted read alignment and identification of spliced variants but also affected differential gene expression analysis significantly. This was due to the occurrence of single nucleotide polymorphisms (SNPs) between the individual used for the transcriptome experiment and the reference as well as differences in annotation, both of which affect the number of sequence reads aligning to a gene. Graph aligners, which utilize genetic variation across a population to create bidirected DNA sequence graphs to serve as references, have been developed to decrease reference bias (Chen et al., 2020; Garrison et al., 2018). Furthermore, the use of "reference flow", which relies on multiple population reference genomes, has been proposed as a means of simultaneously increasing alignment accuracy and reducing reference bias (Chen et al., 2020).

Two major issues associated with using RNA-seq for Differential Gene Expression are deciding the optimal number of biological replicates per treatment and the ideal library sizes (or sequencing depth). Using insufficient number of biological replicates and sequencing depth results in low statistical power in detecting differentially expressed (DE) genes and inefficient use of sequencing resources (Y. Liu et al., 2014). A number of studies that deeply analyzed the impact of number of replicates and library size on the power and FDR of DE analysis have concluded that replicate number is more impactful to the power of DE analysis than library size generally, however for lowly expressed genes both factors have the same level of influence (Ching et al., 2014; Lamarre et al., 2018; Y. Liu et al., 2014; Schurch et al., 2016). Optimal library size for efficient DE analysis has been found to between 10 million (Y. Liu et al., 2014) and 20 million (Ching et al., 2014) reads per sample. (Lamarre et al., 2018) recommended an optimization for the number replicates as it was dependent on the FDR a study wished to achieve to be roughly 2^{-r}, where r is the replicate number. They further found that sensitivity of gene ontology (GO)

enrichment analysis is greatly improved by increasing the replicate number whilst increasing library size was found to enhance specificity. Based on a study with 48 biological replicates per treatment, (Schurch et al., 2016) propose using at least six biological replicates for each condition to generally discover significantly differentially expressed genes, and additionally recommend this number to be increased to at least 12 biological replicates per treatment if the desire is to identify significant DE genes across all fold changes.

1.3 Study of natural variation in transcript accumulation via eQTL

Expression quantitative trait loci (eQTL) are genomic loci that influence the expression of genes. The first genome-wide eQTL mapping was undertaken in yeast. Using expression data of all genes expressed among recombinant offspring of two yeast parental strains genomic regions containing variants that control gene expression were identified(Brem et al., 2002). Examples of eQTL mapping applications in plants can be found in Arabidopsis, maize, soybean, among others (Bolon et al., 2014; Cubillos et al., 2012; X. Wang et al., 2018). Classification of eQTLs is usually based both on their locations relative to the gene or genes they affect and the type of mechanism through which they affect gene expression. The latter classifies eQTLs into *cis*- and *trans*-acting while the former splits eQTLs into local and distant (Figure 1-1). Local eQTLs are located near, linked to, the genes they influence while distant eQTLs are unlinked or poorly linked to the genes they influence. The exact distance for classifying eQTL as local or distant varies by study

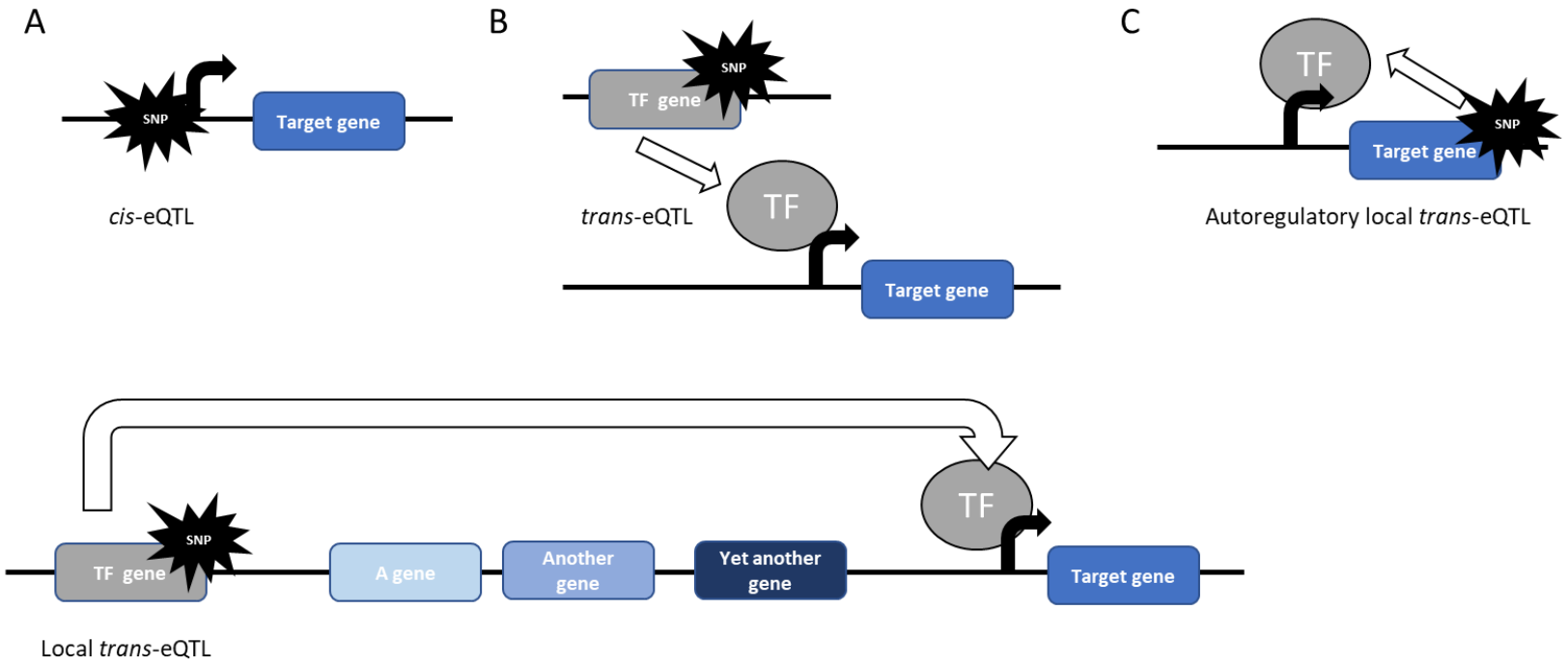


Figure 1-1. eQTL classification. Regulatory variants are usually classified in two ways: first, based on their location relative to the gene(s) they affect into local and distant eQTLs, and second, according to their mode of action into *cis*- and *trans*-eQTLs. Adapted from (Albert & Kruglyak, 2015).

Minimum distances of 2 Mb between eQTL and gene were used to distinguish local and distant eQTL classification in one study (Stranger et al., 2007). In another (Göring et al., 2007), distant eQTLs were only classified as such if they were found on different chromosomes than the genes they affected. With ever improving understanding of the consequences of gene expression at the organismal level eQTL analysis have the potential to uncover the precise biological mechanism through which DNA variation controls traits in the organism. Thus, eQTL studies can further aid prioritization of potential causal polymorphism among several candidates discovered through GWA.

A complementary method for identifying *cis*-regulatory changes is allele specific expression (ASE), which is characterized by differences in the expression of parental alleles within an F_1 hybrid (Edsgärd et al., 2016; Springer & Stupar, 2007; Wittkopp et al., 2004). A cis-regulatory allele can only alter the accumulation of the gene product encoded by the same DNA molecule as the allele (Figure 1-2). When ASE is detected in heterozygous individuals this provides proof for *cis*-regulatory changes between alleles by quantifying expression differences of the two alleles (Albert & Kruglyak, 2015).

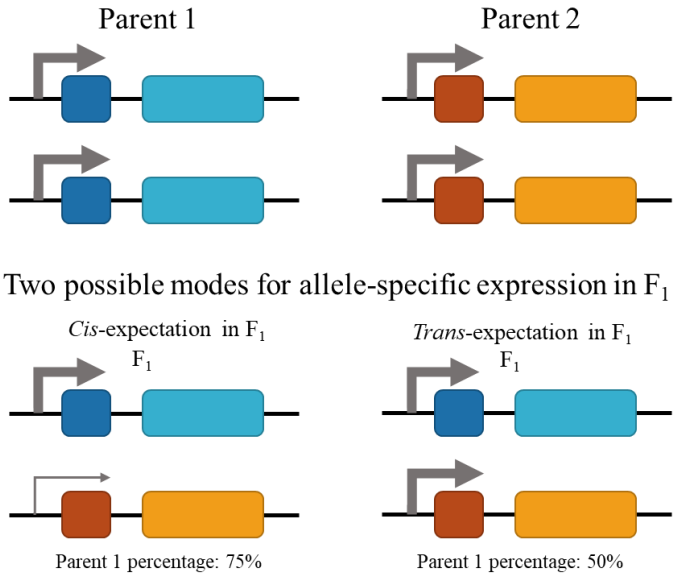


Figure 1-2. Mode of allele-specific gene expression in the F1. The inbred 1 allele is more highly expressed in comparison to the allele of inbred 2, the thickness of the arrows reflects this difference in expression. The allelic ratio in the F_1 would be tilted toward inbred 1 if the gene is under the influence *cis*-eQTLs. On the other hand, if the gene is under *trans*-eQTL control then both parental alleles would be balanced. Adapted from (Waters et al., 2017).

It is possible for genetic interactions to occur between *cis*-acting regions (such as promoters and enhancers) and *trans*-acting elements (such as transcription factors), these interactions are often difficult to distinguish from cis effects (León-Novelo et al., 2018). ASE has been found to play pivotal roles in abiotic stress response (Ereful et al., 2016; Waters et al., 2017), genomic imprinting and parent-of-origin effects (Springer & Stupar, 2007; Zhuo et al., 2017), and divergence in gene expression between species (Wittkopp et al., 2004). Like eQTL, ASE analysis can produce information that can help reduce the number of candidate genes to be investigated from GWA, and hence an important approach towards enhancing our understanding of the influence of genetic variation on molecular processes.

1.4 The hypersensitive response of maize as a case study of transcriptome remodeling

Plant cell death happens during normal growth and development but can also be instigated during pathogen invasion and upon exposure to adverse environmental stimuli such as toxicity. The process which is orderly and determined by a genetically controlled program is known as programmed cell death (PCD). Hypersensitive response (HR) is a subclass of PCD characterized by the killing of cells immediately surrounding the infection point (Coll et al., 2011). It is an effective defense response to several classes of plant pests and pathogens such as insects, nematodes, bacteria, fungi, and viruses (Wu & Baldwin, 2010).

Plants use high-affinity transmembrane pattern-recognition receptors (PRR)—typically leucine-rich repeat and lysine motif (LRR and LysM) kinases—to detect pathogen/microbial-associated molecular patterns (PAMPs or MAMPs). Detection of PAMPs triggers PRR-mediated immunity (PTI) leading to signaling for the transcription of immunity-related genes to inhibit the spread of the microbe. To be successful, pathogens must first circumvent the PTI response. This is achieved through the secretion of virulence effector proteins to plant cell apoplast to inhibit recognition of PAMP/MAMPs. Effectors are also delivered into host cells to block the PTI pathway. Plants use dominant resistance genes (R genes) to code for members of an extremely polymorphic superfamily of nucleotide-binding leucine-rich repeat (NLR) receptors as a second layer of immunity response known as effector-triggered immunity (ETI). ETI re-activates the effector–inhibited PTI pathway resulting in the expression of defense-related genes (Dangl et al., 2013). The process eventually results in localized cell death at the point of infection to hinder

further colonization by pathogens. The genetic control of PTI and ETI is not fully understood, hence further studies are needed to uncover the precise regulatory mechanism that ultimately leads to plant immunity (Chakraborty et al., 2018).

This immune machinery is kept in an inactive state in the absence of effector proteins. However, mutants have been discovered that show constitutive activation of NLRs in a pathogen independent manner resulting in autoimmunity. These autoimmune mutants can be identified by their stunted growth and characteristic disease lesions (patches of dead or damaged cells) on leaves and stalks (Chakraborty et al., 2018). Among early reports of autoactive R genes in plants was that of the *rp1* locus of maize (Chintamanani et al., 2010). The locus found on maize chromosome 10, is comprised of several tandemly.

Repeated R gene paralogs are shown to confer resistance to specific races of *Puccinia sorghi* Schwein, which causes common rust in maize. Frequent unequal crossovers between paralogs have rendered the locus highly unstable meiotically (Sudupak et al., 1993). One such unequal crossover and subsequent recombination event produced *Rp1-D21* (Collins et al., 1999; Smith et al., 2010). Recognition and elicitation functions in the resulting Rp1-D21 protein are separated, leading to spontaneous activation and production of disease-resembling lesions independent of pathogen recognition. The *Rp1-D21* gene exerts partially dominant action in the display of its lesion phenotype, and factors including genetic background, developmental stage, and the environment all affect the intensity of the phenotype. Using autoactive R genes to study etiology and genetics of HR removes the confounding effects from the pathogen (Chaikam et al., 2011; Chintamanani et al., 2010). This approach, named mutant-assistant gene identification and characterization (MAGIC) analysis, is an effective forward genetics method that makes use of the phenotype of a mutant as a reporter to discover and analyze naturally undetectable genetically-controlled variation (Johal et al., 2008).

The nested association mapping (NAM) population is the most extensive public collection of maize lines developed by the Maize Diversity Group to enable the systematic dissection of complex traits (McMullen et al., 2009). The 5000 recombinant inbred line (RIL) population comprising 25 subpopulations, each 200-line strong, was designed to allow linkage analysis and association mapping in a single integrated mapping population. This comprehensive mapping population was developed by crossing a common parent (B73) to 25 other diverse maize founder lines, namely Tzi8, Tx303, P39, Oh7B, Oh43, NC358, NC350, MS71, Mo18W, M162W, M37W,

Ky21, Ki11, Ki3, Il14H, Hp301, CML333, CML322, CML277, CML247, CML228, CML103, CML69, CML52, and B97. Subsequently, the F_1s were selfed to produce 25 individual segregating F2 populations. Each F2 population was then selfed to the F6 generation through single-seed descent to generate 200 RILs (J. Yu et al., 2008). The NAM population has been used as a powerful resource to dissect and map QTL for flowering time control (Buckler et al., 2009), resistance to northern leaf blight (Poland et al., 2011), leaf traits such as angle and size (F. Tian et al., 2011), leaf flecking (Olukolu et al., 2016), as well as drought tolerance (C. Li et al., 2016), to name a few.

Although HR is a vital plant immune mechanism, much remains unknown about how it is triggered and controlled. Attempts to reveal the genetic determinants of HR have been hampered by the rapid and highly localized nature of the response. Previously, (Olukolu et al., 2013) used a combined genome-wide association (GWA) and mutant-assistant gene identification and characterization (MAGIC) approach in maize to assess the phenotypes of segregating F1 progenies created by crossing a collection of 231 diverse inbred lines to an inbred line H95 carrying the *Rp1-D21* mutation (denoted H95;*Rp1-D21/+*). The strategy successfully identified six significant single nucleotide polymorphisms (SNPs) associated with *Rp1-D21*-induced HR. The methodology was subsequently expanded substantially by crossing H95;*Rp1-D21/+* to the 5000 NAM RIL population; 32 additive quantitative trait loci (QTL) were detected through joint linkage analysis, whereas GWA analysis with 7000 SNPs identified 44 significantly associated loci, 36 of which overlapped with the QTL identified with joint linkage analysis (Olukolu et al., 2014).

In the two studies described above, the closest genes to the significant SNPs or QTLs were proposed as candidate genes through which the variants act to generate the phenotype. However, the closest gene to an associated SNP may not always be the causative gene because variants can also act in trans. Thus, while these two previous studies and others have contributed immensely to a greater understanding of the genetic control of plant HR, a mechanistic understanding of how the identified loci influence HR is still lacking. As such, novel approaches are needed to fill these gaps in our knowledge. We propose a novel strategy that combines the MAGIC technique with eQTL and DGE analyses to reveal the precise biological mechanism through which DNA polymorphisms affect HR. This approach of conducting a genome-wide analysis of mRNA accumulation patterns in segregating populations can help re-construct the biochemical pathways of connected genes underlying HR. Since eQTL analysis treats gene expression as quantitative traits in a segregating population the mapped variants directly affect gene expression levels,

making it possible to draw conclusions about the functional consequences of the significant SNP loci. It is possible, with eQTL analysis, to uncover complex connections such as *trans*-acting gene interactions whereby one gene regulates the level of expression of another. We will further utilize allele-specific expression (ASE) analysis to cross-validate the *cis*-acting regulatory variation observed using eQTL analysis.

CHAPTER 2. THE USES AND DETECTION OF ERROR BY REPLICATION IN SEQUENCING EXPERIMENTS

2.1 Introduction to the scientific problem

Isolating genes responsible for mutant phenotypes, a process commonly referred to as forward genetics, has been vital in unearthing a wide array of biological processes. Traditionally, positional cloning was used to localize a causal mutation and then a limited number of genes were sequenced to provide molecular identity of the allele. This technique was expensive, time-consuming, and tedious mainly because as the limits of recombination-based mapping are reached, thousands of individuals must be genotyped to narrow the genomic region of interest containing the causative allele (Lukowitz et al., 2000; Sahu et al., 2020). Shotgun-sequencing and other sequencing-based approaches are relatively less costly and less time-consuming and have been used to successfully identify spontaneous (Ossowski et al., 2010) as well as induced mutations (Schneeberger et al., 2009) responsible for interesting phenotypes. Compared to spontaneous mutagenesis, induced mutagenesis has the added benefit of producing an increased number of mutations providing a source of genetic variation that can be harnessed for fundamental discovery of gene function and crop improvement efforts (Kumawat et al., 2019; Sahu et al., 2020). The ability to assess all positions in a genome for mutations has created a fantastic opportunity for gene function discovery but carries the problem of distinguishing the causative polymorphism for a change in phenotype from all other mutations visible in the data.

The ease of discovering the specific induced mutation responsible for a phenotype can be hampered by the massive number of variants produced, which increases with genome size. Typical treatments of plant genomes with the chemical mutagen ethyl methanesulfonate (EMS) generates thousands of mutations per individual (Comai & Henikoff, 2006). Thus, methodologies that can efficiently distinguish candidate causal variants from noncausal ones are key to the success and efficiency of forward genetics efforts. One approach, similar to positional cloning, is to use mapping information to narrow down the genomic region harboring the causative variant (Sarin et al., 2008). Another approach that uses recombination as well, but does not require the tedium of traditional mapping, relies on bulked segregant analysis to map causative mutation in a single sequencing run through alignments of short reads derived from DNA pooled from a large

segregating population to a reference sequence (Schneeberger et al., 2009). A third approach neither relies on prior mapping information nor a reference sequence. Rather, *k*-mers in whole-genome sequences of backcrossed recombinants are compared to discover homozygous induced mutations behind phenotypes of interest (Nordström et al., 2013). In this case, as with the bulked segregant analysis case, recombination in an experimental population will result in unique k-mers associated with the causative polymorphism in the sequencing data from a pool of affected individuals.

Extensive pedigree information is often available for mutants in crop species and can be recorded in model organisms. Such data can be leveraged in the service of mapping variants to eliminate preexisting variation and thereby exclude large genomic regions as candidate locations for causative mutations. If an organism is self-fertile, then crossing the mutant to a non-mutant plant followed by recurrent selfing of heterozygous F1 plants produces advanced single-lineage families with all loci driven to homozygosity except for the causal variant. This results in phenotypically-affected and unaffected siblings that only differ at the mutant locus and tightly linked sites. An approach that utilizes exclusion of variants common to affected and unaffected siblings as well as the elimination of previously identified variation to remove confounding noncausal mutations has been successfully demonstrated in *Sorghum* (Addo-Quaye et al., 2017) and Arabidopsis (Dolan et al., 2017; Silva-Guzman et al., 2016).

The plant species *Arabidopsis thaliana* has been the model system of choice for plant functional analysis for several reasons. First, it is suitable for classical genetics experiments due to its small stature, quick generation time of 5-6 weeks, ability to produce an abundance of offspring, relative ease of cultivation in controlled environmental conditions, and simplicity of maintaining mutants either through self-fertilization or out-crossing (Page & Grossniklaus, 2002).Second, the fully sequenced genome of Arabidopsis (125 Mb) is among the smallest (Kaul et al., 2000) and harbors relatively few repeats as compared to other plant genetics models (Page & Grossniklaus, 2002). Third, and perhaps most significant, Arabidopsis is amenable to Agrobacterium-mediated transformation. This has allowed for the generation of thousands of induced mutants that are available to the research community, serving as invaluable tools for experimental investigation of gene function. For these reasons Arabidopsis is an exceptional model to test the effectiveness of a strategy for the rapid and efficient detection of causative alleles in phenotypically affected mutants. This approach speeds up forward genetic screens in Arabidopsis

and can potentially be extended to more complex eukaryotic systems to aid the functional characterization of genes.

I leveraged the availability of extensive sequence data and collaborator's pedigree information in *Arabidopsis* to eliminate pre-existing polymorphisms and accelerate the discovery of causal polymorphisms in phenotypically affected mutants. By using whole-genome sequence information from phenotypically unaffected individuals from EMS populations and whole-genome sequence data from previously published unrelated mutants I was able to create a portable bioinformatics approach that can be implemented by anyone to greatly reduce the number of false-positive candidate variants. In addition, the inclusion of pedigree information and the very low rate of false positive variant positions allowed me to test the presumed pedigrees of lines, rather than rely on the reported pedigree as an assumption. This allowed me to discover errors in biological material handling. Early successes in this approach led me to explore the methods originally developed for EMS mutagenesis and point mutations for ionizing radiation mutagenesis and the discovery of insertion and deletion (indel) polymorphisms. The flexibility and robustness of this approach allows a test of the assumptions about the material and to save substantial effort for bench scientists by eliminating from further analysis materials that were assumed to be novel but are not.

2.2 Methods

A standard pipeline that used paired reads as input to produce high-quality variants as output was developed and applied to Case Studies 1 - 6 described below. Briefly, paired-end reads were mapped to the TAIR10.29 reference genome downloaded from The Arabidopsis Information Resource (TAIR) using BWA-MEM (H. Li & Durbin, 2009). Mapping statistics such as the number of QC-passed/failed reads, properly paired reads, and singleton reads were generated with SAMtools (H. Li et al., 2009) using the *flagstat* command with default options. Possible PCR duplicates, which are artifacts that can be introduced into the sample during library construction, were removed with the *rmdup* command of SAMtools. SNP and small indel calling was carried out in a two-step process. Summary information retrieval from the aligned files as well as computation of genotype likelihood was carried out by SAMtools *mpileup* command and subsequently the BCFtools *view* command was used to perform variant calling. Initial variant quality filtering was achieved using the *varFilter* command of the vcfutils.pl script with the *-D100*

option to produce an initial list of variants in a Variant Call Formatted (VCF) file. This option not only excluded variants with coverage depth of greater than 100 reads or less than 2 reads, it also eliminated variants with root mean square quality of less than 10 as well as variants detected within three bases of a gap (R. Li et al., 2009) Additional filtering was performed with SnpSift (Cingolani, Patel, et al., 2012) to further keep only variants that have a phred-scale quality of at least 20.

Several additional case study-specific processing steps were carried out to meet the goals of the specific experiment. For case studies 1 and 2 the causative mutation was expected to be a SNP, as such, indels were eliminated from the list of high-quality variants using the VCFtools *remove-indels* command to retain only SNPs. Contrastingly, in case study 3, an indel was expected to encode the causative mutation, hence VCFtools *--keep-only-indels* command was additionally executed to generate a set of high-quality indel variants. To further keep only homozygous SNPs (case studies 1 – 3, where a homozygous mutation was expected), the SnpSift command *filter "(((DP4[0] = 0) & (DP4[1] = 0)) & ((DP4[2] > 1) & (DP4[3] > 1)))"* was used. On the other hand, when the expected mutation is heterozygous (majority of mutations in case study 2), the SnpSift command *filter "((countHet() > 0) && (DP >= 25))"* was run to retain only heterozygous SNPs with a depth of at least 25 reads.

To create an experiment-specific false-positive SNP list (case study 1 – 3), a custom program was written and used to identify SNPs present in phenotypically unaffected pools of individuals. All individuals descended from seeds that were independently mutagenized with EMS, or phenotypically unaffected individuals were used to construct a collection of non-causative SNPs. These SNPs are considered false positives because SNPs present in multiple lineages or unaffected individuals cannot be causative for a phonotype only observed in one lineage. The VCFtools *exclude-positions* command was then used to exclude these positions from candidate mutations. The TAIR10.29-associated annotation files were retrieved from TAIR and used in predicting the effects of SNPs on gene function using the SnpEff computer program (Cingolani, Platts, et al., 2012).

An analysis workflow was developed to identify likely locations of causative polymorphisms. Plants carrying the induced mutation were outcrossed to derive F2 populations containing pools of phenotypically affected individuals. In case study 1 for example, the mutation was induced in Col background, hence, to narrow down its location a Ler x Col F2 population was created and required analysis. In the pools of phenotypically affected F2s the frequency of the Ler

SNPs will approach zero as their position gets more tightly linked to the causative mutation. DNA from pools of mutant F_2 individuals were sequenced and analyzed as follows. A modification to the variant calling step in the standard pipeline was made to output all positions by piping the output of SAMtools *mpileup* step to *bcftools view* with the *-Ncg* option to generate a new VCF. This new file was then filtered with a list of known SNPs from the outcross genotype to retain only positions polymorphic in the outcross genotype. Allele counts from the VCF's DP4 column were retrieved and processed to compute allele frequencies. Allele frequency per position was computed as (DP4[2]) + (DP4[3]) / (DP4[0]) + (DP4[1]) + (DP4[2]) + (DP4[3]). These values were then used in the construction of allele frequency plots for each chromosome from each mutant.

A comparison of three indel calling tools was carried out in case study 4 to assess their accuracy for detecting indels of varying size ranges. SAMtools, which was our preferred software package for SNP detection, was again used here by virtue of the fact that it can detect small indels (≤10 bp) (Kim et al., 2017). GATK HaplotypeCaller (GATK HC) (Poplin et al., 2018), was chosen for its ability to detect indels of up to 50 bp size (Wang et al., 2022), and LUMPY (Layer et al., 2014) was used for the detection of large (>100 bp) structural variants including insertions, deletions, inversions, duplications, and translocations (Layer et al., 2014; D. X. Liu et al., 2021). Indel calling with SAMtools used the standard pipeline described above followed by exclusion of SNPs from the VCF output.

The GATK HC indel calling pipeline first converted paired reads to unmapped BAM files and marked adapter sequences using picard-tools *FastqToSam* and *MarkIlluminaAdapters* packages, respectively. Mapping was carried out with BWA-MEM using MostDistant option as the primary alignment strategy to designate alignments giving the largest insert size with the mate as primary. Duplicate reads were identified with MarkDuplicates command from picard-tools and excluded from alignment files. The *HaplotypeCaller* from GATK version 3.5 was subsequently utilized for SNP and indel calling with options *"-stand_call_conf 30 -stand_emit_conf 10"* as initial filters for writing variants to VCF. Initial list of indels were used as basis for realignment using *IndelRealinger* and recalibration with *BaseRecalibrator* (both GATK packages) to improve original alignments before re-running the variant calling step.

LUMPY was utilized for variant calling as follows. Paired reads were aligned to the reference with BWA-MEM to produce alignment files. These were then processed with SAMtools *view* command in three separate steps to obtain insert size statistics, and to extract discordant

paired-end and split-read alignments for use in the variant calling step. These three files were used as input for the *lumpy* command which was executed with *"-tt 0"* option to skip read trimming and *"-mw 4"* option to set the minimum weight for a call at 4 across all samples.

To create an experiment-specific false-positive indel list, a custom program was written and used to identify indels present in phenotypically unaffected pools of individuals. All sequenced individuals descended from seeds that were independently mutagenized or phenotypically unaffected individuals were used to construct a collection of non-causative indels. These indels are considered false positives because indels present in multiple lineages or unaffected individuals cannot be causative for a phonotype only observed in one lineage. The VCFtools *exclude-positions* command was then used to exclude these positions from candidate mutations. The TAIR10.29-associated annotation files were retrieved from TAIR and used in predicting the effects of SNPs on gene function using the SnpEff computer program.

2.3 Results

Often, bioinformatics work occurs post-hoc as a consultant or in collaboration with field breeders, molecular biologists, or other owners of biological material. As a result, creativity and flexibility in analysis approach is required to both maintain quality control of the data in a post-hoc scenario or to provide design advice in an advance collaboration. Various implementations of my approach are presented here as a series of case studies with best practices demonstrated by the discoveries. The approaches for informatic processing of genomic data vary dependent on the pedigrees and experimental design. In all cases, the output includes causative polymorphisms and genes for phenotypically affective variation as well as quality control insights into experimental errors or better practices upstream of the analysis.

2.3.1 Case Study 1: Mutant mapping in a multiple mutant pharmacogenomics experiment.

In this case, our collaborators in the Chunhua Zhang Lab at Purdue University utilized a specially constructed line for mutant discovery. Their design involved generating a reporter line carrying the Green fluorescent protein of *Aqueoria* fused to the PIN-FORMED2 (PIN2) auxin transporter. They previously identified novel chemicals that altered the localization of this fusion protein. Their mutant screen was carried out to find *A. thaliana* mutants that modified the PIN2-

GFP localization during drug treatment. The intent was to identify the genes encoding these mutants and use the identity of these genes to develop hypotheses about the drug's function.

To accelerate this, the non-causative SNPs present in the progenitor line were identified by sequencing the PIN2-GFP line (Table 2-1). After removal of error-prone positions the only SNPs altering coding sequences via GA/CT changes in the PIN2-GFP line were in the genes AT1G07110/FKFBP, AT2G17700, and AT2G30500/NET4B. This is a demonstration of the power of the subtraction of error-prone positions as this line began with 4382 high-quality SNP calls but only 137 SNPs remained after subtraction. These SNPs were excluded from consideration as causative polymorphisms for EMS mutants derived from this line with novel phenotypes. These were included along with additional known and previously reported error-prone SNP positions in Col (Silva-Guzman et al., 2016) and independently EMS mutagenized Col (Dolan et al., 2017) in a subtraction file specific to this project.

The steps described above for read mapping, variant calling, indel exclusion and allele frequency mapping were conducted for the F2 samples. However, for D5R, ES3R long root, ES3R no root, *exo70a1-3, rml2-10*, and *rml2-3* very few polymorphisms were identified as possible causative polymorphisms, and a larger number of variants with a small number of reference reads were present. Discussion with the Zhang lab indicated that phenotyping the mutant was difficult, and certainty that the mutant pool only contained mutant individuals was low. As a result, I relaxed the criterion for homozygosity for those samples to allow one read on the forward and/or reverse reference reads due to poor phenotype certainty in collaborators lab. Allele frequency plots were generated per chromosome for each F2 as described above. For sequence data derived from M2 mutant samples, data were processed as described above but the SNP positions present in the PIN2-GFP progenitor line were excluded as possible causative polymorphisms.

Table 2-1. Description of materials used for sequencing in the multiple-mutant pharmacogenomics experiment

Sample ID	Experiment
14.40_MP	DNA form a pool of 500 seedlings from a mapping population of 14.40 mutant
5.23_M2	DNA from a single M2 seedling of 5.23 mutant
5.9_MP	DNA from a pool of 500 seedlings from a mapping population of 5.9 mutant
6.5_M2	DNA from a single M2 seedling of 6.5 mutant
6.5_MP	DNA from a pool of 100 seedlings from a mapping population of 6.5 mutant
PIN2_GFP	DNA from PIN2:GFP, serves as reference sequence as the mutation was generated from this background
D5R_mutant_MP	DNA from a pool of ~100 seedlings from D5R mutant. Mutant was crossed to Ler and resistant seedlings from F2 selected
ES3R_long_root_MP	DNA from a pool of ~100 seedlings from a mapping population of ES3R long root mutant
ES3R_no_root_MP	DNA from a pool of ~100 seedlings from a mapping population of ES3R no root mutant
exo70a1-3	DNA from pooled seedlings of an *exo70a1-3* homozygous mutant.
rml2-10	DNA from a pool of ~100 seedlings from *rml2-10* mutant allele in Col background
rml2-3	DNA from a pool of ~100 seedlings from *rml2-3* mutant allele in Col background

Table 2-2. Summary results from sequencing, variant calling and false-positive SNP removal in the multiple-mutant pharmacogenomics experiment.

Sample ID	PE reads	Coverage (X)	Mapped %	SNPs + Indels	SNPs	HQ SNPs	FP-subtracted SNPs	% (FP/HQ)
14.40_MP	41,830,141	31	99%	595,815	527,735	458,914	98	99.98
5.23_M2	44,843,011	33	90%	8,531	6,587	5,012	534	89.35
5.9_MP	38,234,989	28	99%	629,372	558,836	484,043	117	99.98
6.5_M2	47,302,778	35	97%	7,791	5,877	4,677	372	92.05
6.5_MP	41,884,572	31	90%	568,262	505,732	417,294	46	99.99
PIN2_GFP	45,857,647	34	99%	7,572	5,652	4,393	137	96.88
D5R_mutant_MP	70,176,552	52	99%	679,178	599,743	547,706	2812	99.49
ES3R_long_root_MP	104,484,197	77	99%	590,672	525,147	467,002	682	99.85
ES3R_no_root_MP	63,525,351	47	99%	514,278	459,441	373,654	781	99.79
exo70a1-3	34,993,517	26	99%	8,896	6,535	4,415	593	86.57
rml2-10	50,939,235	38	98%	8,856	6,603	4,412	686	84.45
rml2-3	72,197,625	53	97%	8,304	5,962	4,616	696	84.92

Table 2-3. Predicted effects of variants on gene expression in the multiple-mutant pharmacogenomics experiment.

Sample ID	FP-subtracted SNPs	G/A	C/T	%(G/A or C/T)	Silent	Stop gained	Missense	Splice
14.40_MP	98	22	34	57.1	2	2	8	2
5.23_M2	534	135	287	79	49	11	134	14
5.9_MP	117	18	37	47	1	1	7	0
6.5_M2	372	130	116	66.1	36	4	79	12
6.5_MP	46	14	14	60.9	1	2	0	0
PIN2_GFP	137	7	12	13.9	6	0	23	2
D5R_mutant_MP	2812	388	376	27.2	141	3	160	0
ES3R_long_root_MP	682	57	63	17.6	18	1	56	0
ES3R_no_root_MP	781	79	100	22.9	6	0	7	0
exo70a1-3	593	50	53	17.4	16	1	56	0
rml2-10	686	54	81	19.7	3	0	6	0
rml2-3	696	58	82	20.1	17	1	65	0

Read coverage and mapping rate averaged 41X and 97%, respectively, across the 12 samples analyzed in this experiment (Table 2-2). This was sufficient to call high-confidence SNPs. The number of SNPs discovered in samples derived from mapping populations was nearly 85-fold higher than those of samples from single plants. Filtering and false-positive SNP exclusion produced a reduced number of candidates causative polymorphisms among the high-quality SNP positions by 98% on average.

Seedlings derived from *rml2-10* and *rml2-3* mutant alleles in the Col background (Table 2-1) were utilized for DNA sequencing to identify the causative mutation responsible for the root meristemless phenotype. Reads alignment, variant calling, and variant filtering were all carried out as described in above. Both of these mutant samples' sequence data had a large majority of SNPs

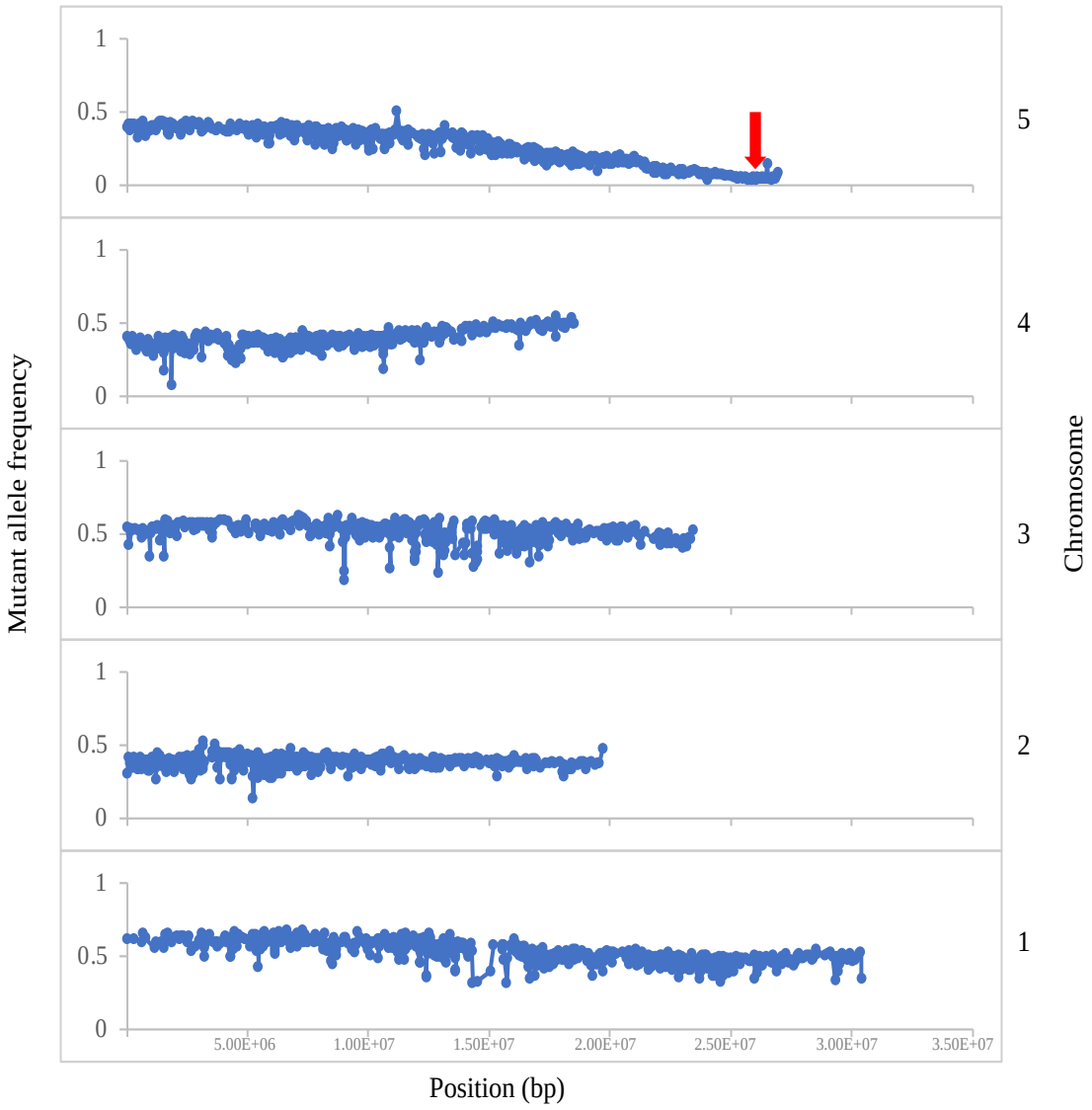


Figure 2-1. Mapping of the causative mutation in D5R to chromosome 5. Plot of allele frequency (y-axis) along chromosomal positions in bp (x-axis). Point where plot touches the x-axis indicates location of the causative polymorphism.

in common. Of the 686 SNPs identified in the *rml2-10* mutant 632 (92.1%) were also found in *rml2-3* (Table 2-3), indicating they are not independently derived EMS mutants, but rather were segregants from the same M1 individual or inadvertent stock duplicates in the lab. This demonstrates, again, that the analysis process is robust to errors in pedigree assignment. Had we subtracted SNPs from within this experiment and proceeded under the assumption that the pedigree

was correct, we would lose true SNPs as a result of the non-independence and subtraction procedure. As the causative SNP must be present in both *rml2-3* and *rm2-10*, we annotated all 632 of the shared SNP positions for effects on protein coding genes. Limiting our consideration to G to A or C to T SNPs, likely induced by EMS, identified only four mutations altering a protein coding gene in the two *rml2* mutants including AT1G17230, AT3G56550 (PCMP-H80), AT4G13490, and AT5G44940.

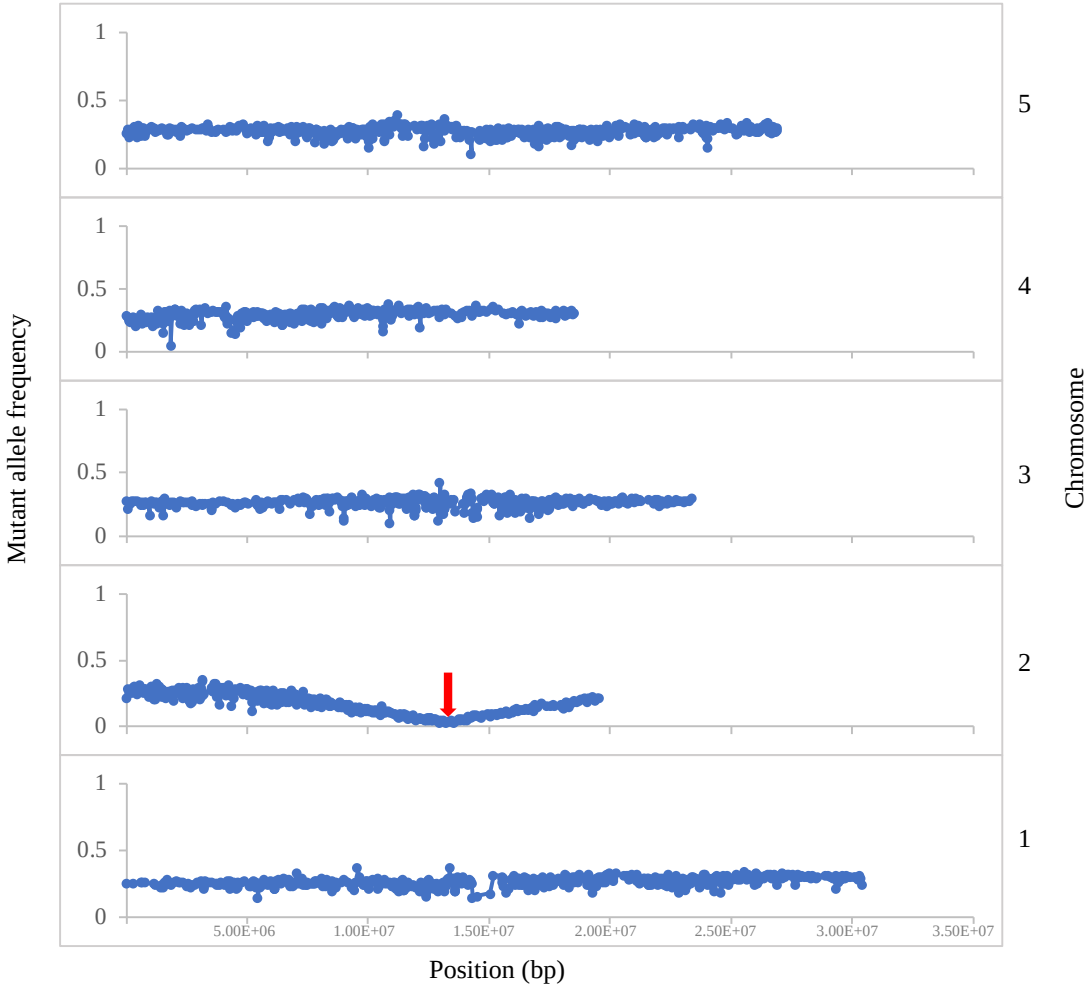


Figure 2-2. Mapping of the causative mutation in ES3R_no_root to chromosome 2. Plot of allele frequency (y-axis) along chromosomal positions in bp (x-axis). Point where plot touches the x-axis indicates location of the causative polymorphism.

To map causative polymorphisms our collaborators produced F2 mapping populations (denoted with the suffix "MP" in Table 2- 6) by outcrossing to Landsberg *erecta* followed selection of affected individuals displaying the respective phenotype. Sequencing was carried out on six DNA pools from about 100-500 seedlings derived from each mapping population (Table 6). The location of likely causative polymorphisms was narrowed using allele frequency plots and only those changes present at or near the maximum Col-0 allele frequency were considered.

Map positions for the other mutants were clearer and provided smaller windows with a limited number of candidate polymorphisms. The causative allele in D5R clearly maps to the tail-end of chromosome 5 (Figure 2-1). This region contained protein-coding changes consistent with EMS mutagenesis in AT5G62890 (NAT6), AT5G63850 (AAP4) or AT5G64740 (CESA6). Thus, all three were equally possibly responsible for the mutant phenotype. The phenotype in D5R is characterized by longer roots in the presence of the drug which causes plants to display a reduced cellulose content, strongly implicating CESA6 as the causative gene. Having identified this as the causative polymorphism, a large number of other mutants with the same drug sensitivity were tested for polymorphisms in CESA6 and fifteen different alleles of CESA6 were identified as responsible for this drug sensitivity (Huang et al., 2020). Of note in this project, requiring only non-reference reads for homozygosity missed the causative SNP. The iterative discussion with the collaborating lab and modification of the SNP calling protocol to match the lab conditions and pedigree certainty was required for the success of this approach

The mutation in ES3R no root maps clearly to chromosome 2 position 1.4×10^7 (Figure 2-2). Within the mapped window mutations affecting protein-coding sequences were identified in AT2G30320, AT2G31190 (RUS2), AT2G31900 (XI-F), AT2G32460 (MYB101), and AT2G33680 (PCMP-E19). Drug ES3 disrupts endoplasmic reticulum (ER)-protein stability (Zhang et al., 2016), however target of the drug is unknown. The phenotype is no roots upon treatment with the ES3 drug. It has also been observed that the line segregates for no roots when not grown in the presence of the drug, indicating that the phenotype is not drug dependent. Although line appears to segregate for long roots, it did not map to any locus and may be a dominant phenotype at any position, linked or unlinked. Presumably, homozygous mutants can grow to maturity on soil albeit small and unhappy.

The causative polymorphism in line 5.9 maps to top of chromosome 1 (Figure 2-3). This region contained protein-coding mutations in AT1G01510 (*Angustifolia*), AT1G02340 (HFR1),

AT1G02590, AT1G03590 (PPC6-6), or AT1G04040. Consistent with the known phenotype of loss-of-function alleles at *Angustifolia*, which affects trichome morphology (Bai et al., 2013), and knock-out allele encoded in line 5.9 this mutant displays the *an1* loss-of-function phenotype. It is conceivable that ANGUSTIFOLIA is required for PIN2-GFP localization, and this mutant should permit testing of that by further linkage analysis in subsequent experiments. The causative polymorphism in line 6.5 maps to the bottom end of chromosome 5 (Figure 2-4) and likely impacts AT5G56369, AT5G57210, AT5G59710 (VIP2), AT5G63570 (GSA1), AT5G64760 (RPN5B), AT5G64950. The causative polymorphism in sample 14.40 has a clear map position at approximately 9.0×10^6 bp on chromosome 4 (Figure 2-5); list of candidates includes AT4G13650 (PCMP-H42), AT4G16650, AT4G16950 (RPP5), AT4G17300 (SYNO) AT4G18670 (LRX5). No map position was identifiable for the 5.23 through sequencing. The list of putative causative genes is as follows: AT1G18335, AT1G22410, AT1G70630, AT3G10490 (ANAC052), AT3G27040, AT3G57420, AT4G02250, AT4G29180, and AT5G32590.

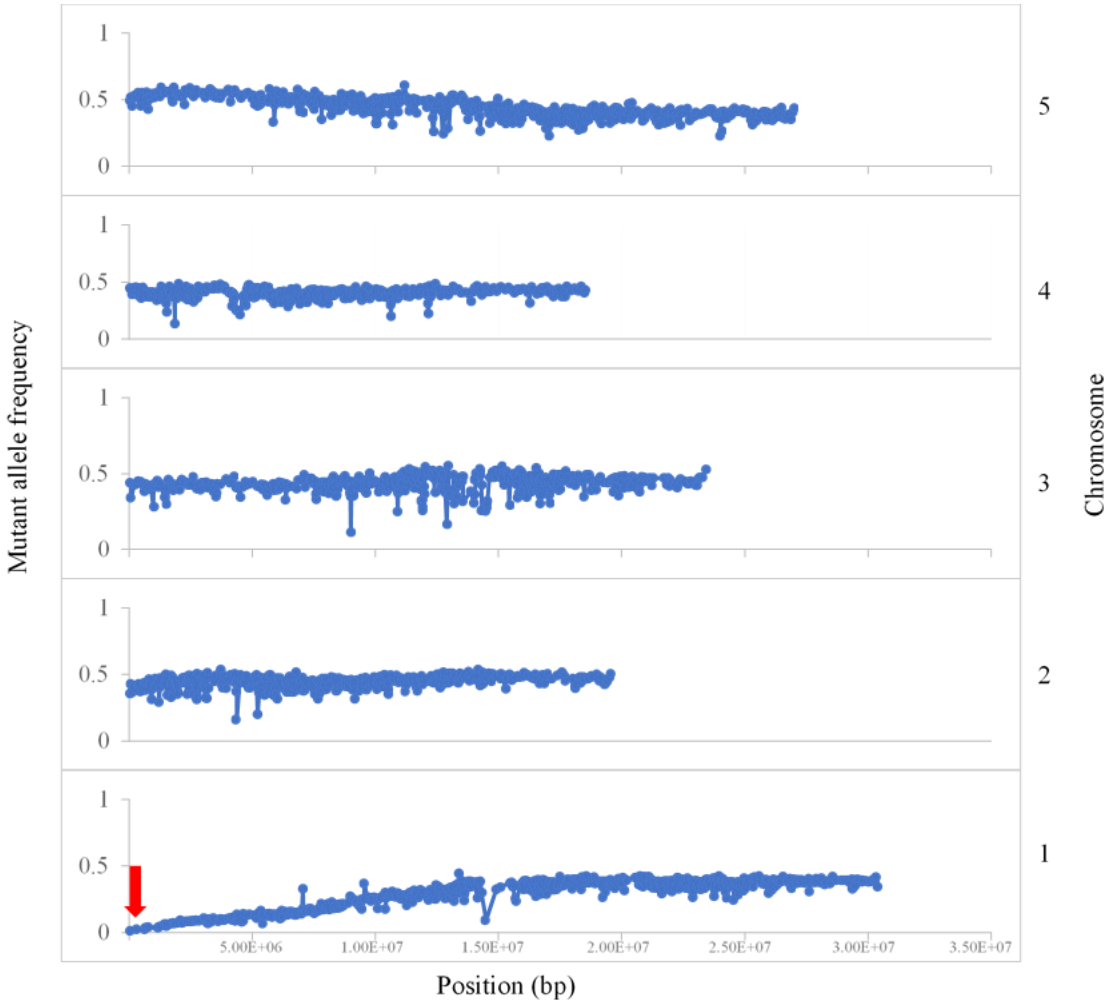


Figure 2-3. Mapping of the causative mutation in 5.9_MP to the chromosome 1. Plot of allele frequency (y-axis) along chromosomal positions in bp (x-axis). Point where plot touches the x-axis indicates location of the causative polymorphism.

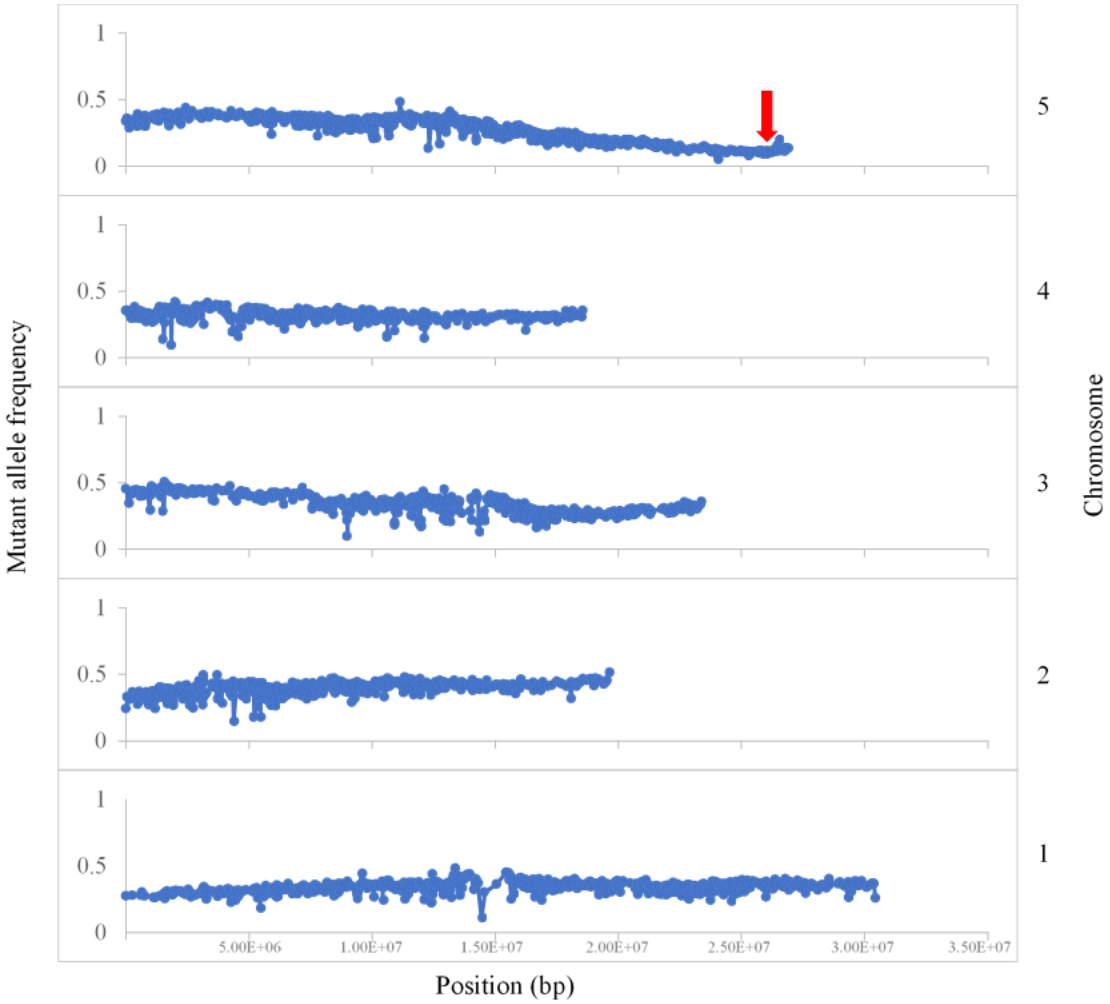


Figure 2-4. Mapping of the causative mutation in 6.5_MP to chromosome 5. Plot of allele frequency (y-axis) along chromosomal positions in bp (x-axis). Point where plot touches the x-axis indicates location of the causative polymorphism.

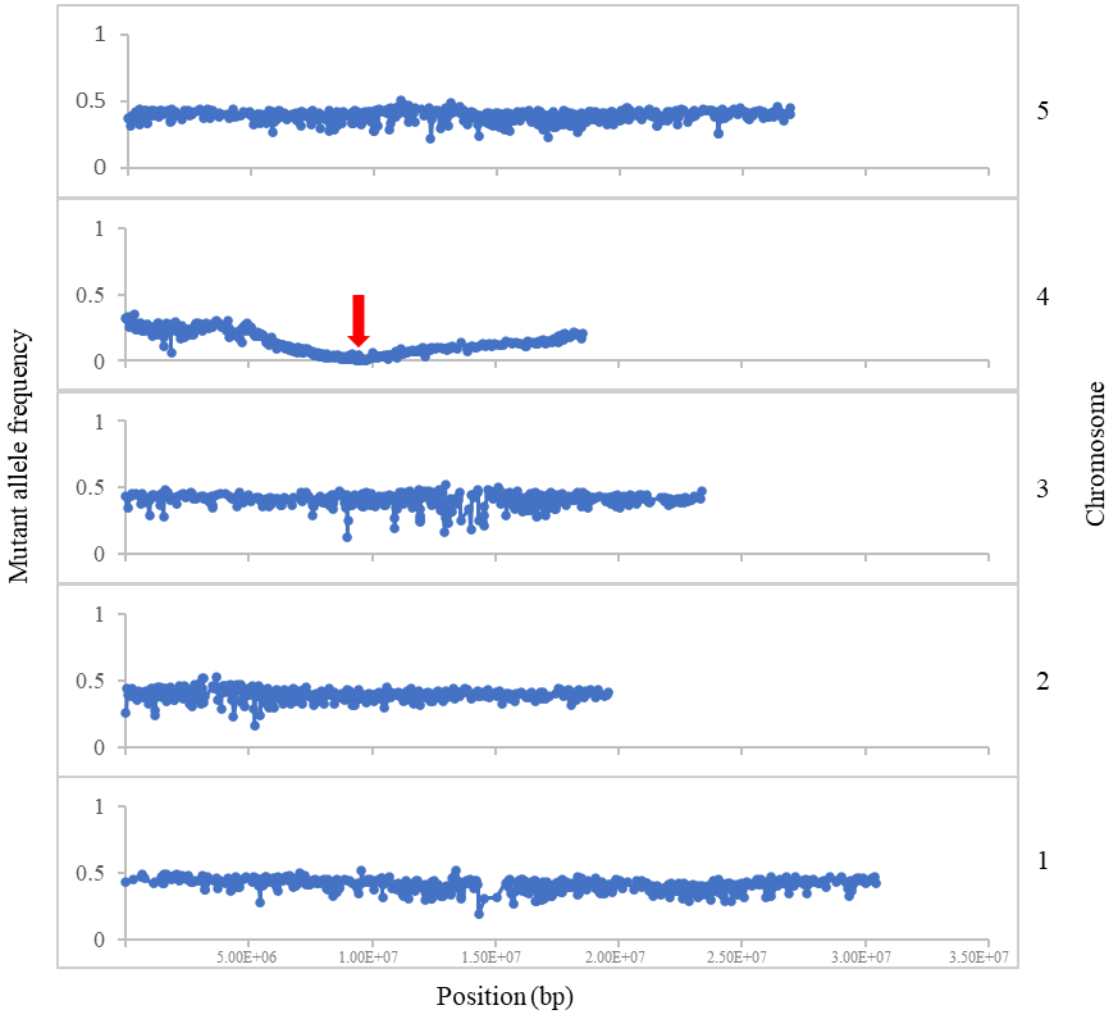


Figure 2-5. Mapping of the causative mutation in 14.40_MP to chromosome 4. Plot of allele frequency (y-axis) along chromosomal positions in bp (x-axis). Point where plot touches the x-axis indicates location of the causative polymorphism.

2.3.2 Case Study 2: Informatic experimental design robust to alternative explanations: CRISPR off – target and pedigree errors in heterozygous mapping experiments

The lab of Dr. Sharon Kessler was seeking to utilize sequencing to permit the study of mutants altering the performance of gametophytes. A set of mutants, independently mutagenized with EMS in Wassilewskija background, were screened in an attempt to map by segregation distortion in a Col x Ws hybrid background. F1 hybrids were used to generate two F1BC1 pools consisting of 150 individuals each, to be analyzed to check for segregation distortion (Table 2-4). A second mutant with an unexpected phenotype was recovered in a CRISPR screen from the same

project. The gene targeted by the guide-RNA construct, MLO8, was not mutated and so the Kessler lab was interested in identifying the causative polymorphism for the loss of stigmatic papillae observed in their mutant. Unlike with EMS mutagenesis, this was likely an indel polymorphism and required a different analysis.

Paired-end reads from all samples were aligned to TAIR10 Arabidopsis reference sequences, followed by variant calling and filtering as described in the above Methods section. High-quality homozygous SNPs common to SK1, SK2, SK4, SK5, SK6, SK23, SK25, SK26, SK27, SK28 were removed from each sample. The unique set of SNPs for each sample were then annotated using SnpEff functional prediction tool. A set of reliable Ws SNPs was generated by compiling a list of SNPs common among SK1, SK2, SK4, SK5, SK6, SK27, and SK28. Read coverage at these reliable Ws positions were extracted for samples SK21, SK22, SK25, SK26, and used to create allele frequency plots. Unfortunately, causative loci could not be clearly seen on any of the plots. High-quality heterozygous SNPs common to SK1, SK2, SK4, SK5, SK6, SK27, and SK28 were compiled and excluded from all samples. Unique heterozygous SNPs for each sample were then annotated with SnpEff.

Table 2-4. Description of plant materials used for sequencing.

