2019

Identification and Characterization of Stress Responsive Genes in Soybean and Sunflower

Surendra Neupane

South Dakota State University

IDENTIFICATION AND CHARACTERIZATION OF STRESS RESPONSIVE GENES

IN SOYBEAN AND SUNFLOWER

BY

SURENDRA NEUPANE

A dissertation submitted in partial fulfillment of the requirements for the

Doctor of Philosophy

Major in Biological Sciences

Specialization in Molecular Biology

South Dakota State University

2019

IDENTIFICATION AND CHARACTERIZATION OF STRESS RESPONSIVE GENES

IN SOYBEAN AND SUNFLOWER

SURENDRA NEUPANE

This dissertation is approved as a creditable and independent investigation by a candidate for the Doctor of Philosophy in Biological Sciences degree and is acceptable for meeting the dissertation requirements for this degree. Acceptance of this does not imply that the conclusions reached by the candidates are necessarily the conclusions of the major department.

Madhav P. Nepal, Ph.D.
Dissertation Advisor   Date

Adam J. Varenhorst, Ph.D.
Dissertation Co-Advisor   Date

Volker Brozel, Ph.D.
Head, Dept. of Biology & Microbiology   Date

Dean, Graduate School   Date

This dissertation is dedicated to my parents Ram Prasad Neupane and Chandra Kala Neupane and to my siblings Deepak Neupane, Sarita Neupane, and Deepa Neupane. I am immensely indebted to them for their infinite encouragement, support and unwavering trust without which there is nothing in my ability to succeed.

ACKNOWLEDGMENTS

I was blessed and lucky to be surrounded with many friends during my stay in Brookings. Especially, I am thankful to Deepak Raj Joshi, Subash Acharya, Krishna Ghimire, Bimal Paudel and Reshma Thapa for their care and support.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

R genes   Resistance genes

avr   Avirulence gene

NBS   Nucleotide-Binding Site

LRR   Leucine-Rich Repeat

CC   Coiled-Coil

NB-ARC   Nucleotide-Binding site found in Apoptotic protease-activating factor 1, R Proteins, and *Caenorhabditis elegans* death-4 protein

CNL   Coiled Coil-NBS-LRR

TNL   Toll-interleukin-1 receptor-like Nucleotide-binding site Leucine-rich repeat

RPW8   Resistance to powdery mildew8

RNL   RPW8-NBS-LRR

RLKs   Receptor-like protein kinases

RLPs   Receptor-like proteins

LRR-RLK   Leucine-rich repeat transmembrane receptor-like kinase

PRR   Pattern Recognition Receptor

PRGs   Pathogen Receptor Genes

SA	Salicylic Acid
JA	Jasmonic Acid
ET	Ethylene
PAMP	Pathogen-Associated Molecular Pattern
PTI	PAMP-Triggered Immunity
ETS	Effector-triggered susceptibility
ETI	Effector-Triggered Immunity
PCD	Programmed Cell Death
NRG1	N-required gene 1
ML	Maximum Likelihood
ADR1	Activated Disease Resistance gene 1
MAPK	Mitogen-Activated Protein Kinase
FLS2	Flagellin Sensitive 2
MYA	Million Years Ago
TILLING	Targeting Induced Local Lesions in Genomes
SNAP	Soluble NSF Attachment Protein
SHMT	Serine Hydroxy Methyl Transferase
WI12	Wound-inducible domain

KASP	Kompetitive Allele-Specific PCR
MPK	MAP Kinase
MKK	MAP Kinase Kinase
MKKK	MAP Kinase Kinase Kinase
HMM	Hidden Markov Model
HR	Hypersensitive Response
PKC	Calcium Dependent Protein Kinases
HMT	Histone Methylase
HAT	Histone Acetylase
SAR	Systemic Acquired Resistance
VIP1	VirE2 Interacting Protein 1
ROS	Reactive Oxygen Species
PR	Pathogenesis-Related
MKP1	MAP Kinase Phosphatase 1
RBOH	Respiratory Burst Oxidase Homolog
VPS52	Vacuolar Protein Sorting-Associated Protein 52
ACE	Angiotensin-Converting-Enzyme-like
CRP	Cysteine-Rich Protein

PPO	Polyphenol oxidase
apN	Aminopeptidase-N
CYPs	Cytochrome p450s
MeJA	Methyl Jasmonate
TF	Transcription Factor
SBA	Soybean Aphid
SCN	Soybean Cyst Nematode
PI	Plant Introduction
Rag	Resistance to *Aphis glycines*
RNA-seq	RNA Sequencing
iDEP	Integrated Differential Expression and Pathway analysis
GWAS	Genome-Wide Association Study
SNPs	Single Nucleotide Polymorphisms
QTL	Quantitative Trait Loci
USDA	United States Department of Agriculture
DEGs	Differentially Expressed Genes
GO	Gene Ontology
THF	Tetrahydrofolate
$\log_2FC$	$\log_2$Fold Change

KEGG   Kyoto Encyclopedia of Genes and Genomes

Rhg   Resistance to *Heterodera glycines*

WGCNA   Weighted gene co-expression network analysis

PCA   Principal Component Analysis

FDR   False Discovery Rate

ANOVA   Analysis of variance

LSD   Least significance difference

dpi   Day post infestation

GEO   Gene Expression Omnibus

SRA   Sequence Read Archive

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

IDENTIFICATION AND CHARACTERIZATION OF STRESS RESPONSIVE GENES

IN SOYBEAN AND SUNFLOWER

SURENDRA NEUPANE

2019

Stress responsive genes encode proteins involved in plants' response to abiotic and biotic stresses. Among such stress responsive proteins, proteins encoded by resistance genes (R genes) or nucleotide binding site-leucine-rich repeats (NBS-LRRs) and mitogen-activated protein kinases (MAPKs) are the major groups of proteins regulating biotic and abiotic stresses, respectively. Previous studies in Nepal's lab at SDSU identified and characterized coiled coil (CC)-NBS-LRRs (CNLs), resistance to powdery mildew8 (RPW8)-NBS-LRRs (RNLs), NBS-LRR (NLs), and MAPK proteins in soybean. This study focuses on R and MAPK genes in the recently sequenced genome of sunflower as well as the toll-interleukin-1 receptor-like nucleotide-binding site leucine-rich repeat (TNL) R genes of soybean.

This study also uses greenhouse experiments and RNA sequencing (RNA-seq) data to characterize stress responsive genes involved in interaction effects of soybean aphid (SBA) and soybean cyst nematode (SCN) interactions on soybean. Thus the major objectives of this dissertation work were to 1) explore the TNL genes in soybean and R (CNL, TNL, RNL) genes in sunflower genomes to assess how they may have evolved and their possible role in resistance against pathogens using available transcriptomic data, 2) identify and characterize MAPK genes in sunflower, and 3) characterize induced susceptibility effects of soybean-soybean aphid and interaction effects of soybean-

soybean aphid-soybean cyst nematode on soybean. In this dissertation, we used *in silico*

approaches to report genome-wide identification and characterization of soybean TNL

proteins as well as sunflower R and MAPK proteins.

In order to achieve these objectives, numerous bioinformatics tools were utilized:

hidden markov model (HMM) profilings were performed, and annotation of protein

domains were conducted. Maximum Likelihood phylogenetic trees were constructed, and

nonsynonymous substitutions per nonsynonymous site to synonymous substitutions per

synonymous site ratios (K_a/K_s) as a proxy for selection pressure of R genes were

calculated. In addition, chromosomal distribution, intron-exon architecture; synteny as

well as gene expression patterns were assessed. In order to characterize stress responsive

genes involved in defense responses, we used soybean aphid (*Aphis glycines*; SBA) and

soybean cyst nematode (*Heterodera glycines*; SCN) to infest soybean cultivars. We

conducted greenhouse experiments to characterize induced susceptibility effects of

soybean-SBA interaction, and three-way interactions among soybean, SBA, and SCN.

We utilized both demographic and genetic (RNA-seq) datasets to characterize the genes

involved in such interactions using biotype 1, biotype 2 soybean aphids and HG type 0

SCN on soybean. FastQC, Btrim, Trimmomatic, Salmon, iDEP, MapMan tools were used

to assess the quality, trim, map, assemble, visualize, pathway analysis and biological

significance of RNA sequencing data to host genome.

We identified an inventory of 117 of 153 regular TNL genes in soybean, and 352

NBS-encoding genes (100 CNLs, 77 TNLs, 13 RNLs, and 162 NLs), 28 MPKs and eight

MKKs in sunflower through *in silico* analyses. R genes in soybean and sunflower formed

several gene clusters suggesting their origin by tandem duplications. The selection

pressure analysis revealed R genes experiencing purifying selection ($K_a/K_s < 1$) in both soybean and sunflower. Sunflower MAP Kinases revealed within and between clade functional divergence, and MKK3 orthologues were highly conserved across the species representing diverse taxonomic groups of the plant kingdom.

Demographic data obtained from greenhouse experiments showed that induced susceptibility as initial feeding with virulent SBA (biotype 2) increased the population of subsequent avirulent SBA (biotype 1) in both susceptible and resistant cultivars. In the three-way interaction among soybean, SBA, and SCN, the number of SCN eggs was significantly greater on the susceptible cultivar and there was no effect in the resistant cultivar in the presence of SBA. The SBA population density was negatively affected by SCN populations. RNA-seq analysis in both studies have revealed differentially expressed genes (DEGs) and transcription factor (TF) binding motifs, which were enriched for various biological processes and pathways at different time points. The DEGs were common and unique in susceptible and resistant cultivars and treatments that were enriched for various biological processes and pathways. These DEGs were also functionally related to known defense mechanisms previously reported in various host-aphid and host-nematode systems. The responses to aphid biotype 1 infestation in the presence or absence of inducer population (biotype 2) at two time points (day1 and 11 post inducer infestation) revealed significant differences on the gene enrichment and regulation in SBA resistant and susceptible cultivars. For instance, enrichment analysis showed 'response to chitin', 'lignin catabolic and metabolic process', 'asparagine metabolic process', 'response to chemical' unique to treatment with no inducer population, whereas, 'response to reactive oxygen species', 'photosynthesis', 'regulation

of endopeptidase activity' unique to treatment with inducer population. Likewise, Soybean-SBA-SCN interaction study showed enrichment of genes in 'Plant Pathogen Interaction' and 'cutin, suberine, and wax biosynthesis' pathways at 5 (days post SBA infestation) dpi; 'isoflavonoid biosynthesis' and 'one carbon pool by folate' pathways enriched at 30 dpi in SCN resistant and susceptible cultivars. Overall, the results from this study have improved the current understanding of diversity and evolution of MAPK and R genes in sunflower and soybean, as well as have first time reported a molecular characterization of induced susceptibility effects due to SBA on soybean, and soybean-SBA-SCN interactions, which has a direct implication in disease and pest management.

CHAPTER 1: LITERATURE REVIEW

1.1 Soybean

Soybean [*Glycine max* (L.) Merr.], a source of high quality sugar, protein, and oil, is one of the most important crops worldwide [1]. Cultivated soybean was domesticated from its wild relative, *Glycine soja* (Sieb. and Zucc.), approximately 5000 years ago in Southern China (primary gene center) [2, 3]. Later, cultivated soybean was introduced into many Asian countries such as Korea and Japan (2,000 years ago), Indonesia, Vietnam, Philippines, Thailand, Malaysia, Burma, Nepal, and India, which are considered as the secondary gene centers [2, 4]. Soybean was first introduced to North America in 1765 for manufacturing soy sauce and vermicelli (soybean noodles) [5]. In 1770, Benjamin Franklin sent soybean seeds from London to John Bartram in Philadelphia, which were used as a forage and ground cover [5]. In 1915, soybean was first used for the production of oil in North Carolina, and thereafter, soybean has been considered as the major oilseed crop [2].

According to the United States Department of Agriculture, soybean is considered as the second major crop in terms of production and acreage (USDA NASS-ERS, 2017). In 2017, the U.S. produced 119.5 million metric tons (MMT) worth $41.01 billion and contributed to 35% of the world soybean production (http://soystats.com). This makes the U.S. the lead producer of soybean followed by Brazil (33%), Argentina (14%), China (4%), India (3%), Paraguay (3%), and Canada (2%) in 2017, which indicates that 85% of the total soybean production in the world is produced in North and South America (http://soystats.com). In 2017, South Dakota produced 2.9 MMT (43 Bushels/acre) worth of $2,147 Million planting soybean in 5,650 thousand acres (http://soystats.com).

1.2. Sunflower

Sunflower (*Helianthus annuus* L.), first domesticated in North America, is the fourth most important oilseed crop in the world (FAO, 2010). According to USDA Reports of 2018, in the U.S., sunflower crop is grown in California, Colorado, Kansas, Minnesota, Nebraska, North Dakota, South Dakota, and Texas. Furthermore, South Dakota is often the leading sunflower producing state (with a production of 1.04 billion pounds in 2017). In 2017, the sunflower yield was 1,613 pounds per acre, the third highest on record even though 118 pounds down from 2016 (as of USDA Jan. 12, 2018; http://www.sunflowernsa.com). Since sunflower has the capacity to maintain stable yields in different environmental conditions such as drought, it has been a model crop species for studying climate change adaptation [6]. The study on diversity analysis of 128 expressed sequenced tag (EST)-based microsatellites in wild *H. annuus* has provided insights into the ability to adapt salt and drought stress and selective sweeps revealing transcription factors as the major group of genes involved in those processes [6]. In addition, studies on wild and cultivated relatives of sunflower on disease resistance [7] and oil content [8] provide insights into the genetic background for these traits to be used in breeding. However, many fungal diseases like charcoal rot (*Macrophomina phseolina*), downy mildew (*Plasmopara halstedii*), *Fusarium* rot and stem rots (*Fusarium* sp.), phoma black stem (*Phoma macdonaldii*), phomopsis stem canker (*Diaporthe helianthi*, *D. gulyae*), *Sclerotinia* mid and basal stem rot (*Sclerotinia scelerotiorum*), *Verticillium* wilt (*Verticillium dahlia*), leaf blight (*Alternariaster helianthi*), leaf spot (*Pseudomonas syringae* pv. *helianthi*), powdery mildew (*Erysiphe cichoracearum*), rust (*Puccinia helianthi*) and many others have caused crop damage resulting in the loss of yield and oil

content [9]. This underlies the growing need of rigorous research into the disease resistance in sunflower.

1.3. Resistance (R) Genes

In response to various biotic and abiotic stresses, plants involve different gene families in signaling networks for the protection [10]. To face different kinds of stresses, plants have developed multifaceted mechanisms to percieve and transfer signals from various stimuli during the course of evolution [11]. Various efforts have been made to study the NBS-LRR group of resistance genes, which are considered as the major disease resistance (R) gene family [12, 13, 14, 15]. These R genes are crucial in recognizing and binding with the effector molecules and trigger downstream signaling in resistance pathways [16, 17]. Two major classes of R genes are toll-interleukin-1 receptor-like nucleotide-binding site (NBS) leucine-rich repeat (LRR) proteins or TNL genes, and coiled coil (CC)-NBS-LRR or CNL genes. Shao et al. (2016) [18] studied NBS-LRR genes of the angiosperms on a large scale, dividing them into three classes [TNLs, CNLs and R (resistance to powdery mildew8) NLs or RNLs]. The NB-ARC [for APAF1 (apoptotic protease-activating factor-1), R (resistance genes), and CED4 (*Caenorhabditis elegans* death-4 protein)] domain hydrolyzes ATP to induce the conformational change in R proteins acting as the nucleotide binding pocket [19]. The LRR domains help in activating or deactivating the defense signaling by interacting with the NB-ARC domain in the presence or absence of pathogen effectors, respectively [20].

1.4. Mitogen Activated Protein Kinase (MAPK) Genes

Often cross-linked with the disease resistance pathways is the MAPK signaling cascade. The MAPK signaling cascade has been the universal module and highly

conserved signal transduction component in the eukaryotes [21, 22]. The MAPK cascade consists of three main sub families based on the structural features [MAPK (MAPK/MPK), MAPK kinase (MAPKK/MKK), and MAPK kinase kinase (MAPKKK/MKKK)], and is involved in a series of phosphorylation events contributing to signaling [23, 24, 25]. The phosphorylation takes place by adding a phosphate group from adenosine triphosphate (ATP) to the downstream substrate proteins [26]. MAPKKK, the largest group of the MAPK cascade are typically serine or threonine protein kinases that phosphorylate MKKs in the conserved S/T-X3-5-S/T motif and possess K/R-K/R-K/R-X1-6-L-X-L/V/S, MAPK-docking domain [10, 22, 26]. The MAPKKK are classified into three subfamilies, MEKK, Raf and Zik on the basis of difference of conserved kinase domain: the MEKKs have G(T/S)Px(W/Y/F)MAPEV domain, the Rafs have GTxx(W/Y)MAPE, and the Zik have GTPEFMAPE(L/V)Y domain [24]. The MEKKs have kinase domain either at C- or N-terminal, the Rafs have N-terminal regulatory domain and C-terminal kinase domain, and the Ziks have N-terminal kinase domain [11]. The MPKs family possess TDY or TEY phosphorylation motifs, which provide a protein-binding domain for MPKs activation in between the VII and VIII kinase domains and consist of 11 conserved kinase domains [22, 27, 28]. An overview of the MAP Kinase signaling pathway and R genes in response to diverse abiotic and biotic stresses in plants is represented in Figure 1.1 [25].
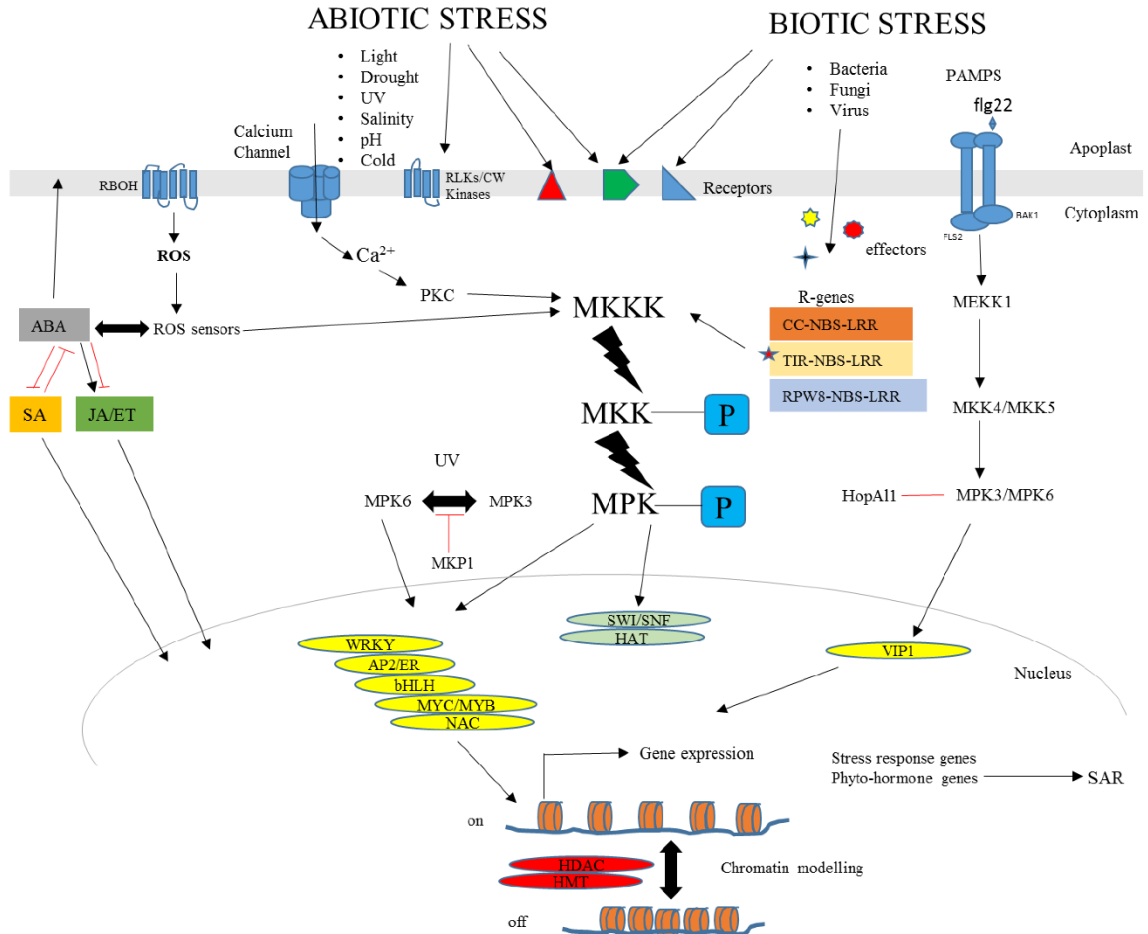
Figure 1.1. MAP Kinase signaling pathway in response to abiotic and biotic stresses in plants (adapted from multiple studies [29, 30, 31, 32, 33, 34]).

1.5. *Aphis glycines* Matsumura

Soybean aphids, *Aphis glycines* Matsumura (Hemiptera: Aphididae), is the most economically important insect pest of soybean [*Glycine max* (L.) Merr] [37]. It is a greenish, pear shaped insect and a size of approximately 1.5mm [38, 39]. It was first reported in North America in Wisconsin (USA) in 2000 [40].

1.5.1 Life Cycle of *A. glycines*

Aphis glycines has a heteroecious (spends life cycle period parasitizing two very different species of host plant) and holocyclic (undergoes sexual reproduction during at

least part of its life cycle) life cycle [41]. The life cycle of *A. glycines* starts on *Rhamnus* spp. in the spring season. It utilizes *Rhamnus* spp. for sexual reproduction and overwintering as an egg [42]. These eggs can withstand temperatures as far as -34 °C [43]. The nymphs emerge on the *Rhamnus* spp. after eggs hatch during spring. These nymphs give rise to wingless fundatrices (mature wingless stem mother, which hatches from over-wintering eggs) and continue to produce few generations of *A. glycines* on the primary host [37, 44]. After a few generations in the primary host, they develop into alates (winged morphs). These alates migrate to soybean plants by late spring or early summer. *A. glycines* becomes almost undetectable on the early season of soybeans as the alternate host, soybean plants are prominently available in the late spring and early summer [45]. During this process, they deposit the nymphs in the soybean plant and reproduce asexually as many as fifteen generations of apterous and alate morphs [46]. After generations, winged offspring arise as the population starts to increase and target other soybean plants for colonization [44]. The aphid population can double in one and half days under favorable conditions, however, doubling time in fields is up to approximately seven days [46]. The optimal temperature for the soybean aphid is 27.8 °C. The reproduction slows down as temperatures increase or decrease and eventually stops when temperatures are greater than 34.9°C or less than 8.6°C [38, 43]. The reduced temperature and photoperiod in the late summer induce the production of gynoparae (winged females). These winged females migrate to the primary host, *Rhamnus* spp. They feed on *Rhamnus* spp. and produce nymphs and developed into apterous oviparae. During the fall season, the male alates, androparae produced in the soybean, start to travel and find the oviparae on the *Rhamnus* spp. These androparae find oviparae to mate, which is

the only sexual reproduction stage in the life cycle of *A. glycines*. The oviparae lay eggs on bud shoots of *Rhamnus* spp. from October to mid-November. They overwinter and hatch the eggs in late March until two parthenogenic generations [44].

Upon an infestation of soybean plants by soybean aphids, they prefer to feed on the ventral side of the leaves mainly in young trifoliate leaves [47]. They feed on the phloem sap and draw assimilates from soybean plants [46, 48]. This results in plant stunting, leaf yellowing and wrinkling with a reduced photosynthesis, poor pod fill, and reduced yield (up to 40%), seed size, and seed quality [46, 49, 50]. Soybean aphids deposits honeydew on soybean leaves that aids as a vector for various viruses such as Soybean mosaic virus*,* Alfalfa mosaic virus, and Bean yellow mosaic virus [51, 52]. The economic loss due to the presence of aphid was estimated at approximately $4 billion annually [51]. For an effective management approach to control *A. glycines*, use of soybean lines that are naturally resistant to aphids can prove one of the best options without disturbing the natural environment [53].

1.5.2. Aphid Effectors

Effector molecules are ejected into the host cells either by type III secretions from bacteria, haustorium by the fungus and parasitic plants or by stylets by the nematodes and insects [16, 54]. These effector molecules help pests/pathogens colonize on the host plant [55]. The survival, growth, and reproduction of pests/pathogens in the host cell depend on the ability of pests/pathogens to escape the recognition event during the host innate immunity [56]. Thus, these pests/pathogens generate variants of *Avr* effector molecules either by transposon insertions or mutation in effector coding genes or alternative splicing in gaining virulence to evade host defense [56, 57]. *A. glycines* uses two types of saliva,

gelling and watery saliva when feeding. The aphid injects the gelling saliva during the

early stages of feeding to form sheaths around the stylets [58]. Later, it injects effector

molecules with watery saliva into both the intra- and intercellular spaces of mesophyll

cells or directly into phloem cells [59]. Since the effector molecules allow each aphid to

sustain and modulate its host plant's immune reaction, they are subject to the scrutiny of

host defense mechanisms and undergo natural selection [60]. Such selection helps

effectors evade the host defense system, maintain their virulence, and evolve new

functions [61].

Transcriptomic and proteomic studies of the pea aphid (*Acyrthosiphon pisum*

Harris) found many salivary proteins undergoing positive selection [62]. Aphid effectors

are host specific so that they can effectively interact with the host proteins for their

virulence [63, 64]. For instance, Rodriguez et al. (2017) [63] reported that Mp1, an

effector molecule produced by the green peach aphid (*Myzus persicae* Sulzer),

specifically targets Vacuolar Protein Sorting-Associated Protein 52 (VPS52) proteins in

their strong hosts. Such interaction is absent in the green peach aphid's poor-hosts.

Furthermore, the reproduction of the green peach aphid did not increase in *Arabidopsis*

that expressed the orthologs of the pea aphid's effectors, including C002, PIntO1 (Mp1),

and PIntO2 (Mp2) [64]. Since the identification and functional characterization of the

first aphid effector molecule, C002 in the pea aphid [65], significant progress has been

made in identifying a wide range of effector molecules in different aphids. The

availability of the whole genome sequences of several aphid species, including the pea

aphid [66], Russian wheat aphid (*Diuraphis noxia* Kurdjumov) [67], green peach aphid

[68], and soybean aphid [69], have allowed the study of various gene families of aphid

salivary effectors. Carolan et al. (2011) [70] identified 324 secretory proteins in the salivary gland of pea aphid. Some proteins, including Glucose dehydrogenase, Glutathione peroxidase, putative sheath protein of aphids, and Angiotensin-converting enzyme-like, showed similarity to some known aphid effectors [71, 72, 73], while others were more similar to nematode effectors, including M1 zinc metalloprotease, Disulfide isomerase, Calreticulin, ARMET, Glutathione peroxidase, and CLIP-domain serine protease [70, 74, 75]. The 'pea aphid effector' proteins were further expanded to 3,603 genes expressed in salivary glands, 740 of which were up-regulated in salivary glands compared to alimentary tract and belonged to the Cysteine-Rich Protein (CRP), Angiotensin-Converting-Enzyme-like (ACE) gene, and Aminopeptidase-N (apN) gene families [62]. Thirty-four salivary genes were identified in the Russian wheat aphid that were similar to the most commonly expressed genes in other aphids, including glucose dehydrogenase and trehalase [67]. An intensive analysis of the genome of the green peach aphid, which can infest plant species belonging to 40 families, demonstrated the role of multigene clusters in colonizing distant plant species [68]. This study suggested the genes belonging to cathepsin B and RR-2 cuticular protein gene families undergo rapid transcriptional plasticity so that the aphids can infest a wide range of plant species belonging to the Brassicaceae and Solanaceae family.

RNA sequencing (RNA-seq) has been a standard tool for studying qualitative and quantitative gene expression [76, 77]. In the context of the soybean aphid, Bansal et al. (2014) [78] studied xenobiotic stress responses in the soybean aphid using RNA-Seq. This study reported 914 significantly expressed genes in the soybean aphid, most of which were related to stress, detoxification [cytochrome p450s (CYPs), glutathione-S-

transferases, carboxyesterases], and ABC transporters. Wenger et al. (2017) [69]

identified 135 putative soybean aphid effector genes, including 68 CYP protein-coding

genes (detoxification genes), 82 genes belonging to ABC transporter subfamilies, 14

glutathione-S transferases, and 17 carboxyl and choline esterases. The detoxification

genes help aphids adapt to host plants [78]. The small number of CYP genes found in the

soybean aphid, the pea aphid (83 CYP genes), and the Russian wheat aphid (48 CYP

genes) might explain why these species are adapted to a limited range of hosts, while the

green peach aphid (115 CYP genes) is adapted to wide host ranges [79]. The availability

of genome sequences of the soybean aphid might explain the species' rapid adaptation to

resistant soybean cultivars despite the lack of both genetic differentiation and selection

pressure between avirulent and virulent biotypes [80].

1.5.3. *Aphis glycines* Biotypes

A biotype is an insect population that can reproduce and survive in cultivars

developed for resistance to that same population [81]. It is a pseudo-taxonomic unit that

classifies insect populations according to their virulence to specific cultivars and shared

phenotype (reviewed in [82]). This term has been used for various insect species,

including *Mayetiola destructor* Say, *Schizaphis graminum, Nilaparvata lugens*, and

Bemisia tabaci [81]. The insect subpopulations capable of surviving in resistant crop

lines, including wheat, barley, melon, and apple, have been studied (reviewed in [83]).

Soybean aphids that are avirulent on any soybean plant that contains the Resistance to

Aphis glycines (*Rag*) gene are attributed to biotype 1 [83]. Biotype 1 is the predominant

biotype of *A. glycines* in North America [84]. Biotype 2 (*Rag1* virulent) was discovered

in Ohio in 2005, five years before the release of commercial *Rag1* cultivars [85]. The

biotype 2 aphids were thought to be the predominant biotype in eastern North America [85], but various field tests found that they were prevalent only in Ohio [86]. Since then, two additional biotypes of soybean aphid have been discovered in the U.S., suggesting the North American populations possess sufficient genetic variability to adapt to the resistant hosts [85]. The biotype 3 aphids discovered in Indiana were able to reproduce on *Rag2* soybean plants but were poorly adapted to *Rag1* soybean plants [87]. Later, biotype 4 aphids were found in Wisconsin that can reproduce in both *Rag1* and *Rag2* soybean plants [88]. Cooper et al. (2015) [83] studied the geographic distribution of the soybean aphid biotypes across 11 states and one Canadian province between 2008 and 2010. The frequency of aphid populations belonging to biotypes 2, 3, and 4 was 54%, 18%, and 7%, respectively. The aphid populations from Wisconsin, the state where the soybean aphid was first reported in the U.S. in 2000, showed higher virulence variability [83]. Additionally, Zhong et al. (2014) [89] reported at least four biotypes of soybean aphid in China. These biotypes were named as China biotype 1 (virulence on host plants with *Rag5* or *Rag6*), China biotype 2 (virulence on host plants with *Rag1*, *Rag3* or *Rag5*), China biotype 3 (virulence on host plants with *Rag1*, *Rag3*, or *Rag6*), and China biotype 4 (virulence on host plants with *Rag1*, *Rag2*, *Rag3*, or *Rag5* genes) [89].

1.5.4. Soybean Cultivars Exhibiting Antibiosis, Antixenosis, and Tolerance as a Resistance Response to Soybean Aphids

According to Painter (1951) [90], plant resistance mechanisms to insects can be grouped into three categories: antibiosis, antixenosis and tolerance. Antibiosis resistance affects the biology, including the mortality or fecundity, of the insect. The soybean cultivar 'Dowling' exhibits antibiosis, and resistance factors are present in the phloem

cells [91]. Antixenosis resistance affects the behavior of the insect. The soybean cultivar
PI200538 exhibits antixenosis [91]. Jesus et al. (2018) [92] studied the physiological
responses of 14 soybean genotypes to aphid infestation in terms of total protein,
peroxidase, chlorophyll, and resistance mechanisms. The genotypes UX 2569-1592-01
(*Rag2* gene; PI243540) and UX 2570-171-04 showed the highest and moderate level of
antibiosis and/or antixenosis, respectively. The chlorophyll content in UX 2569-159-2-01
was reduced at five and 15 days after infestation. Total protein content remained
unchanged between the infested and control plants. Peroxidase activity in UX 2570-171-
04 was higher at 5 and 10 days after infestation, and this cultivar showed a moderate
level of antibiosis and/or antixenosis. Tolerance (experience lower selection pressure than
antibiosis and antixenosis) is the ability of the plant to endure the presence of the insect
without significant impacts on the pest's biology or behavior [93]. The KS4202 cultivar
is tolerant of aphids [94]. The tolerance effect in KS4202 may be attributable to the quick
regulation of RuBP (ribulose-1,5-biphosphate) and the upregulation of detoxification
genes [95].

1.5.5. *Rag* Genes in Soybean Cultivars Provide Resistance to *A. glycines*

Rag (resistance to *Aphis glycines*) loci were first discovered in Dowling, PI71506,
and Jackson cultivars [96], and have since been identified in other soybean lines. The
mapping and inheritance mechanism of the *Rag1* gene have been documented in multiple
soybean cultivars [97, 98, 99, 100]. *Rag1* loci were mapped as a 115 kb interval on
chromosome 7 using the Dowling (PI548663; donor parent of *Rag1*) and Dwight
(PI587386; aphid-susceptible parent) cultivars [101]; *Rag2* loci mapped as a 54 kb
interval on chromosome 13 in the antixenotic PI200538 cultivar [91, 102]; *Rag3* loci

mapped on chromosome 16 (LG J) using PI567543C [103]; and the recessive *rag4* loci were mapped on chromosome 13 (LG F) between markers in PI567541B [104]. The authors also mapped *rag1* provisional (*rag1c*) on chromosome 7 (LG M). *Rag5* (proposed) and *Rag6* have been identified in PI567301B and PI567598B, respectively [105, 106]. Bhusal et al. (2017) [107] identified two major and two minor loci: the major loci were located on chromosome 7 (*qChrom.07.1*) (1Mb distant from *Rag1*) and chromosome 16 (*qChrom.16.1*) (near *Rag3*), and the minor loci were located on chromosome 13 (*qChrom.13.1*) (near *Rag4*) and chromosome 17 (*qChrom.17.1*). The minor loci were associated with aphid resistance in PI603712. Hill et al. (2017) [108] characterized multiple *A. glycines* biotype resistances in five cultivars: PI587663 and PI594592 had resistance genes located in the *Rag1, Rag2*, and *Rag3* regions, PI587677 had resistance genes in the *Rag1, Rag2*, and *rag4* regions, PI587685 had resistance genes in the *Rag1* and *Rag2* regions, and PI587972 had resistance genes only in the *Rag2* region.

More than half of the genetic diversity has been lost in the cultivated soybean [109], but its closest wild relative, *Glycine soja* Siebold & Zucc., may offer opportunities for identifying aphid-resistance genes, studying inheritance patterns, and mapping important resistance loci [84]. Hesler and Tilmon (2018) [110] reported PI135624 and PI65549 were resistant to aphids, and Conzemius (2018) [111] reported PI101404A and PI65549 showed significant high resistance to biotype 4 colonies. *Rag6* and *Rag3c* were mapped in 49-kb (42,146,252–42,195,720 bp) and 150-kb intervals (6,621,540–6,771,675 bp) on chromosome 8 and chromosome 16, respectively, in *G. soja* [112]. The 49–kb interval, where *Rag6* was mapped, contained three clustered NBS–LRR genes

(*Glyma.08g303500*, *Glyma.08g303600*, and *Glyma.08g303700*) and one amine oxidase gene (*Glyma.08g303800*). The 150-kb interval, where *Rag3c* was mapped, contained one LRR gene (*Glyma.16g066800*) and other ten genes belonging to lipase, cytochrome P450, methyltransferases, hydrolases, and Ku70-binding gene families. All identified *Rag* QTLs in various soybean plant introductions (PI) are presented in Table 1.1.

All 1,691 non-redundant genes assessed from the *Rag* QTLs, including *Rag1*[101], *rag1b* [106], *rag1c* [104], *Rag2* [113], *Rag3* [103, 114] , *Rag4* [114], *rag3* [106] , *rag3b* [115], *Rag3c* [112], *rag4* [104], *Rag5* [105] , *Rag6* [112]; *qChrom.07.1*, *qChrom.16.1*, *qChrom.13.1*, and *qChrom.17.1* [107] are significantly associated with 'nutrient reservoir activity' (GO:0045735) and 'binding' (GO:0005488). The 'nutrient reservoir activity' molecular function is important in protecting plant tissues that produce surface waxes [116]. Similarly, the 'binding' molecular function occurring at a higher proportion suggests their important roles in signaling and stress responses. The genes engaged in the process of binding (GO: 0005488) belong to binding to ADP (GO: 0043531), adenyl ribonucleotide (GO: 0032559), calcium-dependent phospholipid (GO: 0005544), adenyl nucleotide (GO: 0030554), purine nucleoside (GO: 0001883), nucleoside (GO: 0001882), pattern (GO: 0001871), and polysaccharide (GO: 0030247) binding gene families (Figure 1.2).
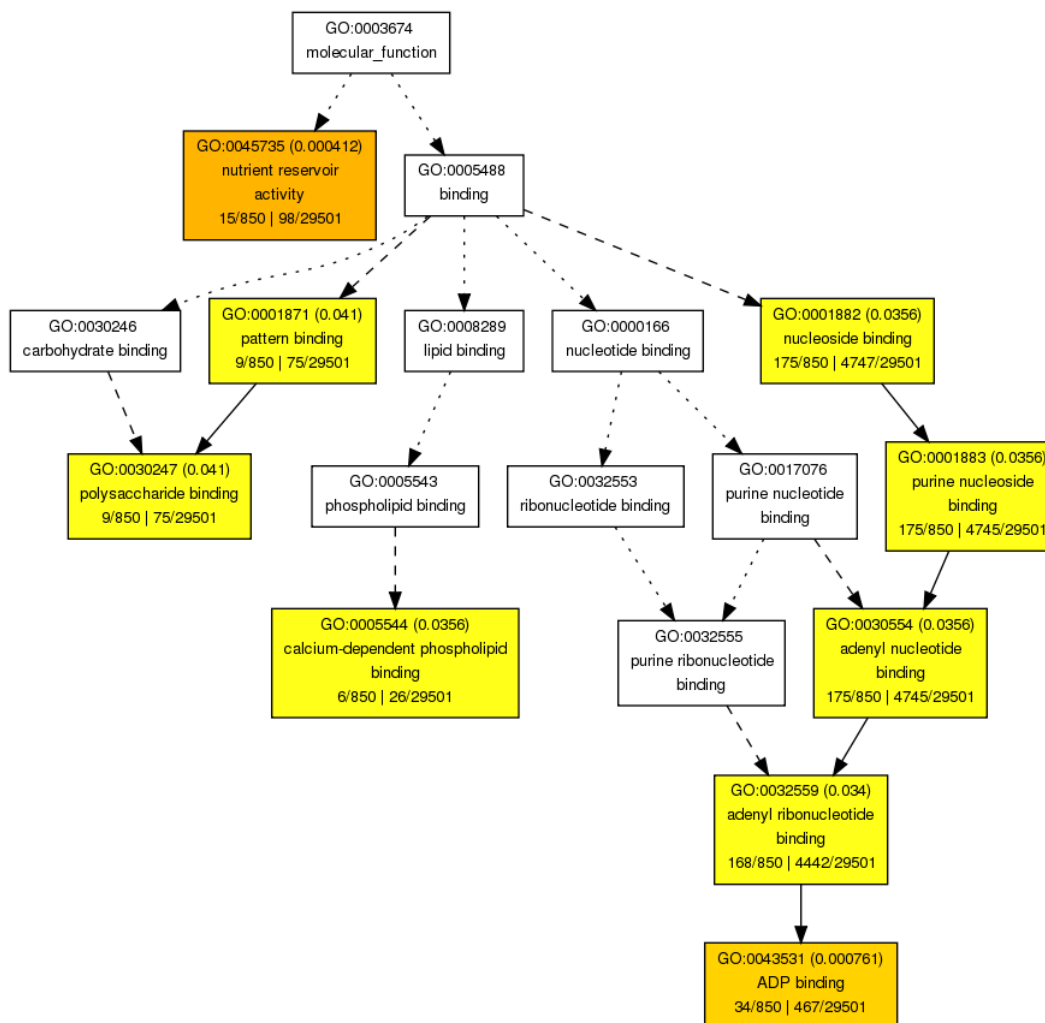
Figure 1.2. Significantly enriched GO molecular function terms of non-redundant 1,691 genes in the *Rag* QTLs: *Rag1* [101], *rag1b* [106], *rag1c* [104], *Rag2* [113], *Rag3* [103, 114], *Rag4* [114], *rag3* [106] , *rag3b* [115], *Rag3c* [112], *rag4* [104], *Rag5* [105] , *Rag6* [112]; *qChrom.07.1*, *qChrom.16.1*, *qChrom.13.1*, *qChrom.17.1* [107] as determined by Fisher's exact test using AgriGO [117]. The same gene can be associated with multiple GO annotations. Only significantly ($P < 0.05$) over-represented GO categories are shown. The stronger color represents the lower *P* value. Information in the box includes GO term, adjusted P value in parentheses, GO description, a number of query list/background mapping GO, and a total number of query list/background.

Table 1.1. Soybean cultivars for mapping *Rag* genes with the information on chromosome location, markers associated, and type of resistance. ($^\gamma$ = Marker positions are based on Glyma 2.0 [114])

QTLs	Soybean Plant Introductions	Chromosome (Linkage group)	Markers associated (Location$^\gamma$)	Type of resistance	References
Rag1	Dowling (PI548663)	7 (M)	46169.7and 21A (5,529,532-5,770,718 bp)	Antibiosis	[97]
	PI71506	7 (M)		Antixenosis	[96]
	PI548663 (cultivar Dowling)	7 (M)	Satt435 and Satt463		[95]
	PI548657 (cultivar Jackson)	7 (M)	Satt435 andSatt463		
	PI587663	7 (M)	Satt567 and Satt245	Antibiosis	[104]
	PI587677	7 (M)	Satt540	Antibiosis	[104]
	PI587685	7 (M)	Satt540	Antibiosis	[104]
	PI594592	7 (M)	Satt540	Antibiosis	[104]
rag1c	PI567541B	7 (M)	sat229-satt435 (2,434,259-8,234,168 bp)		[100]
rag1b	PI567598B	7 (M)	Satt567 and Satt435 (5,523,128-5,909,485 bp)		[102]
Rag2	PI243540	13 (F)	Satt334 and Sct_033(28,415,888–30,739,587 bp)	Antibiosis	[109]
	PI200538	13 (F)	Satt510, Soyhsp176, Satt114, and Sct_033 (29,609,521– 31,802,676 bp)	Antibiosis	[98]
	PI587663, PI587685	13 (F)	Satt114, SNP2, Satt335		[104]
	PI587677	13 (F)	Satt335		[104]
	PI587972	13 (F)	Satt114, Satt510		[104]
	PI594592	13 (F)	Satt114		[104]
Rag3	PI567543C	16 (J)	Sat_339 and Satt414 (4,964,852-7,212,164 bp)	Antixenosis	[99]
	PI587663	16 (J)	Satt285	Antibiosis	[104]
	PI594592	16 (J)	Satt654	Antibiosis	[104]
	PI567543C	16 (J)	ss715625290 and ss715625308 (6,314,060-6,571,305 bp)		[110]
rag3	PI567598B	16 (J)	Satt285 and Satt414 (6,314,120-6,570,336 bp)		[102]
rag3b	PI567537	16 (J)	4,964,852- 7,957,026 bp	Antibiosis	[111]
Rag3c	E12901	16 (J)	Gm16-3 and Gm16-5 (6,621,540–6,771,675 bp)	Antibiosis	[108]
rag4	PI567541B	13 (F)	Satt649-Satt343 (1,225,665-16,340,514 bp)	Antibiosis	[100]
	PI587677	13 (F)	Satt586		[104]
Rag4	PI567543C	13(F)	MSUSNP13-29-ss247923149 (13,691,537-13,626,971 bp)		[110]
Rag5 Proposed	PI567301B	13	4 SSR markers (30,236,183-30,749,047 bp)	Antixenosis	[101]
Rag6	E12901	8	Gm08-15 and Gm08-17 (42,146,252–42,195,720 bp)	Antibiosis	[108]
qChrom.07.1	PI603712	7(M)	ss715598483-ss715598534 (6,444,246-6,819,959 bp)		[103]
qChrom.16.1	PI603712	16(J)	ss715625261-ss715625278 (6,105,250-6,222,257 bp)		[103]
qChrom.13.1	PI603712	13(F)	ss715613721-ss715617240 (13,691,537-13,626,971 bp)		[103]
qChrom.17.1	PI603712	17(D2)	ss715627556-ss715627637 (39,019,814-39,521,449 bp)		[103]

1.5.6. GWAS Studies on *A. glycines* Resistance in Soybean Expanding to a Number of QTLs

Genome-Wide Association Studies (GWAS) have been an important alternative to classical bi-parental QTL mapping [119] for understanding the genetic basis of diseases linked to complex, polygenic traits. While classical QTL mapping is limited in its ability to identify allelic diversity and resolve genomes [120], GWAS can capture all the recombination events occurred during the evolution of sampled genotypes [121]. Different kinds of phenotypes, including quantitative, binary, and ordinal phenotypes, can be studied using GWAS [122] and can be correlated with genotypes using mixed linear models [123]. Chang and Hartman (2017) [124] reported the first GWAS study for aphid-resistance using USDA soybean germplasms. The authors suggested that ss715596142 may be a significant SNP marker and identified three LRR domain containing genes (*Glyma07g13440*, *Glyma07g14810*, and *Glyma07g14791*) along with one MYB transcription factor (*Glyma07g14480*). This marker is close to the *rag1c* gene that was reported in PI567541B [104], but not the *Rag1* locus that contains the candidate LRR genes (*Glyma07g06890* and *Glyma07g06920*) [101]. Hanson et al. (2018) [118] reported significant SNPs on chromosomes 7 (close to *Rag1* and *rag1b* within *rag1c* for biotype 2 resistance), 8 (424 kbp from *Rag6* for aphid biotype 3 resistance), 13 (within range of *Rag2* and *Rag5* for aphid biotype 2 resistance), and 16 (for aphid biotype 1 resistance), where *Rag* genes have been mapped previously, for multiple aphid biotypes. Additionally, they reported markers on chromosomes 1-2, 4-6, 9-11, 12, 14, and 16-20 where *Rag* genes had not been previously reported.