Sample ID	NL number	Mutant	Background/ experiment	Expected mutation type	Genotype
SK1	NL84-23	ntaE 14E-6	*nta-1/nta-1* (Ws) enhancer mutant	EMS	Heterozygous for mutation
SK2	NL87-D	ntaE 13C-5	*nta-1/nta-1* (Ws) enhancer mutant	EMS	Heterozygous for mutation
SK4	NL88-9	ntaE 9D-4	*nta-1/nta-1* (Ws) enhancer mutant	EMS	Heterozygous for mutation
SK5	NL89-11	ntaE 6H-7	*nta-1/nta-1* (Ws) enhancer mutant	EMS	Heterozygous for mutation
SK6	NL90-17	ntaE 11H-1	*nta-1/nta-1* (Ws) enhancer mutant	EMS	Heterozygous for mutation
SK21	NL100-5 X Col-0	ntaE NL-4-27	*nta-1/nta-1* (Ws) enhancer mutant outcrossed with *nta-3* (Col-0)	EMS	150 $F1BC_1$ seedlings pooled
SK22	NL102-4 X Col-0	ntaE NL-1R	*nta-1/nta-1* (Ws) enhancer mutant outcrossed with *nta-3* (Col-0)	EMS	150 $F1BC_1$ seedlings pooled
SK23	TCD 258-1	dsp mutant	Col-0, mutant arose in a CRISPR construct against MLO8	deletion?	Homozygous for mutation
SK25	NL 102-4	ntaE NL-1R	F1 mother plant used for segregation distortion cross (seq sample SK22) (Ws/Col hybrid)	EMS	Heterozygous for mutation
SK26	NL100-5	ntaE NL-4-27	F1 Mother plant used for segregation distortion cross (seq sample SK21) (Ws/Col hybrid)	EMS	Heterozygous for mutation
SK27	ELK56	ntaE 18G-6	*nta-1/nta-1* (Ws) enhancer mutant, pool of 50 BC3 plants	EMS	Heterozygous for mutation
SK28	ELK40	ntaE 3E-1	*nta-1/nta-1* (Ws) enhancer mutant, pool of 50 BC3 plants	EMS	Heterozygous for mutation

Table 2-5. Summary results from sequencing, variant calling and false-positive SNP removal.

Sample ID	PE reads	Coverage (X)	% Mapped	Total Indels	Total SNPs	HQ SNPs	Homoz. HQ SNPs	FP-subtracted Homoz. SNPs	HQ Het. SNPs with DP>=25	FP-subtracted Het. SNPs
SK1	82,705,575	61	98.46	98,859	667,059	631,671	448,416	5,153	46,859	3,927
SK2	53,717,057	40	98.31	99,159	694,251	653,732	463,810	3,069	70,318	3,776
SK4	46,251,002	34	98.37	98,778	697,671	655,717	467,186	3,138	74,488	6,335
SK5	53,321,285	39	98.47	99,279	692,357	653,174	469,723	3,031	69,479	3,807
SK6	52,672,323	39	98.35	99,095	694,902	654,304	463,891	3,173	70,784	4,000
SK21	55,033,628	41	99.19	42,229	341,310	263,641	585		259,196	244,397
SK22	68,811,775	51	99.57	2,086	4,575	3,652	611		2,717	1,885
SK23	54,539,323	40	99.23	2,003	4,965	3,909	682	78	2,870	1,983
SK25	64,865,107	48	97.04	2,042	4,297	3,373	608	18	2,414	1,651
SK26	59,476,834	44	98.61	75,154	552,374	516,004	581	18	507,778	463,245
SK27	55,155,445	41	98.37	99,499	693,657	654,296	464,883	3,292	69,068	3,949
SK28	66,684,774	49	98.41	99,720	680,056	643,120	460,517	3,982	56,360	3,439

An inconsistent number of heterozygous SNPs was observed across the expected Col x Ws hybrids in samples SK21, SK22, SK25, and SK26 (Table 2-5). These indicated that multiple populations were constructed from non-segregating material resulting in a low number of heterozygous SNPs in lines SK22 and SK25. This is indicative of pedigree errors in the mapping and phenotyping of the mutant resources and resulted in the collaborating lab refocusing efforts. Again, this demonstrates the value of informatic processing of sequencing data in a pedigree aware manner and interaction between informatics and bench scientist efforts.

For sample SK23, I took an approach more similar to the identification of recessive mutants described in the previous cases studies. This line carried a recessive mutation that blocked stigmatic papillae production which arose spontaneously in a plant carrying a guide-RNA and Cas9 protein expression cassette targeting the *MLO8* gene (Table 2-4). Alignment of reads and SNPs calling was done as described previously. As this mutant was generated in the Col-0 background, known error-prone SNP positions in Col were excluded from the high-quality homozygous SNPs followed by annotation with SnpEff. Because this mutant was generated in a Cas9 expressing plant, I suspected that genome editing by this nuclease resulted in an indel polymorphism, perhaps at an off-target site, responsible for the phenotype.

Indels were called from the alignments using SAMtools and filtered to exclude indels with a pred-scaled quality score below 20. Indel positions were identified in all previously sequence Col samples in the Dilkes lab by following the same procedure and these positions were excluded as error-prone positions from the high-quality indels. The remaining indels, found only in the SK23 mutant, were then annotated using SnpEff. Two high-impact deletions were observed; the first was an AG -> A on chromosome 4 start position 12059494 affecting the AT4G23000 gene and the second was a GTCCTGCGTCAAAAGGTT -> GT on chromosome 5 start position 18404413 resulting in splice acceptor variant, splice region variant, 3 prime UTR variant and an intron variant in the AT5G45420 gene. Additionally, a moderate impact deletion of TTCCTTCCCGCATTTTATCCAG -> T was detected on chromosome 5 start position 16821238, causing an in-frame change in DL1 (AT5G42080). This last gene is likely to encode the causative polymorphism in the SK23 mutant as stigma papillae also failed to elongate in a previously described loss of function allele (Collings et al., 2008; Kang et al., 2003) . This once again demonstrates the value of informatics approaches to distinguish called polymorphisms that can

and cannot possibly encode causative polymorphisms and then integrating biological information from the literature.

2.3.3 Case Study 3. Experimental design robust to alternative explanations: T-DNA integration induced INDELs

The strategy of false-positive exclusion demonstrably worked well with SNPs and was extended to indels in the previous case study. I extended the SNPs approach and created an analytic pipeline to produce a list of indels found in individuals derived from seeds that were independently mutagenized, or phenotypically unaffected pools of individuals. Just as with the SNPs analyses, these indels are deemed false positives, or not phenotypically relevant, because indels harbored in multiple lineages or unaffected individuals cannot be the determinants of a phenotype only seen in a single lineage. The effectiveness and limitations of this approach was further explored in this and the next case study.

In this experiment, our collaborators in the Kelly lab at Iowa State University sought to identify the causative mutation behind a long hypocotyl phenotype (termed *slim shady*) observed in a T-DNA insertional line SALK_015201. This line had been previously published as a result of it encoding a disruption allele at *IMMUNOREGULATORY RNA-BINDING PROTEIN* (Dressano et al., 2020). However, transgenic plants overexpressing *IRR* retained the long hypocotyl phenotype and additional knock out alleles of *IRR* showed wild-type hypocotyl length, demonstrating that the long hypocotyl phenotype was likely due to an unrelated mutation. This prompted experiments to discover the true genetic basis for the long hypocotyl phenotype. This line was used as a control in an unrelated study of transposon dynamics (Hu et al., 2019). Genomic sequence data for *Arabidopsis thaliana* accessions SALK_015201 and the phenotypically unaffected *Ds* line CS85255 were previously characterized in an unrelated study (Hu et al., 2019). The paired-end 150-bp reads from these lines were retrieved from the Sequence Read Archive and mapped to the TAIR10 reference genome and taken through the variant calling steps described above.

I produced a list of indels present in both SALK_015201 and the phenotypically unaffected CS85255. Variants found in both lines cannot be responsible for the long hypocotyl phenotype observed only in SALK_015201. Exclusion of these false-positive indels from the candidate causative variants and subsequent annotation were carried out as described above. This re-

processing of publicly available whole-genome data for SALK_015201 and subsequent use of our filtering strategy identified a 1-bp deletion of "T" at the *phyB* locus inside the last exon of the gene (Figure 2-6). This deletion at the 3,370 position in the *PHYB* coding sequence results in a nonsense codon that converts a leucine codon to a premature stop codon and truncated the last 48 amino acids. Given the phenotype of the *slim shady* mutant strongly resembled prior *PHYB* loss of function alleles, this candidate gene was reported to the collaborating lab. Additional, targeted, follow-up experiments demonstrated that my informatics approach was successful and the deletion allele in *PHYB* was the causative polymorphism for the long hypocotyl phenotype in this line. Additional details on this collaborative work can be found in (Dash et al., 2021).

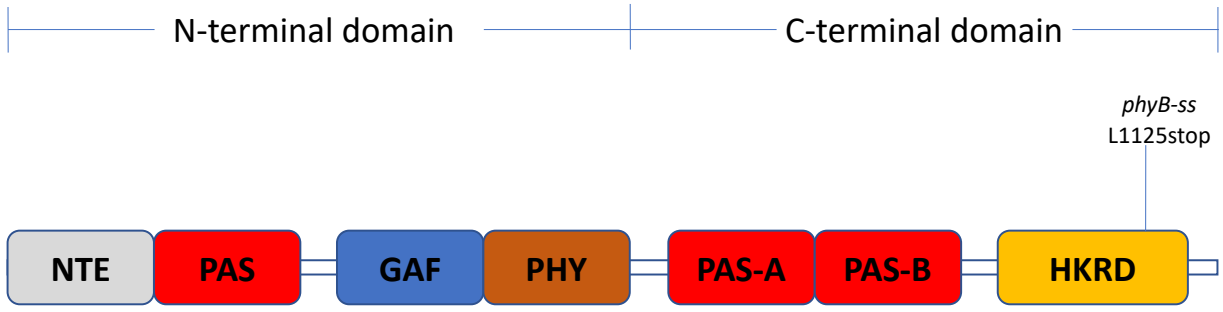


Figure 2-6. Genomic sequencing of Arabidopsis accession SALK_015201 uncovered the *phyB-ss* allele, a 1 bp deletion at position 3,370 of PHYTOCHROME B which leads to a premature stop codon immediately after L1125 amino acid located inside the histidine kinase-related (HKR) domain (Cartoon adapted from (Dash et al., 2021)).

2.3.4 Case Study 4. Robust removal of false positive insertion deletion calls improves mutant allele discovery and comparison of mutagenesis effectiveness across mutagens.

In this case study, I assessed the performance of the false-positive exclusion strategy in detecting indels of varying size ranges. Mutants derived following treatment with three different mutagenic agents with varying mutational patterns were used. Two physical mutagens, fast neutron (FN) and gamma irradiation, and the chemical agent EMS were used. FN bombardment has been reported to generate deletions ranging from a single base to 18 Mb in rice, with a majority falling within the 1 – 4 kb size range (X. Li et al., 2001). In Arabidopsis, it has been demonstrated that these mutagens produce 1 bp substitutions at higher rate than the large deletions (Belfield et al., 2012). These single base substitutions were also observed to be the predominant mutations

when Li et al (2016) sequenced 41 rice lines exposed to FN irradiation. Gamma irradiation is known to be the predominantly used ionizing radiation for mutant production (Hase et al., 2020; F. Li et al., 2019). Even though gamma rays induce mostly small indels and point mutations(Hase et al., 2020; S. Li et al., 2018; Tan et al., 2019), large indels, including whole chromosomes have also been reported (Henry et al., 2015). EMS, a commonly used chemical alkylation agent, is the major means of mutagenesis in Arabidopsis (Page & Grossniklaus, 2002). The overwhelming proportion of the mutations induced by EMS are G/C to A/T transitions randomly distributed across the genome (Greene et al., 2003).

I profiled multiple indel calling software packages to enhance our ability to detect the wide size ranges of indels possibly induced by these three mutagens. Three commonly used indel callers with different size detection ranges were employed: GATK HaplotypeCaller (GATK HC); SAMtools; LUMPY. GATK HC calls variants through local re-assembly and is known to be effective at calling small indels smaller than 50 bp (Wang et al., 2022). SAMtools was developed for manipulating alignment files in the SAM/BAM format and includes both SNP and indel calling capacities. It relies on a Bayesian model for local realignment to call indels mostly ≤ 10 bp (Kim et al., 2017). Of the three callers used here LUMPY calls the largest variations. LUMPY uses a probabilistic framework for detecting large (>100 bp) structural variants (Layer et al., 2014; D. X. Liu et al., 2021).

The largest number of raw variants were discovered by GATK (23,990 on average), whilst SAMtools and LUMPY produced mean counts of 8353 and 384 respectively (Table 2-6 to 2-8). It is worth noting that SNPs accounted for 42% and 27% of raw variant counts for GATK and SAMtools respectively whereas LUMPY did not call SNPs. The proportion of FPs found and removed from LUMPY variants was smallest across the three methods evaluated (average FP in LUMPY = 41.7% compared with 90% for SAMtools and 87.9% for GATK).

Since EMS is known to mostly create single-base changes in contrast to FN and Gamma mutagenesis techniques that additionally produce indels we conducted pairwise t-tests of the number of indels between EMS and the latter two methods to ascertain how well the three indel callers performed. For the indels called by GATK, there was no statistical difference between EMS and the other two mutagenesis techniques post FP exclusion (*p-value* EMS vs. FN = 0.07; EMS vs. Gamma = 0.06) indicating that at these doses the two physical mutagens induce relatively few indel polymorphisms more that alkylation by EMS. Similarly, indels discovered by LUMPY did

not show a statistically significant difference between EMS and the other treatments (*p-value* EMS vs. FN = 0.14; EMS vs. Gamma = 0.1). We only observed a statistically significant difference between EMS and FN for indels produced by SAMtools after FP removal (*p-value* = 0.03) but not between EMS and Gamma (*p-value* = 0.9).

Table 2-6. Summary results from variant calling and false-positive INDEL exclusion using GATK and custom script.

Sample	Treatment	Raw variants	Raw indels	Refined indels	HQ refined indels	Indel count without FPs	% FPs
193	FN	22,859	8,259	6521	4420	495	88.8
194		21,531	8,555	6512	4427	417	90.6
195		22,834	8,643	6859	4599	532	88.4
196		21,454	8,159	6448	4335	475	89.0
197		22,274	9,284	7218	4944	622	87.4
198	EMS	24,913	11,333	8645	6083	840	86.2
199		24,311	9,980	7525	5198	643	87.6
200		25,448	10,488	8154	5714	692	87.9
201		25,802	11,101	8228	5866	721	87.7
202		25,929	11,323	8450	6004	755	87.4
203		24,473	10,680	8002	5643	656	88.4
204		25,416	11,248	8395	5917	755	87.2
205		24,758	11,114	8236	5777	744	87.1
206		25,332	11,729	8670	6144	776	87.4
207	GAMMA	24,398	10,195	8009	5601	738	86.8
208		22,578	9,692	7519	5165	594	88.5
209		25,387	11,929	9101	6529	898	86.2
210		22,119	9,231	6902	4676	486	89.6

Table 2-7. Summary results from variant calling and false-positive INDEL removal using SAMtools and custom script.

Sample	Treatment	Total variants	Indel count	Quality filtered indels	Indel count without FPs	% FPs
193	FN	10196	2431	1835	236	87.1
194		8736	2116	1654	154	90.7
195		9995	2376	1805	216	88.0
196		9877	2391	1854	243	86.9
197		8603	2418	1829	219	88.0
198		8341	2469	1949	258	86.8
199		9277	2524	1897	233	87.7
200		8144	2436	1862	197	89.4
201	EMS	7700	2102	1687	125	92.6
202		7812	2115	1701	134	92.1
203		7987	2175	1708	137	92.0
204		8043	2236	1752	134	92.4
205		7422	2345	1838	183	90.0
206		6921	2178	1730	133	92.3
207	GAMMA	7802	2480	1884	203	89.2
208		8203	2297	1778	168	90.6
209		6898	2060	1674	112	93.3
210		8414	2132	1667	140	91.6

Table 2-8. Summary results from variant calling and false-positive INDEL exclusion using LUMPY and custom script.

Sample	Treatment	Total variants	Indel count without FPs	% FPs
193	FN	497	362	27.2
194		268	149	44.4
195		332	196	41.0
196		181	83	54.1
197		231	108	53.2
198		431	298	30.9
199		372	233	37.4
200	EMS	809	605	25.2
201		615	413	32.8
202		473	293	38.1
203		378	229	39.4
204		347	187	46.1
205		335	185	44.8
206		302	148	51.0
207	GAMMA	585	401	31.5
208		215	100	53.5
209		346	194	43.9
210		204	90	55.9

I estimated the relative expected phenotypic impacts of the different mutagens. To do this, I used SnpEff to examine the effects of the indel variants on genes, transcripts, and protein sequences. Coding sequences are more likely to be unique and are less expected to result in errant calls from reads mismapping and other artifacts. If the lack of indel differences across treatments was the result of false positive indel calling, limiting our view to coding sequences should improve the signal to noise. Because FN and Gamma both produce deletions whereas EMS almost exclusively generates base substitutions, I limited the initial EMS analysis to count of deletions in the EMS-treated samples. As with the total indels observed, no statistical difference was observed between the number of indels detected by GATK with high or moderate effects on coding sequences generated by EMS as compared to FN or Gamma (*p-value* EMS vs. FN = 0.18; EMS vs. Gamma = 0.72) (Table 2-9). Similar results were observed for deletions identified by LUMPY

(*p-value* EMS vs. FN = 0.35; EMS vs. Gamma = 0.6) (Table 2-11). Similar to the observation for total indel counts, the count of deletions with high or moderate effect detected by SAMtools differed between EMS and FN (*p-value* 0.03) but not between EMS and Gamma (*p-value* = 0.6) (Table 2-10).

Across the tools tested the number of high-impact effects of deletions rose with increasing size. Deletions detected by SAMtools had an average length of 3.7 bp and produced on average 4.3 high-impact effects (Table 2-10). Similarly, GATK revealed deletions, with an average length of 7.8 bp, generating an average of 8.9 high-impact effects (Table 2-9). It was no surprise that LUMPY, which identified the largest deletions (2.7 kb on average), detected the largest number of high-impact effects averaging 448.9 (Table 2-11). These high-impact effects outnumbered moderate-impact effects because they were scored large-scale deletions of gene models. We further assessed the performance of the tools used here by testing whether the expected size difference of the deletions produced by the different mutagens could be detected. Only GATK produced deletions with detectable size difference between treatments, however, that was only between EMS and FN (*p-value* = 0.027 for test of mean size) but not between EMS and Gamma (*p-value* = 0.065 for test of mean size). Both SAMtools- and LUMPY-produced deletions did not show differences between mean sizes across the three mutagenesis agents.

Table 2-9. Annotation of GATK Deletions using SnpEff.

Sample	Treatment	Indel count without FPs	Number of Deletions with moderate to high effect	Deletion annotations		Deletion size		
				Number of high impact	Number of moderate impact	Average (bp)	Median (bp)	Range (bp)
193	FN	495	8	6	2	19	2	1 - 97
194		417	15	11	7	19	5	1 - 176
195		532	12	11	3	23	3	1 - 179
196		475	10	11	1	2	1	1 - 3
197	EMS	622	12	9	8	9	4	1 - 35
198		840	16	9	11	6	3	1 - 31
199		643	12	7	5	3	3	1 - 6
200		692	11	6	5	4	3	1 - 16
201		721	12	15	2	2	2	1 - 3
202		755	11	12	2	2	1	1 - 10
203		656	7	6	2	2	2	1 - 3
204		755	5	5	1	1	1	1 - 3
205	GAMMA	744	13	16	2	17	2	1 - 178
206		776	13	9	5	12	1	1 - 111
207		738	5	2	3	9	4	1 - 29
208		594	11	9	4	6	2	1 - 35
209		898	12	11	5	2	1	1 - 3
210		486	8	5	3	2	2	1 - 3

Table 2-10. Annotation of SAMtools Deletions using SnpEff.

Sample	Treatment	Indel count without FPs	Number of Deletions with moderate to high effect	Deletion annotations		Deletion size		
				Number of high impact	Number of moderate impact	Average (bp)	Median (bp)	Range (bp)
193	FN	236	27	4	28	3	3	2 - 17
194		154	30	5	31	4	3	2 - 21
195		216	24	6	30	4	3	1 - 27
196		243	30	7	33	3	3	1 - 15
197		219	20	7	20	5	3	1 - 32
198		258	32	5	37	5	3	2 - 24
199	EMS	233	31	2	38	3	3	1 - 4
200		197	12	3	13	4	3	2 - 16
201		125	16	6	14	3	3	1 - 3
202		134	16	6	17	3	3	2 - 3
203		137	20	2	21	3	3	2 - 3
204		134	9	3	8	2	3	1 - 3
205		183	27	7	26	4	3	1 - 24
206	GAMMA	133	12	2	16	5	3	1 - 27
207		203	19	2	17	3	3	1 - 3
208		168	24	3	28	5	3	2 - 35
209		112	16	3	16	3	3	1 - 3
210		140	20	4	16	4	3	1 - 31

Table 2-11. Annotation of LUMPY Deletions using SnpEff.

Sample	Treatment	Indel count without FPs	Number of deletions with moderate -high impact	Annotation of deletions		Size of deletions		
				Number of high impact changes	Number of moderate impact changes	Ave. (kb)	Median (kb)	Range (kb)
193	FN	362	5	19	0	2.7	1.9	0.1 - 7.0
194		149	4	16	0	6.6	3.1	1.4 - 18.9
195		196	5	605	2	281.7	1.5	0.0 - 1404.3
196		83	6	47	3	7.7	1.1	0.7 - 37.2
197		108	7	600	2	206.4	3.0	0.0 - 1404.4
198		298	7	13	3	3.2	1.1	0.0 - 15.4
199		233	6	20	1	7.4	2.3	0.8 - 22.4
200		605	7	823	0	360.0	3.3	0.7 - 2489.8
201	EMS	413	4	7	0	8.9	8.1	0.7 - 18.8
202		293	8	835	0	320.3	6.2	0.7 - 2489.8
203		229	10	610	4	147.1	3.3	0.7 - 1404.3
204		187	5	599	0	336.3	22.3	3.0 - 1404.4
205		185	11	2242	2	538.9	15.4	0.7 - 2489.8
206		148	6	609	1	238.0	1.7	0.7 - 1404.3
207	GAMMA	401	8	837	0	257.2	9.5	0.7 - 1957.1
208		100	6	14	1	9.2	1.9	0.0 - 35.6
209		194	8	171	1	76.8	17.0	0.7 - 494.9
210		90	5	13	1	8.2	1.4	0.4 - 22.4

2.4 Conclusions

Indel calling from short read data still an issue for most available tools. This is evidenced by the lack of consistency across tools and the detection of substantial indel count in EMS compared to FN and Gamma treated samples regardless of tool used. Size differences and ranges of deletions among the three treatments could also not be consistently detected by the three tools. False positive indel exclusion for SAMtools and GATK indels remove 80-90% of all initially called variants. On the one hand this indicates a very high rate of noisy positions relative to any

signal resulting from true indels. On the other, it indicates that the employed subtraction approach can identify and remove errantly called indels. This case study only relied on data from these 18 samples, and we've previously shown from SNP analysis that the false positive exclusion greatly benefits from inclusion of extensive data from unrelated individuals. Expansion of the analysis to run these tools on a larger number of mutagenized individuals may improve the detection of false positive positions and improve the signal to noise. Greater confidence in the tools, interpretation of their indels, would be warranted if they detected differences between treatments. By this criterion, of the three methods employed GATK showed the best performance as it detected expected deletion size differences between the treatments.

The false positive correction strategy developed here was extremely efficient in reducing noise and amplifying true biological signal from experimental data. The minimal number of candidate genes produced as end-product greatly cuts the cost and time needed for follow-up molecular validation experiments, speeding up the pace of forward genetics projects. The workflow was additionally able to detect pedigree errors as well as errors associated with non-independent mutagenesis and provided explanations to previously failed experiments. Comparison of different mutagenesis agents revealed EMS to be the best. It generated a comparable number of indels, albeit small in size, with similar mutational impact to the two physical mutagens. Together with the large number of transitions generated by DNA alkylation, this mutagen is the most likely to result in a gene disruption or modification and provides the best chance at causing mutations in the greatest number of genes for further studies. The comparatively small numbers of indels produced by FN and gamma suggest these mutagens should only be preferred if alkylating agents are not tolerated by a study organism. This work also brought forth more evidence that indel calling with short read data is still a challenge despite advancements over the years. If indels are of interest in a given experiment, then perhaps new methods or alternative reads technologies are needed.

CHAPTER 3. ELUCIDATING TRANSCRIPTIONAL CONTROL OF *RP1-D21*-INDUCED HR

3.1 Introduction

Plants use an extreme immunity response, the hypersensitive response (HR), to protect themselves from pathogens. HR is defined as a quick death of cells that is limited to the area immediately surrounding the point of pathogen invasion to slow down disease progression (Balint-Kurti, 2019; Goodman & Novacky, 1994; Mur et al., 2008). This defends the individual at the cost of a vital metabolic response and localized cell death. The importance of this phenomena to plant success and crop protection is enormous. A variety of signaling components and responses (el Kasmi, 2021; Lolle et al., 2020; van Wersch et al., 2020)are known, including a growing complement of host-encoded plant pathogen sensors that can initiate HR (Adachi et al., 2020; Leonetti et al., 2021) . Yet, the molecular control of this phenomenon is not fully elucidated. A more complete understanding the genetic control of plant HR, including the mechanisms that attenuate it's spread and carry out the metabolic and transcriptional remodeling of cellular activity, will both deepen our understanding of organismal interactions and be leveraged to further crop improvement via breeding for resistance to diseases.

One set of host-encoded plant pathogen sensors that has been well characterized is the Nucleotide binding-site leucine-rich-repeat coiled-coil domain proteins encoded at the *Resistance to puccina1* (*Rp1*) gene cluster in maize. This cluster encodes a wide diversity of gene complements (Collins et al., 1999) and is responsible for resistance of many maize varieties to *Puccinia sorghii*, or common rust. A mutant version of this NLR cluster, *Rp1-D21* (Sun et al., 2001), derived from unequal crossing over between two paralogs in the gene cluster. *Rp1-D21* encodes a hyperactive NLR that induces HR even in the absence of pathogens (Richter et al., 1996). This hyperactive NLR has been the subject of extensive research into the mechanism, regulation, and consequences of HR, as resistance is triggered in the absence of any additional effects of the pathogen. Multiple studies have explored the metabolic (Ge et al., 2021; G. F. Wang & Balint-Kurti, 2015), transcriptional (Ge et al., 2021; Karre et al., 2021; Murphree et al., 2020), and growth effects (Negeri et al., 2013) of constitutive HR in maize.

Because *Rp1-D21* is semi-dominant, multiple studies have explored the effects of natural variation in modulating or determining the effects of HR in maize. The phenotype of a mutant gene can be used as a reporter to unearth previously undetectable genetic modifiers of that mutant in crosses to diverse genetic backgrounds (Chintamanani et al., 2010). By using this approach with *Rp1-D21*, alleles that control the HR can be discovered without exposure to pathogens, thereby eliminating the confounding effects of pathogen variation. An additional advantage of this approach for the discovery of genes important for the HR is that it relies on a mutant gene that autoactively produces HR throughout the plant. This allows for easy, reliable, detection of a phenotype that is otherwise difficult to measure due to its rapid nature and that is only induced locally in a subset of pathogen-host combinations.

Both RIL and GWAS approaches have identified numerous chromosomal loci encoding natural variants that can modify the *Rp1-D21*-induced HR phenotypes. Candidate genes, resulting from the combination of RNAseq and GWAS data (Olukolu et al., 2014) have been described. Genes that colocalized with the associated SNPs were presented as the most likely causative genes. For a small number of loci, the genes through which these loci manifest their influence have been validated by molecular assays in later studies and they surprisingly encode enzymes. (He et al., 2016; Luan et al., 2021; G. F. Wang et al., 2015; G. F. Wang & Balint-Kurti, 2015). Thus far, all of the successfully validated candidates encode genes that were differentially expressed during *Rp1-D21*-induced HR. Thus, differential expression affected by the hyperactive NLR allele *Rp1-D21* was an effective complementary analysis to GWAS for the severity of *Rp1-D21* lesions to identify genes capable of modulating HR in maize.

Discovering the relationship between the allele at a variant locus and transcription levels of genes, such as through eQTL analysis, provides a hypothesis for a direct link between the observation of a genetic association and an understanding of the molecular mechanisms responsible for trait variation. Past work in the regulation of HR in maize, demonstrated successful validation of candidate genes linked to genetic associations that were also increased in their expression during HR (Kim et al., 2021; Wang et al., 2015). eQTL analysis treats transcription levels as quantitative traits in a segregating population and maps genetic variants that influence transcription *in vivo*, thus allowing the direct biological interpretation of the consequence of variation. In some cases, changes in gene expression are the result of a variant that exerts its influence in *cis*. These only influence the expression of the gene located on the same physical

chromosome with it. One such example is the *cis*-eQTL identified on maize chromosome 4 found to be associated with control of the rubisco activase gene (*ZmRCAβ*) (Y. Zhang et al., 2019). In the case of HR-regulated *cis*-variants, this identifies genes with *cis*-regulatory variation that modulates that gene's response to the transcriptional regulators that control HR. Variants that control signaling, or transcription can also affect an eQTL at a gene but do so in *trans*, meaning that the QTL maps to a position distinct from the gene whose expression is being measured. Indeed, *trans*-eQTL can regulate gene anywhere in the genome. For example, a polymorphism affecting the expression of a transcription factor can affect the expression of one or many genes. Liu et al., (2017) identified a *trans*-eQTL hotspot on maize chromosome 1, immediately upstream of A-type R2R3 Myb-like transcription factor, that regulated the expression of 11 flavonoid metabolism genes. A gene which increases or decreases HR-signal intensity might alter the expression of all genes that respond during HR. Such a scenario would result in what is referred to as a *trans*-regulatory hotspot, because many genes are affected in trans by a polymorphism at one position. I expect that during HR, polymorphisms at transcription factors downstream in the regulatory hierarchy might affect hotspots at a subset of HR-responsive genes while variation in HR signaling acting upstream of transcriptional changes, might affect all HR-responsive genes in an eQTL experiment.

The methodology in eQTL experiments is subject to a special limitation that can confuse the meanings of cis and trans as they are used in molecular biology. Because eQTL mapping relies on linkage mapping, any trans regulator (e.g., transcription factor) that happens to bind and regulate the expression of a tightly linked gene would be identified as a "*cis*-eQTL" according to the criteria in mapping. Such special cases of local trans regulation are best considered as false positive detection of *cis* variation and false negative detection of trans regulation. Given the 10 chromosomes of maize and lack of linkage between chromosome arms, so long as transcription factors and their targets are randomly distributed across the chromosomes 95% of the true *trans*-regulatory changes are unlinked their targets and cannot result in local trans eQTL that would be mistaken placed in a *cis*-eQTL set. The size of this problem is further diminished as the recombination resolution of the mapping population increases. For example, given an average centimorgan length of 75 for each of the 20 chromosome arms, and a conservative resolution of 30 cM for a QTL confidence interval, this would cut the expected number of local trans cases by another factor of 2. So, while local trans cases do happen, and identification of *cis*-eQTL cannot

be taken unequivocally as evidence of *cis*-regulatory polymorphism in the molecular sense, local trans false positives are expected to be rare.

One way to unequivocally test for *cis*-regulatory alleles is by comparing two alleles in the same individual for allele-specific expression (ASE) differences. ASE analyzes gene expression between parental alleles within an F1 hybrid. An imbalance in the relative abundance of transcripts encoded by the two alleles provides a demonstration of *cis*-regulatory differences between the two alleles. I propose that a combination of ASE and traditional eQTL analysis can be used to cross validate *cis*-regulatory variants. ASE will be explored directly in Chapter 5.

I set out to combine, for the first time, the study of the hypersensitive response with eQTL, ASE, and DGE analyses to systematically exploit genetic variation to identify the genetic determinants of this plant disease response. To do this I used the *Rp1-D21*-induced transcriptional phenotype in combination with previously generated crosses of *Rp1-D21*/+ to B73 and NC350 inbred lines which have divergent intensities of HR responses (Chintamanani et al., 2010b; Olukolu et al., 2014b, 2016) and for which a RIL population is available (McMullen et al., 2009; J. Yu et al., 2008). In this chapter, differential gene expression (DGE) analysis was used to identify those genes affected by *Rp1-D21* across and those only altered in one of the two parents and to explore the intensity of expression differences. Identification of genes and pathways through which these genetic variants influence the HR will provide a launching point for the design of future experiments that will uncover the precise molecular mechanism underpinning HR and further allow exploitation for crop improvement.

3.2 Methods

Reads were first assessed for quality with Fast QC (Andrews 2010, available online at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). This helped check for low quality bases as well as sequence length distribution, the level of GC, base N and duplicate sequences present in reads. Since reads already been taken through a preprocessing step by our collaborators, low-quality bases and sequencing adapters had already been removed. This preprocessing step is important since adapter and low-quality sequences can interfere with downstream steps such as mapping to the reference. Reads were subsequently aligned to the B73RefGen_v4 (Jiao et al., 2017) with STAR RNA-seq aligner (Dobin et al., 2013). Duplicate reads within alignment files were

identified and marked with Picard Toolkit *MarkDuplicates* command (Langmead & Salzberg, 2012). HTSeq (Anders et al., 2015) was used to compute raw read count per gene from alignment files and B73RefGen_v4-associated genome annotation.

Raw read counts were imported into DESeq2 for differential gene expression analysis. However, prior to this, a series of exploratory analyses were conducted to ascertain whether global gene expression followed expected patterns based on what was known about the samples. To do this, raw read counts per gene were first transformed with the variance stabilization transformation to produce log_2-transformed data that's also been normalized by library size (Anders & Huber, 2010) These transformed data were then used for hierarchical clustering based on Euclidean distances and visualized with the "plot" function in R. Principal component analysis were also carried out using the transformed data with the "plotPCA" function of ggplot2 (Wickham, 2016).

Differential gene testing using the raw gene counts was carried out with the "DESeq" wrapper script within the "DESeq2" R package, and genes were designated differentially expressed if they had a false discovery rate (FDR) lower than 0.05. Visualization of Differential gene expression between groups was visualized with heatmaps using "pheatmap"(Kolde, 2015) and with volcano plots using the "plot" functions in R. Venn diagram and upset plots used to depict the interaction among differentially expressed genes (DEGs) from the different genetic backgrounds were generated with web-based "InteractiVenn" (http://www.interactivenn.net/index.html) and UpSetR (Conway et al., 2017) respectively.
Gene functional annotations were assigned using Phytozome v13 (https://phytozome-next.jgi.doe.gov/phytomine) before gene ontology (GO) enrichment analysis was carried out with PANTHER (Mi et al., 2019; Thomas et al., 2003), to understand functional relationships among genes.

3.3 Results

3.3.1 Production of biological material and experimental design

In previous work (Olukolu et al., 2013), *Rp1-D21* was introgressed into the H95 maize inbred line to create H95;*Rp1-D21/+*. This was accomplished by repeated backcrossing to H95 to the backcross four (BC4) generation, each time only choosing progenies displaying disease-like

lesion phenotype. Due to the sterility of homozygous mutants, the H95;*Rp1-D21/+* genotype must be kept as a heterozygote. Seeds from this stock were crossed as a pollen-parent to the maize inbred lines B73 and NC350 as well as each member of the 200-line B73 x NC350 Recombinant Inbred Line (RIL) population. Each cross-produced an F1 family segregating 1:1 for wild-type and mutant phenotypes (Figure 3-1). The B73 x NC350 RIL population is a subset of the Nested Association Mapping population of maize (McMullen et al., 2009; Olukolu et al., 2014b, 2016; J. Yu et al., 2008) and has available pedigree and genotypic information. The decision to use this population was informed by the previous work demonstrating that NC350 strongly enhanced the HR phenotype in *Rp1-D21/+* hybrids (Chintamanani et al., 2010). In addition, the B73 x NC350 population was one of the families among the 24 NAM subpopulations previously crossed to *Rp1-D21/+* that showed the highest allelic effect size values for lesion-related traits (Olukolu et al., 2014).

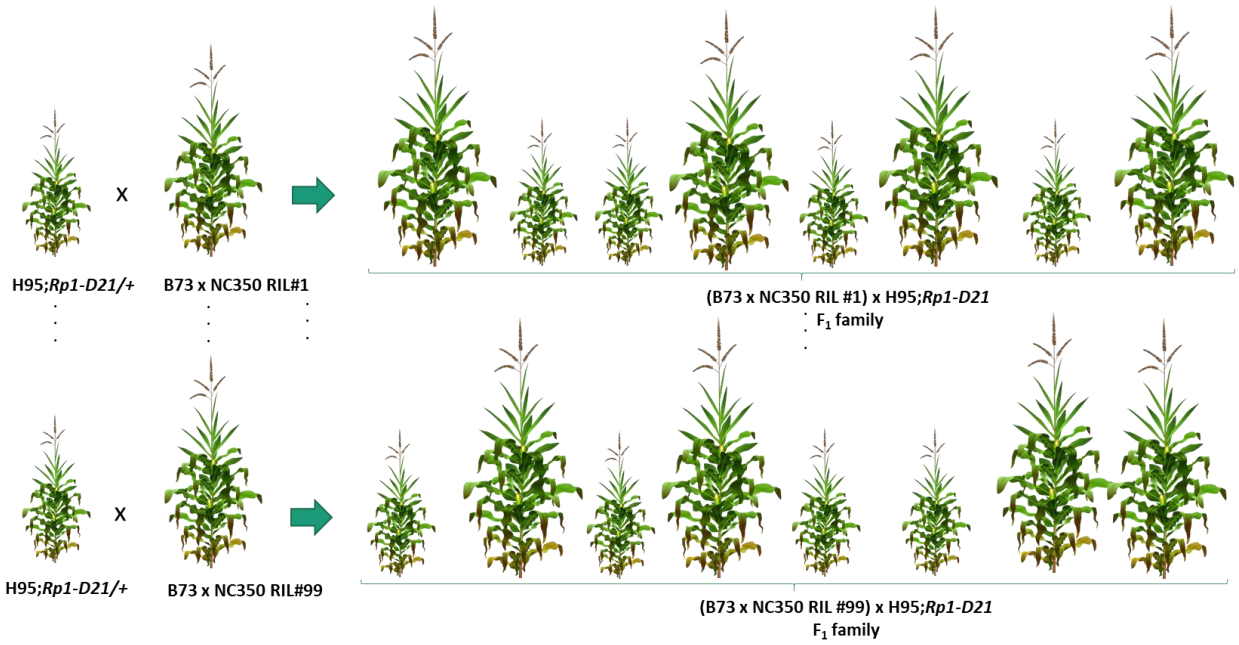


Figure 3-1. Mapping of HR modulators using 99 members of the 200-line NC350 NAM RIL population. The H95;*Rp1-D21*/+ line carrying the *Rp1-D21* mutation in a heterozygous state is crossed to each member of the NC350 NAM RILs. The F1 families generated segregate 1:1 for mutant and wildtype sibs.

3.3.2 Raw read mapping, count-per-gene estimation and differential gene expression analysis

FastQC was used to perform quality control of input reads which were 125 bp in length. This involved assessing GC content, *k*-mer representation as well as sequencing adapter contamination. Reads were mapped to the B73RefGen_v4 (Jiao et al., 2017) with STAR splice-aware aligner. The output was processed with Picard to mark duplicate reads using the *MarkDuplicates* command. Raw counts per gene were computed with HTSeq by projecting de-deduplicated alignment files to B73RefGen_v4 gene annotation using the *htseq-count* function.

An additional QC step was conducted prior to DGE analysis to check whether global expression patterns meet expectations. For example, biological replicates are expected to show similar expression patterns whilst different conditions should be dissimilar. Samples that fail to meet these checks would be excluded from downstream analyses. To improve the performance of these QC analyses and to make results from these easily interpretable, raw reads were read-depth-normalized and transformed to the log_2 scale. The similarity of read counts between replicates was examined by plotting counts in a pairwise manner. Hierarchical clustering was used to ascertain whether samples from different conditions could be separated in an unsupervised manner. This was achieved through pairwise comparison of samples based on Pearson correlation coefficient (r). Principal component analysis (PCA) was carried out to complement hierarchical clustering to determine whether samples from different conditions showed more variability than replicates from the same condition.

Raw read counts per gene from the replicates of each treatment and genotype were analyzed for DEG using DESeq2. This allowed pair-wise identification of differentially expressed genes between mutant/wildtype phenotypes. This assessed whether observed read count differences between samples are more than what would be expected by random chance. DESeq2 estimates this difference by fitting the gene counts to a negative binomial distribution followed by a test of the null hypothesis that the conditions cannot explain differences in gene expression. This tests all genes independently, so I also calculated adjusted p-values to account for multiple hypothesis testing using the Benjamini-Hochberg procedure to reduce false positives (Benjamini and Hochberg, 1995). A gene was considered significantly differentially expressed if its FDR-corrected p-value ≤ 0.05. For each gene a fold change, depicted as a logarithm to base 2 (log_2FC) of the mutant/wildtype expression ratio, was additionally computed. This was estimated using

average read counts across replicates per phenotype, after read counts for all genes for a given sample had been summed and used as basis for sample-specific normalization. Volcano plots were then used to visualize *p*-values against fold change to allow for quick discovery of statistically significant genes that have large fold changes between conditions. As a sanity check, a heatmap was used to visualize the expression values of the top 15 most up-regulated and the top 15 most down-regulated genes across samples. This was done to determine whether the direction of fold change observed matched what was reported in the DEG results table. These 30 most up or down-regulated genes were analyzed using Phytozome v13 to obtain their functional annotations. Lastly, GO enrichment analysis was carried out with PANTHER for all DEGs (upregulated and downregulated analyzed separately) to identify biological processes they are involved in.

3.3.3 Biological insight in HR

Differential gene expression analysis in the parents of the mapping population B73;Rp1-D21/+ versus wildtype

To assess the impact of *Rp1-D21* in the B73 inbred background, a heterozygote for *Rp1-D21* was crossed to B73 to produce F1 plants segregating 1:1 for wild-type and mutant phenotypes. RNA from three biological replicates from each phenotype was converted to cDNA, libraries for sequencing prepared, and sequenced for differential gene expression analysis by our collaborator at the USDA-ARS, Dr. Peter Balint-Kurti, on the campus of North Carolina State University. Reads depths varied from 16.6 – 33.7M per sample. These reads were aligned to the maize reference genome version 4 (Jiao et al., 2017), and 71.1 – 79.67% uniquely mapped (Table 3-1). This indicated high-quality sample preparation and data processing steps.

Table 3-1. Mapping statistics for B73;*Rp1-D21*/+ versus wildtype.

Sample	Background	Phenotype	Raw reads	Average read length (bp)	Uniquely mapped reads %	Multi-mapped reads %	Unmapped reads %
BRwt_rep1	*B73:Rp1-D21*	wildtype	17,366,148	125	71.08%	25.87%	3.04%
BRwt_rep2	*B73:Rp1-D21*	wildtype	17,938,638	125	75.29%	21.38%	3.33%
BRwt_rep3	*B73:Rp1-D21*	wildtype	16,600,257	125	79.67%	18.00%	2.34%
BRmu_rep1	*B73:Rp1-D21*	*Rp1-D21*	30,515,726	125	76.66%	21.17%	2.17%
BRmu_rep2	*B73:Rp1-D21*	*Rp1-D21*	33,755,029	125	75.76%	22.29%	1.95%
BRmu_rep3	*B73:Rp1-D21*	*Rp1-D21*	24,529,243	125	76.71%	21.30%	1.99%

Alignment files were processed together with a reference annotation file to generate read counts per gene for each sample. These counts were higher in the wildtype samples compared to the mutants (Figure 3-2). I do not have a ready explanation for this pattern, but it is consistent with either the normal, healthy wildtype plants exhibiting higher expression of genes or that the HR-affected plant samples were smaller due to the growth-impairment of constitutive disease response and somehow this led to poorer library construction from the mutant plants. Although raw read counts are initially input to DESeq2 for DGE analysis, the assessment of differential expression includes normalization that should remove the effects of these reads distribution effects across the two sample types. The quality control analyses that preceded DGE also relies on transformed counts. The effect of size-factor normalization followed by log_2-transformation on reducing skewness in these samples was clear. This greatly improved the accuracy of algorithms used for clustering and principal component analysis used as a quality control check.

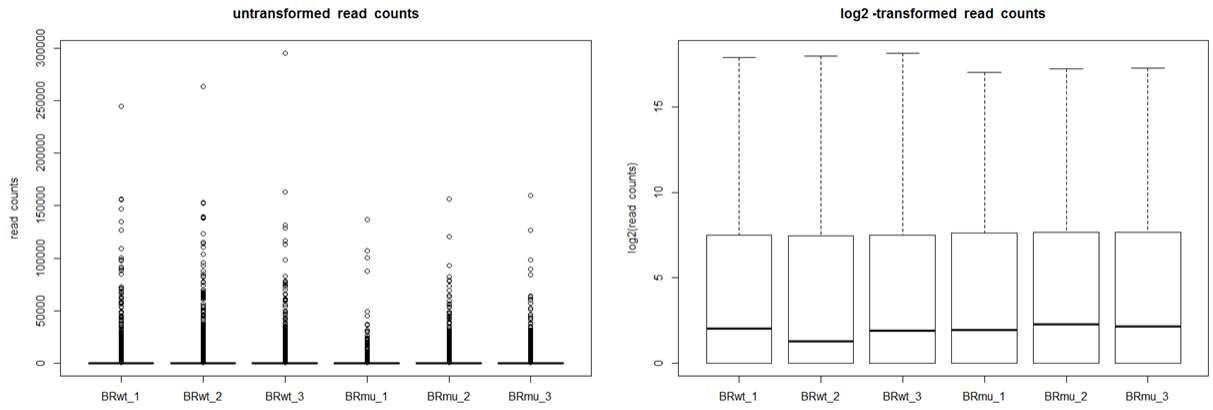


Figure 3-2. Comparison between untransformed and *log*$_2$-transformed read count distribution of B73;*Rp1-D21*/+ versus wildtype (BR) samples showing the effect of transformation in reducing skewness.

It's been long established that replicates are required to provide the statistical power needed to reliably discover differentially expressed genes (Schurch et al., 2016). However, this is only possible if replicates show high correlation of gene expression. Biological replicates with poorly correlated gene expression usually signal mistakes or other problems at some stage during the experiment. In this experiment, replicates showed more than 97% similarity in pairwise comparisons as measured by Pearson correlation, *r* (Table 3-2). This was above the 90% correlation threshold recommended by the ENCODE consortium (Standards, Guidelines and Best Practices for RNA-Seq V1.0) and is an indication of the quality of experimental procedures from tissue collection through the library preparation and sequencing.

Table 3-2. Similarity of read counts between B73;*Rp1-D21*/+ versus wildtype (BR) replicates as measured by Pearson correlation coefficient.

Pairwise comparison	Pearson Correlation coefficient
BRwt_1 vs. BRwt_2	0.979
BRwt_2 vs. BRwt_3	0.979
BRmu_1 vs. BRmu_2	0.977
BRmu_2 vs. BRmu_3	0.984

Hierarchical clustering was performed to group similar samples using the pairwise correlations calculated above. The distance measure (Height) was computed as 1- *r* and the results rendered as a dendrogram. As expected from the Pearson correlations, hierarchical clustering revealed greater gene expression variability between wildtype and mutant phenotypes than among replicates of the same phenotype (Figure 3-3).

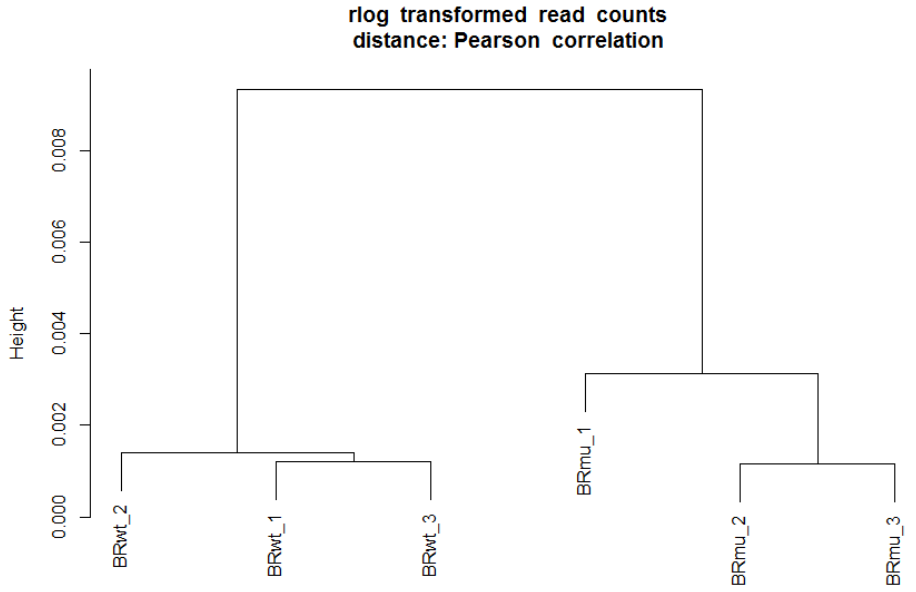


Figure 3-3. Dendrogram showing results of hierarchical clustering of B73;*Rp1-D21*/+ versus wildtype (BR) samples. Replicates displayed greater similarity whereas the two phenotypes were clearly separated.

Principal components confirmed the results of hierarchical clustering and also unearthed greater gene expression variability between wild-type and mutant phenotypes than what was observed among replicates of the same phenotype. In fact, 92% of variation observed across all samples could be explained by differences between mutant and wild type. Differences among replicates only accounted for 4% of total variation (Figure 3-4).

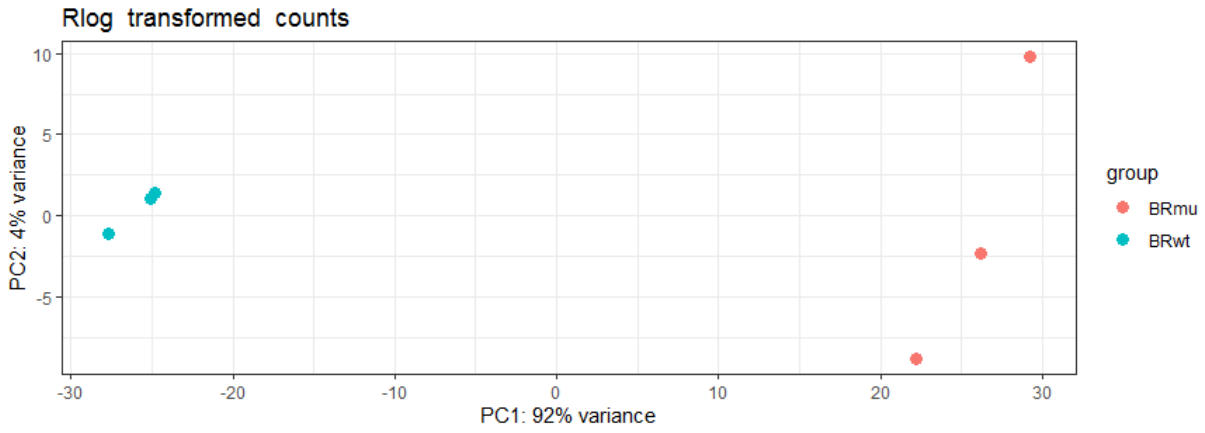


Figure 3-4. PCA on *rlog*-transformed read counts for B73;*Rp1-D21*/+ versus wildtype (BR) samples. Differences between phenotypes account for greater proportion of variance.

Neither hierarchical clustering nor PCA identified outliers, so an analysis of differential gene expression was carried out. The three biological replicates per condition (mutant or wildtype) were analyzed using DESeq2, with wild type set as control, to identify differentially expressed genes (DEG). A total of 6,685 DEG were identified as significantly differently expressed between mutant and wild type. Almost twice as many genes increased expression as part of the hypersensitive response triggered by *Rp1-D21* than were decreased (Figure 3-5); 4,211 (15%) expressed genes were significantly increased in accumulation as opposed to 2,474 (8.9%) that were decreased.

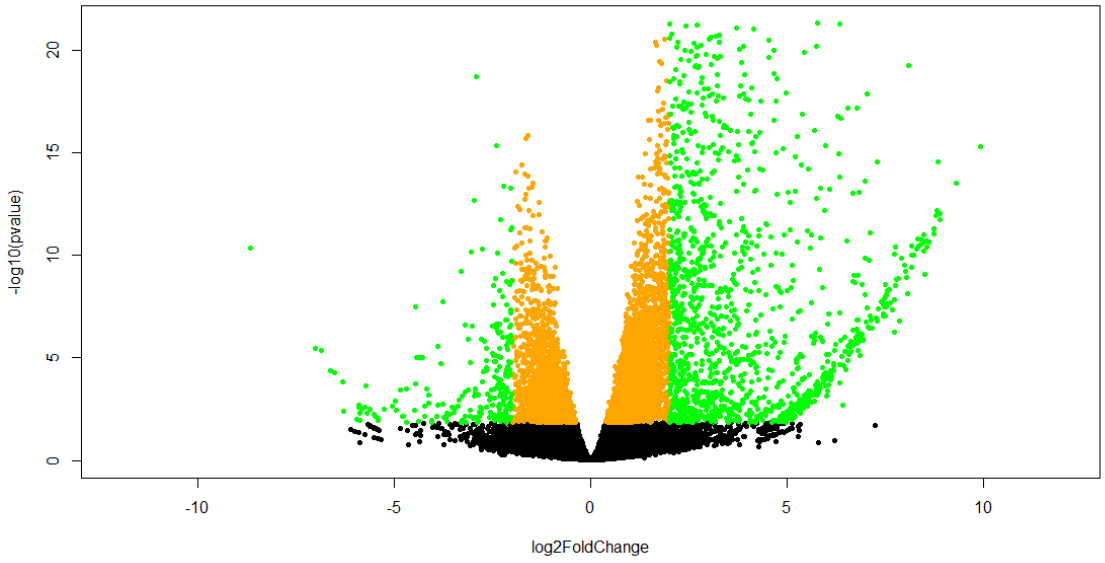


Figure 3-5. Volcano plot of B73;*Rp1-D21*/+ versus wildtype (BR) DGE results depicting statistical significance (p-value) versus magnitude of change (fold change). Black dots are genes that are not statistically significant (adjusted p-value > 0.05), orange dots are statistically significant genes (adjusted p-value < 0.05), green dots are statistically significant genes with absolute log_2 fold change of 2.

The heatmap of log_2-transformed read counts of the top 30 up or down-regulated genes (Figure 3-6) was similar to the direction of fold-change displayed in the DGE results table. Several defense or stress related genes were among the top up-regulated genes (Table 3-3). The most up-regulated gene, Zm00001d025200 (Indolin-2-one monooxygenase), was over a thousand-fold higher in expression in the mutant compared to the wildtype. This enzyme is a critical component of the 2,4-dihydroxy-1,4-benzoxazin-3-one (DIBOA) biosynthesis pathway by converting indolin-2-one to 3-hydroxyindolin-2-one. DIBOA is part of a chemical defense mechanism against insects and pathogenic microbes in the grasses (Frey et al., 1997; Niemeyer, 1988).

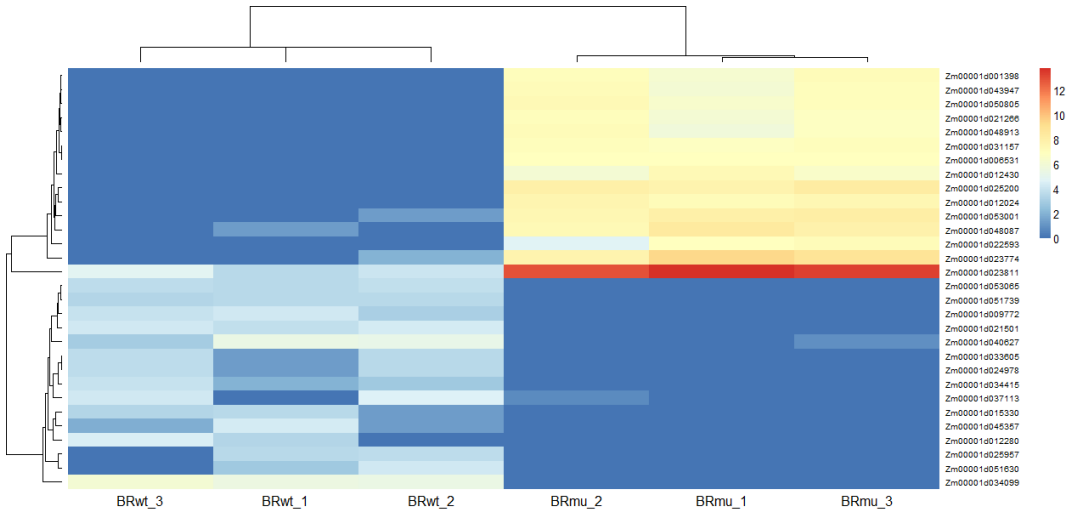


Figure 3-6. Heatmap of *log*$_2$-transformed read counts for top 30 most up or down-regulated genes after DGE analysis of B73;*Rp1-D21* (BR) samples. Genes are sorted based on hierarchical clustering.

Another defense related gene is Zm00001d023811, a gene from the Pathogenesis-related protein Bet v I family, known to be a part of host response triggered upon infection or under stressful conditions (Radauer et al., 2008) Zm00001d048087 (Flavonoid O-methyltransferase 4, FOMT4) was recently characterized in maize to play a role in defense. Similar to another flavonoid OMT2 located on maize chromosome 9, this enzyme is vital to the biosynthesis of xilonenin, which has antifungal activity against maize pathogens, *Fusarium graminearum* and *Fusarium verticillioides* (Förster et al., 2022). Earlier, a closely related enzyme in rice, Naringenin 7-O-methyltransferase (OsNOMT), was reported to be involved in catalyzing the biosynthesis of flavanone phytoalexins, such as sakuranetin, which is accumulated in rice which is produced in response to pathogen invasion and other stresses (Rakwal et al., 2000). Other stress/defense genes identified among the top up-regulated genes include the salt stress response/antifungal gene (Zm00001d006531), Zm00001d021266 (BON1-ASSOCIATED PROTEIN 1-RELATED), Zm00001d022593 (Ornithine decarboxylase / L-ornithine carboxy-lyase). Conversely, several growth-related genes were found in the top 15 most down-regulated genes. One of these was Zm00001d034099 (Ferredoxin NADP reductase or FNR), an enzyme that catalyzes an important step in photosynthesis which is the conversion of light energy to chemical energy (Morales et al.,

2000). Another example in this category is Zm00001d025957 (B-box zinc finger), which encodes a transcription factor known to regulate growth, development, and ripening in grapevine (Wei et al., 2020). These examples suggest that whilst defense-related activities are turned up in the cell as a result of autoimmunity triggered by *Rp1-D21*, normal growth and developmental processes are turned down.

This observation is further supported by the results of the GO enrichment analysis of all down- and up-regulated gene sets (Table 3-4). Down-regulated genes were enriched for growth and developmental processes such as photosynthesis, mitotic cytokinesis, response to light stimulus, among others. Out of the 52 photosynthetic genes present in maize 16 were found in our list of down-regulated genes, more than nine times what was expected by random chance. Genes involved in mitotic cytokinesis and light stimulus were also overrepresented more than five times.

Table 3-3. Annotation for top 30 most up or down-regulated differentially expressed genes in B73;*Rp1-D21*/+ versus wildtype (BR) samples.

Gene	Length	Chromosome	Description	log2FoldChange	padj
Zm00001d025200	1978	10	Indolin-2-one monooxygenase / CYP71C2	9.927645441	3.0576E-14
Zm00001d012024	6401	8	Nucleotide-diphospho-sugar transferase (Nucleotid_trans)	9.30893196	1.53954E-12
Zm00001d023811	1063	10	Pathogenesis-related protein Bet v I family (Bet_v_1)	9.169621403	9.9857E-116
Zm00001d048087	1418	9	Naringenin 7-O-methyltransferase	8.87713028	3.74515E-11
Zm00001d050805	7071	4	Protein kinase domain (Pkinase) // Mlo family (Mlo)	8.872157982	7.54956E-11
Zm00001d023774	958	10	ERGOSTEROL BIOSYNTHETIC PROTEIN 28-RELATED	8.839560206	1.54075E-13
Zm00001d031157	534	1	Thaumatin family (Thaumatin)	8.812819136	4.20238E-11
Zm00001d053001	1257	4	Nepenthesin / Nepenthes aspartic proteinase	8.801794158	2.80339E-11
Zm00001d006531	3407	2	Salt stress response/antifungal (Stress-antifung)	8.775827859	4.7311E-11
Zm00001d043947	1596	3	Phosphatidate phosphatase / Phosphatidic acid phospphatase	8.717184604	3.24612E-10
Zm00001d012430	636	8	C2H2-type zinc finger (zf-C2H2_6)	8.600907369	7.91851E-10
Zm00001d048913	606	4	CALCIUM-BINDING PROTEIN CML30-RELATED	8.549993886	1.31143E-09
Zm00001d021266	588	7	BON1-ASSOCIATED PROTEIN 1-RELATED	8.53644582	6.72476E-10
Zm00001d022593	1300	7	Ornithine decarboxylase / L-ornithine carboxy-lyase	8.496473309	2.13816E-08
Zm00001d037113	3790	6	ANNOTATION UNKOWN	-5.751225654	0.013236692
Zm00001d015330	375	5	ANNOTATION UNKOWN	-5.837793989	0.015357573
Zm00001d025957	1542	10	B-box zinc finger (zf-B_box)	-5.879912291	0.040201114
Zm00001d045357	871	9	ANNOTATION UNKOWN	-5.886704519	0.021188721
Zm00001d034415	1275	1	Probable lipid transfer (LTP_2)	-5.896929462	0.010377257
Zm00001d024978	6071	10	LEUCINE-RICH REPEAT-CONTAINING PROTEIN	-5.90409775	0.012842413
Zm00001d033605	2212	1	uncharacterized protein (K06966)	-5.90409775	0.012842413
Zm00001d051630	2685	4	MPS ONE BINDER KINASE ACTIVATOR-LIKE MOB	-5.960261087	0.038297049
Zm00001d012280	1017	8	Dof domain, zinc finger (zf-Dof)	-6.288788231	0.018412169
Zm00001d051739	453	4	BZIP PROTEIN (ATBZIP48)-RELATED	-6.327099625	0.00114791
Zm00001d053065	5108	4	MLO-LIKE PROTEIN 4	-6.517401825	0.000489942
Zm00001d009772	786	8	CYSTEINE-RICH SECRETORY PROTEIN-RELATED	-6.635951101	0.000384312
Zm00001d040627	2405	3	Cation transport protein (TrkH)	-6.860999772	5.18163E-05
Zm00001d021501	1786	7	ANNOTATION UNKOWN	-7.012486393	4.35218E-05
Zm00001d034099	3966	1	Ferredoxin--NADP(+) reductase	-8.674283109	1.42698E-09

Table 3-4. GO annotations for down-regulated genes in the B73;*Rp1-D21*/+ versus wildtype background.

GO-Slim Biological Process	Zea mays - REFLIST	Actual	Expected	Fold Enrichment	Raw P-value	FDR
photosynthesis (GO:0015979)	52	16	1.77	9.03	5.17E-10	7.71E-07
mitotic cytokinesis (GO:0000281)	35	6	1.19	5.03	2.12E-03	4.72E-02
cytoskeleton-dependent cytokinesis (GO:0061640)	35	6	1.19	5.03	2.12E-03	4.65E-02
response to light stimulus (GO:0009416)	108	16	3.68	4.35	3.39E-06	2.41E-04
response to radiation (GO:0009314)	123	16	4.19	3.82	1.49E-05	6.73E-04
response to abiotic stimulus (GO:0009628)	189	19	6.44	2.95	6.59E-05	2.52E-03
multicellular organism development (GO:0007275)	165	15	5.62	2.67	9.70E-04	2.37E-02
generation of precursor metabolites and energy (GO:0006091)	233	21	7.94	2.65	1.18E-04	3.99E-03
multicellular organismal process (GO:0032501)	169	15	5.76	2.61	1.21E-03	2.78E-02
anatomical structure development (GO:0048856)	195	17	6.64	2.56	7.13E-04	1.90E-02
developmental process (GO:0032502)	262	21	8.93	2.35	5.83E-04	1.70E-02

On the other hand, up-regulated genes were overwhelmingly involved in defense-related processes. Biological processes with the highest fold enrichment from this list notably included defense response to fungus, innate immune responses, immune system process, and programmed cell death. Six out of eight maize genes implicated in defense response to fungi were present in our list of up-regulated genes, almost six times what would be expected by chance. Similarly, half of the genes

involved in programmed cell death or cell death generally were present in our list and were almost four times more enriched in mutants than in wild-type plants (Table 3-5).

Table 3-5. GO annotations for up-regulated genes in the B73;*Rp1-D21*/+ versus wildtype background.

GO-Slim Biological Process	Zea mays - REFLIST	Actual	Expected	Fold Enrichment	Raw P-value	FDR
defense response to fungus (GO:0050832)	8	6	1.06	5.69	3.26E-03	3.83E-02
innate immune response (GO:0045087)	11	8	1.45	5.51	7.77E-04	1.30E-02
immune system process (GO:0002376)	14	8	1.85	4.33	2.38E-03	3.12E-02
immune response (GO:0006955)	14	8	1.85	4.33	2.38E-03	3.09E-02
organic anion transport (GO:0015711)	19	10	2.51	3.99	1.12E-03	1.78E-02
programmed cell death (GO:0012501)	18	9	2.37	3.79	2.58E-03	3.29E-02
cell death (GO:0008219)	18	9	2.37	3.79	2.58E-03	3.26E-02
glutathione metabolic process (GO:0006749)	51	24	6.73	3.57	2.42E-06	1.50E-04
tricarboxylic acid cycle (GO:0006099)	34	15	4.48	3.35	3.15E-04	6.18E-03
maturation of LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA) (GO:0000463)	25	11	3.3	3.34	1.97E-03	2.65E-02
cell surface receptor signaling pathway (GO:0007166)	60	25	7.91	3.16	8.20E-06	3.50E-04

Table 3-5 continued

glucose 6-phosphate metabolic process (GO:0051156)	27	11	3.56	3.09	3.18E-03	3.82E-02
maturation of LSU-rRNA (GO:0000470)	32	13	4.22	3.08	1.42E-03	2.14E-02
ribosomal large subunit assembly (GO:0000027)	32	13	4.22	3.08	1.42E-03	2.11E-02
cellular modified amino acid metabolic process (GO:0006575)	72	29	9.5	3.05	2.76E-06	1.52E-04
response to unfolded protein (GO:0006986)	40	16	5.28	3.03	4.80E-04	8.84E-03
cellular response to unfolded protein (GO:0034620)	40	16	5.28	3.03	4.80E-04	8.73E-03
ribosomal large subunit biogenesis (GO:0042273)	95	37	12.53	2.95	3.22E-07	2.83E-05
aromatic amino acid family metabolic process (GO:0009072)	57	22	7.52	2.93	7.02E-05	2.01E-03
monocarboxylic acid catabolic process (GO:0072329)	37	14	4.88	2.87	1.63E-03	2.38E-02
lipid oxidation (GO:0034440)	44	16	5.8	2.76	1.68E-03	2.42E-02
cytoplasmic translation (GO:0002181)	105	37	13.85	2.67	2.17E-06	1.47E-04
secretion by cell (GO:0032940)	78	27	10.29	2.62	8.55E-05	2.32E-03
export from cell (GO:0140352)	78	27	10.29	2.62	8.55E-05	2.28E-03
secretion (GO:0046903)	78	27	10.29	2.62	8.55E-05	2.24E-03
vesicle fusion to plasma membrane (GO:0099500)	68	22	8.97	2.45	7.19E-04	1.25E-02

Table 3-5 continued

exocytic process (GO:0140029)	68	22	8.97	2.45	7.19E-04	1.23E-02
exocytosis (GO:0006887)	68	22	8.97	2.45	7.19E-04	1.22E-02
cellular amine metabolic process (GO:0044106)	50	16	6.59	2.43	3.47E-03	4.04E-02
cellular biogenic amine metabolic process (GO:0006576)	50	16	6.59	2.43	3.47E-03	4.01E-02
response to external biotic stimulus (GO:0043207)	136	43	17.94	2.4	3.31E-06	1.76E-04
response to other organism (GO:0051707)	136	43	17.94	2.4	3.31E-06	1.70E-04
response to biotic stimulus (GO:0009607)	136	43	17.94	2.4	3.31E-06	1.65E-04
defense response to other organism (GO:0098542)	136	43	17.94	2.4	3.31E-06	1.59E-04
biological process involved in interspecies interaction between organisms (GO:0044419)	137	43	18.07	2.38	3.71E-06	1.73E-04
monocarboxylic acid metabolic process (GO:0032787)	215	64	28.36	2.26	9.58E-08	1.19E-05
carbohydrate catabolic process (GO:0016052)	121	36	15.96	2.26	8.10E-05	2.28E-03
cellular response to topologically incorrect protein (GO:0035967)	68	20	8.97	2.23	3.86E-03	4.33E-02
response to topologically incorrect protein (GO:0035966)	68	20	8.97	2.23	3.86E-03	4.30E-02
defense response (GO:0006952)	219	64	28.88	2.22	1.93E-07	2.22E-05
carboxylic acid catabolic process (GO:0046395)	93	27	12.27	2.2	8.29E-04	1.37E-02

Table 3-5 continued

response to endoplasmic reticulum stress (GO:0034976)	93	27	12.27	2.2	8.29E-04	1.36E-02
amine metabolic process (GO:0009308)	73	21	9.63	2.18	3.16E-03	3.84E-02
fatty acid metabolic process (GO:0006631)	78	22	10.29	2.14	2.75E-03	3.42E-02
small molecule catabolic process (GO:0044282)	174	47	22.95	2.05	5.06E-05	1.64E-03
organic acid catabolic process (GO:0016054)	105	28	13.85	2.02	1.67E-03	2.42E-02
response to external stimulus (GO:0009605)	169	45	22.29	2.02	9.64E-05	2.48E-03
endoplasmic reticulum to Golgi vesicle-mediated transport (GO:0006888)	115	29	15.17	1.91	3.75E-03	4.27E-02
sulfur compound metabolic process (GO:0006790)	176	41	23.21	1.77	1.92E-03	2.60E-02
carboxylic acid metabolic process (GO:0019752)	569	132	75.05	1.76	2.69E-08	4.46E-06
carbohydrate derivative metabolic process (GO:1901135)	411	95	54.21	1.75	2.61E-06	1.49E-04
oxoacid metabolic process (GO:0043436)	572	132	75.44	1.75	4.04E-08	6.03E-06
organic acid biosynthetic process (GO:0016053)	245	56	32.31	1.73	4.01E-04	7.57E-03
protein folding (GO:0006457)	254	58	33.5	1.73	3.57E-04	6.83E-03
carboxylic acid biosynthetic process (GO:0046394)	228	52	30.07	1.73	7.02E-04	1.23E-02
ribosome biogenesis (GO:0042254)	342	78	45.11	1.73	3.25E-05	1.10E-03

Table 3-5 continued

cellular response to chemical stimulus (GO:0070887)	282	64	37.19	1.72	1.93E-04	4.30E-03
organic acid metabolic process (GO:0006082)	591	134	77.95	1.72	6.46E-08	8.77E-06
Golgi vesicle transport (GO:0048193)	199	45	26.25	1.71	1.80E-03	2.46E-02
carbohydrate metabolic process (GO:0005975)	410	92	54.07	1.7	1.25E-05	4.92E-04
purine-containing compound metabolic process (GO:0072521)	210	47	27.7	1.7	1.70E-03	2.39E-02
nucleobase-containing small molecule metabolic process (GO:0055086)	316	70	41.68	1.68	1.68E-04	3.97E-03
cellular response to organic substance (GO:0071310)	226	50	29.81	1.68	1.76E-03	2.43E-02
purine ribonucleotide metabolic process (GO:0009150)	181	40	23.87	1.68	4.38E-03	4.81E-02
purine nucleotide metabolic process (GO:0006163)	195	43	25.72	1.67	3.21E-03	3.83E-02
transmembrane transport (GO:0055085)	306	67	40.36	1.66	3.34E-04	6.47E-03
small molecule metabolic process (GO:0044281)	1020	223	134.53	1.66	4.58E-11	6.83E-08
response to organic substance (GO:0010033)	299	63	39.43	1.6	1.26E-03	1.97E-02
response to chemical (GO:0042221)	401	84	52.89	1.59	2.03E-04	4.46E-03
cellular amide metabolic process (GO:0043603)	666	139	87.84	1.58	2.20E-06	1.43E-04
carbohydrate derivative biosynthetic process (GO:1901137)	255	53	33.63	1.58	4.14E-03	4.57E-02

Table 3-5 continued

cellular carbohydrate metabolic process (GO:0044262)	258	53	34.03	1.56	4.42E-03	4.81E-02
lipid metabolic process (GO:0006629)	492	100	64.89	1.54	1.35E-04	3.26E-03
cellular lipid metabolic process (GO:0044255)	406	82	53.55	1.53	6.41E-04	1.14E-02
peptide metabolic process (GO:0006518)	596	119	78.61	1.51	6.15E-05	1.95E-03
ribonucleoprotein complex biogenesis (GO:0022613)	418	83	55.13	1.51	9.64E-04	1.55E-02
small molecule biosynthetic process (GO:0044283)	356	70	46.95	1.49	3.01E-03	3.68E-02
organonitrogen compound biosynthetic process (GO:1901566)	1108	215	146.13	1.47	4.19E-07	3.47E-05
vesicle-mediated transport (GO:0016192)	516	100	68.06	1.47	6.18E-04	1.11E-02
response to stress (GO:0006950)	919	177	121.21	1.46	7.01E-06	3.07E-04
nitrogen compound transport (GO:0071705)	503	96	66.34	1.45	1.30E-03	2.00E-02
organophosphate metabolic process (GO:0019637)	458	87	60.41	1.44	2.44E-03	3.14E-02
organic substance transport (GO:0071702)	551	103	72.67	1.42	1.69E-03	2.40E-02
amide biosynthetic process (GO:0043604)	573	106	75.57	1.4	1.70E-03	2.37E-02
protein localization (GO:0008104)	558	103	73.59	1.4	2.21E-03	2.95E-02
transport (GO:0006810)	1349	246	177.92	1.38	4.63E-06	2.09E-04
establishment of localization (GO:0051234)	1365	246	180.03	1.37	9.05E-06	3.75E-04

Table 3-5 continued

localization (GO:0051179)	1556	270	205.22	1.32	3.49E-05	1.16E-03
cellular response to stimulus (GO:0051716)	1139	195	150.22	1.3	8.41E-04	1.36E-02
response to stimulus (GO:0050896)	1619	277	213.53	1.3	6.58E-05	1.96E-03
organic substance catabolic process (GO:1901575)	1070	181	141.12	1.28	2.34E-03	3.09E-02
catabolic process (GO:0009056)	1153	191	152.07	1.26	3.79E-03	4.29E-02
organonitrogen compound metabolic process (GO:1901564)	3226	510	425.48	1.2	9.95E-05	2.52E-03
biological_process (GO:0008150)	9681	1491	1276.82	1.17	2.06E-10	1.54E-07
cellular process (GO:0009987)	8516	1283	1123.17	1.14	6.54E-07	5.14E-05

Rp1-D21/+ versus wildtype in the H95 inbred background

DGE analysis was also carried out on H95;*Rp1-D21/+* F1 plants segregating 1:1 for wildtype and mutant phenotypes to examine the impact of *Rp1-D21* in the H95 background. Single-end reads from three biological replicates each from wild-type and mutant phenotypes were processed to identify genes that differ in expression between the two phenotypes. Input read counts ranged from 17.8 – 30.3M per sample were mapped to an H95-anonymized B73RefGen_v4 reference genome (Table 3-6). DNA sequence from the H95 background was used to identify polymorphisms between this inbred background and the B73 reference genome. All polymorphisms were then converted to an ambiguous "N" and a new reference that did not score as mis-matched reads with H95 polymorphisms in the mapping step was saved. This anonymized reference improved mapping rate of the H95 alleles. The proportion of uniquely mapped reads was between 71.2 – 76.02%. These alignment rates were comparable to those of the B73 samples, suggesting that anonymizing the reference prior to mapping was a valuable step. Such a step is not commonplace in RNA-seq experiments, but certainly in species with high rates of nucleotide substitution between accessions, such as maize, reference bias has the potential to alter the effectiveness of alignment-based genomics approaches.

Table 3-6. Mapping statistics for H95;*Rp1-D21*/+ versus wildtype.

Sample	Background	Phenotype	Raw reads	Average read length (bp)	Uniquely mapped reads %	Multi-mapped reads %	Unmapped reads %
HRwt_rep1	H95;*Rp1-D21*	wildtype	22,549,312	125	76.02%	18.96%	5.02%
HRwt_rep2	H95;*Rp1-D21*	wildtype	19,216,656	125	71.23%	23.28%	5.50%
HRwt_rep3	H95;*Rp1-D21*	wildtype	21,524,749	125	72.90%	21.52%	5.58%
HRmu_rep1	H95;*Rp1-D21*	*Rp1-D21*	18,265,368	125	76.88%	19.17%	3.94%
HRmu_rep2	H95;*Rp1-D21*	*Rp1-D21*	30,354,314	125	72.43%	23.40%	4.17%
HRmu_rep3	H95;*Rp1-D21*	*Rp1-D21*	17,899,060	125	75.17%	19.93%	4.90%

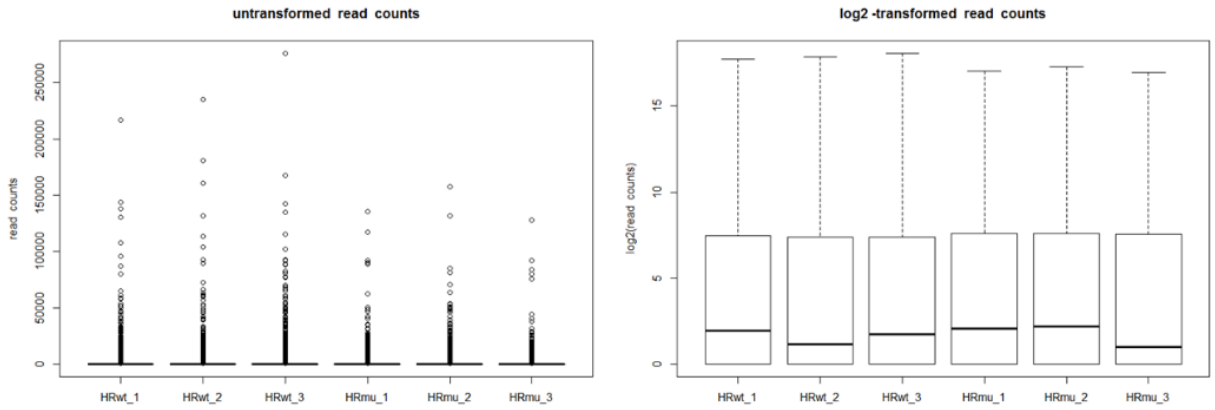


Figure 3-7. Comparison between untransformed and *log*$_2$-transformed read count distribution of *Rp1-D21*/+;H95 samples showing the effect of transformation in reducing skewness.

Table 3-7. Similarity of read counts between H95;*Rp1-D21*/+ versus wildtype replicates as measured by Pearson correlation coefficient.

Pairwise comparison	Pearson Correlation coefficient
HRwt_1 vs. HRwt_2	0.979
HRwt_2 vs. HRwt_3	0.973
HRmu_1 vs. HRmu_2	0.984
HRmu_2 vs. HRmu_3	0.968

Read count per gene was again observed to be higher in the wildtype samples compared to mutants (Figure 3-7), perhaps reflecting that a healthier plant will have higher accumulation of mRNA per ng of total RNA. Replicates showed great similarity (Table 3-7), and hierarchical clustering and PCA did not uncover the presence of outliers (Figures 3-8 and 3-9). Untransformed read counts were analyzed for differential expression with DESeq2 using wild-type samples as control. A greater proportion of the DEGs were up-regulated by *Rp1-D21*/+ (Figure 3-10). Specifically, 3767 (13%) out of the total expressed genes went up in the mutant relative to the wildtype control. In contrast, 2887 (10%) DEGs were downregulated. Normalized read counts for the top 30 up or downregulated genes were extracted and visualized on a heatmap to get an idea of their expression among the mutant and wildtype samples (Figure 3-11). The heatmap also confirmed the direction of fold change reported in the DEG presented in Table 3-8. Analyses of the top 30 up or downregulated genes with Phytozome v13 identified several defense-related genes to be upregulated.

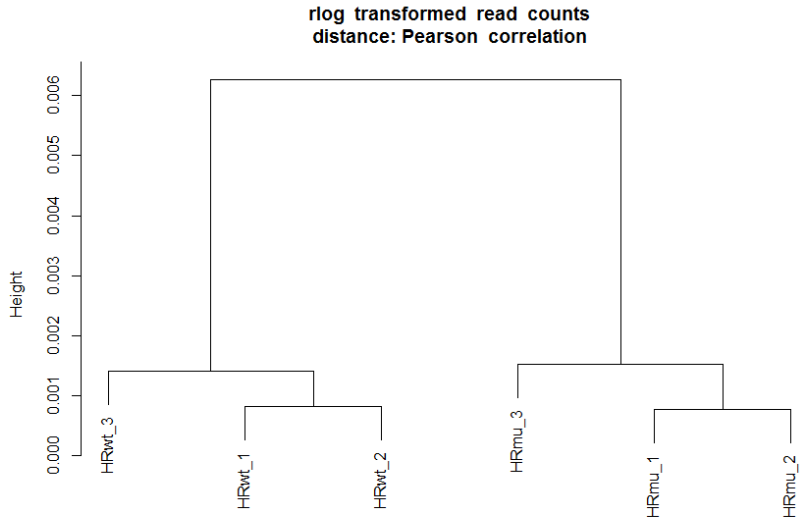


Figure 3-8. Dendrogram showing results of hierarchical clustering of H95;*Rp1-D21*/+ versus wildtype samples. Replicates displayed greater similarity whereas the two phenotypes were clearly separated.

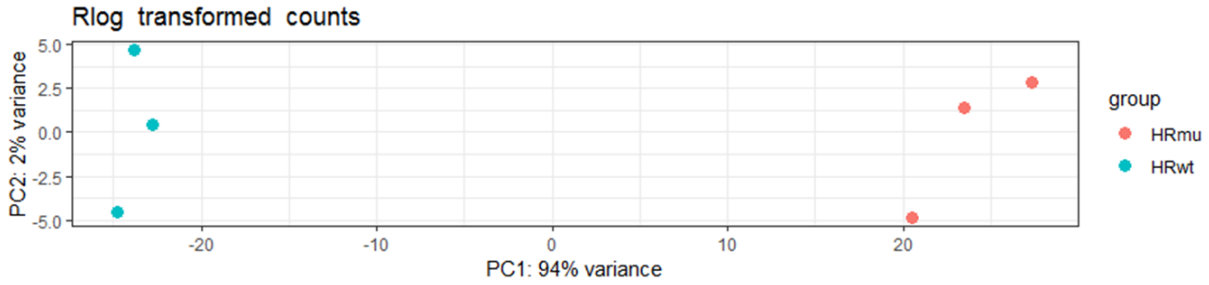


Figure 3-9. PCA on *rlog*-transformed read counts for H95;*Rp1-D21*/+ versus wildtype samples. Differences between phenotypes account for greater proportion of variance.

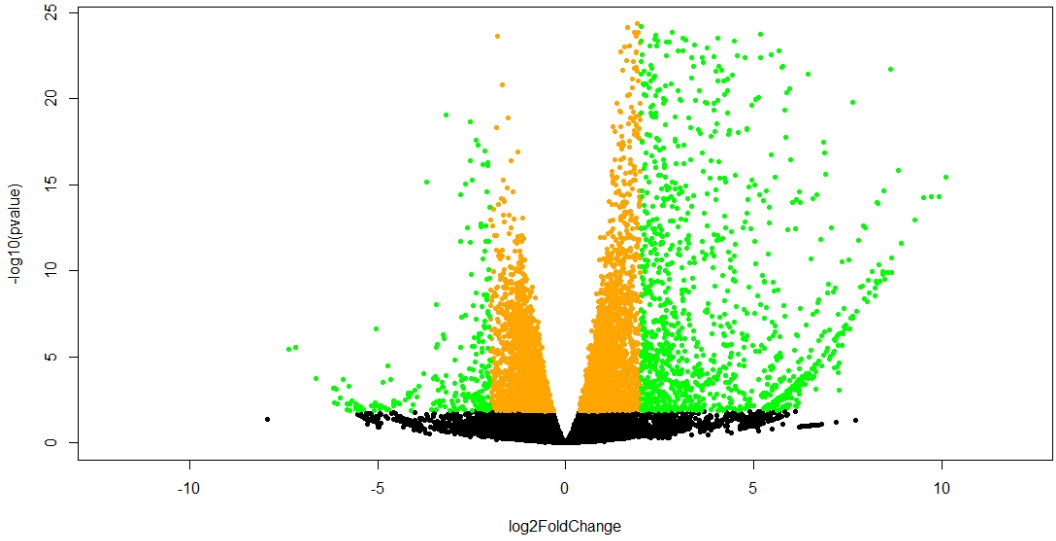


Figure 3-10. Volcano plot of H95;*Rp1-D21*/+ versus wildtype DGE results depicting statistical significance (p-value) versus magnitude of change (fold change). Black dots are genes that are not statistically significant (adjusted p-value > 0.05), orange dots are statistically significant genes (adjusted p-value < 0.05), green dots are statistically significant genes with absolute log_2 fold change of 2.

A number of defense related genes are observed in the upregulated list of genes. One of the genes most increased in accumulation in the mutants compared to the wildtype plants were Zm00001d042934 and Zm00001d045211 (C2H2-type zinc finger, zf-C2H2_6). These zinc finger protein genes play a key role in stress resistance in plants (Han et al 2020). Another notable accumulated gene was Zm00001d048913 (CALCIUM-BINDING PROTEIN). These proteins are Ca2+ sensors that work in conjunction with mitogen activated protein kinases (MAPKs) to regulate

expression of defense genes (Yuan et al., 2022). Zm00001d043622 and Zm00001d011737 (Leucine-rich repeat-containing protein genes) which were accumulated in mutants to 724- and 362-times the levels of wild-type samples, respectively, have been reported to be involved in mediating defense response to fungal pathogens (Block et al., 2021). Down regulated genes included Zm00001d039397 (Very-long-chain 3-oxoacyl-CoA reductase / Very-long-chain beta-ketoacyl-CoA reductase), which plays a role in the formation of cell membranes of higher plants and algae and by extension affect cell division and differentiation (Haslam & Kunst, 2013; Zhukov & Popov, 2022). Zm00001d022098 (tRNA pseudouridine (55) synthase), which converts uridine to pseudouridine is vital for translation efficiency (Xie et al., 2022) was also among the top downregulated genes. Again, as genes implicated in defense response were overexpressed in the *Rp1-D21*/+ plants whereas genes required for normal cellular functions were generally downregulated.

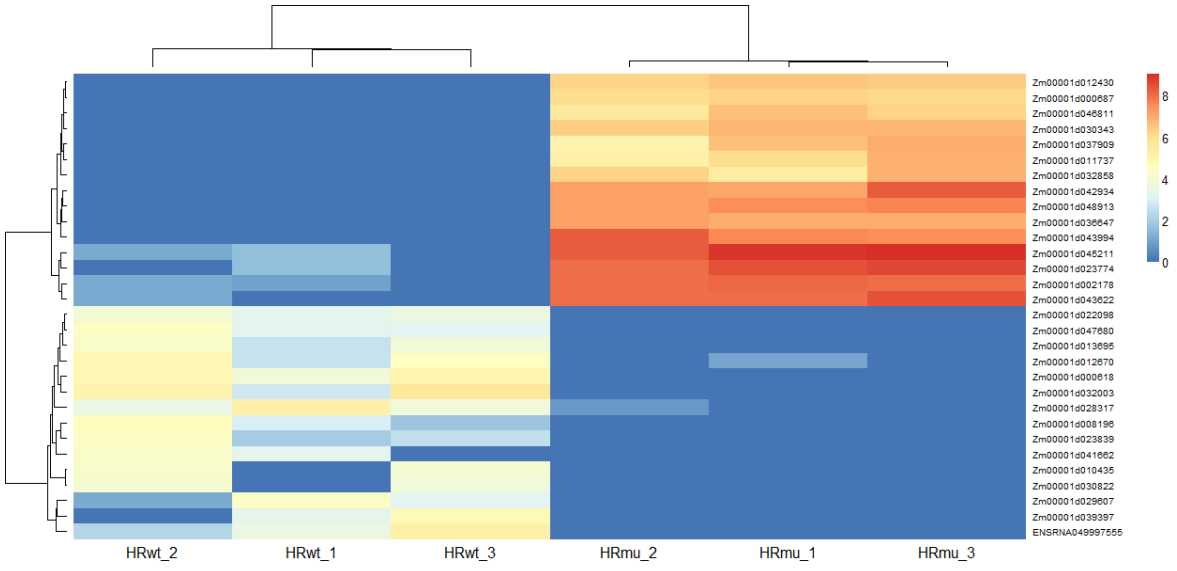


Figure 3-11. Heatmap of log_2-transformed read counts for top 30 most up or down-regulated genes after DGE analysis of H95;*Rp1-D21*/+ versus wildtype samples. Genes are sorted based on hierarchical clustering.

Table 3-8. Annotation for top 30 most up or down-regulated differentially expressed genes in H95;*Rp1-D21*/+ versus wildtype samples.

Gene	Length	Chromosome	Description	log2FoldChange	padj
Zm00001d043994	2431	3	ANNOTATION UNKOWN	10.11111	1.78E-14
Zm00001d042934	1077	3	C2H2-type zinc finger (zf-C2H2_6)	9.922837	2.20E-13
Zm00001d048913	606	4	CALCIUM-BINDING PROTEIN CML30-RELATED	9.726615	2.16E-13
Zm00001d043622	1569	3	LEUCINE-RICH REPEAT-CONTAINING PROTEIN	9.531791	2.64E-13
Zm00001d036647	3883	6	speckle-type POZ protein (SPOP)	9.299148	4.39E-12
Zm00001d030343	1608	1	Very-long-chain 3-oxoacyl-CoA synthase / Very-long-chain beta-ketoacyl-CoA synthase	8.9287	8.21E-11
Zm00001d023774	958	10	ERGOSTEROL BIOSYNTHETIC PROTEIN 28-RELATED	8.858192	7.80E-15
Zm00001d037909	522	6	NPR1 interacting (NPR1_interact)	8.667048	3.35E-09
Zm00001d012430	636	8	C2H2-type zinc finger (zf-C2H2_6)	8.661254	5.36E-10
Zm00001d045211	537	9	C2H2-type zinc finger (zf-C2H2_6)	8.654814	1.60E-20
Zm00001d032858	6183	1	Ent-pimara-8(14),15-diene synthase	8.574602	3.27E-09
Zm00001d011737	1560	8	LEUCINE-RICH REPEAT-CONTAINING PROTEIN	8.482656	8.26E-09
Zm00001d002178	4112	2	Non-specific serine/threonine protein kinase / Threonine-specific protein kinase	8.464001	9.94E-14
Zm00001d046811	2907	9	Dihydrokaempferol 4-reductase / NADPH-dihydromyricetin reductase	8.443807	4.61E-09
Zm00001d041662	982	3	ANNOTATION UNKOWN	-5.726324	0.048819

Table 3-8 continued.

Zm00001d023839	11013	10	condensin-2 complex subunit D3 (NCAPD3)	-5.729652	0.022818
Zm00001d012670	3440	8	ANNOTATION UNKOWN	-5.759316	0.003137
Zm00001d030822	7782	1	LANOSTEROL SYNTHASE	-5.813583	0.041367
Zm00001d029607	1893	1	PROTEIN SHORT-ROOT	-5.813604	0.021851
Zm00001d028317	3478	1	on-specific serine/threonine protein kinase / Threonine-specific protein kinase	-5.928997	0.001453
Zm00001d008196	997	8	ANNOTATION UNKOWN	-5.983376	0.012158
Zm00001d010435	721	8	ANNOTATION UNKOWN	-6.085148	0.022842
Zm00001d022098	4867	7	tRNA pseudouridine(55) synthase / tRNA Psi(55) synthase	-6.09893	0.004397
Zm00001d047680	3370	9	Calmodulin binding protein-like (Calmodulin_bind)	-6.155122	0.004097
Zm00001d039397	2124	3	Very-long-chain 3-oxoacyl-CoA reductase / Very-long-chain beta-ketoacyl-CoA reductase	-6.15786	0.01938
Zm00001d032003	978	1	Arabidopsis protein of unknown function (DUF241)	-7.365104	4.17E-05

Similar to what was observed in the B73 background downregulated genes were enriched for growth and developmental genes whereas upregulated genes were predominantly defense genes (Table 3-9). Five out of the six GO categories enriched in the downregulated genes were related to cell division. Mitosis, cytoskeleton, and cytokinesis were overrepresented almost seven times in the downregulated genes. Out of the 35 genes reportedly involved in cell division in maize, eight were found in our list of downregulated genes, almost five times more than expected. Genes involved in the transport of potassium (K^+)—an ion required for cellular functions such as keeping electrical potential and gradients within cell membranes as well as maintenance of turgor and enzyme activation (Britto & Kronzucker, 2008)—were enriched about six times more in the wildtype plants than mutants. Upregulated genes were mostly involved in defense, as has been noted in the previous experiments (Table 3-10). GO terms with most significant hits included defense response to fungus, immune fungus, program cell death, cell surface signaling.

Table 3-9. GO annotations for down-regulated genes from the H95;*Rp1-D21*/+ versus wildtype background.

GO-Slim Biological Process	Zea mays - REFLIST	Actual	Expected	Fold Enrichment	Raw P-value	FDR
mitotic cytokinesis (GO:0000281)	35	7	1.05	6.68	1.91E-04	1.35E-02
cytoskeleton-dependent cytokinesis (GO:0061640)	35	7	1.05	6.68	1.91E-04	1.29E-02
potassium ion transport (GO:0006813)	38	7	1.14	6.15	2.97E-04	1.71E-02
membrane fission (GO:0090148)	43	7	1.29	5.44	5.77E-04	2.46E-02
cytokinesis (GO:0000910)	43	7	1.29	5.44	5.77E-04	2.39E-02
cell division (GO:0051301)	54	8	1.62	4.95	4.18E-04	1.95E-02

Table 3-10. GO annotations for up-regulated genes from the H95;*Rp1-D21*/+ versus wildtype background.

GO-Slim Biological Process	Zea mays - REFLIST	Actual	Expected	Fold Enrichment	Raw P-value	FDR
defense response to fungus (GO:0050832)	8	7	0.52	13.58	1.23E-05	9.69E-04
innate immune response (GO:0045087)	11	6	0.71	8.47	3.39E-04	9.72E-03
immune system process (GO:0002376)	14	6	0.9	6.65	9.06E-04	2.22E-02
immune response (GO:0006955)	14	6	0.9	6.65	9.06E-04	2.18E-02
programmed cell death (GO:0012501)	18	7	1.16	6.04	5.39E-04	1.46E-02
cell death (GO:0008219)	18	7	1.16	6.04	5.39E-04	1.44E-02
cell surface receptor signaling pathway (GO:0007166)	60	21	3.87	5.43	9.77E-09	1.82E-06
ribosomal large subunit assembly (GO:0000027)	32	11	2.06	5.33	3.70E-05	1.90E-03
double-strand break repair via break-induced replication (GO:0000727)	21	7	1.35	5.17	1.13E-03	2.48E-02

Table 3-10 continued.

endoplasmic reticulum unfolded protein response (GO:0030968)	22	7	1.42	4.94	1.41E-03	2.89E-02
glutathione metabolic process (GO:0006749)	51	14	3.29	4.26	2.80E-05	1.60E-03
ceramide biosynthetic process (GO:0046513)	34	9	2.19	4.11	9.28E-04	2.20E-02
ceramide metabolic process (GO:0006672)	35	9	2.26	3.99	1.10E-03	2.46E-02
xylan biosynthetic process (GO:0045492)	37	9	2.38	3.78	1.54E-03	3.06E-02
cellular modified amino acid metabolic process (GO:0006575)	72	17	4.64	3.66	2.17E-05	1.29E-03
response to external biotic stimulus (GO:0043207)	136	32	8.76	3.65	6.99E-09	3.47E-06
response to other organism (GO:0051707)	136	32	8.76	3.65	6.99E-09	2.61E-06
response to biotic stimulus (GO:0009607)	136	32	8.76	3.65	6.99E-09	2.08E-06
defense response to other organism (GO:0098542)	136	32	8.76	3.65	6.99E-09	1.74E-06
biological process involved in interspecies interaction between organisms (GO:0044419)	137	32	8.83	3.63	8.12E-09	1.73E-06
response to unfolded protein (GO:0006986)	40	9	2.58	3.49	2.44E-03	4.44E-02
cellular response to unfolded protein (GO:0034620)	40	9	2.58	3.49	2.44E-03	4.38E-02
sphingolipid biosynthetic process (GO:0030148)	47	10	3.03	3.3	2.05E-03	3.82E-02
ribosomal large subunit biogenesis (GO:0042273)	95	20	6.12	3.27	1.82E-05	1.18E-03
ubiquitin-dependent ERAD pathway (GO:0030433)	56	11	3.61	3.05	2.17E-03	4.00E-02

Table 3-10 continued.

defense response (GO:0006952)	219	43	14.11	3.05	2.55E-09	3.81E-06
response to endoplasmic reticulum stress (GO:0034976)	93	18	5.99	3	1.20E-04	4.26E-03
response to external stimulus (GO:0009605)	169	32	10.89	2.94	5.32E-07	6.10E-05
cytoplasmic translation (GO:0002181)	105	19	6.77	2.81	1.70E-04	5.41E-03
ribonucleoprotein complex assembly (GO:0022618)	152	26	9.79	2.65	4.28E-05	2.06E-03
non-membrane-bounded organelle assembly (GO:0140694)	137	23	8.83	2.61	1.47E-04	4.87E-03
ribonucleoprotein complex subunit organization (GO:0071826)	157	26	10.12	2.57	5.64E-05	2.55E-03
hormone-mediated signaling pathway (GO:0009755)	154	24	9.92	2.42	2.14E-04	6.53E-03
cellular response to hormone stimulus (GO:0032870)	154	24	9.92	2.42	2.14E-04	6.39E-03
cellular response to endogenous stimulus (GO:0071495)	157	24	10.12	2.37	4.11E-04	1.16E-02
cellular response to organic substance (GO:0071310)	226	34	14.56	2.33	3.41E-05	1.89E-03
response to organic substance (GO:0010033)	299	44	19.27	2.28	3.27E-06	2.87E-04
cellular response to chemical stimulus (GO:0070887)	282	40	18.17	2.2	1.70E-05	1.21E-03
response to hormone (GO:0009725)	177	25	11.4	2.19	8.78E-04	2.18E-02
response to endogenous stimulus (GO:0009719)	180	25	11.6	2.16	9.66E-04	2.25E-02
ribosome biogenesis (GO:0042254)	342	45	22.04	2.04	3.55E-05	1.89E-03

Table 3-10 continued.

signal transduction (GO:0007165)	627	82	40.4	2.03	2.69E-08	4.46E-06
signaling (GO:0023052)	634	82	40.85	2.01	3.31E-08	4.95E-06
cell communication (GO:0007154)	650	82	41.88	1.96	1.12E-07	1.52E-05
response to chemical (GO:0042221)	401	50	25.84	1.94	3.87E-05	1.93E-03
ribonucleoprotein complex biogenesis (GO:0022613)	418	50	26.93	1.86	1.20E-04	4.09E-03
cellular amide metabolic process (GO:0043603)	666	78	42.91	1.82	2.65E-06	2.47E-04
peptide metabolic process (GO:0006518)	596	66	38.4	1.72	7.28E-05	2.94E-03
cellular response to stimulus (GO:0051716)	1139	123	73.39	1.68	2.12E-07	2.64E-05
response to stimulus (GO:0050896)	1619	171	104.32	1.64	3.39E-09	2.53E-06
protein phosphorylation (GO:0006468)	493	52	31.77	1.64	1.44E-03	2.90E-02
response to stress (GO:0006950)	919	96	59.22	1.62	1.79E-05	1.21E-03
amide biosynthetic process (GO:0043604)	573	59	36.92	1.6	9.95E-04	2.28E-02
phosphorylation (GO:0016310)	579	58	37.31	1.55	2.44E-03	4.33E-02
organonitrogen compound biosynthetic process (GO:1901566)	1108	100	71.39	1.4	1.74E-03	3.33E-02
biological_process (GO:0008150)	9681	695	623.79	1.11	1.73E-03	3.35E-02

NC350 x H95;Rp1-D21/+ versus wild type

NC350 x H95;*Rp1-D21/+* wildtype and mutant F1 plants were also analyzed to identify genes differentially expressed in presence of the NC350 genotype. This genotype was previously found to strongly enhance the phenotype of *Rp1-D21/+* relative to B73 (Chintamanani et al., 2010). I expect this to result in the strongest expression effects of the genotypes assessed here, due to the enhanced effects of the genetic background on *Rp1-D21/+* and provide the clearest insight into the pathways and molecular functions altered by the HR affected by *Rp1-D21/+*. Three biological replicates from each phenotype were sequenced for differential gene expression analysis. Again, an anonymized reference was generated. This time DNA sequence from the H95 and NC350 backgrounds were used to identify polymorphisms between these inbred backgrounds and the B73 reference genome. All polymorphisms were then converted to an ambiguous "N" and a new reference that did not bias against reads with either NC350 or H95 polymorphisms at the mapping step was saved. This anonymized reference was used for all alignments. Input reads ranging from 21.3 – 26.7M per sample were aligned to the NC350-anonymized reference genome, out of which 71.4 – 76.91% were uniquely mapped (Table 3-11). This is an indication of high-quality sample preparation and data processing steps and again indicated the value of this novel preprocessing step.

Alignment files were processed together with the reference annotation file to generate read counts per gene for each sample. These counts were higher in the wildtype samples compared to the mutants (Figure 3-12), indicating that the normal, healthy wildtype plants recorded greater reads numbers per library than the growth-impaired mutant plants. Although raw read counts were used for DGE analysis, the quality control analyses that preceded DGE relied on transformed counts. The effect of size-factor normalization followed by log_2-transformation on reducing skewness was clear. This greatly improved the accuracy of algorithms used for clustering and principal component analysis used as a quality control check.

Biological replicates with poorly correlated gene expression usually signal mistakes or other problems at some stage during the experiment. Replicates showed remarkable similarity, more than 97% in pairwise comparisons as measured by Pearson correlation (*r*; Table 3-12). This was above the 90% correlation threshold recommended by the ENCODE consortium (Standards, Guidelines and Best Practices for RNA-Seq V1.0) and is an indication of the quality of experimental procedures from tissue collection through the library preparation and sequencing.

Table 3-11. Mapping statistics for NC350 x H95;*Rp1-D21*/+ versus wildtype.

Sample	Background	Phenotype	Raw reads	Average read length (bp)	Uniquely mapped reads %	Multi-mapped reads %	Unmapped reads %
NHRwt_rep1	NC350 x H95;*Rp1-D21*	wildtype	21,399,539	125	75.62%	19.37%	5.00%
NHRwt_rep2	NC350 x H95;*Rp1-D21*	wildtype	25,645,653	125	76.86%	17.62%	5.52%
NHRwt_rep3	NC350 x H95;*Rp1-D21*	wildtype	22,179,388	125	71.43%	23.28%	5.29%
NHRmu_rep1	NC350 x H95;*Rp1-D21*	*Rp1-D21*	25,567,445	125	72.99%	23.13%	3.88%
NHRmu_rep2	NC350 x H95;*Rp1-D21*	*Rp1-D21*	26,775,218	125	72.63%	23.19%	4.17%
NHRmu_rep3	NC350 x H95;*Rp1-D21*	*Rp1-D21*	26,373,971	125	72.19%	23.95%	3.86%

Hierarchical clustering was performed to group similar samples using the pairwise correlations calculated above. The distance measure (Height) was computed as 1- *r* and the results rendered as a dendrogram. As expected, hierarchical clustering revealed greater gene expression variability between wildtype and mutant phenotypes than among replicates of the same phenotype (Figure 3-13).

Principal components confirmed the results of hierarchical clustering and also unearthed greater gene expression variability between wild-type and mutant phenotypes than what was observed among replicates of the same phenotype. In fact, 99% of variation observed across all samples could be explained by differences between mutant and wild type. Differences among replicates only accounted for less than 1% of total variation (Figure 3-14). This observation is not surprising considering the range of mutant impacts observed in the other backgrounds (Figures 3-

8 and 3-13) and to the fact that mutant phenotype is more severe in the NC350 background than the other backgrounds studied (Chintamanani et al., 2010).

Neither hierarchical clustering nor PCA detected outliers. As a result, the three biological replicates per genotype (mutant or wildtype) were analyzed for DEG using DESeq2. Wildtype samples were set as controls to identify differentially expressed genes. A total of 14,753 genes were identified as significantly differently expressed between mutants and wildtypes. Up-regulated genes marginally outnumbered down-regulated genes (Figure 3-15). Indeed, 7,835 (26%) expressed genes were accumulated to a greater degree by mutants as compared to the 6,918 (23%) that were turned down.

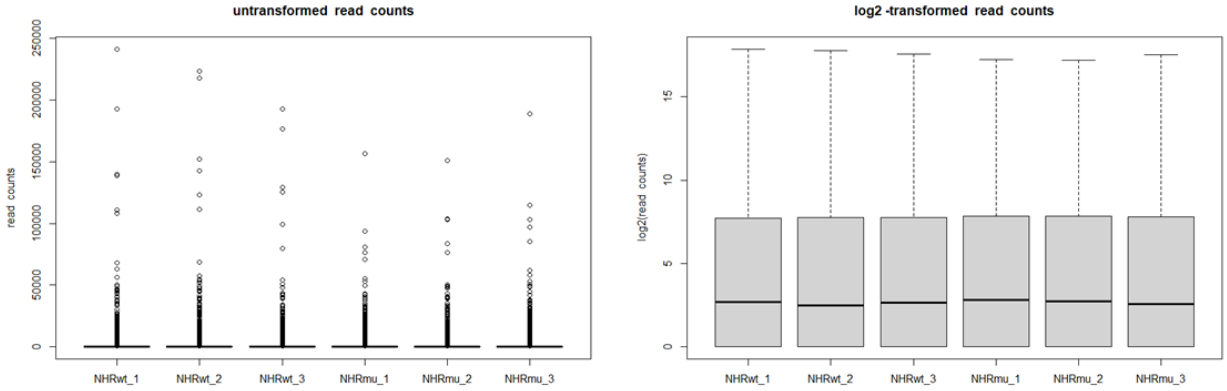


Figure 3-12. Comparison between untransformed and *log*$_2$-transformed read count distribution of NC350 x H95;*Rp1-D21*/+ versus wildtype (NHR) samples showing the effect of transformation in reducing skewness.

Table 3-12. Similarity of read counts between NC350 x H95;*Rp1-D21*/+ versus wildtype (NHR) replicates as measured by Pearson correlation coefficient.

Pairwise comparison	Pearson Correlation coefficient
NHRwt_1 vs. NHRwt_2	0.979
NHRwt_2 vs. NHRwt_3	0.979
NHRmu_1 vs. NHRmu_2	0.985
NHRmu_2 vs. NHRmu_3	0.982

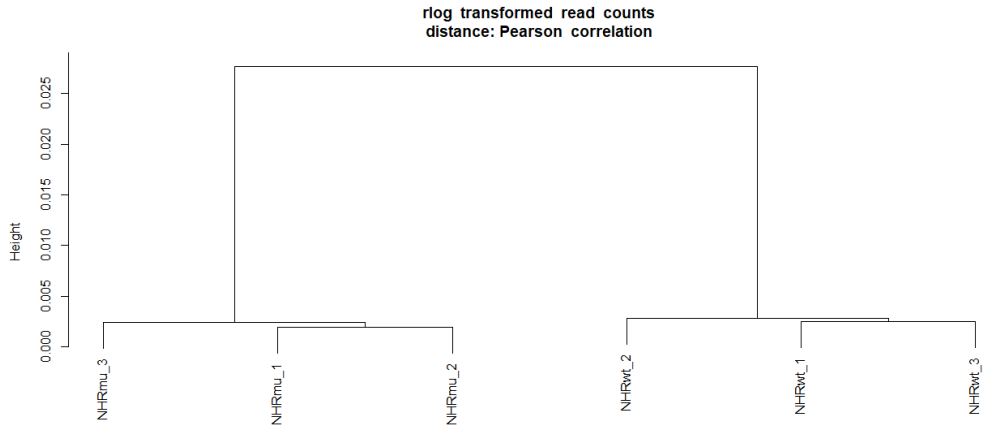


Figure 3-13. Dendrogram showing results of hierarchical clustering of NC350 x H95;*Rp1-D21*/+ versus wildtype (NHR) samples. Replicates displayed greater similarity whereas the two phenotypes were clearly separated.

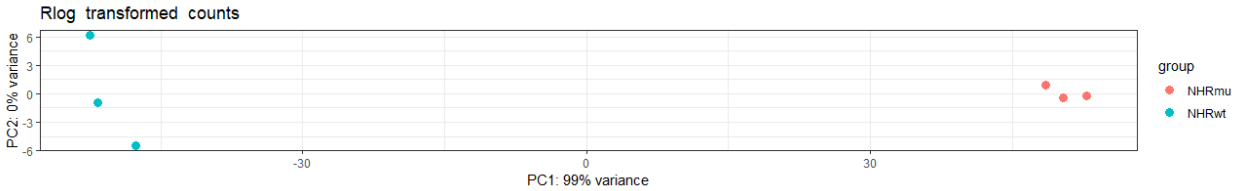


Figure 3-14. PCA on *rlog*-transformed read counts for NC350 x H95;*Rp1-D21*/+ versus wildtype (NHR) samples. Differences between phenotypes account for greater proportion of variance.

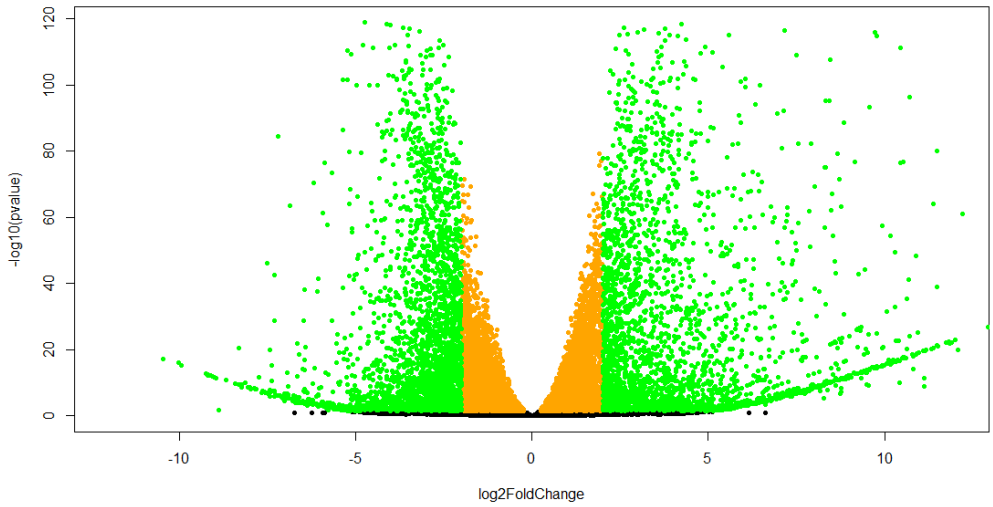


Figure 3-15. Volcano plot of NC350 x H95;*Rp1-D21*/+ versus wildtype (NHR) DGE results depicting statistical significance (p-value) versus magnitude of change (fold change). Black dots are genes that are not statistically significant (adjusted p-value > 0.05), orange dots are statistically significant genes (adjusted p-value < 0.05), green dots are statistically significant genes with absolute log_2 fold change of 2.

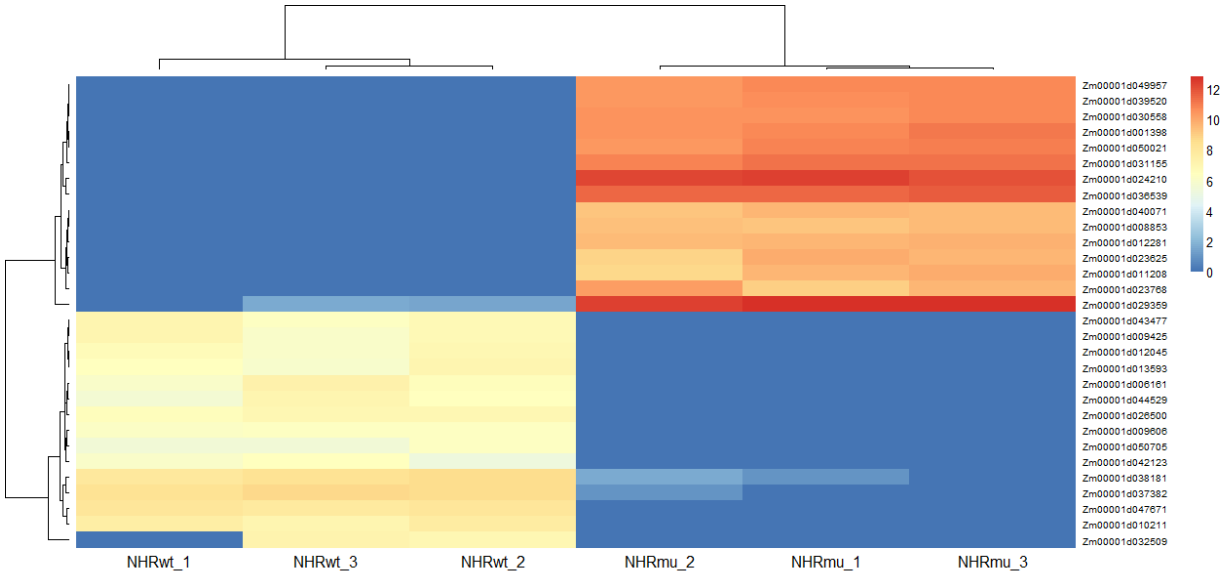


Figure 3-16. Heatmap of log_2-transformed read counts for top 30 most up or down-regulated genes after DGE analysis of NC350 x H95;*Rp1-D21*/+ versus wildtype (NHR) samples. Genes are sorted based on hierarchical clustering.

The heatmap of *log*$_2$-transformed read counts of the top 30 up- or down-regulated genes (Figure 3-16) confirmed the direction of fold-change displayed in the DGE results table. Several defense or stress related genes were among the top up-regulated genes (Table 3-13). Indeed, the upregulated gene with the biggest fold change was Zm00001d024210 ((S)-beta-bisabolene synthase // (S)-beta-macrocarpene synthase). This gene encodes a known terpene in maize β-bisabolene, a secondary metabolite class involved in chemical defense against insect herbivores (Block et al., 2019). Another highly upregulated gene was Zm00001d031155 encoding a Thaumatin family protein. This group of pathogenesis-related (PR) genes play a key function in plant defense (van Loon et al., 2006). They have been known to be key in controlling tolerance to the fungal pathogen *Verticillium dahlia* and drought in several cotton species (Z. Li et al., 2020) Similar studies with transgenic Arabidopsis have revealed that these proteins confer tolerance to two fungi *Scleretonia sclerotiorum* and *Botrytis cinerea* as well as to salinity and dehydration (Misra et al., 2016). Zm00001d049957 which encodes *ent*-kaurene synthase was also overexpressed in the plants showing the mutant phenotype. Although this enzyme is known for the part it plays in the biosynthesis of the phytohormone gibberellin (GA), it also plays additional roles in diterpenoid metabolism (Fu et al., 2016) including secondary metabolites long reported to be overproduced in maize seedlings in response to several pathogenic fungi (Mellon2 & West, 1979). Another gene from this list of *Rp1-D21*/+ upregulated genes was Zm00001d023768 which encodes the omega-6 fatty acid desaturase, FAD2. FAD2 facilitates the biosynthesis of the polyunsaturated linoleic acid from the monounsaturated oleic acid (Mikkilineni & Rocheford, 2003; Ohlroggeav' & Browseb, 1995; Shanklin & Cahoon, 1998) Linoleic acids are known to be used by plants in defense signaling in response to pathogen attack and wounding (Dar et al., 2017; Farmer, 1994). A gene encoding a protein similar to the *Hyoscymus muticus* premnaspirodiene oxygenase (Zm00001d023625), was also highly upregulated. This enzyme is notable for conversion of premnaspirodiene into the phytoalexin solavetivone, known for its strong antifungal properties (Takahashi et al., 2007).

Table 3-13. Annotation for top 30 most up or down-regulated differentially expressed genes in NC350 x H95;*Rp1-D21*/+ versus wildtype (NHR) samples.

Gene	Length	Chromosome	Description	log2FoldChange	padj
Zm00001d024210	3192	10	(S)-beta-bisabolene synthase // (S)-beta-macrocarpene synthase	14.60499	9.06E-34
Zm00001d036539	1023	6	CHITINASE-RELATED	14.01831	3.68E-31
Zm00001d031155	534	1	Thaumatin family (Thaumatin)	13.53141	4.97E-29
Zm00001d050021	5574	4	(+)-abscisic acid 8'-hydroxylase / ABA 8'-hydroxylase	13.16385	2.91E-27
Zm00001d049957	3697	4	Ent-kaurene synthase / Ent-kaurene synthetase B	13.05143	5.35E-27
Zm00001d039520	2267	3	Cytokinin dehydrogenase / Cytokinin oxidase	12.94706	1.36E-26
Zm00001d030558	1490	1	ANNOTATION UNKOWN	12.93233	1.43E-26
Zm00001d029359	2083	1	O-METHYLTRANSFERASE	12.21468	1.97E-60
Zm00001d023768	1155	10	omega-6 fatty acid desaturase (delta-12 desaturase) (FAD2)	12.07258	8.02E-20
Zm00001d012281	753	8	Late embryogenesis abundant protein (LEA_2)	11.99491	7.24E-23
Zm00001d023625	1786	10	Premnaspirodiene oxygenase / Hyoscymus muticus premnaspirodiene oxygenase	11.90385	4.56E-22
Zm00001d011208	1680	8	Aminocyclopropanecarboxylate oxidase / Ethylene-forming enzyme // Non-specific serine/threonine protein kinase / Threonine-specific protein kinase	11.83811	8.67E-22
Zm00001d040071	2612	3	Protein O-GlcNAc transferase / OGTase	11.81396	3.77E-22
Zm00001d008853	3691	8	STEROL REGULATORY ELEMENT-BINDING PROTEIN	11.73954	6.94E-22
Zm00001d050705	1341	4	ANNOTATION UNKOWN	-8.313803	1.22E-09
Zm00001d038181	3297	6	OLIGOPEPTIDE TRANSPORTER-RELATED	-8.31821	2.49E-20

Table 3-13 continued.

Zm00001d042123	2418	3	SULFOTRANSFERASE SULT	-8.408998	7.17E-10
Zm00001d009606	2784	8	ANNOTATION UNKOWN	-8.666946	5.92E-11
Zm00001d032509	1569	1	WD REPEAT DOMAIN 44	-8.875943	0.042973
Zm00001d044529	2628	3	OLIGOPEPTIDE TRANSPORTER-RELATED // PROTEIN NRT1/ PTR FAMILY 6.2	-8.961697	1.70E-11
Zm00001d013593	394	5	Protein of unknown function (DUF1677) (DUF1677)	-9.016516	8.36E-12
Zm00001d012045	900	8	C3HC4 TYPE RING-FINGER PROTEIN-RELATED	-9.063619	4.37E-12
Zm00001d009425	1190	8	ANNOTATION UNKOWN	-9.121169	3.51E-12
Zm00001d043477	5776	3	Cellulose synthase (UDP-forming) / UDP-glucose-cellulose glucosyltransferase	-9.190374	1.52E-12
Zm00001d006161	1676	2	Alpha-L-fucosidase / Alpha-L-fucoside fucohydrolase	-9.200742	2.71E-12
Zm00001d026500	896	10	ANNOTATION UNKOWN	-9.249965	7.42E-13
Zm00001d010211	2871	8	ADENINE NUCLEOTIDE ALPHA HYDROLASE-LIKE PROTEIN	-9.939427	3.62E-15
Zm00001d037382	1428	6	GLUCOSYL/GLUCURONOSYL TRANSFERASES	-10.01569	5.52E-16
Zm00001d047671	4233	9	MYB-LIKE DNA-BINDING PROTEIN MYB // MYB DOMAIN PROTEIN 11-RELATED	-10.45784	4.22E-17

Comparable to what was reported from the GO enrichment analyses in the *Rp1-D21*/+;B73 and H95;*Rp1-D21*/+ backgrounds, NC350 x H95;*Rp1-D21*/+ exhibited downregulated genes enriched for several growth-related biological processes while upregulated genes were mostly involved in defense and immunity related activities. For example, in the downregulated genes photosynthesis terms were enriched 5.81 times, mitotic and cytoskeleton-dependent cytokinesis terms 3.63 times, membrane fission 3.14 times, to name a few (Table 3-14). Among the upregulated genes the category defense response to fungus (GO:0050832) was the most enriched biological process and was enriched 5.81 times than would have been expected by random chance (Table 3-15). Other defense processes overrepresented among the transcripts with increased accumulation included innate immune response (GO:0045087) 5.51 times, immune system process (GO:0002376) 4.33 times, programmed cell death (GO:0012501) 3.79 times more than expected.

Table 3-14. Gene Ontology (GO) annotations for down-regulated genes in the NC350 x H95;*Rp1-D21*/+ versus wildtype.