1.5.7. *Rag* Gene Pyramiding Provides Resistance to all *A. glycines* biotypes

The presence of fitness costs associated with aphid virulence in the *Rag* soybean cultivars could be used to preserve the efficacy of resistance genes in *Rag* soybean cultivars [125, 126]. In addition, the use of refuge susceptible soybean plants might limit the frequency of virulent biotypes [125]. It has been proven that soybean aphids are more virulent in cultivars with a single *Rag* gene than those with pyramided genes [53]. The pyramiding of resistance genes in the soybean cultivars protects the plants from the various aphid biotypes [127, 128]. The first soybean cultivar with both *Rag1* and *Rag2* loci became commercially available in 2012 and was resistant to aphid biotypes 2 and 3 [129]. After aphid Biotype 4 was found, the need for pyramiding more genes became imminent. The pyramiding of *Rag1*, *Rag2*, and *Rag3* resistance genes may provide resistance to all known aphid biotypes [126, 128].

1.5.8. Transcriptomic Studies on Soybean-*A. Glycines* Interaction: Jasmonic Acid (JA) and Abscisic Acid (ABA) Signaling Pathway Plays a Crucial Role in Plant Resistance

Several studies have described differential changes in phytohormones that occur during aphid-feeding in resistant, tolerant, and susceptible cultivars [130, 131, 132, 133, 134]. Cyclical expression patterns of the different marker and responsive genes for salicylic acid observed in aphid-infested plants suggests these hormones play a key role in soybean resistance to aphid feeding [131]. Furthermore, an application of methyl jasmonate (MeJA) on infested plants significantly decreased soybean aphid population, but a similar salicylic acid application did not affect the aphid population; this suggests MeJA may be the elicitor to induce plant defenses [131]. Thus, the JA signaling pathway

that assists the induction of other enzymes, including polyphenol oxidase (PPO), lipoxygenases, peroxidases, and proteinase inhibitors, appears to play a crucial role in aphid-resistance against susceptible soybean cultivar [131, 135].

Brechenmacher et al. (2015) [93] used two *Rag2* and/or *rag2* near-isogenic lines of soybean to identify 396 proteins and 2,361 genes that were differentially regulated in response to soybean infestation. Several genes mapped within the *Rag2* locus, including a gene of unknown function (*Glyma13g25990*), a mitochondrial protease (*Glyma13g26010*), and an NBS-LRR (*Glyma13g25970*), were significantly upregulated in the presence of aphids. Prochaska et al. (2015) [94] identified three and 36 differentially expressed genes (DEGs) at five and 15 days after infestation, respectively, in the resistant (tolerant) KS4202 cultivar but found only zero and 11 DEGs at five and 15 days after infestation, respectively, in the susceptible K-03-4686 cultivar. Most of the DEGs were related to WRKY transcription factors (such as WRKY60), peroxidases [*Peroxidase 52* (*PRX52*) and *Ascorbate peroxidase 4* (*APX4*)], and cytochrome p450s. Aphid-tolerance mostly depended on the constitutive levels of abscisic acid (ABA) and jasmonic acid (JA) and the basal expression of ABA (*NAC19* and *SCOF-1)* and JA [*LOX10*, *LOX2* (a chloroplastic-like linoleate 13S-lipoxygenase 2), *OPDA-REDUCTASE 3* (*OPR3*)]-related transcripts [130]. In addition, the genes *PRX52*, *WRKY60*, and *PATHOGENESIS-RELATED1* (PR1; SA-responsive transcript) were found to be induced by aphid infestation in the tolerant KS4202 cultivar [130]. Lee et al. (2017) [136] evaluated the transcriptomic dynamics of soybean near-isogenic lines (NILs) with either the *Rag5* allele for resistance or the *rag5* allele for susceptibility to the aphid biotype 2. Three genes located near the *Rag5* locus, including *Glyma.13 g190200*, *Glyma.13*

g190500, and *Glyma.13g190600*, were reported to be good candidate genes for imparting

soybean aphid resistance. Li et al. (2008) [133] studied soybean responses to aphid

infestation by using cDNA microarrays to generate transcript profiles and identified 140

genes related to the cell wall, transcription factors, signaling, and secondary metabolism.

Studham and MacIntosh (2013) [134] utilized oligonucleotide microarrays to study

soybean-aphid interactions in the aphid-resistant cultivar LD16060 with *Rag1* gene and

aphid-susceptible cultivar SD01-76R. They identified 49 and 284 differentially expressed

genes (DEGs) at one and seven days after infestation, respectively, in the susceptible

cultivar and found only 0 and 1 DEGs at 1 and 7 days after infestation, respectively, in

the resistant cultivar. They suggested that the expression of defense genes in resistant

plants is constitutive, whereas the defense genes in susceptible plants are expressed only

after aphid infestation.

1.6. *Heterodera glycines* Ichinohe

Soybean cyst nematode (*Heterodera glycines* Ichinohe) (SCN) is the most

distressing pest in the production of soybean (*Glycine max* (L.) Merr.) [137]. SCN

belongs to Kingdom Animalia, phylum Nematoda, class Chromadorea, order Tylenchida,

and family Heteroderidae. All *Heterodera* species belong to the nematodes that form the

cysts, thick walled dead female shielding eggs, on the roots [138]. Also, this genus is

considered as the most economically important group of the plant parasitic nematodes

[139]. It is an obligate, sedentary endoparasitic, soilborne nematode causing $1.3 billion

losses in soybean yield annually in the United States [140, 141]. Soybean yield losses

were approximately 3.4 million metric tons (125 million bushels) worth approximately

$1.6 billion in 2014 [142]. More than 30% yield loss caused by SCN remains unnoticed

because of the unnoticeable aboveground symptoms and sometimes confused with the symptoms caused by viral pathogens [142]. SCN remained on top among ten most destructive diseases in the northern United States and Ontario though 2010 (110 million bushels) to 2014 (108 million bushels) [142]. Thus, SCN is an important pest which unceasingly threatens soybean production [143].

1.6.1. Origin and Distribution of SCN

After SCN being first reported in Japan in 1915 [144], it was later reported in Korea [145], China [146], and U.S. [147]. In the United States, it was first reported in Hanover County of North Carolina in 1954 [147]. The source for the first SCN to arrive in the United States is unknown but believed to have imported via plant or soil material [148]. SCN was reported in Missouri and Tennessee in 1956, Arkansas, Kentucky and Mississippi in 1957, and Virginia in 1958 [149]. Since the detection of SCN in Union County in South Dakota in 1995, it has spread to 30 counties of South Dakota [150]. Now, SCN has been detected in 90% of the soybean producing states in the U.S. [151]. This has caused an estimated yield loss of 1.9 Metric tons annually in South Dakota (https://www.sdsoybean.org.).

1.6.2. Life Cycle of SCN

SCN completes its life cycle mainly in three main stages, the egg, juvenile, and the adult upon parasitizing the soybean roots [148]. The embryogenesis and molting undergo in the egg stage resulting J1, the first juvenile stage in the egg [152]. The J1 stage undergoes molting, and results in the second-stage juveniles (J2) hatched from the eggs under optimal soil conditions near the roots of the soybean plants. Several factors are responsible for egg hatching: soil temperature of approximately 25°C, suitable host

plant, and soil conditions [152]. Soil moisture, soil fertility, and soil type play an important role in the life cycle of the nematode. The light sandier soil allows nematode to move more freely in the soil than in the compact soil that restricts the movement of the nematode. Nonetheless, SCN are reported in all kinds of soil [148, 152]. Other factors can be host root exudates, pH and sometimes age dependent egg hatching [152, 153, 154]. The host plant also plays a major role in the process of egg hatching and releasing J2 in the soil. The organic molecules such as eclepsins, and glycinoeclepin A produced by the host plant assist in egg hatching, which is known as root diffusate based egg hatching [148, 155, 156]. Some other compounds such as solanoeclepin A in tomato and potato, and chemicals such as picloronic acid, sodium thiocyanate, alpha-solanine, and alpha-chaconine help in the hatching process [157, 158]. *Gro-nep-1* has been recently identified as the first gene to be upregulated in eggs treated with host root exudate in *Globodora rostochiensis* [159]. The exudates help J2 find the host plant's root using a form of chemotaxis and infect the root cells using hollow mouth spear called stylet [143, 160]. Inability or failure of J2 in finding the host plant leads to starvation and death [161]. The secretion of the digestive enzymes such as cellulase helps advancing through the epidermal and cortical cells toward the vascular cylinder [143, 162]. The penetration site of the J2 depends on the water status of the soybean plant [163]. Because of their sedentary nature, J2 selects a single cell that undergoes morphological changes and forms the permanent feeding site called syncytium [143, 162]. Syncytium remains intact throughout the remaining time of the life cycle of nematode and draws essential nutrient from the host plant until reaching reproductive maturity [143]. The juveniles molt into a third juvenile state (J3) and undergo sexual differentiation [164]. The ratio of the female

and male remains 1:1 but this is sometimes affected by the milieu and resistance of the host plant [165]. The feeding site swells longitudinally throughout the root dissolving and incorporating numerous cells with dense cytoplasm, hypertrophied nuclei, increased organelle content [143]. During this J3 male metamorphoses to vermiform shape, leaves the root in search of females using sex pheromone, and dies after mating [166, 167]. Concurrently, J3 juvenile molts into J4 stage forming adult females, and changes into lemon-shaped cyst extruding the root surface. Each female in a cyst produces 40-600 eggs with an average of approximately 200 eggs and sometimes eggs are outside in the gelatinous secretions [168, 169]. The cyst produces compounds such as chitinase and polyphenol oxidase in order to save eggs from desiccation and microbial infection [141]. Thus eggs can remain viable up to nine years [141]. Naturally, a SCN completes its life cycle in 3 to 4 weeks, and highly depends on the soil temperature of approximately 25°C, suitable host plant as well as soil conditions [152]. However, the SCN can complete its life cycle in 21 days under controlled conditions with a temperature of 25°C [160]. Depending upon the maturity group of the soybean planted, the SCN completes up to four life cycles during a single soybean growing season [150].

1.6.3. SCN Effectors

Nematode effector molecules are produced in a nematode's esophageal gland before being released into the stylet [170]. The effectors evade and suppress the host plant's defense and reprogram the host cell nucleus, as well as a various cellular process for their suitability [171, 172]. These effector molecules reach into the host cell after dissolving the cell wall through various enzymes and proteins that bind to the components of the cell wall such as cellulose binding proteins and expansins [171]. The

successful parasitism of the nematodes to the plants involves direct or indirect interaction with the host plant targets or immune regulators, cell wall modifications, mimicry of plant peptide hormones, or manipulating hormone transport [reviewed in [173]]. Various nematode effector molecules, including Gr-SPRYSEC (-4,-5, -8, -15, -18, -19), Gp-RBP-1, Gr-VAP1, Hg30C02, Hs10A06, Hs4F01, and Mi-CRT, have been already characterized in different nematodes and hosts [174, 175, 176, 177, 178, 179, 180, 181]. These effectors affect the host immune system by enhancing susceptibility or resistance.

The characteristic cyst nematode effectors, including those found in SCN, are presented in Table 1.2. Gao et al. (2003) [182] identified 51 effector molecules from the esophageal gland of the *H. glycines*. Most of the effector molecules belonged to cellulose genes, pectate lyases, an enzyme in the shikimate pathway, and ubiquitin proteins. The ortholog of *H. glycines* cellulose binding protein (HgCBP) in *H. schachtii* (HsCBP) interacts with pectin methyltransferase protein (PME3) of *Arabidopsis* during the early feeding stage, and exhibits enhanced susceptibility [183]. The function of an ortholog of 25A01-like effector family was studied in *H. schachtii* (Hs25A01) in *Arabidopsis* system [184]. Hs25A01 interacts with *Arabidopsis* F-box-containing protein, chalcone synthase and the translation initiation factor eIF-2 b subunit to increase both root length and susceptibility to *H. schachtii*. Further, 18 more effector molecules showed high similarity to N-acetyltransferases, β-fructofuranosidases, serine proteases, cysteine proteases, an effector for protein degradation in the syncytium, cellulose binding protein, chorismate mutase, and glycosyl hydrolase [173]. Among them, HgGLAND18 secreted in the dorsal gland cell, suppresses basal and hypersensitive cell death innate immune responses in *Nicotiana benthamiana* [185]. The similarity of the N-terminal domain of HgGLAND18

to the domain of effector of *Plasmodium* spp. strongly suggests the role of convergent

evolution of the effector molecules in diverse parasites [185]. Another effector, biotin

synthase (HgBioB) and protein containing protein SNARE domain (HgSLP-1) effector

molecules were reported recently employing allelic imbalance analysis to associate SCN

SNPs [186]. HgSLP-1 interacts with Rhg1 α-SNAP evading the host defense [186].

However, the host defense is evaded on the absence of the HgSLP-1 because of its

avirulence nature like map-1 protein and Mj-Cg-1 effectors [186, 187, 188]. Until

effectors were searched through *de novo* transcriptome assembly of the second stage

juvenile *H. glycines* [57], only 71 effector molecules were reported that were secreted

only from the esophageal glands. Upon use of the joint pipeline that utilizes presence or

absence of signal peptides, altogether 944 effector candidates were predicted, many of

which were homologs to glutathione synthetase, C-type lectins, plants RING/U-box

superfamily, arabinosidase, fructosidase, glycoside hydrolase, expansin and SPRYSEC

family [57].

Table 1.2. Characterized cyst nematode effectors in different plant systems with their targets and susceptibility/resistance effects.

CN effectors	Cyst Nematode	Targets	Host	Susceptibility/Resistance	References
HsCBP	*H. schachtii*	pectin methyltransferase protein (PME3)	*Arabidopsis thaliana*	susceptibility	[183]
Gp-Rbp-1	*G. pallida*	Gpa2	*Nicotiana benthamiana*	hypersensitive response (HR)	[180]
Gr- SPRYSEC (4,5,8,15,18,19)	*G. rostochiensis*	NB-LRR proteins	*Nicotiana benthamiana*	Suppress host defense	[181]
Hs19C07	*H. schachtii*	auxin influx transporter LAX3	*Arabidopsis thaliana*	susceptibility	[189]
Gr-VAP1	*G. rostochiensis*	apoplastic cysteine protease Rcr3pim	*Solanum lycopersicum*	programmed cell death	[179]
Hg30C02	*H. schachtii*	β-1,3-endoglucanase	*Arabidopsis thaliana*	susceptibility	[175]
Hs4D09	*H. schachtii*	14-3-3ε	*Arabidopsis thaliana*	resistance	[190]
Hs10A07	*H. schachtii*	interacting plant kinase (IPK) and IAA16 transcription factor	*Arabidopsis thaliana*	hypersusceptible	[191]
Hs25A01	*H. schachtii*	F-box-containing protein, a chalcone synthase and the translation initiation factor eIF-2 b subunit (eIF-2bs)	*Arabidopsis thaliana*	susceptibility	[184]
Hs30D08	*H. schachtii*	SMU2 (homolog of suppressor of mec-8 and unc-52 2)	*Nicotiana benthamiana*	susceptibility	[192]
Hs10A06	*H. schachtii*	Spermidine Synthase2 (SPDS2)	*Arabidopsis thaliana*	susceptibility	[176]
HgGLAND18	*H. glycines*	---	*Nicotiana benthamiana*	suppresses both canonical basal and HR immune responses	[185]
HgSLP-1	*H. glycines*	*Rhg1* α-SNAP	*Glycine max*	avirulence protein	[186]

1.6.4. *Rhg1* and *Rhg4* as Major QTLs for SCN Resistance

SCN can enter into the roots of susceptible and resistant soybean cultivars equally [193]. Resistant cultivars prevent SCN infection by disrupting syncytium formation interfering its life cycle. Histological experiments have unraveled that syncytia forming in resistant plants undergo a hypersensitive-like response [194]. The sources for the SCN resistance in the commercial soybean cultivars are predominantly Peking (PI548402), PI88788, and PI437654 that carry resistance loci effective against various races of SCN [195, 196]. Up to now, 40 QTLs have been reported in a diverse group of resistant cultivars, which are mapped in 17 of 20 chromosomes [196]. Three recessive resistance *rhg1-rhg3* were initially assigned in the Peking plant introduction [197]. The *rhg1* gene confers resistance to SCN in all germplasms with resistance to SCN and is regarded to be

a significant resistance gene to SCN in soybean cultivars [196]. Moreover, PI437654 and PI88788 each have a different functional SCN resistance allele at or close to *rhg1* [195]. The *rhg1* gene was initially reported as the recessive locus, however, recent studies have reported the occurrence of incomplete dominance [198]. The *rhg1* locus has been present in various resistance plant introductions PI209332, PI437654, PI90763, PI209332, PI89772, PI90763, including Peking (PI548402), PI88788, and PI437654 [196]. *Rhg1* locus has been mapped to chromosome 18's subtelomeric region [199, 200, 201, 202]. *Rhg4,* a dominant locus, is present in PI54840 (Peking) and PI437654 but not in PI88788 or PI209332 [195, 196, 203]. *Rhg4* locus has been mapped to on chromosome 8 (linkage group A2) for SCN resistance [196, 204].

1.6.5. *LRR-RLK* Genes were Considered as the Resistance Genes against *H. glycines*

Rhg1 and *Rhg4* genomic regions in the soybean, and two leucine-rich repeat transmembrane receptor-like kinase (LRR-RLK) genes were patented by two groups [199, 200, 205, 206]. Such claims were based on the similarity of the genes with rice bacterial blight resistance gene *Xa21* [207]. Their claims were accepted by the soybean communities, but their functional assessment was not conducted until 2010. Melito et al. (2010) [198] used artificial microRNA (amiRNA) to study the function of *Glyma18g02680.1* gene (LRR-RLK) at the *Rhg1* locus. The amiRNA used for the reduction of expression of *Glyma18g02680.1* gene from the *Rhg1* locus of Fayette (PI88788 source of *Rhg1*) did not alter the resistance to SCN but instead affected the root development. Later Liu et al. (2011) [208], used the Targeting Induced Local Lesions In Genomes (TILLING) approach to study the role LRR-RLK at the *Rhg4* locus developing

EMS-mutants from the SCN-resistant soybean cultivars Forrest and Essex. The TILLING

tool is the reverse genetic tool to the function of the gene [209]. They concluded that the

Rhg4 LRR-RLK gene is not a gene for SCN resistance. After the availability of the

complete genome sequence of soybean, it has been easier to narrow down these genomic

regions and characterize specific candidate genes that can potentially be involved in the

SCN resistance [210].

1.6.7. Role of *GmSNAP18* (*Rhg1*) and *GmSHMT08* (*Rhg4*) Uncovered for SCN
Resistance

The study by Kim et al. (2010) [211] showed *rhg1-b* within a 67-kb region in

PI88788 genotype. Because of the existence of allelic variants of *rhg1* in the different

soybean genotypes, the *rhg1* in PI88788 was named as *rhg1-b* [195, 211]. This 67-kb

interval from PI88788 does not include the LRR-RLK gene candidate for *rhg1* from

Peking cultivar that was previously patented. Matsye et al. (2011) [212] studied the

expression of the genes within the 67 kb interval of the *rhg1-b* locus. Amino acid

transporter (*Glyma18g02580*) and a soluble NSF attachment protein (α-SNAP;

Glyma18g02590) genes were specifically expressed in syncytia during the SCN defense

in both Peking (PI548402) and PI88788 genotypes. The α-SNAP coding regions are

identical in resistant genotypes Peking (PI548402) and PI437654, but they differ by

numbers in single nucleotide polymorphisms (SNPs) in Williams 82 (PI518671)

genotype [213]. Later, in a 31-kilobase (kb) segment at *rhg1-b* loci, genes

Glyma.18G022400 formerly *Glyma18g02580*, *Glyma.18G022500* formerly

Glyma18g02590, *Glyma.18G022700* formerly *Glyma18g02610* that encodes an amino

acid transporter, an α-SNAP (soluble N-ethylmaleimide–sensitive factor attachment

protein) protein, and a WI12 (wound-inducible domain) protein, respectively were

identified that play a significant role in SCN resistance [214, 215]. The WI12 protein may

involve in producing phenazine like compounds that can be toxic to the nematodes [214,

216]. The α -SNAP protein involves in vesicle trafficking that affects the exocytosis of

food in the syncytium, which in turn affects the nematode physiology [214]. The plant

transporter protein, *Glyma18g02580* consists of a tryptophan/tyrosine permease family

domain [214]. Tryptophan upon catalysis by Trp aminotransferases such as AtTAA1 and

PsTAR1 and subsequent flavin mono-oxygenase such as YUC forms indole-3-acetic

acid, which is a precursor of the hormone auxin [217]. This suggests that

Glyma18g02580 may affect the auxin distribution in the soybean plants [214]. Based on

Glyma18g02590 (*GmSNAP18*) gene, the cultivars Peking-type and PI88788 type can be

differentiated upon selecting the *rhg1* resistance alleles using two specific KASP

(kompetitive allele-specific PCR) SNP markers. [218]. The 31 kb segment is present as a

single copy in the susceptible cultivar, whereas, the resistant variety, PI88788, and

Peking (PI548402) possess 10 and three tandem copies, respectively [214]. Additionally,

Cook et al. (2014) [219] tested *Rhg1* across 41 diverse soybean cultivars using whole-

genome sequencing technique called fiber-FISH (fluorescence in situ hybridization). The

study showed seven *Rhg1* copies in PI548316, nine copies in PI88788, and 10 copies in

PI209332 whereas, both PI437654 and PI548402 (Peking), which show a high level of

SCN resistance, contain three copies of the *Rhg1* with α-SNAP allele [219]. Lee et al.

(2015) [220] genotyped the *Rhg1* locus in 106 SCN-resistant *G. max* and *G. soja*

genotypes developing genomic qPCR assay for the identification of copy number of *Rhg1*

locus and found 2–4, 6, 7, 9 and 10 copies in *G. max* and one three-copy variant in a *G.*

soja genotype. Fayette, derived from PI88788, has ten copies of the repeat that suggested an increased copy number by a single unit during the process of selection.

The use of forward genetics and functional genomics approaches showed the Peking-type *rhg1* resistance in Forrest cultivar depends on the *Rhg4* (*GmSHMT08*) gene, SCN-resistant allele [221]. Such resistance in Forrest cultivar (resistance to SCN requires both *rhg1* and *Rhg4*) differs from the PI88788-type of resistance that only requires *rhg1* [204, 221]. *GmSHMT08* gene was emerged because of the artificial selection during the soybean domestication process accumulating a higher number of non-synonymous mutations [222]. A recent study by Liu et al. (2017) [223] narrowed down the interval to ~14.3 kb in the recombinant lines of Forrest cultivar that contained three genes in three tandem repeats with in *rhg1-a* locus. These genes encode armadillo/β-catenin-like repeat, amino acid transporter (AAT), and soluble N-ethylmelaimide sensitive factor (NSF) attachment protein (*GmSNAP18*). The mapping results and based on SNPs and InDels in Forrest, Peking, and PI88788 cultivars, *GmSNAP18* was identified as an *rhg1* candidate gene for SCN resistance. Additionally, genetic complementation analyses of *GmSNAP18* revealed its different role in PI88788-type *GmSNAP18* and Peking type *GmSNAP18*. Thus both Peking type *GmSHMT08* (*Rhg4*) and *GmSNAP18* (*Rhg1*) play a different role from PI88788-type *GmSHMT08* and *GmSNAP18*. Bayless et al. (2016) [215] confirmed the presence of a dysfunctional variant of resistance-type α-SNAP in the resistant cultivars that impairs the NSF function reducing its interaction during 20S complex formation. This leads to disruption in vesicle trafficking causing an abundance of NSF protein in the syncytium, which is cytotoxic. However, because of the two duplication events that occurred 13 and 59 million years ago (mya) in William 82 soybean genome

[210], soybean encodes other four α-SNAPs *GmSNAP02*, *GmSNAP09*, *GmSNAP11*, and *GmSNAP14*, known as wild-type α-SNAPs [215, 224]. Among them, *GmSNAP11* is a minor contributor to SCN resistance but not *GmSNAP14* and *GmSNAP02* [224]. These wild-type α-SNAPs counteract cytoxicity for the viability of soybeans that carry haplotypes of *Rhg1* for the SCN resistance [215]. In the presence of SCN, the ratio of resistance-type to wild-type α-SNAP increases leading to the hyperaccumulation of resistance-type α-SNAP that reduces the viability of the syncytium [215] (Figure 1.3). Also, some other genes such as ascorbate peroxidase 2, β-1,4-endoglucanase, soybean momilactone A synthase-like, cytochrome b5, DREPP membrane protein-family, plastocyanin –like including serine hydroxymethyltransferase decreased female index of SCN by 50 % or more in SCN susceptible cultivar William 82 upon overexpression [225].
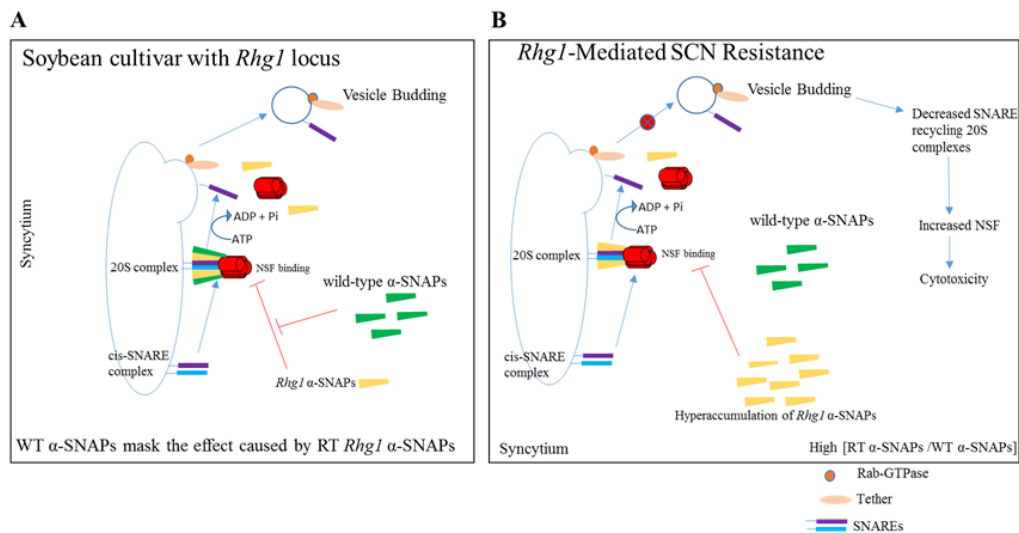
Figure 1.3. Role of α-SNAP in vesicular trafficking. A) Wild-type α-SNAPs counteract the cytoxicity for the viability of soybeans that carry haplotypes of *Rhg1* for SCN resistance. B) In the presence of SCN, the ratio of resistance-type to wild-type α-SNAP increases leading to the hyperaccumulation of resistance-type α-SNAP. The presence of high RT α-SNAPs dysfunctional variants in the resistant cultivars impair the NSF function reducing its interaction during 20S complex formation. This leads to a disruption in vesicle trafficking causing an abundance of NSF protein in the syncytium, which is cytotoxic (Concept adapted from [215] and [226]).

Liu et al. (2012) [227] used two recombinants that carry resistance allele at the *rhg1* and *Rhg4* loci, to study a gene at the *Rhg4* loci. The cultivars used in the study were double recombinants for an 8-kilobase (kb) interval carrying the *Rhg4* resistance allele that carries two important genes serine hydroxymethyltransferase (SHMT) and the other a subtilisin-like protease (SUB1). *SHMT* (*GmSHMT08*) gene was confirmed as the resistance gene at the *Rhg4* locus that catalyzes methylene carbon of glycine to tetrahydrofolate (THF) to form methyleneTHF, that reacts the second glycine to form L-Ser in the glycolate pathway [228]. This reaction produces S-adenosyl-Met (SAM), which is the precursor for the polyamines and plant hormone ethylene [221]. *GmSHMT08*

changes its enzymatic properties because of the changes in two amino acids (P130R and N385Y) in the resistant allele that negatively affects the folate homeostasis in the syncytium resulting hypersensitive responses (HR) leading to programmed cell death (PCD) [222, 227] (Figure 1.4). The alleles of *GmSHMT08* are different between resistant and susceptible plants [227].
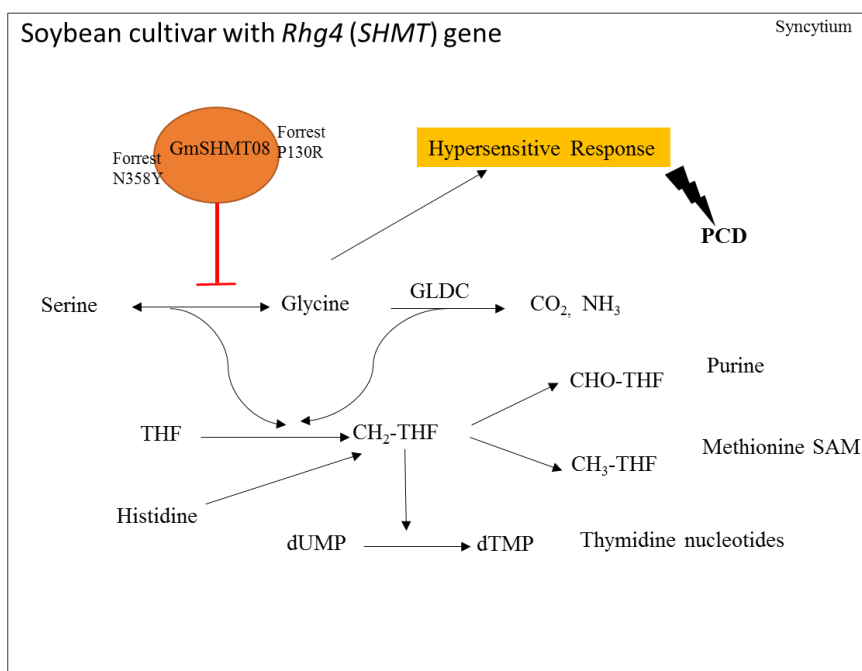
Figure 1.4. Schematic overview of *GmSHMT08* function and C1 metabolism. SHMT, GLDC, and degradation of histidine feed into the pool of C1 units bound by THF. GmSHMT08 with changes in two amino acids (P130R and N385Y) in Forrest cultivar negatively affects the folate homeostasis in the syncytium resulting in hypersensitive responses (HR) leading to programmed cell death (PCD). dTMP, deoxythymidine monophosphate; dUMP, deoxyuridine monophosphate; GLDC, glycine decarboxylase; SAM, S-adenosyl methionine; SHMT, serine hydroxymethyltransferase; THF, tetrahydrofolate (Concept adapted from [222, 227] and pathway modified from [229]).

1.6.8. Minor QTLs/Genes for SCN Resistance

Apart from the major QTLs identified in *Rhg1* and *Rhg4* loci, there are some minor genes or QTLs identified such as *qSCN10* on chromosome 10 in PI567516C cultivar [230]. PI567516C cultivar lacks two major loci *Rhg1* and *Rhg4* and is SCN resistant that implies the importance of other minor genes for SCN resistance [231]. The resistance acquired by the major genes is sometimes not durable and necessitates the use of horizontal or quantitative resistance acquired from the minor genes [232]. Other minor QTLs are *qSCN-003* in PI88788 [233], *qSCN-005* in Hartwig, which has SCN resistance from PI437654 and Peking [234], and *qSCN-11* in PI437654 and PI90763 [235, 236]. The most recent QTLs reported are cqSCN-006 and cqSCN-007 in *G. soja* PI468916 [237], which was mapped finely by Yu and Diers 2017 [238] where *cqSCN-006* was mapped into a 212.1 kb interval and *cqSCN-007* to a 103.2 kb interval on the Williams 82 reference genome in chromosome 15 and 18, respectively. The *cqSCN-006* QTL consists of three major potential candidate genes: *Glyma.15g191200* (Soluble NSF attachment protein), *Glyma.15g191300* (BED-zinc finger related), *Glyma.15g191400* (BED-zinc finger related). *Glyma.15g191200* is predicted to encode soluble N-ethylmaleimide–sensitive factor attachment protein (γ-SNAP) that involves in the same function as α-SNAP, which is one of the important genes in *Rhg1* mediated SCN resistance. Likewise, the potential genes identified in region of *cqSCN-007* are: *Glyma.18g244500* (Lecithin-cholesterol acyltransferase), *Glyma.18g244600* (Apetala 2 transcription factor), *Glyma.18g244700* (Calcineurin-like phosphoesterase), *Glyma.18g244800* (Chromatin assembly factor 1 subunit A), *Glyma.18g244900* (p-Nitrophenyl phosphatase), *Glyma.18g245000* (Rad21/Rec8-like protein), *Glyma.18g245200* (LETM1-like protein),

which are mainly involved in signaling pathways, such as transcription, euchromatin

expression, and membrane receptor detection. These identified potential candidate genes

might be novel SCN resistance genes that should be functionally characterized in the

coming future [238].

Table 1.3. SCN resistance QTLs in soybean cultivars with information on chromosome location, markers associated against SCN HG types or races and refrences.

QTLs	Chromosome and markers associated	SCN HG type or races	Soybean Plant Introductions	References
cqSCN-001 (*Rhg1*)	18	Race 3	PI437654	[239]
		Race 1a, 3a, 3b, 1b, 6	PI209332	[240]
		Races 2, 3 and 5	PI90763	[236, 241]
		Races 1, 3, and 6	PI88788	[241]
		Races 1, 2, were verified in Peking conditioning resistance to SCN 3, 5	PI89772	[242]
		Races 2, 3 and 5	PI404198A	[243]
rhg1-b	18	PA3 (HG type 7) and TN14 (HG type 1.2.5.7)	PI88788	[195]
	18; 67-kb region of the 'Williams 82' genome between BARCSOYSSR_18_0090 and BARCSOYSSR_18_0094	PA3, which originally had an HG type 0 phenotype	PI88788	[211]
cqSCN-002 (*Rhg4*)	8	Race 3	Peking	[193, 204]
		Race 3	PI437654	[239]
cqSCN-003	16	PA3 (HG type 7, race 3) and PA14 (HG type 1.3.5.6.7, race 14)	PI88788	[233]
cqSCN-005	17	HG Type 1.3 (race 14) and HG Type 1.2.5 (race 2)	Hartwig (PI437654 and Peking)	[234]
cqSCN-006	15; (803.4 kb region between SSR markers BARCSOYSSR_15_0886 And BARCSOYSSR15_0903)	HG type 2.5.7 (SCN isolate PA5)	*G. soja* PI468916	[237]
	15; 212.1 kb interval between ss715621232 and ss715621239.	HG type 2.5.7 (SCN isolate PA5)	*G. soja* PI468916	[238]
cqSCN-007	18; (146.5 kb region between the SSR markers BARCSOYSSR_18_1669 and BARCSOYSSR_18_1675)	HG type 2.5.7 (SCN isolate PA5)	*G. soja* PI468916	[237]
	18; 103.2 kb interval between BARCSOYSSR_18_1669 and ss715631888.	HG type 2.5.7 (SCN isolate PA5)	*G. soja* PI468916	[238]
cqSCN 10	10 (Satt592, Satt331, and Sat_274)	LY1 nematode from a mass mating of SCN Race 2 (HG Type 1.2.5) females with Race 5 (HG Type 2.5)	PI567516C	[230]
cqSCN11	11	HG types 0, 2.7, and 1.3.5.6.7 (race 3, 5, and 14)	PI437654	[235]
		Races 2 (HG type 1.2.5.7), 3 (HG type 0) and 5 (HG type 2.5.7)	PI90763	[236]

1.6.9. GWAS Study in SCN Resistance Expands other QTLs on SCN

The GWAS technique has been also used in revealing candidate genes for SCN

resistance relatively in less time and simultaneously verifying QTLs identified by

classical bi-parental mating [119, 120, 121, 244, 245, 246, 247]. Wen et al. (2014) [245]

reported 13 GWAS QTLs for SCN resistance associated with the sudden death syndrome

(SDS) QTLs spanning a physical region of 1.2 Mb (1.2-2.4 Mb) around three *Rhg1*

genes. This might be because of the close linkage of *Rfs2* and *Rhg1* genes that provide

resistance to SDS and SCN resistance, respectively [248]. Han et al. (2015) [247]

reported 19 significant QTLs related to resistance to both SCN HG Type 0 (race 3) and

HG Type 1.2.3.5.7 (race 4) using 440 soybean cultivars. Of the reported SNPs, eight

overlapped to QTLs with *Rhg1* and *Rhg4* genes, eight to other known QTLs and three

were the novel QTLs (on chromosome 2 and 20). The gene, *Glyma.02g161600*, which

encodes the RING-H2 finger domain nearest to the novel loci could be the new source of

SCN resistance. Vuong et al. (2015) [120] utilized 553 soybean PIs and SoySNP50K

iSelect BeadChip (with 45,000 SNP markers) to detect the QTL or genes for HG Type 0

SCN resistance using GWAS study. Fourteen loci with 60 SNPs were significantly

associated with the SCN resistance. Of the 14 detected loci, six QTL that was identified

using bi-parental mapping including *Rhg1* and *Rhg4* were also verified. These GWAS

QTLs contained 161 candidate genes located at significant GWAS loci for SCN

resistance in soybean. Among them, 26 genes were NBS encoding genes. Chang et al.

(2016) [121] reported significant loci to multiple races of SCN using GWAS, of which

one SNP was within *Rhg1* locus for SCN races 1, 3 and 5. Among the five LRR-RLK

genes, *Glyma18g02681* and *Glyma20g33531* were nearest to two significant SNPs

s715629308 and ss715638409, respectively. Additionally, they reported significant SNPs on chromosomes 4, 7, 10, 15, 18, and 19 for SCN races 1 and 5 (HG type 2). However, Li et al. (2016) [244] employed joint linkage mapping and association mapping using 585 informative SNPs across recombinant inbred lines (RILs) bred from the cross Zhongpin03-5373 (ZP; resistant to SCN) × Zhonghuang13 (ZH; susceptible to SCN) to detect alleles associated with SCN race 3. Association mapping revealed three quantitative trait nucleotides (QTNs): *Glyma18g02590* (belonged to locus *rhg1-b*), *Glyma11g35820* and *Glyma11g35810* (an *rhg1-b* paralog); whereas, linkage mapping revealed two QTLs (one mapping to *rhg1-b* and another to *rhg1-b* paralog). Upon combining both linkage and association mapping, six significant markers were detected. Among them, four (Map-5118, Map-5255, Map-5431, and Map-5432) of the significant markers were not identified by an independent study. Map-5431 lies between *rhg1-a* and *rhg1-b* (*Glyma18g02650*), and Map-5432 lies adjacent to *rhg1-a* (*Glyma1802690*) [248].

Zhang et al. (2016) [246] utilized 235 wild soybean (*G. soja* Sieb. & Zucc.) accessions to unravel the genetic basis for HG Type 2.5.7 (race 5). GWAS revealed 10 significant SNPs associated with SCN resistance, among which four SNPs were linked to known QTL, *rhg1* on chromosome 18. Four others were linked to race 5 resistance QTL [249] and remaining two to the 35.5 to 37.8Mb region that overlaps some region identified by Vuong et al. (2015) [120]. Additionally, 58 potential gene candidates were suggested, which included genes encoding NBS-LRR proteins (*Glyma.18G078000*, *Glyma.18G077900*), MAPK proteins (*Glyma.18G106800*), RLPs (*Glyma.18G193800*), a RING/U-box protein (*Glyma.18G063500*), and MYB family transcription factors (*Glyma.19G119300*). Recently, Zhang et al. (2017) [119], performed GWAS in 1032 on

G. soja with 42,000 SNPs to dissect the genetic basis for resistance to race 1. Ten significant SNPs were identified on chromosomes 2, 4, 9, 16 and 18, among which two were within the previously identified QTLs (SCN 18-5 and SCN 19-4; [249] on chromosome 4, and one within QTL SCN 37-2 [231]). This study strongly suggests *R* gene, *Glyma.18G102600*, to be the promising candidate gene for the SCN resistance because of its location in a strong linkage disequilibrium block.

The non-redundant 249 genes assessed from the GWAS SCN QTLs [119, 120, 121, 244, 245, 246, 247] showed most of the genes enriched to binding (GO: 0005488), and catalytic activity (GO: 0003824). The binding category includes binding to nucleoside (GO: 0001882), nucleotide (GO: 0000166), purine ribonucleotide (GO: 0017076), purine nucleoside (GO: 0001883), ribonucleotide (GO: 0032553), adenyl nucleotide (GO: 0030554), adenyl ribonucleotide (GO: 0032559), ATP (GO: 0005524), and ADP (GO: 0043531). Similarly, the catalytic category includes transferase activity (GO: 0016740), transferase activity-transferring phosphorus-containing groups (GO: 0016772), phosphotransferase activity- alcohol group as acceptor (GO: 0016773), kinase (GO: 0016301), protein kinase (GO: 0004672), exopeptidase (GO: 0008238), and serine-type exopeptidase (GO: 0070008) activities (Figure 1.5).
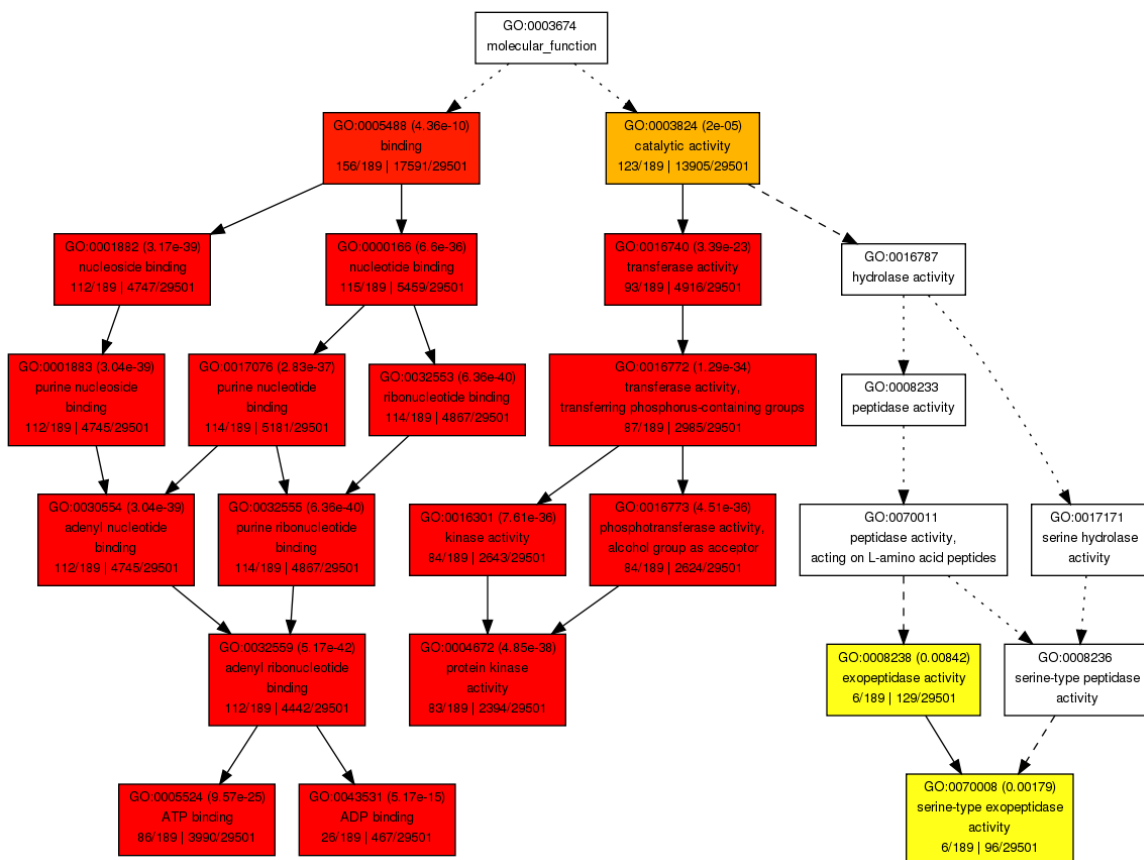
Figure 1.5. Significantly enriched GO molecular function terms of non-redundant 249 genes in the GWAS SCN QTLs [119, 120, 121, 244, 245, 246, 247] as determined by a hypergeometric test using AgriGO [117]. The same gene can be associated with multiple GO annotations. Only significantly ($P < 0.01$) over-represented and Bonferroni adjusted GO categories are shown. The stronger color represent the lower P value. Information in the box includes GO term, adjusted P value in parentheses, GO description, a number of query list/background mapping GO, and a total number of query list/background.