GO-Slim Biological Process	Zea mays - REFLIST	Actual	Expected	Fold Enrichment	Raw P-value	FDR
photosynthesis (GO:0015979)	52	38	6.54	5.81	5.52E-14	8.23E-11
beta-glucan biosynthetic process (GO:0051274)	23	11	2.89	3.8	8.23E-04	1.40E-02
mitotic cytokinesis (GO:0000281)	35	16	4.4	3.63	8.84E-05	2.64E-03
cytoskeleton-dependent cytokinesis (GO:0061640)	35	16	4.4	3.63	8.84E-05	2.59E-03
regulation of multicellular organismal development (GO:2000026)	29	13	3.65	3.56	4.61E-04	8.49E-03
membrane fission (GO:0090148)	43	17	5.41	3.14	2.18E-04	4.58E-03
cytokinesis (GO:0000910)	43	17	5.41	3.14	2.18E-04	4.52E-03

Table 3-14 continued.

response to light stimulus (GO:0009416)	108	42	13.58	3.09	1.21E-08	3.01E-06
pigment biosynthetic process (GO:0046148)	32	12	4.02	2.98	2.55E-03	3.62E-02
pigment metabolic process (GO:0042440)	35	13	4.4	2.95	1.84E-03	2.74E-02
chloroplast organization (GO:0009658)	49	18	6.16	2.92	2.99E-04	6.03E-03
regulation of multicellular organismal process (GO:0051239)	48	17	6.04	2.82	6.11E-04	1.05E-02
plastid organization (GO:0009657)	57	20	7.17	2.79	3.41E-04	6.70E-03
response to radiation (GO:0009314)	123	42	15.47	2.71	2.96E-07	3.16E-05
cell division (GO:0051301)	54	18	6.79	2.65	9.63E-04	1.58E-02
response to abiotic stimulus (GO:0009628)	189	50	23.77	2.1	1.16E-05	5.11E-04
regulation of developmental process (GO:0050793)	95	25	11.95	2.09	1.98E-03	2.87E-02
generation of precursor metabolites and energy (GO:0006091)	233	53	29.3	1.81	2.15E-04	4.58E-03

Table 3-15. Gene Ontology (GO) annotations for up-regulated genes in the NC350 x H95;*Rp1-D21*/+ versus wildtype.

GO-Slim Biological Process	Zea mays - REFLIST	Actual	Expected	Fold Enrichment	Raw P-value	FDR
defense response to fungus (GO:0050832)	8	6	1.06	5.69	3.26E-03	3.83E-02
innate immune response (GO:0045087)	11	8	1.45	5.51	7.77E-04	1.30E-02
immune system process (GO:0002376)	14	8	1.85	4.33	2.38E-03	3.12E-02
immune response (GO:0006955)	14	8	1.85	4.33	2.38E-03	3.09E-02
organic anion transport (GO:0015711)	19	10	2.51	3.99	1.12E-03	1.78E-02
programmed cell death (GO:0012501)	18	9	2.37	3.79	2.58E-03	3.29E-02
cell death (GO:0008219)	18	9	2.37	3.79	2.58E-03	3.26E-02
glutathione metabolic process (GO:0006749)	51	24	6.73	3.57	2.42E-06	1.50E-04
tricarboxylic acid cycle (GO:0006099)	34	15	4.48	3.35	3.15E-04	6.18E-03
maturation of LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA) (GO:0000463)	25	11	3.3	3.34	1.97E-03	2.65E-02
cell surface receptor signaling pathway (GO:0007166)	60	25	7.91	3.16	8.20E-06	3.50E-04
glucose 6-phosphate metabolic process (GO:0051156)	27	11	3.56	3.09	3.18E-03	3.82E-02
maturation of LSU-rRNA (GO:0000470)	32	13	4.22	3.08	1.42E-03	2.14E-02
ribosomal large subunit assembly (GO:0000027)	32	13	4.22	3.08	1.42E-03	2.11E-02
cellular modified amino acid metabolic process (GO:0006575)	72	29	9.5	3.05	2.76E-06	1.52E-04

Table 3-15 continued.

response to unfolded protein (GO:0006986)	40	16	5.28	3.03	4.80E-04	8.84E-03
cellular response to unfolded protein (GO:0034620)	40	16	5.28	3.03	4.80E-04	8.73E-03
ribosomal large subunit biogenesis (GO:0042273)	95	37	12.53	2.95	3.22E-07	2.83E-05
aromatic amino acid family metabolic process (GO:0009072)	57	22	7.52	2.93	7.02E-05	2.01E-03
monocarboxylic acid catabolic process (GO:0072329)	37	14	4.88	2.87	1.63E-03	2.38E-02
lipid oxidation (GO:0034440)	44	16	5.8	2.76	1.68E-03	2.42E-02
cytoplasmic translation (GO:0002181)	105	37	13.85	2.67	2.17E-06	1.47E-04
secretion by cell (GO:0032940)	78	27	10.29	2.62	8.55E-05	2.32E-03
export from cell (GO:0140352)	78	27	10.29	2.62	8.55E-05	2.28E-03
secretion (GO:0046903)	78	27	10.29	2.62	8.55E-05	2.24E-03
vesicle fusion to plasma membrane (GO:0099500)	68	22	8.97	2.45	7.19E-04	1.25E-02
exocytic process (GO:0140029)	68	22	8.97	2.45	7.19E-04	1.23E-02
exocytosis (GO:0006887)	68	22	8.97	2.45	7.19E-04	1.22E-02
cellular amine metabolic process (GO:0044106)	50	16	6.59	2.43	3.47E-03	4.04E-02
cellular biogenic amine metabolic process (GO:0006576)	50	16	6.59	2.43	3.47E-03	4.01E-02
response to external biotic stimulus (GO:0043207)	136	43	17.94	2.4	3.31E-06	1.76E-04

Table 3-15 continued.

response to other organism (GO:0051707)	136	43	17.94	2.4	3.31E-06	1.70E-04
response to biotic stimulus (GO:0009607)	136	43	17.94	2.4	3.31E-06	1.65E-04
defense response to other organism (GO:0098542)	136	43	17.94	2.4	3.31E-06	1.59E-04
biological process involved in interspecies interaction between organisms (GO:0044419)	137	43	18.07	2.38	3.71E-06	1.73E-04
monocarboxylic acid metabolic process (GO:0032787)	215	64	28.36	2.26	9.58E-08	1.19E-05
carbohydrate catabolic process (GO:0016052)	121	36	15.96	2.26	8.10E-05	2.28E-03
cellular response to topologically incorrect protein (GO:0035967)	68	20	8.97	2.23	3.86E-03	4.33E-02
response to topologically incorrect protein (GO:0035966)	68	20	8.97	2.23	3.86E-03	4.30E-02
defense response (GO:0006952)	219	64	28.88	2.22	1.93E-07	2.22E-05
carboxylic acid catabolic process (GO:0046395)	93	27	12.27	2.2	8.29E-04	1.37E-02
response to endoplasmic reticulum stress (GO:0034976)	93	27	12.27	2.2	8.29E-04	1.36E-02
amine metabolic process (GO:0009308)	73	21	9.63	2.18	3.16E-03	3.84E-02
fatty acid metabolic process (GO:0006631)	78	22	10.29	2.14	2.75E-03	3.42E-02
small molecule catabolic process (GO:0044282)	174	47	22.95	2.05	5.06E-05	1.64E-03
organic acid catabolic process (GO:0016054)	105	28	13.85	2.02	1.67E-03	2.42E-02
response to external stimulus (GO:0009605)	169	45	22.29	2.02	9.64E-05	2.48E-03

Table 3-15 continued.

endoplasmic reticulum to Golgi vesicle-mediated transport (GO:0006888)	115	29	15.17	1.91	3.75E-03	4.27E-02
sulfur compound metabolic process (GO:0006790)	176	41	23.21	1.77	1.92E-03	2.60E-02
carboxylic acid metabolic process (GO:0019752)	569	132	75.05	1.76	2.69E-08	4.46E-06
carbohydrate derivative metabolic process (GO:1901135)	411	95	54.21	1.75	2.61E-06	1.49E-04
oxoacid metabolic process (GO:0043436)	572	132	75.44	1.75	4.04E-08	6.03E-06
organic acid biosynthetic process (GO:0016053)	245	56	32.31	1.73	4.01E-04	7.57E-03
protein folding (GO:0006457)	254	58	33.5	1.73	3.57E-04	6.83E-03
carboxylic acid biosynthetic process (GO:0046394)	228	52	30.07	1.73	7.02E-04	1.23E-02
ribosome biogenesis (GO:0042254)	342	78	45.11	1.73	3.25E-05	1.10E-03
cellular response to chemical stimulus (GO:0070887)	282	64	37.19	1.72	1.93E-04	4.30E-03
organic acid metabolic process (GO:0006082)	591	134	77.95	1.72	6.46E-08	8.77E-06
Golgi vesicle transport (GO:0048193)	199	45	26.25	1.71	1.80E-03	2.46E-02
carbohydrate metabolic process (GO:0005975)	410	92	54.07	1.7	1.25E-05	4.92E-04
purine-containing compound metabolic process (GO:0072521)	210	47	27.7	1.7	1.70E-03	2.39E-02
nucleobase-containing small molecule metabolic process (GO:0055086)	316	70	41.68	1.68	1.68E-04	3.97E-03
cellular response to organic substance (GO:0071310)	226	50	29.81	1.68	1.76E-03	2.43E-02

Table 3-15 continued.

purine ribonucleotide metabolic process (GO:0009150)	181	40	23.87	1.68	4.38E-03	4.81E-02
purine nucleotide metabolic process (GO:0006163)	195	43	25.72	1.67	3.21E-03	3.83E-02
transmembrane transport (GO:0055085)	306	67	40.36	1.66	3.34E-04	6.47E-03
small molecule metabolic process (GO:0044281)	1020	223	134.53	1.66	4.58E-11	6.83E-08
response to organic substance (GO:0010033)	299	63	39.43	1.6	1.26E-03	1.97E-02
response to chemical (GO:0042221)	401	84	52.89	1.59	2.03E-04	4.46E-03
cellular amide metabolic process (GO:0043603)	666	139	87.84	1.58	2.20E-06	1.43E-04
carbohydrate derivative biosynthetic process (GO:1901137)	255	53	33.63	1.58	4.14E-03	4.57E-02
cellular carbohydrate metabolic process (GO:0044262)	258	53	34.03	1.56	4.42E-03	4.81E-02
lipid metabolic process (GO:0006629)	492	100	64.89	1.54	1.35E-04	3.26E-03
cellular lipid metabolic process (GO:0044255)	406	82	53.55	1.53	6.41E-04	1.14E-02
peptide metabolic process (GO:0006518)	596	119	78.61	1.51	6.15E-05	1.95E-03
ribonucleoprotein complex biogenesis (GO:0022613)	418	83	55.13	1.51	9.64E-04	1.55E-02
small molecule biosynthetic process (GO:0044283)	356	70	46.95	1.49	3.01E-03	3.68E-02
organonitrogen compound biosynthetic process (GO:1901566)	1108	215	146.13	1.47	4.19E-07	3.47E-05
vesicle-mediated transport (GO:0016192)	516	100	68.06	1.47	6.18E-04	1.11E-02

Table 3-15 continued.

response to stress (GO:0006950)	919	177	121.21	1.46	7.01E-06	3.07E-04
nitrogen compound transport (GO:0071705)	503	96	66.34	1.45	1.30E-03	2.00E-02
organophosphate metabolic process (GO:0019637)	458	87	60.41	1.44	2.44E-03	3.14E-02
organic substance transport (GO:0071702)	551	103	72.67	1.42	1.69E-03	2.40E-02
amide biosynthetic process (GO:0043604)	573	106	75.57	1.4	1.70E-03	2.37E-02
protein localization (GO:0008104)	558	103	73.59	1.4	2.21E-03	2.95E-02
transport (GO:0006810)	1349	246	177.92	1.38	4.63E-06	2.09E-04
establishment of localization (GO:0051234)	1365	246	180.03	1.37	9.05E-06	3.75E-04
localization (GO:0051179)	1556	270	205.22	1.32	3.49E-05	1.16E-03
cellular response to stimulus (GO:0051716)	1139	195	150.22	1.3	8.41E-04	1.36E-02
response to stimulus (GO:0050896)	1619	277	213.53	1.3	6.58E-05	1.96E-03
organic substance catabolic process (GO:1901575)	1070	181	141.12	1.28	2.34E-03	3.09E-02
catabolic process (GO:0009056)	1153	191	152.07	1.26	3.79E-03	4.29E-02
organonitrogen compound metabolic process (GO:1901564)	3226	510	425.48	1.2	9.95E-05	2.52E-03
biological_process (GO:0008150)	9681	1491	1276.82	1.17	2.06E-10	1.54E-07
cellular process (GO:0009987)	8516	1283	1123.17	1.14	6.54E-07	5.14E-05

B73 X H95;Rp1-D21/+ versus wildtype

B73 x H95;*Rp1-D21/+* (BHR) F1 plants segregating 1:1 for wildtype and mutant phenotypes were also analyzed to study the effects of *Rp1-D21/+* in this hybrid's genetic background. These hybrids represent the other appropriate parental control for B73 x NC350 RIL population that was crossed to H95;*Rp1-D21/+* and used for DEG analysis and eQTL mapping (see later section and chapter 4). Three biological replicates from each phenotype were sequenced for differential gene expression analysis. Input reads ranging from 17 – 24.9M per sample were aligned to the H95-anonymized reference genome, out of which 71.3 – 75.8% were uniquely mapped (Table 3-16). This is an indication of high-quality sample preparation and data processing steps.

Table 3-16. Mapping statistics for B73 x H95;*Rp1-D21*/+ versus wildtype.

Sample	Background	Phenotype	Raw reads	Average read length (bp)	Uniquely mapped reads %	Multi-mapped reads %	Unmapped reads %
BHRwt_rep1	B73 x H95;*Rp1-D21*	wildtype	17,093,161	125	75.84%	20.03%	4.14%
BHRwt_rep2	B73 x H95;*Rp1-D21*	wildtype	22,507,024	125	75.41%	20.69%	3.90%
BHRwt_rep3	B73 x H95;*Rp1-D21*	wildtype	22,991,417	125	74.74%	22.11%	3.15%
BHRmu_rep1	B73 x H95;*Rp1-D21*	*Rp1-D21*	24,900,357	125	74.63%	22.06%	3.32%
BHRmu_rep2	B73 x H95;*Rp1-D21*	*Rp1-D21*	21,765,731	125	73.23%	23.46%	3.31%
BHRmu_rep3	B73 x H95;*Rp1-D21*	*Rp1-D21*	23,434,274	125	71.35%	24.96%	3.68%

Alignments files were processed together with reference annotation files to generate read counts per gene for each sample. These counts were higher in the wildtype samples compared to the mutants (Figure 3-17), indicating that the normal, healthy wild-type plants recorded higher expression of genes, most likely related to growth and other developmental activities, than the growth-impaired mutant plants. Even though raw read counts per gene were utilized for DGE analysis, preceding QC steps utilized normalized counts. Size-factor normalization combined with log_2-transformation had the desired effect of reducing skewness. This contributed to increased accuracy of algorithms relied on for PCA and clustering.

Biological replicates that show poor correlation of gene expression indicate likely errors at some point in the experiment. There was notable similarity among replicates, more than 97% in pairwise comparisons as measured by Pearson correlation, *r* (Table 3-17). This level of similarity was above the minimum threshold recommended by ENCODE as best practice and indicated exceptional quality of experimental procedures leading up to the generation of RNA-seq data. The pairwise correlations estimated above were used in hierarchical clustering to group similar samples. Height of clusters were computed with 1-r and the results visualized as a dendrogram. Clusters revealed the expected pattern of gene expression. Variability between wildtype and mutant phenotypes was greater than among replicates of the same phenotype (Figure 3-18).

The results of hierarchical clustering was confirmed by PCA (Figure 3-19), and showed larger gene expression variation between wild type and mutants than among replicates showing the same phenotype. In fact, 95% of variation observed across all samples could be explained by differences between mutant and wild type. Differences among replicates only accounted for about 2% of total variation (Figure 3_19). This observation is not surprising given the results obtained earlier from B73 and H95 inbred comparisons between *Rp1-D21*/+ and wildtype siblings.

As could be observed from both hierarchical clustering and PCA outliers were not detected hence three biological replicates per condition (mutant or wildtype) were analyzed using DESeq2, with wildtype set as control, to identify differentially expressed genes. A total of 5563 were identified as significantly differentially expressed between mutant and wild type. Up-regulated genes outnumbered down-regulated genes (Figure 3-20). Indeed, 3,137 (11%) expressed genes were significantly turned up as opposed to 2,426 (8.4%) that were turned down, indicating that almost 50% more genes had increased expression as part of the hypersensitive response triggered by *Rp1-D21* than were decreased.

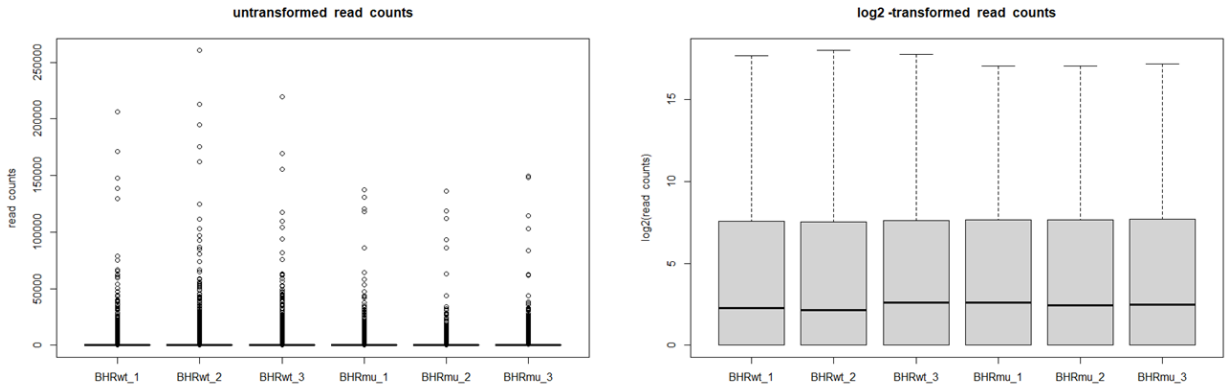


Figure 3-17. Comparison between untransformed and *log*$_2$-transformed read count distribution of B73 x H95;*Rp1-D21*/+ versus wildtype (BHR) samples showing the effect of transformation in reducing skewness.

Table 3-17. Similarity of read counts between B73 x H95;*Rp1-D21*/+ versus wildtype (BHR) replicates as measured by Pearson correlation coefficient.

Pairwise comparison	Pearson Correlation coefficient
BHRwt_1 vs. BHRwt_2	0.978
BHRwt_2 vs. BHRwt_3	0.983
BHRmu_1 vs. BHRmu_2	0.975
BHRmu_2 vs. BHRmu_3	0.976

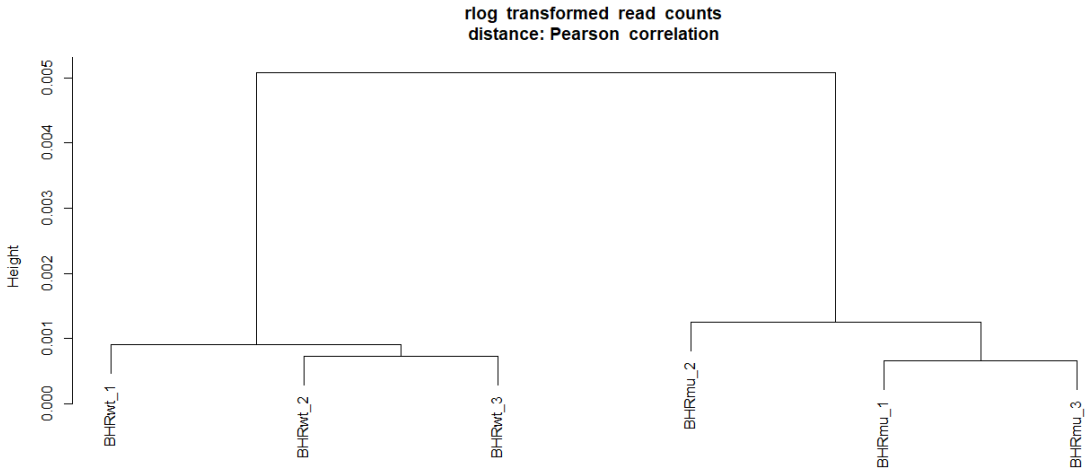


Figure 3-18. Dendrogram showing results of hierarchical clustering of B73 x H95;*Rp1-D21*/+ F1 mutants versus wildtype sibling samples. Replicates displayed greater similarity whereas the two phenotypes were clearly separated.

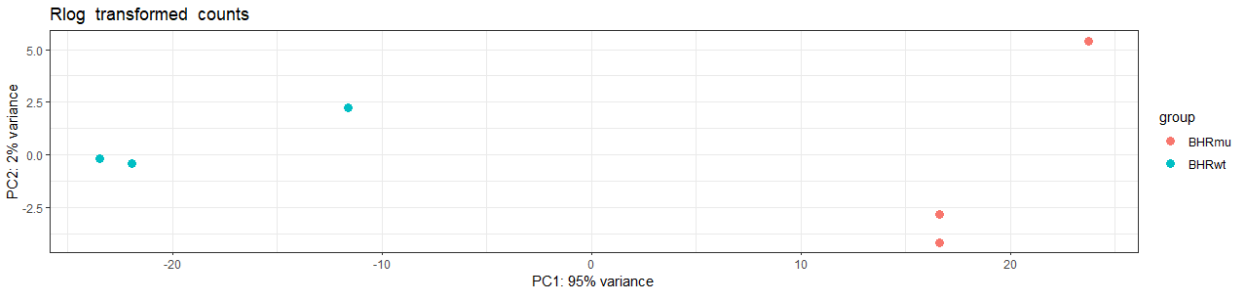


Figure 3-19. PCA on *rlog*-transformed read counts for B73 x H95;*Rp1-D21*/+ versus wildtype sibling samples. Differences between phenotypes account for greater proportion of variance.

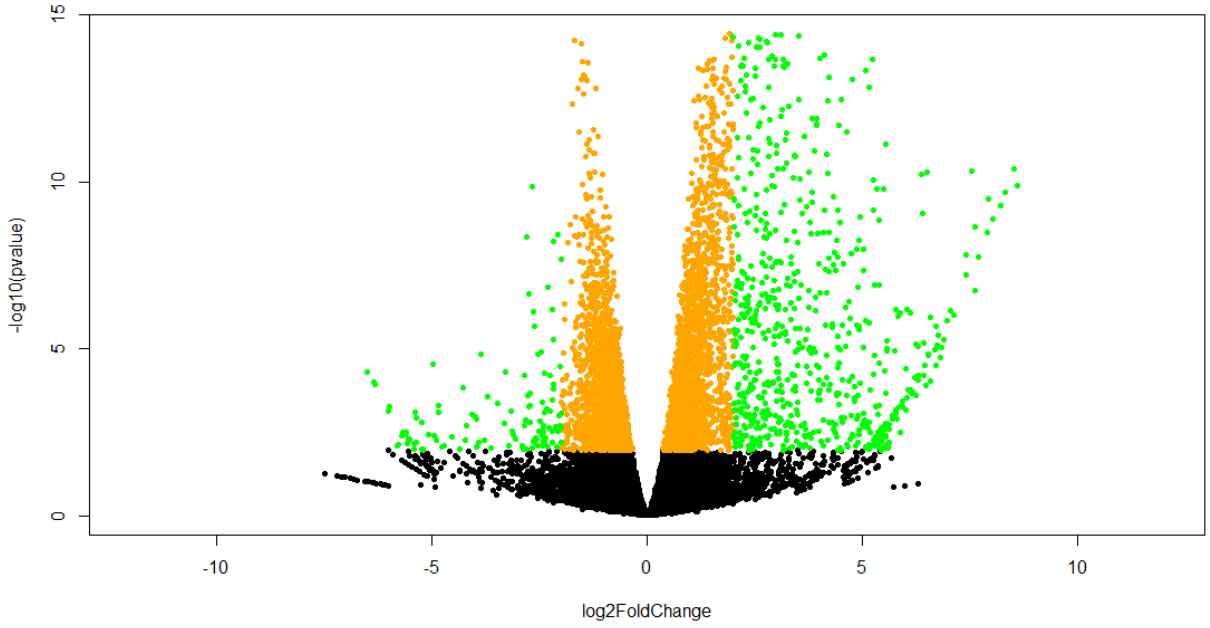


Figure 3-20. Volcano plot of B73 x H95;*Rp1-D21*/+ versus wildtype DGE results depicting statistical significance (p-value) versus magnitude of change (fold change). Black dots are genes that are not statistically significant (adjusted p-value > 0.05), orange dots are statistically significant genes (adjusted p-value < 0.05), green dots are statistically significant genes with absolute log_2 fold change of 2.

The heatmap of log_2-transformed read counts of the top 30 up or down-regulated genes (Figure 3-21) confirmed the direction of fold-change displayed in the DGE results table. Several defense or stress-related genes were among the top up-regulated genes (Table 3-18). Notably, Zm00001d047440 which encodes Alpha-humulene synthase / ZSS1, a vital enzyme for α-humulene biosynthesis (F. Yu et al., 2008) was turned up >200 times in the mutant plants than in the wildtype. Repellency of α-humulene to insects or pathogens is well known (Chang et al., 2017; Suga et al., 1993). Genes coding for C2H2-type zinc finger proteins, which play a vital role in stress resistance in plants (Han et al., 2020), were upregulated just as they were in the B73*;Rp1-D21* and H95*;Rp1-D21* backgrounds, signifying their importance in the hypersensitive response occasioned by *Rp1-D21*. Differential expression was also observed for Zm00001d005056 (WRKY DNA -binding domain). There have been reports of WRKYs facilitating salicylic acid-induced defense in tobacco (van Verk et al., 2008) and conferring drought and heat tolerance in transgenic

Arabidopsis (C. T. Wang et al., 2018). More recently, (Tang et al., 2021) suggested their involvement in response to attack by herbivore *Ostrinia furnacalis* in maize

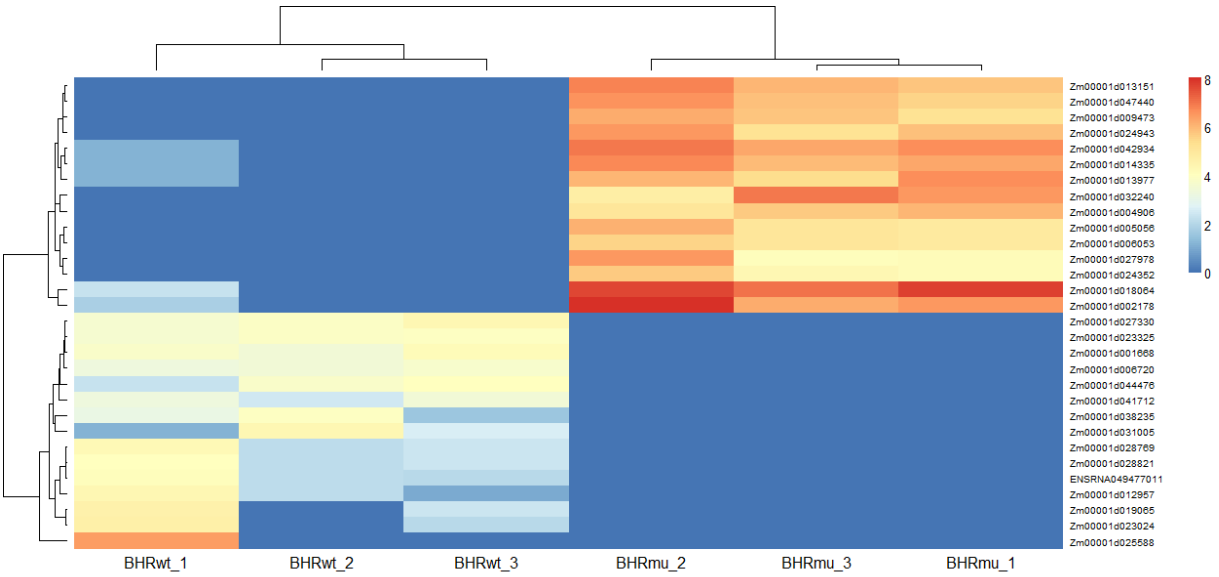


Figure 3-21. Heatmap of log_2-transformed read counts for top 30 most up or down-regulated genes after DGE analysis of B73 x H95;*Rp1-D21*/+ versus wildtype (BHR) samples. Genes are sorted based on hierarchical clustering.

Interestingly, Zm00001d038235 which encodes Superoxide dismutase (SOD) was among the top downregulated genes. The conversion of superoxide into oxygen and hydrogen peroxide─two reactive oxygen species (ROS)─within cells is controlled by SODs. Although ROS are central signaling molecules during hypersensitive response to pathogenic attacks in plants their overproduction is deleterious and can result in damage to DNA and other cellular macromolecules such as lipids and proteins (Ali et al., 2018; Das & Roychoudhury, 2014). The SOD gene is therefore turned down in the *Rp1-D21*/+ mutants to maintain ROS homeostasis. We also observe the transcription of Zm00001d019065 which codes for a homolog of Cucumisin, the first purified plant subtilisin (Kaneda & Tominaga, 1975), to be downregulated. In addition to their role in protein turnover and plant development, subtilisins have been linked to defense response to pathogenic attack in plants (Granell et al., 1987; Schaller et al., 2012) (Figueiredo et al., 2018) I hypothesize that cucumisin production is turned up initially, however, over time the propeptide

produced as by-product accumulates and serves as a negative regulator of cucumisin to turn down its further synthesis.

Table 3-18. Annotation for top 30 most up or down-regulated differentially expressed genes in B73 x H95;*Rp1-D21*/+ versus wildtype (BHR) samples.

Gene	Length	Chromosome	Description	log2FoldChange	padj
Zm00001d032240	1293	1	MYB-LIKE DNA-BINDING PROTEIN MYB // SUBFAMILY NOT NAMED	8.591554	6.73E-09
Zm00001d013151	1466	5	FAMILY NOT NAMED // NAC DOMAIN-CONTAINING PROTEIN 6-RELATED	8.528303	2.20E-09
Zm00001d047440	1190	9	Alpha-humulene synthase / ZSS1	8.307973	9.59E-09
Zm00001d024943	1900	10	Cytochrome P450 CYP2 subfamily	8.212377	2.27E-08
Zm00001d009473	3548	8	Non-specific serine/threonine protein kinase / Threonine-specific protein kinase	8.020034	5.57E-08
Zm00001d042934	1077	3	C2H2-type zinc finger (zf-C2H2_6)	7.919239	1.55E-08
Zm00001d004906	1636	2	Non-specific serine/threonine protein kinase / Threonine-specific protein kinase	7.8918	1.27E-07
Zm00001d005056	1451	2	WRKY DNA -binding domain (WRKY)	7.689155	5.92E-07
Zm00001d014335	1581	5	Cytochrome P450 CYP2 subfamily	7.624764	8.99E-08
Zm00001d027978	1164	1	Purine nucleobase transmembrane transport (PUNUT)	7.608118	4.65E-06
Zm00001d002178	4112	2	Non-specific serine/threonine protein kinase / Threonine-specific protein kinase	7.539325	2.61E-09

Table 3-18 continued.

Zm00001d006053	1814	2	FAMILY NOT NAMED // NAC DOMAIN CONTAINING PROTEIN 38	7.418086	1.79E-06
Zm00001d013977	1110	5	ANNOTATION UNKOWN	7.395456	5.00E-07
Zm00001d018064	2725	5	Laccase / Urishiol oxidase	7.336179	7.20E-14
Zm00001d024352	1509	10	Non-specific serine/threonine protein kinase / Threonine-specific protein kinase	7.114913	2.04E-05
Zm00001d028821	1206	1	F-box domain (F-box)	-5.538032	0.026542
Zm00001d012957	2431	5	FAMILY NOT NAMED // EXPANSIN-A16-RELATED	-5.586401	0.033722
Zm00001d041712	1334	3	ANNOTATION UNKOWN	-5.594714	0.018584
Zm00001d038235	1465	6	Superoxide dismutase	-5.642482	0.021125
Zm00001d028769	4363	1	ETO1-LIKE PROTEIN 2-RELATED	-5.691131	0.018692
Zm00001d031005	7665	1	E3 ubiquitin ligase involved in syntaxin degradation	-5.714519	0.022493
Zm00001d019065	3528	7	Cucumisin	-5.786646	0.036175
Zm00001d006720	798	2	ANNOTATION UNKOWN	-6.002681	0.004518
Zm00001d044476	1830	3	ATP-BINDING CASSETTE TRANSPORTER // ABC TRANSPORTER G FAMILY MEMBER 10	-6.014236	0.005772
Zm00001d023325	18680	10	LEUCINE-RICH REPEAT-CONTAINING PROTEIN // SUBFAMILY NOT NAMED	-6.33005	0.001216
Zm00001d027330	1449	1	Remorin, N-terminal region (Remorin_N)	-6.506029	0.000595
Zm00001d025588	2438	10	Pectinesterase / Pectin methylesterase	-20.99175	2.26E-06

As was noted from the GO analysis in the earlier discussed backgrounds (i.e., B73;*Rp1-D21*/+, H95;*Rp1-D21*/+, NC350 x H95;*Rp1-D21*/+), downregulated genes were principally involved in developmental processes within the plant, while upregulated genes mostly participated in defense, and with generally higher fold enrichment. Processes such as photosynthesis (GO:0015979), response to light stimulus (GO:0009416), regulation of multicellular organismal development (GO:2000026), cytokinesis (GO:0061640) and membrane fission (GO:0090148) were overrepresented 3.44 – 7.9 times more than would have been expected by random chance (Table 3-19). On the other hand, defense related processes including defense response to fungus (GO:0050832), immune response (GO:0006955), programmed cell death (GO:0012501), defense response to other organism (GO:0098542) were overrepresented 4.78 – 9.63 times in the list of all upregulated genes (Table 3-20).

Table 3-19. GO annotations for down-regulated genes from the B73 x H95;*Rp1-D21*/+ versus wildtype background.

GO-Slim Biological Process	Zea mays - REFLIST	Actual	Expected	Fold Enrichment	Raw P-value	FDR
photosynthesis (GO:0015979)	52	25	3.17	7.9	5.21E-13	7.77E-10
beta-glucan biosynthetic process (GO:0051274)	23	7	1.4	5	1.28E-03	3.25E-02
response to light stimulus (GO:0009416)	108	31	6.58	4.71	5.57E-11	4.16E-08
regulation of multicellular organismal development (GO:2000026)	29	8	1.77	4.53	1.00E-03	2.83E-02
mitotic cytokinesis (GO:0000281)	35	9	2.13	4.22	7.56E-04	2.26E-02
cytoskeleton-dependent cytokinesis (GO:0061640)	35	9	2.13	4.22	7.56E-04	2.21E-02
response to radiation (GO:0009314)	123	31	7.49	4.14	8.41E-10	4.18E-07

Table 3-19 continued.

hexose biosynthetic process (GO:0019319)	32	8	1.95	4.11	1.72E-03	3.77E-02
membrane fission (GO:0090148)	43	9	2.62	3.44	2.60E-03	4.97E-02
cytokinesis (GO:0000910)	43	9	2.62	3.44	2.60E-03	4.91E-02
amine metabolic process (GO:0009308)	73	15	4.44	3.38	1.37E-04	6.84E-03
cellular amine metabolic process (GO:0044106)	50	10	3.04	3.29	2.06E-03	4.26E-02
cellular biogenic amine metabolic process (GO:0006576)	50	10	3.04	3.29	2.06E-03	4.21E-02
response to abiotic stimulus (GO:0009628)	189	36	11.51	3.13	2.50E-08	7.45E-06
electron transport chain (GO:0022900)	66	12	4.02	2.99	1.58E-03	3.68E-02
generation of precursor metabolites and energy (GO:0006091)	233	38	14.19	2.68	4.61E-07	9.82E-05
hexose metabolic process (GO:0019318)	86	14	5.24	2.67	1.71E-03	3.81E-02
regulation of response to stimulus (GO:0048583)	90	14	5.48	2.56	2.49E-03	4.83E-02
small molecule biosynthetic process (GO:0044283)	356	40	21.67	1.85	6.61E-04	2.10E-02
carbohydrate metabolic process (GO:0005975)	410	42	24.96	1.68	2.16E-03	4.29E-02
small molecule metabolic process (GO:0044281)	1020	93	62.1	1.5	3.11E-04	1.29E-02

Table 3-20. GO annotations for up-regulated genes from the B73 x H95;*Rp1-D21*/+ versus wildtype background.

GO-Slim Biological Process	Zea mays - REFLIST	Actual	Expected	Fold Enrichment	Raw P-value	FDR
protein N-linked glycosylation via asparagine (GO:0018279)	6	5	0.47	10.7	6.26E-04	6.72E-03
defense response to fungus (GO:0050832)	8	6	0.62	9.63	2.56E-04	3.21E-03
innate immune response (GO:0045087)	11	7	0.86	8.17	1.59E-04	2.16E-03
immune system process (GO:0002376)	14	7	1.09	6.42	4.80E-04	5.38E-03
immune response (GO:0006955)	14	7	1.09	6.42	4.80E-04	5.34E-03
maturation of LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA) (GO:0000463)	25	12	1.95	6.16	6.64E-06	1.32E-04
ribosomal large subunit assembly (GO:0000027)	32	15	2.49	6.02	6.00E-07	1.72E-05
programmed cell death (GO:0012501)	18	8	1.4	5.71	3.52E-04	4.27E-03
endonucleolytic cleavage of tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA) (GO:0000479)	18	8	1.4	5.71	3.52E-04	4.24E-03
endonucleolytic cleavage involved in rRNA processing (GO:0000478)	18	8	1.4	5.71	3.52E-04	4.20E-03
cell death (GO:0008219)	18	8	1.4	5.71	3.52E-04	4.17E-03
maturation of LSU-rRNA (GO:0000470)	32	14	2.49	5.62	2.69E-06	5.66E-05
ribosomal large subunit biogenesis (GO:0042273)	95	40	7.4	5.41	6.01E-15	4.72E-13
establishment of protein localization to extracellular region (GO:0035592)	12	5	0.93	5.35	5.81E-03	4.66E-02
cell surface receptor signaling pathway (GO:0007166)	60	25	4.67	5.35	8.19E-10	4.07E-08

Table 3-20 continued.

protein secretion (GO:0009306)	12	5	0.93	5.35	5.81E-03	4.64E-02
protein localization to extracellular region (GO:0071692)	12	5	0.93	5.35	5.81E-03	4.61E-02
retrograde transport, endosome to Golgi (GO:0042147)	22	9	1.71	5.25	2.47E-04	3.18E-03
cytosolic transport (GO:0016482)	25	10	1.95	5.14	1.31E-04	1.84E-03
protein glycosylation (GO:0006486)	42	16	3.27	4.89	2.27E-06	5.14E-05
macromolecule glycosylation (GO:0043413)	42	16	3.27	4.89	2.27E-06	5.06E-05
glycosylation (GO:0070085)	42	16	3.27	4.89	2.27E-06	4.99E-05
cytoplasmic translation (GO:0002181)	105	40	8.18	4.89	8.36E-14	5.94E-12
intra-Golgi vesicle-mediated transport (GO:0006891)	53	20	4.13	4.85	1.45E-07	4.80E-06
response to external biotic stimulus (GO:0043207)	136	51	10.59	4.82	5.80E-17	7.86E-15
response to other organism (GO:0051707)	136	51	10.59	4.82	5.80E-17	7.21E-15
response to biotic stimulus (GO:0009607)	136	51	10.59	4.82	5.80E-17	6.65E-15
defense response to other organism (GO:0098542)	136	51	10.59	4.82	5.80E-17	6.18E-15
biological process involved in interspecies interaction between organisms (GO:0044419)	137	51	10.67	4.78	7.40E-17	7.36E-15
protein N-linked glycosylation (GO:0006487)	30	11	2.34	4.71	1.14E-04	1.62E-03
glycoprotein biosynthetic process (GO:0009101)	44	16	3.43	4.67	3.70E-06	7.57E-05
endoplasmic reticulum unfolded protein response (GO:0030968)	22	8	1.71	4.67	1.02E-03	1.05E-02

Table 3-20 continued.

glycoprotein metabolic process (GO:0009100)	59	21	4.59	4.57	1.59E-07	5.15E-06
retrograde vesicle-mediated transport, Golgi to endoplasmic reticulum (GO:0006890)	47	16	3.66	4.37	7.37E-06	1.43E-04
COPII-coated vesicle budding (GO:0090114)	27	9	2.1	4.28	8.31E-04	8.86E-03
tricarboxylic acid cycle (GO:0006099)	34	11	2.65	4.16	2.80E-04	3.45E-03
response to external stimulus (GO:0009605)	169	51	13.16	3.88	7.31E-14	5.46E-12
defense response (GO:0006952)	219	65	17.05	3.81	5.79E-17	8.65E-15
secretion by cell (GO:0032940)	78	23	6.07	3.79	6.70E-07	1.88E-05
export from cell (GO:0140352)	78	23	6.07	3.79	6.70E-07	1.85E-05
secretion (GO:0046903)	78	23	6.07	3.79	6.70E-07	1.82E-05
cleavage involved in rRNA processing (GO:0000469)	38	11	2.96	3.72	6.14E-04	6.64E-03
vesicle fusion to plasma membrane (GO:0099500)	68	19	5.29	3.59	1.16E-05	2.06E-04
exocytic process (GO:0140029)	68	19	5.29	3.59	1.16E-05	2.04E-04
exocytosis (GO:0006887)	68	19	5.29	3.59	1.16E-05	2.02E-04
xylan biosynthetic process (GO:0045492)	37	10	2.88	3.47	1.66E-03	1.55E-02
ribosome biogenesis (GO:0042254)	342	91	26.63	3.42	1.76E-20	6.58E-18
ribosomal small subunit biogenesis (GO:0042274)	129	34	10.04	3.38	1.74E-08	7.63E-07

Table 3-20 continued.

endoplasmic reticulum to Golgi vesicle-mediated transport (GO:0006888)	115	30	8.95	3.35	1.42E-07	4.80E-06
RNA phosphodiester bond hydrolysis, endonucleolytic (GO:0090502)	31	8	2.41	3.31	5.99E-03	4.73E-02
Golgi vesicle transport (GO:0048193)	199	51	15.5	3.29	1.29E-11	6.86E-10
response to unfolded protein (GO:0006986)	40	10	3.11	3.21	2.71E-03	2.38E-02
cellular response to unfolded protein (GO:0034620)	40	10	3.11	3.21	2.71E-03	2.36E-02
aromatic amino acid family metabolic process (GO:0009072)	57	14	4.44	3.15	4.89E-04	5.41E-03
receptor-mediated endocytosis (GO:0006898)	49	12	3.82	3.15	1.23E-03	1.22E-02
lipid transport (GO:0006869)	37	9	2.88	3.12	5.07E-03	4.18E-02
response to endoplasmic reticulum stress (GO:0034976)	93	22	7.24	3.04	2.31E-05	3.78E-04
glutathione metabolic process (GO:0006749)	51	12	3.97	3.02	1.65E-03	1.56E-02
non-membrane-bounded organelle assembly (GO:0140694)	137	32	10.67	3	4.64E-07	1.38E-05
maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA) (GO:0000462)	74	17	5.76	2.95	2.53E-04	3.20E-03
translational initiation (GO:0006413)	57	13	4.44	2.93	1.38E-03	1.32E-02
ribonucleoprotein complex biogenesis (GO:0022613)	418	95	32.55	2.92	2.01E-17	4.27E-15
ribonucleoprotein complex assembly (GO:0022618)	152	34	11.84	2.87	4.89E-07	1.43E-05
maturation of SSU-rRNA (GO:0030490)	92	20	7.16	2.79	1.44E-04	2.01E-03

Table 3-20 continued.

ribonucleoprotein complex subunit organization (GO:0071826)	157	34	12.22	2.78	1.51E-06	3.68E-05
RNA phosphodiester bond hydrolysis (GO:0090501)	51	11	3.97	2.77	4.52E-03	3.79E-02
cellular response to lipid (GO:0071396)	76	16	5.92	2.7	8.64E-04	9.08E-03
cellular modified amino acid metabolic process (GO:0006575)	72	15	5.61	2.68	1.35E-03	1.31E-02
rRNA processing (GO:0006364)	229	47	17.83	2.64	5.74E-08	2.20E-06
peptide metabolic process (GO:0006518)	596	122	46.41	2.63	8.71E-19	2.17E-16
cellular amide metabolic process (GO:0043603)	666	135	51.86	2.6	2.55E-20	7.60E-18
cellular response to oxygen-containing compound (GO:1901701)	99	20	7.71	2.59	4.66E-04	5.30E-03
translational elongation (GO:0006414)	518	104	40.33	2.58	9.89E-16	9.22E-14
translation (GO:0006412)	518	104	40.33	2.58	9.89E-16	8.68E-14
amide biosynthetic process (GO:0043604)	573	115	44.62	2.58	3.79E-17	6.28E-15
peptide biosynthetic process (GO:0043043)	519	104	40.41	2.57	1.09E-15	9.03E-14
rRNA metabolic process (GO:0016072)	236	47	18.38	2.56	9.12E-08	3.17E-06
nucleic acid phosphodiester bond hydrolysis (GO:0090305)	73	14	5.68	2.46	5.18E-03	4.23E-02
organelle assembly (GO:0070925)	178	34	13.86	2.45	1.21E-05	2.07E-04
vesicle fusion (GO:0006906)	110	21	8.57	2.45	5.18E-04	5.68E-03

Table 3-20 continued.

response to lipid (GO:0033993)	84	16	6.54	2.45	2.69E-03	2.37E-02
protein autophosphorylation (GO:0046777)	90	17	7.01	2.43	2.03E-03	1.83E-02
vesicle-mediated transport (GO:0016192)	516	97	40.18	2.41	4.19E-13	2.84E-11
response to nitrogen compound (GO:1901698)	82	15	6.38	2.35	4.73E-03	3.92E-02
organelle membrane fusion (GO:0090174)	115	21	8.95	2.35	1.20E-03	1.20E-02
organelle fusion (GO:0048284)	116	21	9.03	2.32	1.25E-03	1.22E-02
vesicle organization (GO:0016050)	256	46	19.93	2.31	1.68E-06	4.04E-05
membrane fusion (GO:0061025)	118	21	9.19	2.29	1.37E-03	1.32E-02
hormone-mediated signaling pathway (GO:0009755)	154	27	11.99	2.25	4.08E-04	4.80E-03
cellular response to hormone stimulus (GO:0032870)	154	27	11.99	2.25	4.08E-04	4.76E-03
aerobic respiration (GO:0009060)	103	18	8.02	2.24	3.88E-03	3.32E-02
vesicle budding from membrane (GO:0006900)	145	25	11.29	2.21	8.36E-04	8.85E-03
cellular response to endogenous stimulus (GO:0071495)	157	27	12.22	2.21	4.65E-04	5.34E-03
response to organic substance (GO:0010033)	299	51	23.28	2.19	1.98E-06	4.62E-05
cellular respiration (GO:0045333)	106	18	8.25	2.18	4.49E-03	3.79E-02
protein folding (GO:0006457)	254	43	19.78	2.17	1.36E-05	2.30E-04

Table 3-20 continued.

ncRNA metabolic process (GO:0034660)	391	66	30.44	2.17	8.23E-08	2.92E-06
cellular response to organic substance (GO:0071310)	226	38	17.6	2.16	6.03E-05	9.09E-04
energy derivation by oxidation of organic compounds (GO:0015980)	110	18	8.57	2.1	5.70E-03	4.59E-02
cellular response to chemical stimulus (GO:0070887)	282	46	21.96	2.09	1.47E-05	2.46E-04
organonitrogen compound biosynthetic process (GO:1901566)	1108	179	86.27	2.07	2.19E-17	4.09E-15
ncRNA processing (GO:0034470)	323	52	25.15	2.07	7.48E-06	1.43E-04
localization within membrane (GO:0051668)	169	27	13.16	2.05	1.23E-03	1.22E-02
response to hormone (GO:0009725)	177	28	13.78	2.03	1.04E-03	1.07E-02
response to endogenous stimulus (GO:0009719)	180	28	14.02	2	1.70E-03	1.57E-02
protein localization to membrane (GO:0072657)	148	23	11.52	2	4.23E-03	3.59E-02
signal transduction (GO:0007165)	627	95	48.82	1.95	1.61E-08	7.29E-07
protein transport (GO:0015031)	397	60	30.91	1.94	7.04E-06	1.38E-04
establishment of protein localization (GO:0045184)	407	61	31.69	1.92	8.53E-06	1.61E-04
signaling (GO:0023052)	634	95	49.37	1.92	2.87E-08	1.22E-06
response to chemical (GO:0042221)	401	60	31.22	1.92	1.11E-05	1.99E-04
protein localization (GO:0008104)	558	82	43.45	1.89	4.48E-07	1.36E-05

Table 3-20 continued.

cell communication (GO:0007154)	650	95	50.61	1.88	6.67E-08	2.43E-06
membrane organization (GO:0061024)	397	58	30.91	1.88	2.44E-05	3.96E-04
carbohydrate derivative metabolic process (GO:1901135)	411	60	32	1.87	2.11E-05	3.50E-04
carbohydrate derivative biosynthetic process (GO:1901137)	255	37	19.86	1.86	8.85E-04	9.23E-03
macromolecule localization (GO:0033036)	656	94	51.08	1.84	2.28E-07	7.23E-06
organic substance transport (GO:0071702)	551	78	42.9	1.82	3.79E-06	7.63E-05
cellular macromolecule localization (GO:0070727)	517	73	40.26	1.81	8.69E-06	1.62E-04
nitrogen compound transport (GO:0071705)	503	71	39.17	1.81	9.81E-06	1.81E-04
cellular protein localization (GO:0034613)	503	71	39.17	1.81	9.81E-06	1.78E-04
intracellular protein transport (GO:0006886)	360	50	28.03	1.78	3.42E-04	4.18E-03
response to stress (GO:0006950)	919	127	71.56	1.77	9.91E-09	4.62E-07
transport (GO:0006810)	1349	185	105.04	1.76	5.74E-12	3.43E-10
establishment of localization (GO:0051234)	1365	186	106.28	1.75	7.67E-12	4.40E-10
protein phosphorylation (GO:0006468)	493	67	38.39	1.75	5.03E-05	7.81E-04
response to stimulus (GO:0050896)	1619	216	126.06	1.71	8.86E-13	5.75E-11
phosphorylation (GO:0016310)	579	77	45.08	1.71	2.63E-05	4.18E-04

Table 3-20 continued.

intracellular transport (GO:0046907)	597	79	46.49	1.7	2.46E-05	3.95E-04
cellular localization (GO:0051641)	802	106	62.45	1.7	1.03E-06	2.70E-05
establishment of localization in cell (GO:0051649)	607	80	47.26	1.69	2.78E-05	4.37E-04
localization (GO:0051179)	1556	205	121.16	1.69	9.66E-12	5.34E-10
cellular response to stimulus (GO:0051716)	1139	147	88.69	1.66	3.70E-08	1.49E-06
carboxylic acid metabolic process (GO:0019752)	569	73	44.3	1.65	1.51E-04	2.09E-03
oxoacid metabolic process (GO:0043436)	572	73	44.54	1.64	1.57E-04	2.16E-03
transmembrane transport (GO:0055085)	306	39	23.83	1.64	6.23E-03	4.89E-02
cellular component biogenesis (GO:0044085)	1059	133	82.46	1.61	6.86E-07	1.83E-05
organic acid metabolic process (GO:0006082)	591	74	46.02	1.61	2.67E-04	3.32E-03
small molecule metabolic process (GO:0044281)	1020	126	79.42	1.59	2.53E-06	5.39E-05
phosphorus metabolic process (GO:0006793)	1146	140	89.23	1.57	1.32E-06	3.33E-05
phosphate-containing compound metabolic process (GO:0006796)	1126	137	87.68	1.56	1.90E-06	4.49E-05
protein metabolic process (GO:0019538)	2503	289	194.89	1.48	3.77E-10	1.94E-08
cellular protein metabolic process (GO:0044267)	2407	277	187.42	1.48	1.28E-09	6.16E-08
organonitrogen compound metabolic process (GO:1901564)	3226	369	251.19	1.47	2.65E-12	1.65E-10

Table 3-20 continued.

cellular component organization or biogenesis (GO:0071840)	2104	235	163.83	1.43	2.58E-07	8.03E-06
biological_process (GO:0008150)	9681	1048	753.81	1.39	3.60E-30	2.69E-27
cellular process (GO:0009987)	8516	906	663.09	1.37	7.25E-23	3.61E-20
cellular biosynthetic process (GO:0044249)	2792	291	217.4	1.34	2.12E-06	4.87E-05
biosynthetic process (GO:0009058)	2884	300	224.56	1.34	1.50E-06	3.72E-05
organic substance biosynthetic process (GO:1901576)	2817	293	219.34	1.34	2.33E-06	5.04E-05
macromolecule biosynthetic process (GO:0009059)	2029	211	157.99	1.34	6.65E-05	9.63E-04
cellular macromolecule biosynthetic process (GO:0034645)	2005	206	156.12	1.32	1.69E-04	2.25E-03
cellular nitrogen compound biosynthetic process (GO:0044271)	2072	208	161.33	1.29	4.71E-04	5.33E-03
cellular nitrogen compound metabolic process (GO:0034641)	3262	318	253.99	1.25	1.03E-04	1.47E-03
organic substance metabolic process (GO:0071704)	6357	615	494.98	1.24	3.86E-08	1.52E-06
cellular metabolic process (GO:0044237)	6194	599	482.29	1.24	6.62E-08	2.47E-06
metabolic process (GO:0008152)	6623	638	515.7	1.24	3.62E-08	1.50E-06
nitrogen compound metabolic process (GO:0006807)	5331	509	415.1	1.23	3.64E-06	7.55E-05
primary metabolic process (GO:0044238)	5922	564	461.11	1.22	1.23E-06	3.16E-05
gene expression (GO:0010467)	2415	228	188.04	1.21	5.17E-03	4.24E-02

Table 3-20 continued.

cellular macromolecule metabolic process (GO:0044260)	3980	364	309.9	1.17	2.42E-03	2.15E-02
macromolecule metabolic process (GO:0043170)	4911	443	382.39	1.16	1.72E-03	1.58E-02

B73:NC350RIL x H95;Rp1-D21/+ versus wildtype (BNRIL_HR)

Greater biological replicates permits the robust detection of differentially expressed genes at all fold changes (Pan et al., 2002; Schurch et al., 2016). To further explore the consequences of *Rp1-D21*/+ on gene expression, we performed DEG analysis using F1 families produced by crossing 99 members of the B73 x NC350 RIL subpopulation and *Rp1-D21*/+;H95. Because *Rp1-D21*/+;H95is maintained in a heterozygous state for the *Rp1-D21* allele the F1 offspring segregate 1:1 ratio for those carrying the *Rp1-D21* allele (mutant F1) and those carrying the wildtype H95 allele at the *Rp1* locus (wildtype F1). The 99 F1 families showing the *Rp1-D21* phenotype were treated as "biological replicates" and analyzed against those showing the wild-type phenotype for differential expression.

As observed in the genetic backgrounds discussed earlier, read counts per gene were higher in the wildtype group than in the mutants (Figure 3-22). Hierarchical clustering showed a clear separation between the mutant and wildtype classes (Figure 3-23), meaning global gene expression patterns were markedly different between the two groups. This was reinforced by principal component analysis (Figure 3-24). PC1 was clearly the *Rp1-D21*/+ effect and separated mutants and wildtype groups, albeit with a single wild-type outlier sample. Since this sample was only one out of 99 samples in that group its impact on overall DE analysis was expected to be minimal and was thus not excluded from further analysis. PC2, which represented differences among the F1 families only accounted for about 1% of the total variation across samples. DE analysis identified a remarkable 23,911 genes to be differentially expressed between mutant and wildtype. With 99 "biological replicates" our statistical analysis had enough power to detect genes with even subtle fold changes between the two groups hence the large number of DEGs. The volcano plot visualizing fold changes depicted a huge skew (Figure 3-25). Upregulated genes with a log_2 fold

change of at least 2 totaled 2,872, outnumbering downregulated genes at the same threshold (573 in sum) more than five times. Indicating that genes that saw the biggest difference in expression between wildtype and mutants were mostly those that were induced in the mutants in response to *Rp1-D21*/+.

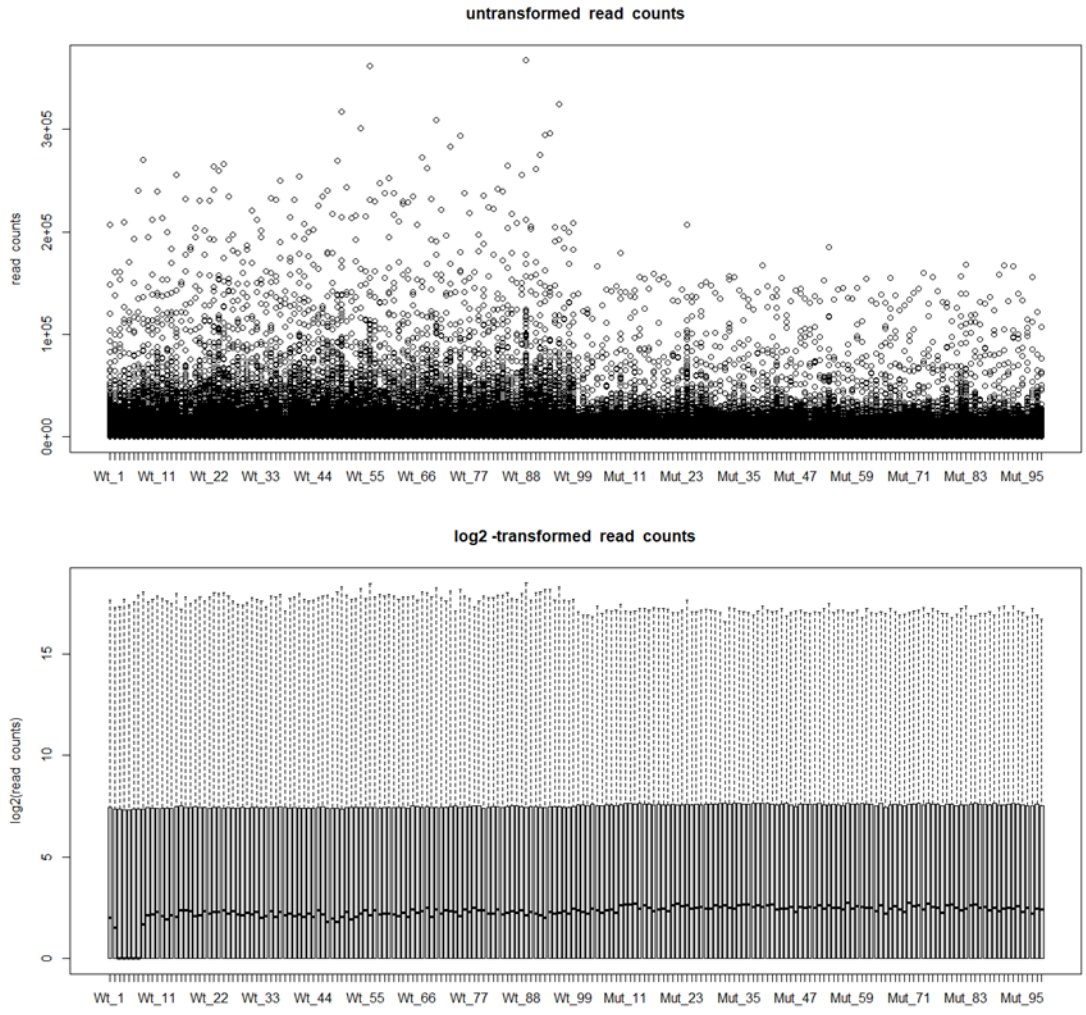


Figure 3-22. Comparison between untransformed and *log*$_2$-transformed read count distribution of B73:NC350RIL x H95;*Rp1-D21*/+ versus wildtype (BNRIL_HR) samples showing the effect of transformation in reducing skewness.