1.7. Plant-aphid Interactions

A series of cell signaling events such as plasma membrane potential variation, calcium signaling, and generation of reactive oxygen species leads to the production of hormones and metabolites during plant-aphid interactions [250]. In most cases, a hormone release is specific to a stimulus. For example, jasmonic acid (JA) is released in response to chewing herbivores, cell content feeders and necrotrophic pathogens whereas,

salicylic acid (SA) is released in response to piercing-sucking herbivores [251]. However, ethylene (ET) is produced synergistically with JA, and modulate JA and SA signaling pathways [252]. The change in metabolite products during the herbivore feeding occurs both in local and systemic tissues [253].

Approximately, half of one million known insect species along with aphids get their nutrition from plants [254]. These insects are grouped into the family Aphididae of the order Hemiptera. Over 4,000 aphid species are identified as harmful to plants [255, 256]. Many aphids are specific to their host, and attack plants of a single family, for example, *Acyrthosiphon pisum* attacks hosts belonging to *Fabaceae* family, however, there are species such as *Myzus persicae*, which can infest dicot plants of more than 40 families [257]. Asexual life cycle in aphids allows prompt population growth and infestation in a expedite manner [257]. Up to this time, many research studies have been done in understanding plant-aphid interactions. This has led to concrete findings on both plant and aphid side, increasing insights into plant defense mechanisms against aphids. The plant can sense aphid effector molecules, which are mostly expressed in salivary glands, secreted into saliva and eventually released inside the host at the time of feeding and probing [258]. There are also chewing insects such as beetles and lepidopteran larvae that can cause damage in various plant tissues and feed through stylet penetration consuming phloem sap [259]. The effector molecules can be either cell wall degrading enzymes such as pectinases, glucanases, amylases or detoxifying enzymes such as oxidoreductases, phenol oxidases, peroxidases (reviewed in [260]). The manipulation of host responses by aphids depends on the capacity to alter host morphology, affect the nutrient distribution and destroying host defense responses (reviewed in [260]). To avoid

such attacks from aphids, plants have developed their own defense strategies such as the presence of preformed barriers, chemical defenses that are constitutive in nature, and employing direct and indirect inducible defenses [254]. Like in any coevolutionary interaction, in plant-parasite interactions evolutionary arms race takes place [260]. There are various models that describe plant-pathogen interactions such as the gene for gene model, guard model, decoy model, bait and switch model and zig-zag model [16, 261, 262, 263]. The zig-zag model depicts the interaction between the plant and parasites [16]. It is still unknown if aphid and other insects interaction follow this particular model [260]. According to the zig-zag model, aphids possess conserved molecular pattern called Pathogen Associated Molecular Patterns (PAMPs) and are recognized by various receptors present in the external face called pattern recognition receptors (PRRS) [264]. These PRRS on plasma membrane recognize PAMPs when challenged by the pathogens, and plant basal immune response called PAMP-triggered immunity (PTI) is triggered. Effector-triggered susceptibility (ETS) is triggered by the successful aphids that deliver effectors capable for pathogen virulence; ETI (Effector-triggered Immunity) results from the successful recognition of effectors by the NBS-LRR proteins; Natural selection helps pathogen to dodge ETI by shedding or varying the effector gene or adding some effectors that suppress ETI, which eventually results in new R specificities so that ETI can be triggered again. As reviewed by Wu and Baldwin, 2010 [254] early defense signaling events take place in a cell of insect attacked leaf. Major events are described as: elicitors that are perceived by the receptors on plasma membrane trigger Ca^{2+} channels and produces Ca^{2+}. Ca^{2+} binds with NADPH oxidase, which gets enhanced through phosphorylation by CDPKs eventually producing reactive oxygen species (ROS).

MAPKs are also activated quickly among which SIPK and WIPK trigger the synthesis of

Jasmonic acid (JA) and JA-Ile (JA-isoleucine). JA-Ile binds to COI1 receptor causing

degradation of JASMONATE ZIM-DOMAIN (JAZ) proteins and releases MYC2 and

MYC2-like transcription factors. SIPK phosphorylates ACS proteins and increases

ethylene production, which leads to increased activity of ERF (Ethylene Responsive

Elements). These series of signaling get translated into metabolites that are responsible

for plant defense.

1.8. Plant-nematode Interactions

Plant parasitic nematodes are obligatory parasites and are sedentary endoparasites

[*Heterodera* and *Globodera* (cyst nematode) and *Meloidogyne* (root-knot nematodes)]

[139]. Cyst nematodes get their way to vascular cylinders by use of stylets through the

root and form the feeding site coupled with multinucleate syncytium formation [265].

These cyst nematodes go through three molt stages and eventually become adults. The

infected cells around the feeding site of nematodes divide and swell forming root knots

[139], and after the infection, endoglucanase and polygalacturonase genes in the host are

upregulated [266]. In *Arabidopsis*, a homolog of pectin acetyltransferase gene is

upregulated in both syncytia and pre-giant cells [267]. Various experiments have reported

upregulation of auxin-response genes and an increase in ethylene (reviewed in [139]).

Nematodes are also involved in upregulating genes such as *ENOD40*, involved in

nodulation and *CCS52a*, involved in cell-cycle [268].

Numerous plant resistance genes involved in defense mechanism against aphids

and nematodes encode proteins containing a nucleotide-binding site (NBS) and a leucine-

rich repeat (LRR) motifs [269]. Root knot nematode (*Meloidogyne incognita*) resistance gene in tomato, *Mi*, is an NBS-LRR gene [270] and is involved in resistance to potato aphid (*Macrosiphum euphorbiae* Thomas) [271]. *Vat* gene, which confers resistance to *Aphis gossypii* in melon (*Cucumis melo*) is also an NBS-LRR gene [272]. Besides these, an aphid resistance gene, *AKR* (*Acyrthosiphon kondoi* resistance) was mapped on to a CNL cluster in *Medicago truncatula* [273]. Other nematode resistance genes, *Gpa2,* and *Hero* belong to the NBS-LRR family and have been cloned (reviewed in [139]).

1.9. Plant-aphid-nematode Interactions

Both above- and belowground herbivores, although spatially segregated, share the host plant through the systemic tissues and can influence each other [274]. Such herbivory has increased diversification across the insects [275]. Numerous belowground organisms such as nematodes, microbes, fungi, and insects that feed on plant roots can fluctuate the concentration of defense compounds such as phenolics, terpenoids or glucosinolates, both in belowground and aboveground plant tissues [276]. The impact of root-feeders on shoot defense and effects of aboveground herbivory on root defense was remained unnoticed for a long time [277].

There have been several previous studies toward understanding plant-aphid-nematode interactions [172, 276, 278, 279, 280, 281, 282, 283, 284, 285, 286, 287, 288, 289, 290] (Table 1.4). The nematode, *Pratylenchus penetrans* infection on *Brassica nigra* caused a decreased infestation of shoot herbivore, *Pieris rapae* [276]. Bezemer et al. (2005) [286] reported decreased fertility of aphids *Rhopalosiphum padi* infesting *Agrostis capillaris* and *Anthoxanthum odoratum* because of decreased amino acid in the phloem

sap of nematode infected plants. A similar type of effect was seen in the offspring of aphid *Myzus persicae* on *Plantago lanceolata* infected with nematode *P. penetrans* [287]. Hol et al. (2010) [291] reported a detrimental effect on aphids, *Brevicoryne brassicae* in the presence of nematodes (*H. schachtii*) in *B. oleracea*. This might be because of the disturbance on feeding relations between plants and aphids as nematodes reduced amino acid and sugar in the phloem and reduced glucapin concentration and increased gluconapoleiferin and 4-methoxyglucobrassicin concentration in leaves [280]. Also, the reproduction of aphid (*Schizaphis rufula*) was reduced in the presence of three nematodes (*Pratylenchus*, *Meloidogyne*, and *Heterodera* spp.) in the plant, *Ammophila arenaria* in laboratory conditions [292]. The possible reason might be mechanical factors such as changes in waxes of the cuticle, leaf toughness or water content in the presence of nematodes [293]. The water stress in the aerial part of the host plant might affect the insects that rely on phloem feeding [294]. Also, decreased shoot herbivory could be because of the accumulation of phenolics and glucosinolates [276, 277]. However, the changes in the concentration of plant metabolites in the host plant are independent of the presence of another herbivore [280]. A recent study by Hoysted et al. (2017) [279] reported the positive effect on the reproduction of aphids, *Myzus persicae*, on the presence of endoparasitic nematode (*Globodera pallida*) in *Solanum tuberosum*, which contrasts with the previous studies. The increased SA in the leaves and suppression of JA, when co-infected with the nematodes, played a positive effect in the *M. persicae*. There are also been some studies to show shoot aphids, in turn, possess the ability to affect nematode infections. The abundance of nematode, *Tylenchorhynchus* was decreased on aphid infested plants and there was no effect on *Pratylenchus* in *N. tabaccum* [295]. On

the same way, the abundance of *H. glycines* and *Meloidogyne incognita* was found to be increased when soybean plants were infested by *Pseudoplusia includes* or *Helicoverpa zea* caterpillars [288]. Another study by Kutyniok and Müller (2012) [285], showed that the presence of aphids reduced the number of nematodes, *Heterodera schachtii* and *Brevicoryne brassicae* when all added at the same time in *A. thaliana*. *Ostrinia nubilalis* caused a decreased abundance of *Meloidogyne incognita* infecting maize [289]. Hoysted et al. (2018) [296] reported the increase in the inoculum of *M. persicae*, inhibited the hatching of eggs of *G. pallida* as the content of fructose and glucose was decreased in the root exudates of aphid infested potato plant.

The feeding habit of nematodes and aphids, the sequence of the herbivory (which arrives first on the plant), duration of infestation by aphids, the extent of susceptibility to herbivores and identification of insects are considered as crucial factors in understanding interactions between nematodes and insects [284, 285, 297, 298]. It is expected that the above ground herbivore that arrives first on the plant negatively affects the subsequent below ground herbivore [298]. The presence of aboveground herbivore, *Spodoptera frugiperda* on the maize had negatively affected the colonization of below ground herbivore, *Diabrotica virgifera* if maize is infested by *S. frugiperda* first [299]. However, the interaction effects between aboveground and belowground herbivores can be positive, negative or neutral [292]. The feeding tomato plants by chewing caterpillars (*Spodoptera exigua*) and sucking aphids (*Myzus persicae*) did not show a negative effect on the root-knot nematodes (*Meloidogyne incognita*) [284]. However, plants showed compensatory growth of shoots in the form of tolerance response that was reduced by *S. exigua* upon nematode herbivory. Also, the plant responses can vary to subsequent herbivores

depending on the feeding mode of the above ground herbivores [284]. Not only the feeding mode of aboveground herbivore affect the belowground herbivore, but also the feeding habits of the belowground herbivore affect the performance of the aboveground herbivore. The study on the effect of cyst nematode *Heterodera schachtii,* and the root-knot nematode *Meloidogyne hapla* showed the differential performance of cabbage aphid, *Brevicoryne brassicae* in black mustard (*Brassica nigra)* plants [283]. The preference and population of *B. brassicae* were negatively affected in *H. schachtii* infested plants whereas, opposite effect in the plants infested with *M. hapla*. *H. schachtii* enhanced aphid induced-resistance through SA pathway whereas, *M. hapla* enhanced through the JA pathway. This suggests the cross-talk of different hormonal signaling pathways during an infestation of the plant with aboveground aphid and belowground nematodes with different feeding strategies [283].

1.10. Induced Susceptibility

The interaction between insect herbivores with their host plant creates a condition called induced susceptibility that assists other subsequent herbivores [300]. This type of susceptibility takes place between conspecifics on susceptible as well as resistant plants [300, 301]. The phenotype of conspecific can be both virulent and avirulent biotype. This can be explained by the increased survival capacity of avirulent *Myzus persicae* (Sulzer) on the initially fed resistant plant by avirulent *M. persicae* [302]. Hence, the diverse populations of both virulent and avirulent insects that appear phenotypically similar can stimulate induced susceptibility on the resistant plants [303]. Such effect of soybean aphid infestation on other pests colonizing soybean plants at the same time would be related to the suppression of host plant defense blocking jasmonate-dependent metabolic

pathways [304]. For instance, *A. glycines* reduces the activity of fatty acid desaturase 2 (FAD2) and fatty acid desaturase 6 (FAD6) in the fatty acid pathway, thus reducing polyunsaturated fatty acids such as linolenic acid, the precursor of jasmonate [304]. Varenhorst et al. (2015) [305] concluded that virulent *A. glycines* increase the suitability of resistant soybean for avirulent conspecifics. Induced susceptibility arises through two different ways in *A. glycines*: feeding facilitation and obviation of resistance [301, 306]. This was demonstrated in the experiment by Varenhorst et al. (2015) [307] on finding the duration of induced susceptibility to monitor the durability of *A. glycines* resistance in soybean and this effect, persistent till the inducer population. The authors suggested that further studies of virulent aphid and soybean with *Rag* gene should be conducted considering the amount of time in which plant is allowed to *A. glycines* for the only obviation of resistance (i.e., 120 h post-infections). The influence of cyst nematode, *H. glycines* on aphid, *Aphis glycines* infestation or *vice versa* has been studied on soybean [281, 282, 290, 308, 309]. The study of the interaction effect of SCN and SBA on 'Williams' soybean cultivar revealed that SBA choose the plants that are uninfected with SCN and the population growth of aphids remained unaffected by SCN infection [281] in laboratory conditions. Further, this study was validated in the natural field conditions (both open plots and experimental cages), where aphids preferably colonized uninfected soybean plants with SCN. Also, the population growth of the aphids remained almost the same in SCN infected and uninfected soybean plants. Further, the independent effect was observed in soybean yield in the presence of SBA and SCN in the field. The effect of SCN was related to a decline in soybean yield, whereas SBA was related to a decline in seed weight depending on their respective population densities. Heeren et al. (2012) [309]

utilized resistant and susceptible lines with respect to both SBA and SCN in order to study the interaction effect of SBA and SCN in the field conditions. The effect of SBA feeding on soybean on the SCN reproduction was not observed in any of the soybean cultivars as the SCN eggs and aphid densities, less than 100 SCN eggs per 100 cc of soil and less than 10 aphids per plant for <10 days, respectively, were too low in some of the cultivars. McCarville et al. (2012) [290] conducted experiment on different SCN susceptible (DK 28-52, IA 3018, IA 3041) and SCN resistant (DK 27-52, AG 2821 V, IA 3028) soybean cultivars to understand the effect of multiple pests/pathogens (SBA, SCN, and the fungus *Cadophora gregata*) interaction. The study showed that the SCN reproduction was increased (5.24 times) in the presence of SBA and *C. gregata*. In contrast, the aphid population decreased by 26.4% in the presence of SCN and *C. gregata* and the SCN resistant cultivars (derived from PI88788) reduced aphid exposure by 19.8%. Later, McCarville et al. (2014) [282] demonstrated the relationship between the aboveground feeding of SBA and reproduction belowground of SCN in the SCN resistant (Dekalb 27-52, PI88788 derived) and SCN susceptible (Kenwood 94) soybean cultivars. The authors concluded that SBA feeding improved the quality of soybean as a host for SCN, but this result varied significantly with the cultivar and length of the experiment. In 30- days, the SCN eggs and females increased by 33% (1.34 times) in SCN-resistant cultivar and reduced by 50% in the SCN-susceptible cultivar. In 60-days, the numbers of SCN eggs and females remained unaffected in the resistant cultivar but decreased in the susceptible cultivar.

Table 1.4 Major host-nematode-aphid interaction studies in diverse host systems

Host	Nematode	Aphids	Effect	Chemistry	References
Brassica nigra	*Pratylenchus penetrans*	*Pieris rapae*	Negative effect on aphids	Increased phenolics and glucosinate levels	[276]
Agrostis capillaris, Anthoxanthum odoratum	Paratylenchidae, Pratylenchidae, and Dolichodoridae	*Rhopalosiphum padi* plus *Aphidius coleman*	Negative effect on aphid population/Reduced parasitoid mortality	Decreased foliar phenolic content and amino acid in phloem sap	[286]
Plantago lanceolata	*Pratylenchus Penetrans*	*Myzus persicae*	Negative effect on aphid population	-	[287]
Brassica oleracea	*Heterodera Schachtii*	*Brevicoryne brassicae*	Reduced body size of aphids	-	[291]
Ammophila arenaria	*Pratylenchus*, *Meloidogyne*, and *Heterodera spp*	*Schizaphis rufula*	Nematodes and aphids negatively affect each other	Reduction of foliar nitrogen and amino acid	[292]
Nicotiana tabacum	*Meloidogyne incognita*	*Trichoplusia ni* and *Manduca sexta*	Positive effects on aboveground aphids	Change of foliar nicotine dynamics	[277]
Brassica oleracea	Nematode species dominant of Cephalobidae and Rhabditidae families	*Brevicoryne brassicae*	Negative effect on aphid density	-	[310]
Arabidopsis thaliana	*Heterodera schachtii*	*Brevicoryne brassicae*	No effect on aphid growth in presence of nematode/reduced number of nematodes in presence of aphids	Reduced glucosinolates in shoots	[285]
Brassica oleracea	*Heterodera schachtii*	*Brevicoryne brassicae*	Increase in aphid doubling time from 3.8 to 6.7 days	Reduced glucapin /Increased gluconapoleiferin and 4-methoxyglucobrassicin in leaves/Decreased amino acid and sugar in phloem	[280]
Solanum tuberosum	*Globodera pallida*	*Myzus persicae*	Positive effect on the reproduction of aphids	Increased SA in the leaves and suppression of JA	[279]
Brassica nigra	*Heterodera schachtii*	*Brevicoryne brassicae*	Lower preference of aphids/ lower reproduction of aphids	Induced PR1 (SA pathway) Reduced VSP2 and MYC2 (JA pathway)	[283]
Brassica nigra	*Meloidogyne hapla*	*Brevicoryne brassicae*	Higher preference of ahids/higher reproduction	No PR1 expression/ High VSP2 and MYC2 expression	[283]
Nicotiana tabacum	*Meloidogyne incognita*, *Tylenchorhynchus* and *Pratylenchus*	*Myzus persicae*	Reduced the abundance of aphids/ *Tylenchorhynchus* was decreased on aphid	-	[295]

			infested plants/no effect on *Pratylenchus*		
Zea mays	*Meloidogyne incognita*	*Ostrinia nubilalis*	Reduced nematode reproduction	-	[289]
Solanum tuberosum	*Globodera pallida*	*Myzus persicae*	Inhibited the hatching of eggs of the nematode	Decreased fructose and glucose in the root exudates	[296]
Solanum tuberosum	*Meloidogyne incognita*	*Myzus persicae*	No effect on the nematodes	Decreased the root SA content	[284]
Glycine max	*Heterodera glycines*	*Aphis glycines*	Aphids unaffected/aphid preference	-	[281, 308]
Glycine max	*Heterodera glycines*	*Aphis glycines*	No effect of aphid on SCN reproduction	-	[309]
Glycine max	*Heterodera glycines* plus *Cadophora gregata*	*Aphis glycines*	SCN reproduction increased (5.24 times) in the presence of SBA and *C. gregata/* aphid population decreased by 26.4% in the presence of SCN and *C. gregata* in PI88788 derived cultivar	-	[290]
Glycine max	*Heterodera glycines*	*Aphis glycines*	SCN eggs and females increased by 33% (1.34 times) in SCN-resistant cultivar/reduced by 50% in the SCN-susceptible cultivar.	-	[282]

In summary, this literature review provides insights into the molecular mechanisms of how R genes and MAPK genes are involved in regulating abiotic and biotic stresses including soybean-SBA-SCN interactions. Most of the previous studies agree that both SBA and SCN do not depend on a single gene or do not rely on just R gene-mediated resistance. The resistance in the soybean is controlled by several genes such as *Rag* genes for soybean aphid and *Rhg* genes for SCN, and in fact, soybean resistance to these pests isquantitative resistance. Gene pyramiding and integration of integrated pest management (IPM) could prove promising for soybean crop improvement with durable resistance. This dissertation aids this effort by unraveling stress responsive genes in soybean, including those involved in soybean-SBA-SCN interactions.

References

1.   Yu, X.; Yuan, F.; Fu, X.; Zhu, D. Profiling and relationship of water-soluble sugar and protein compositions in soybean seeds. *Food chemistry* 2016, *196*, 776-782.
2.   Wilson, R.F. Soybean: market driven research needs. In *Genetics and genomics of soybean*; Springer: 2008; pp. 3-15.
3.   Gai, J.; Guo, W. In *History of Maodou production in China*, Proceedings Second International Vegetable Soybean Conference, Washington State University, Pullman, WA, 2001; 2001; pp. 41-47.
4.   Hymowitz, T. Soybeans: The success story. *Advances in new crops* 1990, 159-163.
5.   Hymowitz, T.; Harlan, J.R. Introduction of soybean to North America by Samuel Bowen in 1765. *Economic Botany* 1983, *37*, 371-379.
6.   Kane, N.C.; Rieseberg, L.H. Selective sweeps reveal candidate genes for adaptation to drought and salt tolerance in common sunflower, Helianthus annuus. *Genetics* 2007, *175*, 1823-1834.
7.   Fernández-Martínez, J.; Melero-Vara, J.; Muñoz-Ruz, J.; Ruso, J.; Domínguez, J. Selection of wild and cultivated sunflower for resistance to a new broomrape race that overcomes resistance of the gene. *Crop Science* 2000, *40*, 550-555.
8.   Seiler, G. Wild annual *Helianthus anomalus* and *H. deserticola* for improving oil content and quality in sunflower. *Industrial Crops and Products* 2007, *25*, 95-100.

9.   Markell, S.; Harveson, R.; Block, C.; Gulya, T. Sunflower disease diagnostic series. *North Dakota State University Extension Series Publications PP1727-19* 2005.

10.   Baluška, F.; Mancuso, S., Signaling in Plants, Signaling and Communication in Plants. In Springer-Verlag, Berlin/Heidelberg: 2009.

11.   Wang, J.; Pan, C.; Wang, Y.; Ye, L.; Wu, J.; Chen, L.; Zou, T.; Lu, G. Genome-wide identification of MAPK, MAPKK, and MAPKKK gene families and transcriptional profiling analysis during development and stress response in cucumber. *BMC Genomics* 2015, *16*, 386.

12.   Plocik, A.; Layden, J.; Kesseli, R. Comparative analysis of NBS domain sequences of NBS-LRR disease resistance genes from sunflower, lettuce, and chicory. *Molecular Phylogenetics and Evolution* 2004, *31*, 153-163.

13.   Hewezi, T.; Mouzeyar, S.; Thion, L.; Rickauer, M.; Alibert, G.; Nicolas, P.; Kallerhoff, J. Antisense expression of a NBS-LRR sequence in sunflower (*Helianthus annuus* L.) and tobacco (Nicotiana tabacum L.): evidence for a dual role in plant development and fungal resistance. *Transgenic Research* 2006, *15*, 165-180.

14.   Radwan, O.; Gandhi, S.; Heesacker, A.; Whitaker, B.; Taylor, C.; Plocik, A.; Kesseli, R.; Kozik, A.; Michelmore, R.W.; Knapp, S.J. Genetic diversity and genomic distribution of homologs encoding NBS-LRR disease resistance proteins in sunflower. *Molecular Genetics and Genomics* 2008, *280*, 111-125.

15.   Radwan, O.; Mouzeyar, S.; Nicolas, P.; Bouzidi, M. Induction of a sunflower CC-NBS-LRR resistance gene analogue during incompatible interaction with Plasmopara halstedii. *Journal of Experimental Botany* 2004, *56*, 567-575.

16.   Jones, J.D.; Dangl, J.L. The plant immune system. *Nature* 2006, *444*, 323-329.

17.   Zipfel, C. Pattern-recognition receptors in plant innate immunity. *Current Opinion in Immunology* 2008, *20*, 10-16.

18.   Shao, Z.-Q.; Xue, J.-Y.; Wu, P.; Zhang, Y.-M.; Wu, Y.; Hang, Y.-Y.; Wang, B.; Chen, J.-Q. Large-scale analyses of angiosperm nucleotide-binding site-leucine-rich repeat (NBS-LRR) genes reveal three anciently diverged classes with distinct evolutionary patterns. *Plant Physiology* 2016, pp. 01487.02015.

19.   Lee, H.A.; Yeom, S.I. Plant NB-LRR proteins: tightly regulated sensors in a complex manner. *Brief Functional Genomics* 2015, *14*, 233-242.

20.   Michelmore, R.W.; Christopoulou, M.; Caldwell, K.S. Impacts of resistance gene genetics, function, and evolution on a durable future. *Annual Review of Phytopathology* 2013, *51*, 291-319.

21.   Tena, G.; Asai, T.; Chiu, W.-L.; Sheen, J. Plant mitogen-activated protein kinase signaling cascades. *Current Opinion in Plant Biology* 2001, *4*, 392-400.

22.   Ichimura, K.; Shinozaki, K.; Tena, G.; Sheen, J.; Henry, Y.; Champion, A.; Kreis, M.; Zhang, S.; Hirt, H.; Wilson, C. Mitogen-activated protein kinase cascades in plants: a new nomenclature. *Trends in Plant Science* 2002, *7*, 301-308.

23.   Cristina, M.S.; Petersen, M.; Mundy, J. Mitogen-activated protein kinase signaling in plants. *Annual Review of Plant Biology* 2010, *61*, 621-649.

24.   Jonak, C.; Ökrész, L.; Bögre, L.; Hirt, H. Complexity, cross talk and integration of plant MAP kinase signalling. *Current Opinion in Plant Biology* 2002, *5*, 415-424.

25.	Neupane, S.; Schweitzer, S.E.; Neupane, A.; Andersen, E.J.; Fennell, A.; Zhou, R.; Nepal, M.P. Identification and Characterization of Mitogen-Activated Protein Kinase (MAPK) Genes in Sunflower (*Helianthus annuus* L.). *Plants* 2019, *8*, 28.

26.	Mohanta, T.K.; Arora, P.K.; Mohanta, N.; Parida, P.; Bae, H. Identification of new members of the MAPK gene family in plants shows diverse conserved domains and novel activation loop variants. *BMC Genomics* 2015, *16*, 58.

27.	Nadarajah, K.; Sidek, H.M. The green MAPKS. *Asian Journal of Plant Sciences* 2010, *9*, 1.

28.	Rohila, J.S.; Yang, Y. Rice mitogen-activated protein kinase gene family and its role in biotic and abiotic stress response. *Journal of Integrative Plant Biology* 2007, *49*, 751-759.

29.	Kissoudis, C.; van de Wiel, C.; Visser, R.G.; van der Linden, G. Enhancing crop resilience to combined abiotic and biotic stress through the dissection of physiological and molecular crosstalk. *Frontiers in Plant Science* 2014, *5*, 207.

30.	Ranty, B.; Aldon, D.; Cotelle, V.; Galaud, J.-P.; Thuleau, P.; Mazars, C. Calcium sensors as key hubs in plant responses to biotic and abiotic stresses. *Frontiers in Plant Science* 2016, *7*, 327.

31.	Pitzschke, A.; Schikora, A.; Hirt, H. MAPK cascade signalling networks in plant defence. *Current Opinion in Plant Biology* 2009, *12*, 421-426.

32.	Müller-Xing, R.; Xing, Q.; Goodrich, J. Footprints of the sun: memory of UV and light stress in plants. *Frontiers in Plant Science* 2014, *5*, 474.

33.	Davis, P.K.; Brachmann, R.K. Chromatin remodeling and cancer. *Cancer Biology & Therapy* 2003, *2*, 23-30.

34.	Rejeb, I.B.; Pastor, V.; Mauch-Mani, B. Plant responses to simultaneous biotic and abiotic stress: molecular mechanisms. *Plants* 2014, *3*, 458-475.

35.	Meyers, B.C.; Kozik, A.; Griego, A.; Kuang, H.; Michelmore, R.W. Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. *Plant Cell* 2003, *15*.

36.	Ramirez-Prado, J.S.; Abulfaraj, A.A.; Rayapuram, N.; Benhamed, M.; Hirt, H. Plant immunity: from signaling to epigenetic control of defense. *Trends in Plant Science* 2018.

37.	Ragsdale, D.W.; Landis, D.A.; Brodeur, J.; Heimpel, G.E.; Desneux, N. Ecology and management of the soybean aphid in North America. *Annual Review of Entomology* 2011, *56*, 375-399.

38.	Wu, Z.; Schenk-Hamlin, D.; Zhan, W.; Ragsdale, D.W.; Heimpel, G.E. The soybean aphid in China: a historical review. *Annals of. Entomological Society of America* 2004, *97*, 209-218.

39.	Matsumura, S. A list of the Aphididae of Japan, with description of new species and genera. *The journal of the College of Agriculture, Tohoku Imperial University, Sapporo, Japan= 東北帝國大學農科大學紀要* 1917, *7*, 351-414.

40.	Hartman, G.; Domier, L.; Wax, L.; Helm, C.; Onstad, D.; Shaw, J.; Solter, L.; Voegtlin, D.; d'Arcy, C.; Gray, M. Occurrence and distribution of Aphis glycines on soybeans in Illinois in 2000 and its potential control. *Plant Health Progr. doi* 2001, *10*.

41.	Takahashi, S.; Inaizumi, M.; Kawakami, K. Life cycle of the soybean aphid Aphis glycines Matsumura, in Japan. *Japanese Journal of Applied Entomology and Zoology* 1993, *37*, 207-212.

42. Voegtlin, D.J.; O'neil, R.J.; Graves, W.R.; Lagos, D.; Yoo, H.J.S. Potential winter hosts of soybean aphid. *Annals of. Entomological Society of America* 2005, *98*, 690-693.

43. McCornack, B.; Carrillo, M.; Venette, R.; Ragsdale, D.J.E.e. Physiological constraints on the overwintering potential of the soybean aphid (Homoptera: Aphididae). *Environmental Entomology* 2005, *34*, 235-240.

44. Ragsdale, D.W.; Voegtlin, D.J.; O'neil, R.J. Soybean aphid biology in North America. *Annals of. Entomological Society of America* 2004, *97*, 204-208.

45. Clark, T.L.; Puttler, B.; Bailey, W.C. Is horsenettle, Solanum carolinense L.(Solanaceae), an alternate host for soybean aphid, Aphis glycines Matsumura (Hemiptera: Aphididae)? *Journal of the Kansas. Entomological. Society* 2006, *79*, 380-383.

46. Ragsdale, D.W.; McCornack, B.; Venette, R.; Potter, B.; MacRae, I.V.; Hodgson, E.W.; O'Neal, M.E.; Johnson, K.D.; O'neil, R.; DiFonzo, C. Economic threshold for soybean aphid (Hemiptera: Aphididae). *Journal of Economic Entomology* 2007, *100*, 1258-1267.

47. McCornack, B.P.; Costamagna, A.; Ragsdale, D. Within-plant distribution of soybean aphid (Hemiptera: Aphididae) and development of node-based sample units for estimating whole-plant densities in soybean. *Journal of Economic Entomology* 2008, *101*, 1488-1500.

48. Chandran, P.; Reese, J.C.; Khan, S.A.; Wang, D.; Schapaugh, W.; Campbell, L.R. Feeding behavior comparison of soybean aphid (Hemiptera: Aphididae) biotypes on different soybean genotypes. *Journal of Economic Entomology* 2013, *106*, 2234-2240.

49. Fox, C.M.; Kim, K.-S.; Cregan, P.B.; Hill, C.B.; Hartman, G.L.; Diers, B.W. Inheritance of soybean aphid resistance in 21 soybean plant introductions. *Theoretical and Applied Genetics* 2014, *127*, 43-50.

50. Beckendorf, E.A.; Catangui, M.A.; Riedell, W.E. Soybean aphid feeding injury and soybean yield, yield components, and seed composition. *Agronomy Journal* 2008, *100*, 237-246.

51. Hill, C.; Chirumamilla, A.; Hartman, G. Resistance and virulence in the soybean-Aphis glycines interaction. *Euphytica* 2012, *186*, 635-646.

52. Pedersen, P.; Grau, C.; Cullen, E.; Koval, N.; Hill, J.H. Potential for integrated management of soybean virus disease. *Plant Dis.* 2007, *91*, 1255-1259.

53. Smith, C.M. *Plant resistance to insects. A fundamental approach*; John Wiley and Sons Ltd.: 1989; p.^pp.

54. Saucet, S.B.; Shirasu, K. Molecular parasitic plant–host interactions. *PLoS Pathogens* 2016, *12*, e1005978.

55. Hogenhout, S.A.; Bos, J.I. Effector proteins that modulate plant–insect interactions. *Current Opinion in Plant Biology* 2011, *14*, 422-428.

56. Na, R.; Gijzen, M. Escaping host immunity: new tricks for plant pathogens. *PLoS Pathology* 2016, *12*, e1005631.

57. Gardner, M.; Dhroso, A.; Johnson, N.; Davis, E.L.; Baum, T.J.; Korkin, D.; Mitchum, M.G. Novel global effector mining from the transcriptome of early life stages of the soybean cyst nematode Heterodera glycines. *Scientific Reports* 2018, *8*, 2505.

58.   Moreno, A.; Garzo, E.; Fernandez-Mata, G.; Kassem, M.; Aranda, M.; Fereres, A. Aphids secrete watery saliva into plant tissues from the onset of stylet penetration. *Entomologia Experimentalis et Applicata* 2011, *139*, 145-153.

59.   Züst, T.; Agrawal, A.A. Mechanisms and evolution of plant resistance to aphids. *Nature Plants* 2016, *2*, 15206.

60.   Will, T.; Furch, A.C.; Zimmermann, M.R. How phloem-feeding insects face the challenge of phloem-located defenses. *Frontiers in Plant Science* 2013, *4*, 336.

61.   Varden, F.A.; De la Concepcion, J.C.; Maidment, J.H.; Banfield, M.J. Taking the stage: effectors in the spotlight. *Current Opinion in Plant Biology* 2017, *38*, 25-33.

62.   Boulain, H.; Legeai, F.; Guy, E.; Morlière, S.; Douglas, N.E.; Oh, J.; Murugan, M.; Smith, M.; Jaquiéry, J.; Peccoud, J. Fast Evolution and Lineage-Specific Gene Family Expansions of Aphid Salivary Effectors Driven by Interactions with Host-Plants. *Genome Biology and Evolution* 2018, *10*, 1554-1572.

63.   Rodriguez, P.; Escudero-Martinez, C.; Bos, J. An aphid effector targets trafficking protein VPS52 in a host-specific manner to promote virulence. *Plant Physiology* 2017, pp. 01458.02016.

64.   Pitino, M.; Hogenhout, S.A. Aphid protein effectors promote aphid colonization in a plant species-specific manner. *Molecular Plant-Microbe Interactions* 2013, *26*, 130-139.

65.   Mutti, N.S.; Park, Y.; Reese, J.C.; Reeck, G.R. RNAi knockdown of a salivary transcript leading to lethality in the pea aphid, *Acyrthosiphon pisum*. *Journal of Insect Science* 2006, *6*.

66.   IAGC, I.A.G.C. Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLoS Biology* 2010, *8*, e1000313.

67.   Nicholson, S.J.; Nickerson, M.L.; Dean, M.; Song, Y.; Hoyt, P.R.; Rhee, H.; Kim, C.; Puterka, G.J. The genome of Diuraphis noxia, a global aphid pest of small grains. *BMC Genomics* 2015, *16*, 429.

68.   Mathers, T.C.; Chen, Y.; Kaithakottil, G.; Legeai, F.; Mugford, S.T.; Baa-Puyoulet, P.; Bretaudeau, A.; Clavijo, B.; Colella, S.; Collin, O. Rapid transcriptional plasticity of duplicated gene clusters enables a clonally reproducing aphid to colonise diverse plant species. *Genome Biology and Evolution* 2017, *18*, 27.

69.   Wenger, J.A.; Cassone, B.J.; Legeai, F.; Johnston, J.S.; Bansal, R.; Yates, A.D.; Coates, B.S.; Pavinato, V.A.; Michel, A. Whole genome sequence of the soybean aphid, Aphis glycines. *Insect Biochemistry and Molecular Biology* 2017.

70.   Carolan, J.C.; Caragea, D.; Reardon, K.T.; Mutti, N.S.; Dittmer, N.; Pappan, K.; Cui, F.; Castaneto, M.; Poulain, J.; Dossat, C. Predicted effector molecules in the salivary secretome of the pea aphid (*Acyrthosiphon pisum*): a dual transcriptomic/proteomic approach. *Journal of proteome research* 2011, *10*, 1505-1518.

71.   Harmel, N.; Létocart, E.; Cherqui, A.; Giordanengo, P.; Mazzucchelli, G.; Guillonneau, F.; De Pauw, E.; Haubruge, E.; Francis, F. Identification of aphid salivary proteins: a proteomic investigation of *Myzus persicae*. *Insect Biochemistry and Molecular Biology* 2008, *17*, 165-174.

72. Carolan, J.C.; Fitzroy, C.I.; Ashton, P.D.; Douglas, A.E.; Wilkinson, T.L. The secreted salivary proteome of the pea aphid *Acyrthosiphon pisum* characterised by mass spectrometry. *Proteomics* 2009, *9*, 2457-2467.

73. Wang, W.; Luo, L.; Lu, H.; Chen, S.; Kang, L.; Cui, F. Angiotensin-converting enzymes modulate aphid–plant interactions. *Scientific Reports* 2015, *5*, 8885.

74. Bellafiore, S.; Shen, Z.; Rosso, M.-N.; Abad, P.; Shih, P.; Briggs, S.P. Direct identification of the Meloidogyne incognita secretome reveals proteins with host cell reprogramming potential. *PLoS Pathogens* 2008, *4*, e1000192.

75. Jones, J.; Reavy, B.; Smant, G.; Prior, A. Glutathione peroxidases of the potato cyst nematode *Globodera rostochiensis*. *Gene* 2004, *324*, 47-54.

76. Griffith, M.; Walker, J.R.; Spies, N.C.; Ainscough, B.J.; Griffith, O.L. Informatics for RNA sequencing: a web resource for analysis on the cloud. *PLoS Computational Biology* 2015, *11*, e1004393.

77. Shan, X.; Li, Y.; Jiang, Y.; Jiang, Z.; Hao, W.; Yuan, Y. Transcriptome profile analysis of maize seedlings in response to high-salinity, drought and cold stresses by deep sequencing. *Plant Molecular Biology Reporter* 2013, *31*, 1485-1491.

78. Bansal, R.; Mian, M.; Mittapalli, O.; Michel, A.P. RNA-Seq reveals a xenobiotic stress response in the soybean aphid, Aphis glycines, when fed aphid-resistant soybean. *BMC Genomics* 2014, *15*, 972.

79. Ramsey, J.S.; Rider, D.S.; Walsh, T.K.; De Vos, M.; Gordon, K.; Ponnala, L.; Macmil, S.; Roe, B.; Jander, G. Comparative analysis of detoxification enzymes in *Acyrthosiphon pisum* and *Myzus persicae*. *Insect Molecular Biology* 2010, *19*, 155-164.

80. Wenger, J.A.; Michel, A.P. Implementing an evolutionary framework for understanding genetic relationships of phenotypically defined insect biotypes in the invasive soybean aphid (*Aphis glycines*). *Evolutionary applications* 2013, *6*, 1041-1053.

81. Downie, D. Baubles, bangles, and biotypes: a critical review of the use and abuse of the biotype concept. *Journal of Insect Science* 2010, *10*.

82. Varenhorst, A.J. Beyond biotypes: *Aphis glycines* (Hemiptera: Aphididae) biology and the durability of aphid-resistant soybean. *Graduate Theses and Dissertations*. 14202. 2015. https://lib.dr.iastate.edu/etd/14202

83. Cooper, S.G.; Concibido, V.; Estes, R.; Hunt, D.; Jiang, G.-L.; Krupke, C.; McCornack, B.; Mian, R.; O'Neal, M.; Poysa, V. Geographic distribution of soybean aphid biotypes in the United States and Canada during 2008–2010. *Crop Science* 2015, *55*, 2598-2608.

84. Michel, A.P.; Mittapalli, O.; Mian, M.R. Evolution of soybean aphid biotypes: understanding and managing virulence to host-plant resistance. In *Soybean-Molecular Aspects of Breeding*; InTech: 2011.

85. Kim, K.-S.; Hill, C.B.; Hartman, G.L.; Mian, M.; Diers, B.W. Discovery of soybean aphid biotypes. *Crop Science* 2008, *48*, 923-928.

86. Hesler, L.S.; Chiozza, M.V.; O'neal, M.E.; MacIntosh, G.C.; Tilmon, K.J.; Chandrasena, D.I.; Tinsley, N.A.; Cianzio, S.R.; Costamagna, A.C.; Cullen, E.M. Performance and prospects of R ag genes for management of soybean aphid. *Entomologia Experimentalis et Applicata* 2013, *147*, 201-216.

87. Hill, C.B.; Crull, L.; Herman, T.K.; Voegtlin, D.J.; Hartman, G.L. A new soybean aphid (Hemiptera: Aphididae) biotype identified. *Journal of Economic Entomology* 2010, *103*, 509-515.

88. Alt, J.; Ryan-Mahmutagic, M. Soybean aphid biotype 4 identified. *Crop Science* 2013, *53*, 1491-1495.

89. Zhong, Y.P.; Xiao, L.; Wang, B.; Jiang, Y.N.; Yan, J.H.; Cheng, L.J.; Wu, T.L. Biotypic variation among soybean aphid isolates from four provinces in China. *Crop Science* 2014, *54*, 2023-2029.

90. Painter, R.H. *Insect resistance in crop plants*; The Macmillan Company New York: 1951; p.^pp.

91. Baldin, E.L.; Stamm, M.D.; Bentivenha, J.P.; Koch, K.G.; Heng-Moss, T.M.; Hunt, T.E. Feeding Behavior of Aphis glycines (Hemiptera: Aphididae) on Soybeans Exhibiting Antibiosis, Antixenosis, and Tolerance Resistance. *Florida Entomology* 2018, *101*, 223-228.

92. Jesus, F.; Marchi-Werle, L.; Fischer, H.; Posadas, L.; Graef, G.; Heng-Moss, T. Documenting Resistance and Physiological Changes in Soybean Challenged by *Aphis glycines* Matsumura (Hemiptera: Aphididae). *Neotropical entomology* 2018, 1-8.

93. Brechenmacher, L.; Nguyen, T.H.N.; Zhang, N.; Jun, T.-H.; Xu, D.; Mian, M.R.; Stacey, G. Identification of soybean proteins and genes differentially regulated in near isogenic lines differing in resistance to aphid infestation. *Journal of Proteome Research* 2015, *14*, 4137-4146.

94. Prochaska, T.J.; Donze-Reiner, T.; Marchi-Werle, L.; Palmer, N.; Hunt, T.E.; Sarath, G.; Heng-Moss, T. Transcriptional responses of tolerant and susceptible soybeans to soybean aphid (*Aphis glycines* Matsumura) herbivory. *Arthropod-Plant Interactions* 2015, *9*, 347-359.

95. Pierson, L.; Heng-Moss, T.; Hunt, T.; Reese, J. Physiological responses of resistant and susceptible reproductive stage soybean to soybean aphid (Aphis glycines Matsumura) feeding. *Arthropod-Plant Interactions* 2011, *5*, 49-58.

96. Hill, C.B.; Li, Y.; Hartman, G.L. Resistance to the soybean aphid in soybean germplasm. *Crop Science* 2004, *44*, 98-106.

97. Hill, C.B.; Li, Y.; Hartman, G.L. A single dominant gene for resistance to the soybean aphid in the soybean cultivar Dowling. *Crop Science* 2006, *46*, 1601-1605.

98. Hill, C.B.; Li, Y.; Hartman, G.L. Soybean aphid resistance in soybean Jackson is controlled by a single dominant gene. *Crop Science* 2006, *46*, 1606-1608.

99. Li, Y.; Hill, C.B.; Carlson, S.R.; Diers, B.W.; Hartman, G.L. Soybean aphid resistance genes in the soybean cultivars Dowling and Jackson map to linkage group M. *Molecular Breeding* 2007, *19*, 25-34.

100. Van Nurden, A.; Scott, R.; Hesler, L.; Tilmon, K.; Glover, K.; Carter, C. Inheritance of soybean aphid resistance from PI71506. *Journal of Crop Improvement* 2010, *24*, 400-416.

101. Kim, K.-S.; Bellendir, S.; Hudson, K.A.; Hill, C.B.; Hartman, G.L.; Hyten, D.L.; Hudson, M.E.; Diers, B.W. Fine mapping the soybean aphid resistance gene Rag1 in soybean. *Theoretical and Applied Genetics* 2010, *120*, 1063-1071.

102. Kim, K.-S.; Hill, C.B.; Hartman, G.L.; Hyten, D.L.; Hudson, M.E.; Diers, B.W. Fine mapping of the soybean aphid-resistance gene Rag2 in soybean PI 200538. *Theoretical and Applied Genetics* 2010, *121*, 599-610.
103. Zhang, G.; Gu, C.; Wang, D. A novel locus for soybean aphid resistance. *Theoretical and Applied Genetics* 2010, *120*, 1183-1191.
104. Zhang, G.; Gu, C.; Wang, D. Molecular mapping of soybean aphid resistance genes in PI 567541B. *Theoretical and Applied Genetics* 2009, *118*, 473-482.
105. Jun, T.-H.; Mian, M.R.; Michel, A.P. Genetic mapping revealed two loci for soybean aphid resistance in PI 567301B. *Theoretical and Applied Genetics* 2012, *124*, 13-22.
106. Bales, C.; Zhang, G.; Liu, M.; Mensah, C.; Gu, C.; Song, Q.; Hyten, D.; Cregan, P.; Wang, D. Mapping soybean aphid resistance genes in PI 567598B. *Theoretical and Applied Genetics* 2013, *126*, 2081-2091.
107. Bhusal, S.J.; Jiang, G.-L.; Song, Q.; Cregan, P.B.; Wright, D.; Gonzalez-Hernandez, J.L. Genome-wide detection of genetic loci associated with soybean aphid resistance in soybean germplasm PI 603712. *Euphytica* 2017, *213*, 144.
108. Hill, C.B.; Shiao, D.; Fox, C.M.; Hartman, G.L. Characterization and genetics of multiple soybean aphid biotype resistance in five soybean plant introductions. *Theoretical and Applied Genetics* 2017, 1-14.
109. Zhou, Z.; Jiang, Y.; Wang, Z.; Gou, Z.; Lyu, J.; Li, W.; Yu, Y.; Shu, L.; Zhao, Y.; Ma, Y. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature Biotechnology* 2015, *33*, 408.
110. Hesler, L.S.; Tilmon, K.J. Resistance to Aphis glycines among wild soybean accessions in laboratory experiments. *Crop Protection* 2018, *112*, 74-82.
111. Conzemius, S.R. Soybean Aphid Biotype 4 Resistance in Soja and Soybean Plant Introductions. *Electronic Theses and Dissertations*. 2483. 2018. https://openprairie.sdstate.edu/etd/2483
112. Zhang, S.; Zhang, Z.; Wen, Z.; Gu, C.; An, Y.-Q.C.; Bales, C.; DiFonzo, C.; Song, Q.; Wang, D. Fine mapping of the soybean aphid-resistance genes Rag6 and Rag3c from Glycine soja 85-32. *Theoretical and Applied Genetics* 2017, *130*, 2601-2615.
113. Mian, M.R.; Kang, S.-T.; Beil, S.E.; Hammond, R.B. Genetic linkage mapping of the soybean aphid resistance gene in PI 243540. *Theoretical and Applied Genetics* 2008, *117*, 955-962.
114. Wang, D.; Bales, C.; Yuan, J.; Zhang, Z., Aphid resistant soybean plants. In Google Patents: 2015.
115. Zhang, G.; Gu, C.; Wang, D. Mapping and validation of a gene for soybean aphid resistance in PI 567537. *Molecular Breeding* 2013, *32*, 131-138.
116. Yeats, T.H.; Rose, J.K. The formation and function of plant cuticles. *Plant Physiology* 2013, pp. 113.222737.
117. Tian, T.; Liu, Y.; Yan, H.; You, Q.; Yi, X.; Du, Z.; Xu, W.; Su, Z. agriGO v2. 0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Research* 2017, *45*, W122-W129.

118. Hanson, A.A.; Lorenz, A.J.; Hesler, L.S.; Bhusal, S.J.; Bansal, R.; Michel, A.P.; Jiang, G.-L.; Koch, R.L. Genome-wide association mapping of host-plant resistance to soybean aphid. *The Plant Genome* 2018.

119. Zhang, H.; Song, Q.; Griffin, J.D.; Song, B.-H. Genetic architecture of wild soybean (Glycine soja) response to soybean cyst nematode (*Heterodera glycines*). *Molecular Genetics Genomics* 2017, *292*, 1257-1265.

120. Vuong, T.D.; Sonah, H.; Meinhardt, C.G.; Deshmukh, R.; Kadam, S.; Nelson, R.L.; Shannon, J.G.; Nguyen, H.T. Genetic architecture of cyst nematode resistance revealed by genome-wide association study in soybean. *BMC Genomics* 2015, *16*, 593.

121. Chang, H.-X.; Lipka, A.E.; Domier, L.L.; Hartman, G.L. Characterization of disease resistance loci in the USDA soybean germplasm collection using genome-wide association studies. *Phytopathology* 2016, *106*, 1139-1151.

122. Rincker, K.; Lipka, A.E.; Diers, B.W. Genome-wide association study of brown stem rot resistance in soybean across multiple populations. *The Plant Genome* 2016, *9*.

123. Yu, J.; Pressoir, G.; Briggs, W.H.; Bi, I.V.; Yamasaki, M.; Doebley, J.F.; McMullen, M.D.; Gaut, B.S.; Nielsen, D.M.; Holland, J.B. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* 2006, *38*, 203.

124. Chang, H.-X.; Hartman, G.L. Characterization of Insect Resistance Loci in the USDA Soybean Germplasm Collection Using Genome-Wide Association Studies. *Frontiers in Plant Science* 2017, *8*.

125. Varenhorst, A.J.; McCarville, M.T.; O'Neal, M.E. Reduced fitness of virulent *Aphis glycines* (Hemiptera: Aphididae) biotypes may influence the longevity of resistance genes in soybean. *PloS one* 2015, *10*, e0138252.

126. Varenhorst, A.; Pritchard, S.; O'Neal, M.; Hodgson, E.; Singh, A. Determining the effectiveness of three-gene pyramids against Aphis glycines (Hemiptera: Aphididae) biotypes. *Journal of Economic Entomology* 2017, *110*, 2428-2435.

127. McCarville, M.T.; O'Neal, M.E.; Potter, B.; Tilmon, K.J.; Cullen, E.M.; McCornack, B.P.; Tooker, J.F.; Prischmann-Voldseth, D. One gene versus two: a regional study on the efficacy of single gene versus pyramided resistance for soybean aphid management. *Journal of Economic Entomology* 2014, *107*, 1680-1687.

128. Ajayi-Oyetunde, O.; Diers, B.; Lagos-Kutz, D.; Hill, C.; Hartman, G.; Reuter-Carlson, U.; Bradley, C. Differential reactions of soybean isolines with combinations of aphid resistance genes Rag1, Rag2, and Rag3 to four soybean aphid biotypes. *Journal of Economic Entomology* 2016, *109*, 1431-1437.

129. McCarville, M.T.; Hodgson, E.W.; O'Neal, M.E. Soybean aphid-resistant soybean varieties for Iowa. 2012.

130. Chapman, K.M.; Marchi-Werle, L.; Hunt, T.E.; Heng-Moss, T.M.; Louis, J. Abscisic and Jasmonic Acids Contribute to Soybean Tolerance to the Soybean Aphid (*Aphis glycines* Matsumura). *Scientific Reports* 2018, *8*, 15148.

131. Selig, P.; Keough, S.; Nalam, V.J.; Nachappa, P. Jasmonate-dependent plant defenses mediate soybean thrips and soybean aphid performance on soybean. *Arthropod-Plant Interactions* 2016, *10*, 273-282.

132. Studham, M.E.; MacIntosh, G.C. Phytohormone signaling pathway analysis method for comparing hormone responses in plant-pest interactions. *BMC research notes* 2012, *5*, 392.

133. Li, Y.; Zou, J.; Li, M.; Bilgin, D.D.; Vodkin, L.O.; Hartman, G.L.; Clough, S.J. Soybean defense responses to the soybean aphid. *New Phytologist* 2008, *179*, 185-195.

134. Studham, M.E.; MacIntosh, G.C. Multiple phytohormone signals control the transcriptional response to soybean aphid infestation in susceptible and resistant soybean plants. *Mol. Plant-Microbe Interactions* 2013, *26*, 116-129.

135. Howe, G.A.; Jander, G. Plant immunity to insect herbivores. *Annual Review of Plant Biology* 2008, *59*, 41-66.

136. Lee, S.; Cassone, B.J.; Wijeratne, A.; Jun, T.-H.; Michel, A.P.; Mian, M.R. Transcriptomic dynamics in soybean near-isogenic lines differing in alleles for an aphid resistance gene, following infestation by soybean aphid biotype 2. *BMC Genomics* 2017, *18*, 472.

137. Wrather, J.A.; Koenning, S.R. Estimates of disease effects on soybean yields in the United States 2003 to 2005. *Journal of Nematology* 2006, *38*, 173.

138. Agrios, G.J.B., Ma. USA Plant pathology 5th Edition: Elsevier Academic Press. 2005, 79-103.

139. Williamson, V.M.; Gleason, C.A. Plant–nematode interactions. *Current Opinion in Plant Biology* 2003, *6*, 327-333.

140. Koenning, S.R.; Wrather, J.A. Suppression of soybean yield potential in the continental United States by plant diseases from 2006 to 2009. *Plant Health Progress* 2010, *10*.

141. Niblack, T.; Lambert, K.; Tylka, G. A model plant pathogen from the kingdom animalia: Heterodera glycines, the soybean cyst nematode. *Annu. Rev. Phytopathol.* 2006, *44*, 283-303.

142. Allen, T.W.; Bradley, C.A.; Sisson, A.J.; Byamukama, E.; Chilvers, M.I.; Coker, C.M.; Collins, A.A.; Damicone, J.P.; Dorrance, A.E.; Dufault, N.S. Soybean yield loss estimates due to diseases in the United States and Ontario, Canada, from 2010 to 2014. *Plant Health Progress* 2017, *18*, 19-27.

143. Mitchum, M.G. Soybean resistance to the soybean cyst nematode Heterodera glycines: an update. *Phytopathology* 2016, *106*, 1444-1450.

144. Hori, S. Sick soil of soybean caused by nematode. *Journal Plant Protection* 1916, *2*, 927-930.

145. Yokoo, T. Host plants of Heterodera schachtii Schmidt and some instructions. *Korea Agricultural Experiment Station Bulletin* 1936, *8*, 47-174.

146. Nakata, K.; Asuyana, H. Survey of the principal diseases of crops in Manchuria. *Bureau Industry Report* 1938, *32*.

147. Winstead, N.; Skotland, C.; Sasser, J. Soybean cyst nematode in North Carolina. *Plant Disease Reporter* 1955, *39*, 9-11.

148. Davis, E.L.; Tylka, G.L. Soybean cyst nematode disease. *The Plant Health Instructor* 2000.

149. Riggs, R. Worldwide distribution of soybean-cyst nematode and its economic importance. *Journal of Nematology* 1977, *9*, 34.

150. Acharya, K.; Tande, C.; Byamukama, E.J.P.H.P. Assessment of Commercial Soybean Cultivars for Resistance Against Prevalent *Heterodera glycines* Populations of South Dakota. 2017, *18*, 156-161.
151. Tylka, G.L.; Marett, C.C.Distribution of the soybean cyst nematode, Heterodera glycines, in the United States and Canada: 1954 to 2014. *Plant Health Progress* 2014, *15*, 85.
152. Niblack, T. Soybean cyst nematode management reconsidered. *Plant disease* 2005, *89*, 1020-1026.
153. Tefft, P.M.; Bone, L.W. Plant-induced hatching of eggs of the soybean cyst nematode Heterodera glycines. *Journal of Nematology* 1985, *17*, 275.
154. TEFFT, P.M.; RENDE, J.F.; BONE, L.W. In *Factors Influencing Egg Hatching of the Soybean Cyst Nematode*, *Proceedings of the Helminthological Society of Washington*, 1982; 1982; pp. 258-265.
155. Rasmann, S.; Ali, J.G.; Helder, J.; van Der Putten, W.H. Ecology and evolution of soil nematode chemotaxis. *Journal of Chemical Ecology* 2012, *38*, 615-628.
156. Masamune, T.; Anetai, M.; Takasugi, M.; Katsui, N. Isolation of a natural hatching stimulus, glycinoeclepin A, for the soybean cyst nematode. *Nature* 1982, *297*, 495.
157. Byrne, J.; Maher, N.; Jones, P. Comparative responses of Globodera rostochiensis and G. pallida to hatching chemicals. *Journal of Nematology* 2001, *33*, 195.
158. Schenk, H.; Driessen, R.A.; de Gelder, R.; Goubitz, K.; Nieboer, H.; Brüggemann-Rotgans, I.E.; Diepenhorst, P. Elucidation of the structure of solanoeclepin A, a natural hatching factor of potato and tomato cyst nematodes, by single-crystal X-ray diffraction. *Croatica Chemica Acta* 1999, *72*, 593-606.
159. Duceppe, M.-O.; Lafond-Lapalme, J.; Palomares-Rius, J.E.; Sabeh, M.; Blok, V.; Moffett, P.; Mimee, B. Analysis of survival and hatching transcriptomes from potato cyst nematodes, *Globodera rostochiensis* and *G. pallida*. *Scientific Reports* 2017, *7*, 3882.
160. Lauritis, J.; Rebois, R.; Graney, L. Development of *Heterodera glycines* Ichinohe on soybean, *Glycine max* (L.) Merr., under gnotobiotic conditions. *Journal Nematology* 1983, *15*, 272.
161. Hershman, D. Soybean Cyst Nematode: Soybean Thief and Public Enemy Number One. *Cooperative Extension Service* 1997.
162. Davis, E.L.; Hussey, R.S.; Baum, T.J. Getting to the roots of parasitism by nematodes. *Trends in Parasitology* 2004, *20*, 134-141.
163. Johnson, A.; Kim, K.; Riggs, R.; Scott, H. Location of Heterodera glycines-induced syncytia in soybean as affected by soil water regimes. *Journal of Nematology* 1993, *25*, 422.
164. Riggs, R.D.; Wrather, J.A. *Biology and management of the soybean cyst nematode*; 1992; p.^pp.
165. Colgrove, A.; Niblack, T. The effect of resistant soybean on male and female development and adult sex ratios of Heterodera glycines. *Journal of nematology* 2005, *37*, 161.
166. Jaffe, H.; Huettel, R.N.; Demilo, A.B.; Hayes, D.K.; Rebois, R.V. Isolation and identification of a compound from soybean cyst nematode, Heterodera glycines, with sex pheromone activity. *Journal of Chemical Ecology* 1989, *15*, 2031-2043.

167. Triantaphyllou, A.; Hirschmann, H. Oogenesis and Mode of Reproduction in the Soybean Cyst Nematode, *Heterodera Glycines1*. *Nematologica* 1962, *7*, 235-241.

168. Sipes, B.; Schmitt, D.; Barker, K. Fertility of three parasitic biotypes of *Heterodera glycines*. *PHYTOPATHOLOGY-NEW YORK BALTIMORE THEN ST PAUL-*1992, *82*, 999-999.

169. Koenning, S. Resistance of soybean cultivars to field populations of Heterodera glycines in North Carolina. *Plant Dis.* 2004, *88*, 942-950.

170. Mitchum, M.G.; Hussey, R.S.; Baum, T.J.; Wang, X.; Elling, A.A.; Wubben, M.; Davis, E.L. Nematode effector proteins: an emerging paradigm of parasitism. *New Phytologist* 2013, *199*, 879-894.

171. Gheysen, G.; Mitchum, M.G. How nematodes manipulate plant development pathways for infection. *Current Opinion in Plant Biology* 2011, *14*, 415-421.

172. Hofmann, J.; El Ashry, A.E.N.; Anwar, S.; Erban, A.; Kopka, J.; Grundler, F. Metabolic profiling reveals local and systemic responses of host plants to nematode parasitism. *The Plant Journal* 2010, *62*, 1058-1071.

173. Noon, J.B.; Hewezi, T.; Maier, T.R.; Simmons, C.; Wei, J.-Z.; Wu, G.; Llaca, V.; Deschamps, S.; Davis, E.L.; Mitchum, M.G. Eighteen new candidate effectors of the phytonematode Heterodera glycines produced specifically in the secretory esophageal gland cells during parasitism. *Phytopathology* 2015, *105*, 1362-1372.

174. Rehman, S.; Postma, W.; Tytgat, T.; Prins, P.; Qin, L.; Overmars, H.; Vossen, J.; Spiridon, L.-N.; Petrescu, A.-J.; Goverse, A. A secreted SPRY domain-containing protein (SPRYSEC) from the plant-parasitic nematode Globodera rostochiensis interacts with a CC-NB-LRR protein from a susceptible tomato. *Molecular Plant-Microbe Interactions* 2009, *22*, 330-340.

175. Hamamouch, N.; Li, C.; Hewezi, T.; Baum, T.J.; Mitchum, M.G.; Hussey, R.S.; Vodkin, L.O.; Davis, E.L. The interaction of the novel 30C02 cyst nematode effector protein with a plant β-1, 3-endoglucanase may suppress host defence to promote parasitism. *Journal of Experimental Botany* 2012, *63*, 3683-3695.

176. Hewezi, T.; Howe, P.J.; Maier, T.R.; Hussey, R.S.; Mitchum, M.G.; Davis, E.L.; Baum, T.J. *Arabidopsis* spermidine synthase is targeted by an effector protein of the cyst nematode *Heterodera schachtii*. *Plant physiology* 2010, *152*, 968-984.

177. Patel, N.; Hamamouch, N.; Li, C.; Hewezi, T.; Hussey, R.S.; Baum, T.J.; Mitchum, M.G.; Davis, E.L. A nematode effector protein similar to annexins in host plants. *Journal of Experimental Botany* 2009, *61*, 235-248.

178. Jaouannet, M.; Magliano, M.; Arguel, M.J.; Gourgues, M.; Evangelisti, E.; Abad, P.; Rosso, M.-N. The root-knot nematode calreticulin Mi-CRT is a key effector in plant defense suppression. *Molecular Plant-Microbe Interactions* 2013, *26*, 97-105.

179. Lozano-Torres, J.L.; Wilbers, R.H.; Gawronski, P.; Boshoven, J.C.; Finkers-Tomczak, A.; Cordewener, J.H.; America, A.H.; Overmars, H.A.; Van't Klooster, J.W.; Baranowski, L. Dual disease resistance mediated by the immune receptor Cf-2 in tomato requires a common virulence target of a fungus and a nematode. *Proceedings of the National Academy of Sciences* 2012, *109*, 10119-10124.

180. Sacco, M.A.; Koropacka, K.; Grenier, E.; Jaubert, M.J.; Blanchard, A.; Goverse, A.; Smant, G.; Moffett, P. The cyst nematode SPRYSEC protein RBP-1 elicits

Gpa2-and RanGAP2-dependent plant cell death. *PLoS Pathology* 2009, *5*, e1000564.

181.	Moffett, P.; Ali, S.; Magne, M.; Chen, S.; Obradovic, N.; Jamshaid, L.; Wang, X.; Bélair, G. Analysis of Globodera rostochiensis effectors reveals conserved functions of SPRYSEC proteins in suppressing and eliciting plant immune responses. *Frontiers in Plant Science* 2015, *6*, 623.

182.	Gao, B.; Allen, R.; Maier, T.; Davis, E.L.; Baum, T.J.; Hussey, R.S. The parasitome of the phytonematode *Heterodera glycines*. *Molecular Plant-Microbe Interactions* 2003, *16*, 720-726.

183.	Hewezi, T.; Howe, P.; Maier, T.R.; Hussey, R.S.; Mitchum, M.G.; Davis, E.L.; Baum, T.J. Cellulose binding protein from the parasitic nematode Heterodera schachtii interacts with *Arabidopsis* pectin methylesterase: cooperative cell wall modification during parasitism. *The Plant Cell* 2008, *20*, 3080-3093.

184.	Pogorelko, G.; Juvale, P.S.; Rutter, W.B.; Hewezi, T.; Hussey, R.; Davis, E.L.; Mitchum, M.G.; Baum, T.J. A cyst nematode effector binds to diverse plant proteins, increases nematode susceptibility and affects root morphology. *Molecular plant pathology* 2016, *17*, 832-844.

185.	Noon, J.B.; Qi, M.; Sill, D.N.; Muppirala, U.; Eves-van den Akker, S.; Maier, T.R.; Dobbs, D.; Mitchum, M.G.; Hewezi, T.; Baum, T.J. A Plasmodium-like virulence effector of the soybean cyst nematode suppresses plant innate immunity. *New Phytologist* 2016, *212*, 444-460.

186.	Bekal, S.; Domier, L.L.; Gonfa, B.; Lakhssassi, N.; Meksem, K.; Lambert, K.N. A SNARE-like protein and biotin are implicated in soybean cyst nematode virulence. *PLoS One* 2015, *10*, e0145601.

187.	Castagnone-Sereno, P.; Semblat, J.-P.; Castagnone, C. Modular architecture and evolution of the map-1 gene family in the root-knot nematode Meloidogyne incognita. *Molecular Genetics Genomics* 2009, *282*, 547.

188.	Gleason, C.A.; Liu, Q.L.; Williamson, V.M. Silencing a candidate nematode effector gene corresponding to the tomato resistance gene Mi-1 leads to acquisition of virulence. *Molecular Plant-Microbe Interactions* 2008, *21*, 576-585.

189.	Lee, C.; Chronis, D.; Kenning, C.; Peret, B.; Hewezi, T.; Davis, E.L.; Baum, T.J.; Hussey, R.; Bennett, M.; Mitchum, M.G. The novel cyst nematode effector protein 19C07 interacts with the Arabidopsis auxin influx transporter LAX3 to control feeding site development. *Plant Physiology* 2011, *155*, 866-880.

190.	Maier, T.R.; Hewezi, T.; Peng, J.; Baum, T.J. Isolation of whole esophageal gland cells from plant-parasitic nematodes for transcriptome analyses and effector identification. *Molecular Plant-Microbe Interactions* 2013, *26*, 31-35.

191.	Hewezi, T.; Juvale, P.S.; Piya, S.; Maier, T.R.; Rambani, A.; Rice, J.H.; Mitchum, M.G.; Davis, E.L.; Hussey, R.S.; Baum, T.J. The cyst nematode effector protein 10A07 targets and recruits host posttranslational machinery to mediate its nuclear trafficking and to promote parasitism in *Arabidopsis*. *The Plant Cell* 2015, tpc. 114.135327.

192.	Verma, A.; Lee, C.; Morriss, S.; Odu, F.; Kenning, C.; Rizzo, N.; Spollen, W.G.; Lin, M.; McRae, A.G.; Givan, S.A. The novel cyst nematode effector protein

30D08 targets host nuclear functions to alter gene expression in feeding sites. *New Phytologist* 2018, *219*, 697-713.

193. Mahalingam, R.; Skorupska, H. Cytological expression of early response to infection by *Heterodera glycines* Ichinohe in resistant PI 437654 soybean. *Genome* 1996, *39*, 986-998.

194. Sobczak, M.; Avrova, A.; Jupowicz, J.; Phillips, M.S.; Ernst, K.; Kumar, A. Characterization of susceptibility and resistance responses to potato cyst nematode (*Globodera* spp.) infection of tomato lines in the absence and presence of the broad-spectrum nematode resistance *Hero* gene. *Molecular Plant-Microbe Interactions* 2005, *18*, 158-168.

195. Brucker, E.; Carlson, S.; Wright, E.; Niblack, T.; Diers, B. Rhg1 alleles from soybean PI 437654 and PI 88788 respond differentially to isolates of Heterodera glycines in the greenhouse. *Theoretical and Applied Genetics* 2005, *111*, 44-49.

196. Concibido, V.C.; Diers, B.W.; Arelli, P.R. A decade of QTL mapping for cyst nematode resistance in soybean. *Crop Science* 2004, *44*, 1121-1131.

197. Caldwell, B.E.; Brim, C.; Ross, J. Inheritance of resistance of soybeans to the cyst nematode, Heterodera glycines. *Agronomy Journal* 1960, *52*.

198. Melito, S.; Heuberger, A.L.; Cook, D.; Diers, B.W.; MacGuidwin, A.E.; Bent, A.F. A nematode demographics assay in transgenic roots reveals no significant impacts of the Rhg1 locus LRR-Kinase on soybean cyst nematode resistance. *BMC Plant Biology* 2010, *10*, 104.

199. Lightfoot, D.; Meksem, K., Isolated polynucleotides and polypeptides relating to loci underlying resistance to soybean cyst nematode and soybean sudden death syndrome and methods employing same. In Google Patents: 2001.

200. Ruben, E.; Jamai, A.; Afzal, J.; Njiti, V.; Triwitayakorn, K.; Iqbal, M.; Yaegashi, S.; Bashir, R.; Kazi, S.; Arelli, P. Genomic analysis of the rhg1 locus: candidate genes that underlie soybean resistance to the cyst nematode. *Molecular Genetics Genomics* 2006, *276*, 503-516.

201. Hyten, D.L.; Choi, I.-Y.; Song, Q.; Shoemaker, R.C.; Nelson, R.L.; Costa, J.M.; Specht, J.E.; Cregan, P.B. Highly variable patterns of linkage disequilibrium in multiple soybean populations. *Genetics* 2007, *175*, 1937-1944.

202. Concibido, V.; Denny, R.; Boutin, S.; Hautea, R.; Orf, J.; Young, N. DNA marker analysis of loci underlying resistance to soybean cyst nematode (*Heterodera glycines* Ichinohe). *Crop Science* 1994, *34*, 240-246.

203. Colgrove, A.; Niblack, T. Correlation of female indices from virulence assays on inbred lines and field populations of *Heterodera glycines*. *Journal of Nematology* 2008, *40*, 39.

204. Meksem, K.; Pantazopoulos, P.; Njiti, V.; Hyten, L.; Arelli, P.; Lightfoot, D. 'Forrest'resistance to the soybean cyst nematode is bigenic: saturation mapping of the Rhg1and Rhg4 loci. *Theoretical and Applied Genetics* 2001, *103*, 710-717.

205. Hauge, B.M.; Wang, M.L.; Parsons, J.D.; Parnell, L.D., Nucleic acid molecules and other molecules associated with soybean cyst nematode resistance. In Google Patents: 2006.

206. HAUGE, B. Nucleic acid molecules and other molecules associated with soybean cyst nematode resistance. *US Patent Application Publication Number 20030005491* 2001.

207. Ronald, P.C.; Albano, B.; Tabien, R.; Abenes, L.; Wu, K.-s.; McCouch, S.; Tanksley, S.D. Genetic and physical analysis of the rice bacterial blight disease resistance locus, Xa21. *Molecular and General Genetics* 1992, *236*, 113-120.

208. Liu, X.; Liu, S.; Jamai, A.; Bendahmane, A.; Lightfoot, D.A.; Mitchum, M.G.; Meksem, K. Soybean cyst nematode resistance in soybean is independent of the Rhg4 locus LRR-RLK gene. *Functional & Integrative Genomics* 2011, *11*, 539-549.

209. Cooper, J.L.; Till, B.J.; Laport, R.G.; Darlow, M.C.; Kleffner, J.M.; Jamai, A.; El-Mellouki, T.; Liu, S.; Ritchie, R.; Nielsen, N. TILLING to detect induced mutations in soybean. *BMC Plant Biology* 2008, *8*, 9.

210. Schmutz, J.; Cannon, S.B.; Schlueter, J.; Ma, J.; Mitros, T.; Nelson, W.; Hyten, D.L.; Song, Q.; Thelen, J.J.; Cheng, J. Genome sequence of the palaeopolyploid soybean. *Nature* 2010, *463*, 178-183.

211. Kim, M.; Hyten, D.L.; Bent, A.F.; Diers, B.W. Fine mapping of the SCN resistance locus rhg1-b from PI 88788. *The Plant Genome* 2010, *3*, 81-89.

212. Matsye, P.D.; Kumar, R.; Hosseini, P.; Jones, C.M.; Tremblay, A.; Alkharouf, N.W.; Matthews, B.F.; Klink, V.P. Mapping cell fate decisions that occur during soybean defense responses. *Plant Molecular Biology* 2011, *77*, 513.

213. Matsye, P.D.; Lawrence, G.W.; Youssef, R.M.; Kim, K.-H.; Lawrence, K.S.; Matthews, B.F.; Klink, V.P. The expression of a naturally occurring, truncated allele of an α-SNAP gene suppresses plant parasitic nematode infection. *Plant Molecular Biology* 2012, *80*, 131-155.

214. Cook, D.E.; Lee, T.G.; Guo, X.; Melito, S.; Wang, K.; Bayless, A.M. Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. *Science* 2012, *338*.

215. Bayless, A.M.; Smith, J.M.; Song, J.; McMinn, P.H.; Teillet, A.; August, B.K.; Bent, A.F. Disease resistance through impairment of α-SNAP–NSF interaction and vesicular trafficking by soybean Rhg1. *Proceedings of the National Academy of Sciences* 2016, *113*, E7375-E7382.

216. Kavitha, K.; Mathiyazhagan, S.; Sendhilvel, V.; Nakkeeran, S.; Chandrasekar, G.; Dilantha Fernando, W. Broad spectrum action of phenazine against active and dormant structures of fungal pathogens and root knot nematode. *Archives of Phytopathology and Plant Protection* 2005, *38*, 69-76.

217. Cook, S.D.; Nichols, D.S.; Smith, J.; Chourey, P.S.; McAdam, E.L.; Quittenden, L.J.; Ross, J.J. Auxin biosynthesis: Are the indole-3-acetic acid and phenylacetic acid biosynthesis pathways mirror images? *Plant physiology* 2016, pp. 00454.02016.

218. Shi, Z.; Liu, S.; Noe, J.; Arelli, P.; Meksem, K.; Li, Z. SNP identification and marker assay development for high-throughput selection of soybean cyst nematode resistance. *BMC Genomics* 2015, *16*, 314.

219. Cook, D.; Bayless, A.; Wang, K.; Guo, X.; Song, Q.; Jiang, J.; Bent, A. Distinct copy number, coding sequence and locus methylation patterns underlie Rhg1-mediated soybean resistance to soybean cyst nematode. *Plant Physiology* 2014, pp. 114.235952.

220. Lee, T.G.; Kumar, I.; Diers, B.W.; Hudson, M.E. Evolution and selection of Rhg1, a copy-number variant nematode-resistance locus. *Molecular Ecology* 2015, *24*, 1774-1791.

221. Kandoth, P.K.; Liu, S.; Prenger, E.; Ludwig, A.; Lakhssassi, N.; Heinz, R.; Zhou, Z.; Howland, A.; Gunther, J.; Eidson, S. Systematic mutagenesis of serine hydroxymethyltransferase reveals essential role in nematode resistance. *Plant Physiology* 2017, pp. 00553.02017.

222. Wu, X.-Y.; Zhou, G.-C.; Chen, Y.-X.; Wu, P.; Liu, L.-W.; Ma, F.-F.; Wu, M.; Liu, C.-C.; Zeng, Y.-J.; Chu, A.E. Soybean cyst nematode resistance emerged via artificial selection of duplicated serine hydroxymethyltransferase genes. *Frontiers in Plant Science* 2016, *7*, 998.

223. Liu, S.; Kandoth, P.K.; Lakhssassi, N.; Kang, J.; Colantonio, V.; Heinz, R.; Yeckel, G.; Zhou, Z.; Bekal, S.; Dapprich, J. The soybean GmSNAP18 gene underlies two types of resistance to soybean cyst nematode. *Nature Communications* 2017, *8*, 14822.

224. Lakhssassi, N.; Liu, S.; Bekal, S.; Zhou, Z.; Colantonio, V.; Lambert, K.; Barakat, A.; Meksem, K. Characterization of the soluble NSF attachment protein gene family identifies two members involved in additive resistance to a plant pathogen. *Scientific Reports* 2017, *7*, 45226.

225. Matthews, B.F.; Beard, H.; MacDonald, M.H.; Kabir, S.; Youssef, R.M.; Hosseini, P.; Brewer, E. Engineered resistance and hypersusceptibility through functional metabolic studies of 100 genes in soybean to its major pathogen, the soybean cyst nematode. *Planta* 2013, *237*, 1337-1357.

226. Ungermann, C.; Langosch, D. Functions of SNAREs in intracellular membrane fusion and lipid bilayer mixing. *Journal of Cell Science* 2005, *118*, 3819-3828.

227. Liu, S.; Kandoth, P.K.; Warren, S.D.; Yeckel, G.; Heinz, R.; Alden, J. A soybean cyst nematode resistance gene points to a new mechanism of plant resistance to pathogens. *Nature* 2012, *492*.

228. Ros, R.; Muñoz-Bertomeu, J.; Krueger, S. Serine in plants: biosynthesis, metabolism, and functions. *Trends in Plant Science* 2014, *19*, 564-569.

229. Winkler, F.; Kriebel, M.; Clever, M.; Gröning, S.; Großhans, J. Essential function of the serine hydroxymethyl transferase (SHMT) gene during rapid syncytial cell cycles in Drosophila. *G3: Genes, Genomes, Genetics* 2017, g3. 117.043133.

230. Arelli, P.R.; Concibido, V.C.; Young, L.D. QTLs associated with resistance in soybean PI567516C to synthetic nematode population infecting cv. Hartwig. *Journal of Crop Science Biotechnology* 2010, *13*, 163-167.

231. Vuong, T.D.; Sleper, D.A.; Shannon, J.G.; Nguyen, H.T. Novel quantitative trait loci for broad-based resistance to soybean cyst nematode (Heterodera glycines Ichinohe) in soybean PI 567516C. *Theoretical and Applied Genetics* 2010, *121*.

232. Kadam, S.; Vuong, T.D.; Qiu, D.; Meinhardt, C.G.; Song, L.; Deshmukh, R.; Patil, G.; Wan, J.; Valliyodan, B.; Scaboo, A.M. Genomic-assisted phylogenetic analysis and marker development for next generation soybean cyst nematode resistance breeding. *Plant Science* 2016, *242*, 342-350.

233. Glover, K.; Wang, D.; Arelli, P.; Carlson, S.; Cianzio, S.; Diers, B. Near isogenic lines confirm a soybean cyst nematode resistance gene from PI 88788 on linkage group J. *Crop Science* 2004, *44*, 936-941.

234. Kazi, S.; Shultz, J.; Afzal, J.; Hashmi, R.; Jasim, M.; Bond, J.; Arelli, P.R.; Lightfoot, D.A. Iso-lines and inbred-lines confirmed loci that underlie resistance from cultivar 'Hartwig'to three soybean cyst nematode populations. *Theoretical and Applied Genetics* 2010, *120*, 633-644.

235. Wu, X.; Blake, S.; Sleper, D.A.; Shannon, J.G.; Cregan, P.; Nguyen, H.T. QTL, additive and epistatic effects for SCN resistance in PI 437654. *Theor. Appl. Genet.* 2009, *118*.

236. Guo, B.; Sleper, D.; Arelli, P.; Shannon, J.; Nguyen, H. Identification of QTLs associated with resistance to soybean cyst nematode races 2, 3 and 5 in soybean PI 90763. *Theoretical and Applied Genetics* 2005, *111*, 965-971.

237. Kim, M.; Diers, B.W. Fine mapping of the SCN resistance QTL cqSCN-006 and cqSCN-007 from Glycine soja PI 468916. *Crop Science* 2013, *53*, 775-785.

238. Yu, N.; Diers, B.W.J.E. Fine mapping of the SCN resistance QTL cqSCN-006 and cqSCN-007 from *Glycine soja* PI 468916. *Euphytica* 2017, *213*, 54.

239. Webb, D.; Baltazar, B.; Rao-Arelli, A.; Schupp, J.; Clayton, K.; Keim, P.; Beavis, W. Genetic mapping of soybean cyst nematode race-3 resistance loci in the soybean PI 437.654. *Theoretical and Applied Genetics* 1995, *91*, 574-581.

240. Concibido, V.; Denny, R.; Lange, D.; Orf, J.; Young, N. RFLP mapping and marker-assisted selection of soybean cyst nematode resistance in PI 209332. *Crop Science* 1996, *36*, 1643-1650.

241. Concibido, V.C.; Lange, D.A.; Denny, R.L.; Orf, J.H.; Young, N.D. Genome mapping of soybean cyst nematode resistance genes in 'Peking', PI 90763, and PI 88788 using DNA markers. *Crop Science* 1997, *37*, 258-264.

242. Yue, P.; Sleper, D.A.; Arelli, P.R. Mapping resistance to multiple races of Heterodera glycines in soybean PI 89772. *Crop Science* 2001, *41*, 1589-1595.

243. Guo, B.; Sleper, D.; Nguyen, H.; Arelli, P.; Shannon, J. Quantitative trait loci underlying resistance to three soybean cyst nematode populations in soybean PI 404198A. *Crop Science* 2006, *46*, 224-233.

244. Li, Y.-h.; Shi, X.-h.; Li, H.-h.; Reif, J.C.; Wang, J.-j.; Liu, Z.-x.; He, S.; Yu, B.-s.; Qiu, L.-j. Dissecting the genetic basis of resistance to soybean cyst nematode combining linkage and association mapping. *The plant Genome* 2016, *9*.

245. Wen, Z.; Tan, R.; Yuan, J.; Bales, C.; Du, W.; Zhang, S.; Chilvers, M.I.; Schmidt, C.; Song, Q.; Cregan, P.B. Genome-wide association mapping of quantitative resistance to sudden death syndrome in soybean. *BMC Genomics* 2014, *15*, 809.

246. Zhang, H.; Li, C.; Davis, E.L.; Wang, J.; Griffin, J.D.; Kofsky, J.; Song, B.-H. Genome-Wide Association Study of Resistance to Soybean Cyst Nematode (*Heterodera glycines*) HG Type 2.5. 7 in Wild Soybean (*Glycine soja*). *Frontiers in Plant Science* 2016, *7*.

247. Han, Y.; Zhao, X.; Cao, G.; Wang, Y.; Li, Y.; Liu, D.; Teng, W.; Zhang, Z.; Li, D.; Qiu, L. Genetic characteristics of soybean resistance to HG type 0 and HG type 1.2. 3.5. 7 of the cyst nematode analyzed by genome-wide association mapping. *BMC Genomics* 2015, *16*, 598.

248. Srour, A.; Afzal, A.J.; Blahut-Beatty, L.; Hemmati, N.; Simmonds, D.H.; Li, W.; Liu, M.; Town, C.D.; Sharma, H.; Arelli, P. The receptor like kinase at *Rhg1-a/Rfs2* caused pleiotropic resistance to sudden death syndrome and soybean cyst

nematode as a transgene by altering signaling responses. *BMC genomics* 2012, *13*, 368.

249. Yue, P.; Arelli, P.; Sleper, D. Molecular characterization of resistance to Heterodera glycines in soybean PI 438489B. *Theoretical and Applied Genetics* 2001, *102*, 921-928.

250. Maffei, M.E.; Mithöfer, A.; Boland, W. Before gene expression: early events in plant–insect interaction. *Trends in Plant Science* 2007, *12*, 310-316.

251. Bostock, R.M. Signal crosstalk and induced resistance: straddling the line between cost and benefit. *Annual Review of Phytopathology* 2005, *43*, 545-580.

252. Leon-Reyes, A.; Spoel, S.H.; De Lange, E.S.; Abe, H.; Kobayashi, M.; Tsuda, S.; Millenaar, F.F.; Welschen, R.A.; Ritsema, T.; Pieterse, C.M. Ethylene modulates the role of NONEXPRESSOR OF PATHOGENESIS-RELATED GENES1 in cross talk between salicylate and jasmonate signaling. *Plant Physiology* 2009, *149*, 1797-1809.

253. Baldwin, I.T.; Schmelz, E.A.; Ohnmeiss, T.E. Wound-induced changes in root and shoot jasmonic acid pools correlate with induced nicotine synthesis inNicotiana sylvestris spegazzini and comes. *Journal of Chemical Ecology* 1994, *20*, 2139-2157.

254. Wu, J.; Baldwin, I.T. New insights into plant responses to the attack from insect herbivores. *Annual Review of Genetics* 2010, *44*, 1-24.

255. Blackman, R.L.; Eastop, V.F. *Aphids on the world's crops. An identification and information guide*; John Wiley: 1984; p.^pp.

256. Dixon, A.F.G. *Aphid ecology an optimization approach*; Springer Science & Business Media: 2012; p.^pp.

257. Blackman, R.; Eastop, V. Aphids on the world's crops: an identification and information guide. 2000.

258. Elzinga, D.A.; Jander, G. The role of protein effectors in plant–aphid interactions. *Current Opinion in Plant Biology* 2013, *16*, 451-456.

259. Tjallingii, W.; Esch, T.H. Fine structure of aphid stylet routes in plant tissues in correlation with EPG signals. *Physiological Entomology* 1993, *18*, 317-328.

260. Jaouannet, M.; Rodriguez, P.A.; Thorpe, P.; Lenoir, C.J.; MacLeod, R.; Escudero-Martinez, C.; Bos, J.I. Plant immunity in plant–aphid interactions. *Frontiers in Plant Science* 2014, *5*.

261. Flor, H.H. Current status of the gene-for-gene concept. *Annual Review of Phytopathology* 1971, *9*, 275-296.

262. van der Hoorn, R.A.; Kamoun, S. From Guard to Decoy: a new model for perception of plant pathogen effectors. *Plant Cell* 2008, *20*, 2009-2017.

263. Collier, S.M.; Moffett, P. NB-LRRs work a "bait and switch" on pathogens. *Trends Plant Sci.* 2009, *14*, 521-529.

264. Boller, T.; Felix, G. A renaissance of elicitors: perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors. *Annual Review Plant Biology* 2009, *60*, 379-406.

265. WYSS, U.; ZUNKE, U. Observations on the behaviour of second stage juveniles of Hetero inside host roots. *Revue de Nematologie* 1986, *9*, 153-165.

266. Goellner, M.; Wang, X.; Davis, E.L. Endo-β-1, 4-glucanase expression in compatible plant–nematode interactions. *The Plant Cell* 2001, *13*, 2241-2255.

267. Vercauteren, I.; de Almeida Engler, J.; De Groodt, R.; Gheysen, G. An Arabidopsis thaliana pectin acetylesterase gene is upregulated in nematode feeding sites induced by root-knot and cyst nematodes. *Molecular Plant-Microbe Interactions* 2002, *15*, 404-407.

268. Koltai, H.; Dhandaydham, M.; Opperman, C.; Thomas, J.; Bird, D. Overlapping plant signal transduction pathways induced by a parasitic nematode and a rhizobial endosymbiont. *Molecular Plant-Microbe Interactions* 2001, *14*, 1168-1177.

269. Dangl, J.L.; Jones, J.D. Plant pathogens and integrated defence responses to infection. *Nature* 2001, *411*, 826-833.

270. Milligan, S.B.; Bodeau, J.; Yaghoobi, J.; Kaloshian, I.; Zabel, P.; Williamson, V.M. The root knot nematode resistance gene Mi from tomato is a member of the leucine zipper, nucleotide binding, leucine-rich repeat family of plant genes. *The Plant Cell* 1998, *10*, 1307-1319.

271. Rossi, M.; Goggin, F.L.; Milligan, S.B.; Kaloshian, I.; Ullman, D.E.; Williamson, V.M. The nematode resistance gene Mi of tomato confers resistance against the potato aphid. *Proceedings of the National Academy of Sciences* 1998, *95*, 9750-9754.

272. Dogimont, C.; Bendahmane, A.; Pitrat, M.; Burget-Bigeard, E.; Hagen, L.; Le Menn, A.; Pauquet, J.; Rousselle, P.; Caboche, M.; Chovelon, V., Transfer the resistance-promoting Vat allele, by transgenesis, to *Aphis gossypii*-sensitive melon varieties; aphicides. In Google Patents: 2009.

273. Klingler, J.; Creasy, R.; Gao, L.; Nair, R.M.; Calix, A.S.; Jacob, H.S.; Edwards, O.R.; Singh, K.B. Aphid resistance in Medicago truncatula involves antixenosis and phloem-specific, inducible antibiosis, and maps to a single locus flanked by NBS-LRR resistance gene analogs. *Plant Physiology* 2005, *137*, 1445-1455.

274. Megías, A.G.; Müller, C. Root herbivores and detritivores shape above-ground multitrophic assemblage through plant-mediated effects. *Journal of Animal Ecology* 2010, *79*, 923-931.

275. Wiens, J.J.; Lapoint, R.T.; Whiteman, N.K. Herbivory increases diversification across insect clades. *Nature communications* 2015, *6*, 8370.

276. Van Dam, N.M.; Raaijmakers, C.E.; Van Der Putten, W.H. Root herbivory reduces growth and survival of the shoot feeding specialist Pieris rapae on Brassica nigra. *Entomologia Experimentalis et Applicata* 2005, *115*, 161-170.

277. Kaplan, I.; Halitschke, R.; Kessler, A.; Rehill, B.J.; Sardanelli, S.; Denno, R.F. Physiological integration of roots and shoots in plant defense strategies links above-and belowground herbivory. *Ecology Letters* 2008, *11*, 841-851.