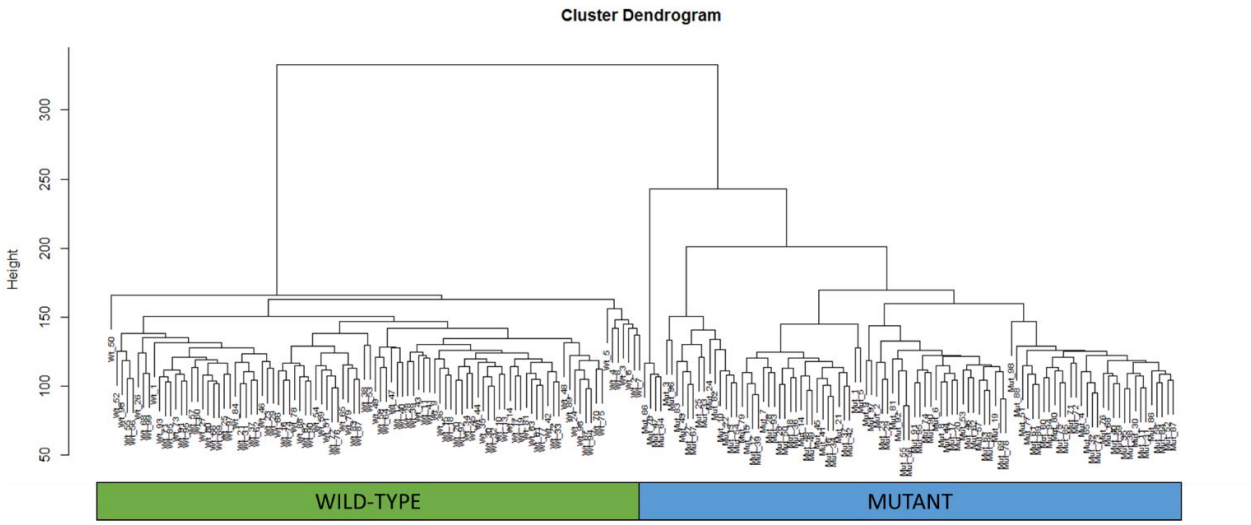


Figure 3-23. Dendrogram showing results of hierarchical clustering of B73:NC350RIL x H95;*Rp1-D21*/+ versus wildtype (BNRIL_HR). Each leaf of the tree is a single RIL hybrid either WT at the *Rp1* locus or carrying *Rp1-D21* as a heterozygote. The entire set of WT RIL hybrids grouped tother (green bar) and the set of *Rp1-D21*/+ RIL hybrids group together (blue bar).

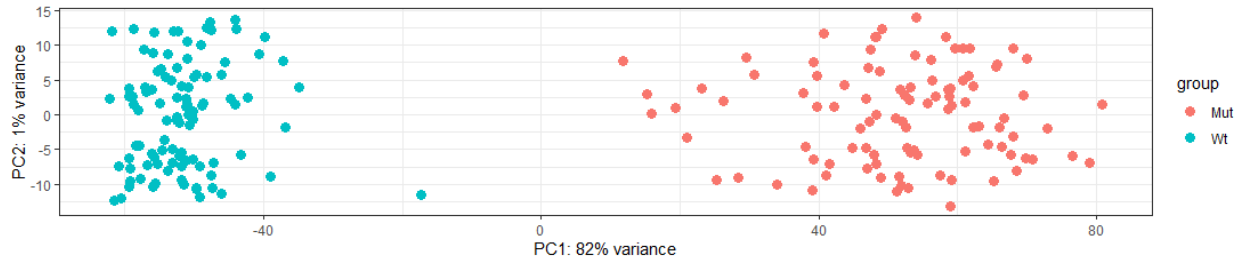


Figure 3-24. PCA on *rlog*-transformed read counts for B73:NC350RIL x H95;*Rp1-D21*/+ versus wildtype (BNRIL_HR) samples. Differences between phenotypes account for greater proportion of variance.

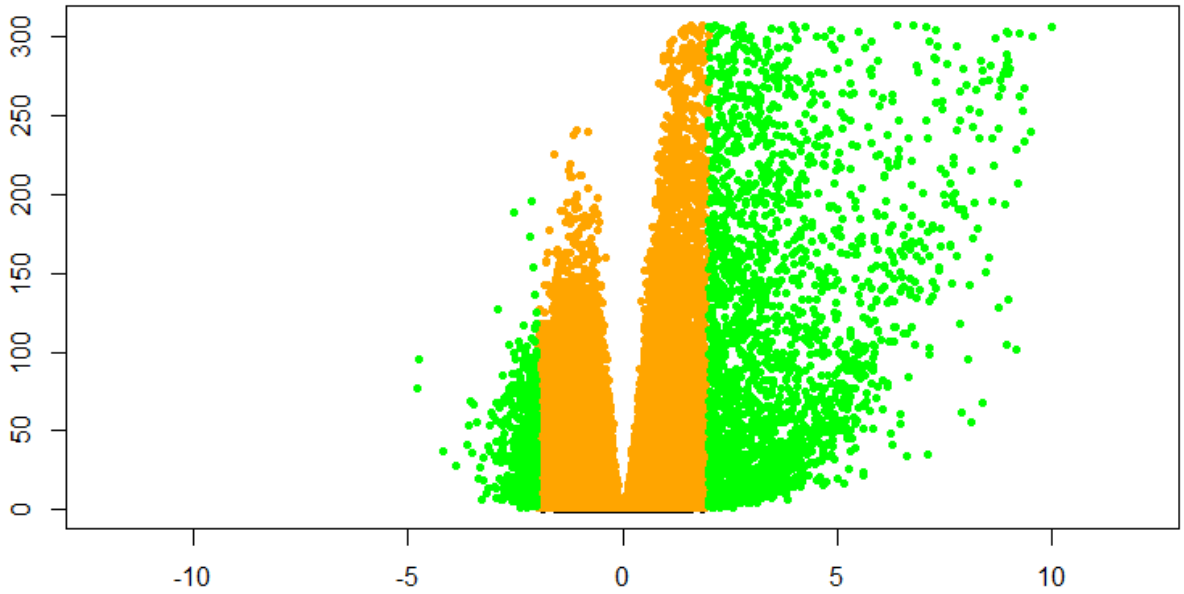


Figure 3-25. Volcano plot of B73:NC350RIL x H95;*Rp1-D21*/+ versus wildtype (BNRIL_HR) DGE results depicting statistical significance (p-value) versus magnitude of change (fold change). Black dots are genes that are not statistically significant (adjusted p-value > 0.05), orange dots are statistically significant genes (adjusted p-value < 0.05), green dots are statistically significant genes with absolute log_2 fold change of 2.

Fold-change direction observed on the heatmap of the log_2-transformed read counts for the top 30 most up or downregulated genes (Figure 3-26) matched up what was displayed in the results table (Table 3-21). Furthermore, heatmap showed foldchange differences between wildtype and mutant classes to be more extreme for upregulated genes compared to downregulated genes, indicating that when genes were turned up in response to *Rp1* the magnitude of expression was much greater than when they were turned down. Several upregulated genes identified in earlier-discussed backgrounds were found in this genetic background. Among these were genes encoding omega-6 fatty acid desaturase (delta-12 desaturase, FAD2), Pathogenesis-related protein Bet v I family, Thaumatin family and (S)-beta-bisabolene synthase. FAD2 functions in the production of linoleic acids from oleic acid which are in turn utilized to signal plant defense machinery in response to pathogenic attack and wounding (Dar et al., 2017; Farmer, 1994; Mikkilineni & Rocheford, 2003). Pathogenesis-related protein Bet v I also plays a role in defense. It is a key component of the host response elicited upon pathogen invasion and other abiotic stresses

(Radauer et al., 2008). The Thaumatin gene family, another group of PR genes, has been reported to confer tolerance to several fungal species and abiotic stresses such as salinity and dehydration (Z. Li et al., 2020b; Misra et al., 2016; van Loon et al., 2006). Beta-bisabolene synthase, which was also detected previously to be upregulated in the mutants, codes for a maize terpene beta-bisabolene. This secondary metabolite is known to function in chemical defense against insect herbivores (Block et al., 2019; Köllner et al., 2008; Meihls et al., 2012)

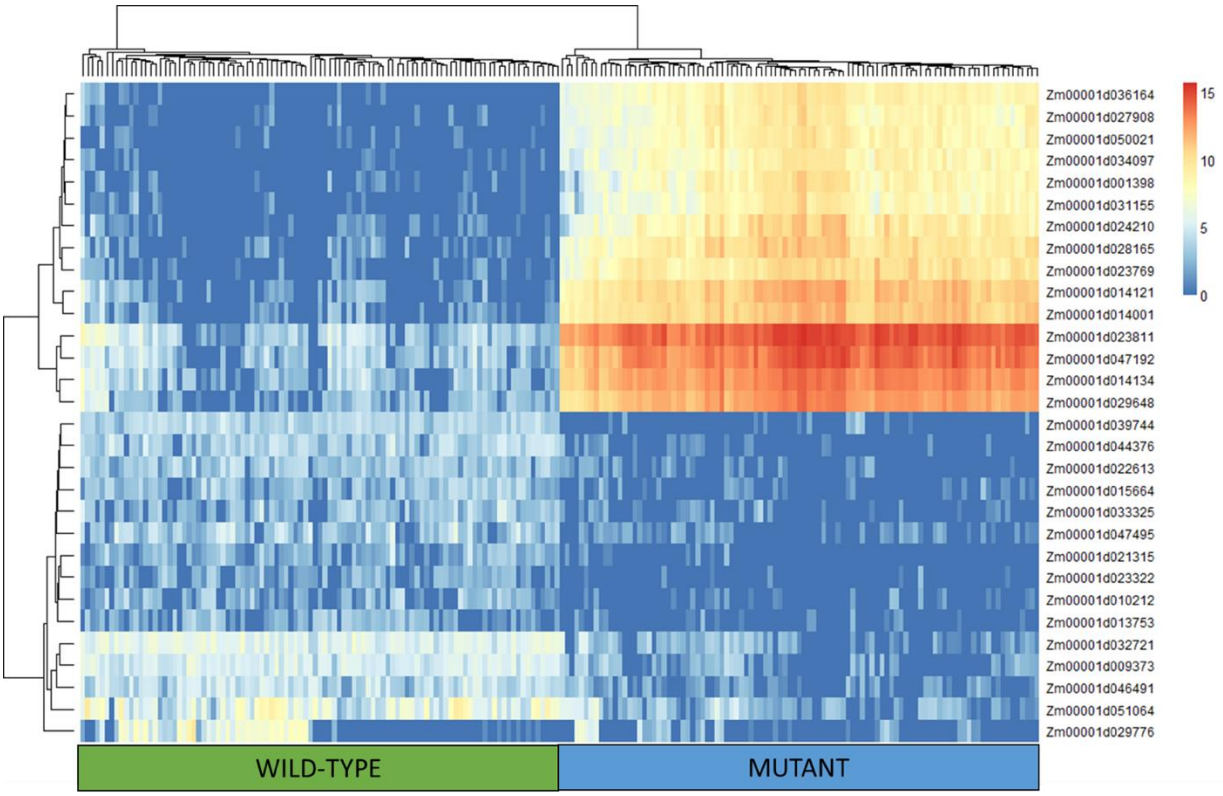


Figure 3-26. Heatmap of *log*$_2$-transformed read counts for top 30 most up or down-regulated genes after DGE analysis of B73:NC350RIL x H95;*Rp1-D21*/+ versus wildtype (BNRIL_HR) samples. Genes are sorted based on hierarchical clustering.

Table 3-21. Annotation for top 30 most up or down-regulated differentially expressed genes in B73:NC350RIL x H95;*Rp1-D21*/+ versus wildtype (BNRIL_HR) samples.

Gene	Length	Chromosome	Description	log2FoldChange	padj
Zm00001d034097	1699	1	Isoflavone 2'-hydroxylase / Isoflavone 2'-monooxygenase	10.06976	0
Zm00001d050021	5574	4	(+)-abscisic acid 8'-hydroxylase / ABA 8'-hydroxylase	10.01578	7.04E-305
Zm00001d047192	1630	9	O-METHYLTRANSFERASE	9.980479	0
Zm00001d023769	1661	10	omega-6 fatty acid desaturase (delta-12 desaturase) (FAD2)	9.932205	0
Zm00001d023811	1063	10	Pathogenesis-related protein Bet v I family (Bet_v_1)	9.761622	0
Zm00001d036164	1576	6	Theobromine synthase / MXMT	9.545794	4.34E-299
Zm00001d027908	3019	1	Flavin-containing monooxygenase / Ziegler's enzyme	9.35757	3.94E-266
Zm00001d031155	534	1	Thaumatin family (Thaumatin)	9.355874	1.34E-232
Zm00001d024210	3192	10	(S)-beta-bisabolene synthase // (S)-beta-macrocarpene synthase	9.329124	3.94E-252
Zm00001d014134	1803	5	Ent-isokaurene C2-hydroxylase / CYP71Z6	9.322145	0
Zm00001d014001	5208	5	flotillin (FLOT)	9.311704	0
Zm00001d028165	4513	1	LL-diaminopimelate aminotransferase / LL-diaminopimelate transaminase	9.255103	1.28E-300
Zm00001d029648	3816	1	Ent-copalyl diphosphate synthase / Ent-kaurene synthetase A	9.248359	0
Zm00001d014121	1940	5	CYTOCHROME P450 76C1	9.233724	0
Zm00001d013753	2036	5	Tyrosine N-monooxygenase / Tyrosine N-hydroxylase	-3.245896	7.41E-19
Zm00001d047495	2677	9	Premnaspirodiene oxygenase / Hyoscymus muticus premnaspirodiene oxygenase	-3.257003	1.61E-33

Table 3-21 continued

Zm00001d029776	2751	1	Carboxypeptidase D / Carboxypeptidase S1	-3.287136	1.62E-07
Zm00001d033325	945	1	Dof domain, zinc finger (zf-Dof)	-3.341792	5.34E-27
Zm00001d010212	6601	8	Beta-galactosidase / Lactase	-3.375495	5.00E-20
Zm00001d046491	1934	9	2-alkenal reductase (NAD(P)(+)) / NADPH:2-alkenal alpha,beta-hydrogenase	-3.395784	1.12E-55
Zm00001d009373	2495	8	Peroxidase / Lactoperoxidase	-3.475656	5.82E-67
Zm00001d015664	507	5	CAMP-RESPONSE ELEMENT BINDING PROTEIN-RELATED	-3.520202	3.98E-36
Zm00001d032721	5431	1	Long-chain-alcohol oxidase	-3.563823	7.52E-69
Zm00001d022613	1148	7	CAMP-RESPONSE ELEMENT BINDING PROTEIN-RELATED	-3.580436	3.41E-53
Zm00001d051064	744	4	Predicted E3 ubiquitin ligase	-3.642922	3.79E-41
Zm00001d023322	1473	10	LEUCINE-RICH REPEAT-CONTAINING PROTEIN	-3.90035	5.45E-28
Zm00001d021315	1344	7	RNA uridylyltransferase / TUT	-4.171407	1.04E-37
Zm00001d039744	8566	3	ANNOTATION UNKOWN	-4.764812	3.89E-95
Zm00001d044376	1207	3	Non-specific serine/threonine protein kinase / Threonine-specific protein kinase	-4.776343	3.60E-77

The trend seen in earlier GO analysis results continued; while downregulated genes were overwhelmingly involved in growth and developmental processes, upregulated genes were in biological processes relating to defense and immunity. Photosynthesis (GO:0015979) saw the highest enrichment (5.58 times) among the downregulated genes. Other processes turned down in response to *Rp1* included chlorophyll metabolic process (GO:0015994), response to light stimulus (GO:0009416), flower development (GO:0009908), and regulation of multicellular organismal development (GO:2000026) were overrepresented 3.06 – 5.13 times (Table 3-22). Upregulated genes were enriched for defense signaling processes (Table 3-23); for example, regulation of salicylic acid mediated signaling pathway (GO:2000031), transmembrane receptor protein serine/threonine kinase signaling pathway (GO:0007178), and enzyme linked receptor protein signaling pathway (GO:0007167) were enriched. Similarly, response to oomycetes (GO:0002239), defense response to oomycetes (GO:0002229), defense response to bacterium (GO:0042742), defense response to other organism (GO:0098542) were overrepresented among the upregulated genes.

Table 3-22. GO annotation for down-regulated genes from the B73:NC350RIL x H95;*Rp1-D21*/+ versus wildtype.

GO-Slim Biological Process	Zea mays - REFLIST	Actual	Expected	Fold Enrichment	Raw P-value	FDR
photosynthesis (GO:0015979)	52	36	6.45	5.58	5.55E-13	8.28E-11
chlorophyll metabolic process (GO:0015994)	11	7	1.37	5.13	2.10E-03	1.99E-02
beta-glucan biosynthetic process (GO:0051274)	23	11	2.85	3.85	7.44E-04	8.54E-03
mitotic cytokinesis (GO:0000281)	35	15	4.34	3.45	2.22E-04	2.91E-03
cytoskeleton-dependent cytokinesis (GO:0061640)	35	15	4.34	3.45	2.22E-04	2.89E-03
response to light stimulus (GO:0009416)	108	42	13.4	3.13	8.55E-09	3.99E-07
flower development (GO:0009908)	34	13	4.22	3.08	1.34E-03	1.40E-02

Table 3-22 continued

regulation of multicellular organismal development (GO:2000026)	29	11	3.6	3.06	3.23E-03	2.81E-02
membrane fission (GO:0090148)	43	15	5.34	2.81	1.25E-03	1.34E-02
cytokinesis (GO:0000910)	43	15	5.34	2.81	1.25E-03	1.33E-02
pigment biosynthetic process (GO:0046148)	32	11	3.97	2.77	5.89E-03	4.26E-02
hexose biosynthetic process (GO:0019319)	32	11	3.97	2.77	5.89E-03	4.24E-02
pigment metabolic process (GO:0042440)	35	12	4.34	2.76	4.18E-03	3.33E-02
response to radiation (GO:0009314)	123	42	15.26	2.75	2.51E-07	7.33E-06
regulation of multicellular organismal process (GO:0051239)	48	16	5.96	2.69	1.85E-03	1.84E-02
glucose metabolic process (GO:0006006)	59	19	7.32	2.6	7.85E-04	8.94E-03
cell division (GO:0051301)	54	16	6.7	2.39	3.76E-03	3.06E-02
response to abiotic stimulus (GO:0009628)	189	55	23.45	2.35	2.87E-07	8.24E-06
regulation of developmental process (GO:0050793)	95	27	11.79	2.29	3.95E-04	4.83E-03
hexose metabolic process (GO:0019318)	86	23	10.67	2.16	1.96E-03	1.91E-02
shoot system development (GO:0048367)	84	22	10.42	2.11	2.97E-03	2.64E-02
monosaccharide metabolic process (GO:0005996)	98	24	12.16	1.97	5.28E-03	3.96E-02
multicellular organismal process (GO:0032501)	169	37	20.97	1.76	3.45E-03	2.96E-02

Table 3-22 continued

multicellular organism development (GO:0007275)	165	36	20.48	1.76	3.21E-03	2.82E-02
anatomical structure development (GO:0048856)	195	42	24.2	1.74	2.37E-03	2.21E-02
generation of precursor metabolites and energy (GO:0006091)	233	50	28.91	1.73	8.01E-04	9.05E-03
developmental process (GO:0032502)	262	53	32.51	1.63	2.02E-03	1.93E-02
Unclassified (UNCLASSIFIED)	29708	3908	3686.63	1.06	1.25E-12	1.69E-10

Table 3-23. GO annotations for up-regulated genes from the B73:NC350RIL x H95;*Rp1-D21*/+ versus wildtype.

GO biological process complete	Zea mays - REFLIST	Actual	Expected	Fold Enrichment	Raw P-value	FDR
neutral amino acid transport (GO:0015804)	9	7	1.26	5.56	1.73E-03	4.71E-02
chorismate biosynthetic process (GO:0009423)	21	14	2.94	4.77	3.18E-05	1.59E-03
biological process involved in interaction with symbiont (GO:0051702)	15	9	2.1	4.29	1.44E-03	4.09E-02
regulation of salicylic acid mediated signaling pathway (GO:2000031)	15	9	2.1	4.29	1.44E-03	4.07E-02
maturation of LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA) (GO:0000463)	31	18	4.33	4.15	1.00E-05	5.69E-04
transmembrane receptor protein serine/threonine kinase signaling pathway (GO:0007178)	59	34	8.25	4.12	1.61E-09	2.02E-07
enzyme linked receptor protein signaling pathway (GO:0007167)	59	34	8.25	4.12	1.61E-09	1.97E-07
response to oomycetes (GO:0002239)	60	34	8.39	4.05	2.22E-09	2.65E-07
defense response to oomycetes (GO:0002229)	60	34	8.39	4.05	2.22E-09	2.59E-07
glucosamine-containing compound catabolic process (GO:1901072)	23	13	3.22	4.04	2.07E-04	8.17E-03
chitin catabolic process (GO:0006032)	23	13	3.22	4.04	2.07E-04	8.10E-03
chitin metabolic process (GO:0006030)	23	13	3.22	4.04	2.07E-04	8.04E-03
aminoglycan catabolic process (GO:0006026)	23	13	3.22	4.04	2.07E-04	7.97E-03
amino sugar catabolic process (GO:0046348)	23	13	3.22	4.04	2.07E-04	7.90E-03
chorismate metabolic process (GO:0046417)	29	16	4.05	3.95	4.92E-05	2.36E-03

Table 3-23 continued

cell wall macromolecule catabolic process (GO:0016998)	33	17	4.61	3.68	5.57E-05	2.59E-03
ribosomal large subunit assembly (GO:0000027)	37	19	5.17	3.67	2.14E-05	1.11E-03
double-strand break repair via break-induced replication (GO:0000727)	28	14	3.91	3.58	3.14E-04	1.12E-02
DNA replication initiation (GO:0006270)	35	17	4.89	3.47	9.75E-05	4.28E-03
aminoglycan metabolic process (GO:0006022)	31	15	4.33	3.46	2.55E-04	9.27E-03
aromatic amino acid family catabolic process (GO:0009074)	25	12	3.5	3.43	1.10E-03	3.38E-02
ribosome assembly (GO:0042255)	81	38	11.32	3.36	1.43E-08	1.38E-06
cell surface receptor signaling pathway (GO:0007166)	150	70	20.97	3.34	1.89E-14	6.75E-12
glucosamine-containing compound metabolic process (GO:1901071)	28	13	3.91	3.32	8.75E-04	2.77E-02
maturation of LSU-rRNA (GO:0000470)	65	30	9.09	3.3	5.94E-07	4.25E-05
ribosomal large subunit biogenesis (GO:0042273)	142	65	19.85	3.27	3.14E-13	7.68E-11
cytoplasmic translation (GO:0002181)	146	65	20.41	3.18	8.34E-13	1.94E-10
ribosomal small subunit assembly (GO:0000028)	36	16	5.03	3.18	3.40E-04	1.19E-02
recognition of pollen (GO:0048544)	68	30	9.51	3.16	1.24E-06	8.38E-05
cell recognition (GO:0008037)	68	30	9.51	3.16	1.24E-06	8.26E-05
L-phenylalanine metabolic process (GO:0006558)	30	13	4.19	3.1	1.44E-03	4.12E-02

Table 3-23 continued

erythrose 4-phosphate/phosphoenolpyruvate family amino acid metabolic process (GO:1902221)	30	13	4.19	3.1	1.44E-03	4.10E-02
ceramide biosynthetic process (GO:0046513)	40	17	5.59	3.04	3.40E-04	1.20E-02
benzene-containing compound metabolic process (GO:0042537)	45	19	6.29	3.02	1.67E-04	6.68E-03
defense response to bacterium (GO:0042742)	109	45	15.24	2.95	1.72E-08	1.64E-06
pollen-pistil interaction (GO:0009875)	76	31	10.63	2.92	3.88E-06	2.34E-04
response to bacterium (GO:0009617)	138	55	19.29	2.85	1.45E-09	1.87E-07
positive regulation of growth (GO:0045927)	41	16	5.73	2.79	1.62E-03	4.49E-02
dicarboxylic acid biosynthetic process (GO:0043650)	44	17	6.15	2.76	1.15E-03	3.49E-02
aromatic amino acid family metabolic process (GO:0009072)	103	37	14.4	2.57	6.32E-06	3.72E-04
defense response to other organism (GO:0098542)	253	87	35.37	2.46	1.87E-11	3.11E-09
multi-multicellular organism process (GO:0044706)	110	37	15.38	2.41	2.07E-05	1.10E-03
pollination (GO:0009856)	110	37	15.38	2.41	2.07E-05	1.09E-03
response to external biotic stimulus (GO:0043207)	292	97	40.82	2.38	5.64E-12	1.19E-09
response to other organism (GO:0051707)	292	97	40.82	2.38	5.64E-12	1.14E-09
biological process involved in interspecies interaction between organisms (GO:0044419)	310	101	43.34	2.33	6.77E-12	1.31E-09
response to biotic stimulus (GO:0009607)	326	104	45.58	2.28	7.64E-12	1.42E-09

Table 3-23 continued

aromatic amino acid family biosynthetic process (GO:0009073)	76	24	10.63	2.26	1.09E-03	3.39E-02
Golgi to vacuole transport (GO:0006896)	91	27	12.72	2.12	1.11E-03	3.39E-02
glutathione metabolic process (GO:0006749)	124	36	17.34	2.08	2.55E-04	9.34E-03
ribosomal small subunit biogenesis (GO:0042274)	142	41	19.85	2.07	1.07E-04	4.64E-03
defense response (GO:0006952)	570	163	79.69	2.05	3.12E-14	1.04E-11
ribonucleoprotein complex assembly (GO:0022618)	200	57	27.96	2.04	8.69E-06	5.06E-04
ribonucleoprotein complex subunit organization (GO:0071826)	204	57	28.52	2	1.60E-05	8.85E-04
non-membrane-bounded organelle assembly (GO:0140694)	158	44	22.09	1.99	1.40E-04	5.85E-03
cell wall macromolecule metabolic process (GO:0044036)	192	51	26.84	1.9	1.13E-04	4.83E-03
ribosome biogenesis (GO:0042254)	527	139	73.68	1.89	2.79E-10	4.06E-08
translation (GO:0006412)	955	251	133.52	1.88	2.04E-17	1.58E-14
peptide metabolic process (GO:0006518)	1130	295	157.99	1.87	5.58E-20	8.66E-17
peptide biosynthetic process (GO:0043043)	972	251	135.9	1.85	1.12E-16	6.53E-14
polysaccharide catabolic process (GO:0000272)	186	48	26	1.85	2.97E-04	1.06E-02
amide biosynthetic process (GO:0043604)	1074	274	150.16	1.82	1.80E-17	1.67E-14
cellular amide metabolic process (GO:0043603)	1283	325	179.38	1.81	3.22E-20	7.48E-17

Table 3-23 continued

response to external stimulus (GO:0009605)	463	117	64.73	1.81	4.94E-08	4.26E-06
hydrogen peroxide catabolic process (GO:0042744)	173	43	24.19	1.78	1.20E-03	3.51E-02
protein phosphorylation (GO:0006468)	1650	410	230.69	1.78	6.28E-24	2.92E-20
hydrogen peroxide metabolic process (GO:0042743)	174	43	24.33	1.77	1.69E-03	4.65E-02
ribonucleoprotein complex biogenesis (GO:0022613)	648	159	90.6	1.76	1.39E-09	1.85E-07
organic anion transport (GO:0015711)	188	45	26.28	1.71	1.82E-03	4.91E-02
carbohydrate catabolic process (GO:0016052)	365	87	51.03	1.7	1.89E-05	1.03E-03
detoxification (GO:0098754)	334	76	46.7	1.63	2.60E-04	9.40E-03
response to toxic substance (GO:0009636)	357	80	49.91	1.6	2.35E-04	8.69E-03
anion transport (GO:0006820)	326	71	45.58	1.56	1.18E-03	3.48E-02
rRNA processing (GO:0006364)	373	81	52.15	1.55	5.30E-04	1.78E-02
phosphorylation (GO:0016310)	2331	505	325.9	1.55	1.12E-18	1.30E-15
rRNA metabolic process (GO:0016072)	384	83	53.69	1.55	4.92E-04	1.66E-02
organonitrogen compound biosynthetic process (GO:1901566)	2078	441	290.53	1.52	4.28E-15	1.81E-12
carboxylic acid biosynthetic process (GO:0046394)	532	105	74.38	1.41	1.54E-03	4.30E-02
organic acid biosynthetic process (GO:0016053)	552	108	77.18	1.4	1.84E-03	4.92E-02

Table 3-23 continued

cellular response to chemical stimulus (GO:0070887)	908	175	126.95	1.38	1.41E-04	5.87E-03
cell communication (GO:0007154)	1482	283	207.2	1.37	2.06E-06	1.35E-04
cellular macromolecule biosynthetic process (GO:0034645)	2159	411	301.85	1.36	1.13E-08	1.14E-06
macromolecule biosynthetic process (GO:0009059)	2223	423	310.8	1.36	6.12E-09	6.63E-07
phosphorus metabolic process (GO:0006793)	3417	650	477.73	1.36	2.69E-13	7.36E-11
phosphate-containing compound metabolic process (GO:0006796)	3372	636	471.44	1.35	1.77E-12	3.93E-10
response to chemical (GO:0042221)	1424	267	199.09	1.34	1.34E-05	7.54E-04
oxoacid metabolic process (GO:0043436)	1135	212	158.69	1.34	1.43E-04	5.91E-03
organic acid metabolic process (GO:0006082)	1167	217	163.16	1.33	1.48E-04	6.05E-03
signal transduction (GO:0007165)	1319	245	184.41	1.33	5.21E-05	2.45E-03
carboxylic acid metabolic process (GO:0019752)	1120	208	156.59	1.33	2.15E-04	8.08E-03
signaling (GO:0023052)	1337	245	186.93	1.31	1.10E-04	4.76E-03
cellular protein metabolic process (GO:0044267)	5284	957	738.76	1.3	9.39E-15	3.64E-12
cellular nitrogen compound biosynthetic process (GO:0044271)	2091	377	292.34	1.29	5.34E-06	3.19E-04
cellular biosynthetic process (GO:0044249)	3796	672	530.72	1.27	6.70E-09	7.08E-07
transmembrane transport (GO:0055085)	1687	298	235.86	1.26	2.04E-04	8.13E-03

Table 3-23 continued

organonitrogen compound metabolic process (GO:1901564)	7378	1297	1031.52	1.26	1.03E-16	6.86E-14
biosynthetic process (GO:0009058)	4060	708	567.63	1.25	2.03E-08	1.89E-06
organic substance biosynthetic process (GO:1901576)	3911	678	546.8	1.24	8.93E-08	6.93E-06
protein metabolic process (GO:0019538)	6008	1036	839.98	1.23	2.88E-11	4.62E-09
response to stimulus (GO:0050896)	4358	751	609.29	1.23	3.60E-08	3.22E-06
organic substance transport (GO:0071702)	1677	287	234.46	1.22	1.53E-03	4.30E-02
small molecule metabolic process (GO:0044281)	1971	336	275.57	1.22	6.99E-04	2.29E-02
protein modification process (GO:0036211)	3871	656	541.21	1.21	2.48E-06	1.58E-04
cellular protein modification process (GO:0006464)	3871	656	541.21	1.21	2.48E-06	1.56E-04
macromolecule modification (GO:0043412)	4223	696	590.42	1.18	2.74E-05	1.39E-03
cellular macromolecule metabolic process (GO:0044260)	7343	1172	1026.63	1.14	3.74E-06	2.29E-04
organic substance metabolic process (GO:0071704)	12384	1959	1731.41	1.13	9.85E-10	1.39E-07
primary metabolic process (GO:0044238)	11753	1854	1643.19	1.13	9.98E-09	1.03E-06
cellular metabolic process (GO:0044237)	11945	1873	1670.04	1.12	3.73E-08	3.27E-06
metabolic process (GO:0008152)	13715	2145	1917.5	1.12	2.49E-09	2.82E-07
nitrogen compound metabolic process (GO:0006807)	9947	1541	1390.7	1.11	1.68E-05	9.19E-04
biological_process (GO:0008150)	22488	3463	3144.06	1.1	2.58E-16	1.20E-13
cellular process (GO:0009987)	17046	2607	2383.21	1.09	1.42E-08	1.40E-06

DEG intersections

Beyond the assessment of *Rp1-D21*/+-induced transcriptomic changes within each genetic background, I carried out an interaction analysis of DEGs across all five genetic backgrounds. This identified genes that were only responsive to *Rp1-D21/+* in particular genetic backgrounds as candidates for targets of genetic-background effects. In addition, this highlights genes that were reliably differentially expressed independently of genetic background. The result from this analysis was depicted with two separate visualization methods for ease of intepretation; both with a Venn diagram (Figure 3-27) and with an Upset plot (Figure 3-28).

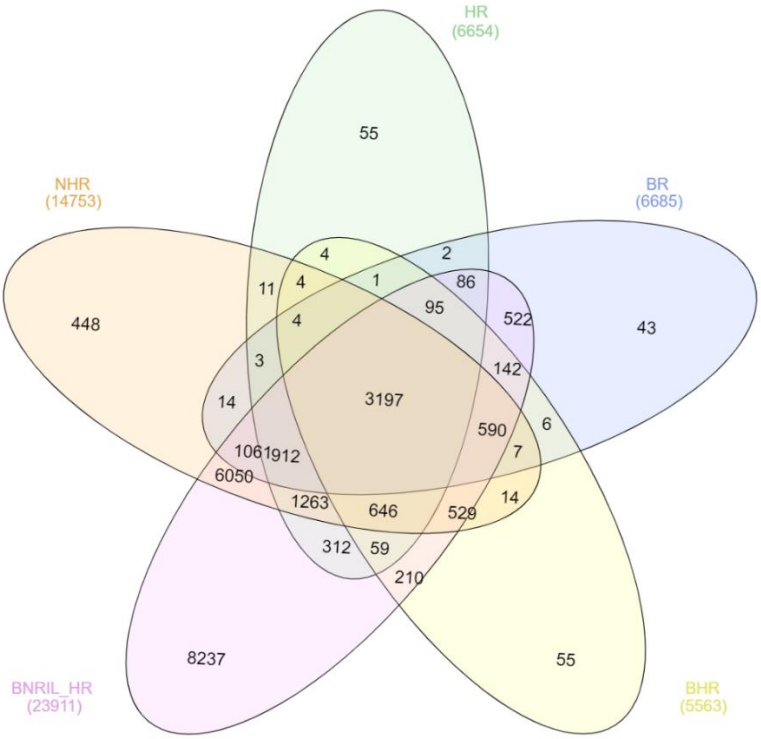


Figure 3-27. Venn diagram of differentially expressed genes among the assessed genotypes.

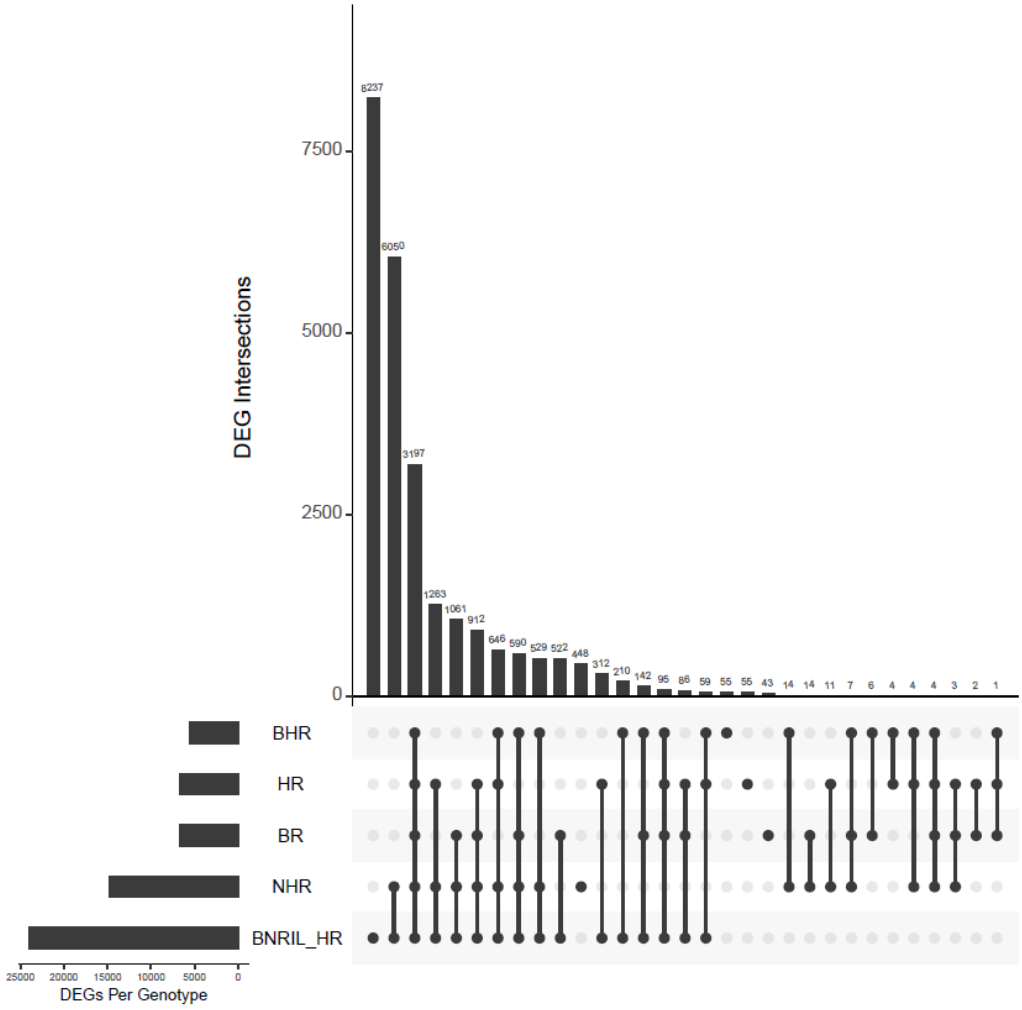


Figure 3-28. Upset plot of differentially expressed genes across genotypes showing interactions among groups.

Genes that were differentially expressed across all five backgrounds totaled 3197 (Figures 3-27 & 3-28), this represents the core set of genes that respond to *Rp1* no matter the genetic background within which it was induced. Fascinatingly, almost all these genes were moved in the same direction in all genotypes (Table 3-24). This suggests the hypersensitive response triggered by Rp1is under similar transcriptional control from genotype to genotype. Among the genes whose expression direction was opposite relative to BNRIL_HR only two shared a common direction and those were between BHR and BR (Figure 3-29), suggesting that they may not be involved in the Rp1-triggered response but rather some other genotype-specific processes.

Table 3-24. Expression direction of most consistent genes.

Genotype	Same direction with BNRIL_HR	Opposite BNRIL_HR
BHR	3194	3
NHR	3189	8
HR	3195	2
BR	3195	2

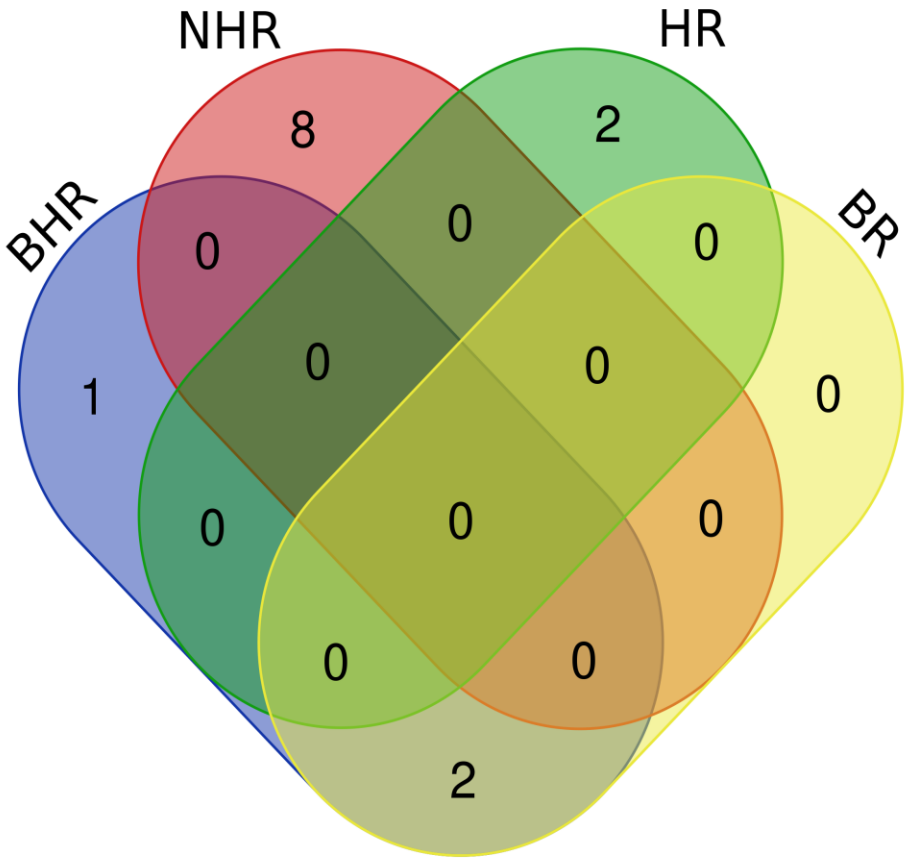


Figure 3-29. Overlap in expression direction among genes whose expression was opposite to BNRIL_HR.

3.4 Conclusions

Differential gene expression analyses carried in these experiments were successful in capturing expression changes brought about by *Rp1-D21*-induced hypersensitive response. Experimental procedure leading up to data generation were void of errors that could have adversely impacted data analysis. Gene expression differences between wild-type and mutant plants across different genetic backgrounds corresponded with HR severity. For example, the worst affected genotype, NC350, displayed the most extreme gene expression variation whilst mildly affected backgrounds such as B73 showed less extreme changes. I also demonstrated the power of replication in discovering even subtle changes in expression, as has been previously reported by others.

Analysis techniques developed and applied in these experiments permitted discovery of several defense-responsive genes whose expression was elevated as a result of *Rp1-D21*-induced HR. Conversely, many with roles in growth and developmental processes with the cell were uncovered to be turned down as a consequence of the *Rp1-D21* effect, confirming the huge cost of the defense response. Interaction analysis of DEGs from different genetic backgrounds revealed that while the *Rp1-D21* effect may be under similar control across genotypes, significant background effects exist and may be responsible for varying levels of severity of the phenotype.

CHAPTER 4. DISSECTION OF REGULATORY VARIATION AFFECTING *RP1-D21*/+ INDUCED HR VIA EQTL ANALYSIS

4.1 Introduction

Expression quantitative trait loci (eQTL) are genomic regions harboring sequence variants that influence transcription of one or more genes. These expression variants are identical to other QTL approaches, but the quantitative trait of interest is the transcript abundance of a gene (Albert & Kruglyak, 2015a; Nica & Dermitzakis, 2013). Here, expression of a gene is viewed as a quantitative assessment of its activity and change between individuals. Detection of eQTL requires comparison of transcript abundance to the genotype at molecular markers previously determined for all individuals in a mapping population. A change in transcript abundance that is significantly associated with the genotype at a molecular marker indicates genetic control of expression variation and reflects the underlying genetic variation linked to the molecular marker. As such, they can be associated with genomic regions just like any QTL affecting any trait variation (Doerge, 2002). Genome-wide eQTL, first undertaken in yeast (Brem et al., 2002), has recently been used to highlight genetic factors responsible for controlling several biological traits including anthocyanin biosynthesis in sweet potato roots (Zhang et al., 2020), fatty acid composition in rapeseed (R. Li et al., 2018), benzoxazinoid biosynthesis in maize (X. Wang et al., 2018), among others.

Since the beginning of our understanding of gene regulation (McClintock, 1949; Jacob & Monod, 1961; McClintock, 1956a, 1956b; McClintock, 1961; Peterson, 1953), gene expression variation has been classified as being encoded in *cis* or in *trans*. *Cis*-regulation is mediated by changes in DNA encoding the control elements at the gene itself, while *trans* regulation is mediated by gene products. An important distinction here is that *cis* and *trans* effects are distinguished by their molecular mechanisms. This can lead to some confusion in the interpretation of eQTL and necessitates some careful vocabulary. eQTL are either classified by their proximity to the gene or group of genes they affect into “local” or “distant” eQTL, or by the mechanism through which they influence their target genes as *cis*-acting or *trans*-acting eQTL (Albert & Kruglyak, 2015) *cis*-acting regulatory variants are changes that impact only those genes located on the same physical chromosome with it, affecting their expression initiation, rate, and stability in an allele-specific

way. *Trans*-eQTL, on the other hand, influences their target genes by changing the transcription of diffusible factors such as transcription factors or microRNAs that interact with *cis*-regulatory elements of the target genes (Wittkopp et al., 2004). Although *trans*-eQTL are typically encoded by a gene located at an unlinked position, such as a different chromosome, to the target gene they can be encoded by a linked gene. A genetically linked *trans*-acting QTL would be "local", but not *cis*. As a result, all *cis*-eQTL are local eQTL but not all local eQTL are *cis*-acting. This becomes important when co-locality of an eQTL and the target gene are used to interpret the molecular nature of the allele. While local eQTL are likely to be encoded by *cis*-variants, the eQTL itself is not definitive demonstration of this. Unlike true *cis*-eQTL these local *trans*-eQTL do not affect their target genes in an allele-specific manner and true *cis*-eQTL can be determined by additionally carrying out a test of the molecular mechanism via allele-specific expression analysis (ASE; see chapter 5 for more on this).

eQTL mapping is similar in many ways to traditional QTL analysis and identifies loci through the study of genetically diverse individuals. These individuals, drawn either from an outbred population or from an experimental cross between genetically different individuals, vary at thousands to millions of loci many of which have no impact on gene expression. Two pieces of information are needed to identify those variants that affect gene expression. First, the genotype of each individual in the population is required. Second, the phenotype, in this case, the expression of each gene in each individual within the population, is required. Genotype can be in the form of pre-existing molecular marker data, such as SNP data, or the RNA-seq data used for expression assessment can be used for variant discovery. Quantitation of gene expression is robustly and cost-effectively generated by RNA-seq experiments. Genotypes are subsequently compared to expression levels in statistical association or linkage analysis. A molecular marker is assessed for its effect on the expression of a given gene by first comparing the population of individuals according to the allele they carry. The statistical test for differences in the gene's expression between the groups indicates that an allele linked to the molecular marker is responsible for regulating the expression of this gene. Each gene is evaluated against all molecular markers in this manner to identify expression variants genome-wide (Albert & Kruglyak, 2015).

Alleles at key regulators of transcription can affect expression at multiple genes and thus encode *trans*-acting eQTL for multiple genes. e-QTL studies often characterize *trans*-eQTL "hotspots", sometimes referred to as "hubs" or "clusters." These are segments within the genome

that are enriched for *trans*-acting expression variants (de Koning & Haley, 2005; J. Tian et al., 2016). There is no accepted threshold for when a location is a "hotspot," and description of hotspots across papers are arbitrary and posthoc. Despite this, they are often quite striking and obvious after graphing the number of genes with eQTL linked to each position in the genome. These regions are thought to harbor master regulators involved in coordinated control of multiple downstream genes (Kliebenstein, 2009; L. Li et al., 2013). While this is one explanation for such hotspots, it is formally possible that hotspots identify positions where many linked genes encode multiple *trans*-acting eQTL. Biological annotation of the genes affected in *trans* can be used as a weak test of the master regulator hypothesis, as coordinated expression of a set of genes of known co-regulation is only expected if the hotspot results from many targets affected by a single regulator.

Several thresholds have been set for defining *trans*-eQTL hotspots. Some authors have used an estimate of the background distribution of eQTL for the average location in the genome. Some proportion over a number of eQTLs is identified as the basis for determining the location as a hotspot. Studies have variously defined *trans*-eQTL hotspots as regions containing more than 1% of all eQTL identified across the genome (Schadt et al., 2003) to 5% of total eQTL discovered as the cutoff for deciding a hotspot (Shi et al., 2007). Another technique for defining a hotspot threshold permutes the global distribution of eQTL randomly across the genome and identifies *trans*-eQTL hotspots as sliding windows that deviate from this at some threshold (Bolon et al., 2014; L. Li et al., 2013). A third approach aggregates the number of genes affected by *trans*-eQTL within a specified bin size (e.g., 10 cM) across the genome. Peaks within each bin then define a hotspot (Tian et al 2016). Still, other investigators have used an arbitrary numerical threshold (e.g., >10 target genes per *trans*-eQTL) to categorize hotspots (Zhou et al., 2020).

Using these strategies, the number of genes under the influence of most *trans*-eQTL hotspots identified in maize have ranged from tens to several hundred. In their study into transcription regulation in response and tolerance to drought in maize, (Y. Liu et al., 2020), identified *trans*-eQTL hotspots that controlled 21-24 genes. (X. Wang et al., 2018) identified hotspots that controlled about 60 target genes on average that were instrumental in a variety of metabolic pathways in a maize-teosinte experimental population. Other hotspots have had a wider spectrum of influence. The four *trans*-eQTL hotspots identified for grey leaf spot (GLS) were found to each influence between 141-407 genes (Christie et al., 2017). Similarly, 10 *trans*-eQTL

hotspots were highlighted in a study of gene expression variation in shoot apices in a maize biparental population which affected the expression of 289 genes on average (L. Li et al., 2013).

I carried out eQTL mapping using a novel approach (Figure 4-1) that allowed me to associate natural variation in gene expression to those changes brought about by *Rp1-D21*-induced hypersensitive response in maize. Our collaborators took a recombinant inbred line population resulting from a cross between two parents with divergent sensitivity to *Rp1-D21* and crossed it to a common parent carrying the *Rp1-D21* mutation (introduced in Chapter 3). The common parent was the H95 inbred carrying *Rp1-D21/+* as a heterozygote (Chintamanani et al., 2010). The recombinant inbred population was the 200-line B73 x NC350 subpopulation of the maize association mapping (NAM) recombinant inbred lines (RILs)(McMullen et al., 2009; J. Yu et al., 2008). By carrying out RNA-seq on the Wildtype and Mutant siblings of each RIL x H95;*Rp1-D21/+* F1 family, I was able to map eQTL affecting expression in wildtype maize hybrids, *Rp1-D21/+* mutant hybrids. In addition, by using the difference in transcript abundance between wildtype and mutant siblings, I was able to eliminate eQTL resulting from background processes unrelated too immunity and identify eQTL specifically affecting expression variation during the hypersensitive response. This identified, for the first time, the genome-wide scale of expression variation during plant immune signaling and the largest eQTL hotspots in any organism.

4.2 Methods

4.2.1 Plant material and RNA sequencing data

I used RNA-seq data from F1 families produced by crossing 99 members of the B73 x NC350 RIL subpopulation and H95*;Rp1-D21/+* (Figure 4-2; also see Chapter 3) to map variants controlling gene expression in response to HR. The size of this subpopulation coupled with the extensive high-density genotype data available enables accurate identification of QTL (F. Tian et al., 2011). Since H95*;Rp1-D21/+* is heterozygous for the *Rp1-D21* allele, the F1 progeny segregate 1:1 for those carrying *Rp1-D21/+* (mutant F1 progeny) and those homozygous for wildtype *Rp1* alleles from the parents (non-autoactive phenotype). With the exception of the *Rp1* locus, F1 siblings are nearly isogenic. RNA was extracted from pools of wildtype and mutant siblings from

99 F1 families, converted to cDNA libraries and prepared for sequencing by our collaborator at the USDA-ARS, Dr. Peter Balint-Kurti lab, on the campus of North Carolina State University.

4.2.2 Genotypic data

I extracted genotype data for the 7,386 SNP positions corresponding to the 200 NC350 RILs from the set of 4,892 NAM RILs (Olukolu et al., 2014) for use in the eQTL analysis. Since these SNPs were originally anchored on the maizeAGPv2 reference genome, I converted their locations to the more recent AGPv4 assembly using CrossMap (Zhao et al., 2014) Some 6,685 SNPs were successfully converted from AGPv2 and were used in conjunction with gene expression data for eQTL mapping. The 701 SNPs that could not be converted were located on contigs that only existed in the previous reference version and could thus not be placed on the new genome.

4.2.3 Reads mapping and processing of expression data

RNA-seq reads were mapped to the B73 v4 reference genome and read counts per gene were computed for each of the 99 F1 sibling progeny pairs (Figure 4-1). To reduce mapping bias due to the reference genome, the maize AGPv4 reference genome was anonymized with NC350 SNPs using the bcftools consensus package from SAMtools v1.8 (H. Li, 2011). Single-end RNA-seq reads were aligned to the anonymized reference using STAR v2.7.9a with a transcript annotation (Dobin et al., 2013). Raw read counts mapped to a given gene were computed with HTSeq v0.6.1 (Anders et al., 2015). These counts are not comparable among samples due to varying library sizes or sequencing depths among samples. Still, raw counts of different genes within a single sample cannot be compared due to different transcript lengths; whist longer transcript have more reads aligned to them shorter transcripts with similar expression level will have fewer reads mapped. For these reasons, raw read count normalization is key to guaranteeing accurate comparison of gene expression data (Dillies et al., 2013; Zhao et al., 2020). The normalized expression unit, reads per kilobase of transcript per million reads mapped (RPKM), was used rather than the raw counts to limit the effects of technical bias introduced during the sequencing step. RPKM corrects for library size and transcript length differences and thus allows for comparison of expression levels within and across samples (Mortazavi et al., 2008). Raw read

counts per gene were normalized with the RPKM normalization procedure as implemented in DESeq2 (Love et al., 2014)package for use as input in eQTL analysis.

4.2.4 eQTL Mapping

Genetic association was assessed between each SNP and each normalized gene count using MatrixEQTL (Shabalin, 2012) under the linear model,

$G = \beta_0 + \beta_1 * SNP + \epsilon$,

where G is the normalized expression value of the gene being influenced, β_1 is the SNP allele substitution effect, SNP is the genotype covariate, β_0 is the intercept, ϵ is the residual or error term. The genotype covariate is coded as 0 for homozygous B73, 1 for the heterozygote, and 2 for homozygous NC350. This means a positive effect at an eQTL ($\beta_1 > 0$) indicates that the NC350 allele increased, and B73 allele decreased the expression level of the gene. A negative effect ($\beta_1 < 0$) indicates that the NC350 allele decreased the expression of the gene and the B73 allele increased that gene's gene expression.

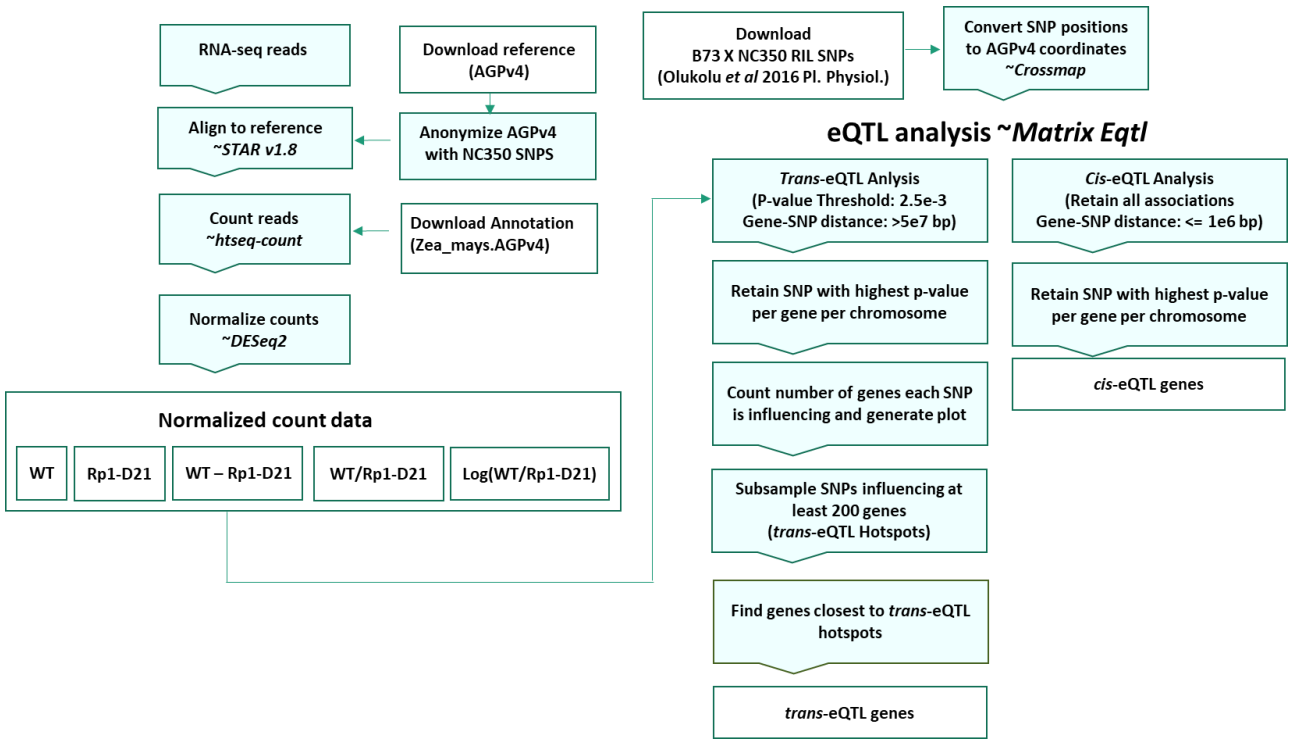


Figure 4-1. Overview of eQTL analysis workflow. RNA-seq reads was mapped to the B73 v4 reference genome and read count per gene computed. Raw count data was normalized prior eQTL analyses.

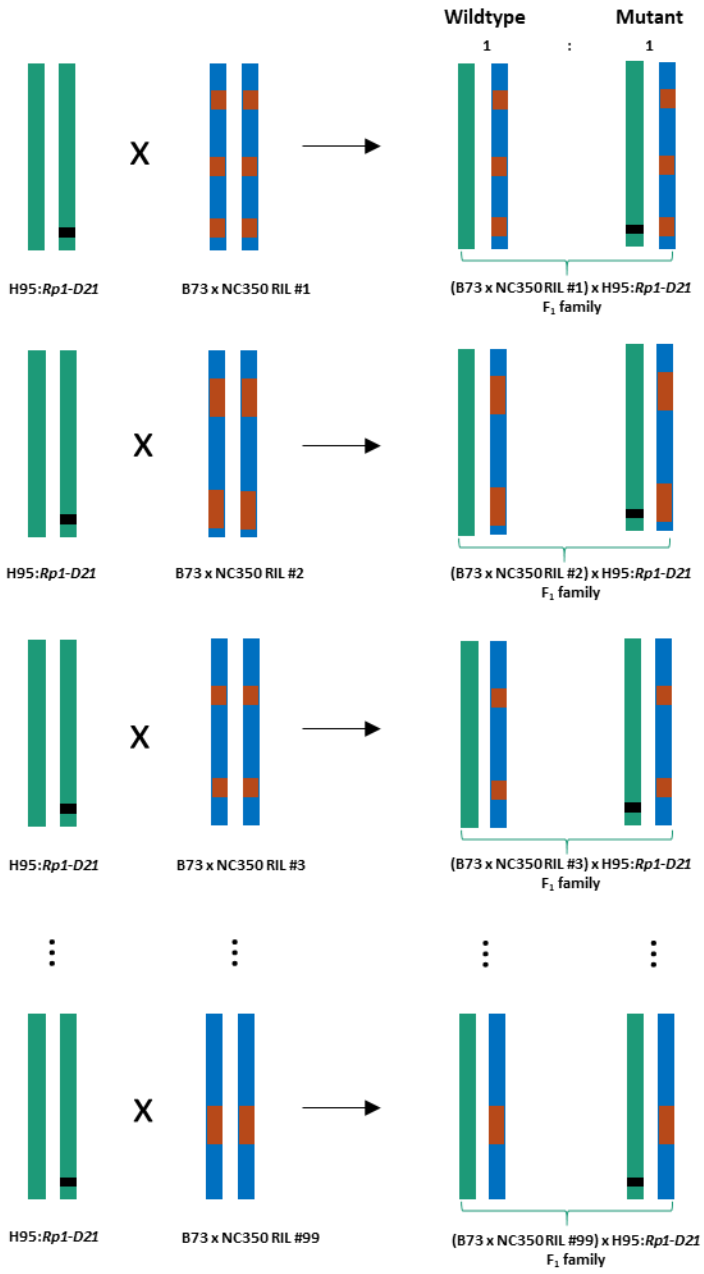


Figure 4-2. Illustration of the cross between H95:*Rp1-D21* and 99 members of B73 x NC350 recombinant inbred lines (RILs). F1 offspring from H95:*Rp1-D21* and B73 x NC350 cross segregate 1:1 ratio for F1 offspring carrying *Rp1-D21* allele (mutant constitutive HR F_1 progeny) and F_1 offspring carrying the wildtype H95 allele at the *Rp1* locus (non-autoactive phenotype). The F_1 offspring are nearly isogenic except at the *Rp1* locus.

Cis-eQTL for each gene were defined as the marker with the maximum QTL test statistic within 1 megabase (Mb). The window included all sequence from 1 Mb upstream of the transcription start site to 1 Mb downstream of the gene's termination site (Figure 3A). For each gene, I retained and tabulated the most significantly associated SNP located within 1 Mb the *cis*-eQTL. Subsequent filtering of this table at any significance threshold permits identification of significant associations due to a linked allele altering each target gene's expression. By only retaining the SNP with the lowest p-value, a single *cis*-eQTL is tabulated for each gene. Since *cis*-eQTL mapping is carried out within a 1 Mb window, it is formally possible that locally-encoded *trans*-acting eQTL linked to this region (see introduction and Figure 4-3) may be responsible for the association. Identification of *trans*-eQTL was also undertaken. To avoid errantly scoring *cis*-eQTL as *trans*, no *trans* eQTL were retained within a window of 50 Mb around each target gene (Figure 4-3B). To reduce data tabulated while retaining flexibility in downstream analysis, *trans*-eQTLs test statistics were retained for all gene-marker tests with p-values below 2.5×10^{-3}. This approach identifies all marker-transcript associations at this threshold with a low likelihood of spurious linkage with *cis*-acting variants. This set of associations can be further filtered for significance.

Cis-eQTL and *trans*-eQTL were performed on RNA-seq estimates of transcript abundance from F1 progeny showing the mutant or wildtype phenotype separately. Additionally, to remove the effects of background effects on eQTL detection, a series of expression phenotypes were derived from these data and used as phenotypic input for eQTL analyses. First, the difference between the mutant and wildtype normalized counts was obtained for each gene in each RIL F1 family. Next, a ratio of wild type to mutant normalized counts was computed per gene in each RIL F1 family. Lastly, a log transformation was applied to the ratio of wildtype to mutant normalized counts, to reduce the effects of non-normality in the data on eQTL detection. These modified input data were then used separately to perform eQTL analyses. *Trans*-eQTL were identified as described above with the same locus definitions and thresholds for all input data, and *cis*-eQTL analysis was carried out only using input data from the wild type, mutant and difference between mutant and wild-type expression.

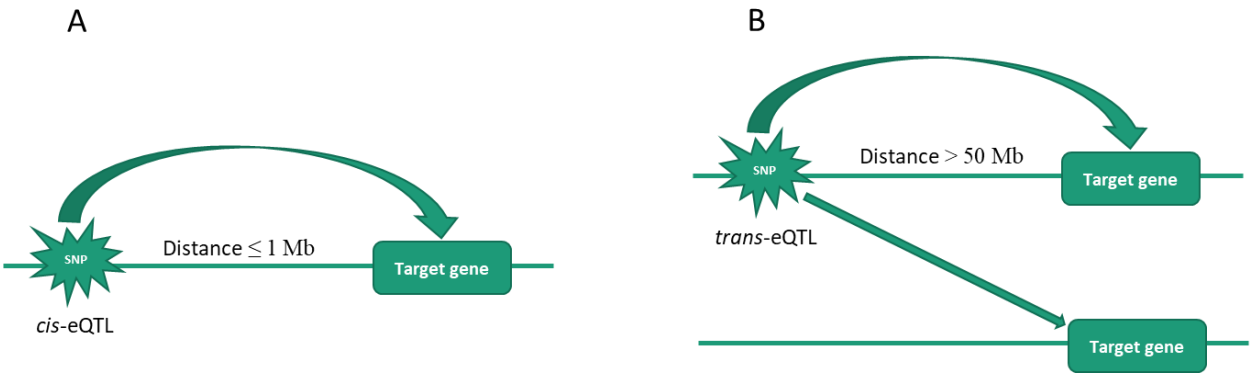


Figure 4-3. Illustration of eQTL mapping criteria for defining *cis*-eQTL (A) and *trans*-eQTL (B). *cis*-eQTL analysis searches within 1 Mb of SNP for significant associations to target genes whereas *trans*-eQTL were only identified if the target gene and SNP were on different chromosomes or more than 50 Mb when encoded on the same chromosome.

4.3 Results

4.3.1 Analysis of *cis*-eQTL in the RIL x H95;*Rp1-D21*/+ families identifies HR modulated genes as targets for cis – regulatory variation

Abundance levels were computed for all maize AGPv4 reference genes for all 198 samples, mutant and wildtype, from all 99 RIL x H95;*Rp1-D21*/+ families. Following quality control steps after alignment, 41,474 genes had non-zero read counts for all individuals and were kept for further analysis. 6,685 SNPs were used in conjunction with the gene expression data to evaluate 2.8×10^8 SNP-gene associations for eQTL discovery.

eQTL mapping tested 454,326 gene-SNP associations for local/*cis*-eQTL discovery. This tests expression variants within a 1 Mb window of each target gene. A quantile-quantile (QQ - plot) of observed versus expected p-values showed that p-values for local eQTL broke from the diagonal much earlier than those of distant eQTL, demonstrating that distant eQTL were relatively more difficult to detect (Figure 4-4). After retaining only the most significantly associated gene-SNP pairs per chromosome 10,317 and 10,407 significant *cis*-eQTL were identified for wildtype (WT) and mutant (MU) F1 respectively. The test with the difference in expression values between wildtype and mutant (MIN) produced 6,209 significant *cis*-gene-SNP associations.

Table 4-1. Summary results from *cis*-eQTL and their relationship with genes affected by *Rp1-D21*/+ in NC350 x H95 (NH) hybrids.

cis-eQTL	Gene count	DE NH	Not DE	DE NH expected proportion[a]	Not DE expected proportion[a]	DE NH expected	Not DE expected	χ2 pval enrich
WT	10317	5775	4542	0.495	0.505	5106.92	5210.08	1.59E-39
MU	10407	5528	4879	0.495	0.505	5151.47	5255.53	1.56E-13
MIN	6209	3567	2642	0.495	0.505	3073.46	3135.54	5.28E-36

[a] Proportion is calculated from data in Chapter 3.

Overlap analysis between *cis*-eQTL target genes and genes differentially expressed between wildtype and *Rp1-D21*/+ mutants in the NC350 x H95 (NH) F1 hybrid backgrounds was assessed. Of the genes differentially expressed, 5775 overlapped with *cis*-eQTL targets in the WT, 5528 in MU, and 3567 in MIN. All of these were greater than the 5107, 5151 and 3073 expected by random chance in the WT, MU, and MIN *cis*-eQTL targets respectively. Tests of enrichment assessed by Chi squared test showed significance for all three groups (Table 4-1).

Intersection of DEG between wildtype and mutant F1 hybrids from the B73 x H95 cross and *cis*-eQTL target genes was also carried out. There is a lesser impact of *Rp1-D21*/+ on gene expression (see chapter 3) and a suppression of the HR phenotype in these genetic backgrounds. Out of the DEGs identified, the number overlapping with *cis*-eQTL target genes were 2199 in the WT, 2049 in the MU, and 1543 in the MIN. These overlapping genes outnumber the expectation of 1980.66 in the WT, 1997.93 in the MU, and 1192 in the MIN *cis*-eQTL targets. However, chi-square enrichment tests only revealed significant enrichment of DEGs in *cis*-eQTL gene targets for the WT and MIN but not for MU analysis (Table 4-2).

Since differential expression analysis was additionally carried out between wildtype and mutant F1 progeny from the H95*;Rp1-D21*/+ and B73 x NC350 RIL (BNRIL) cross, there was interest in ascertaining how many of these DEGs overlapped with the *cis*-eQTL targets. Similar to our earlier observation in the NH background the *cis*-eQTL targets were enriched for DEGs. The number of DEGs found to be overlapping with *cis*-eQTL were 8677 for WT, 8668 for MU, and 5211 for MIN, significantly more than the 6584.89, 6642.33, and 3962.93 expected by random chance in the WT, MU, and MIN *cis*-eQTL targets respectively (Table 4-3).

Table 4-2. Summary results from *cis*-eQTL and their relationship with genes affected by *Rp1-D21*/+ in B73 x H95 (BH) hybrids.

cis-eQTL	Gene count	DE BH	Not DE	DE BH expected proportion[a]	Not DE expected proportion[a]	DE BH expected	Not DE expected	$\chi 2$ enrich
WT	10317	2199	8118	0.192	0.808	1980.66	8336.34	4.82E-08
MU	10407	2049	8358	0.192	0.808	1997.93	8409.07	2.04E-01
MIN	6209	1543	4666	0.192	0.808	1192.00	5017.00	1.17E-29

[a] Proportion is calculated from data in Chapter 3.

It is not surprising that only a fraction of *cis*-eQTL target genes overlapped with DEGs. We do not expect every gene that is differentially expressed to also encode a *cis*-acting variant allele in the B73 x NC350 cross. There was more consistent enrichment of DEGs in the NH and BNRIL backgrounds compared to BH, indicating that the better you are at detecting differentially expressed genes the more enriched they are in *cis*-eQTL. The fact that *cis*-eQTL targets were enriched for differentially expressed genes even in the WT is intriguing. This test may not be indicative of an underlying biological phenomenon connected to gene expression changes brought by *Rp1-D21*-induced HR. Rather the slight enrichment observed my indicate that the tests used for both *cis*-eQTL and DEG analyses are more sensitive at the most abundantly expressed genes and as a result are non-randomly distributed among the set of expressed genes. That said, the DEG analysis using the entire B73xNC350 recombinant inbred line F1 population has 99 replicates, and as a result is far less sensitive to such a reads-depth artifact. Yet, it is this overlap list that provided the strongest test statistics. This favors a biological explanation, such as genes modulating plant immune responses and affected by HR have accumulated more regulatory variation than the average gene. This would result in more *cis*-eQTL, even in the wildtype siblings eQTL experiment, being found at HR-affected DEGs than expected by random chance.

Table 4-3. Summary results from *cis*-eQTL and their relationship with DEGs in F1 progeny from the H95;*Rp1-D21* and B73xNC350 recombinant inbred lines (BNRIL) cross progenies.

cis-eQTL	Gene count	DE RIL	Not DE	DE BH expected proportion[a]	Not DE expected proportion[a]	DE BH expected	Not DE expected	$\chi 2$ enrich
WT	10317	8677	1640	0.638	0.362	6584.89	3732.11	0.00E+00
MU	10407	8668	1739	0.638	0.362	6642.33	3764.67	0.00E+00
MIN	6209	5211	998	0.638	0.362	3962.93	2246.07	2.74E-238

[a] Proportion is calculated from data in Chapter 3.

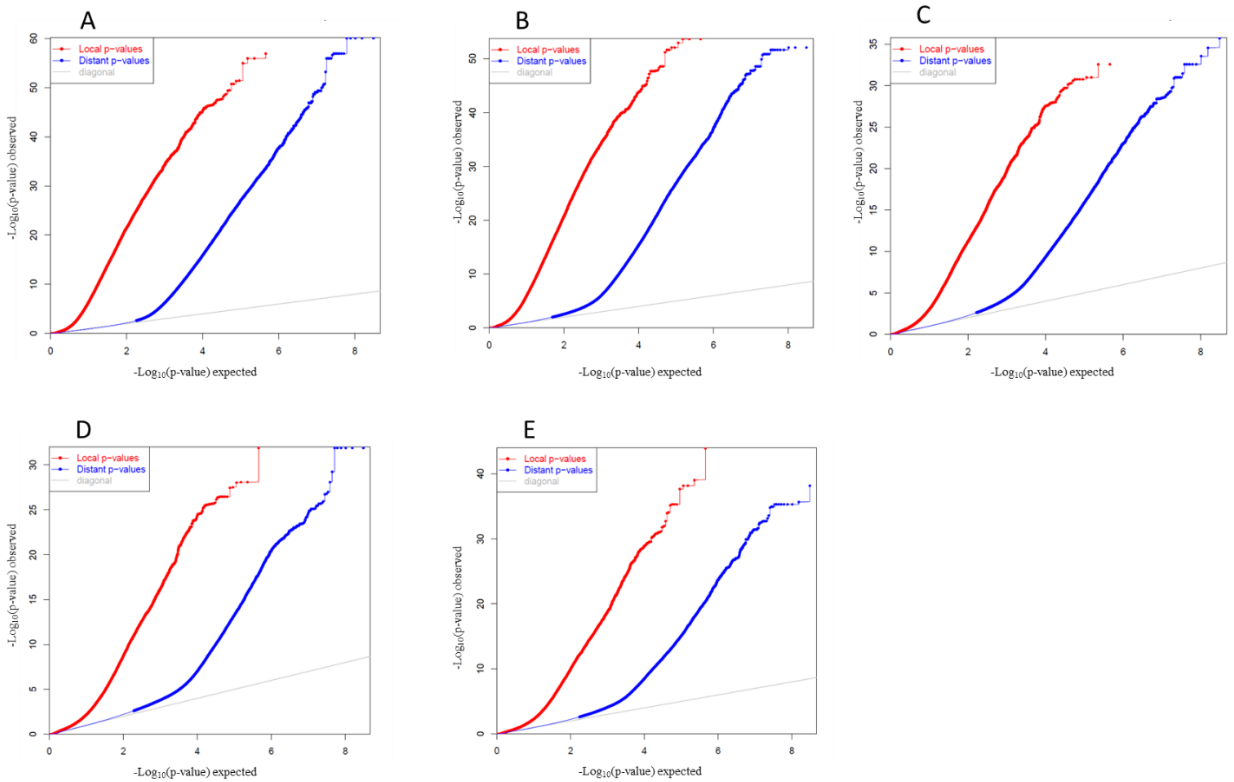


Figure 4-4. Quantile-quantile (QQ -plot) of observed against expected p-values from *cis*-eQTL analysis in wild-type RIL F1s (A), *Rp1-D21*/+ RIL F1s (B), difference between wildtype and *Rp1-D21*/+ RIL F1s (C), ratio of the wildtype to *Rp1-D21*/+ RIL F1s (D) and the Log of the ratio of wildtype to *Rp1-D21*/+ RIL F1s (E). The x-axis denotes the theoretical p-value whilst the y-axis shows observed p-value. Local p-values are from SNP-gene associations within 1 Mb; distant p-values are from SNP associations with genes more than 50 Mb away.

Integrated analyses using effect direction of DEGs and *cis*-eQTL were carried out to assess which allele (B73 or NC350) more closely resembles the *Rp1-D21*/+ effect. Any bias would indicate a differential accumulation of *cis*-acting variants promoting greater or lesser immune responses in a particular line. For each gene that was significantly differentially expressed, the direction of effect observed at each eQTL was compared to the effect of *Rp1-D21* had on that gene's expression (analysis in Chapter 3). For WT *cis*-eQTL, the number of times the NC350 allele shared the same effect direction with *Rp1-D21* (2818) did not significantly differ from the number of times the B73 allele's effect direction was the same as *Rp1-D21* (2957) in the NH background (Table 4-4 and Figure 5-5). In the comparison between WT *cis* eQTL in the mapping experiments and DEG in the BH background, the number of times the NC350 allele shared the same effect direction as *Rp1-D21* (1163) was statistically different from that of *Rp1-D21* (1036), though the absolute number of differences was quite modest (Table 4-4 and Figure 4-6). The observation of no difference and a weak effect are somewhat expected as WT samples did not harbor the *Rp1-D21* allele and consequently sensitivity to *Rp1* was not being measured by this set of eQTL.

In contrast, for *cis*-eQTL detected in the RIL F1 *Rp1-D21*/+ mutant samples, statistically significant differences could be observed between the number of times the NC350 allele shared the same effect direction as *Rp1-D21* (3196) and that of the B73 allele (2332) when differential expression was assessed in the NH background (Table 4-4 and Figure 4-5). For the BH DEGs the same trend was observed, 1081 and 968 for the number of times *Rp1-D21* effect direction was the same as NC350 and B73 respectively in the MU *cis*-eQTL targets (Table 4-4 and Figure 4-6). Clearly, NC350 alleles more often matched the direction of the *Rp1-D21* effect on gene expression than the B73 alleles. This suggests that some of the sensitivity to *Rp1-D21* previously reported (Chintamanani et al., 2010) may result from an accumulation of *cis*-regulatory variants that enhance the effect of *Rp1-D21* on gene expression. Alternatively, there may be a substantial and surprising contribution of local-*trans* effects to the set eQTL investigated here.

Table 4-4. Comparison of effect direction of *cis*-eQTL and differentially expressed genes.

cis-eQTL	Gene count	*Rp1-D21* in NH effect direction[a]			*Rp1-D21* in BH effect direction[b]		
		Effect of NC350 allele and *Rp1* the same	Effect of B73 allele and *Rp1* the same	χ2 enrich	Effect of NC350 allele and *Rp1* the same	Effect of B73 allele and *Rp1* the same	χ2 enrich
WT	10317	2818	2957	0.067	1163	1036	0.0068
MU	10407	3196	2332	3.24E-31	1081	968	0.013
MIN	6209	1856	1711	0.015	754	789	0.372

[a] Proportion is calculated from data in Chapter 3.

[b] Proportion is calculated from data in Chapter 3.

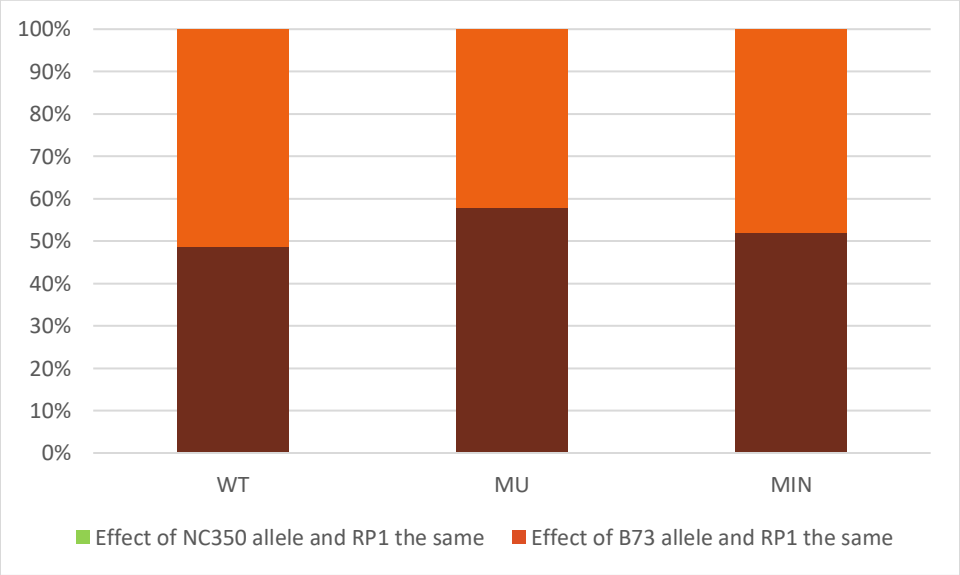


Figure 4-5. Effect direction of differentially expressed genes between wildtype and mutant progeny from NC350 x H95;*Rp1-D21* (NH) in *cis*-eQTL analysis. *Cis*-eQTL analyses were performed in wildtype (WT) and *Rp1-D21* (MU) F1 progeny from the H95*;Rp1-D21* and B73 x NC350 recombinant inbred lines (RIL) cross. MIN corresponds to *cis*-eQTL analysis using the difference in each gene's expression between wildtype and *Rp1-D21* F1. Dark brown denotes proportion of DEGs between wildtype and mutant NH F1, for which the NC350 allele has the same effect direction as *Rp1*. Orange represents the proportion of DEGs, between wildtype and mutant NH F1, for which the B73 allele has the same effect as *Rp1*. The numbers used in this chart are drawn from Table 4-4.

The results from the MIN *cis*-eQTL targets were mixed between the NH and BH backgrounds. For the NH DEGs that were also targets for *cis*-eQTL, a statistically significant difference was observed between the number of times the effect direction of the NC350 allele was the same as *Rp1-D21* (1856) and the number of times the effect direction of the B73 allele's was the same as *Rp1-D21* (1711) (Table 4-4 and Figure 4-5). However, for the DEGs in the BH background that were also targets for MIN *cis*-eQTL, a statistically-significant difference was not seen between NC350 and B73 (Table 4-4 and Figure 4-6). This suggests that controlling for the changes in gene expression in the wild-type samples removed the effect observed in the mutant samples rather than enhance it. This would be consistent with the observation and interpretation of Table 4-3, that genes involved in the HR have accumulated more *cis*-regulatory variation in general.

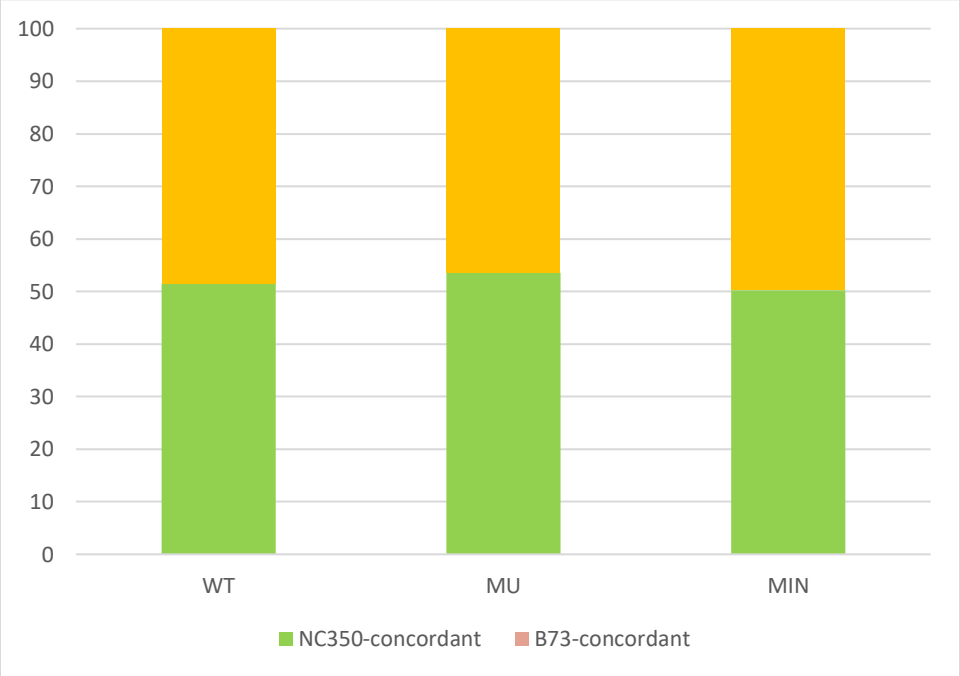


Figure 4-6. Effect direction of differentially expressed genes between wildtype and mutant progeny from B73 x H95;*Rp1-D21* (BH) in *cis*-eQTL analysis. *Cis*-eQTL analyses were performed in wildtype (WT) and *Rp1-D21* (MU) F1 progeny from the H95*;Rp1-D21* and B73 x NC350 recombinant inbred lines (RIL) cross. MIN corresponds to *cis*-eQTL analysis using the difference in each gene's expression between wildtype and *Rp1-D21* F1. Green denotes proportion of DEGs between wildtype and mutant BH F1, for which the NC350 allele has the same effect direction as *Rp1*. Yellow represents the proportion of DEGs, between wildtype and mutant BH F1, for which the B73 allele has the same effect as *Rp1*. The numbers used in this chart are drawn from Table 4-4.

4.3.2 Analysis of *trans*-eQTL in the RIL x H95;*Rp1-D21*/+ families identifies an outsized role for HR – modulated regulatory hotspots at the top of the regulatory hierarchy

Trans-eQTL analysis is useful to identify the long-range effects of an allele. This analysis was carried out to discover variants that regulate transcription of remote genes or network of such genes as a consequence of HR induced by *Rp1-D21/+*. I assessed gene-SNP associations at greater than 50 Mb from each target gene in the RNA seq data from RIL F1 plants. Wild-type and *Rp1-D21/+* mutant sibling sets from the cross between H95*;Rp1-D21/+* and B73xNC350 RIL cross were analyzed separately. I also derived values from the expression data so the analysis would minimize the effect of background eQTL and to maximize statistical power. The same transformations were used as described above: differences between the mutant and wild type normalized counts; a ratio of wild type to mutant normalized counts; and a log transformation of the ratio of wildtype to mutant normalized counts. All calculated values were also utilized for *trans*-eQTL analysis.

Table 4-5. Summary of *trans*-eQTL analysis results in different phenotypic backgrounds.

Background	Number of significant SNP-gene associations	Number of hotspots	Average number of genes per hotspot
WT	904,211	23	353.0
MU	2,069,311	17	1,589.9
MIN	1,436,895	15	1,056.7
DIV	1,338,677	15	1,118.2
LOG	1,466,213	16	1,062.2

I assessed associations between SNPs and gene-level expression for 295,060,714 *trans*-SNP-gene pairs. This large number of SNP-gene pair associations being tested makes *trans*-eQTL discovery more difficult in comparison with *cis*-eQTL. To limit the problem of linked loci affecting the same gene and reduce the number of false positives retained by the process, for each gene only the most significantly associated SNP per chromosome was kept for downstream analyses. A threshold of 200 target genes (dotted line in Figures 4-7 to 4-11) was set as an arbitrary threshold for the identification of *trans*-regulatory hotspots and likely to encode alleles at regulators controlling the expression of many genes.

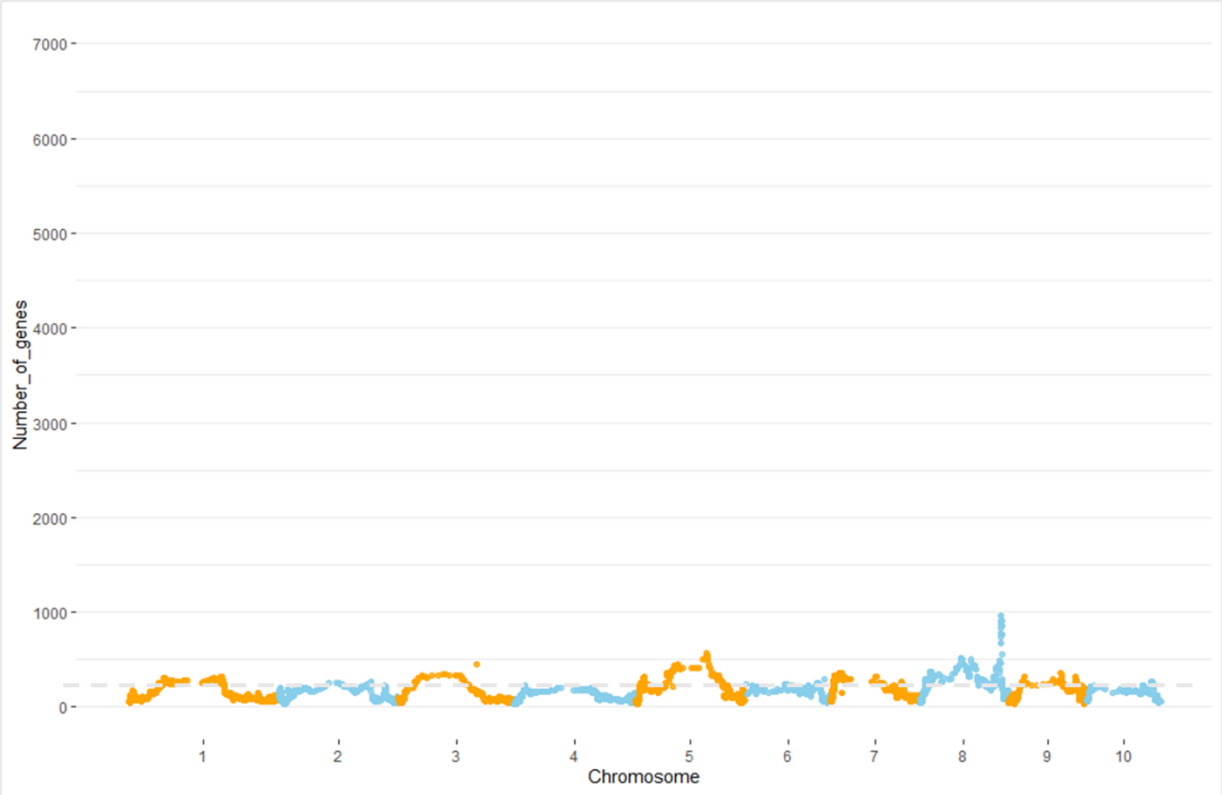


Figure 4-7. *Trans*-eQTL results in F1 progeny from the cross between H95;*Rp1-D21*/+ and B73 x NC350 recombinant inbred lines (RIL) showing wildtype (WT) phenotype. X-axis represents chromosome number; y-axis is the number of genes influenced by each SNP. Dotted red line is at 200 and represents the minimum number of genes a SNP must influence to be considered a hotspot.

The *trans*-eQTL analysis of gene expression variation in the wildtype background detected 904,211 significant SNP-gene associations (Table 4-5). Using the threshold indicated above, 23 *trans*-eQTL hotspots were identified (Figure 4-7). Notably, only one of these *trans*-eQTL, located on chromosome 8, was shared with a *trans*-eQTL from the mutant background (Figure 4-8). This common hotspot, at marker m5923, was the strongest eQTL hotspot in the wildtype siblings experiment and significantly affected the expression of 964 genes. On average the hotspots detected in the wildtype samples influenced the expression of 353 genes. This demonstrates that this cross is a rich source of both *cis* and *trans* regulation, including a number of loci with influence on a very large number of transcripts consistent with alleles encoded by critical regulators of transcription. These wild-type F1 progeny, however, do not in themselves allow insight into the regulation of the HR.

Trans-eQTL analysis in the F1 progeny showing the mutant (MU) phenotype identified 2,069,311 significant SNP-gene associations. This was more than twice what was found in the wildtype analysis. This suggests that *trans*-regulatory variation affecting expression differences affects the expression of more genes when the hypersensitive response is triggered. One mechanism for this would be allelic variation in HR-determinants, perhaps encoded by the alleles that altered lesion severity in the NAM population (Olukolu et al., 2016). Based on the minimum threshold of 200 target genes, 17 *trans*-eQTL hotspots were characterized. Only the previously mentioned single hotspot at m5923) was shared between the wildtype and mutant samples. This hotspot influenced 304 genes in the mutant background (Table 4-6). The fact that this *trans*-eQTL is shared between wildtype and mutant suggests that it is not influencing its target genes in response to the *Rp1-D21*-induced HR but rather maybe acting as part of transcription control of normal growth and developmental related activities in the plant. This idea is supported by the fact that in the mutant analysis, the number of genes affected by this hotspot is 3-fold lower in comparison to the wildtype. We know, from the stunted habit and lesion-like structures on plants showing the *Rp1-D21* phenotype that normal growth and development is sacrificed to make way for a strong hypersensitive response and that GO annotations associated with growth, development, and primary metabolism are among the down regulated genes in *Rp1-D21* mutants (see chapter 3).

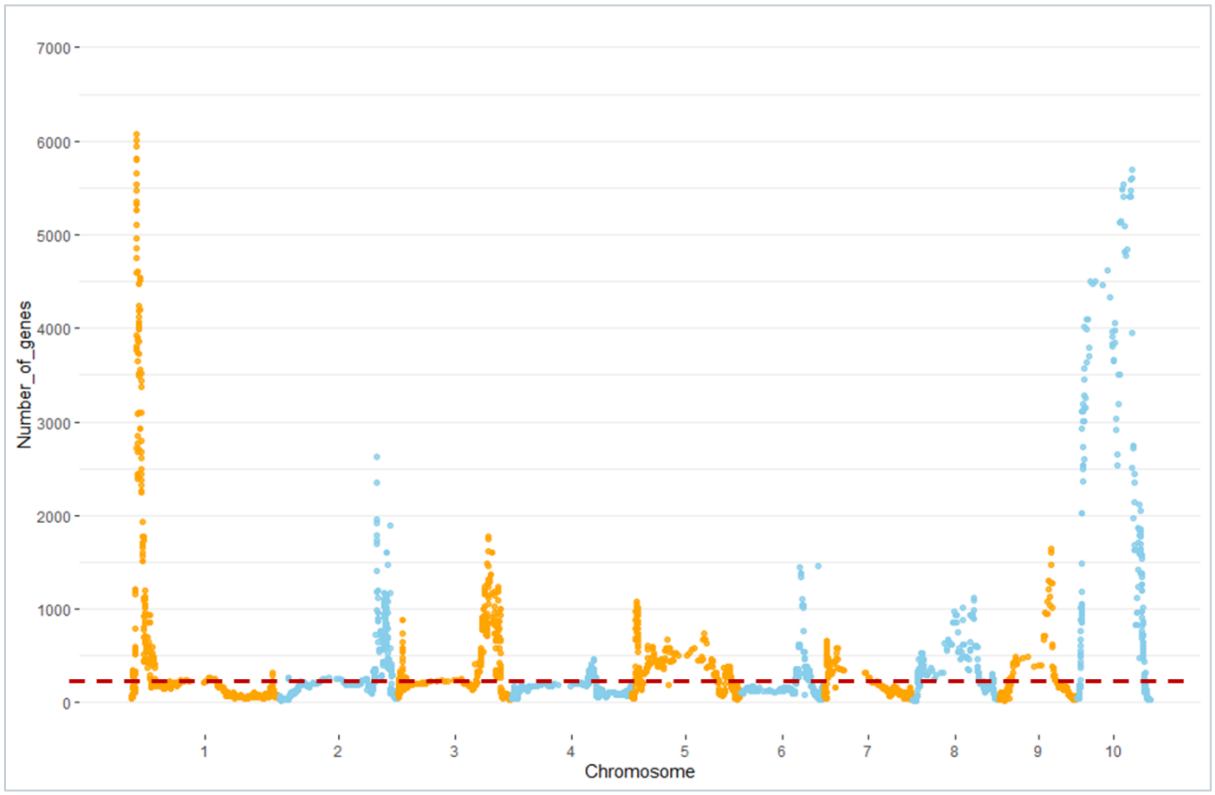


Figure 4-8. *Trans*-eQTL results in F1 progeny from the cross between H95;*Rp1-D21*/+ and B73 x NC350 recombinant inbred lines (RIL) showing *Rp1-D21* (MU) phenotype. X-axis represents chromosome number; y-axis is the number of genes influenced by each SNP. Dotted red line represents the minimum number of genes a SNP must influence to be considered a hotspot.

Trans-eQTL hotspots were identified on all chromosomes except for chromosome 4. The identified hotspots influenced an average of 1,589.9 genes, 5X the average number of *trans*-eQTL target genes found in the wildtype analysis (Table 4-5). Indeed, the two hotspots with the highest number of target genes, m139 located on chromosome 1 and m7076 located on chromosome 10, controlled the expression of 6,077 and 5,700 genes respectively (Table 4-6 and Figure 4-8). This is indicative of an extremely strong transcription response brought about by these alleles and is reminiscent of the numbers of genes detected as affected by *Rp1-D21* auto activation (see Chapter 3). One possible explanation for the extremely large number of genes affected by these *trans*-eQTL hotspots is that they sit upstream in the defense response pathway. Alleles that encode upstream modulators of HR would be expected to coordinately modify the accumulation of all HR-responsive transcripts by increasing, or decreasing, the intensity of the entire host response.

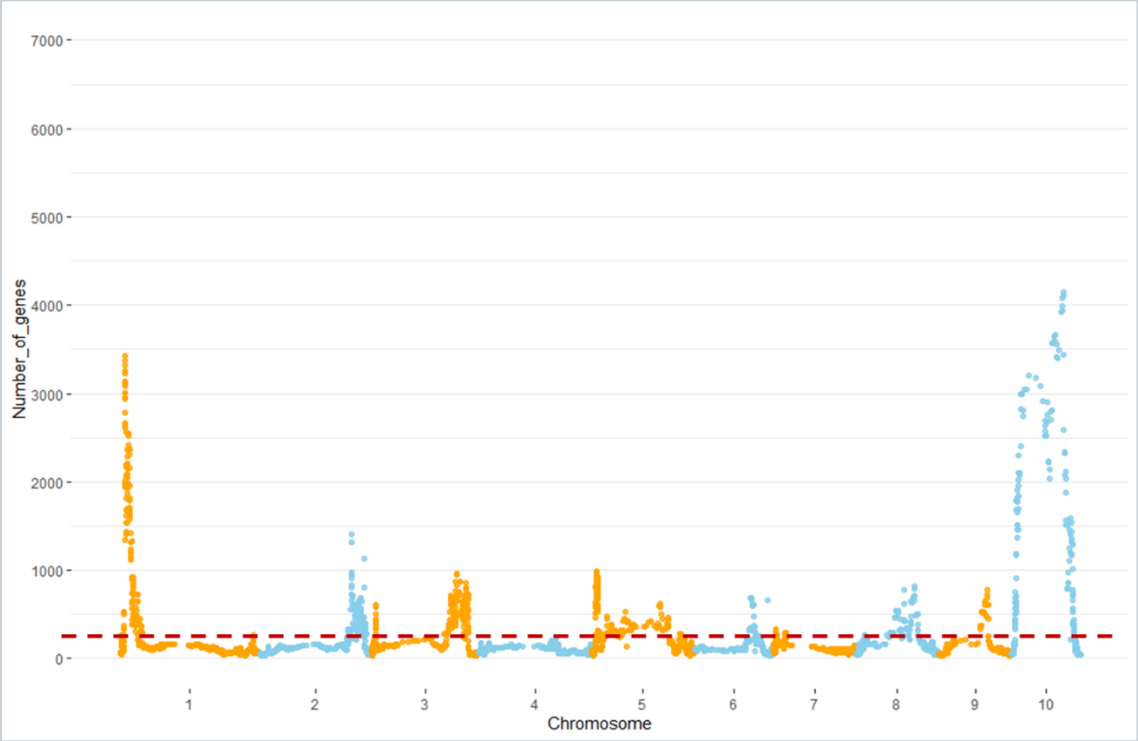


Figure 4-9. *Trans*-eQTL results using the difference (MIN) between the expression values of *Rp1-D21* and wildtype F1 progeny from the cross between H95:*Rp1-D21* and B73 x NC350 recombinant inbred lines (RIL). X-axis represents chromosome number; y-axis is the number of genes influenced by each SNP. Dotted red line represents the minimum number of genes a SNP must influence to be considered a hotspot.

Trans-eQTL analysis results from the mutant plants remained largely unchanged, in terms of the number of hotspots identified, when the analysis was repeated with transformed expression data. The number of genes influenced by two hotspots m1970 and m5923 fell below the threshold to 199 and 183 respectively in the MIN analysis (Figure 4-9), and to 195 and 129 respectively in the DIV analysis (Figure 4-10), hence 15 hotspots were identified in contrast to the MU analysis. Similarly, in the LOG analysis, the number of hotspots was reduced to 16 (Figure 4-11); the number of genes influenced by m5923 was reduced to 130, hence it was no longer considered a hotspot (Table 4-5). Taken together, these are further pieces of evidence that m5923 is background *trans*-eQTL and is not specific to the *Rp1-D21* response. The number of significant SNP-gene associations and the average number of genes per hotspot reduced in the analyses in the MIN, DIV, LOG analysis in comparison to the MU analysis. It appears the additional data transformation steps, helped remove a background *trans*-eQTL, resulted in fewer genes detected overall, but also

demonstrate that that *trans*-eQTL hotspots are not the result of *Rp1-D21* amplification of regulatory variation visible in the mutant. Rather it is consistent with HR-specific regulation driving changes in gene expression at a very large number of transcripts specifically in the *Rp1-D21/+* mutant siblings. All major *trans*-eQTL detected in the MU analysis were consistently detected across transformed data. The consistent discovery of the 16 major *trans*-eQTL hotspots across the different analyses with *Rp1-D21* plants provides further proof of the authenticity of these hotspots. Some discontinuity in the test values and number of affected genes are visible on chromosomes 2 and 10 in the MU, MIN, DIV, and LOG analyses could be the result of bad mapping or the presence of multiple QTL. A large proportion of the clustered diverse nucleotide-binding leucine-rich repeat (NLR) genes in maize are located on chromosome 10 (Y. Cheng et al., 2012). Because of this the "hotspot" that makes up the majority of that chromosome may be the result of multiple alleles at NLR loci across that chromosome that work with *rp1*, rather than a single QTL and unusual linkage.

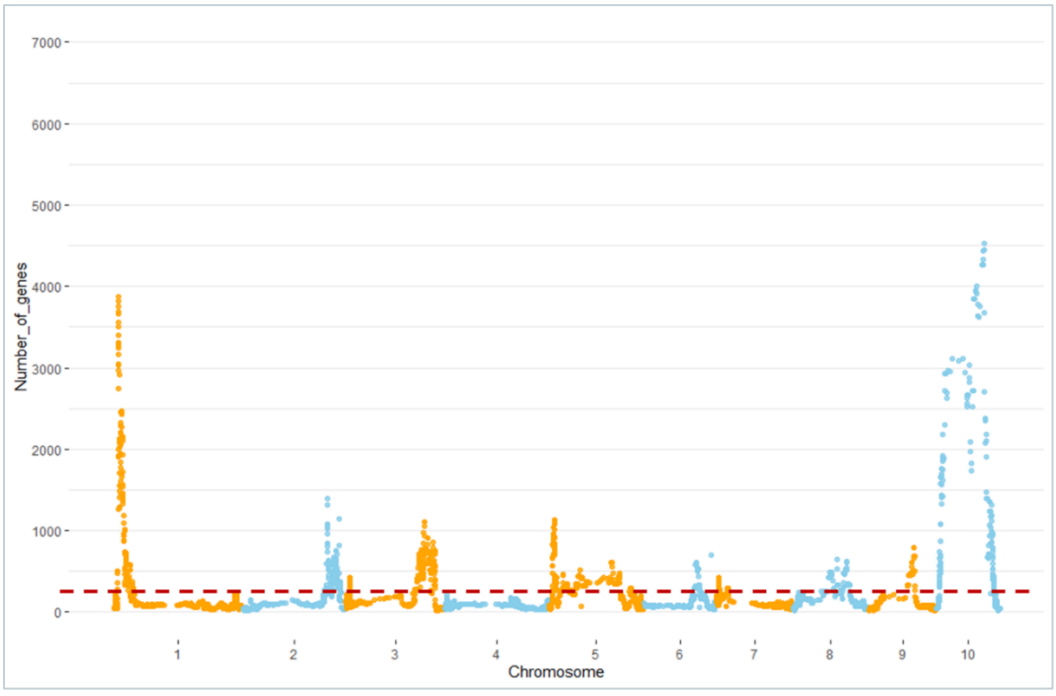


Figure 4-10. *Trans*-eQTL results using the ratio (DIV) between the expression values of wildtype and *Rp1-D21* F1 progeny from the cross between H95;*Rp1-D21*/+ and B73 x NC350 recombinant inbred lines (RIL). X-axis represents chromosome number; y-axis is the number of genes influenced by each SNP. Dotted red line represents the minimum number of genes a SNP must influence to be considered a hotspot.

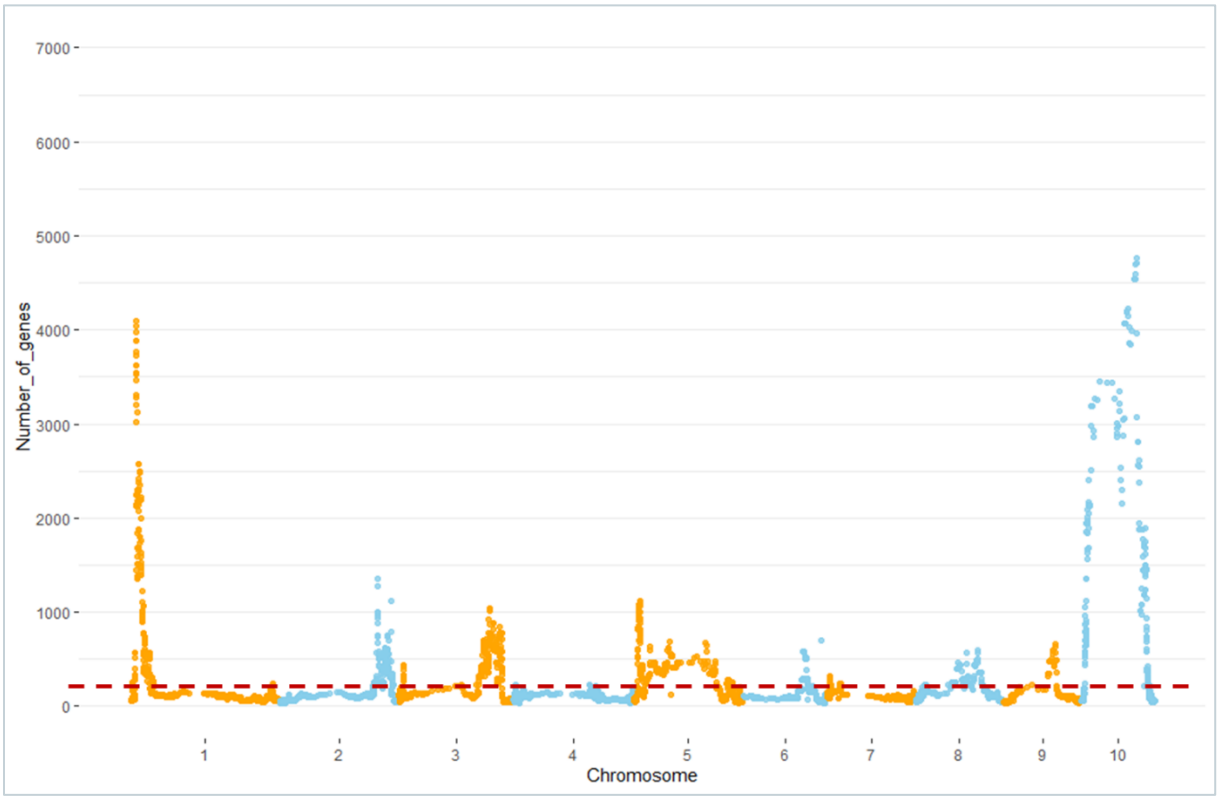


Figure 4-11. *Trans*-eQTL results using the log (LOG) of the ratio between the expression values of wildtype and *Rp1-D21* F1 progeny from the cross between H95;*Rp1-D21*/+ and B73 x NC350 recombinant inbred lines (RIL). X-axis represents chromosome number; y-axis is the number of genes influenced by each SNP. Dotted red line represents the minimum number of genes a SNP must influence to be considered a hotspot.

Intersection analysis between *trans*-eQTL hotspot target genes and wildtype versus mutant DEGs from the NH and BH backgrounds was carried out. A majority of the target genes at the mutant-specific *trans*-eQTL hotspots were also identified as DEG in the NH and BH F1 hybrid experiments (Table 4-6). As expected, due to the greater intensity of the *Rp1-D21*/+ phenotype in NH as compared to BH, the degree of *trans*-eQTL hotspot target gene enrichment was greater among the NH DEG than the BH DEG. As an example, of the 6077 genes under the control of m139, 5710 (93.9 %) were DE in the NH background as opposed to 2620 (43.1%) in the BH background (Table 4-6).

If these hotspots are the result of alleles affecting HR, or working at the top of the immune regulatory hierarchy, then their effects on expression should resemble the induction of HR. On the

other hand, if they alter a subset of the HR response, relatively few of the genes affected by HR should be altered by the alleles encoding these *trans*-acting hotspots. To test this, the direction of the effect of each hotspot on gene expression was compared with the effect of *Rp1-D21* on gene expression (chapter 3). Remarkably, the genes identified as DEG in response to *Rp1-D21/+* in the NC350 x H95 F1 hybrid families were coordinately regulated by 16 of the 17 *trans*-eQTL hotspots (Table 4-6 and Figure 4-12). More than 80% of the DEG were affected in the same direction in each case (Table 4-6). For example, at marker m139 on chromosome one, the NC350 affected the direction of expression in the same manner as *Rp1-D21* at 9919 of the 10307 DEG detected (96%) in NC350 x H95 F1 hybrid families (Table 4-6 and Figure 4-12). Consistent with all previous analyses, a repeat of this process using the DEG from BH uncovered an identical pattern (Table 4-6 and Figure 4-13). For example, at marker m139, 2545 of the 2757 genes (92%) that were DEG in BH were affected in the same direction by *Rp1-D21* and the NC350 allele at this *trans*-eQTL hotspot (Table 4-6 and Figure 4-13). The one exception to this pattern was the non-HR related hotspot on chromosome 8 which showed nearly equal split between which allele more closely resembled the *Rp1-D21* effect (4854 to 5453; Table 4-6). As indicated above, this hotspot was not involved in *Rp1-D21*-specific transcription responses. In 12 out of the 16 HR-specific hotspots it was the NC350 allele that was responsible for the enhanced response (Table 4-6). This indicates that the increased severity of *Rp1-D21* in NC350 is affected in part by the presence of alleles that strongly increase the transcription response to HR, process-wide. These massive *trans*-eQTL with effects on gene expression at nearly all HR-induced DEG indicate upstream master regulators of HR. The ability to detect this via eQTL, and thereby investigate the mode of action of these effects provides unprecedented detail to this natural variation in immune response.

Table 4-6. *Trans*-eQTL hotspots in mutants and their relationship with DEGs in both BH and NH backgrounds.

Hotspot	Chr.	Position	Gene count	DE NH	Not DE	DE BH	Not DE	NC350 or B73 enhancement	*Rp1-D21* in NH effect direction		*Rp1-D21* in BH effect direction	
									Effect of NC350 allele and *Rp1* the same	Effect of B73 allele and *Rp1* the same	Effect of NC350 allele and *Rp1* the same	Effect of B73 allele and *Rp1* the same
m139	chr1	11719235	6077	5710	367	2620	3457	NC350	9,919	388	2545	212
m925	chr1	291653311	312	250	62	84	228	NC350	9322	985	2401	356
m1590	chr2	201856260	1955	1810	145	1068	887	NC350	9725	582	2605	152
m1726	chr2	200891453	2620	2459	161	1306	1314	NC350	9850	456	2590	167
m2061	chr3	9761665	888	748	140	299	589	NC350	9434	873	2406	351
m1970	chr3	3838931	331	246	85	107	224	NC350	9608	699	2555	202
m2353	chr3	186138473	1771	1617	154	824	947	B73	603	9704	187	2570
m3611	chr5	10198879	1085	886	199	342	743	NC350	8865	1442	2209	548
m3813	chr5	148816139	740	506	234	189	551	NC350	8328	1979	1999	758
m3972	chr5	201868522	388	280	108	122	266	B73	1012	9295	229	2528
m4454	chr6	124533055	1444	1277	167	627	817	NC350	9759	547	2561	195
m4908	chr7	5928090	653	538	115	217	436	B73	775	9532	261	2496
m5022	chr7	27435303	585	417	168	160	425	B73	959	9348	299	2458
m5707	chr8	21567679	532	371	161	100	432	NC350	9122	1185	2318	439
m5923	chr8	166652339	304	127	177	60	244	Not Rp1-specific	4854	5453	1173	1584
m6457	chr9	105751521	1644	1260	384	509	1135	NC350	9080	1227	2323	434
m7076	chr10	112662547	5700	5396	304	2772	2928	NC350	9978	329	2608	149

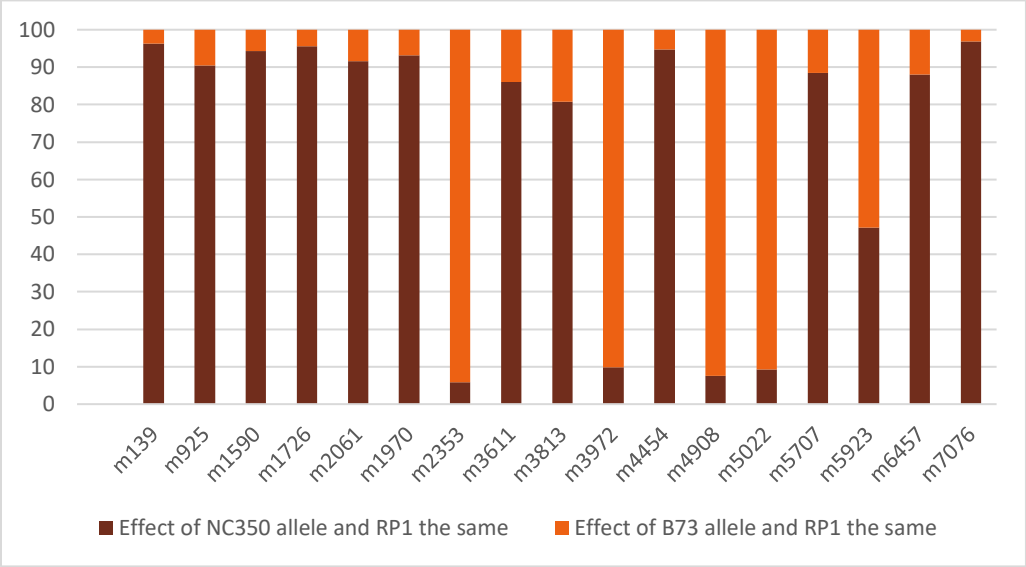


Figure 4-12. Effect direction of differentially expressed genes (DEGs) between wildtype and mutant F1 progeny from cross between NC350 and H95;*Rp1-D21*/+ (NH) by *trans*-eQTL hotspots. *Trans*-eQTL analyses were performed in *Rp1-D21* F1 progeny from the cross between H95*;Rp1-D21*/+ and B73 x NC350 recombinant inbred lines (RIL). Dark brown denotes proportion of DEGs between wildtype and mutant NH F1, for which the NC350 allele has the same effect direction as *Rp1-D21*. Orange represents the proportion of DEGs, between wildtype and mutant NH F1, for which the B73 allele has the same effect as *Rp1-D21*. The numbers used in this chart are drawn from Table 4-6.

Figure 4-13. Effect direction of differentially expressed genes (DEGs) between wildtype and mutant progeny from cross between B73 and H95;*Rp1-D21*/+ (BH) by *trans*-eQTL hotspots. *Trans*-eQTL analyses were performed in *Rp1-D21* F1 progeny from the cross between H95*;Rp1-D21/+* and B73 x NC350 recombinant inbred lines (RIL). Green denotes proportion of DEGs between wildtype and mutant BH F1, for which the NC350 allele has the same effect direction as *Rp1-D21*. Yellow represents the proportion of DEGs, between wildtype and mutant BH F1, for which the B73 allele has the same effect as *RP1-D21*. The numbers used in this chart are drawn from Table 4-6.

4.4 Conclusions

Overall, this experiment uncovered an avalanche of expression variation, as evidenced by the QQ-plots, and presents powerful tools for describing response to disease signaling. *cis*-eQTL, or at least local eQTL, were detected at about a third of all genes. Given the size of the population and variation in maize, such a proportion is perhaps not surprising. Yet, not every gene within the genome will be polymorphic between NC350 and B73. The extent of enrichment of DEGs within the genes under the influence of *cis*-eQTL is also not unexpected. We do not expect every DEG to be polymorphic between the pair of genotypes. However, a couple of questions arise. Should we expect more polymorphisms within the promoters of HR-induced genes? Is selection diversifying expression effects for disease-responsive loci? This appears to be the case, and our results provided modest evidence in support of this. More *cis*-eQTL genes were detected at genes that were differentially expressed than would be expected by random chance for BH, NH, and BNRIL backgrounds. The strength of the *Rp1-D21* effect increases as the experiment becomes better

powered, with the massive replication and large number of DEG detected in the BNRIL having the strongest signal of overlap with the eQTL experiment. This suggests that this is not the result of the most responsive genes being differentially subject to selection for expression polymorphisms nor an effect disproportionately affecting the most abundant genes. In addition, the greater signal of overlap between the stronger NH and BNRIL experiments and the eQTL results argues against the overlap being an artifact of a lower-powered experiment. Were this an artifact of low power, more abundant genes are more often differentially expressed, and the enrichment should have favored the BH and NH comparisons. Rather, thousands of genes appear to be contributing to HR severity, some with very tiny effects. Likewise this cannot really be a case of "Omnigenetics" (Boyle et al., 2017) in which the entire genome affects variation in a trait because enrichment is observed. Still, the large number of genes that are present in this overlap present an opportunity for breeding, as selection should enrich for disease modulating alleles, and a unique challenge, as so many genes are involved that segregation and linkage will never allow selection to accumulate all of the "favorable" alleles. How do you select for thousands of loci within a breeding program? Maize breeding for disease resistance is extremely facile and has been very successful, consistent with selection favoring beneficial alleles.

In this context, the *trans*-eQTL results are quite extraordinary. The number of genes under the control of the hotspots were often an order of magnitude more than has been reported in previous eQTL studies in maize. The allele effect directions conformed to the expectation for NC350 encoding alleles that affect *Rp1-D21* sensitivity. This was one possible explanation for the greater differential expression and it would appear that these hotspots are contributed by the lesion-increasing loci in this accession. Indeed, 12:4 of the hotspots are due to NC350 contributing the HR-enhancing allele, consistent with a the lesion enhancing QTL study by (Olukolu et al., 2016) that used the same RIL population. Strikingly, the whole of chromosome 10 encodes markers with low p-values for a very large number of genes. One possible explanation for this is the fact that chromosome 10 is replete with many NLR gene clusters.

Each *Rp1-D21*-specific hotspot induced nearly all *Rp1-D21* responsive genes. Had these been encoded by downstream transcription factors, I would expect that the subset of promoters bound by that specific transcription factor would be affected by the trans-eQTL hotspot. Instead, the entire *Rp1-D21* response was either turned up or down. This fits a model where these hotspots act high up in the HR regulatory hierarchy, for instance modulating the severity of the lesion

response. In addition, the 16 *Rp1-D21* specific hotspots require this allele to be observed, indicating that *Rp1-D21* is epistatic to the variation encoded by the eQTL. I cannot say, however, whether the *trans*-eQTL acts downstream or upstream of *Rp1-D21*. Rather, they appear to be acting in concert with *Rp1-D21* to modify HR severity. Previous work has demonstrated exactly this for multiple proteins that are both induced by *Rp1-D21* and negatively regulated *Rp1-D21*-mediated HR via protein-protein interactions (Wang & Balint-Kurti, 2016; Wang et al., 2015).

There are two artifacts present in the data that are worthy of comment. The first are the displacement of SNP from hotspots and patterns of hotspots not consistent with linkage. This suggests there are some inconsistencies between the SNP-chromosome map and reality. These could be the result of either NC350 vs B73 rearrangements or bad SNP positions present in the marker data that was lifted over to the B73 v4 reference used in these analyses. The second is the very large number of genes affected by *trans*-eQTL hotspots and the problem of distinguishing local *trans*-eQTL from *cis*-eQTL. Since I detected an unprecedented number of *trans*-eQTL in my analysis it seems likely that I will also have a proportionally greater number of local *trans* eQTL currently annotated as "*cis*" in my analysis. The most efficient way to distinguish these is molecularly via an analysis of allele-specific expression. This is the subject of the final chapter of my dissertation.

CHAPTER 5. TESTING THE MOLECULAR MECHANISM OF *CIS*-EQTL THROUGH ALLELE SPECIFIC EXPRESSION AS VALIDATION OF *CIS*-REGULATORY VARIANTS AFFECTED BY *RP1-D21*/+ INDUCED HR

5.1 Introduction

Allele-specific expression (ASE), the differential accumulation of mRNA from alleles at a locus (Knight, 2004; Wittkopp et al., 2004) is a powerful approach to understand the genetic basis for gene expression variation. ASE can only result from the action of *cis*-acting variation since each allele only affects the expression of the copy of the gene residing on the same physical chromosome. This difference can be detected by measuring the relative expression of the parental alleles in a heterozygous individual (Figure 5-1). *Trans*-eQTL, on the other hand, do not influence the expression of their target genes in an allele-specific manner. Some combination (e.g. *cis* x *trans* interactions) can result in *trans* mediated observation of ASE, but only in the presence of an underlying *cis*-regulatory difference distinguishing the alleles at the target. This is because the diffusible element (e.g., transcription factor) whose function is altered by the eQTL is equally available to both alleles of the target gene in the heterozygote (Albert et al., 2018; Albert & Kruglyak, 2015b; Castel et al., 2015) .

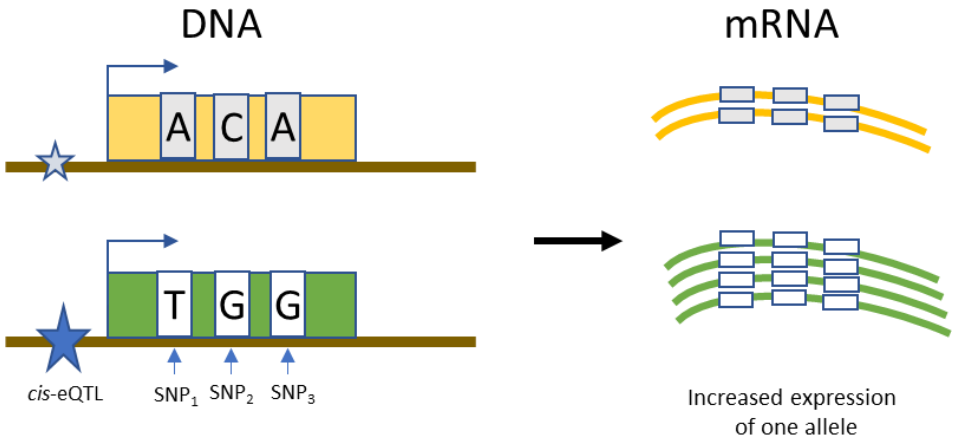


Figure 5-1. Allele-specific expression effects from *cis*-regulatory variants. Heterozygous *cis*-eQTL generates transcript-level differences between the two haplotypes which is detectable by counting of reads contained in the SNP position. SNP, single nucleotide polymorphism; *cis*-eQTL, *cis*-acting expression quantitative trait locus.