278. Clifton, E.H.; Tylka, G.L.; Gassmann, A.J.; Hodgson, E.W. Interactions of effects of host plant resistance and seed treatments on soybean aphid (*Aphis glycines* Matsumura) and soybean cyst nematode (*Heterodera glycines* Ichinohe). *Pest Management Science* 2018, *74*, 992-1000.

279. Hoysted, G.A.; Lilley, C.J.; Field, K.J.; Dickinson, M.; Hartley, S.E.; Urwin, P.E. A Plant-Feeding Nematode Indirectly Increases the Fitness of an Aphid. *Frontiers in Plant Science* 2017, *8*, 1897.

280. Hol, W.G.; De Boer, W.; Termorshuizen, A.J.; Meyer, K.M.; Schneider, J.H.; Van Der Putten, W.H.; Van Dam, N.M. *Heterodera schachtii* nematodes interfere

with aphid-plant relations on Brassica oleracea. *Journal of Chemical Ecology* 2013, *39*, 1193-1203.

281. Hong, S.; Donaldson, J.; Gratton, C. Soybean cyst nematode effects on soybean aphid preference and performance in the laboratory. *Environmental Entomology* 2010, *39*, 1561-1569.

282. McCarville, M.T.; Soh, D.H.; Tylka, G.L.; O'Neal, M.E. Aboveground feeding by soybean aphid, *Aphis glycines*, affects soybean cyst nematode, *Heterodera glycines*, reproduction belowground. *PloS one* 2014, *9*.

283. Van Dam, N.M.; Gossa, M.W.; Mathur, V.; Tytgat, T.O. Differences in hormonal signaling triggered by two root-feeding nematode species result in contrasting effects on aphid population growth. *Frontiers in Ecology and Evolution* 2018, *6*, 88.

284. Kafle, D.; Hänel, A.; Lortzing, T.; Steppuhn, A.; Wurst, S. Sequential above-and belowground herbivory modifies plant responses depending on herbivore identity. *BMC Ecology* 2017, *17*, 5.

285. Kutyniok, M.; Müller, C. Crosstalk between above-and belowground herbivores is mediated by minute metabolic responses of the host Arabidopsis thaliana. *Journal of Experimental Botany* 2012, *63*, 6199-6210.

286. Bezemer, T.; De Deyn, G.; Bossinga, T.; Van Dam, N.; Harvey, J.; Van der Putten, W. Soil community composition drives aboveground plant–herbivore–parasitoid interactions. *Ecology Letters* 2005, *8*, 652-661.

287. Wurst, S.; van der Putten, W.H. Root herbivore identity matters in plant-mediated interactions between root and shoot herbivores. *Basic Applied Ecology* 2007, *8*, 491-499.

288. Russin, J.; McGawley, E.; Boethel, D. Population development of Meloidogyne incognita on soybean defoliated by Pseudoplusia includens. *Journal of Nematology* 1993, *25*, 50.

289. Tiwari, S.; Youngman, R.; Lewis, E.; Eisenback, J. European corn borer (Lepidoptera: Crambidae) stalk tunneling on root-knot nematode (Tylenchida: Heteroderidae) fitness on corn. *Journal of Economic Entomology* 2009, *102*, 602-609.

290. McCarville, M.; O'Neal, M.; Tylka, G.; Kanobe, C.; MacIntosh, G. A nematode, fungus, and aphid interact via a shared host plant: implications for soybean management. *Entomologia Experimentalis et Applicata* 2012, *143*, 55-66.

291. Hol, W.G.; De Boer, W.; Termorshuizen, A.J.; Meyer, K.M.; Schneider, J.H.; Van Dam, N.M.; Van Veen, J.A.; Van Der Putten, W.H. Reduction of rare soil microbes modifies plant–herbivore interactions. *Ecology Letters* 2010, *13*, 292-301.

292. Vandegehuchte, M.L.; De La Peña, E.; Bonte, D. Interactions between root and shoot herbivores of Ammophila arenaria in the laboratory do not translate into correlated abundances in the field. *Oikos* 2010, *119*, 1011-1019.

293. Kuhlmann, F.; Müller, C. UV-B impact on aphid performance mediated by plant quality and plant changes induced by aphids. *Plant Biology* 2010, *12*, 676-684.

294. Huberty, A.F.; Denno, R.F. Plant water stress and its consequences for herbivorous insects: a new synthesis. *Ecology* 2004, *85*, 1383-1398.

295. Kaplan, I.; Sardanelli, S.; Denno, R.F. Field evidence for indirect interactions between foliar-feeding insect and root-feeding nematode communities on Nicotiana tabacum. *Ecological Entomology* 2009, *34*, 262-270.

296. Hoysted, G.A.; Bell, C.A.; Lilley, C.; Urwin, P.E. Aphid colonisation affects potato root exudate composition and the hatching of a soil borne pathogen. *Frontiers in Plant Science* 2018, *9*, 1278.

297. Wondafrash, M.; Van Dam, N.M.; Tytgat, T.O. Plant systemic induced responses mediate interactions between root parasitic nematodes and aboveground herbivorous insects. *Frontiers in Plant Science* 2013, *4*.

298. Johnson, S.N.; Clark, K.E.; Hartley, S.E.; Jones, T.H.; McKenzie, S.W.; Koricheva, J. Aboveground–belowground herbivore interactions: a meta-analysis. *Ecology* 2012, *93*, 2208-2215.

299. Erb, M.; Robert, C.A.; Hibbard, B.E.; Turlings, T.C. Sequence of arrival determines plant-mediated interactions between herbivores. *Journal of Ecology* 2011, *99*, 7-15.

300. Robert, C.A.; Erb, M.; Hibbard, B.E.; Wade French, B.; Zwahlen, C.; Turlings, T.C. A specialist root herbivore reduces plant resistance and uses an induced plant volatile to aggregate in a density-dependent manner. *Functional Ecology* 2012, *26*, 1429-1440.

301. Baluch, S.D.; Ohm, H.W.; Shukle, J.T.; Williams, C.E. Obviation of wheat resistance to the hessian Fly through systemic induced susceptibility. *Journal of Economic Entomology* 2012, *105*, 642-650.

302. Sauge, M.H.; Mus, F.; Lacroze, J.P.; Pascal, T.; Kervella, J.; Poëssel, J.L. Genotypic variation in induced resistance and induced susceptibility in the peach-Myzus persicae aphid system. *Oikos* 2006, *113*, 305-313.

303. Claridge, M.; Den Hollander, J. The biotype concept and its application to insect pests of agriculture. *Crop Protection* 1983, *2*, 85-95.

304. Kanobe, C.; McCarville, M.T.; O'Neal, M.E.; Tylka, G.L.; MacIntosh, G.C. Soybean aphid infestation induces changes in fatty acid metabolism in soybean. *PloS one* 2015, *10*, e0145660.

305. Varenhorst, A.; McCarville, M.; O'Neal, M. An Induced Susceptibility Response in Soybean Promotes Avirulent Aphis glycines (Hemiptera: Aphididae) Populations on Resistant Soybean. *Environmental Entomology* 2015, nvv051.

306. Price, P.W.; Denno, R.F.; Eubanks, M.D.; Finke, D.L.; Kaplan, I. *Insect ecology: behavior, populations and communities*; Cambridge University Press: 2011; p.^pp.

307. Varenhorst, A.J.; McCarville, M.T.; O'Neal, M.E. Determining the duration of Aphis glycines (Hemiptera: Aphididae) induced susceptibility effect in soybean. *Arthropod-Plant Interactions* 2015, *9*, 457-464.

308. Hong, S.; MacGuidwin, A.; Gratton, C. Soybean aphid and soybean cyst nematode interactions in the field and effects on soybean yield. *Journal of Economic Entomology* 2011, *104*, 1568-1574.

309. Heeren, J.; Steffey, K.; Tinsley, N.; Estes, R.; Niblack, T.; Gray, M. The interaction of soybean aphids and soybean cyst nematodes on selected resistant and susceptible soybean lines. *Journal of Applied Entomology* 2012, *136*, 646-655.

310. Kabouw, P.; Kos, M.; Kleine, S.; Vockenhuber, E.; Van Loon, J.; Van der Putten, W.; Van Dam, N.; Biere, A. Effects of soil organisms on aboveground multitrophic interactions are consistent between plant genotypes mediating the interaction. *Entomologia Experimentalis et Applicata* 2011, *139*, 197-206.

CHAPTER 2: EVOLUTIONARY DIVERGENCE OF TNL DISEASE-RESISTANCE PROTEINS IN SOYBEAN (*GLYCINE MAX*) AND COMMON BEAN (*PHASEOLUS VULGARIS*)

This chapter has been published in the Journal *Biochemical Genetics*:

Neupane, S.; Ma, Q.; Mathew, F.M.; Varenhorst, A.J.; Andersen, E.J.; Nepal, M.P.

Abstract

Disease-resistant genes (R genes) encode proteins that are involved in protecting plants from their pathogens and pests. Availability of complete genome sequences from soybean and common bean allowed us to perform a genome-wide identification and analysis of the Toll interleukin-1 receptor-like nucleotide-binding site leucine-rich repeat (TNL) proteins. Hidden Markov model (HMM) profiling of all protein sequences resulted in the identification of 117 and 77 regular TNL genes in soybean and common bean, respectively. We also identified TNL gene homologs with unique domains, and signal peptides as well as nuclear localization signals. The TNL genes in soybean formed 28 clusters located on 10 of the 20 chromosomes, with the majority found on chromosome 3, 6 and 16. Similarly, the TNL genes in common bean formed 14 clusters located on five of the 11 chromosomes, with the majority found on chromosome 10. Phylogenetic analyses of the TNL genes from Arabidopsis, soybean and common bean revealed less divergence within legumes relative to the divergence between legumes and Arabidopsis. Syntenic blocks were found between chromosomes Pv10 and Gm03, Pv07 and Gm10, as well as Pv01 and Gm14. The gene expression data revealed basal level expression and tissue

specificity, while analysis of available microRNA data showed 37 predicted microRNA

families involved in targeting the identified TNL genes in soybean and common bean.

Keywords: Comparative genomics; Gene duplication; Legume disease-resistant genes;

Purifying selection; R gene targeting MicroRNAs; Synteny

2.1. Introduction

Plant defense strategies have coevolved with their natural enemies such as pests

and pathogens (Jones and Dangl 2006). Interactions among plants, pathogens, and pests

have been recently discussed in various models including zig-zag and multicomponent

models (Andolfo and Ercolano 2015; Jones and Dangl 2006). These models largely

involve proteins encoded by clustered disease resistance (R) genes in plant genomes

(Hulbert et al. 2001). The R gene encoded proteins were classified in a previous study

into eight major groups based on an amino acid motif organization and localization in the

cell (Gururani et al. (2012). Among these groups, two major types of R gene proteins are

Toll interleukin-1 receptor-like Nucleotide-binding site Leucine-rich repeat (TNL)

proteins and Coiled Coil (CC)-NBS-LRR or CNL proteins. In a recent Angiosperm wide

study, Shao et al. (2016) have classified NBS-LRR genes into three classes (TNLs, CNLs

and R [resistance to powdery mildew8] NLs). The TNL genes encode proteins similar to

Drosophila melanogaster Toll and human interleukin-1 receptor's domain sequences at

the N-terminal and hence given the name TIR (DeYoung and Innes 2006), whereas CNL

genes encode a coiled-coil (CC) domain at the N-terminal (Meyers et al. 2003). The TIR

domain has mainly three conserved motifs: TIR1, TIR2, and TIR3 and one variable TIR4

motif (Meyers et al. 2002). Both groups are believed to have vital roles in the plant

defense system (Marone et al. 2013). The whole genome sequences of plant species at

diverse taxonomic levels allow us to study the diversity and evolution of the R-genes (Schatz et al. 2012). Increasing number of recent studies has used the whole genome sequences to study R genes in both monocots and dicots including legumes (Andersen et al. 2016; Christie et al. 2016; Luo et al. 2012; Meyers et al. 2003; Nepal and Benson 2015; Zhang et al. 2016b).

Legumes constitute the third largest group of Angiosperms after Orchidaceae and Asteraceae (http://www.theplantlist.org/1.1/browse/A/Leguminosae/). They contribute approximately 27% of world's major crops and supply one-third of dietary protein to humans along with fodder (Duc et al. 2015; Zhu et al. 2005). Legumes also play important role in biological nitrogen fixation through their symbiosis with rhizobia: such symbiosis is almost 60 million-year-old (MYA), and root nodulation is believed to be almost 58 MYA (Sprent 2007). The production of legumes is limited by both biotic and abiotic factors. The abiotic factors include water deficit, flooding, salinity, cold, heat, UV-B radiation, ozone etc. and biotic factors include diseases such as rusts, mildews, root rot diseases etc. (Rathi et al. 2016). The major biotic threats to the legume crops consist of bacteria, viruses and fungi and the crops employ NBS-LRR genes to confer resistance against them. Resistance to *Phytophthora sojae*, a major root-rot causing disease in legumes is conferred by *Rps* (*Rps1-k-2*, *Rps1-k-1*; Resistance to *P. sojae*), an NBS-LRR disease resistance gene in soybean (Gao and Bhattacharyya 2008; Gao et al. 2005). Bacterial blight disease, caused by *Pseudomonas syringae* in soybean with their pathogen avrA, avrB, avrC, avrD, avrE, avrF, and avrG effectors interacting against host's *Rpg* genes, the CNL type of R-genes (Chen et al. 2010; Milos et al. 2013). *Phakopsora pachyrhizi,* a basidiomycete that causes soybean rust (SBR), considered to

be the most destructive foliar disease in soybean, resistance for which is conferred by

Rpp genes (Goellner et al. 2010). Two dominant R- genes *Phg-1* and *Phg-2* were

identified to confer resistance to *Pseudocercospora griseola* (Namayanja et al. 2006;

Sartorato et al. 2000). *Uromyces appendiculatus*, fungus causing common bean rust

produces effectors that interact against 14 major dominant rust R-genes (*Ur-1* to *Ur-14*)

in common bean (Souza et al. 2013). Various resistance loci (*Rag*, resistance to *Aphis*

glycines) in soybean including *Rag1*, *Rag2*, *Rag3*, *Rag4*, *Rag5*, and *Rag6* have been

identified conferring resistance to *A. glycines* Matsumura (Hemiptera: Aphididae) (Bales

et al. 2013; Hartman et al. 2001; Hill et al. 2012; Jun et al. 2012; Kim et al. 2010a; Kim

et al. 2010b; Zhang et al. 2009; Zhang et al. 2010). Studies have shown that these *Rag*

genes encode NBS-LRR proteins (Kim et al. 2010a; Kim et al. 2010b). The recently

studied comparative transcriptomic analysis of common bean genotypes revealed that

TNL protein Phvul.010G054400 was highly expressed high in soybean cyst nematode

(SCN) resistant genotype of common bean (Jain et al. 2016).

Among the two major classes of NBS-LRR genes in plants, TNL genes are absent

in monocots (Li et al. 2015) although origin of both CNLs and TNLs dates back to

bryophytes, one of the oldest groups of land plants (Yue et al. 2012). Even though both

CNL and TNL genes are present in dicot species, the absence of TNLs in some eudicot

species such as *Aquilegia coerulea* (Collier et al. 2011), and *Beta vulgaris* (Tian et al.

2004) has been reported. Various studies on phylogenetic analyses have suggested that

TNL group expanded after the monocots and dicots diverged from each other and are

mostly involved in resistance to species-specific pathogens (Yang et al. 2008). In

Arabidopsis, TNLs require functionally enhanced disease susceptibility (EDS1) allele to

activate hypersensitive response (HR), whereas CNLs require a functional non-race specific disease resistance (NDR1) gene for the activation of disease resistance (Glazebrook 2001). Collier et al. (2011) have reported the absence of NRG1 (N Requirement Gene 1) genes, typical CNL genes absent in plant species lacking TNL genes. Interestingly, a TNL type protein N, activated against tobacco mosaic virus (TMV), partners with NRG1 (N Requirement Gene 1), which belongs to CNL type of protein (Peart et al. 2005). This implies the involvement of overlapping signaling pathways involving CNL and TNL genes (Tian et al. 2004). The diversity and distribution of the TNLs and CNLs vary from plant species to species: 161 CNLs and 54 TNLs in *Malus x domestica* (Arya et al. 2014); 119 CNLs and 64 TNLs in *Populus trichocarpa* (Kohler et al. 2008); 203 CNLs and 97 TNLs in *Vitis vinifera* (Yang et al. 2008); 118 CNLs and 18 TNLs in *Solanum lycopersicum* (Andolfo et al. 2014); 25 CNLs and 19 TNLs in *Cucumis sativus* (Yang et al. 2013); 65 CNLs and 37 TNLs in *Solanum tuberosum* (Lozano et al. 2012); 152 CNLs and 118 TNLs in *Medicago truncatula* (Yu et al. 2014); 128 CNLs and 34 TNLs in *Manihot esculenta* (Lozano et al. 2015); 64 CNLs and 57 TNLs in *Capsella rubella* (Zhang et al. 2016b); 167 CNLs and 112 TNLs in *Cajanus cajan* (Shao et al. 2014); 126 CNLs and 27 TNLs in *Gossypium raimondii* (Wei et al. 2013). However, *A. thaliana* has 94 TNLs and 55 CNLs (Meyers et al. 2003); *A. lyrata* has 103 TNLs and 21 CNLs (Guo et al. 2011); *Brassica rapa* has 52 TNLs and 28 CNLs (Mun et al. 2009); *Eucalyptus grandis* has 162 TNLs and 128 CNLs (Christie et al. 2016); *Thellungiella salsuginea* has 50 TNLs and 33 CNLs (Zhang et al. 2016b) suggesting TNLs are abundant than CNLs in some plant species.

Increased availability of the legume genome sequences (*Glycine max*, *Phaseolus vulgaris*, *Medicago truncatula*, *Arachis ipaensis*, *A. duranensis*, *Arachis hypogaea*, *Trifolium pratense*, *Vigna radiata*, *V. angularis*, *V. unguiculata*, *Lupinus angustifolius*, *Cicer arietinum*, *Cajanus cajan*, *Lotus japonicus*) provides opportunities for comparative genomic analyses, particularly in enhancing our understanding of R genes and ability to develop durable resistance in cultivars (https://legumeinfo.org/species). In this study, we analyzed the genomes of *G. max* (Schmutz et al. 2010) and *P. vulgaris* (Schmutz et al. 2014), with the genome sizes of approximately 1,100 and 588-637 million base pairs (Mbp), respectively (Arumuganathan and Earle 1991; Bennett and Smith 1976). The two legume species diverged almost 19 MYA at the time soybean genome got the last duplication event (Lavin et al. 2005; Schlueter et al. 2004). Analyses of *Rpg1b* (for Resistance to *Pseudomonas glycinea* 1b) showed a potential involvement of this locus in speciation event in the two species (Ashfield et al. 2012). The recombination rates twinned with loss and retention of the redundant regions have caused differences in the number of genes in the two species (Ashfield et al. 2012; Du et al. 2012). Previously, bioinformatics analyses of NBS-LRR genes in soybean and other legumes were conducted by many groups (Benson 2014; Kang et al. 2012; Nepal and Benson 2015; Shao et al. 2016; Shao et al. 2014; Zhang et al. 2016a; Zheng et al. 2016). It was beyond the scope of these large-scale studies to focus on specific group(s) of R genes, lacking the clear picture of protein domains, their function and evolutionary significance. Nonetheless, detailed evolutionary relationships, structural as well as functional divergence of the CNL R-genes in soybean and common bean were revealed by Nepal and Benson (2015) and Benson (2014). In this study, our objectives were to identify TNL

R-genes in soybean (*G. max*; 2n=40) and common bean (*P. vulgaris*; 2n=22) and assess

their structural and functional divergence. The results from this study shed light on

evolutionary relationships of the TNL genes with potential implication in crop

improvement.

2.2. Materials and Methods

   2.2.1. Hidden Markov Model (HMM) Search and TNL Gene Identification

   Genome-wide identification of TNL genes in soybean and common bean was

carried out using methods used in *A. thaliana*, *Setaria italica*, and *Hordeum vulgare*

(Andersen et al. 2016; Andersen and Nepal 2017; Meyers et al. 2003). *G. max* protein

sequences accessed through Ensembl Genomes database and *P. vulgaris* protein

sequences obtained from Phytozome.net were used to construct local protein database for

HMM profiling (Finn et al. 2015). *Arabidopsis* TNL protein sequences

(http://niblrrs.ucdavis.edu/data_protein.php) were used as reference or seed sequences to

search for TNL protein sequences in soybean and common bean. Reference sequences in

the stockholm file format was used for HMM profiling in the program HMMER version

3.1b2 (Finn et al. 2015) at a threshold expectation value of 0.05. NB-ARCs [for APAF1

(apoptotic protease activating factor-1), R (Resistance genes), and CED4 (*Caenorhabditis

elegans* death-4 protein)] were further identified using Pfam (Finn et al. 2013) database

integrated in InterProScan (Jones et al. 2014). Proteins with match in accordance to Pfam

with the TIR domain (PF01582), NB-ARC (NB) domain (PF00931), and LRR domains

with 'LxxLxxLxx' signatures were selected. Genes with NB-ARCs were aligned to create

reference for second HMM profile to scan the respective genomes with threshold

expectation value of 0.001. Further Interproscan database was searched using the

program Geneious [(Kearse et al. 2012) https://www.geneious.com/] to confirm the NBS

proteins with TIR domains. Multiple expectation maximization for motif elicitation

(MEME) (Bailey and Elkan 1994) analysis was performed to confirm the presence of P-

loop, Kinase-2, and GLPL motifs in NBS domain and TIR1, TIR2, TIR3, TIR4 motifs in

TIR domain. SignalP 4.1 (Petersen et al. 2011) was employed to analyze the presence of

signal peptides in the identified TNL genes. Subcellular localization of the putative TNL

genes were analyzed using TargetP 1.1 (Emanuelsson et al. 2007). NLStradamus (Ba et

al. 2009) was used to examine whether the TNL genes contain nuclear localization

signals (NLS).

2.2.2. Phylogenetic Analysis

The NBS domains in TNL proteins identified in *A. thaliana*, *G. max* and *P.*

vulgaris were aligned using ClustalW (Larkin et al. 2007) with default settings. MEGA

(version 7.0.14) (Kumar et al. 2016) was used to perform maximum likelihood (ML)

phylogenetic analysis with 100 bootstrap replicates. The trees were rooted with

Streptomyces coelicolor accession P25941 as an out-group as previously used in

Arabidopsis study (Meyers et al. 2003). The protein model selection for phylogenetic

analysis was carried out in MEGA, resulting in the selection of the JTT+G+I (Jones–

Taylor–Thornton with gamma distribution and invariant sites). The sequences with

bootstrap confidence ≥70% (0.7) are assigned to be orthologous sequences (Peele et al.

2014). In order to find some orthologs for manually curated TNL genes, one additional

tree was constructed using the same method but including manually curated TNL genes

obtained from PRGdb (http://prgdb.crg.eu/wiki/Main_Page) with TNL genes of soybean

and common bean.

2.2.3. Chromosomal Locations, Clustering and Syntenic Analysis

Entire chromosome sequences of *G. max* and *P. vulgaris* obtained from Ensembl Genomes and Phytozome.net were uploaded into the program Geneious. The TNL gene locations and clustering were visualized by matching the locations of genes in their respective chromosomes. The clustering of TNL genes was quantified on the basis of nucleotide intervals between the genes following Jupe et al. (2012), which used two criteria: 1) distance between two TNL genes is less than 200kb, and 2) presence of no more than eight annotated non-TNL sequences between two consecutive TNL sequences. Coding sequences of the TNL genes were used to calculate the nonsynonymous substitutions per nonsynonymous site (K_a) and synonymous substitutions per synonymous site (K_s) values. K_a/K_s ratios for each clade were calculated using the program DnaSP 5.10.1(Rozas 2009). Relative age of duplication events was inferred from average K_s values. Syntenic map of TNL genes of *G. max* and *P. vulgaris* were created using SyMAP version 4.2 (Soderlund et al. 2011) using South Dakota State University High Performance Computing Cluster (HPC Cluster). The input files for syntenic map of *G. max* and *P. vulgaris* were whole chromosome sequences and TNL R-gene annotation files.

2.2.4. Expression and microRNA (miRNA) Analysis

Expression profiles of the soybean and common bean TNL genes were studied using RNA-Seq data available at Soybase.org (Severin et al. 2010) and http://plantgrn.noble.org/PvGEA/ (O'Rourke et al. 2014), respectively. Raw data (number of mapped reads per gene per tissue) and normalized data were used for the

respective expression profile studies. Heatmaps were generated using deseq normalized data through MeV package available in http://mev.tm4.org/ (Howe et al. 2011). The expression data were further studied through K-means clustering method dividing data into three clusters (moderate to minimal expression, minimal expression to no expression and no expression at all) based on Euclidean distance under 1000 iterations (Howe et al. 2011). The mature microRNA (miRNA) sequences of soybean and common bean were acquired from miRBase (Kozomara and Griffiths-Jones 2014). There were 639 and ten miRNAs for soybean and common bean, respectively. microRNAs and regular TNL gene sequences identified in this study, were used in Plant Small RNA Target Analysis Server to predict miRNA-targeting sites (Dai and Zhao 2011).

2.3. Results

2.3.1. Identification of TNL Genes in Soybean and Common Bean

The first HMM analysis of 46,430 soybean protein sequences resulted in 741 protein sequences orthologous to *A. thaliana* TNL reference sequences, at a threshold expectation value of 0.05. After InterProscan annotation, the NB-ARC domains extracted from 644 sequences were employed in the second HMM analysis at a threshold expectation value of 0.001. This resulted in 153 unique protein sequences containing TIR and NB-ARC regions in soybean. Among them, 117 sequences containing three signature motifs: P-loop, Kinase-2, and GLPL were identified as regular TNL genes, and included for further analysis. Similarly, the first HMM analysis of 27,197 common bean protein sequences resulted in 465 protein sequences orthologous to the reference TNL protein sequences in *Arabidopsis*, at stringency of 0.05. After Interproscan annotation, the NB-ARC domains extracted from 395 sequences were used in the second HMM analysis at

stringency of 0.001. Among 93 sequences containing TIR, NB-ARCs regions, 77

sequences containing three signature motifs were identified as regular TNL proteins in

common bean and were included for further analysis. The LxxLxxLxx signatures were

present in 126 out of 153 protein sequences ranging from 1 to 18 in soybean, and 81 out

of 93 protein sequences ranging from one to 27 in common bean. Further, the TNL genes

were classified into subgroups: TNL, TN (truncated)L, TLTN, TN (lacking LRRs),

TNTN, TNT and TX (lacking both NBS and LRRs) (Supplementary Fig. 1 and

Supplementary Fig. 2; Supplementary Table 1). Soybean and common bean TNL

proteins possess some unique domains in some members. The gene members

GLYMA16G33971 and GLYMA16G33961 of soybean possess basic secretory proteins

domain (BSP, PF04450) at the C-terminal instead of leucine rich domains. Other gene

members GLYMA08G41270, GLYMA09G29050, GLYMA16G23790,

GLYMA16G33590, GLYMA16G33616, Phvul.001G128200, and Phvul.002G171400

possessed zinc binding in reverse transcriptase (zf-RVT, PF13966) and reverse

transcriptase like (RVT_3, PF13456) domains. SignalP analysis of the identified TNL

genes showed 17 and ten N-terminal signal peptides, in soybean and common bean,

respectively (Supplementary Table 2). Out of the 153 and 93 soybean and common bean

proteins, 19 and nine proteins were identified to contain a putative NLS using

NLStrdamus (Supplementary Table 3).

In soybean, P-loop, Kinase-2, GLPL, and RNBS C motifs are present in all 117

regular TNL genes, whereas RNBS A, RNBS B and RNBS D are present in 115 (except

GLYMA08G41560 and GLYMA13G26450), 116 (except GLYMA06G40690) and 52

members, respectively. Of the total 153 genes annotated to have TIR domain, only 123

genes have all TIR1, TIR2, TIR3, and TIR4 motifs. The TIR1, TIR2, TIR3, TIR4 are present in 139, 143, 142 and 145 members, respectively. Altogether, 108 gene members possess all four TIR domains and three NBS signature motifs. Also, MHDL motif is present in 108 gene members (Supplementary Fig. 3). Similarly, in common bean, P-loop, Kinase-2, GLPL, and RNBS C motifs are present in all 77 genes whereas RNBS A, RNBS B, and RNBS D are present in 75 (excluding Phvul.010G028500 and Phvul.010G136800), 76 (excluding Phvul.011G140300) and 55 members respectively. Of the 93 genes annotated by Interproscan to have TIR domain, only 80 genes have all TIR1, TIR2, TIR3, and TIR4 motifs. TIR1 and TIR2 motifs are present in 88 members whereas TIR3 and TIR4 motifs are present in 90 members. Altogether, 70 gene members possess all four TIR domains and three NBS signature motifs. Also, MHDL motif is present in 73 gene members (Supplementary Fig. 4). The sequences of the conserved motifs of soybean and common bean are represented in Table 2.1.

2.3.2. Gene Clustering and Structural Variation

Figure 2.1 visualizes the TNL gene clustering in soybean and common bean. Since Phvul.L003500 and GLYMA0220S50 were present in the scaffold_220 of common bean genome and scaffold_40 of soybean respectively, they were excluded from cluster analysis. Seventy four of the 117 TNL genes identified in soybean formed 28 clusters located on 15 of 20 chromosomes, and most of them were located on chromosomes Gm03, Gm06 and Gm16. Approximately, 40% of the clustered genes were present in Gm16 (Figure 2.1A). Likewise, approximately 48 of 77 TNL genes identified in common bean formed 14 clusters located on five of 11 chromosomes, and mostly clustered on chromosome Pv10. Approximately, 70% of the clustered genes are present in Pv10

(Figure 2.1B). Exon analysis showed presence of average of 5.5 exons in both soybean

(ranging from two to 12 exons) and common bean (ranging from two to 13 exons) (See

Fig. 2.1)
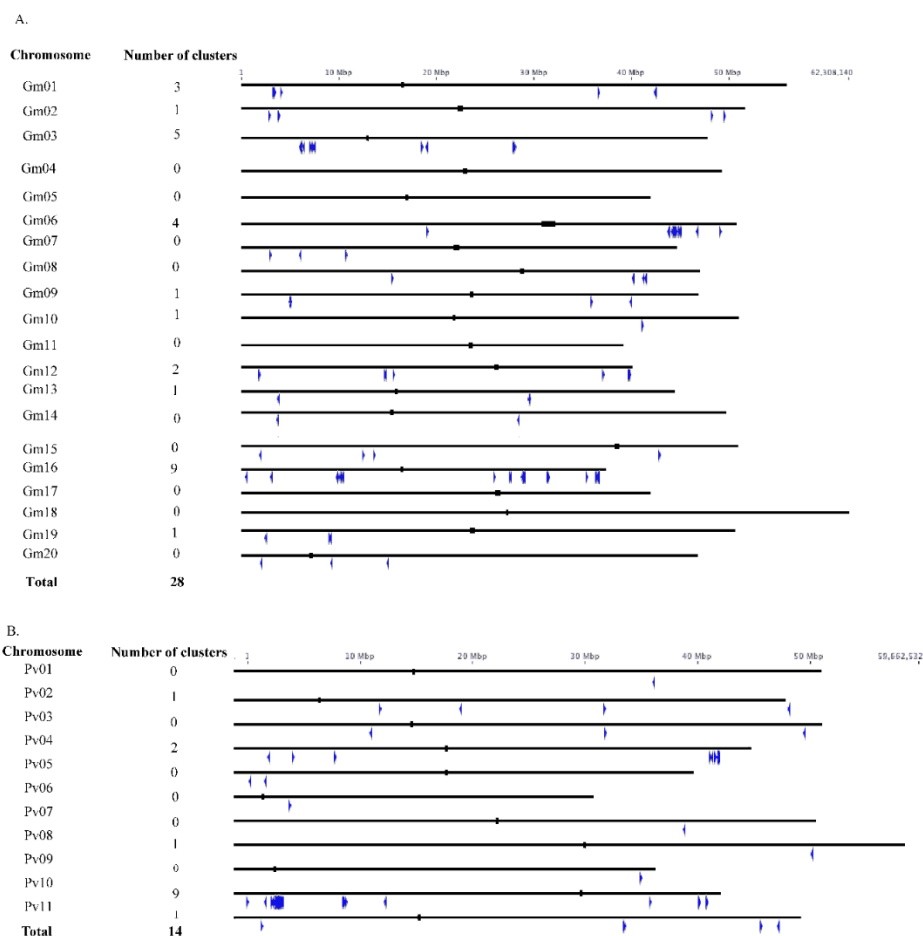
Figure 2.1. Chromosomal distribution of TNL gene clusters in A) soybean (N=20) and in B) common bean (N=11) genomes. Each blue arrow represents a TNL gene location and orientation on a chromosome represented by the black line. A black rectangle on the chromosome represents a centromere position.

2.3.4. K_s Values

Estimated synonymous substitutions per synonymous site (K_s) values were used

to infer the relative age of the gene clusters. In soybean, average K_s values were highest

for cluster 1_1 (GLYMA01G03921, GLYMA01G03980, GLYMA01G04000) and lowest

for the cluster 3_2 (GLYMA03G06854, GLYMA03G06976). Likewise, in common

bean, average Ks values were highest for the cluster 4_1 (Phvul.004G134300,

Phvul.004G135100) and lowest for the cluster 8_1 (Phvul.008G195100,

Phvul.009G195300) (Supplementary Table 4). The average K_a/K_s ratios of the clades

inferred from phylogenetic tree showed the values less than 1 except for the pair

GLYMA06G41714, GLYMA06G41896 having values greater than 1 accessions in L11

clade (See Figure 2.2).

2.3.5. Phylogenetic and Syntenic Relationships

Phylogenetic relationships of TNL genes of soybean and common bean were

examined and compared with those of *A. thaliana*. We used amino acid sequences of

NBS domain with P-loop, Kinase 2 and GLPL motifs from these species for the

phylogenetic analyses. Among the eight clades reported in *Arabidopsis*, only TNL-C

gene members were nested with GLYMA01G04590, GLYMA08G40501,

Phvul.002G079200, and Phvul.003G072500 (TNL-C clade) (Fig. 2). Although soybean

and common bean diverged about 19MYA, most of the TNL genes appear to be

conserved in these plants. Based on the clade support and orthologous relationships, we

named the clades (I-XIV and TNL-C) of soybean and common bean TNL genes on the

phylogenetic tree. Soybean and common bean TNL genes are concentrated on

chromosome Gm16 and Pv10 respectively. Hence, we examined if chromosome Gm16

and Pv10 were clustered in the same clade in the phylogenetic tree. The TNL genes on

chromosome Gm16 were found in 11 different clades and those on chromosome Pv10

were found in four different clades. For comparative analysis of orthologs of disease

resistance genes, we constructed the second phylogenetic tree including characterized and

manually curated TNL and RPW8-NL type genes from *Solanum tuberosum*, *S.*

tuberosum subsp. *andigena*, *A. thaliana*, *Nicotiana benthamiana*, *N. glutinosa*, *Linum*

usitatissimum, and *G. max* (Supplementary Fig. 5). The BLAST result of Rj2/Rfg1

protein sequence within the Ensembl Genomes database showed that GLYMA16G33780,

GLYMA16G23800, GLYMA19G02670, Phvul.004G028900 are the most likely gene

accessions for *Rj2/Rfg1* gene. These accessions were found in the same clade in the

second phylogenetic tree as well. Thus, we assigned Phvul.004G028900 gene accession

as homolog for *Rj2/Rfg1* gene in common bean (BS=97%). The *KR1*, gene resistant to

soybean mosaic virus (SMV), clustered together with GLYMA19G07680,

GLYMA19G07650, GLYMA19G07700 and Phvul.004G058700 with high bootstrap

support and are the most likely orthologs of *KR1* gene. Likewise, flax rust resistance

proteins L6 and M were clustered with GLYMA01G04590, Phvul.002G079200,

GLYMA08G40501 and Phvul.003G072500 with strong bootstrap support suggesting that

these are the most likely orthologs of L6 and M proteins. Other group of TNLs of plants

belonging to Solanaceae family (Gro 1.4 from *S. tuberosum*, RY-1 from *S. tuberosum*

subsp. *andigena*, N from *N. glutinosa*) formed their own cluster. Likewise, *RPP5, RPP4,*

RAC1, SSI4, RPP1 genes from *Arabidopsis* formed their own clade and were with sub

clade formed by many soybean and common bean TNL genes. The RPW8 group of

resistance genes containing *ADR1*, *ADR1-L1* from *Arabidopsis* and the NRG1 from *N.*

benthamiana formed their own clusters suggesting these groups might have different

evolutionary history than the TNL resistance proteins. The syntenic map of TNL genes of

G. max and *P. vulgaris*, created using SyMAP, showed high synteny between

chromosomes Pv10 and Gm03; Pv07 and Gm10, Pv01 and Gm14 (Supplementary Fig. 6). It also showed that most of the fragments of multiple chromosomes of common bean had similarity to the single chromosome of soybean. For instance, soybean chromosome Gm08 possesses the chromosomal fragments from Pv02, Pv04, Pv01, Pv03, Pv05, Pv06, Pv08 and Pv10. Likewise, soybean chromosome Gm05 possess the fragments from Pv03, Pv09, Pv04, Pv02, Pv01.
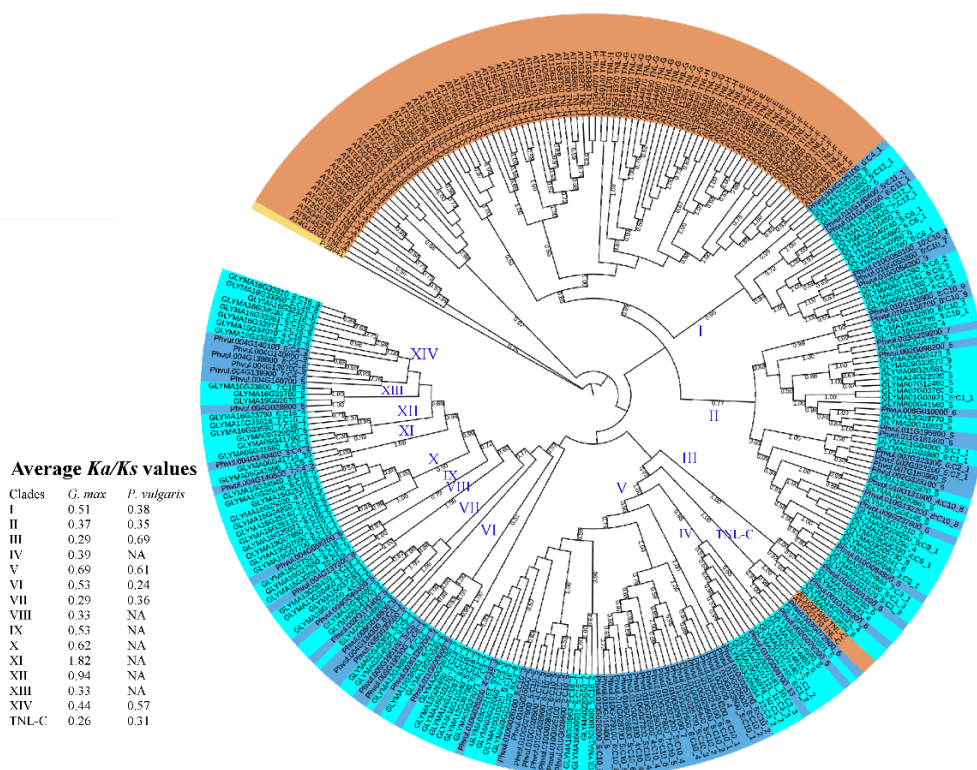
Figure 2.2. Phylogenetic relationships of the NB-ARC amino acid sequences of the TNL genes from *A. thaliana* (AT; orange), *P. vulgaris* (Phvul; blue) and *G. max* (GLYMA; light blue). The JTT+G+I (Jones–Taylor–Thornton with gamma distribution and invariant sites) model was used for the Maximum-Likelihood tree construction using 100 bootstrap replicates. The tree was rooted using *Streptomyces coelicolor* (P25941) as an outgroup. The accessions of soybean and common bean are followed by the number of exons and genomic cluster. The clades are named I-XIV and TNL-C. Average K_a/K_s values of each clade are represented on the left side of the figure.

2.3.6. Expression and miRNA Analysis

Available gene expression data for soybean and common bean are visualized as heatmaps (Figure 2.3). Eighty-nine of the 153 TNL genes of soybean had unique mappable reads whereas we obtained all 93 common bean TNL genes with unique mappable reads. Soybean had reads ranging from 1351 to 222 reads in the upper quartile while common bean had reads ranging from 28460 to 4124 in the upper quartile. GLYMA16G33590 (1351 reads in diverse set of tissues under different conditions) and Phvul.002G323800 (28460 reads in diverse sets of tissues under different conditions) had highest number of reads and mostly expressed in all tissues. The dataset revealed GLYMA02G03760 and GLYMA03G05730 had zero expression (zero reads) in all tissues whereas in common bean the minimum number of reads was three for Phvul.010G025700. The dataset and heatmap revealed some of the genes were tissue specific (Supplementary Table 5). We observed 15 genes in Cluster I representing moderate to minimal expression, 29 genes in Cluster II representing minimal expression to no expression and 45 genes in Cluster III representing no expression at all (except significantly high expression of GLYMA12G03040, GLYMA16G25140 in nodule and GLYMA01G31550 in root) in all tissues in soybean. Similarly, in common bean, 11 genes in Cluster I representing moderate to minimal expression (Phvul.010G054400 being highly expressed in pre-fixing and ineffectively fixing nodules), 40 genes in Cluster II representing minimal expression to no expression (except significantly high expression of Phvul.008G195300, Phvul.004G046400, Phvul.008G19510 in whole roots and Phvul.010G054600 in ineffectively fixing nodules) and 42 genes in Cluster III representing no expression at all in all tissues was observed (Supplementary Table 6). Six

hundred thirty nine and 10 miRNAs for soybean and common bean respectively were utilized to discover potential regulators of identified regular TNL gene sequences. We identified 35 soybean microRNAs and two common bean microRNAs to be involved in TNL genes regulation (Supplementary Table 7). Among 35 soybean miRNAs involved, 16 seemed to regulate soybean TNL genes, nine seemed to regulate common bean TNL genes, and ten shared by both species. In case of two common bean microRNAs (pvu-miR482-3p and pvu-miR2118), they seemed to regulate both soybean and common bean TNL genes. The pvu-miR482-3p regulates GLYMA01G31520 and GLYMA08G41270 soybean TNL proteins and pvu-miR2118 regulates GLYMA13G03770, GLYMA01G04590, GLYMA12G03040, GLYMA20G06780, GLYMA01G27455, GLYMA03G14888 and GLYMA12G36841 soybean TNL proteins.
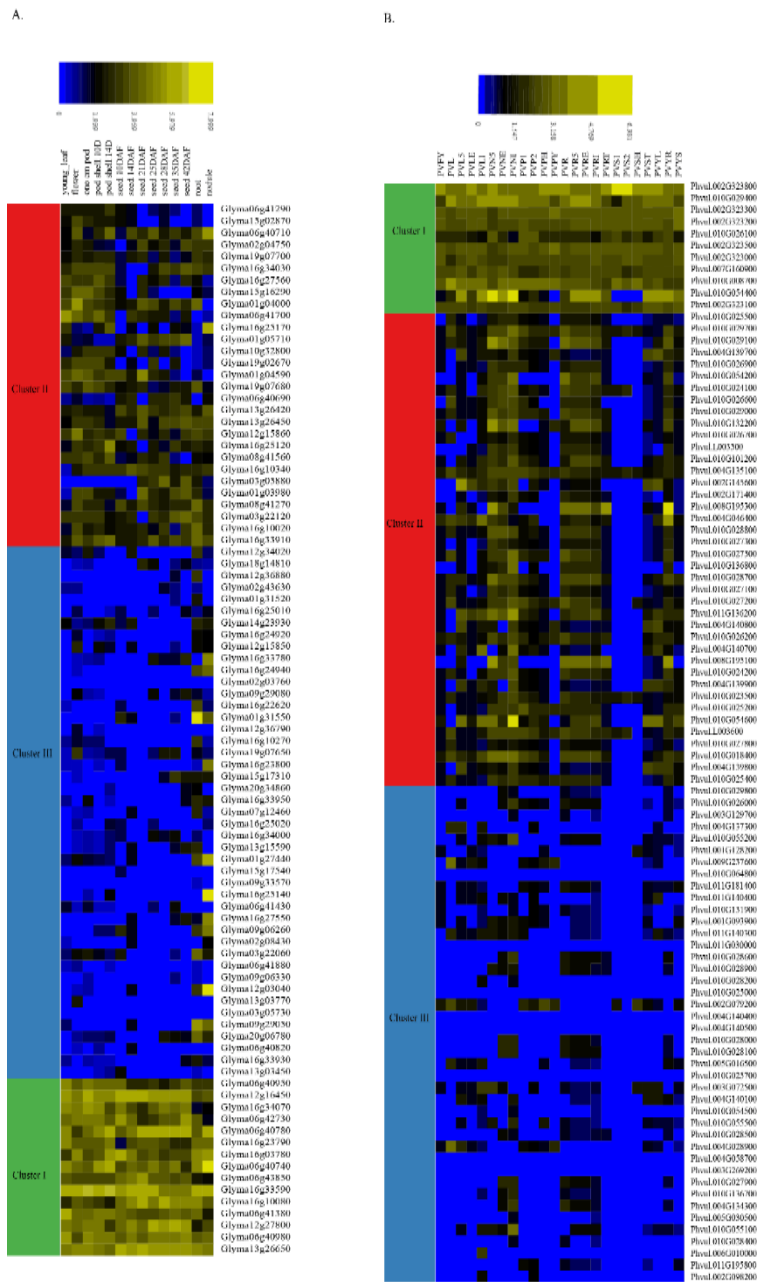
Figure 2.3. Expression profile for A) soybean TNL genes and B) common bean TNL genes visualized as heatmaps. Deseq normalized data were employed to generate the heatmap for soybean TNL gene expression in different tissues. Clustering (I, II and III) was based on K means Clustering method.
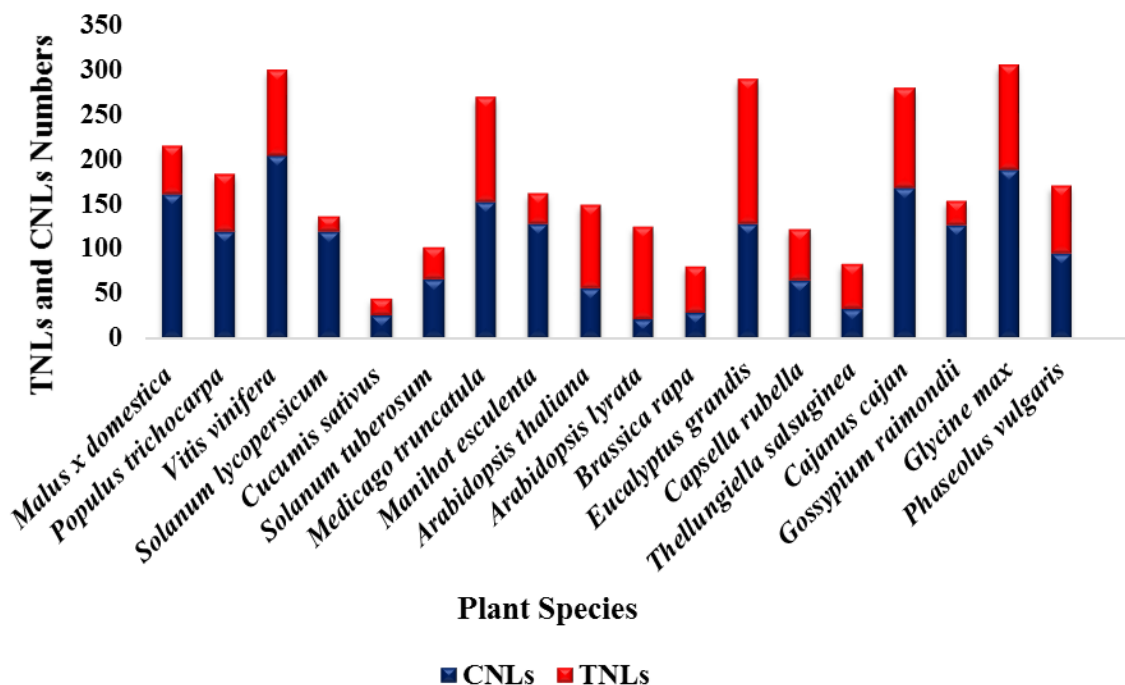
Figure 2.4. Diversity of putative TNL and CNL genes in the genomes of 18 different plant species.

2.4. Discussion

2.4.1. Diversity of TNL Genes

Whole genome sequencing of multiple plant species at various taxonomic levels has implications in revealing the details of genome architecture (Goff et al. 2002) and shedding light on the processes leading to functional divergences (Caicedo and Purugganan 2005). In the present study, we carried out comparative genomics of the TNL R genes from soybean and common bean to understand their diversity, structure and functions. We identified 117 and 77 regular TNL genes in soybean and common bean, respectively. Shao et al. (2014) reported 178 (124 TNL, 24 TN, 17 N_{TIR}, 12 N_{TIR} only, 1 others) in soybean and 103 (78 TNL, six TN, 15 N_{TIR} ,11 N_{TIR} only and two others) in common bean. Zheng et al. (2016) reported 237 TNLs subclass (112 TIR, 76 TN and 49 TNL) in soybean and 66 TNLs subclass (57 TIR, eight TN and one TNL) in common

bean. Our identification is based on the presence of three signature motifs (P-loop, Kinase-2, and GLPL) in the NBS domain (Andersen et al. 2016; Nepal and Benson 2015). We observed these motifs manually and verified the presence of these motifs. Since the threshold expectation value for second HMM profiling was 0.001, the false positive results and a few functional TNLs might have been omitted during HMM profiling. We compared our findings on the TNL gene diversity with respect to CNL gene diversity with the previous findings from other plant species (Figure 2.4, Supplementary Table 8). The numbers of TNLs in soybean and common bean were lower than CNL genes- 188 and 94 CNLs, respectively (Benson 2014; Nepal and Benson 2015). One recent study has reported a wide variation (0.55% to 54.17%) of the proportion of TNLs across eudicots (Zhang et al. 2016a). The factors determining the numbers of TNL or CNL genes in the genome can be the pool of pathogens that infect the plant and the different patterns of evolution that drive the success and failure of R genes (Lozano et al. 2012). These R genes evolve in plants by divergent selection and that is explained by birth and death model (Michelmore and Meyers 1998). The soybean TNLs number was found almost similar to TNLs of *Medicago truncatula* (118 TNL genes) despite that the soybean genome experienced two whole genome duplication (WGD) events (59 and 13 million years ago) (Schmutz et al. 2010). It could be attributable to low exposure to pathogenic environment and a longer domestication history in soybean than *M. truncatula* (Kang et al. 2012).

Soybean and common bean have gone through WGD events 56.5 million years ago and then diverged 19.2 million years ago (Lavin et al. 2005). After diverging, soybean underwent an independent WGD 13 million years ago (Schmutz et al. 2010).

The WGD events have important role in evolution of protein-coding genes, as genes could be produced and mutated causing no harm to organism (Taylor and Raes 2004). We expected in soybean twice the number of genes as in common bean, whereas soybean retained 64% more TNL genes present in common bean. The reason might be the genes that shared the first WGD might not be involved in disease resistance or may have been purged through purifying selection. An equally plausible explanation can be that common bean has evolved more rapidly than soybean since K_s value for common bean (8.46×10^{-9} substitutions/year) is 1.4 times higher than that of soybean (5.85×10^{-9} substitutions/year) (Schmutz et al. 2014). Subfunctionalization of certain R genes in soybean possibly facilitated by artificial selection (during domestication) and R genes prone to diploidization events might have contributed to the reduction in number of R-genes (Zheng et al. 2016).

Analysis of the identified TNL R genes in soybean and common bean showed 117 and 77 gene members containing full length and conserved signature motifs (P-loop, Kinase 2 and GLPL). We observed conserved sequence "DDVD" of Kinase-2 and 'TTRD' in the RNBS-B motif, the distinguished feature of TNL subgroup of NBS-LRR genes (Shao et al. 2016). Further classification of TNL genes has revealed 12 TN, four TX, eight TN(truncated)L, one TNTN, one TLTNT subgroup gene members in soybean and 14 TN, four TN(truncated)L, one TNT, one TNTN subgroup gene members in common bean. Similar classification within the TNL group was reported in previous studies (Lozano et al. 2012; Meyers et al. 2002; Yu et al. 2014). Previous findings about the occurrence of TX, TN and TNL genes in dicots and conifers suggest that these were present approximately 300MYA when the species diverged (Savard et al. 1994). The

function of TX and TN type proteins in plants is unknown but have been compared to some Toll-like receptor (TLR) family proteins such as MyD88 and Mal, which act as adaptor proteins involved in mammalian and *Drosophila* immune responses (Meyers et al. 2002). Nandety et al. (2013) inferred the role of TN and TX proteins in plant defense responses as the adapters or guard complexes interacting with TNL proteins. Apart from the conserved domains in NBS-LRR proteins, these genes also contain unique and variable domains which are involved in resistance mechanism (Cesari et al. 2014). The unique domains such as basic secretory proteins domain (BSP, PF04450) zinc binding in reverse transcriptase (zf-RVT, PF13966) and reverse transcriptase like (RVT_3, PF13456) domains were identified in the TNL genes of soybean and common bean. Specific functions of the BSP domain are not clear but believed to be involved in the defense mechanism against pathogens (Kuwabara et al. 1999). The zf-RVT domain is the zinc-binding region of putative reverse transcriptase and RVT_3 domain is found in plants and appear to be the part of a retrotransposon (Marchler-bauer et al. 2015). The TNL gene, *RRS1*, encodes a protein having an additional WRKY domain in the C-terminal that plays an important role in plant defense by acting as integrated decoys (Cesari et al. 2014). Among TNL genes with N-terminal signal peptides, six in soybean and four in common bean, were predicted to be transmembrane type. TargetP analysis predicted 17 proteins in soybean, 15 of them were predicted to enter secretory pathway, one (GLYMA15G37276) was predicted to enter mitochondria and one (GLYMA16G10020) was predicted to enter chloroplast. Likewise, of the 10 predicted proteins in common bean, eight were predicted to enter the secretory pathway and remaining two (Phvul.010G028200 and Phvul.002G098200) were predicted to enter the

chloroplast. When flax TNL proteins, such as L6 and M interact with fungal effectors AvrL567 and AvrM, respectively, L6 gets localized to Golgi endomembrane and M gets localized to vacuolar endomembrane (Takemoto et al. 2012). Another TNL protein, RRS1 gets localized to the nucleus upon detection of broad spectrum avr proteins from *Ralstonia solanacearum* strains (Lahaye 2002).

2.4.2. Gene Clustering and Structural Variation

Clustering of the NBS-LRR genes have been observed in many previous studies (Asai et al. 2002; Meyers et al. 2003; Mun et al. 2009; Wan et al. 2013). The cluster arrangement assists their evolution through mispairing during recombination, which aids in exchange of sequences (Friedman and Baker 2007; Hulbert et al. 2001). This process is assisted by several other processes such as gene conversions, unequal crossovers, and tandem duplications (Leister 2004). In this study, 74 out of 117 TNL genes (approximately 63%) were involved in forming 28 clusters in soybean and 48 out of 77 TNL genes (approximately 62%) were involved in forming 14 clusters in common bean. In *Arabidopsis*, 43 clusters were formed by 109 out of 149 (approximately 76%) NBS-LRR genes (Meyers et al. 2003). The largest gene cluster was present in chromosome 16 in soybean (30 genes; approximately 40%) and chromosome 10 in common bean (34 genes; approximately 70%). The formation of big clusters would be as a result of tandem duplications or chromosomal rearrangements, and transposases activities conferring benefits of co-amplification of clustered genes and aiding adaptation to the changing environment (Pontes et al. 2004; Reams and Neidle 2004). Kang et al. (2012) reported the presence of *Mutator*-like element (MULE) transposase and the *MuDR* (*Mutator* autonomous element) family transposase domain in TNL gene

GLYMA16G23790 located on chromosome Gm16. In this study, we report the presence

of retrotransposon domains (reverse transcriptase like; RVT_3) in *GLYMA16G23790*,

which might contribute to tandem duplications of R genes as reported in a previous study

(Ratnaparkhe et al. 2011). Large numbers of R genes in a single chromosome have been

observed in other plant genomes as well (Ameline-Torregrosa et al. 2008; Meyers et al.

2003; Nepal and Benson 2015). The eight gene members of the cluster 16_5 are located

within a 0.24Mb section of chromosome Gm16 in soybean, and are nested together in

phylogenetic tree suggesting their evolution through tandem duplication. Likewise, the 34

TNL genes that are distributed in nine clusters in chromosome 10Pv of common bean

genome show evidence of R gene expansion through tandem duplications. Similar

expansion of NBS-LRR genes in *Arabidopsis* was believed to be due to tandem and

large-scale block duplications (Leister 2004). Overall, tandem duplications we observed

in TNL R genes in this research are consistent with those observed in previous studies,

and our inferences are: tandem duplications should be the source of genetic variation

(Dangl and Jones 2001), frequent sequence exchanges and high copy number suggest

their rapid evolution (Li et al. 2010), and high diversity introduced through tandem

duplication could guarantee the resistance to rapidly evolving pathogen effectors (Kuang

et al. 2008).

2.4.3. K_s Values as a Proxy of Gene Duplication History

The K_s values infers the history of gene duplication events when WGD and

polyploidy are taken into account (Pfeil et al. 2005; Schmutz et al. 2010). In soybean,

cluster 1_1 (GLYMA01G03921, GLYMA01G03980, GLYMA01G04000) has the

highest average K_s value of 1.968 and the cluster 3_2 (GLYMA03G06854,

GLYMA03G06976) has the lowest average K_s value of 0.013 which suggest that cluster 1_1 formed before the cluster 3_2. Likewise, in common bean, the cluster 4_1 (Phvul.004G134300, Phvul.004G135100) has the highest average K_s value of 1.518 and the cluster 8_1 (Phvul.008G195100, Phvul.009G195300) has the lowest average K_s value of 0.036 suggesting cluster 4_1 formed before the cluster 8_1 was formed. The selection pressure was detected using K_a/K_s ratios with the interpretation that the value greater than 1 indicates positive selection, less than 1 indicates negative or stabilizing or purifying selection, and equal to 1 as neutral selection. Although overall average K_a/K_s ratios of the clades inferred from the phylogenetic tree showed value less than 1 suggesting a purifying selection for the TNL family, the pair GLYMA06G41714 and GLYMA06G41896 accessions showed K_a/K_s ratio value >1 in the XI clade, indicating that genes have undergone positive selection. The mean K_a/K_s ratios for the TNL genes in soybean (0.57) were slightly higher than for common bean (0.44). The TNL exons in both soybean and common bean were similar (5.5 exons per gene) consistent with the number of TNL exons reported in *Arabidopsis* (5.25 on average) (Meyers et al. 2003), however, are not consistent with the number of TNL exons in grapevine (7.68 on average) or in poplar (3.5 on average) (Yang et al. 2008). The number of TNL exons we found in this research is greater than the number of CNL exons in soybean (3.6 on average), common bean (4.0 on average), grapevine (3.22 on average), poplar (2.23 on average) and *Arabidopsis* (2.17 on average) (Benson 2014; Meyers et al. 2003; Nepal and Benson 2015; Yang et al. 2008). Increased number of exons in TNL genes might have implication in alternate splicing, a mechanism of making diverse defense proteins by host plants in response to rapidly evolving pathogen effectors.

2.4.4. Phylogenetic Relationships of Identified TNL Genes and Their Orthologs

Phylogenetic analysis of the NBS protein sequences of *Arabidopsis*, soybean and common bean showed evidence of gene duplications and revealed orthologs. Species-specific nesting patterns are common in the phylogenetic tree except *Arabidopsis* TNL-C members AT1G27180 and AT1G27180, which were nested with their orthologs in soybean and common bean. Similar patterns were reported in previous studies for TIR-NBS sequences from distantly related taxa (Pan et al. 2000; Plocik et al. 2004; Wan et al. 2013). These patterns are often reported to be useful in differentiating species (Wei et al. 2013). As expected, due to relatively recent divergence, soybean and common bean did not have as many species-specific clades of TNL genes as compared with the *Arabidopsis*. Interspecies clades or mixed clades allowed us to identify TNL orthologs in the two species. The phylogenetic analyses of the manually curated TNL genes revealed homologs for *Rj2/Rfg1*, *KR1*, *L6* and *M* proteins. The TNL protein Phvul.010G054400, a member of the TN subgroup, was expressed in SCN resistant genotypes (Jain et al. 2016). Our prediction that the TNL orthologs of Phvul.010G054400 protein (in the TN subgroup) are GLYMA12G16450 (E-value: 1.4E-127), GLYMA06G41241 (Evalue: 4E-126), GLYMA06G41380 (Evalue: 2.5E-124), GLYMA06G40710 (Evalue: 3.7E-127) and GLYMA06G40780 (Evalue: 8.5E-119) in soybean, and are the potential gene accessions for conferring resistance to SCN infection. These predictions should be validated through functional characterization, such as hairy root transformation, over expressing and silencing of defense related genes. The TNLs *RPP1* (Botella et al. 1998), *RPP4* (Van Der Biezen et al. 2002), *RPP5* (Noël et al. 1999) confer resistance to *Pernospora parasitica*, also *RPS4* resist against *Pseudomonas syringae* (Gassmann et al.

1999). In addition, *Linum usitatissimum* TNLs L, L6, M, P proteins have shown resistance against *Melampsora lini* (Anderson et al. 1997; Dodds et al. 2001; Ellis et al. 1999; Lawrence et al. 1995) and *Nicotiana tabacum N* protein has shown specific protein interactions with effector proteins of Tobacco Mosaic Virus (TMV) (Whitham et al. 1994). The *KR1* gene encodes a TNL protein that confers resistance against soybean mosaic virus (SMV) and was isolated from SMV resistant variety Kefeng-1 soybean (He et al. 2003). Among the characterized TNL genes, *Rj2* and *Rfg1* encoded proteins restrict the nodulation in soybean (Yang et al. 2010) suggesting their role in biological nitrogen fixation, an example of broader role of the TNLs in biotic interactions. To date only a few important TNL genes have been characterized, and our discussion was limited to the comparison with the characterized genes. Functions of the majority of the TNL genes we identified are unknown suggesting their involvement in unknown resistance pathways or non-host resistance responses (Schulze-Lefert and Panstruga 2011) warranting experimental studies leading to the characterization of these genes.

Many studies have been done to understand the syntenic relationship in plants such as tomato, potato, sorghum and maize (Bonierbale et al. 1988; Whitkus et al. 1992). The purpose of syntenic map data (Choi et al. 2004) is to provide a reference point for ortholog comparison between species. McClean et al. (2010) have shown the syntenic relationship between common bean and soybean. They have suggested that soybean and common bean shared loci in syntenic blocks. Another study of syntenic relationships between the two showed single region of the chromosomal blocks of common bean mapped onto the two chromosomal blocks in soybean (Schmutz et al. 2014).This might be due to independent WGD along with extensive breakage and rearrangement in the

soybean genome that correspond to fragments from several common bean chromosomes (McClean et al. 2010; Schmutz et al. 2010). We observed that all of the soybean chromosomes contained fragments from multiple chromosomes of common bean.

2.4.5. TNL Gene Expression and Role of microRNA (miRNA)

The expression profiles showed differential expression of genes in tissues with moderate, minimal to no expression in respective tissues of soybean and common bean. The basal level expression and tissue-specificity of some TNL genes showed functional divergence. For example, GLYMA16G33780 (Rj2/Rfg1 protein) restricts nodulation and is highly expressed only in nodules. The expression of R genes in plants are regulated through utilization of microRNA (miRNA) which helps in creating balance between fitness cost and benefits of resistance (Shivaprasad et al. 2012). These miRNAs are single stranded hairpin RNAs (hpRNAs) which yield few functional small RNAs (Axtell 2013). Previous studies reported that expression of *RPM1* and *RPW8* genes have potential fitness costs in *Arabidopsis* (Orgil et al. 2007; Tian et al. 2003) and are lethal to plant cells (Stokes et al. 2002). In *M. truncatula*, the NBS-LRR genes are controlled by five miRNAs namely miR2118a, b, and c, miR2109, and miR1507 (Fei et al. 2015; Zhai et al. 2011a; Zhai et al. 2011b). The TNL gene involved in symbiotic specificity in soybean, is targeted by microRNA miR482 (Fei et al. 2013). We identified microRNAs belonging to 37 families that may be involved in targeting TNL genes of soybean and common bean. Among the identified microRNAs, some were predicted to recognize the TNL genes of both legumes and some were predicted to target TNL genes of unrelated species. The major miRNAs involved belonged to gma-miR2118a-3p, gma-miR2118b-3p, gma-miR5668, gma-miR2109-5p, gma-miR1510a-3p, gma-miR1510b-3p, pvu-miR2118

family. Among these, miR482/2118 family is considered to be ancient, huge and complex

family that target NBS-LRR genes in wide variety of plant genomes except in

Brassicaceae and Poaceae family (González et al. 2015; Shivaprasad et al. 2012; Zhang et

al. 2016a). This family down regulates NSB-LRR proteins when there are no pathogens

(González et al. 2015). Presence of such potential microRNAs depicts the establishment

of microRNA-NBS gene relationship and providing proof for common ancestry of

soybean and common bean. Such relationship was also shown between soybean and *M.*

truncatula (Shao et al. 2014). Previous studies have mentioned of microRNAs that can

target NBS-LRR genes of different species (Shivaprasad et al. 2012; Zhai et al. 2011b).

For instance, nine soybean microRNAs that can potentially target TNL genes of common

bean only were identified in this study. Recently, Cui et al., (2017) have shown that

microRNA gma-miR1510a/b plays a crucial role in cleavage of the TNL protein

Glyma.16G135500 (GLYMA16G24940 in this study) that is upregulated in response to

Phytophthora sojae in soybean. This suggests the role of microRNAs in regulation of

TNL R genes in response to pathogens.

2.5. Conclusions

We identified 117 and 77 regular TNL R genes in soybean and common bean,

respectively, and assessed their structural and functional divergence. The presence of

unique domains such as BSP, zinc binding in reverse transcriptase, and reverse

transcriptase-like domains and signal peptides identified in some TNL genes provides

insights into their evolution and sub-cellular localization. Most of the TNL genes

identified in soybean and common bean have undergone purifying selection rather than

positive selection except for a few accessions. Approximately 63% of the regular TNL

genes were found in clusters in both soybean and common bean, which signifies the

occurrence of tandem duplications. We also identified microRNAs potentially targeting

TNL genes in both soybean and common bean, and involving in balancing fitness costs

and resistance advantages. Characterization of these TNL genes is warranted for

understanding resistance pathways paving avenues toward crop improvement.

Acknowledgements

Supplementary Files: https://doi.org/10.1007/s10528-018-9851-z

References

Ameline-Torregrosa C, Wang BB, O'Bleness MS, Deshpande S, Zhu H, Roe B, Young ND, Cannon SB (2008) Identification and characterization of nucleotide-binding site-leucine-rich repeat genes in the model plant *Medicago truncatula*. Plant Physiol 146 doi:10.1104/pp.107.104588

Andersen EJ, Ali S, Reese RN, Yen Y, Neupane S, Nepal MP (2016) Diversity and Evolution of Disease Resistance Genes in Barley (Hordeum vulgare L.). Evol Bioinformatics Online 12:99

Andersen EJ, Nepal MP (2017) Genetic diversity of disease resistance genes in foxtail millet (*Setaria italica* L.). Plant Gene 10:8-16

Anderson PA, Lawrence GJ, Morrish BC, Ayliffe MA, Finnegan EJ, Ellis JG (1997) Inactivation of the flax rust resistance gene *M* associated with loss of a repeated unit within the leucine-rich repeat coding region. Plant Cell 9:641-651

Andolfo G, Ercolano MR (2015) Plant innate immunity multicomponent model. Front Plant Sci 6

Andolfo G, Sanseverino W, Aversano R, Frusciante L, Ercolano M (2014) Genome-wide identification and analysis of candidate genes for disease resistance in tomato. Mol Breed 33:227-233

Arumuganathan K, Earle E (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Report 9:208-218

Arya P, Kumar G, Acharya V, Singh AK (2014) Genome-wide identification and expression analysis of NBS-encoding genes in *Malus x domestica* and expansion of NBS genes family in Rosaceae. PLoS One 9:e107987 doi:10.1371/journal.pone.0107987

Asai T, Tena G, Plotnikova J, Willmann MR, Chiu WL, Gomez-Gomez L, Boller T, Ausubel FM, Sheen J (2002) MAP kinase signalling cascade in *Arabidopsis* innate immunity. Nature 415:977-983 doi:10.1038/415977a

Ashfield T, Egan AN, Pfeil BE, Chen NW, Podicheti R, Ratnaparkhe MB, Ameline-Torregrosa C, Denny R, Cannon S, Doyle JJ (2012) Evolution of a complex disease resistance gene cluster in diploid *Phaseolus* and tetraploid *Glycine*. Plant Physiol 159:336-354

Axtell MJ (2013) Classification and comparison of small RNAs from plants. Annu Rev Plant Biol 64:137-159

Ba ANN, Pogoutse A, Provart N, Moses AM (2009) NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction. BMC Bioinformatics 10:1

Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in bipolymers. 28-36

Bales C, Zhang G, Liu M, Mensah C, Gu C, Song Q, Hyten D, Cregan P, Wang D (2013) Mapping soybean aphid resistance genes in PI 567598B. Theor Appl Genet 126:2081-2091

Bennett MD, Smith J (1976) Nuclear DNA amounts in angiosperms. Phil Trans Soc B 274:227-274

Benson B (2014) Disease Resistance Genes and their Evolutionary History in Six Plant Species. South Dakota State University; Brookings, SD

Bonierbale MW, Plaisted RL, Tanksley SD (1988) RFLP maps based on a common set of clones reveal modes of chromosomal evolution in potato and tomato. Genetics 120:1095-1103

Botella MA, Parker JE, Frost LN, Bittner-Eddy PD, Beynon JL, Daniels MJ, Holub EB, Jones JD (1998) Three genes of the *Arabidopsis RPP1* complex resistance locus recognize distinct *Peronospora parasitica* avirulence determinants. Plant Cell 10:1847-1860

Caicedo AL, Purugganan MD (2005) Comparative plant genomics. Frontiers and prospects. Plant Physiol 138:545-547

Cesari S, Bernoux M, Moncuquet P, Kroj T, Dodds PN (2014) A novel conserved mechanism for plant NLR protein pairs: the "integrated decoy" hypothesis. Front Plant Sci 5:10.3389

Chen NW, Sévignac M, Thareau V, Magdelenat G, David P, Ashfield T, Innes RW, Geffroy V (2010) Specific resistances against *Pseudomonas syringae* effectors AvrB and AvrRpm1 have evolved differently in common bean (*Phaseolus vulgaris*), soybean (*Glycine max*), and *Arabidopsis thaliana*. New Phytol 187:941-956

Choi H-K, Mun J-H, Kim D-J, Zhu H, Baek J-M, Mudge J, Roe B, Ellis N, Doyle J, Kiss GB (2004) Estimating genome conservation between crop and model legume species. Proc Natl Acad Sci USA 101:15289-15294

Christie N, Tobias PA, Naidoo S, Külheim C (2016) The *Eucalyptus grandis* NBS-LRR gene family: physical clustering and expression hotspots. Front Plant Sci 6 doi:10.3389/fpls.2015.01238

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

Cui X, Yan Q, Gan S, Xue D, Dou D, Guo N, Xing H (2017) Overexpression of gma-miR1510a/b suppresses the expression of a NB-LRR domain gene and reduces resistance to *Phytophthora sojae*. Gene 621:32-39

Dai X, Zhao PX (2011) psRNATarget: a plant small RNA target analysis server. Nucleic acids Res 39:W155-W159

Dangl JL, Jones JD (2001) Plant pathogens and integrated defense responses to infection. Nature 411:826-833

DeYoung BJ, Innes RW (2006) Plant NBS-LRR proteins in pathogen sensing and host defense. Nat Immunol 7:1243-1249

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

Du J, Tian Z, Sui Y, Zhao M, Song Q, Cannon SB, Cregan P, Ma J (2012) Pericentromeric effects shape the patterns of divergence, retention, and expression of duplicated genes in the paleopolyploid soybean. Plant Cell 24:21-32

Duc G, Agrama H, Bao S, Berger J, Bourion V, De Ron AM, Gowda CL, Mikic A, Millot D, Singh KB (2015) Breeding annual grain legumes for sustainable agriculture: new methods to approach complex traits and target new cultivar ideotypes. Crit Rev Plant Sci 34:381-411

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

Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2:953-971

Fei Q, Li P, Teng C, Meyers BC (2015) Secondary siRNAs from *Medicago* NB-LRRs modulated via miRNA–target interactions and their abundances. Plant J 83:451-465

Fei Q, Xia R, Meyers BC (2013) Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks. Plant Cell 25:2400-2415

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J (2013) Pfam: the protein families database. Nucleic Acids Res 42, D222–D231.

Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, Bateman A, Eddy SR (2015) HMMER web server: 2015 update. Nucleic Acids Res 43, W30–W38.

Friedman AR, Baker BJ (2007) The evolution of resistance genes in multi-protein plant resistance systems. Curr Opin Genet Dev 17 doi:10.1016/j.gde.2007.08.014

Gao H, Bhattacharyya MK (2008) The soybean-*Phytophthora* resistance locus *Rps1-k* encompasses coiled coil-nucleotide binding-leucine rich repeat-like genes and repetitive sequences. BMC Plant Biol 8:29

Gao H, Narayanan NN, Ellison L, Bhattacharyya MK (2005) Two classes of highly similar coiled coil-nucleotide binding-leucine rich repeat genes isolated from the *Rps1-k* locus encode Phytophthora resistance in soybean. Mol Plant Microbe Interact 18:1035-1045

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

Glazebrook J (2001) Genes controlling expression of defense responses in *Arabidopsis*—2001 status. Curr Opin Plant Biol 4:301-308

Goellner K, Loehrer M, Langenbach C, Conrath U, Koch E, Schaffrath U (2010) *Phakopsora pachyrhizi*, the causal agent of Asian soybean rust. Mol Plant Pathol 11:169-177 doi:10.1111/j.1364-3703.2009.00589.x

Goff SA, Ricke D, Lan T-H, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296:92-100

González VM, Müller S, Baulcombe D, Puigdomènech P (2015) Evolution of NBS-LRR gene copies among dicot plants and its regulation by members of the miR482/2118 superfamily of miRNAs. Mol Plant 8:329-331

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

Gururani MA, Venkatesh J, Upadhyaya CP, Nookaraju A, Pandey SK, Park SW (2012) Plant disease resistance genes: Current status and future directions. Physiol Mol Plant Pathol 78:51-65 doi:10.1016/j.pmpp.2012.01.002

Hartman G, Domier L, Wax L, Helm C, Onstad D, Shaw J, Solter L, Voegtlin D, d'Arcy C, Gray M (2001) Occurrence and distribution of *Aphis glycines* on soybeans in Illinois in 2000 and its potential control. Plant Health Progr

He C-Y, Tian A-G, Zhang J-S, Zhang Z-Y, Gai J-Y, Chen S-Y (2003) Isolation and characterization of a full-length resistance gene homolog from soybean. Theor Appl Genet 106:786-793

Hill C, Chirumamilla A, Hartman G (2012) Resistance and virulence in the soybean-*Aphis glycines* interaction. Euphytica 186:635-646

Howe EA, Sinha R, Schlauch D, Quackenbush J (2011) RNA-Seq analysis in MeV. Bioinformatics 27:3209-3210

Hulbert SH, Craig A. Webb, Shavannor M. Smith, Sun Q (2001) Resistance Gene Complexes. Annu Rev Phytopathol 39:285-312

Jain S, Chittem K, Brueggeman R, Osorno JM, Richards J, Nelson Jr BD (2016) Comparative Transcriptome Analysis of Resistant and Susceptible Common Bean Genotypes in Response to Soybean Cyst Nematode Infection. PloS one 11:e0159338

Jones JD, Dangl JL (2006) The plant immune system. Nature 444:323-329 doi:10.1038/nature05286

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236-1240

Jun T-H, Mian MR, Michel AP (2012) Genetic mapping revealed two loci for soybean aphid resistance in PI 567301B. Theor Appl Genet 124:13-22

Jupe F, Pritchard L, Etherington GJ, MacKenzie K, Cock PJ, Wright F, Sharma SK, Bolser D, Bryan GJ, Jones JD (2012) Identification and localisation of the NB-LRR gene family within the potato genome. BMC Genomics 13:1

Kang YJ, Kim KH, Shim S, Yoon MY, Sun S, Kim MY, Van K, Lee S-H (2012) Genome-wide mapping of NBS-LRR genes and their association with disease resistance in soybean. BMC Plant Biol 12:1

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647-1649

Kim K-S, Bellendir S, Hudson KA, Hill CB, Hartman GL, Hyten DL, Hudson ME, Diers BW (2010a) Fine mapping the soybean aphid resistance gene *Rag1* in soybean. Theor Appl Genet 120:1063-1071

Kim K-S, Hill CB, Hartman GL, Hyten DL, Hudson ME, Diers BW (2010b) Fine mapping of the soybean aphid-resistance gene *Rag2* in soybean PI 200538. Theor Appl Genet 121:599-610

Kohler A, Rinaldi C, Duplessis S, Baucher M, Geelen D, Duchaussoy F, Meyers BC, Boerjan W, Martin F (2008) Genome-wide identification of NBS resistance genes in *Populus trichocarpa*. Plant Mol Biol 66 doi:10.1007/s11103-008-9293-9

Kozomara A, Griffiths-Jones S (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 42: D68-D73

Kuang H, Caldwell KS, Meyers BC, Michelmore RW (2008) Frequent sequence exchanges between homologs of *RPP8* in *Arabidopsis* are not necessarily associated with genomic proximity. Plant J 54:69-80

Kumar S, Stecher G, Tamura K (2016) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol Biol Evol: 33, 1870-1874

Kuwabara C, Arakawa K, Yoshida S (1999) Abscisic acid-induced secretory proteins in suspension-cultured cells of winter wheat. Plant Cell Physiol 40:184-191

Lahaye T (2002) The *Arabidopsis RRS1-R* disease resistance gene–uncovering the plant's nucleus as the new battlefield of plant defense? Trends Plant Sci 7:425-427

Larkin MA, Blackshields G, Brown N, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947-2948

Lavin M, Herendeen PS, Wojciechowski MF (2005) Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the Tertiary. Syst Biol 54:575-594

Lawrence GJ, Finnegan EJ, Ayliffe MA, Ellis JG (1995) The L6 gene for flax rust resistance is related to the *Arabidopsis* bacterial resistance gene *RPS2* and the tobacco viral resistance gene *N*. Plant Cell 7:1195-1206

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

Li J, Ding J, Zhang W, Zhang Y, Tang P, Chen JQ, Tian D, Yang S (2010) Unique evolutionary pattern of numbers of gramineous NBS-LRR genes. Mol Genet Genomics 283:427-438 doi:10.1007/s00438-010-0527-6

Li X, Kapos P, Zhang Y (2015) NLRs in plants. Curr Opin Immunol 32:114-121 doi:10.1016/j.coi.2015.01.014

Lozano R, Hamblin MT, Prochnik S, Jannink JL (2015) Identification and distribution of the NBS-LRR gene family in the Cassava genome. BMC Genomics 16:360 doi:10.1186/s12864-015-1554-9

Lozano R, Ponce O, Ramirez M, Mostajo N, Orjeda G (2012) Genome-wide identification and mapping of NBS-encoding resistance genes in *Solanum tuberosum* group phureja. PLoS One 7:e34775

Luo S, Zhang Y, Hu Q, Chen J, Li K, Lu C, Liu H, Wang W, Kuang H (2012) Dynamic nucleotide-binding site and leucine-rich repeat-encoding genes in the grass family. Plant Physiol 159:197-210 doi:10.1104/pp.111.192062

Marchler-bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI (2015) CDD: NCBI's conserved domain database. Nucleic Acids Res 43:D222-226

Marone D, Russo MA, Laido G, De Leonardis AM, Mastrangelo AM (2013) Plant nucleotide binding site-leucine-rich repeat (NBS-LRR) genes: active guardians in host defense responses. Int J Mol Sci 14:7302-7326 doi:10.3390/ijms14047302

McClean PE, Mamidi S, McConnell M, Chikara S, Lee R (2010) Synteny mapping between common bean and soybean reveals extensive blocks of shared loci. BMC Genomics 11:184

Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW (2003) Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. Plant Cell 15 doi:10.1105/tpc.009308

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

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

Milos V, Vuk D, Kristina P, Jegor M (2013) Review of soybean resistance to pathogens. Ratarstvo i povrtarstvo 50:52-61 doi:10.5937/ratpov50-4038

Mun J-H, Yu H-J, Park S, Park B-S (2009) Genome-wide identification of NBS-encoding resistance genes in *Brassica rapa*. Mol Genet Genomics 282:617-631

Namayanja A, Buruchara R, Mahuku G, Rubaihayo P, Kimani P, Mayanja S, Eyedu H (2006) Inheritance of resistance to angular leaf spot in common bean and validation of the utility of resistance linked markers for marker assisted selection out side the mapping population. Euphytica 151:361-369

Nandety RS, Caplan JL, Cavanaugh K, Perroud B, Wroblewski T, Michelmore RW, Meyers BC (2013) The role of TIR-NBS and TIR-X proteins in plant basal defense responses. Plant Physiol 162:1459-1472

Nepal MP, Benson BV (2015) CNL disease Resistance Genes in Soybean and Their Evolutionary Divergence. Evol Bioinform Online 11:49-63 doi:10.4137/EBO.S21782

Noël L, Moores TL, Van Der Biezen EA, Parniske M, Daniels MJ, Parker JE, Jones JD (1999) Pronounced intraspecific haplotype divergence at the *RPP5* complex disease resistance locus of *Arabidopsis*. Plant Cell 11:2099-2111

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

Pan Q, Liu Y-S, Budai-Hadrian O, Sela M, Carmel-Goren L, Zamir D, Fluhr R (2000) Comparative genetics of nucleotide binding site-leucine rich repeat resistance gene homologues in the genomes of two dicotyledons: tomato and *Arabidopsis*. Genetics 155:309-322

Peart JR, Mestre P, Lu R, Malcuit I, Baulcombe DC (2005) NRG1, a CC-NB-LRR protein, together with N, a TIR-NB-LRR protein, mediates resistance against tobacco mosaic virus. Curr Biol 15:968-973

Peele HM, Guan N, Fogelqvist J, Dixelius C (2014) Loss and retention of resistance genes in five species of the Brassicaceae family. BMC Plant Biol 14:1

Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8:785-786

Pfeil B, Schlueter JA, Shoemaker R, Doyle J (2005) Placing paleopolyploidy in relation to taxon divergence: a phylogenetic analysis in legumes using 39 gene families. Sys Biol 54:441-454

Plocik A, Layden J, Kesseli R (2004) Comparative analysis of NBS domain sequences of NBS-LRR disease resistance genes from sunflower, lettuce, and chicory. Mol Phylogenet Evol 31:153-163

Pontes O, Neves N, Silva M, Lewis MS, Madlung A, Comai L, Viegas W, Pikaard CS (2004) Chromosomal locus rearrangements are a rapid response to formation of the allotetraploid *Arabidopsis suecica* genome. Proc Natl Acad Sci USA 101:18240-18245

Rathi D, Gayen D, Gayali S, Chakraborty S, Chakraborty N (2016) Legume proteomics: Progress, prospects, and challenges. Proteomics 16:310-327

Ratnaparkhe MB, Wang X, Li J, Compton RO, Rainville LK, Lemke C, Kim C, Tang H, Paterson AH (2011) Comparative analysis of peanut NBS-LRR gene clusters suggests evolutionary innovation among duplicated domains and erosion of gene microsynteny. New Phytol 192:164-178

Reams AB, Neidle EL (2004) Selection for gene clustering by tandem duplication. Annu Rev Microbiol 58:119-142

Rozas J (2009) DNA sequence polymorphism analysis using DnaSP. Bioinformatics for DNA sequence analysis:337-350

Sartorato A, Nietsche S, Barros EG, Moreira MA (2000) RAPD and SCAR markers linked to resistance gene to angular leaf spot in common beans. Fitopatol Bras 25:637-642

Savard L, Li P, Strauss SH, Chase MW, Michaud M, Bousquet J (1994) Chloroplast and nuclear gene sequences indicate late Pennsylvanian time for the last common ancestor of extant seed plants. Proc Natl Acad Sci USA 91 doi:10.1073/pnas.91.11.5163

Schatz MC, Witkowski J, McCombie WR (2012) Current challenges in de novo plant genome sequencing and assembly. Genome Biol 13:243

Schlueter JA, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Shoemaker RC (2004) Mining EST databases to resolve evolutionary events in major crop species. Genome 47:868-876

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178-183

Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, Jenkins J, Shu S, Song Q, Chavarro C (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet 46:707-713

Schulze-Lefert P, Panstruga R (2011) A molecular evolutionary concept connecting nonhost resistance, pathogen host range, and pathogen speciation. Trends Plant Sci 16:117-125

Shao Z-Q, Xue J-Y, Wu P, Zhang Y-M, Wu Y, Hang Y-Y, Wang B, Chen J-Q (2016) Large-scale analyses of angiosperm nucleotide-binding site-leucine-rich repeat (NBS-LRR) genes reveal three anciently diverged classes with distinct evolutionary patterns. Plant Physiol: 01487

Shao Z-Q, Zhang Y-M, Hang Y-Y, Xue J-Y, Zhou G-C, Wu P, Wu X-Y, Wu X-Z, Wang Q, Wang B (2014) Long-term evolution of nucleotide-binding site-leucine-rich repeat genes: understanding gained from and beyond the legume family. Plant Physiol 166:217-234

Shivaprasad PV, Chen H-M, Patel K, Bond DM, Santos BA, Baulcombe DC (2012) A microRNA superfamily regulates nucleotide binding site–leucine-rich repeats and other mRNAs. Plant Cell 24:859-874

Soderlund C, Bomhoff M, Nelson WM (2011) SyMAP v3. 4: a turnkey synteny system with application to plant genomes. Nucleic Acids Res:39, e68–e76.

Souza TLP, Faleiro FG, Dessaune SN, Paula-Junior TJd, Moreira MA, Barros EGd (2013) Breeding for common bean (*Phaseolus vulgaris* L.) rust resistance in Brazil. Trop Plant Pathol 38:361-374

Sprent JI (2007) Evolving ideas of legume evolution and diversity: a taxonomic perspective on the occurrence of nodulation. New Phytol 174:11-25

Stokes TL, Kunkel BN, Richards EJ (2002) Epigenetic variation in *Arabidopsis* disease resistance. Genes & Development 16:171-182

Takemoto D, Rafiqi M, Hurley U, Lawrence GJ, Bernoux M, Hardham AR, Ellis JG, Dodds PN, Jones DA (2012) N-terminal motifs in some plant disease resistance proteins function in membrane attachment and contribute to disease resistance. Mol Plant-Microbe Interact 25:379-392

Taylor JS, Raes J (2004) Duplication and divergence: the evolution of new genes and old ideas. Annu Rev Genet 38:615-643

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

Tian Y, Fan L, Thurau T, Jung C, Cai D (2004) The absence of TIR-type resistance gene analogues in the sugar beet (*Beta vulgaris* L.) genome. J Mol Evol 58:40-53

Van Der Biezen EA, Freddie CT, Kahn K, Jones JD (2002) *Arabidopsis RPP4* is a member of the *RPP5* multigene family of TIR-NB-LRR genes and confers downy mildew resistance through multiple signalling components. Plant J 29:439-451

Wan H, Yuan W, Bo K, Shen J, Pang X, Chen J (2013) Genome-wide analysis of NBS-encoding disease resistance genes in *Cucumis sativus* and phylogenetic study of NBS-encoding genes in Cucurbitaceae crops. BMC Genomics 14:1

Wei H, Li W, Sun X, Zhu S, Zhu J (2013) Systematic analysis and comparison of nucleotide-binding site disease resistance genes in a diploid cotton *Gossypium raimondii*. PLoS One 8:e68435 doi:10.1371/journal.pone.0068435

Whitham S, Dinesh-Kumar S, Choi D, Hehl R, Corr C, Baker B (1994) The product of the tobacco mosaic virus resistance gene *N*: similarity to toll and the interleukin-1 receptor. Cell 78:1101-1115

Whitkus R, Doebley J, Lee M (1992) Comparative genome mapping of Sorghum and maize. Genetics 132:1119-1130

Yang L, Li D, Li Y, Gu X, Huang S, Garcia-Mas J, Weng Y (2013) A 1,681-locus consensus genetic map of cultivated cucumber including 67 NB-LRR resistance gene homolog and ten gene loci. BMC Plant Biol 13:1

Yang S, Tang F, Gao M, Krishnan HB, Zhu H (2010) R gene-controlled host specificity in the legume–rhizobia symbiosis. Proc Natl Acad Sci USA 107:18735-18740

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

Yu J, Tehrim S, Zhang F, Tong C, Huang J, Cheng X, Dong C, Zhou Y, Qin R, Hua W, Liu S (2014) Genome-wide comparative analysis of NBS-encoding genes between *Brassica* species and *Arabidopsis thaliana*. BMC Genomics 15:1-18 doi:10.1186/1471-2164-15-3

Yue JX, Meyers BC, Chen JQ, Tian D, Yang S (2012) Tracing the origin and evolutionary history of plant nucleotide-binding site–leucine-rich repeat (NBS-LRR) genes. New Phytol 193:1049-1063

Zhai J, Jeong D-H, De Paoli E, Park S, Rosen BD, Li Y, González AJ, Yan Z, Kitto SL, Grusak MA (2011a) MicroRNAs as master regulators of the plant NB-LRR defense gene family via the production of phased, trans-acting siRNAs. Genes & development 25:2540-2553

Zhai J, Jeong DH, De Paoli E, Park S, Rosen BD, Li Y, Gonzalez AJ, Yan Z, Kitto SL, Grusak MA, Jackson SA, Stacey G, Cook DR, Green PJ, Sherrier DJ, Meyers BC (2011b) MicroRNAs as master regulators of the plant NB-LRR defense gene family via the production of phased, trans-acting siRNAs. Genes Dev 25:2540-2553 doi:10.1101/gad.177527.111

Zhang G, Gu C, Wang D (2009) Molecular mapping of soybean aphid resistance genes in PI 567541B. Theor Appl Genet 118:473-482

Zhang G, Gu C, Wang D (2010) A novel locus for soybean aphid resistance. Theor Appl Genet 120:1183-1191

Zhang Y, Xia R, Kuang H, Meyers BC (2016a) The diversification of plant NBS-LRR defense genes directs the evolution of microRNAs that target them. Mol Biol Evol 33:2692-2705

Zhang YM, Shao ZQ, Wang Q, Hang YY, Xue JY, Wang B, Chen JQ (2016b) Uncovering the dynamic evolution of nucleotide-binding site-leucine-rich repeat (NBS-LRR) genes in Brassicaceae. J Integr Plant Biol 58:165-177

Zheng F, Wu H, Zhang R, Li S, He W, Wong F-L, Li G, Zhao S, Lam H-M (2016) Molecular phylogeny and dynamic evolution of disease resistance genes in the legume family. BMC Genomics 17:1

Zhu H, Choi H-K, Cook DR, Shoemaker RC (2005) Bridging model and crop legumes through comparative genomics. Plant Physiol 137:1189-1196

CHAPTER 3: GENOME-WIDE IDENTIFICATION OF NBS-ENCODING

RESISTANCE GENES IN SUNFLOWER (*HELIANTHUS ANNUUS* L.)