Sequencing-based expression analyses can be used to detect ASE in the presence of variation in the transcript sequences. Analyses assign RNA-seq reads spanning heterozygous single nucleotide variants to each parental allele and evaluate statistical significance of any imbalance in reads from each allele with a binomial test (Castel et al., 2015). Since the detection of eQTL in traditional mapping experiments (See Chapter 4) considers markers linked to a gene, a *cis*-eQTL may be confounded by local QTL encoding trans-regulatory determinants of expression. For example, in my work in Chapter 3, all significant estimates of expression variants within a 1 Mb region of the affected gene were considered as *cis*-eQTL. While this eliminated 99.95% of the genome, it is formally possible that the variation detected may not exert influence in *cis*, but rather be a linked, or local, *trans*-acting factor. Local trans-acting effects can be distinguished from true *cis*-eQTL because only the latter will produce allele-specific expression. ASE analysis is a simultaneous test of mechanism and can be used to validate *cis*-eQTL. Validated eQTL and their genes can be prioritized for further studies or use in plant breeding

Using RNA-seq to assess ASE presents some interesting computational analysis bottlenecks. Chiefly that biased mapping of reads to the allele that resembles the genome reference can result in deviation of read numbers from the trivial null hypothesis of equal reads from each allele. This is often referred to as reference bias. Transcript abundance estimation from RNA-seq involves a number of steps. First, reads are assigned to chromosomal regions showing the highest sequence similarity followed by a comparison to an annotated genomic reference. A gene's expression level is subsequently quantified by counting the number of reads mapped to it (Mortazavi et al., 2008). Since each polymorphism between an allele and reference decreases the mapping score, reads harboring variant positions are frequently imprecisely aligned and more frequently fail to meet minimum mapping criteria (Degner et al., 2009; Salavati et al., 2019; Stevenson et al., 2013). When reads are assigned to genotype and counted, this leads to systematic bias towards the reference allele and consequently high false-positive rates for detecting ASE (Degner et al., 2009).

Sequencing both parental genomes and aligning reads from the F1 to each genome separately is one robust approach to eliminating this bias (Shen et al., 2013). This provides an estimate of the counts obtained when a 1:1 mix of reads from the two parental alleles are processed by the analytical pipeline for ASE. This approach is costly for organisms with large genomes, as reliable estimates of reads proportions requires sequencing to high depth. Alternative methods rely

on phased genotype data to supplement or modify the reference (Vijaya Satya et al., 2012) or performing mapping and alignment with SNP-tolerant software such as GSNAP (Wu & Baldwin, 2010) in an effort to reduce the effect of polymorphisms on this bias. These strategies may work for organisms with relatively few polymorphisms between individuals, like humans. Attempting only to minimize the effects of variants in a species with greater intraspecific variation, such as maize, might be challenging as the number of likely haplotypes scales exponentially with a number of variants (Stevenson et al., 2013). A third strategy to minimize reference allele bias uses pseudo reference sequences with N-masking at the heterozygous positions and has been benchmarked to be very robust to the challenges elaborated above (Degner et al., 2009; Salavati et al., 2019; van de Geijn et al., 2015). I used such N-masked references in my DEG analyses (Chapter 3) and eQTL experiments (Chapter 4) to improve reads mapping. The improvement in reads mapping is exactly what we would expect if they were reducing reference bias.

The relatively few studies that have described integrated *cis*-eQTL and ASE analyses have found overlaps between the loci identified but also notable differences. For instance, in their investigation into expression variation controlling meat quality traits in pigs, Liu et al 2020 identified a 12.4% overlap between genes discovered with the two approaches. Specifically, among the 2098 genes targeted by *cis*-eQTL only 540 were shared with the 2253 genes that showed significant allelic imbalance. Another study that leveraged ASE and *cis*-eQTL analyses to identify determinants of gene expression variation behind complex traits in cattle showed only about 50% overlap (Khansefid et al., 2018). Differences in detection power between the two analysis methods have been cited as an underlying reason for this observed limited overlap (Khansefid et al., 2018). But it is possible that a larger than expected number of local *trans*-eQTL are encoded in genomes.

In maize, genome-wide ASE analyses using RNA-seq have been carried, but none in combination with eQTL. ASE has contributed towards elucidating the genetic architecture behind complex phenomena such as heterosis (Z. Li et al., 2021; Springer & Stupar, 2007; Wan et al., 2022). ASE has also enabled a deeper understanding of genetic control underpinning response to abiotic stress in maize (Waters et al., 2017). With recent advancements in single-molecule sequencing technologies, techniques have been developed to use long reads in ASE analysis. These have been used to enhance our understanding of parent of origin effects within different tissues of reciprocal hybrids derived from a temperate x tropical maize line cross (T. Wang et al., 2020).

I reanalyzed the data from the F1 hybrids from the B73 X H95 cross and NC350 X H95 cross for ASE. This allowed me to detect *cis*-regulatory differences between the alleles of these parent-pairs in the presence and absence of HR as induced by *Rp1-D21*. Comparison of these data to the cis eQTL identified in chapter 4 provides remarkable validation of a subset of these QTL. In addition, in an effort to make a comparison between NC350 and B73 I created a common-reference ASE analysis method. While successful at validating a number of *cis*-eQTL, all experiments that included B73 parentage displayed substantial reference bias that was not fully corrected by using N-masked reference.

5.2 Methods

5.2.1 SNP calling

Variant calling from H95 and NC350 whole-genome shotgun data was carried out using the standard pipeline described in Chapter 2 to generate SNPs for anonymized reference creation. In summary, paired reads were aligned to B73RefGen_v4 (Jiao et al., 2017) using BWA-MEM (H. Li & Durbin, 2009; R. Li et al., 2009). SAMtools *flagstat* command (H. Li et al., 2009) was then used to retrieve alignment quality statistics including number of QC-passed/failed reads, number of properly paired reads, as well as a number of singletons. This was followed by running of the SAMtools rmdup command on alignment files to exclude PCR duplicates which are amplification artifacts that could have been introduced during library construction. Combined SNP and small indel discovery was performed with BCFtools *view* command using genotype likelihoods computed from de-duplicated alignment files with SAMtools *mpileup*. First round of variant filtering was carried out using *varFilter* command of the vcfutils.pl script with the *– D100* option which excludes polymorphisms by the following criteria: coverage less than 2 reads but not exceeding 100 reads, root mean square quality less than 10, as well as variants located within three bases of a gap (H. Li et al., 2009). Indels were then removed from the list of variants with VCFtools to create a VCF with only SNPs. This file was further processed using SnpSift (Cingolani, Patel, et al., 2012) to only retain homozygous SNPs with minimum phred-scale quality of 20 and a read depth of at least 4.

5.2.2 B73 x H95*;Rp1-D21*/+ ASE analysis

The RNA-seq reads used in differential gene expression analysis (see Chapter 3) from B73 x H95*;Rp1-D21*/+ (BH) F1 hybrids were used to examine ASE. These single-end reads were derived from six individuals (3 biological replicates per wildtype or mutant). I used the H95 SNPs generated from variant calling from whole genome sequencing to create an anonymized reference by converting the variant positions within AGPv4 to ambiguous (N) bases. Read mapping to the SNP-anonymized reference was conducted using STAR aligner (Dobin et al., 2013) to produce alignment files which were further processed with Picard tools to assign read group information, sort, mark duplicates, and create indices. SAMtools *merge* was then used to combine alignment files from 3 biological replicates prior to the allele counting step.

ASEReadCounter within the GATK suite of tools (Depristo et al., 2011; McKenna et al., 2010) was used to compute B73 (reference) and H95 (alternative) allele counts at each bi-allelic heterozygous H95 variant within the merged alignments. The settings used for counting were as follows: minimum depth of 10, mapping quality of 10 and a base quality of at least 2. Options to count reads only once and filter out duplicate reads were also specified. Allele counts per gene were computed by adding allele counts for SNPs within a gene and used to estimate ASE via a binomial test carried out under the null hypothesis that each allele is expressed equally (Figure 5-2). P-values were adjusted using the Benjamini-Hochberg procedure to control for multiple tests (Benjamini and Hochberg, 1995). Significance was determined at a False Discovery Rate (FDR) of 5%. To identify genes that co-occur in ASE and *cis*-eQTL experiments, significant ASE genes from mutant BH analysis were overlapped with the significant *cis*-eQTL target genes from the mutant RIL analysis. A similar comparison was carried out between wild-type BH significant ASE genes and significant *cis*-eQTL from the wild-type RIL experiment.

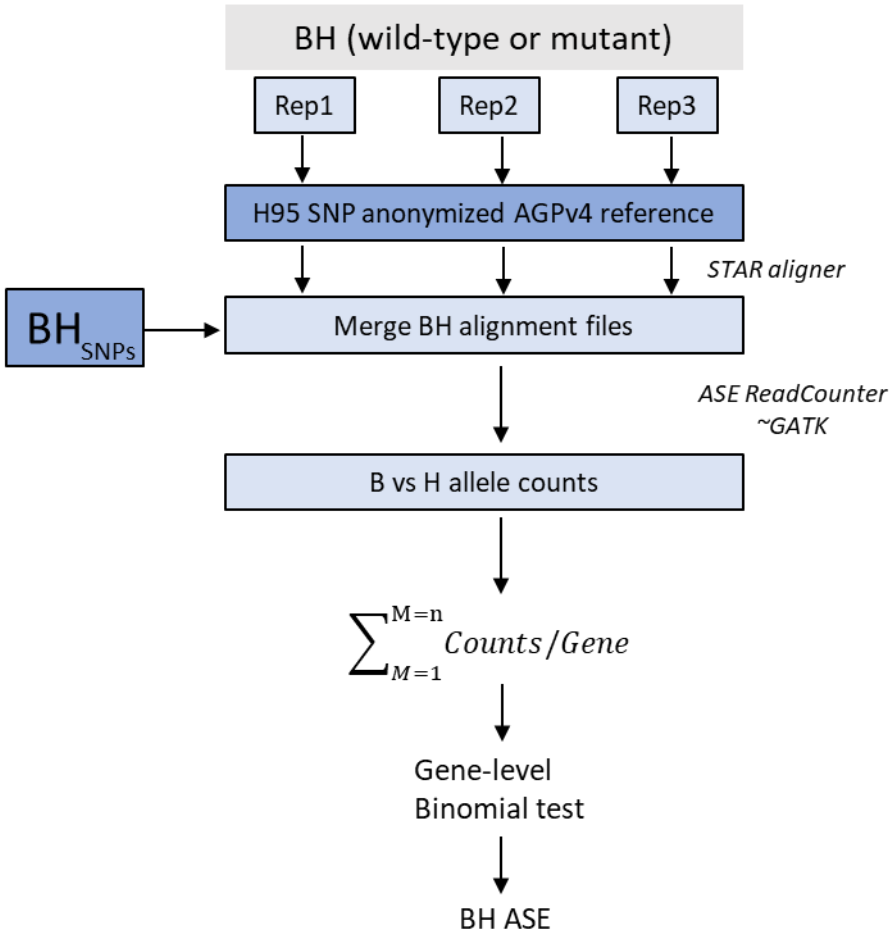


Figure 5-2. Overview of ASE analysis in B73 x H95:*Rp1-D21*/+ (BH) F1 hybrids. RNA-seq reads are aligned to a H95 SNP anonymized AGPv4 reference genome. H95 homozygous SNPs were used to generate read counts at each SNP position from merged BH alignment files generated by mapping RNA-seq reads to H95-anonymized AGPv4 reference. Allele read counts per gene were then computed and used to assess ASE via a binomial test.

5.2.3 NC350 x H95*;Rp1-D21*/+ ASE analysis

Patterns for ASE within the NC350 x H95*;Rp1-D21*/+ (NH) background were assayed by first creating an anonymized AGPv4 reference with H95 and NC350 SNP positions converted to ambiguous bases. RNA-seq reads produced from six F1 hybrid individuals (3 biological replicates per wildtype or mutant) that were used to perform differential gene expression analysis in the NH background (see Chapter 3) were again used for this analysis. Single-end read alignment and post-alignment processing were conducted as described above for the BH analysis.

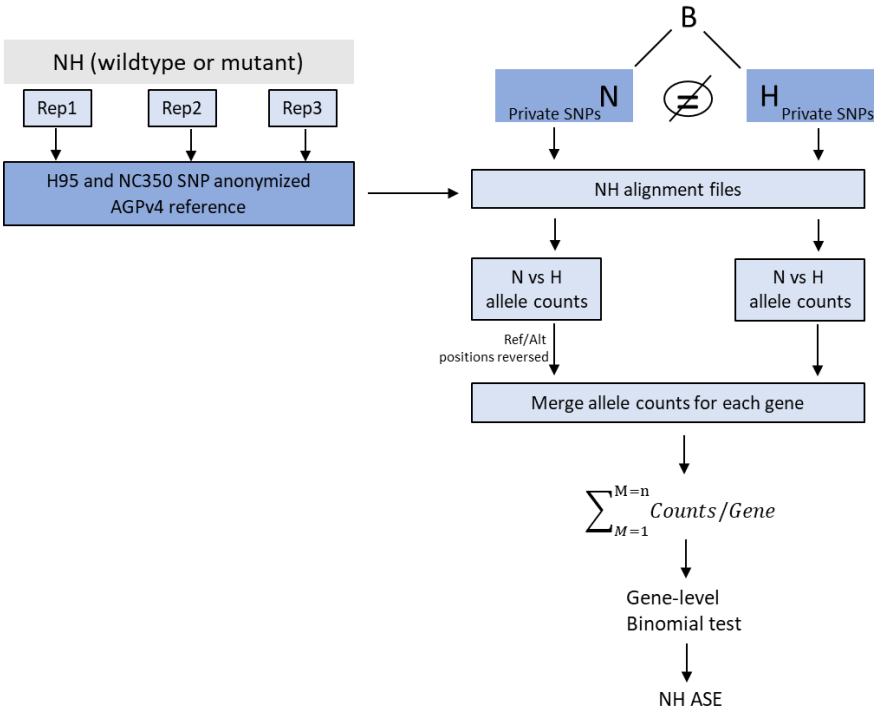
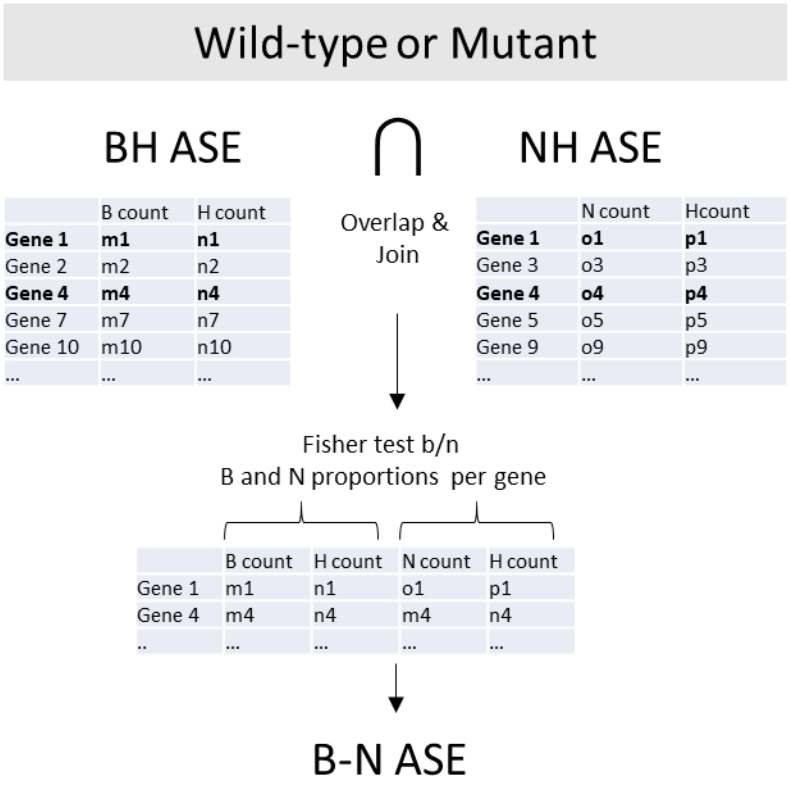


Figure 5-3. Schematic showing an overview of NC350 x H95:*Rp1-D21*/+ (NH) ASE analysis. B73:NC350 (BN) and B73:H95 (BH) homozygous SNPs were compared to exclude common SNPs from each. Private BN and BH SNPs were separately used to generate allele counts from NH alignment files produced by mapping RNA-seq reads to H95 and NC350 anonymized AGPv4 reference genome. Allele read counts per gene were then computed and used to assess ASE via a binomial test.

A modified approach was, however, implemented for allele counting as follows. The H95 and NC350 SNPs discovered from the variant calling step were compared to exclude common positions. The non-overlapping SNP positions were then fed into ASEReadCounter to count reads per allele using the same parameters specified above. Reference and alternate allele read counts in the NC350-private list were reversed before being merged with the counts from H95. The combined list was then sorted and processed to produce allele counts per gene. A binomial test between allele read counts was used to measure ASE at a significance level of 0.05 FDR (Figure 5-3). Genes showing ASE from wildtype and mutant experiments were overlapped with *cis*-eQTL target genes from respective eQTL experiments to isolate co-occurring genes.

5.2.4 Relative ASE by comparison to a common reference

Using H95 (H) as the control, B73-NC350 (B-N) ASE analysis was conducted by first comparing ASE genes from BH and NH mutant experiment to identify overlapping genes. Allele read counts for these common genes were retrieved and used to create a combined count file. Fisher-exact test was subsequently performed between the pairs of reference and alternative allele counts for each gene (Figure 5-4). Nominal p-values were FDR corrected with the Benjamini-Hochberg transformation; significance was set at the 5% threshold. The analysis was repeated for significant ASE genes identified from the BH and NH wildtype experiment.



	B count	H count
Gene 1	**m1**	**n1**
Gene 2	m2	n2
Gene 4	**m4**	**n4**
Gene 7	m7	n7
Gene 10	m10	n10
...	...	...

	N count	Hcount
Gene 1	**o1**	**p1**
Gene 3	o3	p3
Gene 4	**o4**	**p4**
Gene 5	o5	p5
Gene 9	o9	p9
...	...	...

	B count	H count	N count	H count
Gene 1	m1	n1	o1	p1
Gene 4	m4	n4	m4	n4
..	...	...	...	...

Figure 5-4. Overview of the B73-NC350 (B-N) ASE analysis. Significant ASE genes from BH and NH were compared to identify common genes. Allele counts for overlapping genes were joined to produce a single file. A Fisher-exact test was then conducted between the pairs of reference and alternative allele read counts for each overlapping gene.

5.3 Results

5.3.1 B73 x H95*;Rp1-D21*/+ ASE analysis

Allelic specific expression (ASE) was assessed between the B73 and H95 alleles in the progeny of B73 x H95*;Rp1-D21*/+ (BH). These F1 hybrids were either heterozygous for the wildtype alleles at *rp1* from B73 and H95 or carried the *Rp1-D21* autoactive HR-inducing allele and the wildtype B73 *rp1* allele. The RNA-seq reads used in the differential gene expression analysis (see Chapter 3) consisting of single-end reads from six individuals (3 biological replicates per wildtype or mutant phenotype) were used for this analysis. An anonymized reference, in which all known polymorphisms between H95 and B73 were converted to ambiguous bases, was used to reduce bias towards B73 alleles during alignment. Reads were mapped to this H95 single nucleotide polymorphism (SNP)-anonymized AGPv4 reference genome using the STAR splice-aware aligner (Dobin et al., 2013). This produced alignment files that were processed with Picard tools to add read group information, sort, mark duplicates, and create indices. Alignment files from three biological replicates were then merged to boost read depth and used as input for the allele count retrieval step. This step involved the use of ASEReadCounter within the GATK suite of tools (Depristo et al., 2011; McKenna et al., 2010) to count reference and alternative allele reads at each bi-allelic heterozygous H95 variant with minimum depth of 10, mapping quality of 10 and a base quality of at least 2. Each read fragment was counted only once and duplicate reads are filtered out. Removal of duplicates and counting reads only once were carried out so that the result from the downstream process of aggregating read counts per SNP to produce haplotype counts for each gene would not be inflated.

Reads from each sample ranged from 17,093,161 to 24,900,357 and were aligned to a specially prepared version of maize genome (version 4). An average of 74.28% aligned to a single location in the anonymized reference (Table 5-1). To prepare SNPs for allele counting the initial H95 SNP call set were filtered to retain only SNPs located within genes. These genic SNPs totaling 1,793,959 allowed assessments of ASE in 13,954 (30.3% of known maize genes) and 15,470 genes (33.7% of known maize genes) in plants showing wildtype and the lesions induced by *Rp1-D21* respectively.

Table 5-1. Results from aligning B73 x H95;*Rp1-D21*/+ (BH) RNA-seq reads to H95-anonymized AGPv4 reference genome.

Sample	Background	Phenotype	Raw reads	Uniquely mapped reads	Uniquely mapped reads %
BHwt_rep1	B73 x H95*;Rp1-D21*	wildtype	17,093,161	12,974,886	75.91
BHwt_rep2	B73 x H95*;Rp1-D21*	wildtype	22,507,024	16,992,485	75.50
BHwt_rep3	B73 x H95*;Rp1-D21*	wildtype	22,991,417	17,205,304	74.83
BHmu_rep1	B73 x H95*;Rp1-D21*	*Rp1-D21*	24,900,357	18,600,871	74.70
Bhmu_rep2	B73 x H95*;Rp1-D21*	*Rp1-D21*	21,765,731	15,956,516	73.31
Bhmu_rep3	B73 x H95*;Rp1-D21*	*Rp1-D21*	23,434,274	16,738,254	71.43

Allele counts for variants within a gene were summed up to produce reference and alternative allele counts per gene. To quantify ASE, a binomial test was conducted under the null hypothesis that each allele is equally expressed. Nominal *P-values* were corrected using Benjamini-Hochberg adjustment, and genes were considered significant ASE at False Discovery Rate (FDR) of 0.05. Of the genes tested, 8590 in wildtype (61.5%) and 10656 in mutant (68.9%) displayed significant allelic imbalance (Table 5-2 and Figure 5-5). A majority of the genes showing ASE exhibited bias in the direction of B73; 62.3% for wildtype and 59.7 % for mutants (Table 5-2).

Table 5-2. Number of genes showing allelic imbalance and direction of bias. Within the B73 x H95;*Rp1-D21*/+ background.

Sample	B73-bias	H95-bias	Sum
BHwt	5,350	3,240	8,590
Bhmu	6,363	4,293	10,656

The genes identified as significantly imbalanced from a 1:1 ratio were compared to the differentially expressed genes (DEGs) described in Chapter 3. To make the closest comparison possible, DEG and ASE sets were compared from analyses carried out on the same RNAseq data. The DEG from the comparison of B73 x H95 wildtype to B73 x H95*;Rp1-D21/*+ (aka BH) F1 hybrids were intersected with the genes exhibiting ASE. Among the genes identified as showing

allelic imbalance in the wildtype samples, only 922 (10%) were differentially expressed (DE) between wild type and mutant (Figure 5-6). This was more than the 786 expected by random chance (χ^2 p-value 1.15E-08). In the mutant samples 1281 (12%) of the ASE genes showed differential expression; several hundred more genes than the 964 expected by random chance (χ^2 p-value 1.06E-36) (Figure 5-6). This greater than expected overlap, though modest, indicates some *cis* by *trans* interaction (Becker et al., 2012; Yang et al., 2017) where the transcriptional mechanisms affected by HR interact with *cis* regulatory differences between the two alleles.

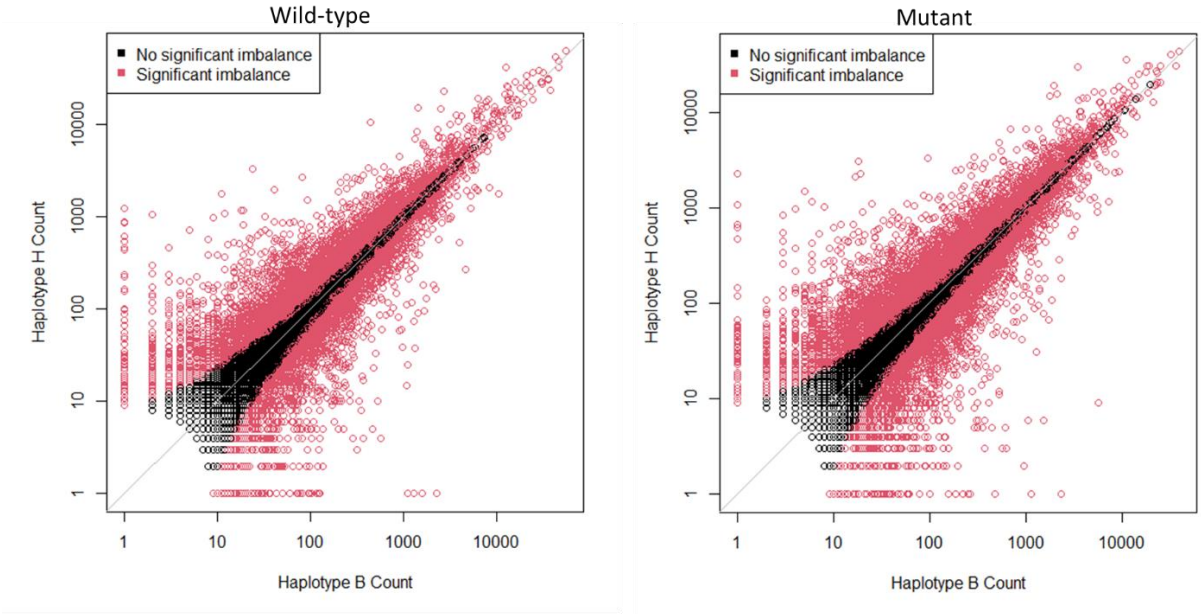


Figure 5-5. ASE analysis results from B73 x H95;*Rp1-D21*/+ (BH) hybrid F1 plants showing wildtype (left) or *Rp1-D21* (right) phenotype. The x-axis represents allele counts for B haplotype for each gene tested whilst the y-axis denotes the counts for H haplotype. Red dots are genes showing significant allelic imbalance (FDR ≤ 0.05), whereas black dots represent genes with balanced expression.

The genes identified as exhibiting ASE were also compared to the *cis*-eQTL identified in Chapter 4. As *cis*-eQTL, defined in Chapter 3, either result from *cis*-regulatory differences or from local trans effects, ASE can be used a validation and molecular mechanism test. A greater proportion of *cis*-eQTL genes assessed for ASE were validated; 3729 (66.3 %) for the wildtype and 4189 (72.2%) for the mutants (Figure 5-7). Since the regulatory variants affecting these genes act in an allele-specific manner they are in fact *cis*-acting and can be considered the true *cis*-eQTL.

Non-validated *cis*-eQTL could fall into one of two categories. The first group comprises local regulatory variants that are *trans*-acting but were detected as *cis*-eQTL by virtue of the being located within 1 Mb of their target gene. The second category are true *cis*-eQTL that could not be detected in ASE due to a lack of power to detect. To investigate this further, the magnitude of gene expression was compared among groups of genes common to ASE and *cis*-eQTL analyses and those that are unique to either methodology. Log_2-normalized gene expression counts were averaged across the 3 replicates (wildtype or mutant). Two pairwise t-tests were then carried out between ASE-validated *cis*-eQTL genes versus non-validated *cis*-eQTL genes and ASE genes not found in *cis*-eQTL genes. This confirmed that indeed ASE-validated *cis*-eQTL genes had significantly higher expression than their non-validated counterparts in both wild type and mutant (p-value < 2.22E-16 in both) (Figure 5-8). This suggests ASE analysis may have less detection sensitivity under lower gene expression and confirm that while some *cis*-eQTL may rather be local *tran*-eQTL, others could not be validated because ASE may just lack the power to detect them due to their low expression.

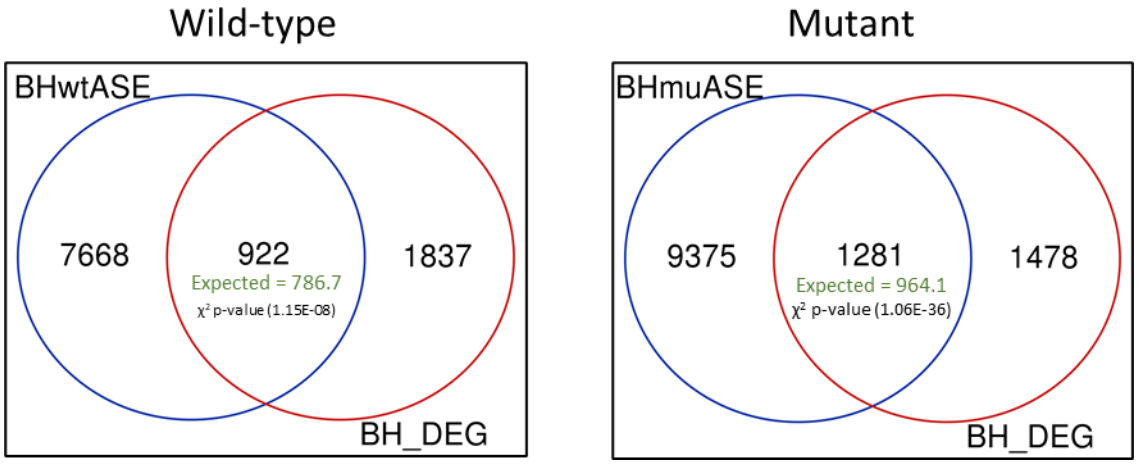


Figure 5-6. Overlap between genes identified from ASE analysis in B73 x H95;*Rp1-D21*/+ (BH) hybrid F1 plants showing wildtype (left) or *Rp1-D21*/+ (right) phenotype versus differentially expressed genes between plants showing wildtype or *Rp1-D21* phenotype.

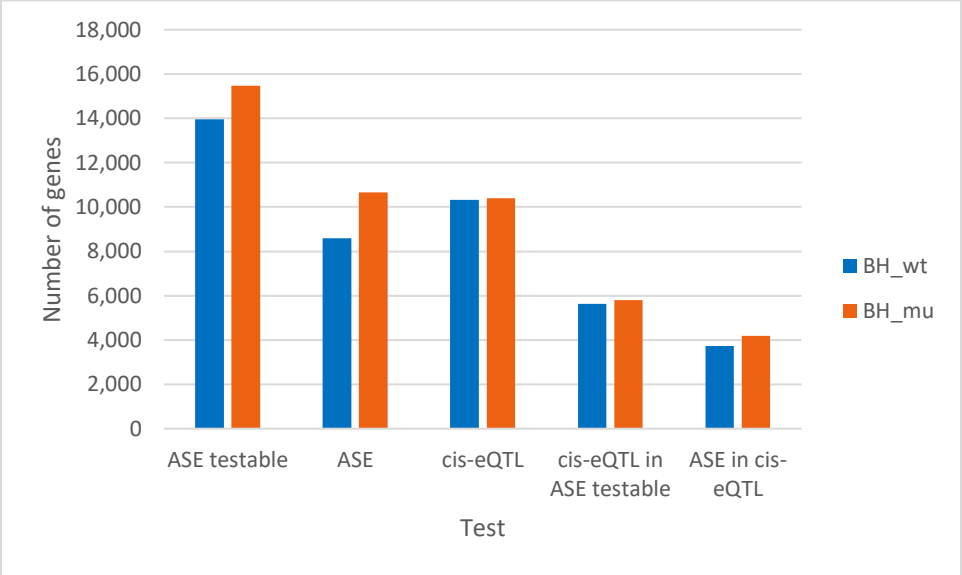


Figure 5-7. Integrating results from ASE in B73 x H95;*Rp1-D21*/+ (BH) hybrid F1 and *cis*-eQTL analyses. Orange bars indicate results from plants showing *Rp1-D21* phenotype, and blue bars are results from plants showing wildtype phenotype. X-axis shows the different comparisons carried out whereas the y-axis shows the number of genes identified.

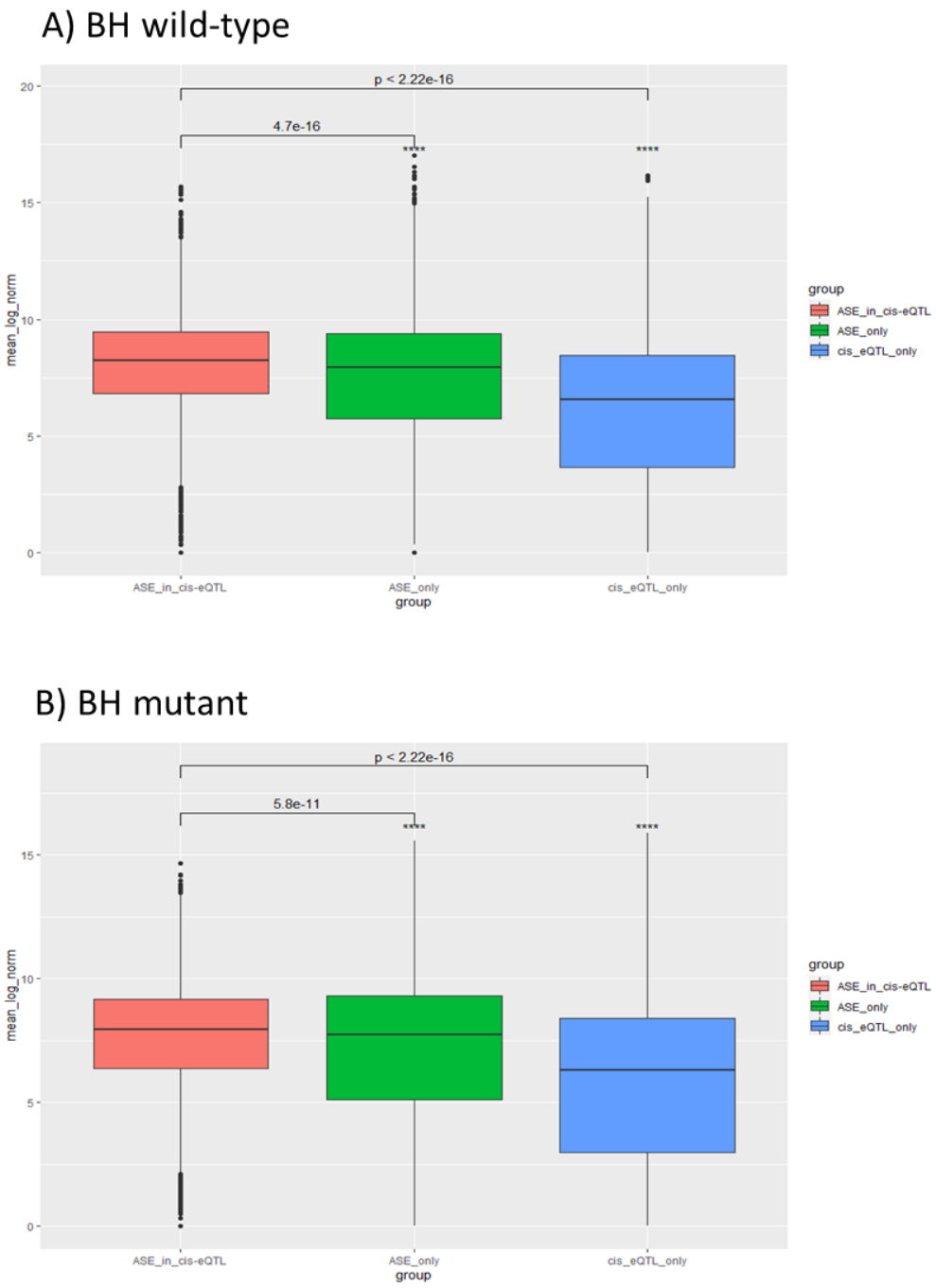


Figure 5-8. Comparison of gene expression among *cis*-eQTL and ASE genes in B73 x H95;*Rp1-D21* (BH) hybrid F1 plants displaying wilt-type (A) or RPI-D21 (B) phenotype. X-axis represents unique or overlapping genes from ASE and eQTL analyses. Y-axis is the mean of log_2-normalized expression values across three replicates. ASE_in_*cis*-eQTL group denote significant genes from ASE and *cis*-eQTL analyses; ASE_only represent genes significant in ASE but not significant in *cis*-eQTL analysis; *cis*-eQTL_only designate genes significant in *cis*-eQTl analysis but not significant in ASE analysis. A t-test was performed between ASE_in_*cis*-eQTL group and the other two groups to assess whether a significant difference in mean expression values exists.

5.3.2 NC350 x H95*;Rp1-D21/+* (NH) ASE analysis

The AGPv4 reference was prepared for the NC350 x H95 (NH) ASE analysis by first anonymizing H95 and NC350 SNP positions. Mapping to the anonymized reference and post alignment processing of single-end reads from six F1 hybrid individuals (3 biological replicates per wildtype or mutant phenotype) were carried out as previously described for BH hybrids. Retrieval of allele counts was done by first, comparing H95 and NC350 SNPs to remove shared positions. ASEReadCounter was then used, with same parameters as indicated above, to generate counts for the non-overlapping sets of SNPs. The reference and alternate allele read counts in the NC350-private list were flipped prior to merging with the H95 counts, sorted and used to compute allele counts per gene. Binomial test was used to assess ASE with significance determined at 0.05 FDR. To identify genes that co-occur in single ASE and *cis*-eQTL experiments significant ASE genes from mutant BH analysis were overlapped with the significant *cis*-eQTL genes from the mutant RIL analysis. A similar comparison was carried out between wildtype BH significant ASE genes and significant *cis*-eQTL from the wildtype RIL experiment.

The H95-anonymized AGPv4 reference was further masked using 1,312,196 high-quality NC350 homozygous SNPs to produce a H-95-, NC350-anonymized AGPv4 reference. Between 21,399,539 and 26,775,218 raw reads were aligned, with uniquely mapped rate of 73.6% on average (Table 5-3). Further filtering of homozygous NC350 SNPs yielded 1,147,309 genic SNPs. Comparison of the genic SNPs between H95 and NC350 resulted in 1,742,425 (97.1%) and 1,003,144 (87.4%) private SNPs respectively. These were then used separately to generate allele counts at each position. Allele counts per gene were then computed and merged for subsequent use in gene-level binomial tests. A total of 18,046 genes (39.3% of known maize genes) in wildtype and 17,777 genes (38.7% of known maize genes) in the mutant were examined for allele-specific expression. Out of these, 14, 018 (77.7%) and 13,916 (78.2%) displayed significant imbalanced allelic expression in wildtype and mutant respectively (Table 5-4 and Figure 5-9). A greater proportion of the genes showing ASE exhibited bias in the direction of H95; 63.3% for wildtype and 63.4 % for mutants (Table 5-4).

Table 5-3. Results from aligning NC350 x H95;*Rp1-D21* (NH) RNA-seq reads to an NC350-anonymized AGPv4 reference genome.

Sample	Genotype	Phenotype	Raw reads	Uniquely mapped reads	Uniquely mapped reads %
NHwt_rep1	NC350 x H95;*Rp1-D21*	wildtype	21,399,539	16,178,489	75.6
NHwt_rep2	NC350 x H95;*Rp1-D21*	wildtype	25,645,653	19,707,310	76.84
NHwt_rep3	NC350 x H95;*Rp1-D21*	wildtype	22,179,388	15,838,290	71.41
NHmu_rep1	NC350 x H95;*Rp1-D21*	*Rp1-D21*	25,567,445	18,656,614	72.97
NHmu_rep2	NC350 x H95;*Rp1-D21*	*Rp1-D21*	26,775,218	19,422,746	72.54
NHmu_rep3	NC350 x H95;*Rp1-D21*	*Rp1-D21*	26,373,971	19,008,342	72.07

Table 5-4. Number of genes showing allelic imbalance and direction of bias within the NC350 x H95;*Rp1-D21*/+ background.

Sample	NC350-bias	H95-bias	Sum
NHwt	5,147	8,871	14,018
NHmu	5,085	8,831	13,916

Of the genes with imbalanced allelic expression in the wildtype, 5362 (38.3%) were differentially expressed between wildtype and mutant, more than the 4587.8 expected by random chance (χ2 p-value 4.00E-53). Likewise, 5517 (39.6%) of the ASE genes in the mutant showed differential expression; almost a thousand more genes than the 4590.1 expected by random chance (χ^2 p-value 2.00E-75) (Figure 5-10). The proportion of ASE genes that were differentially expressed here was more than three-fold larger (in both wildtype and mutant) than those identified in the BH background. This lends support to the suggestion that detection power might be contributing to limited overlap between ASE and DEG. NC350, having an enhanced phenotype due to *Rp1-D21* (in contrast to B73's suppressed phenotype) consequently showed enhanced gene expression, and led to better detection sensitivity of DEGs. More DEGs identified between wildtype and mutant NH increased the chance for overlap with ASE Genes. A majority of *cis*-eQTL genes assessed for ASE were validated; 5,669 (79.7%) for the wildtype and 5520 (81.3%) for the mutants (Figure 5-11). *Cis*-eQTL affecting these genes represent the true *cis*-eQTL set since

they act in an allele-specific manner to influence their target gene. As indicated earlier, non-validated *cis*-eQTL could either be local *trans*-acting regulatory variants that are but were erroneously detected as *cis*-eQTL or true *cis*-eQTL that could not be detected due to low sensitivity of ASE analysis.

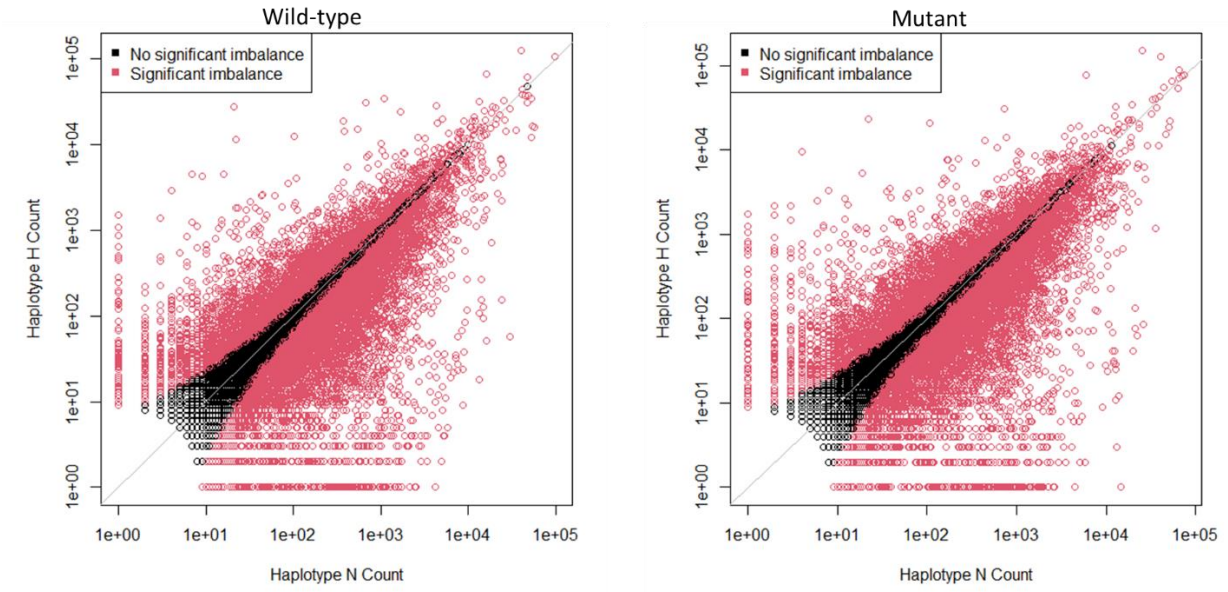


Figure 5-9. ASE analysis results from NC350 x H95;*Rp1-D21*/+ (NH) hybrid F1 plants showing wildtype (left) or *Rp1-D21* (right) phenotype. The x-axis represents allele counts for N haplotype for each gene tested whilst the y-axis denotes the counts for H haplotype. Red dots are genes showing significant allelic imbalance (FDR ≤ 0.05), whereas black dots represent genes with balanced expression.

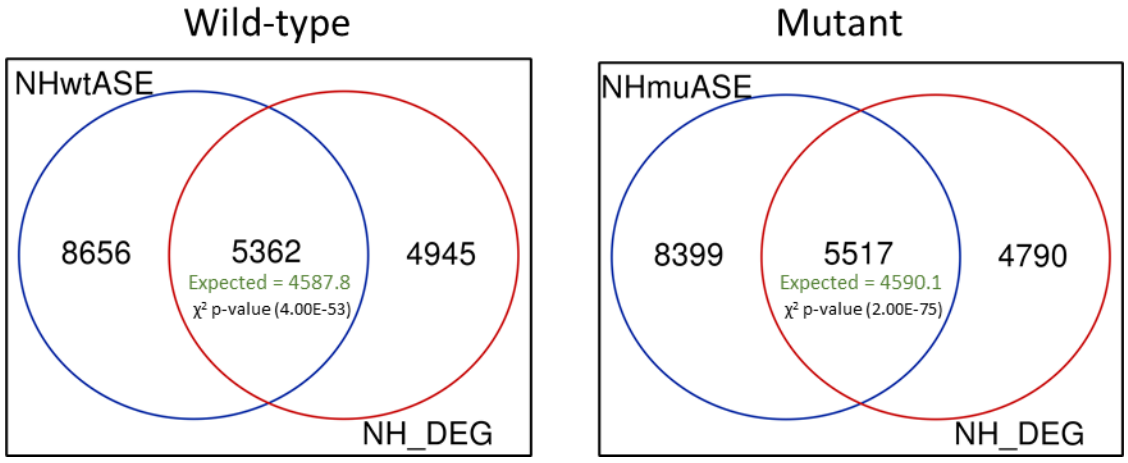


Figure 5-10. Overlap between genes identified from ASE analysis in NC350 x H95;*Rp1-D21*/+ (NH) hybrid F1 plants showing wildtype (left) or RPI-D21 (right) phenotype and differentially expressed genes between plants showing wildtype versus RPI-D21 phenotype.

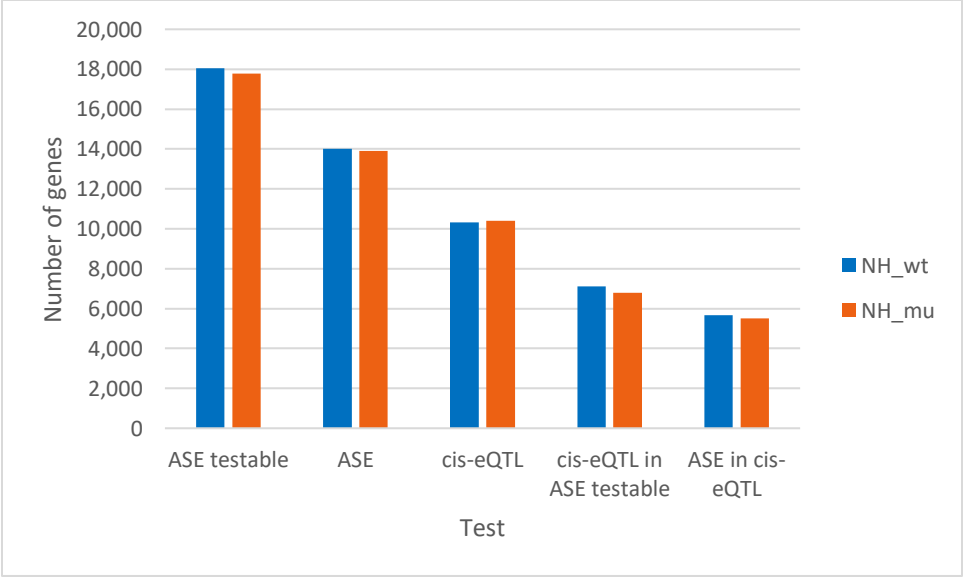


Figure 5-11. Integrating results from ASE in NC350 x H95;*Rp1-D21*/+ (NH) hybrid F1 and *cis*-eQTL analyses. Orange bars indicate results from plants showing *Rp1-D21* (right) phenotype, and blue bars are results from plants showing wildtype phenotype. X-axis shows the different comparisons carried out whereas the y-axis shows the number of genes identified.

Similar to ASE analysis in the BH background, a comparison of gene expression across groups of genes common to ASE and *cis*-eQTL analyses and those that are unique to either methodology was carried out. Pairwise t-test of log_2-normalized expression values showed that ASE-validated *cis*-eQTL genes had significantly higher expression than non-validated *cis*-eQTL genes in both wildtype and mutant plants (p-value < 2.22E-16) (Figure 5-12). This is further evidence that ASE analysis may have less detection sensitivity under lower gene expression and further confirms that while some *cis*-eQTL may rather be local *trans*-eQTL, others could not be validated because ASE may just lack the power to detect them due to their low expression.

Figure 12. Comparison of gene expression among *cis*-eQTL and ASE genes in NC350 x H95*;Rp1-D21*/+ (NH) hybrid F1 plants displaying wilt-type (A) or RPI-D21 (B) phenotype. X-axis represents unique or overlapping genes from ASE and eQTL analyses. Y-axis is the mean of log_2-normalized expression values across three replicates. ASE_in_*cis*-eQTL group denote significant genes from ASE and *cis*-eQTL analyses; ASE_only represent genes significant in ASE but not significant in *cis*-eQTL analysis; *cis*-eQTL_only designate genes significant in *cis*-eQTl analysis but not significant in ASE analysis. A t-test was performed between ASE_in_*cis*-eQTL group and the other two groups to assess whether a significant difference in mean expression values exists.

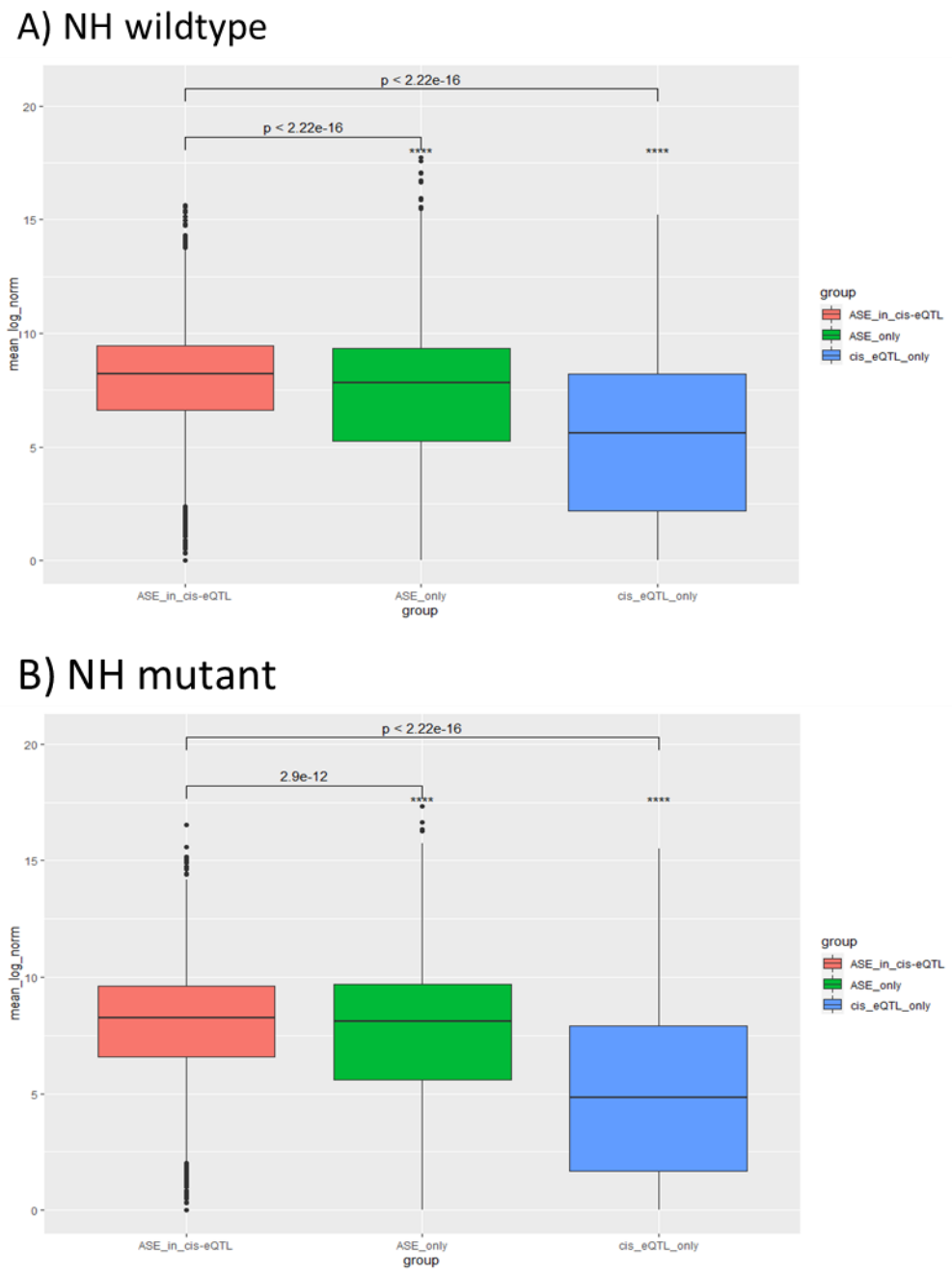


Figure 5-12. Comparison of gene expression among *cis*-eQTL and ASE genes in NC350 x H95:*Rp1-D21*/+ (NH) hybrid F1 plants displaying wilt-type (A) or RPI-D21 (B) phenotype. X-axis represents unique or overlapping genes from ASE and eQTL analyses. Y-axis is the mean of log_2-normalized expression values across three replicates. ASE_in_*cis*-eQTL group denote significant genes from ASE and *cis*-eQTL analyses; ASE_only represent genes significant in ASE but not significant in *cis*-eQTL analysis; *cis*-eQTL_only designate genes significant in *cis*-eQTl analysis but not significant in ASE analysis. A t-test was performed between ASE_in_*cis*-eQTL group and the other two groups to assess whether a significant difference in mean expression values exists.

5.3.3 Three-way ASE, by comparison to a common reference, matches B73-NC350 eQTL with ASE as a test of *cis*-eQTL mechanism

To reconstruct B73-NC350 (B-N) ASE a set of comparisons using H as the common reference, or control, were done. Reads counts from BH and NH mutant backgrounds were compared and allele counts for common genes combined in a separate file. Different approaches were taken to combine the BH and NH ASE data to permit downstream hypothesis testing. One approach was to take those genes with significant ASE in both comparisons. A Fisher's exact test was then conducted between the pairs of reference and alternative allele counts for each gene using the allele-specific read counts from the NC350 vs H95 comparison and from the independent B73 vs H95 comparison. *P*-values were FDR corrected with the Benjamini and Hochberg transformation. Deviation from the null hypothesis of a Fisher's exact test indicates a different relative expression between B73 and NC350, relative to the H95 common reference. Lists of these types were constructed from the wildtype siblings and mutant siblings separately and all tests were carried out independently. Genes with differing expression between the B and N alleles, relative to H95 were tabulated as Fisher test significant genes (Figure 5-13).

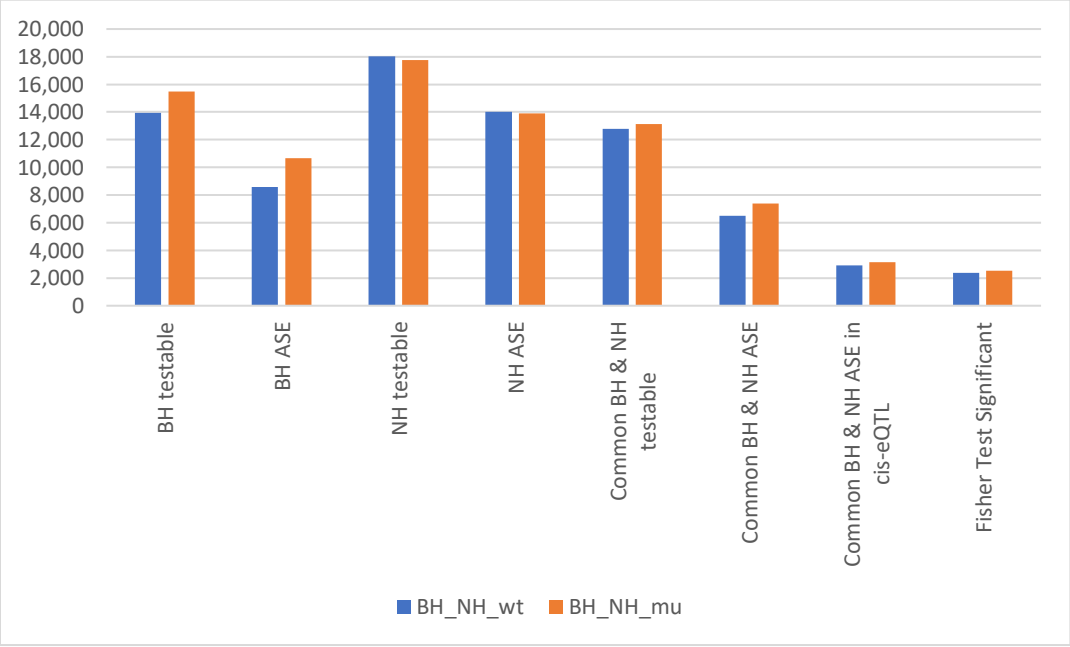


Figure 5-13. Comparison between B73-NC350 (B-N) ASE and *cis*-eQTL analysis results. Orange bars indicate results from plants showing *Rp1-D21* phenotype, and blue bars are results from plants showing wildtype phenotype. X-axis shows the different comparisons carried out whereas the y-axis shows the number of genes identified.

If local eQTL, detected in chapter 4, are the result of true *cis*-regulatory variation then they should affect ASE when NC350 and B73 alleles are compared. However, to deliver *Rp1-D21* in the same genetic background as used for the RIL F1 family mapping experiment, they need to be in the presence of H95;*Rp1-D21/+* as one of the parents. We used the tabulated data described above to compare the relative abundance of B73 and NC350 transcripts to the common H95 reference allele. If a *cis*-eQTL is encoded at a gene, then the direction of the effect in the *cis*-eQTL and ASE should be the same. For example, if the NC350 allele at the *cis*-eQTL increased transcript abundance, the ASE analysis should demonstrate a greater relative transcript abundance from the NC350 allele as compared to the B73 allele. If no ASE is detected, or the ASE occurs in the opposite direction as the eQTL, then the eQTL is more likely to be a local *trans*-acting variant.

Given the very large number of *trans*-eQTL detected in this experiment, a large number of local *trans*-eQTL seems possible. I used the sets genes identified as having ASE in both BH and NH in wildtype and mutant F1 progenies. These lists consisted of 6496 genes exhibiting ASE in the wildtype families and 7392 genes exhibiting ASE in the mutants. Out of these, 2930 (45.1%) in the wildtype results and 3144 (42.5%) in the mutant results were also affected by *cis*-eQTL. The ASE data were compared between BH and NH and the favored direction of expression for these genes were tabulated and compared to the direction of the effect observed at these genes among the *cis*-eQTL. For both wild-type (Table 5-5) and mutant (Table 5-6) comparisons, the number of genes with ASE that favored the NC350 allele and also had a *cis*-eQTL in the same direction was nearly two-fold greater that the number of genes with NC350-favoring ASE that had *cis*-eQTL affecting expression in the opposite direction. The same pattern was observed for the genes with ASE that favored the B73 allele and their eQTL. This pattern held true for both the wild-type (Table 5-5) and mutant (Table 5-6) expression data. This finding that a majority of the genes at which both eQTL and ASE were affected in the same direction, indicates that most of the local eQTL detected in Chapter 4 are indeed *cis*-eQTL and validates the utility of this double test of molecular mechanism to characterize *cis*-regulatory variants.

One caveat must be mentioned in this analysis. There was a fold-difference greater detection of ASE favoring the B73 allele in both mutant and wildtype samples. The direction of this effect, favoring the reference genome allele, strongly implicates reference bias as a confounding factor in these data. This reference bias appears to remain my efforts in genome

anonymization. For genes that were affected in the same direction in both ASE and *cis*-eQTL in wildtype the B allele was preferred almost 3-fold more (488 were N-allele favoring whilst 1508 favored the B allele) (Table 5-5). In the same vein, 550 genes were N favoring in contrast to 1491 B-favoring genes mutant plants (Table 5-6). It is clear from these experiments that for efficient ASE analysis an actual hybrid reference genome created from sequencing the F1 parental genomic DNA is invaluable.

Table 5-5. Assessing effect direction of *cis*-eQTL in plants with wildtype phenotype using common ASE and *cis*-eQTL genes.

		Relative ASE	
	Favoring	N	B
Cis-eQTL	N	488	666
	B	265	1508

Table 5-6. Assessing effect direction of *cis*-eQTL in plants with mutant phenotype using common ASE and *cis*-eQTL genes.