This chapter has been published in the Journal *Genes*:

Neupane, S., Andersen, E., Neupane, A., and Nepal, M.P. (2018). Genome-Wide

Identification of NBS-Encoding Resistance Genes in Sunflower (*Helianthus annuus* L.).

Genes 9(8)

Abstract

Nucleotide Binding Site-Leucine-Rich Repeat (NBS-LRR) genes encode disease resistance proteins involved in plants' defense against their pathogens. Although sunflower is affected by many diseases, only a few molecular details have been uncovered regarding pathogenesis and resistance mechanisms. Recent availability of sunflower whole genome sequences in publicly accessible databases allowed us to accomplish a genome-wide identification of Toll-interleukin-1 receptor-like Nucleotide-binding site Leucine-rich repeat (TNL), Coiled Coil (CC)-NBS-LRR (CNL), Resistance to powdery mildew8 (RPW8)-NBS-LRR (RNL) and NBS-LRR (NL) protein encoding genes. Hidden Markov Model (HMM) profiling of 52,243 putative protein sequences from sunflower resulted in 352 NBS-encoding genes, among which 100 genes belong to CNL group including 64 genes with RX_CC like domain, 77 to TNL, 13 to RNL, and 162 belong to NL group. We also identified signal peptides and nuclear localization signals present in the identified genes and their homologs. We found that NBS genes were located on all chromosomes and formed 75 gene clusters, one-third of which were located on chromosome 13. Phylogenetic analyses between sunflower and *Arabidopsis* NBS genes revealed a clade-specific nesting pattern in CNLs, with RNLs nested in the

CNL-A clade, and species-specific nesting pattern for TNLs. Surprisingly, we found a moderate bootstrap support (BS = 50%) for CNL-A clade being nested within TNL clade making both the CNL and TNL clades paraphyletic. *Arabidopsis* and sunflower showed 87 syntenic blocks with 1049 synteny hits and high synteny between chromosome 5 of *Arabidopsis* and chromosome 6 of sunflower. Expression data revealed functional divergence of the NBS genes with basal level tissue-specific expression. This study represents the first genome-wide identification of NBS genes in sunflower paving avenues for functional characterization and potential crop improvement.

Keywords: coiled coil, disease resistance, nucleotide binding site encoding genes, gene clustering, plant defense, resistance pathways, resistance to powdery mildew 8, R genes, sunflower, synteny

3.1. Introduction

Plants employ different gene families in signaling networks in response to numerous biotic and abiotic stresses [1]. In order to deal with these stresses, during the course of evolution, plants have developed multifaceted processes to recognize the stress stimuli, transfer them to the plant's own message(s) and complete the signal transduction pathways [2, 3]. In response to the stresses due to pathogens, plants have developed race-specific and race non-specific resistance, known as qualitative and quantitative resistance, respectively [4]. Plants recruit proteins encoded by disease resistance (R) genes that recognize or interact with specific pathogen avirulence (*avr*) gene products [5] or effector molecules triggering a downstream signaling in resistance pathways [6, 7]. Various models, such as Zig-zag and multicomponent models, propose a dynamic relationship

between a host and its pathogen and explain how incompatible interactions between the

hosts and pathogens lead to a selection of new R genes in response to co-evolutionary

pressure due to pathogen races [6, 8]. Host R genes can vary within a species, and their

variation is correlated with that of the corresponding pathogen effectors [9]. For example,

host polymorphic to R genes is found to provide partial resistance against pathogens [10].

Such partial resistance accumulates throughout the plant development and eventually

provide quantitative resistance in the form of broad spectrum resistance [10].

Identification of R genes and their pathogen effectors is essential for understanding host–

pathogen interactions and disease resistance pathways in order to develop durable

resistance in crop species.

The Pathogen Recognition Genes database (PRGdb, http://prgdb.org) listed 153 R

genes that have been cloned and characterized, and 177,072 annotated candidate

Pathogen Receptor Genes (PRGs) [11]. These R genes encode mostly nucleotide binding

site (NBS) leucine-rich repeat (LRR) proteins and have been classified into categories

based upon the domains and motifs organization in the proteins [12, 13]. Most commonly

recognized categories are Toll-interleukin-1 receptor-like-NBS-LRR (TNL), Coiled-Coil-

NBS-LRR (CNL), and Resistance to powdery mildew8 (RPW8)-NBS-LRR (RNL) [13,

14]. All TNL, CNL and RNL genes are present in dicots, whereas TNL genes are absent

in monocots [14, 15]. Analysis of NBS genes in Fabaceae and Brassicaceae revealed that

CNLs and RNLs diverged prior to divergence of Rosid I and Rosid II lineages of

Angiosperms, and, in both plant families, the two clades are sister to each other [15, 16].

The NBS domain, also known as NB-ARC where ARC stands for APAF1 (apoptotic

protease-activating factor-1), R genes, and CED4 (*Caenorhabditis elegans* death-4

protein), hydrolyzes ATP to induce the conformational change in R proteins by acting as the nucleotide binding pocket [17]. The NBS domain mainly consists of P-loop, Kinase-2, RNBS A, GLPL and MHDL motifs [14]. The LRR domains at the C-terminus help activate or deactivate the defense signaling by interacting with the NBS domain in the presence or absence of pathogen effectors, respectively [18]. A diverse number of NBS genes have been reported in various plant species since the first study in *Arabidopsis thaliana* was published in 2003 [14]. With the increasing availability of plant genome sequences, R gene proteins have been identified in many plant species, such as *A. thaliana* [5, 14]; *Vaccinium* spp. [19]; *Amborella trichopoda*, *Musa acuminata*, *Phyllostachys heterocycla*, *Capsicum annuum*, and *Sesamum indicum* by Shao et al. 2016 [13]; *Cicer arietinum* [20]; *Glycine max* [21, 22, 23, 24]; *Oryza sativa* [25, 26]; *Medicago truncatula* [27]; *Vitis vinifera* and *Populus trichocarpa* by Yang et al. 2008 [28]; *Solanum tuberosum* [29]; *Brassica rapa* and *B. oleracea* by Zhang et al. 2016 [30]; *Hordeum vulgare* [31]; *Setaria italica [32]*; *Theobroma cacao* [5]; *Populus trichocarpa* [5]; *V. vinifera* [5]; *Cucumis sativus* [33]; *Phaseolus vulgaris* [16, 24], *Lotus japonicas*, *Cajanus cajan*, *Glycine soja* by Zheng et al. 2016, *Gossypium arboretum* [34], etc. A recent study by Li et al. 2016 [35] has identified NBS-encoding genes as well as receptor-like protein kinases (RLKs) and receptor-like proteins (RLPs), collectively called as Resistance Gene Analogs (RGAs), for 50 plant genomes using a RGAugury pipeline.

According to a report by Food and Agriculture Organization (FAO) in 2010 (http://www.fao.org), domesticated *Helianthus annuus* L. (Family Asteraceae), is the fourth most important oilseed crop in the world. Since sunflower has the capacity to maintain stable yields in different environmental conditions such as drought, it has been a

model crop species for studying climate change adaptation [36]. The study on diversity

analysis of 128 expressed sequenced tag (EST)-based microsatellites in wild *H. annuus*

has provided insights into the ability to adapt salt and drought stress and selective sweeps

revealing transcription factors as the major group of genes involved in those processes

[36]. In addition, studies on wild and cultivated relatives of sunflower on disease

resistance [37] and oil content [38] aspects have played great roles in understanding the

genetic background for these traits. However, many fungal diseases like charcoal rot

(*Macrophomina phseolina*), downy mildew (*Plasmopara halstedii*), *Fusarium* rot and

stem rots (*Fusarium* sp.), phoma black stem (*Phoma macdonaldii*), phomopsis stem

canker (*Diaporthe helianthi*, *D. gulyae*), *Sclerotinia* mid and basal stem rot (*Sclerotinia*

scelerotiorum), *Verticillium* wilt (*Verticillium dahlia*), leaf blight (*Alternariaster*

helianthi), leaf spot (*Pseudomonas syringae* pv. *helianthi*), powdery mildew (*Erysiphe*

cichoracearum), rust (*Puccinia helianthi*) and many others have caused crop damage

resulting in the loss of yield and oil content [39].

Previously, various studies have contributed their findings about the NBS group

of R genes in sunflower [40, 41, 42, 43]. Plocik et al. 2004 [40] identified nine sunflower

resistance gene candidates with coiled-coil (CC) domains in the N-terminal region using

degenerate primer sets. Sunflower showed diverse structures in CC subfamily, while

lettuce and chicory, closely related species, showed high similarity in structure. Radwan

et al. 2008 [42] used degenerate primers to identify 630 NBS-LRR homologs in wild

species of sunflower (*Helianthus annuus*, *H. argophyllus*, *H. deserticola*, *H. paradoxus*,

and *H. tuberosus*). In addition, *Radwan et al. 2004 [43]* isolated R gene analogs

belonging to the CNL class of R genes from the inbred sunflower line QIR8 containing

Pl8I locus against *P. halstedii*, which causes downy mildew. Later, Hewezi et al. 2006 [41] cloned partial antisense *PLFOR48*, which showed homology to the TNL family, in mildew resistant sunflower line, RHA 266 and *Nicotiana tabacum* L. The recent availability of the *H. annuus* genome [44] has now made it possible for studying the diversity and evolution of gene families in sunflower. The main objectives of this research were to conduct a genome-wide search for *H. annuus* NBS genes and analyze their genomic structure and functions. A proper identification of the R genes is crucial to elucidate their roles against various diseases in sunflower.

3.2. Materials and Methods

3.2.1. Retrieval and Identification of Sunflower NBS-Encoding Genes

The genome of sunflower (INRA inbred genotype XRQ described in [44]; *H. annuus* r1.2) was accessed from the sunflower genome database (https://www.sunflowergenome.org) as well as Phytozome (https://phytozome.jgi.doe.gov). The sunflower genome is 3.6 gigabases and its genes distributed over 17 chromosomes encode 52,243 proteins (https://phytozome.jgi.doe.gov). *A. thaliana* TNL and CNL (= nonTNL or nTNL) protein sequences were used as reference for the identification of NBS-LRR proteins in sunflower, and were obtained from http://niblrrs.ucdavis.edu. The multiple sequence alignment file of these reference sequences in Stockholm format were employed in hmmbuild and hmmsearch for HMM profiling using the program HMMER version 3.1b2 [45] at a cut-off value of 0.01. InterProScan Version 5.27 (EMBL-European Bioinformatics Institute, UK) [46] and Pfam ID [47] and PROSITE ID (http://prosite.expasy.org/) were used to search for the conserved domains. The proteins

with PfamID of TIR (PF01582), NBS (PF00931), RPW8 (PF05659), CC and LRR

domains with 'LxxLxxLxx' signatures were selected to determine the NBS proteins in

sunflower. Further verification of the CC domains at the N-terminus was carried out

using the MARCOIL server [48] with a 9FAM matrix having probability between 0.4–1.

Multiple expectation maximization for motif elicitation (MEME) [49] analysis was

performed to confirm the presence of P-loop, Kinase-2, GLPL, MHDL, RNBS A, RNBS

B, RNBS C, and RNBS D motifs in the NBS domain, TIR1, TIR2, TIR3 motifs in TIR

domain and RPW8 motifs in RPW8 domains. A set of parameters used in MEME

analysis included maxsize: 100,000, mod: zoops, nmotifs: 20, minw: 6, and maxw: 50 to

25. Subcellular localization of the putative NBS genes were analyzed using TargetP 1.1

[50]. The program NLStradamus [51] was used to examine nuclear localization signals

(NLS) in identified NBS genes of sunflower using a two-state HMM static model with

Viterbi and posterior prediction methods (with 0.5 cut-off).

3.2.1. Phylogenetic Tree Construction

The NBS protein sequences from *A. thaliana* and *H. annuus* were aligned using

CLUSTAL W [52] and MUSCLE [53] integrated in the program Geneious [54].

Phylogenetic analysis of the aligned data matrix was performed using Maximum

Likelihood (ML) method (1000 replicates) in the program MEGA Version 7.0.14 [55].

The phylogenetic analysis employed the best evolutionary model (resulted from the

ModelTest analysis using MEGA7) JTT + G + I (Jones–Taylor–Thornton with γ

distribution and invariant sites), and *Streptomyces coelicolor* accession P25941 as an

outgroup [14]. Additional phylogenetic trees of the NBS domains of predicted TNL and

CNL proteins of sunflower and all reference proteins obtained from http://prgdb.crg.eu

were reconstructed using the methods and models described above. Thus, the obtained Newick format of phylogenetic trees were employed in the Interactive tree of life (iTOL) Version 3 (Biobyte solutions GmbH, Bothestr, Germany) for their visual enhancement [56].

3.2.2. Chromosomal Locations, Clustering and Gene Structure

All 17 chromosome sequences of *H. annuus* were obtained from https://www.sunflowergenome.org and uploaded in the program Geneious [54]. The chromosome locations of the respective gene families were visualized using an annotation file in Generic Feature Format (GFF). The NBS gene locations, NBS types and clustering were visualized on their respective chromosomes. Gene clustering of the NBS genes was carried out following Jupe et al. 2012 [57], using two major criteria: (a) distance between two NBS genes is less than 200 kb, and (b) presence of no more than eight annotated non-NBS sequences between two consecutive NBS sequences. The exon-intron distribution pattern was obtained by the Gene Structure Display Server (http://gsds.cbi.pku.edu.cn).

3.2.3. *Ka/Ks* and Syntenic Analysis

Coding sequences (CDS) of the NBS genes were used in calculating nonsynonymous substitutions per nonsynonymous site (K_a) and synonymous substitutions per synonymous site (K_s) in the program DnaSP 6.11.01 [58]. Syntenic map of the NBS genes of *H. annuus* and *A. thaliana* was created using SyMAP Version 4.2 (Arizona Genomics Computational Lab, Tucson, AZ, USA) [59], executed within South Dakota State University's High-Performance Computing Cluster. Whole chromosome

sequences and gene annotation files were used as input files for syntenic mapping of *H. annuus* and *A. thaliana*.

3.2.4. Gene Homology and Expression Analysis

Putative homologs of the predicted sunflower NBS genes were accessed using BLAST tool available in http://prgdb.org with reference genes of PRGdb and a cutoff E-value of 0.01. The filtering included sequences with E-values less than 0.01 and identity percentage of greater than 50%. Expression profiles of the putative NBS genes were downloaded from https://www.sunflowergenome.org. A heatmap was generated using deseq normalized data through the MeV package, available at http://mev.tm4.org/ [60]. The heatmap clustering was performed based on Euclidean distance under 1000 iterations using the K-means Clustering Method. The clustering classification used these categories: moderate to minimal expression, minimal expression to no expression, and no expression at all.

3.3. Results

3.3.1. Diversity of the NBS-Encoding Genes in Sunflower

The HMM analysis of all sunflower protein-coding genes using the reference sequences of *A. thaliana* resulted in 485 NBS proteins, using a filtering threshold expectation value of 0.01. These sequences were further annotated with InterProscan, and evaluated for the presence of NBS domains in each sequence. After a careful examination, 352 protein sequences were confirmed to have an NBS domain. Among these, 100 genes belonging to CNL group (after verification using MARCOIL server omitted ten false positives), 77 to TNL, 13 to RNL group, as well as 162 genes possess neither CC nor TIR domains thus classified as an NL group. Among 100 CNL types, 64

possesses a CC domain similar to *S*. tuberosum disease resistance protein (Rx).

Furthermore, Leucine-rich repeats (LxxLxxLxx signatures) were examined to classify

CNLs, TNLs, RNLs and NLs into their subgroups. Following the classification of NBS-

encoding genes in *Brassica* species and *A. thaliana* [5], the NBS genes were classified

into: CC-NBS-LRR (CNL), CC-NBS (CN), CC-NBS-NBS-LRR (CNNL), CC-NBS-

NBS (CNN), RPW8-NBS-LRR (RNL), RPW8-NBS (RN), RPW8-CC-NBS-LRR

(RCNL), TIR-NBS-LRR (TNL), TIR-NBS (TN), TIR-TIR-NBS-LRR (TTNL), TIR-

NBS-LRR-TIR-NBS-LRR (TNLTNL), TIR-CC-NBS-LRR (CTNL), TIR-CC-NBS

(CTN), NBS (N), NBS-LRR (NL), NBS-NBS (NN), and NBS-NBS-LRR (NNL) (see

Table 3.1, Figures S1–S4). The LxxLxxLxx (=LRRs) signatures were present in 97 (out

of 100) CNL genes with their LRRs ranging from two to 22, 12 (of 13) RNL genes with

one to eight LRRs, 55 (of 77) TNL genes with two to 26 LRRs, and 131 (of 162) NL

genes with two to 30 LRRs. Among them, HanXRQChr02g0052061, a TNL protein

sequence contained a unique Kelch motif sequence (PF01344). TargetP analysis showed

that 20 NBS proteins were predicted to localize to the chloroplast, 14 to mitochondria, 80

enter the secretory pathway, and 238 were predicted to enter other subcellular locations

other than mitochondria or the chloroplast (Table S1). Thirteen CNLs, seven TNLs, one

RNL, and eight NLs were identified to contain a putative NLS using NLStradamus

(Table S2).

 Three major signature motifs: P-loop, Kinase-2, and GLPL of the NBS domain of

disease resistance proteins were present in 57 out of 100 CNLs, 69 out of 77 TNLs, all 13

RNLs and 58 out of 162 NLs (Supplementary File S1, Figures S5–S7). Other important

motifs RNBS A, RNBS B, RNBS C and RNBS D, and MHDL were also present in the

NBS proteins (Tables S3–S5). Motifs TIR1, TIR2, TIR3, and TIR4 varied in number across the TNL genes: among the 77 TNLs, 76 had TIR1, 76 had TIR2, 75 had TIR3 and 76 had TIR4 motifs. Only two TNLs (HanXRQChr05g0136351 and HanXRQChr06g0184071) did not have all four TIR motifs. Of the 100 CNLs, 81 had the characteristic conserved amino acid sequence 'DDVW' in the Kinase-2 motif. Remaining CNLs had either Isoleucine (I), Methionine (M), or Leucine (L) in the place of Valine (V) amino acid in the sequence 'DDVW'. Of the 77 TNLs, 50 shared the characteristic 'DDVD' amino acid sequence in the Kinase-2 motif. Of the 162 NLs, 83 had 'DDVW' and 18 had 'DDVD', hence classified as N_{CC} and N_{TIR} group of the NLs, respectively. All of the 13 RNLs had 'DDVW' sequence in the Kinase-2 motif except for HanXRQChr03g0067681 with 'DDVR' sequence. Another key characteristic found within the RNBS B motif was that the majority of the CNLs had 'TSR', TNLs had 'TTRD', and RNLs had 'TSR' residues. The sequence alignments illustrating all the conserved motifs of the CNLs, TNLs, and RNLs are presented in Supplementary File S2.

3.3.2. Gene Location, Clustering, K_a/K_s Values and Structural Variation

The NBS genes are located on each of the chromosomes, with only four (*HanXRQChr00c0003g0570971*, *HanXRQChr00c0003g0570951*, *HanXRQChr00c0004g0571011*, and *HanXRQChr00c0037g0571241*) were not assigned to any chromosome (Figure S8). The number of the NBS genes located on each chromosome ranged from three (chromosome Ha12) to 99 (chromosome Ha13). Chromosomal distribution of the CNL, TNL, RNL, and NL genes and their clusters are shown in Figure 3.1. The CNL genes were absent in chromosomes Ha3, Ha5, and Ha16, whereas, TNL genes were absent in chromosomes Ha7 and Ha11. Most of the TNL genes

were uniformly distributed across the chromosomes, whereas most of the CNL and NL genes were densely represented on chromosome Ha13 (approximately 28%). The smallest number of RNL genes (thirteen) were present in chromosomes Ha2, Ha3, Ha4, Ha5, Ha7, Ha11, Ha14, and Ha15 (see Figure 3.1). Among the 352 NBS genes, 200 (~57%) genes formed 75 clusters (4.4 clusters per chromosome and 2.7 genes per cluster) with chromosome Ha13 hosting 25 clusters of 73 genes (~37%; Table S6). The gene clusters were present in all chromosomes except for Ha5 and Ha12. Gene positions and clusters on chromosomes of *H. annuus* are shown in Figure 3.2. The average K_a/K_s values for the clades of CNLs, TNLs, and RNLs were 0.68, 0.89, and 0.31, respectively. The number of exons in the genes is shown in Table S1 and Figures S9–S12. The number of exons for CNLs, TNLs, RNLs, and NLs ranged from 1 to 11, 2 to 18, 4 to 9, and 1 to 19, respectively. In average CNLs, TNLs, RNLs, and NLs had 2.7, 6.1, 6.2, and 2.9 exons per gene, respectively.

	Ha1	Ha2	Ha3	Ha4	Ha5	Ha6	Ha7	Ha8	Ha9	Ha10	Ha11	Ha12	Ha13	Ha14	Ha15	Ha16	Ha17
■ Clusters	6	6	1	6	0	3	1	5	4	4	3	0	25	1	3	2	5
■ NL group	5	10	4	12	3	5	2	10	10	6	10	0	50	7	5	6	13
■ RNL group	0	2	2	1	1	0	3	0	0	0	1	0	0	1	2	0	0
■ TNL group	4	4	8	12	3	5	0	6	19	3	0	1	2	1	1	7	1
■ CNL group	8	7	0	2	0	1	1	3	1	7	4	2	47	2	12	0	3

■ CNL group   ■ TNL group   ■ RNL group   ■ NL group   ■ Clusters
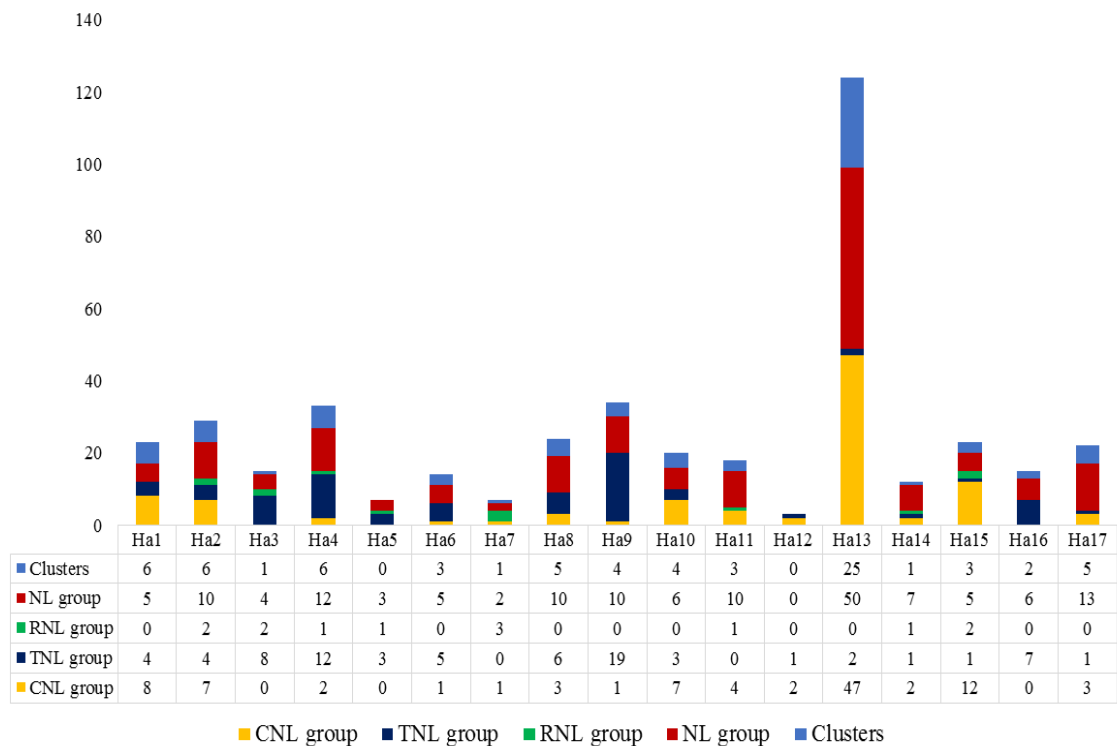
Figure 3.1. Chromosomal distribution (Ha1–Ha17) of the NBS genes and gene clusters in sunflower. Different NBS groups and gene clusters are color coded. CNL: Coiled-Coil-NBS-LRR; TNL: Toll-interleukin-1 receptor-like-NBS-LRR, RNL: Resistance to powdery mildew8 (RPW8)-NBS-LRR; NL: Nucleotide Binding Site—Leucine-Rich Repeat (NBS-LRR).

Table 3.1 Nucleotide Binding Site (NBS)-encoding proteins in sunflower in relation to 15 other plant species.

Predicted protein domains	Letter code	Number of proteins															
		Ha*	At[a]	Gm[b,c]	Mt[a]	Bo[a]	Br[a]	Tc[a]	Pt[a]	Vv[a]	Ca[d]	Cs[e]	Pv[f,c]	Lj[f]	Cc[f]	Gs[f]	Ga[h]
CC-NBS-LRR	CNL	92	17	95	152	6	19	82	120	203	19	17	31	11	37	47	80
CC-NBS	CN	5	8		25	5	15	46	14	26	33	1	40	26	41	62	44
CC-NBS-NBS-LRR	CNNL	4	-	5			2				1						
CC-NBS-NBS	CNN	1															
RPW8-NBS-LRR	RNL	10	2	6		1	4				2	2					3
RPW8NBS	RN	1	3			2	1				2						0
RPW8-CC-NBS-LRR	RCNL	2															
TIR-NBS-LRR	TNL	52	79	126	118	40	93	8	78	97	6	11	81	16	47	49	5
TIR-NBS	TN	21	17	22	38	29	23	4	10	14	7	2	11	53	36	76	2
TIR-NBS-NBS-LRR	TNNL	0	1			1	4										
TIR-TIR-NBS-LRR	TTNL	1									1						
TIR-NBS-LRR-TIR-NBS-LRR	TNLT-NL	1															
TIR-CC-NBS-LRR	CTNL	1															
TIR-CC-NBS	CTN	1															
NBS	N	29	26	4	328	53	29	53	62	36	14	1	59	82	136	213	59
NBS-LRR	NL	125	20	73		24	27	104	132	159	12	23	20	18	56	58	53
NBS-NBS	NN	2				3	2				1						
NBS-NBS-LRR	NNL	6					3										

Note: Ha: *Helianthus annuus*; At: *Arabidopsis thaliana*; Gm: *Glycine max*; Mt: *Medicago truncatula*; Bo: *Brassica oleracea*; Br*: Brassica rapa*; Tc*: Theobroma cacao*; Pt*: Populus trichocarpa*; Vv*: Vitis vinifera*; Ca*: Cicer arietinum*; Cs*: Cucumus sativus*; Pv*: Phaseolus vulgaris*; Lj*: Lotus japonicas*; Cc*: Cajanus cajan*; Gs*: Glycine soja*; Ga*: Gossypium arboretum* (* = this study, [a] = [5], [b] = [23], [c] = [24], [d] = [20], [e] = [33], [f] = [16], [g] = [34]).

3.3.3. Phylogenetic and Syntenic Analysis

The data matrix with the NBS aligned sequences (NBS domain region is more conserved than remaining 5' and 3' regions) was used in phylogenetic analyses. Phylogenetic relationships among the sunflower NBS sequences are shown in Figure 3.3, and those of the sunflower and *Arabidopsis* NBS sequences are shown in Figure 3.4; each tree reveals distinct clades of CNLs, RNLs and TNLs. The RNL clade was surprisingly nested within the TNL clade. As shown in Figure 3.3, the CNLs and TNLs formed six subclades each. The TNL subclades are named TIR (I), TIR (II), TIR (III), TIR (IV), TIR (V), and TIR (VI), whereas CNL subclades are named CC (I), CC (II), CC (III), CC (IV), CC (V), and CC (VI). The phylogenetic tree reconstructed using sunflower and *Arabidopsis* NBS sequences revealed clade-specific nesting patterns in the CNL group (Figure 3.4). The nesting of all sunflower RNL genes within CNL-A clade (with *Arabidopsis* RPW8 genes) was strongly supported (bootstrap support = 96%). CNL-C (I) clade constituted six CNL genes (*HanXQRChr14g0440091*, *HanXQRChr17g0562451*, *HanXQRChr12g0374601*, *HanXQRChr08g0224171*, *HanXQRChr13g0417971*, and *HanXQRChr13g0417981*) with a weak support [bootstrap support (BS) = 57%]. CNL-C (I) clade, sister clade to CNL-C (II) and CNL-D constituted 79 genes. CNL-B clade constituted three genes (*HanXQRChr02g0046161*, *HanXQRChr11g0333001*, and *HanXQRChr11g0333091*). The remaining 12 genes did not belong to any clade of *Arabidopsis* CNL genes. The TNL group formed a species-specific clade, except ten genes that formed a small clade with AT5G36930, named TNL-D clade with strong bootstrap support of 100%. We found a moderate bootstrap support (BS = 50%) for

CNL-A clade being nested within TNL clade making both the CNL and TNL clades paraphyletic. Another tree constructed using RNL genes of *A. thaliana* and *H. annuus* showed two distinct clades for two lineages: activated disease resistance gene 1 (*ADR1*) and N-required gene 1 (*NRG1*) (Figure 3.5). The Newick files related to phylogenetic trees in Figures 3.3–3.5 are provided in Supplementary File S3. For the comparative study, all the manually curated TNL and CNL reference proteins obtained from http://prgdb.crg.eu were phylogenetically compared with sunflower TNL and CNL NBS proteins. The sunflower NBS proteins formed clades with various reference proteins such as Pi36, Pl8, Rps2, VAT, RPG1, Gro1.4, RY-1, and N proteins suggesting their homologs (Figure S13). The syntenic relationship between the *Arabidopsis*'s 119,146 kb genome and sunflower's 3,641,596 kb genome showed 87 syntenic blocks with 1049 synteny hits. The chromosome 2 of *Arabidopsis* was highly syntenic to chromosome Ha1, Ha2, Ha3, and Ha15 chromosomes of sunflower. Similarly, the highest syntenic region was observed between chromosomes 5 of *Arabidopsis* and chromosome 6 of sunflower. The sunflower chromosomes Ha2, Ha5, Ha11, Ha13, Ha15, and Ha17 are least syntenic to any of the chromosome of *Arabidopsis*. The pericentromeric region of the sunflower chromosomes Ha3, Ha9, and Ha14 were highly syntenic to the chromosomes of *Arabidopsis*. The chromosome Ha13 that contains 99 NBS genes contains fragments from only chromosome 2 of *Arabidopsis* (Figure S14).
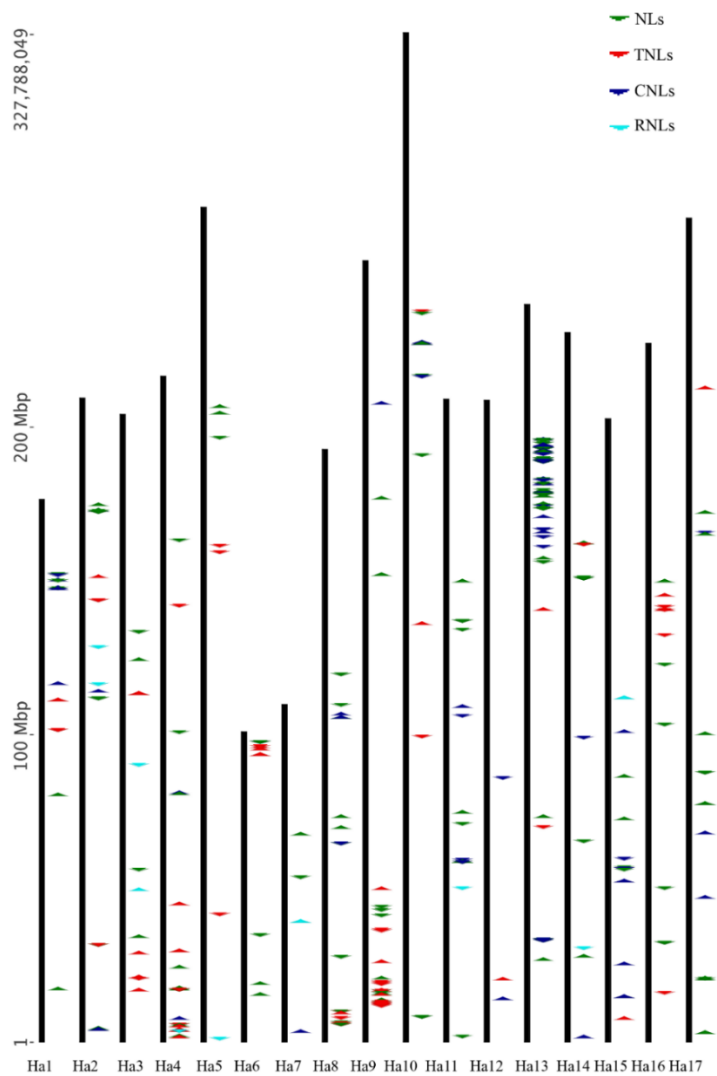
Figure 3.2. Chromosomal distribution of sunflower NBS gene clusters (*n* = 17). Each arrow color represents an NBS gene type and orientation, and the thick vertical line represents a chromosome.
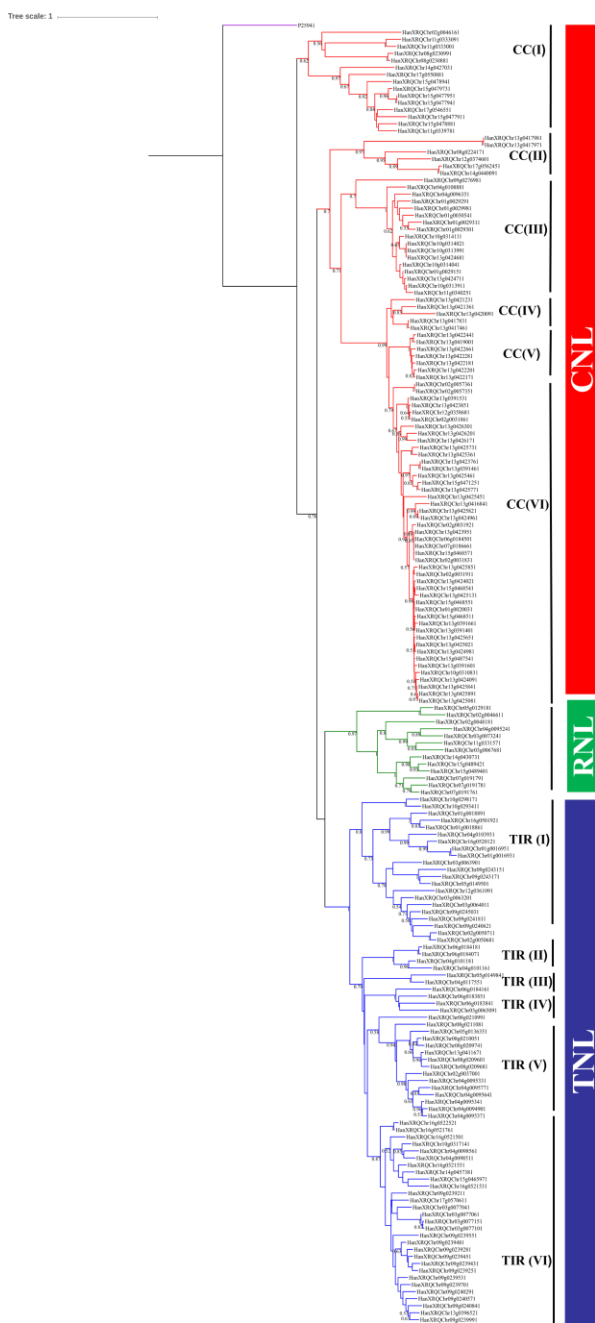
Figure 3.3. Maximum likelihood (ML) tree featuring NBS groups based on the conserved domains of the CNL, TNL, and RNL genes from *Helianthus annuus*. The ML tree was constructed using the JTT + G + I (Jones–Taylor–Thornton with γ distribution and invariant sites) model with 1000 bootstrap replicates. The ML tree was rooted using a *Streptomyces coelicolor* NBS containing protein, P25941, as an outgroup. The clades TNL (blue), CNL (red), and RNL (green) and outgroup (purple) are color-coded. Subclades are mentioned as TIR (I) to TIR (VI) and CC (I) to CC (VI).
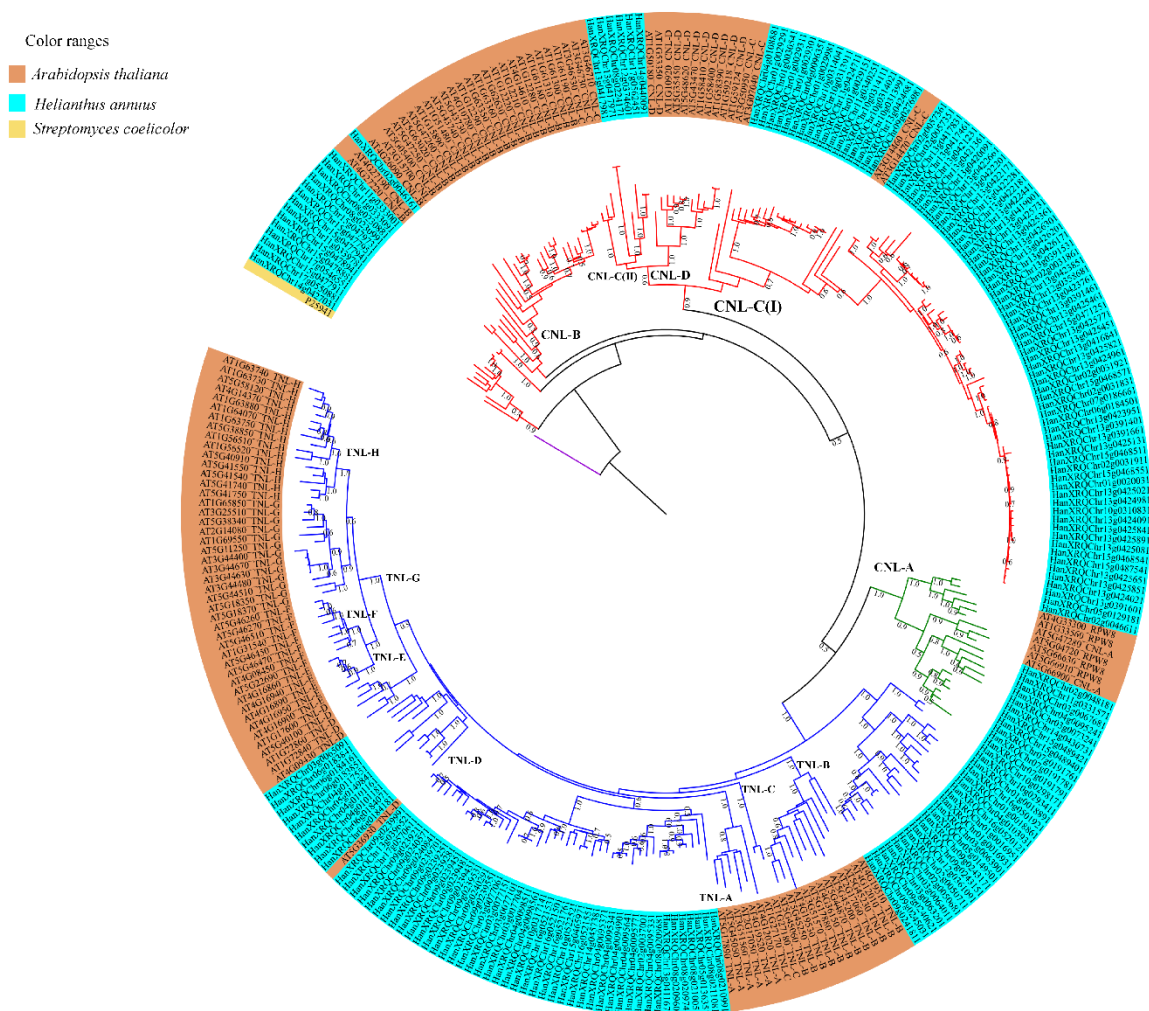
Figure 3.4. Maximum likelihood (ML) tree featuring NBS domain amino acid sequences of the CNL, TNL, and RNL genes from Arabidopsis thaliana (AT; orange) and *Helianthus annuus* (light blue). The ML tree was reconstructed using JTT + G + I (Jones–Taylor–Thornton with γ distribution and invariant sites) evolutionary model with 1000 bootstrap replicates. The ML tree was rooted using *Streptomyces coelicolor* NBS-containing protein, P25941, as an outgroup (yellow). The clades are color-coded: TNL in blue, CNL in red, RNL clade in green, and outgroup in purple. Subclades are labeled as CNL-A to CNL-D and TNL-A to TNL-H.

3.3.4 Homologs and Expression Analysis

The predicted 352 NBS proteins of sunflower showed homology, with identity greater than 50% and E-value less than 0.01, to 39 genes among 153 reference genes on the Plant Resistance Genes database (Table S7). Among them, 21 proteins showed greater than 70% identity to the *H. annuus* clone Ha-NTIR11g CC-NBS-LRR gene (*Pl8*). HanXRQChr13g0425411, HanXRQChr13g0425361, and HanXRQChr13g0425431 showed more than an 80% identity to the *Pl8* gene suggesting the probable homologs to that gene. HanXRQChr04g0123041, belonging to the NL group has shown homology to *Lycopersicon esculentum* EIX receptor 1 (*LeEIX1*), a gene that encodes receptor-like proteins (RLPs). Similarly, HanXRQChr17g0552491 showed homology to MLA10, HanXRQChr13g0420141 to N, HanXRQChr17g0552491 to both MLA12 and MLA13 and HanXRQChr17g0552491 to Sr33 protein with greater than 60% identity. Sunflower Genome Database with *H. annuus* r1 annotations was employed to obtain expression data for predicted NBS genes. We compared accessions of *H. annuus* r1.2 annotations to *H. annuus* r1 to obtain the expression data for NBS proteins. Since there were many duplicates for *H. annuus* r1.2 annotations, we used only the sequences with the unique names. The raw Read Per Kilobase Million) (RPKM) values of gene expression were downloaded separately. The expression values were from bract, corolla, leaves, ligule, ovary, pollen, seed, stamen and stem. Only expression data for 9 CNL type, 33 TNL type, 23 NL type and 6 RNL type genes were retrieved from the database and employed to generate heatmap after deseq normalization of the data using MeV package (Figure 3.6). Cluster I consists of 13 genes representing moderate to minimal expression, cluster II

with 43 genes representing basal to no expression and cluster III with 15 genes

representing minimal expression to basal expression (Figure S15).
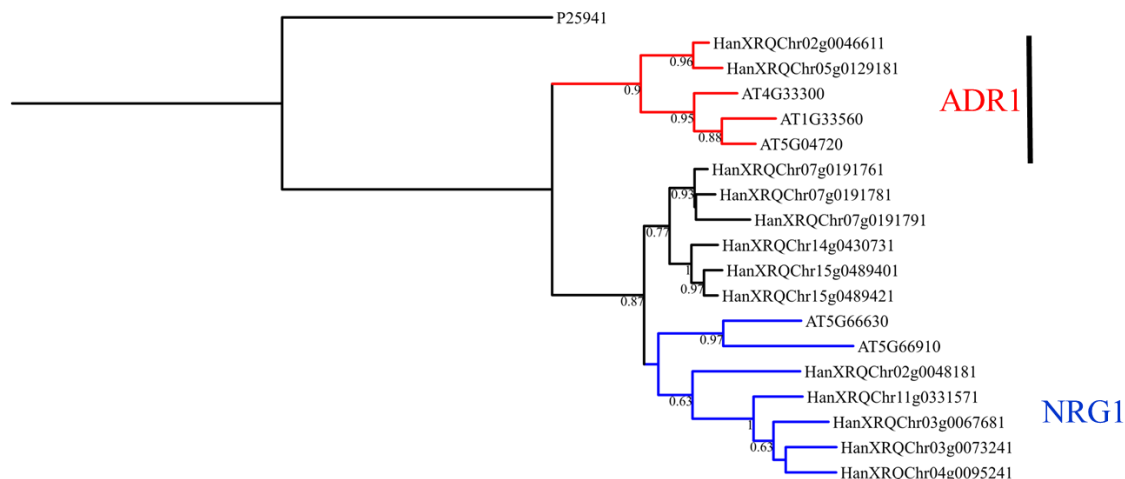
Figure 3.5. Phylogenetic relationships of RNL proteins in *Arabidopsis thaliana* and *Helianthus annuus*. The clades N-required gene 1 (*NRG1*) and activated disease resistance gene 1 (*ADR1*) are color-coded in blue and red, respectively. The tree was rooted using *Streptomyces coelicolor* NBS-containing protein, P25941, as an outgroup.