		Relative ASE	
	Favoring	N	B
Cis-eQTL	N	550	773
	B	327	1491

5.4 Conclusions

Several factors could be responsible for the limited overlap observed between differentially expressed genes and genes showing allelic expression imbalance. This could be caused by differences in the detection power between differential gene expression and ASE analyses. For instance, a gene with minimal expression (i.e., 4 – 8 reads) could still be detected as differentially expressed but not ASE because of the high read depth threshold (minimum 10 reads) of ASE. In this instance the gene will not even be assessed for ASE. As such, low number of ASE showing DE could be the result of the low detection power of ASE analysis. On the other hand, the lack of concordance between ASE and DEG gene sets may reveal a general lack of *cis*-regulatory variation

on gene expression variation in the two crosses. DEGs are identified as a difference between B73 x H95 versus B73 x H95 *Rp1-D21/+*, in the case of BH or between NC350 x H95 versus NC350 x H95 *Rp1-D21/+* as is the case for NH. If the changes in gene expression are affected by the HR induced by *Rp1-D21* and not by an interaction between the *trans*-regulatory mechanisms and the *cis*-acting polymorphisms we do not expect to observe substantial overlap between DEG and genes displaying ASE.

REFERENCES

Adachi, H., Białas, A., & Kamoun, S. (2020). *4.0 (CC BY-NC-ND) How to trick a plant pathogen? Plant Genomics How do plants detect pathogens?* http://portlandpress.com/biochemist/article-pdf/42/4/14/890986/bio20200043.pdf

Addo-Quaye, C., Buescher, E., Best, N., Chaikam, V., Baxter, I., & Dilkes, B. P. (2017). Forward genetics by sequencing EMS variation-induced inbred lines. *G3: Genes, Genomes, Genetics*, *7*(2), 413–425. https://doi.org/10.1534/g3.116.029660

Albert, F. W., Bloom, J. S., Siegel, J., Day, L., & Kruglyak, L. (2018). *Genetics of trans-regulatory variation in gene expression*. https://doi.org/10.7554/eLife.35471.001

Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. In *Nature Reviews Genetics* (Vol. 16, Issue 4, pp. 197–212). Nature Publishing Group. https://doi.org/10.1038/nrg3891

Ali, M., Cheng, Z., Ahmad, H., & Hayat, S. (2018). Reactive oxygen species (ROS) as defenses against a broad range of plant fungal infections and case study on ros employed by crops against verticillium dahlia wilts. In *Journal of Plant Interactions* (Vol. 13, Issue 1, pp. 353–363). Taylor and Francis Ltd. https://doi.org/10.1080/17429145.2018.1484188

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10). https://doi.org/10.1186/gb-2010-11-10-r106

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

Bai, Y., Vaddepalli, P., Fulton, L., Bhasin, H., Hülskamp, M., & Schneitz, K. (2013). ANGUSTIFOLIA is a central component of tissue morphogenesis mediated by the atypical receptor-like kinase STRUBBELIG. *BMC Plant Biology*, *13*, 16. https://doi.org/10.1186/1471-2229-13-16

Balint-Kurti, P. (2019). The plant hypersensitive response: concepts, control and consequences. In *Molecular Plant Pathology* (Vol. 20, Issue 8, pp. 1163–1178). Blackwell Publishing Ltd. https://doi.org/10.1111/mpp.12821

Becker, J., Wendland, J. R., Haenisch, B., Nöthen, M. M., & Schumacher, J. (2012). A systematic eQTL study of *cis*-trans epistasis in 210 HapMap individuals. *European Journal of Human Genetics*, *20*(1), 97–101. https://doi.org/10.1038/ejhg.2011.156

Belfield, E. J., Gan, X., Mithani, A., Brown, C., Jiang, C., Franklin, K., Alvey, E., Wibowo, A., Jung, M., Bailey, K., Kalwani, S., Ragoussis, J., Mott, R., & Harberd, N. P. (2012). Genome-wide analysis of mutations in mutant lineages selected following fast-neutron irradiation mutagenesis of Arabidopsis thaliana. *Genome Research*, *22*(7), 1306–1315. https://doi.org/10.1101/gr.131474.111

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological), 57(1), 289–300. http://www.jstor.org/stable/2346101

Block, A. K., Vaughan, M. M., Schmelz, E. A., & Christensen, S. A. (2019). Biosynthesis and function of terpenoid defense compounds in maize (Zea mays). In *Planta* (Vol. 249, Issue 1, pp. 21–30). Springer Verlag. https://doi.org/10.1007/s00425-018-2999-2

Block, A. K., Tang, H. V., Hopkins, D., Mendoza, J., Solemslie, R. K., du Toit, L. J., & Christensen, S. A. (2021). A maize leucine-rich repeat receptor-like protein kinase mediates responses to fungal attack. Planta, 254(4), 73. https://doi.org/10.1007/s00425-021-03730-0

Bolon, Y.-T., Hyten, D., Orf, J. H., Vance, C. P., Muehlbauer, G. J., Bolon, Y.-T. ;, Hyten, D. ;, Orf, J. H. ;, & Vance, C. P. ; (2014). *eQTL Networks Reveal Complex Genetic Architecture in the Immature Soybean Seed*. *817*. https://doi.org/10.3835/plantgenome2013.08.0027

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell*, *169*(7), 1177. https://doi.org/10.1016/J.CELL.2017.05.038

Brandt, D. Y. C., Aguiar, V. R. C., Bitarello, B. D., Nunes, K., Goudet, J., & Meyer, D. (2015). Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 genomes project phase I data. *G3: Genes, Genomes, Genetics*, *5*(5), 931–941. https://doi.org/10.1534/g3.114.015784

Brem, R. B., Yvert, G., Clinton, R., & Kruglyak, L. (2002). Genetic Dissection of Transcriptional Regulation in Budding Yeast. *Science*, *296*(5568), 752–755. https://doi.org/10.1126/science.1069516

Britto, D. T., & Kronzucker, H. J. (2008). Cellular mechanisms of potassium transport in plants. *Physiologia Plantarum*, *133*(4), 637–650. https://doi.org/10.1111/j.1399-3054.2008.01067.x

Buckler, E. S., Holland, J. B., Bradbury, P. J., Acharya, C. B., Brown, P. J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J. C., Goodman, M. M., Harjes, C., Guill, K., Kroon, D. E., Larsson, S., Lepak, N. K., Li, H., Mitchell, S. E., Pressoir, G., … McMullen, M. D. (2009). The Genetic Architecture of Maize Flowering Time. *Science*, *325*(5941), 714–718. https://doi.org/10.1126/science.1174276

Candela, H., & Hake, S. (2008). The art and design of genetic screens: maize. *Nature Reviews. Genetics*, *9*(3), 192–203. https://doi.org/10.1038/NRG2291

Carneiro, M. O., Russ, C., Ross, M. G., Gabriel, S. B., Nusbaum, C., & DePristo, M. A. (2012). Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics*, *13*(1), 375. https://doi.org/10.1186/1471-2164-13-375

Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E., & Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. *Genome Biology*, *16*(1). https://doi.org/10.1186/s13059-015-0762-6

Chaikam, V., Negeri, A., Dhawan, R., Puchaka, B., Ji, J., Chintamanani, S., Gachomo, E. W., Zillmer, A., Doran, T., Weil, C., Balint-Kurti, P., & Johal, G. (2011). Use of mutant-assisted gene identification and characterization (MAGIC) to identify novel genetic loci that modify the maize hypersensitive response. *Theoretical and Applied Genetics*, *123*(6), 985–997. https://doi.org/10.1007/s00122-011-1641-5

Chakraborty, J., Ghosh, P., & Das, S. (2018). Autoimmunity in plants. *Planta*, *248*(4), 751–767. https://doi.org/10.1007/s00425-018-2956-0

Chang, K. S., Jeon, J. H., Kim, G. H., Jang, C. W., Jeong, S. J., Ju, Y. R., & Ahn, Y. J. (2017). Repellency of zerumbone identified in Cyperus rotundus rhizome and other constituents to Blattella germanica. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-16099-6

Chen, N.-C., Solomon, B., Mun, T., Iyer, S., & Langmead, B. (2020). *Reducing reference bias using multiple population reference genomes*. https://doi.org/10.1101/2020.03.03.975219

Cheng, A. Y., Teo, Y.-Y., & Ong, R. T.-H. (2014). Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. *Bioinformatics*, *30*(12), 1707–1713. https://doi.org/10.1093/bioinformatics/btu067

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 Journal*, *279*(13), 2431–2443. https://doi.org/10.1111/j.1742-4658.2012.08621.x

Ching, T., Huang, S., & Garmire, L. X. (2014). Power analysis and sample size estimation for RNA-Seq differential expression. *RNA*, *20*(11), 1684–1696. https://doi.org/10.1261/rna.046011.114

Chintamanani, S., Hulbert, S. H., Johal, G. S., & Balint-Kurti, P. J. (2010). Identification of a maize locus that modulates the hypersensitive defense response, using mutant-assisted gene identification and characterization. *Genetics*, *184*(3), 813–825. https://doi.org/10.1534/genetics.109.111880

Christie, N., Myburg, A. A., Joubert, F., Murray, S. L., Carstens, M., Lin, Y. C., Meyer, J., Crampton, B. G., Christensen, S. A., Ntuli, J. F., Wighard, S. S., van de Peer, Y., & Berger, D. K. (2017). Systems genetics reveals a transcriptional network associated with susceptibility in the maize–grey leaf spot pathosystem. *Plant Journal*, *89*(4), 746–763. https://doi.org/10.1111/tpj.13419

Cingolani, P., Patel, V. M., Coon, M., Nguyen, T., Land, S. J., Ruden, D. M., & Lu, X. (2012). Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Frontiers in Genetics*, *3*, 35. https://doi.org/10.3389/fgene.2012.00035

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Coll, N. S., Epple, P., & Dangl, J. L. (2011). Programmed cell death in the plant immune system. *Cell Death & Differentiation*, *18*(8), 1247–1256. https://doi.org/10.1038/cdd.2011.37

Collings, D. A., Gebbie, L. K., Howles, P. A., Hurley, U. A., Birch, R. J., Cork, A. H., Hocart, C. H., Arioli, T., & Williamson, R. E. (2008). Arabidopsis dynamin-like protein DRP1A: a null mutant with widespread defects in endocytosis, cellulose synthesis, cytokinesis, and cell expansion. *Journal of Experimental Botany*, *59*(2), 361–376. https://doi.org/10.1093/JXB/ERM324

Collins, N., Drake, J., Ayliffe, M., Sun, Q., Ellis, J., Hulbert, S., & Pryor, T. (1999). Molecular Characterization of the Maize Rp1-D Rust Resistance Haplotype and Its Mutants. In *The Plant Cell* (Vol. 11). www.plantcell.org

Comai, L., & Henikoff, S. (2006). TILLING: Practical single-nucleotide mutation discovery. In *Plant Journal* (Vol. 45, Issue 4, pp. 684–694). https://doi.org/10.1111/j.1365-313X.2006.02670.x

Conway, J. R., Lex, A., & Gehlenborg, N. (2017). UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics*, *33*(18), 2938–2940. https://doi.org/10.1093/bioinformatics/btx364

Cubillos, F. A., Yansouni, J., Khalili, H., Balzergue, S., Elftieh, S., Martin-Magniette, M. L., Serrand, Y., Lepiniec, L., Baud, S., Dubreucq, B., Renou, J. P., Camilleri, C., & Loudet, O. (2012). Expression variation in connected recombinant populations of Arabidopsis thaliana highlights distinct transcriptome architectures. *BMC Genomics*, *13*(1). https://doi.org/10.1186/1471-2164-13-117

Dangl, J. L., Horvath, D. M., & Staskawicz, B. J. (2013). Pivoting the plant immune system from dissection to deployment. *Science (New York, N.Y.)*, *341*(6147), 746–751. https://doi.org/10.1126/science.1236011

Dar, A. A., Choudhury, A. R., Kancharla, P. K., & Arumugam, N. (2017). The FAD2 gene in plants: Occurrence, regulation, and role. *Frontiers in Plant Science*, *8*. https://doi.org/10.3389/fpls.2017.01789

Das, K., & Roychoudhury, A. (2014). Reactive oxygen species (ROS) and response of antioxidants as ROS-scavengers during environmental stress in plants. In *Frontiers in Environmental Science* (Vol. 2, Issue DEC). Frontiers Media S.A. https://doi.org/10.3389/fenvs.2014.00053

Dash, L., McEwan, R. E., Montes, C., Mejia, L., Walley, J. W., Dilkes, B. P., & Kelley, D. R. (2021). slim shady is a novel allele of PHYTOCHROME B present in the T-DNA line SALK_015201 . *Plant Direct*, *5*(6). https://doi.org/10.1002/pld3.326

de Koning, D.-J., & Haley, C. S. (2005). *Genetical genomics in humans and model organisms*. http://www.cephb.fr/cephdb/

Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., & Pritchard, J. K. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, *25*(24), 3207–3212. https://doi.org/10.1093/bioinformatics/btp579

Depristo, M. A., Banks, E., Poplin, R. E., Garimella, K. v, Maguire, J. R., Hartl, C., Rivas, M. A., Hanna, M., Mckenna, A., Fennell, T. J., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., Genetics, P., Hospital, M. G., & Simches, R. B. (2011). *HHS Public Access*. *43*(5), 491–498. https://doi.org/10.1038/ng.806.A

Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, N. S., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloë, D., le Gall, C., Schaëffer, B., le Crom, S., Guedj, M., & Jaffrézic, F. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, *14*(6), 671–683. https://doi.org/10.1093/bib/bbs046

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Doerge, R. W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. In *Nature Reviews Genetics* (Vol. 3, Issue 1, pp. 43–52). https://doi.org/10.1038/nrg703

Doitsidou, M., Jarriault, S., & Poole, R. J. (2016). Next-Generation Sequencing-Based Approaches for Mutation Mapping and Identification in Caenorhabditis elegans. *Genetics*, *204*(2), 451–474. https://doi.org/10.1534/genetics.115.186197

Doitsidou, M., Poole, R. J., Sarin, S., Bigelow, H., & Hobert, O. (2010). C. elegans Mutant Identification with a One-Step Whole-Genome-Sequencing and SNP Mapping Strategy. *PLoS ONE*, *5*(11), e15435. https://doi.org/10.1371/journal.pone.0015435

Dolan, W. L., Dilkes, B. P., Stout, J. M., Bonawitz, N. D., & Chapple, C. (2017). Mediator complex subunits MED2, MED5, MED16, and MED23 genetically interact in the regulation of Phenylpropanoid biosynthesis. *Plant Cell*, *29*(12), 3269–3285. https://doi.org/10.1105/tpc.17.00282

Dressano, K., Weckwerth, P. R., Poretsky, E., Takahashi, Y., Villarreal, C., Shen, Z., Schroeder, J. I., Briggs, S. P., & Huffaker, A. (2020). Dynamic regulation of Pep-induced immunity through post-translational control of defence transcript splicing. *Nature Plants*, *6*(8), 1008–1019. https://doi.org/10.1038/s41477-020-0724-1

Edsgärd, D., Iglesias, M. J., Reilly, S. J., Hamsten, A., Tornvall, P., Odeberg, J., & Emanuelsson, O. (2016). GeneiASE: Detection of condition-dependent and static allele-specific expression from RNA-seq data without haplotype information. *Scientific Reports*, *6*. https://doi.org/10.1038/srep21134

el Kasmi, F. (2021). How activated NLRs induce anti-microbial defenses in plants. In *Biochemical Society Transactions* (Vol. 49, Issue 5, pp. 2177–2188). Portland Press Ltd. https://doi.org/10.1042/BST20210242

Ereful, N. C., Liu, L. Y., Tsai, E., Kao, S. M., Dixit, S., Mauleon, R., Malabanan, K., Thomson, M., Laurena, A., Lee, D., Mackay, I., Greenland, A., Powell, W., & Leung, H. (2016). Analysis of Allelic Imbalance in Rice Hybrids Under Water Stress and Association of Asymmetrically Expressed Genes with Drought-Response QTLs. *Rice*, *9*(1). https://doi.org/10.1186/s12284-016-0123-4

Farmer, E. E. (1994). Fatty acid signalling in plants and their associated microorganisms. In *Plant Molecular Biology* (Vol. 26).

Figueiredo, J., Sousa Silva, M., & Figueiredo, A. (2018). Subtilisin-like proteases in plant defence: the past, the present and beyond. In *Molecular Plant Pathology* (Vol. 19, Issue 4, pp. 1017–1028). Blackwell Publishing Ltd. https://doi.org/10.1111/mpp.12567

Forsburg, S. L. (2001). The art and design of genetic screens: yeast. *Nature Reviews Genetics*, *2*(9), 659–668. https://doi.org/10.1038/35088500

Förster, C., Handrick, V., Ding, Y., Nakamura, Y., Paetz, C., Schneider, B., Castro-Falcón, G., Hughes, C. C., Luck, K., Poosapati, S., Kunert, G., Huffaker, A., Gershenzon, J., Schmelz, E. A., & Köllner, T. G. (2022). Biosynthesis and antifungal activity of fungus-induced O-methylated flavonoids in maize. *Plant Physiology*, *188*(1), 167–190. https://doi.org/10.1093/plphys/kiab496

Frey, M., Chomet, P., Glawischnig, E., Stettner, C., Grun, S., Winklmair, A., Eisenreich, W., Bacher, A., Meeley, R. B., Briggs, S. P., Simcox, K., & Gierl, A. (1997). Analysis of a Chemical Plant Defense Mechanism in Grasses. *Science*, 277, 696–699. https://doi.org/10.1126/science.277.5326.696

Fu, J., Ren, F., Lu, X., Mao, H., Xu, M., Degenhardt, J., Peters, R. J., & Wang, Q. (2016). A tandem array of ent-Kaurene synthases in maize with roles in gibberellin and more specialized metabolism. *Plant Physiology*, *170*(2), 742–751. https://doi.org/10.1104/pp.15.01727

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., Paten, B., & Durbin, R. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, *36*(9), 875–881. https://doi.org/10.1038/nbt.4227

Ge, C., Wang, Y. G., Lu, S., Zhao, X. Y., Hou, B. K., Balint-Kurti, P. J., & Wang, G. F. (2021). Multi-Omics Analyses Reveal the Regulatory Network and the Function of ZmUGTs in Maize Defense Response. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/fpls.2021.738261

Goodman, R. N., & Novacky, A. J. (Anton J.). (1994). *The hypersensitive reaction in plants to pathogens : a resistance phenomenon*. American Phytopathological Society.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Göring, H. H. H., Curran, J. E., Johnson, M. P., Dyer, T. D., Charlesworth, J., Cole, S. A., Jowett, J. B. M., Abraham, L. J., Rainwater, D. L., Comuzzie, A. G., Mahaney, M. C., Almasy, L., MacCluer, J. W., Kissebah, A. H., Collier, G. R., Moses, E. K., & Blangero, J. (2007). Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nature Genetics*, *39*(10), 1208–1216. https://doi.org/10.1038/ng2119

Granell, A., Belles, J. M., & Conejero, V. (1987). Induction of pathogenesis-related proteins in tomato by citrus exocortis viroid, silver ion and ethephon. In *Physiological and Molecular Plant Pathology* (Vol. 31).

Greene, E. A., Codomo, C. A., Taylor, N. E., Henikoff, J. G., Till, B. J., Reynolds, S. H., Enns, L. C., Burtner, C., Johnson, J. E., Odden, A. R., Comai, L., & Henikoff, S. (2003). *Spectrum of Chemically Induced Mutations From a Large-Scale Reverse-Genetic Screen in Arabidopsis*. www.proweb.org/parsesnp

Han, G., Lu, C., Guo, J., Qiao, Z., Sui, N., Qiu, N., & Wang, B. (2020). C2H2 Zinc Finger Proteins: Master Regulators of Abiotic Stress Responses in Plants. *Frontiers in Plant Science*, *11*. https://doi.org/10.3389/FPLS.2020.00115/FULL

Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., Schork, N. J., Murray, S. S., Topol, E. J., Levy, S., & Frazer, K. A. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*, *10*(3), R32. https://doi.org/10.1186/gb-2009-10-3-r32

Hase, Y., Satoh, K., Seito, H., & Oono, Y. (2020). Genetic Consequences of Acute/Chronic Gamma and Carbon Ion Irradiation of Arabidopsis thaliana. *Frontiers in Plant Science*, *11*. https://doi.org/10.3389/fpls.2020.00336

He, X., Ma, H., Zhao, X., Nie, S., Li, Y., Zhang, Z., Shen, Y., Chen, Q., Lu, Y., Lan, H., Zhou, S., Gao, S., Pan, G., & Lin, H. (2016). Comparative RNA-Seq analysis reveals that regulatory network of maize root development controls the expression of Genes in response to N stress. *PLoS ONE*, *11*(3). https://doi.org/10.1371/journal.pone.0151697

Henry, I. M., Zinkgraf, M. S., Groover, A. T., & Comaia, L. (2015). A system for dosage-based functional genomics in poplar. *Plant Cell*, *27*(9), 2370–2383. https://doi.org/10.1105/tpc.15.00349

Hu, K., Xu, K., Wen, J., Yi, B., Shen, J., Ma, C., Fu, T., Ouyang, Y., & Tu, J. (2019). Helitron distribution in Brassicaceae and whole Genome Helitron density as a character for distinguishing plant species. *BMC Bioinformatics*, *20*(1). https://doi.org/10.1186/s12859-019-2945-8

Huang, L., Li, X., Zhang, W., Ung, N., Liu, N., Yin, X., Li, Y., McEwan, R. E., Dilkes, B., Dai, M., Hicks, G. R., Raikhel, N. v., Staiger, C. J., & Zhang, C. (2020). Endosidin20 targets the cellulose synthase catalytic domain to inhibit cellulose biosynthesis. *Plant Cell*, *32*(7), 2141–2157. https://doi.org/10.1105/tpc.20.00202

Jacob, F., & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. In *Journal of Molecular Biology* (Vol. 3, Issue 3, pp. 318–356). https://doi.org/10.1016/S0022-2836(61)80072-7

Jain, M., Fiddes, I. T., Miga, K. H., Olsen, H. E., Paten, B., & Akeson, M. (2015). Improved data analysis for the MinION nanopore sequencer. *Nature Methods*, *12*(4), 351–356. https://doi.org/10.1038/nmeth.3290

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M. C., Wang, B., Campbell, M. S., Stein, J. C., Wei, X., Chin, C.-S., Guill, K., Regulski, M., Kumari, S., Olson, A., Gent, J., Schneider, K. L., Wolfgruber, T. K., May, M. R., Springer, N. M., … Ware, D. (2017). Improved maize reference genome with single-molecule technologies. *Nature*, *546*(7659), 524. https://doi.org/10.1038/nature22971

Johal, G. S., Balint-Kurti, P., & Weil, C. F. (2008). Mining and Harnessing Natural Variation: A Little MAGIC. *Crop Science*, *48*(6), 2066. https://doi.org/10.2135/cropsci2008.03.0150

Jorgensen, E. M., & Mango, S. E. (2002). The art and design of genetic screens: Caenorhabditis elegans. *Nature Reviews Genetics*, *3*(5), 356–369. https://doi.org/10.1038/nrg794

Kaneda, M., & Tominaga, N. (1975). Isolation and Characterization of a Proteinase from the Sarcocarp of Melon Fruit. In *J, Biochem* (Vol. 78).

Kang, B. H., Busse, J. S., & Bednarek, S. Y. (2003). Members of the arabidopsis dynamin-like gene family, ADL1, are essential for plant cytokinesis and polarized cell growth. *Plant Cell*, *15*(4), 899–913. https://doi.org/10.1105/tpc.009670

Karre, S., Kim, S. B., Kim, B. S., Khangura, R. S., Sermons, S. M., Dilkes, B., Johal, G., & Balint-Kurti, P. (2021). Maize Plants Chimeric for an Autoactive Resistance Gene Display a Cell-Autonomous Hypersensitive Response but Non Cell Autonomous Defense Signaling. *Molecular Plant-Microbe Interactions*, *34*(6). https://doi.org/10.1094/MPMI-04-20-0091-R

Kaul, S., Koo, H. L., Jenkins, J., Rizzo, M., Rooney, T., Tallon, L. J., Feldblyum, T., Nierman, W., Benito, M. I., Lin, X., Town, C. D., Venter, J. C., Fraser, C. M., Tabata, S., Nakamura, Y., Kaneko, T., Sato, S., Asamizu, E., Kato, T., … Somerville, C. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature 2000 408:6814*, *408*(6814), 796–815. https://doi.org/10.1038/35048692

Khansefid, M., Pryce, J. E., Bolormaa, S., Chen, Y., Millen, C. A., Chamberlain, A. J., vander Jagt, C. J., & Goddard, M. E. (2018). Comparing allele specific expression and local expression quantitative trait loci and the influence of gene expression on complex trait variation in cattle. *BMC Genomics*, *19*(1). https://doi.org/10.1186/s12864-018-5181-0

Kile, B. T., & Hilton, D. J. (2005). The art and design of genetic screens: mouse. *Nature Reviews Genetics*, *6*(7), 557–567. https://doi.org/10.1038/nrg1636

Kim, B. Y., Park, J. H., Jo, H. Y., Koo, S. K., & Park, M. H. (2017). Optimized detection of insertions/deletions (INDELs) in whole-exome sequencing data. *PLoS ONE*, *12*(8). https://doi.org/10.1371/journal.pone.0182272

Kim, S. B., Van den Broeck, L., Karre, S., Choi, H., Christensen, S. A., Wang, G. F., . . . Balint-Kurti, P. (2021). Analysis of the transcriptomic, metabolomic, and gene regulatory responses to Puccinia sorghi in maize. *Molecular Plant Pathology*, *22*(4), 465-479. https://doi.org/10.1111/mpp.13040

Kliebenstein, D. (2009). Quantitative genomics: Analyzing intraspecific variation using global gene expression polymorphisms or eqtls. *Annual Review of Plant Biology*, *60*, 93–114. https://doi.org/10.1146/annurev.arplant.043008.092114

Knight, J. C. (2004). Allele-specific gene expression uncovered. In *Trends in Genetics* (Vol. 20, Issue 3, pp. 113–116). Elsevier Ltd. https://doi.org/10.1016/j.tig.2004.01.001

Kolde R. (2015) Package "pheatmap.".

Köllner, T. G., Schnee, C., Li, S., Svatoš, A., Schneider, B., Gershenzon, J., & Degenhardt, J. (2008). Protonation of a neutral (S)-β-bisabolene intermediate is involved in (S)-β-macrocarpene formation by the maize sesquiterpene synthases TPS6 and TPS11. *Journal of Biological Chemistry*, *283*(30), 20779–20788. https://doi.org/10.1074/jbc.M802682200

Kumawat, S., Rana, N., Bansal, R., Vishwakarma, G., Mehetre, S. T., Das, B. K., Kumar, M., Kumar Yadav, S., Sonah, H., Raj Sharma, T., & Deshmukh, R. (2019). Expanding avenue of fast neutron mediated mutagenesis for crop improvement. In *Plants* (Vol. 8, Issue 6). MDPI AG. https://doi.org/10.3390/plants8060164

Lamarre, S., Frasse, P., Zouine, M., Labourdette, D., Sainderichin, E., Hu, G., le Berre-Anton, V., Bouzayen, M., & Maza, E. (2018). Optimization of an RNA-seq differential gene expression analysis depending on biological replicate number and library size. *Frontiers in Plant Science*, *9*. https://doi.org/10.3389/fpls.2018.00108

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology*, *15*(6). https://doi.org/10.1186/gb-2014-15-6-r84

Leonetti, P., Stuttmann, J., & Pantaleo, V. (2021). Regulation of plant antiviral defense genes via host RNA-silencing mechanisms. In *Virology Journal* (Vol. 18, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s12985-021-01664-3

León-Novelo, L., Gerken, A. R., Graze, R. M., McIntyre, L. M., & Marroni, F. (2018). Direct Testing for Allele-Specific Expression Differences Between Conditions. *G3 (Bethesda, Md.)*, *8*(2), 447–460. https://doi.org/10.1534/g3.117.300139

Li, C., Sun, B., Li, Y., Liu, C., Wu, X., Zhang, D., Shi, Y., Song, Y., Buckler, E. S., Zhang, Z., Wang, T., & Li, Y. (2016). Numerous genetic loci identified for drought tolerance in the maize nested association mapping populations. *BMC Genomics*, *17*(1). https://doi.org/10.1186/s12864-016-3170-8

Li, F., Shimizu, A., Nishio, T., Tsutsumi, N., & Kato, H. (2019). Comparison and characterization of mutations induced by gamma-ray and carbon-ion irradiation in rice (Oryza sativa L.) using whole-genome resequencing. *G3: Genes, Genomes, Genetics*, *9*(11), 3743–3751. https://doi.org/10.1534/g3.119.400555

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987. https://doi.org/10.1093/BIOINFORMATICS/BTR509

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, L., Petsch, K., Shimizu, R., Liu, S., Xu, W. W., Ying, K., Yu, J., Scanlon, M. J., Schnable, P. S., Timmermans, M. C. P., Springer, N. M., & Muehlbauer, G. J. (2013). Mendelian and Non-Mendelian Regulation of Gene Expression in Maize. *PLoS Genetics*, *9*(1). https://doi.org/10.1371/journal.pgen.1003202

Li, R., Jeong, K., Davis, J. T., Kim, S., Lee, S., Michelmore, R. W., Kim, S., & Maloof, J. N. (2018). Integrated QTL and eQTL mapping provides insights and candidate genes for fatty acid composition, flowering time, and growth traits in a F2 population of a novel synthetic allopolyploid brassica napus. *Frontiers in Plant Science*, *871*. https://doi.org/10.3389/fpls.2018.01632

Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K., & Wang, J. (2009). SNP detection for massively parallel whole-genome resequencing. *Genome Research*, *19*(6), 1124–1132. https://doi.org/10.1101/GR.088013.108

Li, S., Liu, S. mei, Fu, H. wei, Huang, J. zhong, & Shu, Q. yao. (2018). High-resolution melting-based TILLING of γ ray-induced mutations in rice. *Journal of Zhejiang University: Science B*, *19*(8), 620–629. https://doi.org/10.1631/jzus.B1700414

Li, X., Song, Y., Century, K., Straight, S., Ronald, P., Dong, X., Lassner, M., & Zhang, Y. (2001). A fast neutron deletion mutagenesis-based reverse genetics system for plants. *Plant Journal*, *27*(3), 235–242. https://doi.org/10.1046/j.1365-313X.2001.01084.x

Li, Z., Wang, X., Cui, Y., Qiao, K., Zhu, L., Fan, S., & Ma, Q. (2020). Comprehensive Genome-Wide Analysis of Thaumatin-Like Gene Family in Four Cotton Species and Functional Identification of GhTLP19 Involved in Regulating Tolerance to Verticillium dahlia and Drought. *Frontiers in Plant Science*, *11*. https://doi.org/10.3389/fpls.2020.575015

Li, Z., Zhou, P., della Coletta, R., Zhang, T., Brohammer, A. B., H. O'Connor, C., Vaillancourt, B., Lipzen, A., Daum, C., Barry, K., de Leon, N., Hirsch, C. D., Buell, C. R., Kaeppler, S. M., Springer, N. M., & Hirsch, C. N. (2021). Single-parent expression drives dynamic gene expression complementation in maize hybrids. *Plant Journal*, *105*(1), 93–107. https://doi.org/10.1111/tpj.15042

Liu, D. X., Rajaby, R., Wei, L. L., Zhang, L., Yang, Z. Q., Yang, Q. Y., & Sung, W. K. (2021). Calling large indels in 1047 Arabidopsis with IndelEnsembler. *Nucleic Acids Research*, *49*(19), 10879–10894. https://doi.org/10.1093/nar/gkab904

Liu, H., Luo, X., Niu, L., Xiao, Y., Chen, L., Liu, J., Wang, X., Jin, M., Li, W., Zhang, Q., & Yan, J. (2017). Distant eQTLs and Non-coding Sequences Play Critical Roles in Regulating Gene Expression and Quantitative Trait Variation in Maize. *Molecular Plant*, *10*(3), 414–426. https://doi.org/10.1016/j.molp.2016.06.016

Liu, Y., Liu, X., Zheng, Z., Ma, T., Liu, Y., Long, H., Cheng, H., Fang, M., Gong, J., Li, X., Zhao, S., & Xu, X. (2020). Genome-wide analysis of expression QTL (eQTL) and allele-specific expression (ASE) in pig muscle identifies candidate genes for meat quality traits. *Genetics Selection Evolution*, *52*(1), 59. https://doi.org/10.1186/s12711

Liu, Y., Zhou, J., & White, K. P. (2014). RNA-seq differential expression studies: More sequence or more replication? *Bioinformatics*, *30*(3), 301–304. https://doi.org/10.1093/bioinformatics/btt688

Lolle, S., Stevens, D., & Coaker, G. (2020). Plant NLR-triggered immunity: from receptor activation to downstream signaling. In *Current Opinion in Immunology* (Vol. 62, pp. 99–105). Elsevier Ltd. https://doi.org/10.1016/j.coi.2019.12.007

Loman, N. J., Misra, R. v, Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., & Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, *30*(5), 434–439. https://doi.org/10.1038/nbt.2198

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12). https://doi.org/10.1186/s13059-014-0550-8

Luan, Q. L., Zhu, Y. X., Ma, S., Sun, Y., Liu, X. Y., Liu, M., Balint-Kurti, P. J., & Wang, G. F. (2021). Maize metacaspases modulate the defense response mediated by the NLR protein *Rp1-D21* likely by affecting its subcellular localization. *Plant Journal*, *105*(1), 151–166. https://doi.org/10.1111/TPJ.15047

Lukowitz, W., Gillmor, C. S., & Diger Scheible, W.-R. (2000). *Breakthrough Technologies Positional Cloning in Arabidopsis. Why It Feels Good to Have a Genome Initiative Working for You 1*. http://www.arabidopsis.org/cgi-bin/maps/Riintromap

McClintock, B. (1949). THE ORIGIN AND BEHAVIOR OF MUTABLE LOCI IN MAIZE. In *Am. J. Botany* (Vol. 35, Issue 2). Cambridge University Press. https://www.pnas.org

McClintock, B. (1956a). Controlling elements and the gene. *Cold Spring Harbor Symposia on Quantitative Biology*, *21*, 197–216. https://doi.org/10.1101/SQB.1956.021.01.017

McClintock, B. (1956b). Intranuclear systems controlling gene action and mutation. *Brookhaven Symposia in Biology*, *8*, 58–74.

McClintock, B. (1961). Some Parallels Between Gene Control Systems in Maize and in Bacteria. *The American Naturalist*, *95*(884), 265–277. https://doi.org/10.1086/282188

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

McMullen, M. D., Kresovich, S., Villeda, H. S., Bradbury, P., Li, H., Sun, Q., Flint-Garcia, S., Thornsberry, J., Acharya, C., Bottoms, C., Brown, P., Browne, C., Eller, M., Guill, K., Harjes, C., Kroon, D., Lepak, N., Mitchell, S. E., Peterson, B., … Buckler, E. S. (2009). Genetic Properties of the Maize Nested Association Mapping Population. *Science*, *325*(5941), 737–740. https://doi.org/10.1126/science.1174320

Meihls, L. N., Kaur, H., & Jander, G. (2012). Natural variation in maize defense against insect herbivores. *Cold Spring Harbor Symposia on Quantitative Biology*, *77*, 269–283. https://doi.org/10.1101/sqb.2012.77.014662

Mellon2, J. E., & West, C. A. (1979). Diterpene Biosynthesis in Maize Seedlings in Response to Fungal Infection'. In *Plant Physiol* (Vol. 64).

Mi, H., Muruganujan, A., Ebert, D., Huang, X., & Thomas, P. D. (2019). PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research*, *47*(D1), D419–D426. https://doi.org/10.1093/nar/gky1038

Mikkilineni, V., & Rocheford, T. R. (2003). Sequence variation and genomic organization of fatty acid desaturase-2 (fad2) and fatty acid desaturase-6 (fad6) cDNAs in maize. *Theoretical and Applied Genetics*, *106*(7), 1326–1332. https://doi.org/10.1007/s00122-003-1190-7

Minevich, G., Park, D. S., Blankenberg, D., Poole, R. J., & Hobert, O. (2012). CloudMap: a cloud-based pipeline for analysis of mutant genome sequences. *Genetics*, *192*(4), 1249–1269. https://doi.org/10.1534/genetics.112.144204

Minoche, A. E., Dohm, J. C., & Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biology*, *12*(11), R112. https://doi.org/10.1186/gb-2011-12-11-r112

Misra, R. C., Sandeep, Kamthan, M., Kumar, S., & Ghosh, S. (2016). A thaumatin-like protein of Ocimum basilicum confers tolerance to fungal pathogen and abiotic stress in transgenic Arabidopsis. *Scientific Reports*, *6*. https://doi.org/10.1038/srep25340

Morales, R., Charon, M. H., Kachalova, G., Serre, L., Medina, M., Gómez-Moreno, C., & Frey, M. (2000). A redox-dependent interaction between two electron-transfer partners involved in photosynthesis. *EMBO Reports*, *1*(3), 271. https://doi.org/10.1093/EMBO-REPORTS/KVD057

Moresco, E. M. Y., Li, X., & Beutler, B. (2013). Going Forward with Genetics. *The American Journal of Pathology*, *182*(5), 1462–1473. https://doi.org/10.1016/j.ajpath.2013.02.002

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621–628. https://doi.org/10.1038/nmeth.1226

Mur, L. A. J., Kenton, P., Lloyd, A. J., Ougham, H., & Prats, E. (2008). The hypersensitive response; The centenary is upon us but how much do we know? *Journal of Experimental Botany*, *59*(3), 501–520. https://doi.org/10.1093/jxb/erm239

Murphree, C., Kim, S. B., Karre, S., Samira, R., & Balint-Kurti, P. (2020). Use of virus-induced gene silencing to characterize genes involved in modulating hypersensitive cell death in maize. *Molecular Plant Pathology*, *21*(12), 1662–1676. https://doi.org/10.1111/mpp.12999

Negeri, A., Wang, G. F., Benavente, L., Kibiti, C. M., Chaikam, V., Johal, G., & Balint-Kurti, P. (2013). Characterization of temperature and light effects on the defense response phenotypes associated with the maize *Rp1-D21* autoactive resistance gene. *BMC Plant Biology*, *13*(1). https://doi.org/10.1186/1471-2229-13-106

Nica, A. C., & Dermitzakis, E. T. (2013). Expression quantitative trait loci: Present and future. In *Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 368, Issue 1620). Royal Society of London. https://doi.org/10.1098/rstb.2012.0362

Niemeyer, H. M. (1988). *HYDROXAMIC ACIDS (4-HYDROXY-1,4-BENZOXAZIN-3-ONES), DEFENCE CHEMICALS IN THE GRAMINEAE* (Vol. 27, Issue 11).

Nordström, K. J. V., Albani, M. C., James, G. V., Gutjahr, C., Hartwig, B., Turck, F., Paszkowski, U., Coupland, G., & Schneeberger, K. (2013). Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k-mers. *Nature Biotechnology*, *31*(4), 325–330. https://doi.org/10.1038/nbt.2515

Obholzer, N., Swinburne, I. A., Schwab, E., Nechiporuk, A. v., Nicolson, T., & Megason, S. G. (2012). Rapid positional cloning of zebrafish mutations by linkage and homozygosity mapping using whole-genome sequencing. *Development*, *139*(22), 4280–4290. https://doi.org/10.1242/dev.083931

Ohlroggeav', J., & Browseb, J. (1995). Lipid Biosynthesis. In *The Plant Cell* (Vol. 7). American Society of Plant Physiologists. https://academic.oup.com/plcell/article/7/7/957/5985048

Olukolu, B. A., Bian, Y., de Vries, B., Tracy, W. F., Wisser, R. J., Holland, J. B., & Balint-Kurti, P. J. (2016). The genetics of leaf flecking in maize and its relationship to plant defense and disease resistance. *Plant Physiology*, *172*(3), 1787–1803. https://doi.org/10.1104/pp.15.01870

Olukolu, B. A., Negeri, A., Dhawan, R., Venkata, B. P., Sharma, P., Garg, A., Gachomo, E., Marla, S., Chu, K., Hasan, A., Ji, J., Chintamanani, S., Green, J., Shyu, C. R., Wisser, R., Holland, J., Johal, G., & Balint-Kurti, P. (2013). A connected set of genes associated with programmed cell death implicated in controlling the hypersensitive response in maize. *Genetics*, *193*(2), 609–620. https://doi.org/10.1534/genetics.112.147595

Olukolu, B. A., Wang, G. F., Vontimitta, V., Venkata, B. P., Marla, S., Ji, J., Gachomo, E., Chu, K., Negeri, A., Benson, J., Nelson, R., Bradbury, P., Nielsen, D., Holland, J. B., Balint-Kurti, P. J., & Johal, G. (2014). A Genome-Wide Association Study of the Maize Hypersensitive Defense Response Identifies Genes That Cluster in Related Pathways. *PLoS Genetics*, *10*(8). https://doi.org/10.1371/journal.pgen.1004562

Ossowski, S., Schneeberger, K., Lucas-Lledó, J. I., Warthmann, N., Clark, R. M., Shaw, R. G., Weigel, D., & Lynch, M. (2010). The Rate and Molecular Spectrum of Spontaneous Mutations in Arabidopsis thaliana. *Science*, *327*(5961).

Page, D. R., & Grossniklaus, U. (2002). The art and design of genetic screens: Arabidopsis thaliana. *Nature Reviews Genetics*, *3*(2), 124–136. https://doi.org/10.1038/nrg730

Pan, W., Lin, J., & Le, C. T. (2002). *How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach*. http://genomebiology.com/2002/3/5/research/0022.1

Patton, E. E., & Zon, L. I. (2001). The art and design of genetic screens: zebrafish. *Nature Reviews Genetics*, *2*(12), 956–966. https://doi.org/10.1038/35103567

Peterson, P. A. (1953). *A STUDY OF A MUTABLE PALE GREEN LOCUS IN MAIZE*.

Poland, J. A., Bradbury, P. J., Buckler, E. S., & Nelson, R. J. (2011). Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(17), 6893–6898. https://doi.org/10.1073/pnas.1010894108

Poplin, R., Ruano-Rubio, V., Depristo, M. A., Fennell, T. J., Carneiro, M. O., van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M. J., Neale, B., Macarthur, D. G., & Banks, E. (2018). *Scaling accurate genetic variant discovery to tens of thousands of samples*. https://doi.org/10.1101/201178

Radauer, C., Lackner, P., & Breiteneder, H. (2008). The Bet v 1 fold: an ancient, versatile scaffold for binding of large, hydrophobic ligands. *BMC Evolutionary Biology*, *8*(1), 286. https://doi.org/10.1186/1471-2148-8-286

Rakwal, R., Kumar Agrawal, G., Yonekura, M., & Kodama, O. (2000). Naringenin 7-O-methyltransferase involved in the biosynthesis of the flavanone phytoalexin sakuranetin from rice (Oryza sati6a L.). In *Plant Science* (Vol. 155). www.elsevier.com/locate/plantsci

Richter, T. E., Hulbert, S. H., & Pryorb, T. (1996). Disease Lesion Mimicry Caused by Mutations in the Rust Resistance Gene rpl. In *The Plant Cell* (Vol. 8). American Society of Plant Physiologists.

Sahu, P. K., Sao, R., Mondal, S., Vishwakarma, G., Gupta, S. K., Kumar, V., Singh, S., Sharma, D., & Das, B. K. (2020). Next generation sequencing based forward genetic approaches for identification and mapping of causal mutations in crop plants: A comprehensive review. In *Plants* (Vol. 9, Issue 10, pp. 1–47). MDPI AG. https://doi.org/10.3390/plants9101355

Salavati, M., Bush, S. J., Palma-Vera, S., McCulloch, M. E. B., Hume, D. A., & Clark, E. L. (2019). Elimination of Reference Mapping Bias Reveals Robust Immune Related Allele-Specific Expression in Crossbred Sheep. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.00863

Sarin, S., Prabhu, S., O'Meara, M. M., Pe'er, I., & Hobert, O. (2008). Caenorhabditis elegans mutant allele identification by whole-genome sequencing. *Nature Methods*, *5*(10), 865–867. https://doi.org/10.1038/nmeth.1249

Schadt, E. E., Monks, S. A., Drake, T. A., Lusis, A. J., Che, N., Colinayo, V., Ruff, T. G., Milligan, S. B., Lamb, J. R., Cavet, G., Linsley, P. S., Mao, M., Stoughton, R. B., & Friend, S. H. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature 2003 422:6929*, *422*(6929), 297–302. https://doi.org/10.1038/nature01434

Schaller, A., Stintzi, A., & Graff, L. (2012). Subtilases - versatile tools for protein turnover, plant development, and interactions with the environment. In *Physiologia Plantarum* (Vol. 145, Issue 1, pp. 52–66). https://doi.org/10.1111/j.1399-3054.2011.01529.x

Schneeberger, K. (2014). Using next-generation sequencing to isolate mutant genes from forward genetic screens. *Nature Reviews Genetics*, *15*(10), 662–676. https://doi.org/10.1038/nrg3745

Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A. H., Nielsen, K. L., Jørgensen, J.-E., Weigel, D., & Andersen, S. U. (2009). SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nature Methods*, *6*(8), 550–551. https://doi.org/10.1038/nmeth0809-550

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G. G., Owen-Hughes, T., Blaxter, M., & Barton, G. J. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, *22*(6), 839–851. https://doi.org/10.1261/rna.053959.115

Shabalin, A. A. (2012). Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, *28*(10), 1353–1358. https://doi.org/10.1093/bioinformatics/bts163

Shanklin, J., & Cahoon, E. B. (1998). DESATURATION AND RELATED MODIFICATIONS OF FATTY ACIDS 1. In *Annu. Rev. Plant Physiol. Plant Mol. Biol* (Vol. 49). www.annualreviews.org

Shen, Y., Garcia, T., Pabuwal, V., Boswell, M., Pasquali, A., Beldorth, I., Warren, W., Schartl, M., Cresko, W. A., & Walter, R. B. (2013). Alternative strategies for development of a reference transcriptome for quantification of allele specific expression in organisms having sparse genomic resources. *Comparative Biochemistry and Physiology - Part D: Genomics and Proteomics*, *8*(1), 11–16. https://doi.org/10.1016/j.cbd.2012.10.006

Shi, C., Uzarowska, A., Ouzunova, M., Landbeck, M., Wenzel, G., & Lübberstedt, T. (2007). Identification of candidate genes associated with cell wall digestibility and eQTL (expression quantitative trait loci) analysis in a Flint × Flint maize recombinant inbred line population. *BMC Genomics*, *8*. https://doi.org/10.1186/1471-2164-8-22

Shuman, H. A., & Silhavy, T. J. (2003). The art and design of genetic screens: Escherichia coli. *Nature Reviews Genetics*, *4*(6), 419–431. https://doi.org/10.1038/nrg1087

Silva-Guzman, M., Addo-Quaye, C., & Dilkes, B. P. (2016). Re-evaluation of reportedly metal tolerant Arabidopsis thaliana accessions. *PLoS ONE*, *11*(7). https://doi.org/10.1371/journal.pone.0130679

Slabaugh, E., Desai, J. S., Sartor, R. C., Mae Lawas, L. F., Krishna Jagadish, S., & Doherty, C. J. (2019). *Analysis of differential gene expression and alternative splicing is significantly influenced by choice of reference genome*. https://doi.org/10.1261/rna

Smith, S. M., Steinau, M., Trick, H. N., & Hulbert, S. H. (2010). Recombinant Rp1 genes confer necrotic or nonspecific resistance phenotypes. *Molecular Genetics and Genomics*, *283*(6), 591–602. https://doi.org/10.1007/s00438-010-0536-5

Springer, N. M., & Stupar, R. M. (2007). Allele-specific expression patterns reveal biases and embryo-specific parent-of-origin effects in hybrid maize. *Plant Cell*, *19*(8), 2391–2402. https://doi.org/10.1105/tpc.107.052258

St Johnston, D. (2002). The art and design of genetic screens: Drosophila melanogaster. *Nature Reviews Genetics*, *3*(3), 176–188. https://doi.org/10.1038/nrg751

Stevenson, K. R., Coolon, J. D., & Wittkopp, P. J. (2013). Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. *BMC Genomics*, *14*(1). https://doi.org/10.1186/1471-2164-14-536

Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., Ingle, C. E., Dunning, M., Flicek, P., Koller, D., Montgomery, S., Tavaré, S., Deloukas, P., & Dermitzakis, E. T. (2007). Population genomics of human gene expression. *Nature Genetics*, *39*(10), 1217–1224. https://doi.org/10.1038/ng2142

Sudupak, M. A., Bennetzen, J. L., & Hulbert, S. H. (1993). Unequal exchange and meiotic instability of disease-resistance genes in the Rp1 region of maize. *Genetics*, *133*(1), 119–125. http://www.ncbi.nlm.nih.gov/pubmed/8417982

Suga, T., Ohta, S., Munesada, K., Ide, N., Kurokawa, M., Shimizu, M., & Ohta, E. (1993). *ENDOGENOUS PINE WOOD NEMATICIDAL SUBSTANCES IN PINES, PINUS MASSONIANA, P. STROBUS AND P. PALUSTRIS* (Vol. 33, Issue 6).

Sun, Q., Collins, N. C., Ayliffe, M., Smith, S. M., Drake, J., Pryor, T., & Hulbert, S. H. (2001). *Recombination Between Paralogues at the rp1 Rust Resistance Locus in Maize*.

Takahashi, S., Yeo, Y.-S., Zhao, Y., O'maille, P. E., Greenhagen, B. T., Noel, J. P., Coates, R. M., & Chappell, J. (2007). Functional Characterization of Premnaspirodiene Oxygenase, a Cytochrome P450 Catalyzing Regio-and Stereo-specific Hydroxylations of Diverse Sesquiterpene Substrates*,s. In *J Biol Chem* (Vol. 282, Issue 43). http://www.jbc.org

Tan, C., Zhang, X. Q., Wang, Y., Wu, D., Bellgard, M. I., Xu, Y., Shu, X., Zhou, G., & Li, C. (2019). Characterization of genome-wide variations induced by gamma-ray radiation in barley using RNA-Seq. *BMC Genomics*, *20*(1). https://doi.org/10.1186/s12864-019-6182-3

Tang, Y., Guo, J., Zhang, T., Bai, S., He, K., & Wang, Z. (2021). Genome-wide analysis of WRKY gene family and the dynamic responses of key WRKY genes involved in Ostrinia furnacalis attack in Zea mays. *International Journal of Molecular Sciences*, *22*(23). https://doi.org/10.3390/ijms222313045

Teo, S. M., Pawitan, Y., Ku, C. S., Chia, K. S., & Salim, A. (2012). Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics*, *28*(21), 2711–2718. https://doi.org/10.1093/bioinformatics/bts535

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., & Narechania, A. (2003). PANTHER: A library of protein families and subfamilies indexed by function. *Genome Research*, *13*(9), 2129–2141. https://doi.org/10.1101/gr.772403

Tian, F., Bradbury, P. J., Brown, P. J., Hung, H., Sun, Q., Flint-Garcia, S., Rocheford, T. R., McMullen, M. D., Holland, J. B., & Buckler, E. S. (2011). Genome-wide association study of leaf architecture in the maize nested association mapping population. In *Nature Genetics* (Vol. 43, Issue 2, pp. 159–162). https://doi.org/10.1038/ng.746

Tian, J., Keller, M. P., Broman, A. T., Kendziorski, C., Yandell, B. S., Attie, A. D., & Broman, K. W. (2016). The dissection of expression quantitative trait locus hotspots. *Genetics*, *202*(4), 1563–1574. https://doi.org/10.1534/genetics.115.183624

van de Geijn, B., Mcvicker, G., Gilad, Y., & Pritchard, J. K. (2015). WASP: Allele-specific software for robust molecular quantitative trait locus discovery. In *Nature Methods* (Vol. 12, Issue 11, pp. 1061–1063). Nature Publishing Group. https://doi.org/10.1038/nmeth.3582

van Loon, L. C., Rep, M., & Pieterse, C. M. J. (2006). Significance of inducible defense-related proteins in infected plants. In *Annual Review of Phytopathology* (Vol. 44, pp. 135–162). https://doi.org/10.1146/annurev.phyto.44.070505.143425

van Verk, M. C., Pappaioannou, D., Neeleman, L., Bol, J. F., & Linthorst, H. J. M. (2008). A novel WRKY transcription factor is required for induction of PR-1a gene expression by salicylic acid and bacterial elicitors. *Plant Physiology*, *146*(4), 1983–1995. https://doi.org/10.1104/pp.107.112789

van Wersch, S., Tian, L., Hoy, R., & Li, X. (2020). Plant NLRs: The Whistleblowers of Plant Immunity. In *Plant Communications* (Vol. 1, Issue 1). Cell Press. https://doi.org/10.1016/j.xplc.2019.100016

Vijaya Satya, R., Zavaljevski, N., & Reifman, J. (2012). A new strategy to reduce allelic bias in RNA-Seq readmapping. *Nucleic Acids Research*, *40*(16). https://doi.org/10.1093/nar/gks425

Wan, J., Wang, Q., Zhao, J., Zhang, X., Guo, Z., Hu, D., Meng, S., Lin, Y., Qiu, X., Mu, L., Ding, D., & Tang, J. (2022). Gene expression variation explains maize seed germination heterosis. *BMC Plant Biology*, *22*(1). https://doi.org/10.1186/s12870-022-03690-x

Wang, C. T., Ru, J. N., Liu, Y. W., Li, M., Zhao, D., Yang, J. F., Fu, J. D., & Xu, Z. S. (2018). Maize wrky transcription factor zmwrky106 confers drought and heat tolerance in transgenic plants. *International Journal of Molecular Sciences*, *19*(10). https://doi.org/10.3390/ijms19103046

Wang, N., Lysenkov, V., Orte, K., Kairisto, V., Aakko, J., Khan, S., & Elo, L. L. (2022) Tool evaluation for the detection of variably sized indels from next generation whole genome and targeted sequencing data. PLoS Comput Biol 18(2):e1009269. https://doi.org/10.1371/journal.pcbi.1009269

Wang, G. F., & Balint-Kurti, P. J. (2015). Cytoplasmic and nuclear localizations are important for the hypersensitive response conferred by maize autoactive Rp1-D21 protein. *Molecular Plant-Microbe Interactions*, *28*(9), 1023–1031. https://doi.org/10.1094/MPMI-01-15-0014-R

Wang, G. F., & Balint-Kurti, P. J. (2016). Maize Homologs of CCoAOMT and HCT, Two Key Enzymes in Lignin Biosynthesis, Form Complexes with the NLR Rp1 Protein to Modulate the Defense Response. *Plant Physiol*, *171*(3), 2166-2177. https://doi.org/10.1104/pp.16.00224

Wang, G. F., He, Y., Strauch, R., Olukolu, B. A., Nielsen, D., Li, X., & Balint-Kurti, P. J. (2015). Maize homologs of hydroxycinnamoyltransferase, a key enzyme in lignin biosynthesis, bind the nucleotide binding leucine-rich repeat rp1 proteins to modulate the defense response. *Plant Physiology*, *169*(3), 2230–2243. https://doi.org/10.1104/pp.15.00703

Wang, T., Peng, Q., Liu, B., Liu, X., Liu, Y., Peng, J., & Wang, Y. (2020). eQTLMAPT: Fast and Accurate eQTL Mediation Analysis With Efficient Permutation Testing Approaches. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.01309

Wang, X., Chen, Q., Wu, Y., Lemmon, Z. H., Xu, G., Huang, C., Liang, Y., Xu, D., Li, D., Doebley, J. F., & Tian, F. (2018). Genome-wide Analysis of Transcriptional Variability in a Large Maize-Teosinte Population. *Molecular Plant*, *11*(3), 443–459. https://doi.org/10.1016/j.molp.2017.12.011

Waters, A. J., Makarevitch, I., Noshay, J., Burghardt, L. T., Hirsch, C. N., Hirsch, C. D., & Springer, N. M. (2017). Natural variation for gene expression responses to abiotic stress in maize. *Plant Journal*, *89*(4), 706–717. https://doi.org/10.1111/tpj.13414

Wei, H., Wang, P., Chen, J., Li, C., Wang, Y., Yuan, Y., Fang, J., & Leng, X. (2020). Genome-wide identification and analysis of B-BOX gene family in grapevine reveal its potential functions in berry development. *BMC Plant Biology*, *20*(1). https://doi.org/10.1186/s12870-020-2239-3

Wickham, H. (2016) ggplot2 Elegant Graphics for Data Analysis Second Edition. http://www.springer.com/series/6991

Wittkopp, P. J., Haerum, B. K., & Clark, A. G. (2004). Evolutionary changes in cis and trans gene regulation. *Nature*, *430*(6995), 85–88. https://doi.org/10.1038/nature02698

Wu, J., & Baldwin, I. T. (2010). New Insights into Plant Responses to the Attack from Insect Herbivores. *Annual Review of Genetics*, *44*(1), 1–24. https://doi.org/10.1146/annurev-genet-102209-163500

Yang, F., Wang, J., Pierce, B. L., Chen, L. S., Aguet, F., Ardlie, K. G., Cummings, B. B., Gelfand, E. T., Getz, G., Hadley, K., Handsaker, R. E., Huang, K. H., Kashin, S., Karczewski, K. J., Lek, M., Li, X., MacArthur, D. G., Nedzel, J. L., Nguyen, D. T., … Zhu, J. (2017). Identifying *cis*-mediators for *trans*-eQTLs across many human tissues using genomic mediation analysis. *Genome Research*, *27*(11), 1859–1871. https://doi.org/10.1101/gr.216754.116

Yu, F., Okamto, S., Nakasone, K., Adachi, K., Matsuda, S., Harada, H., Misawa, N., & Utsumi, R. (2008). Molecular cloning and functional characterization of α-humulene synthase, a possible key enzyme of zerumbone biosynthesis in shampoo ginger (Zingiber zerumbet Smith). *Planta*, *227*(6), 1291–1299. https://doi.org/10.1007/s00425-008-0700-x

Yu, J., Holland, J. B., McMullen, M. D., & Buckler, E. S. (2008). Genetic design and statistical power of nested association mapping in maize. *Genetics*, *178*(1), 539–551. https://doi.org/10.1534/genetics.107.074245

Yuan, P., Tanaka, K., & Poovaiah, B. W. (2022). Calcium/Calmodulin-Mediated Defense Signaling: What Is Looming on the Horizon for AtSR1/CAMTA3-Mediated Signaling in Plant Immunity. In *Frontiers in Plant Science* (Vol. 12). Frontiers Media S.A. https://doi.org/10.3389/fpls.2021.795353

Zhang, C., Brown, M. Q., van de Ven, W., Zhang, Z. M., Wu, B., Young, M. C., Synek, L., Borchardt, D., Harrison, R., Pan, S., Luo, N., Huang, Y. M. M., Ghang, Y. J., Ung, N., Li, R., Isley, J., Morikis, D., Song, J., Guo, W., … Raikhel, N. v. (2016). Endosidin2 targets conserved exocyst complex subunit EXO70 to inhibit exocytosis. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(1), E41–E50. https://doi.org/10.1073/pnas.1521248112

Zhang, L., Yu, Y., Shi, T., Kou, M., Sun, J., Xu, T., Li, Q., Wu, S., Cao, Q., Hou, W., & Li, Z. (2020). Genome-wide analysis of expression quantitative trait loci (eQTLs) reveals the regulatory architecture of gene expression variation in the storage roots of sweet potato. *Horticulture Research*, *7*(1). https://doi.org/10.1038/s41438-020-0314-4

Zhang, Y., Zhou, Y., Sun, Q., Deng, D., Liu, H., Chen, S., & Yin, Z. (2019). Genetic determinants controlling maize rubisco activase gene expression and a comparison with rice counterparts. *BMC Plant Biology*, *19*(1). https://doi.org/10.1186/s12870-019-1965-x

’Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J. P., & Wang, L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics. 2014 Apr 1;30(7):1006-7. doi: 10.1093/bioinformatics/btt730. Epub 2013 Dec 18. PMID: 24351709; PMCID: PMC3967108.

Zhao, S., Ye, Z., & Stanton, R. (2020). *Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols*. https://doi.org/10.1261/rna

Zhou, P., Li, Z., Magnusson, E., Cano, F. G., Crisp, P. A., Noshay, J. M., Grotewold, E., Hirsch, C. N., Briggs, S. P., & Springer, N. M. (2020). Meta gene regulatory networks in maize highlight functionally relevant regulatory interactions. *Plant Cell*, *32*(5), 1377–1396. https://doi.org/10.1105/tpc.20.00080

Zhuo, Z., Lamont, S. J., & Abasht, B. (2017). RNA-Seq Analyses Identify Frequent Allele Specific Expression and No Evidence of Genomic Imprinting in Specific Embryonic Tissues of Chicken. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-12179-9