3.4. Discussion

3.4.1. Diversity of NBS-Encoding Genes

Our findings on the NBS-encoding genes in this study is based on recently

sequenced sunflower genome [44]. Previously, Gedil et al. 2001 [61] identified RGC

fragments with the NBS domains and assigned to 11 groups among which Ha4W2A was

linked to Pl1, a downy mildew resistance gene. Plocik et al. 2004 [40] identified nine

unique NBS domain sequences using degenerate primers in sunflower and compared

them to lettuce, chicory and *A. thaliana*. They concluded that NBS gene sequences of

Asteraceae family are ancestral to the Brassicaceae family. Later, Radwan et al. 2008

[42] identified 118 and 95 NBS domain sequences in RHA373 and ANN-1811 germplasm of H. annuus, respectively. In this study, we identified 352 NBS-encoding genes that constitute 0.67% of the total predicted proteins in sunflower, which shows similarity to *M. truncatula* (~0.66%) [27]. This number is higher than that of Arabidopsis *(~0.43%)* [14], *C. sativus* (~0.21%) [33], Carica papaya *(~0.21%)* [62] and lower than that of P. vulgaris (~1.19%) [63], *Manihot esculenta* (~0.9%) [64], *V. vinifera* (~1.3%) [28], and *G. max* (~0.73%) [23, 24]. We performed protein blast (BLASTp) analyses using 352 NBS domains of NBS-encoding genes identified in this study against a database with previously studied NBS domain sequences. The BLASTp analyses against a database comprised of sequences from Gedil et al. 2001 [61], Plocik et al. 2004 [40], and Radwan et al. 2008 [42] showed 70 to 100% identity to 143, 68 and 100 NBS domain sequences identified in this study, respectively (Supplementary File S4).
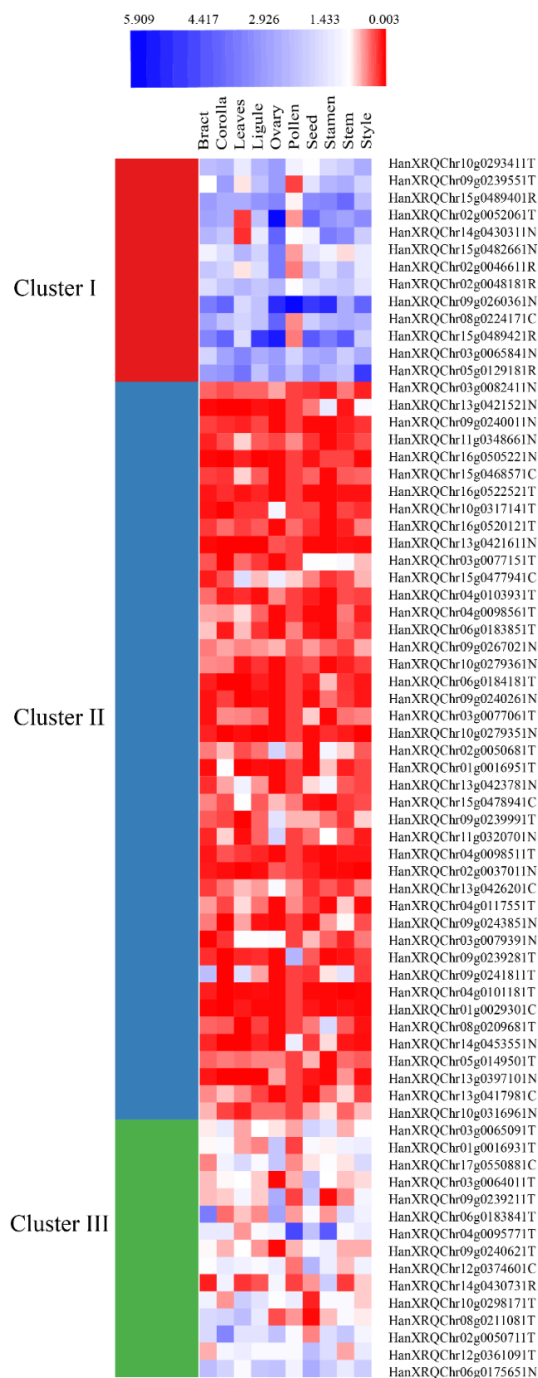
Figure 3.6. Expression profile of NBS genes from sunflower visualized as heatmap. The heatmap was generated using deseq normalized data for sunflower NBS genes expression in different tissues. K-means Clustering Method was employed for clustering (I, II and III). Gene IDs are followed by NBS type (C: CNLs; T: TNLs; N: NLs and R: RNLs).

Following the classification of NBS genes by Shao et al. 2016 [13] and Yu et al. 2014 [5], we classified NBS genes of sunflower into CNL, TNL, RNL and NL groups and their subgroups. We identified 100 genes belonging to the CNL group, with 64 possessing RX_CC like domain, 77 to the TNL group, 13 to the RNL group, and 162 to the NL group. In sunflower, the number of CNLs was found to be higher than that of TNLs, and the ratio of CNLs to TNLs was 1.3:1. The CNL:TNL ratio in the current study is not consistent with the findings observed in some other dicot species such as *A. thaliana* (1:2), *A. lyrata* (1:2), *B. rapa* (1:2), *Eucalyptus grandis* (1:1.25), and *Thellungiella salsuginea* (1:1.5) as numbers of TNLs were higher than CNLs in these species [14, 30, 65, 66, 67]. However, grapevine, chickpea, and potato genomes constituted CNL:TNL in a ratio of 4:1 [20, 28, 57]. The higher number of CNLs in sunflower might suggest the higher contribution of these genes providing resistance against pest or pathogen attack, which warrants future investigation. Furthermore, these groups are classified into subgroups as CNLs were classified into four subgroups [CNL (90), CN (5), CNN (1), CNNL (4)], TNLs into six subgroups [TNL (52), TN (21), TTNL (1), TNLTNL (1), CTNL (1), CTN (1)], RNLs into three subgroups [RNL (10), RN (1), RCNL (2)], and NLs into four subgroups [N (29), NL (125), NN (2), NNL (6)]. The classification is based on the presence of the CC domain named as 'C', the presence of TIR domain as 'T', the presence of RPW8 domain as 'R', the presence of the NBS domain as 'N', the presence of two NBS domains as 'NN', and the presence of LxxLxxLxx signatures as 'L' in the amino acid sequences of the proteins. The CNL type constituted approximately 92% of the genes belonging to CNL subgroup, 67% of the genes belonging to TNL subgroup in the TNL type, 76% of the genes belonging to RNL

subgroup in RNL type and 77% of the NL types genes are comprised of NL subgroup genes. The subgroups CN, CNNL, N, NN, and TTNL were also observed in *M. truncatula*, *A. thaliana*, and *B. rapa* [5, 19, 27]. HanXRQChr03g0067681 and HanXRQChr03g0073241 constituted both RPW8 and coiled-coil domains in the N-terminal and named RCNL, which were also reported in *A. thaliana* and *B. rapa* [5]. HanXRQChr05g0136351 and HanXRQChr06g0184071 possessed both TIR and coiled coil domain in the N-terminal of NBS proteins of sunflower and named CTN and CTNL, respectively. Such subgroups have been previously reported in many legumes and blueberries [16, 19].

NBS-encoding genes also called NBS-LRR genes encode proteins having TIR/CC at the N-terminal, NBS domain in the center and LRR at the C-terminal [14]. Among the identified NBS groups, genes belonging to NLs possessed less conserved NBS domain, as only 32% of the genes possessed all three signature motifs, while 57% of the CNLs, 89% of TNLs, and 100% of RNLs possessed all three signature motifs. Of the 100 CNLs, 64 genes possessed Rx_CC like domain in their N-terminal region. The disease resistance protein Rx possess CC domain in the N-terminal, and is expressed against potato virus X in *S. tuberosum* [68]. All TIR1, TIR2, TIR3 and TIR4 were detected in the TNLs of sunflower, which shows the consistency of TIR domain as described in other plant species such as *A. thaliana*, *P. vulgaris*, *G. max*, and *P. trichocarpa* [14, 24, 63, 69]. The characteristic 'DDVW' sequence was conserved in kinase-2 motifs of RNL and CNL genes, whereas 'DDVD' sequence was frequently found in TNL genes. The 'TSR' sequence was highly conserved in RNBS B motifs of the RNLs, while it slightly varies as 'TTR' and 'TTRD' in the CNLs and TNLs, respectively. This was found to be consistent

with the large scale study of NBS proteins in angiosperms [13]. All of the identified NBS

proteins possessed MHDL motifs, except for the RNL genes, frequently possessing

QHDL motif. Such QHDL motifs were observed in NBS proteins of *P. trichocarpa* [69].

A unique Kelch motif sequence was observed in HanXRQChr02g0052061 protein.

Previously, Kelch motifs were reported in the NBS proteins of *B. rapa* [5]. Kelch motif

sequences are considered to be signature motif for positive selection mostly found at the

C-terminal of F-Box proteins and are well studied in plant species such as *A. thaliana*, *P.

trichocarpa*, and O. *sativa* [70].

We further compared our pipeline with another pipeline, RGAugury [35], for the

identification of NBS-encoding genes. RGAugury is the integrative pipeline that

facilitates the prediction of NBS-encoding genes, RLKs, and RLPs [35]. RGAugury

predicted all 352 NBS proteins identified in this study plus five more proteins

[HanXRQChr02g0037021 (TN), HanXRQChr09g0240471 (TN),

HanXRQChr11g0340171 (CNL), HanXRQChr13g0394521 (TN), and

HanXRQChr16g0515381 (CN)] and 25 belonging TX (absence of NBS domain)

subclass. These missed proteins were manually checked and NBS domain (PF00931) in

HanXRQChr09g0240471, HanXRQChr11g0340171, HanXRQChr13g0394521, and

HanXRQChr16g0515381 were absent except in HanXRQChr02g0037021 (could belong

to TN subgroup). In addition, we suggest HanXRQChr09g0240471 to be classified as a

TX subclass. We found some discrepancies in the CNL group counts between two

pipelines. The use of a MARCOIL tool in our pipeline helped with filtering false

positives from the CNL group counts, and we could not observe any discrepancies in the

TNL group counts between the two pipelines. Furthermore, the RGAugury pipeline could

not identify an RNL group of genes that were identified in this study and majorly

categorized them to NL group (N and NL subclasses) of genes. The study and

identification of TX proteins were beyond the scope of our study as these proteins were

filtered out because of the absence of NBS domains. The differences and discrepancies

between identification and classification of predicted NBS-encoding genes using our and

RGAugury pipelines are represented in Supplementary File S5. In addition, RGAugury

was employed to predict proteins belonging to RLP, RLK and Transmembrane-coiled-

coil (TM-CC) proteins. A total of 257 RLPs [255-LRR type, 2-lysin motif (LysM) type],

1086 RLKs (368-LRR type, 12-LysM type and 706 Other-receptor type) and 173 TM-CC

proteins were predicted in the sunflower (Supplementary File S5). Both RLKs and RLPs

play important role in plant development and defense mechanism [4, 71]. RLKs such as

FLAGELLIN SENSITIVE 2 (FLS2) [72], elongation factor Tu receptor (EFR) [73],

systemin cell-surface receptor (SR160) [74], Xa21 [75], ERECTA RLK [76] and many

more are well characterized that are mainly involved in detection of pathogen associated

molecular patterns (PAMPs). On the other hand, RLP (lacking Kinase-2 domain) such as

Arabidopsis CLAVATA2 (CLV2, AtRLP10) [77] is involved in the development of

meristem and Cf is involved in pathogenesis against *Cladosporium fulvum* in tomato [78].

3.4.2. Gene Location, Clustering, K_a/K_s Values and Structural Variation

A variety of clustering patterns of NBS-encoding genes, frequently observed in

almost all plant species, is one of the major reasons for rapid evolution of the NBS genes

[14, 79]. The NBS genes of sunflower formed 75 clusters, 25 of which reside in

chromosome Ha13, 73 out of 200 (~37%) genes. In *M. esculenta*, 143 NBS genes

positioned in 39 clusters [64]. In *C. sativus*, 33 *NBS* genes were located in nine clusters

[33]. The average number of NBS proteins per cluster in sunflower was approximately

2.7, lesser than ratios in Solanaceae species such as tomato (3.48), potato (4.65), pepper

(3.44) [80], Brassicaceae species such as *B. oleracea* (3.04), *B. rapa* (2.7), *A. thaliana*

(2.8) [5], Fabaceae species such as *G. max* (4), *V. vinifera* (6), *M. truncatula* (5) [16],

Gossypium species such as *G. arboretum* (3.4), *G. raimondii* (5.5), *G. hirsutum* (5.3), and

G. barbadense (3.5) [34]. Both segmental and tandem duplications are responsible for the

formation of new clusters that generate intraspecific variation by processes such as

unequal crossing over [9, 14, 81]. However, NBS-encoding genes do not undergo high

rates of mutation and maintain both intra- and inter-specific variation [9]. The average

exon counts of sunflower CNLs (2.7 exons per gene) and TNLs (6.1 exons per gene)

were consistent with CNLs (2.7 exons per gene) and TNLs (5.1) of Arabidopsis [14].

This implies a high number of exons of TNLs and RNLs could help with generating

diverse resistance proteins through alternative splicing. All NBS types showed K_a/K_s

values of less than one, indicating that these genes are under the influence of purifying

selection.

3.4.3. Phylogenetic Relationships, Homology, Synteny and Expression Analysis

Sunflower CNL genes were similar to *C. sativus* CNL genes while compared to

their respective TNL genes [33]. However, the CNL clade size in sunflower is different

from *Arabidopsis*, as TNL clades constitute larger numbers of genes than CNL clade

[14]. Subclades CC (I) possessed gene members with introns in range of one to ten, and

CC (II) constituted gene members with introns in the range of zero to one. Other

subclades, CC (III) and CC (IV) constituted gene members with introns in the range of

zero to two and CC (V) and CC (VI) constituted gene members with introns in the range

of zero to four. Only *HanXRQChr02g0057361*, *HanXRQChr02g0057351*, and

HanXRQChr13g0425771 in the subclade CC (VI) possessed in the range of five to seven.

Similarly, subclade TIR (II) possessed gene members with introns in the lowest range

(three to five). TIR (I), TIR (III), TIR (IV), TIR (V) and TIR (VI) gene members

possessed introns in range of 3 to 17, 2 to 7, 1 to six, 1 to 15, and 1 to 13, respectively.

Similar patterns were also observed in the phylogenetic tree of CNL and TNL in *C.

sativus* [33]. The differences in the clade pattern with correlation to introns in two gene

families suggest the role of intron loss and gain in the structural evolution of the NBS

genes as suggested by Wan et al. 2013 [33]. In addition, the position, presence or

absence, and phase of introns often play important roles in phylogeny [82].

We found that RNLs were nested within the clade of TNLs in sunflower (a member of

the Asterids lineage) although RNLs in the families Brassicaceae and Fabaceae (Rosids

lineage) were found to be related to CNLs [15, 30]. The lineage of Asterids is believed to

have evolved from the rest of Angiosperms (Rosids + monocots + basal Angiosperms)

little over 100 million years ago (MYA) [83]. A large-scale study of Angiosperms NBS

genes also concluded that RNLs were sister to the CNLs [13]. However, these earlier

studies did not include *H. annuus* in the analysis as the genome was not available by then.

Our results indicate a surprising position of RNLs within TNLs in sunflower making the

clades of TNL and CNL potentially paraphyletic. Upon reconstruction of the

phylogenetic tree with *Arabidopsis* NBS genes, RNL genes of sunflower were observed

in a CNL-A clade (although it is consistent with the previous study) [14]. The CNL-A

clade did not consist of any sunflower CNL gene members besides RNLs. Further study

on comparative genomics or transcriptomes across the Asterids lineage can confirm

whether CNL genes are completely absent in the lineage. Shao et al. 2016 [13] suggested that RNLs were derived from *ADR1* and *NRG1*, and two ancient lineages separated before the Angiosperms diversified. The RNL genes, *ADR1* and *NRG1*, have been characterized in *Arabidopsis* and *Nicotiana*, respectively. A separate tree, constructed to observe the relationships among sunflower RNLs and *Arabidopsis* RNLs, formed two clades. The sunflower RNL genes *HanXRQChr02g0046611* and *HanXRQChr05g0129181* were nested with *AT4G3330* (*ADR1-L1*), *AT1G33560* (*ADR1*) and *AT5G04720* (*ADR1-L2* or *PHX21*), with bootstrap support of 90%. On the other hand, *HanXRQChr02g0048181*, *HanXRQChr11g0331571*, *HanXRQChr03g0067681*, *HanXRQChr0073241*, and *HanXRQChr04g0095241* were nested with *AT5G66630* (RNL) and *AT5G66910* (homologous to *NRG1*), with bootstrap support of 63%. This suggests that the sunflower RNLs mentioned above are orthologous to the ADR1 and NRG1 homologs of *Arabidopsis*. ADR1 proteins play a role as helper genes for receiving signals from the R genes in downstream signaling of effector-triggered immunity [84]. Similarly, NRG1 proteins help the N protein during the pathogenesis by the tobacco mosaic virus [85]. Since they are not directly involved in detecting the pathogen effectors, they are not much influenced by a selection pressure due to the pathogens [13]. Only 5.8% of the total NBS genes in sunflower are RNL genes which is consistent with other species, such as *A. lyrata* (2.5%), *A. thaliana* (4.2%), *B. rapa* (4.4%), *Capsella rubella* (4.7%) and *T. salsuginea* (5.7%) [30]. Other results from this study that separate RNLs from the rest of the NBS genes include their highest average number of exons per gene and lowest average K_a/K_s ratios values for the clade. This supports the hypothesis of high conservation and slow evolutionary rates among the RNL genes [86].

Sunflower NBS proteins identified in this study formed clades with reference proteins such as Pi36, Pl8, Rps2, VAT, RPG1, Gro1.4, RY-1, and N proteins, suggesting their homologous relationships (Figure S13). The sunflower TNL proteins are inferred to be orthologous to *S. tuberosum* nematode resistance protein (Gro1.4) [87], *S. tuberosum* subsp. *andigena* RY-1 (conferring resistance to potato virus Y) [88], and *N. glutinosa* Tobacco Mosaic Virus resistance (N) gene [89]. Similarly, sunflower CNL proteins are inferred to be orthologous to *A. thaliana* RPS2 (Resistant to *P. syringae* 2) [90], *Cucumis melo* VAT (resistance to *Aphis gossypii*) [91], *H. annuus* Pl8 [43], *O. sativa* Pi36 (conferring resistance to Blast fungus) [92], and *H. vulgare* subsp. *vulgare* RPG1 (conferring resistance to stem rust fungus) [93]. The BLAST investigation of sunflower NBS proteins with reference proteins available on http://www.prgdb.org has shown some of them to be the possible homologs of the reference proteins (Table S7). Sunflower NBS proteins such as HanXRQChr13g0425411, HanXRQChr13g0425361, and HanXRQChr13g0425431 showed greater than 80% sequence identity to the *H. annuus* gene, *Pl8* gene (CNL). The *Pl8* gene is involved in conferring resistance to *P. halstedii*, a causative agent to downy mildew [43]. HanXRQChr04g0123041, belonging to the NL group has shown homology to *L. esculentum* EIX receptor 2 (*Eix2*), a gene that encodes receptor-like proteins (RLPs) involved in detecting ethylene-inducing xylanase, a fungus elicitor [94]. Other inferred homologs include HanXRQChr17g0552491 to MLA10, HanXRQChr13g0420141 to N, HanXRQChr17g0552491 to both MLA12 as well as MLA13, and HanXRQChr17g0552491 to Sr33. The *MLA* locus is highly polymorphic and encode allelic CNL type resistance proteins such as MLA1, MLA2, and MLA3 that confer resistance to barley powdery mildew fungus (*Blumeria graminis* f. sp. Hordei,

Bgh) [95]. Another protein, Sr33, which belongs to the CNL type, confers resistance to a

wheat stem rust pathogen, *Puccinia graminis* f. sp. *tritici* [96]. We were able to access

expression profiles for only a few unique sunflower NBS proteins because of the

duplicated names found for corresponding *H. annuus* r1.2 annotations compared to *H.*

annuus r1 annotations. From the available expression data, it can be deduced that NBS

genes can be expressed at a basal level with tissue specificity in unchallenged conditions

[97]. In the expression dataset, most of the NBS genes were found to have a minimal to

no expression value possibly as a result of low sequencing coverage, or their expression

dependent on infection of pathogens or due to a pseudogenization, which was also noted

by Frazier et al. 2016 [98]. Thus, detailed transcriptomic and proteomics studies are

warranted to functionally characterize the sunflower NBS genes, particularly challenging

the plant by various pests and pathogens through carefully crafted experimental designs.

3.5. Conclusions

We identified 352 NBS genes in sunflower and studied their clustering,

phylogenetic relationships, gene homology and functional divergence. These genes

formed clusters and showed structural conservation in signature domains and exon/intron

architecture in CNL, TNL and RNL types of NBS genes. The RNLs belonged to the

CNL-A clade, which in turn was found nested within the TNL clade, making both CNL

and TNL clades paraphyletic. This warrants further rigorous analysis. All of the NBS-

encoding genes have undergone purifying selection and available expression data have

revealed their functional divergence. We confirmed homology of sunflower NBS genes

to multiple previously characterized *Pl8*, *LeEIX1*, *MLA10-13*, *Sr33* resistance genes.

Further characterization of the NBS genes will help us to understand resistance pathways and to develop durable resistance necessary for crop improvement in sunflower, one of the major oilseed crops in the world.

Supplementary Materials: The following are available online at https://www.mdpi.com/2073-4425/9/8/384/s1. Figure S1. Predicted protein domains in sunflower CNL protein sequences with number of LxxLxxLxx signatures (in parentheses). Each black line represents amino acid sequence lengths that correspond to the scale provided at the top of the figure. The name and signs of the domains are presented below the black line: CC: Coiled Coil; NBS: Nucleotide Binding Site; RX_CC: Coiled coil domain present in disease resistance protein, Rx; LxxLxxLxx signatures are represented by the purple shapes under the black lines. Figure S2. Predicted protein domains in sunflower TNL protein sequences with the number of LxxLxxLxx signatures (in parentheses). The black lines represent the protein lengths that correspond to the scale provided at the top of the figure. The name and signs of the domains are presented below the black lines: TIR: Toll-interleukin-1 Receptor; NBS: Nucleotide Binding Site; CC: Coiled Coil; LxxLxxLxx signatures are represented by the purple shapes under the black lines. Figure S3. Predicted protein domains in sunflower RNL protein sequences with number of LxxLxxLxx signatures (in parentheses). The name and signs of the domains are presented below the color-coded lines: RPW8: Resistance to powdery mildew 8; NBS: Nucleotide Binding Site; CC: Coiled Coil; LxxLxxLxx signatures are represented by the purple shapes under the color coded lines. Figure S4. Predicted protein domains in sunflower NL protein sequences with the number of LxxLxxLxx signatures (in parentheses). The black lines represent the protein length that corresponds to the scale

provided at the top of the figure. The name and signs of the domains are presented just below the black lines: NBS: Nucleotide Binding Site; LxxLxxLxx signatures are represented by the purple triangular shapes under each black line. Figure S5. Conserved domains of sunflower CNL genes predicted by MEME analysis. The parameters used were the number of motif-20; minimum width-6; and maximum width-50. The logos of numbered color-coded motifs are shown in Supplementary File S1a. Figure S6. Conserved domains of sunflower TNL genes predicted by MEME analysis. The parameters used were- the number of motif-20; minimum width-6; maximum width-25. The logos of numbered color-coded motifs are shown in Supplementary File S1b. Figure S7. Conserved domains of sunflower RNL genes predicted by MEME analysis. The parameters used were-the number of motif-20; minimum width-6; maximum width-25. The logos of numbered color-coded motifs are shown in File S1c. Figure S8. Chromosomal distribution of NBS genes in a sunflower ($n = 17$). Figure S9. Exon–intron architecture of the coding sequences of CNL genes in sunflower. Figure S10. Exon–intron architecture of the coding sequences of TNL genes in sunflower. Figure S11. Exon–intron architecture of the coding sequences of RNL genes in sunflower. Figure S12. Exon–intron architecture of the coding sequences of NL genes in sunflower. Figure S13. Maximum likelihood (ML) tree of the NBS amino acid sequences of the CNL, TNL and RNL genes from sunflower along with those of previously characterized CNL, TNL and RPW8 type genes. The tree was reconstructed using JTT + G + I (Jones–Taylor–Thornton with gamma distribution and invariant sites) model with 1000 bootstrap replicates. The tree was rooted with *Streptomyces coelicolor* (P25941) as an outgroup. Figure S14. Syntenic relationships between chromosomes of *Arabidopsis* and sunflower.

Chromosomal blocks of *Arabidopsis* (color coded) are mapped onto the chromosome of sunflower (Ha1 to Ha17). Figure S15. Identified genes in different clusters showing differential expression on all tissues in sunflower. Gene IDs are followed by NBS type (C: CNL type; T: TNL type; N: NL type and R: RNL type). Table S1. List of NBS gene accessions, their type, number of LxxLxxLxx signatures, exon/introns number, protein sequence length, gene orientation, amino acid length and amino acid sequences. Table S2. List of *NBS* genes of sunflower with nuclear localization signal (NLS) peptides. Table S3. MEME predicted conserved motifs in CNL proteins of sunflower. The motif logos represented in column three are shown in Supplementary File S1a. Table S4. MEME predicted conserved motifs in TNL proteins of sunflower. The motif logos represented in column three are shown in Supplementary File S1b. Table S5. MEME predicted conserved motifs in RPW8 proteins of sunflower. The motif logos represented in column three are shown in Supplementary File S1c. Table S6. List of gene accessions, their corresponding gene clusters and chromosomal location. Table S7. BLAST result of NBS genes against reference genes of Plant Resistance Genes database (PRGdb; http://prgdb.org) with a cutoff E-value of 0.01. Supplementary File S1. Motif sequence logos for the sunflower NBS proteins. Supplementary File S2. Sequence alignment of the NBS domains belonging to different groups in fasta format. Supplementary File S3. Newick files for phylogenetic trees shown in Figure 3, Figure 4 and Figure 5. Supplementary File S4. BLASTp result of NBS domains of the genes identified in this study and those previously identified by Gedil et al. 2001 [61], Plocik et al. 2004 [40], and Radwan et al. 2008 [42]. Supplementary File S5. Identification and classification of

NBS-encoding genes using current pipeline (this study) compared to those predicted by

RGAugury pipeline, and a list of RLPs, RLKs and TM-CC proteins in sunflower.

References

1. Baluška, F.; Mancuso, S., *Signaling in Plants, Signaling and Communication in Plants*; Springer: Berlin/Heidelberg, Germany, 2009.
2. Wang, J.; Pan, C.; Wang, Y.; Ye, L.; Wu, J.; Chen, L.; Zou, T.; Lu, G. Genome-wide identification of MAPK, MAPKK, and MAPKKK gene families and transcriptional profiling analysis during development and stress response in cucumber. *BMC Genom.* 2015, *16*, 386.
3. Andersen, E.; Ali, S.; Byamukama, E.; Yen, Y.; Nepal, M. Disease resistance mechanisms in plants. *Genes* 2018, *9*, doi:10.3390/genes9070339.
4. Sekhwal, M.K.; Li, P.; Lam, I.; Wang, X.; Cloutier, S.; You, F.M. Disease resistance gene analogs (RGAs) in Plants. *Int. J. Mol. Sci.* 2015, *16*, 19248–19290.
5. Yu, J.; Tehrim, S.; Zhang, F.; Tong, C.; Huang, J.; Cheng, X.; Dong, C.; Zhou, Y.; Qin, R.; Hua, W.; et al. Genome-wide comparative analysis of NBS-encoding genes between *Brassica* species and *Arabidopsis thaliana*. *BMC Genom.* 2014, *15*, 1–18.
6. Jones, J.D.; Dangl, J.L. The plant immune system. *Nature* 2006, *444*, 323–329.
7. Zipfel, C. Pattern-recognition receptors in plant innate immunity. *Curr. Opin. Immunol.* 2008, *20*, 10–16.
8. Andolfo, G.; Ercolano, M.R. Plant innate immunity multicomponent model. *Front. Plant Sci.* 2015, *6*, 987.

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

10. Vergne, E.; Grand, X.; Ballini, E.; Chalvon, V.; Saindrenan, P.; Tharreau, D.; Notteghem, J.-L.; Morel, J.-B. Preformed expression of defense is a hallmark of partial resistance to rice blast fungal pathogen *Magnaporthe oryzae*. *BMC Plant Biol.* 2010, *10*, 206.

11. Osuna-Cruz, C.M.; Paytuvi-Gallart, A.; Di Donato, A.; Sundesha, V.; Andolfo, G.; Aiese Cigliano, R.; Sanseverino, W.; Ercolano, M.R. PRGdb 3.0: A comprehensive platform for prediction and analysis of plant disease resistance genes. *Nucleic Acids Res.* 2017, *46*, D1197–D1201.

12. Gururani, M.A.; Venkatesh, J.; Upadhyaya, C.P.; Nookaraju, A.; Pandey, S.K.; Park, S.W. Plant disease resistance genes: Current status and future directions. *Physiol. Mol. Plant Pathol.* 2012, *78*, 51–65.

13. Shao, Z.-Q.; Xue, J.-Y.; Wu, P.; Zhang, Y.-M.; Wu, Y.; Hang, Y.-Y.; Wang, B.; Chen, J.-Q. Large-scale analyses of angiosperm nucleotide-binding site-leucine-rich repeat (NBS-LRR) genes reveal three anciently diverged classes with distinct evolutionary patterns. *Plant Physiol.* 2016, doi:10.1104/pp.15.0148.

14. Meyers, B.C.; Kozik, A.; Griego, A.; Kuang, H.; Michelmore, R.W. Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. *Plant Cell* 2003, *15*, 809–834.

15. Shao, Z.-Q.; Zhang, Y.-M.; Hang, Y.-Y.; Xue, J.-Y.; Zhou, G.-C.; Wu, P.; Wu, X.-Y.; Wu, X.-Z.; Wang, Q.; Wang, B. Long-term evolution of nucleotide-binding site-leucine-rich repeat genes: Understanding gained from and beyond the legume family. *Plant Physiol.* 2014, *166*, 217–234.

16. Zheng, F.; Wu, H.; Zhang, R.; Li, S.; He, W.; Wong, F.-L.; Li, G.; Zhao, S.; Lam, H.-M. Molecular phylogeny and dynamic evolution of disease resistance genes in the legume family. *BMC Genom.* 2016, *17*, 402.

17. Lee, H.A.; Yeom, S.I. Plant NB-LRR proteins: Tightly regulated sensors in a complex manner. *Brief Funct. Genom.* 2015, *14*, 233–242.

18. Michelmore, R.W.; Christopoulou, M.; Caldwell, K.S. Impacts of resistance gene genetics, function, and evolution on a durable future. *Annu. Rev. Phytopathol.* 2013, *51*, 291–319.

19. Die, J.V.; Román, B.; Qi, X.; Rowland, L.J. Genome-scale examination of NBS-encoding genes in blueberry. *Sci. Rep.* 2018, *8*, 3429.

20. Sharma, R.; Rawat, V.; Suresh, C. Genome-wide identification and tissue-specific expression analysis of nucleotide binding site-leucine rich repeat gene family in *Cicer arietinum* (kabuli chickpea). *Genom. Data* 2017, *14*, 24–31.

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

22. Nepal, M.P.; Benson, B.V. CNL Disease resistance genes in soybean and their evolutionary divergence. *Evol. Bioinform. Online* 2015, *11*, 49–63.

23. Nepal, M.P.; Andersen, E.J.; Neupane, S.; Benson, B.V. Comparative genomics of non-TNL Disease resistance genes from six plant species. *Genes* 2017, *8*, 249.

24. Neupane, S.; Ma, Q.; Mathew, F.M.; Varenhorst, A.J.; Andersen, E.J.; Nepal, M.P. Evolutionary divergence of TNL disease-resistant proteins in soybean (*Glycine max*) and common bean (*Phaseolus vulgaris*). *Biochem. Genet.* 2018, *56*, 397–442.

25. Monosi, B.; Wisser, R.J.; Pennill, L.; Hulbert, S.H. Full-genome analysis of resistance gene homologues in rice. *Theor. Appl. Genet.* 2004, *109*, 1434–1477.

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

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

28. Yang, S.; Zhang, X.; Yue, J.-X.; Tian, D.; Chen, J.-Q. Recent duplications dominate NBS-encoding gene expansion in two woody species. *Mol. Genet. Genom.* 2008, *280*, 187–198.

29. Lozano, R.; Ponce, O.; Ramirez, M.; Mostajo, N.; Orjeda, G. Genome-wide identification and mapping of NBS-encoding resistance genes in *Solanum tuberosum* group *phureja*. *PLoS ONE* 2012, *7*, e34775.

30. Zhang, Y.M.; Shao, Z.Q.; Wang, Q.; Hang, Y.Y.; Xue, J.Y.; Wang, B.; Chen, J.Q. Uncovering the dynamic evolution of nucleotide-binding site-leucine-rich repeat (NBS-LRR) genes in Brassicaceae. *J. Integr. Plant Biol.* 2016, *58*, 165–177.

31. Andersen, E.J.; Ali, S.; Reese, R.N.; Yen, Y.; Neupane, S.; Nepal, M.P. Diversity and evolution of disease resistance genes in Barley (*Hordeum vulgare* L.). *Evol. Bioinform. Online* 2016, *12*, 99.

32. Andersen, E.J.; Nepal, M.P. Genetic diversity of disease resistance genes in foxtail millet (*Setaria italica* L.). *Plant Gene* 2017, *10*, 8–16.

33. Wan, H.; Yuan, W.; Bo, K.; Shen, J.; Pang, X.; Chen, J. Genome-wide analysis of NBS-encoding disease resistance genes in *Cucumis sativus* and phylogenetic study of NBS-encoding genes in Cucurbitaceae crops. *BMC Genom.* 2013, *14*, 109.

34. Xiang, L.; Liu, J.; Wu, C.; Deng, Y.; Cai, C.; Zhang, X.; Cai, Y. Genome-wide comparative analysis of NBS-encoding genes in four *Gossypium* species. *BMC Genom.* 2017, *18*, 292.

35. Li, P.; Quan, X.; Jia, G.; Xiao, J.; Cloutier, S.; You, F.M. RGAugury: A pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genom.* 2016, *17*, 852.

36. Kane, N.C.; Rieseberg, L.H. Selective sweeps reveal candidate genes for adaptation to drought and salt tolerance in common sunflower, *Helianthus annuus*. *Genetics* 2007, *175*, 1823–1834.

37. Fernández-Martínez, J.; Melero-Vara, J.; Muñoz-Ruz, J.; Ruso, J.; Domínguez, J. Selection of wild and cultivated sunflower for resistance to a new broomrape race that overcomes resistance of the *Or*$_5$ gene. *Crop Science* 2000, *40*, 550–555.

38. Seiler, G. Wild annual *Helianthus anomalus* and *H. deserticola* for improving oil content and quality in sunflower. *Ind. Crops Prod.* 2007, *25*, 95–100.

39. Markell, S.; Harveson, R.; Block, C.; Gulya, T. Sunflower Disease Diagnostic Series. 2005. *North Dakota State Univ. Ext. Serv. Publ.* 1727-19.

40. Plocik, A.; Layden, J.; Kesseli, R. Comparative analysis of NBS domain sequences of NBS-LRR disease resistance genes from sunflower, lettuce, and chicory. *Mol. Phylogenet. Evol.* 2004, *31*, 153–163.

41. Hewezi, T.; Mouzeyar, S.; Thion, L.; Rickauer, M.; Alibert, G.; Nicolas, P.; Kallerhoff, J. Antisense expression of a NBS-LRR sequence in sunflower (*Helianthus annuus* L.) and tobacco (*Nicotiana tabacum* L.): Evidence for a dual role in plant development and fungal resistance. *Transgenic Res.* 2006, *15*, 165–180.

42. Radwan, O.; Gandhi, S.; Heesacker, A.; Whitaker, B.; Taylor, C.; Plocik, A.; Kesseli, R.; Kozik, A.; Michelmore, R.W.; Knapp, S.J. Genetic diversity and genomic distribution of homologs encoding NBS-LRR disease resistance proteins in sunflower. *Mol. Genet. Genom.* 2008, *280*, 111–125.

43. Radwan, O.; Mouzeyar, S.; Nicolas, P.; Bouzidi, M. Induction of a sunflower CC-NBS-LRR resistance gene analogue during incompatible interaction with *Plasmopara halstedii*. *Journal of Experimental Botany* 2004, *56*, 567–575.

44. Badouin, H.; Gouzy, J.; Grassa, C.J.; Murat, F.; Staton, S.E.; Cottret, L.; Lelandais-Brière, C.; Owens, G.L.; Carrère, S.; Mayjonade, B. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* 2017, *546*, 148.

45. Finn, R.D.; Clements, J.; Arndt, W.; Miller, B.L.; Wheeler, T.J.; Schreiber, F.; Bateman, A.; Eddy, S.R. HMMER web server: 2015 update. *Nucleic Acids Res.* 2015, *43*, W30–W38.

46. Jones, P.; Binns, D.; Chang, H.-Y.; Fraser, M.; Li, W.; McAnulla, C.; McWilliam, H.; Maslen, J.; Mitchell, A.; Nuka, G. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* 2014, *30*, 1236–1240.

47. Finn, R.D.; Bateman, A.; Clements, J.; Coggill, P.; Eberhardt, R.Y.; Eddy, S.R.; Heger, A.; Hetherington, K.; Holm, L.; Mistry, J. Pfam: The protein families database. *Nucleic Acids Res.* 2013, *42*, D222–D231.

48. Delorenzi, M.; Speed, T. An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics* 2002, *18*, 617–625.

49. Bailey, T.L.; Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 1994, *2*, 28–36.

50. Emanuelsson, O.; Brunak, S.; von Heijne, G.; Nielsen, H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* 2007, *2*, 953–971.

51. Ba, A.N.N.; Pogoutse, A.; Provart, N.; Moses, A.M. NLStradamus: A simple Hidden Markov Model for nuclear localization signal prediction. *BMC Bioinform.* 2009, *10*, 202.

52. Thompson, J.D.; Higgins, D.G.; Gibson, T.J. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 1994, *22*, 4673–4680.

53. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004, *32*, 1792–1797.

54. Kearse, M.; Moir, R.; Wilson, A.; Stones-Havas, S.; Cheung, M.; Sturrock, S.; Buxton, S.; Cooper, A.; Markowitz, S.; Duran, C. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 2012, *28*, 1647–1649.

55. Kumar, S.; Stecher, G.; Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 2016, *33*, 1870–1874.
56. Letunic, I.; Bork, P. Interactive tree of life (iTOL) v3: An online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* 2016, *44*, W242–W245.
57. Jupe, F.; Pritchard, L.; Etherington, G.J.; MacKenzie, K.; Cock, P.J.; Wright, F.; Sharma, S.K.; Bolser, D.; Bryan, G.J.; Jones, J.D. Identification and localisation of the NB-LRR gene family within the potato genome. *BMC Genom.* 2012, *13*, 75.
58. Rozas, J.; Ferrer-Mata, A.; Sánchez-DelBarrio, J.C.; Guirao-Rico, S.; Librado, P.; Ramos-Onsins, S.E.; Sánchez-Gracia, A. DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evol.* 2017, *34*, 3299–3302.
59. Soderlund, C.; Bomhoff, M.; Nelson, W.M. SyMAP v3. 4: A turnkey synteny system with application to plant genomes. *Nucleic Acids Res.* 2011, *39*, e68–e68.
60. Howe, E.A.; Sinha, R.; Schlauch, D.; Quackenbush, J. RNA-Seq analysis in MeV. *Bioinformatics* 2011, *27*, 3209–3210.
61. Gedil, M.A.; Slabaugh, M.B.; Berry, S.; Johnson, R.; Michelmore, R.; Miller, J.; Gulya, T.; Knapp, S.J. Candidate disease resistance genes in sunflower cloned using conserved nucleotide-binding site motifs: Genetic mapping and linkage to the downy mildew resistance gene *Pl1*. *Genome* 2001, *44*, 205–212.
62. Ming, R.; Hou, S.; Feng, Y.; Yu, Q.; Dionne-Laporte, A.; Saw, J.H.; Senin, P.; Wang, W.; Ly, B.V.; Lewis, K.L. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 2008, *452*, 991.
63. Wu, J.; Zhu, J.; Wang, L.; Wang, S. Genome-wide association study identifies NBS-LRR-encoding genes related with anthracnose and common bacterial blight in the common Bean. *Front. Plant Sci.* 2017, *8*, 1398.
64. Lozano, R.; Hamblin, M.T.; Prochnik, S.; Jannink, J.L. Identification and distribution of the NBS-LRR gene family in the Cassava genome. *BMC Genom.* 2015, *16*, 360.
65. Guo, Y.-L.; Fitz, J.; Schneeberger, K.; Ossowski, S.; Cao, J.; Weigel, D. Genome-wide comparison of nucleotide-binding site-leucine-rich repeat-encoding genes in *Arabidopsis*. *Plant Physiol.* 2011, *157*, 757–769.
66. Mun, J.-H.; Yu, H.-J.; Park, S.; Park, B.-S. Genome-wide identification of NBS-encoding resistance genes in *Brassica rapa*. *Mol. Genet. Genom.* 2009, *282*, 617–631.
67. Christie, N.; Tobias, P.A.; Naidoo, S.; Külheim, C. The Eucalyptus grandis NBS-LRR gene family: Physical clustering and expression hotspots. *Front. Plant Sci.* 2016, *6*, 1238.
68. Hao, W.; Collier, S.M.; Moffett, P.; Chai, J. Structural basis for the interaction between the potato virus X resistance protein (Rx) and its cofactor Ran GTPase-activating protein 2 (RanGAP2). *J. Biol. Chem.* 2013, *288*, 35868–35876.
69. Kohler, A.; Rinaldi, C.; Duplessis, S.; Baucher, M.; Geelen, D.; Duchaussoy, F.; Meyers, B.C.; Boerjan, W.; Martin, F. Genome-wide identification of NBS resistance genes in *Populus trichocarpa*. *Plant Mol. Biol.* 2008, *66*, 619–636.
70. Schumann, N.; Navarro-Quezada, A.; Ullrich, K.; Kuhl, C.; Quint, M. Molecular evolution and selection patterns of plant F-box proteins with C-terminal kelch repeats. *Plant Physiol.* 2011, *155*, 835–850.
71. Afzal, A.J.; Wood, A.J.; Lightfoot, D.A. Plant receptor-like serine threonine kinases: Roles in signaling and plant defense. *Mol. Plant-Microbe Interact.* 2008, *21*, 507–517.

72. Gómez-Gómez, L.; Boller, T. Flagellin perception: A paradigm for innate immunity. *Trends Plant Sci.* 2002, *7*, 251–256.

73. Zipfel, C.; Kunze, G.; Chinchilla, D.; Caniard, A.; Jones, J.D.; Boller, T.; Felix, G. Perception of the bacterial PAMP EF-Tu by the receptor EFR restricts *Agrobacterium*-mediated transformation. *Cell* 2006, *125*, 749–760.

74. Macho, A.P.; Zipfel, C. Plant PRRs and the activation of innate immune signaling. *Mol. Cell* 2014, *54*, 263–272.

75. Song, W.-Y.; Wang, G.-L.; Chen, L.-L.; Kim, H.-S.; Pi, L.-Y.; Holsten, T.; Gardner, J.; Wang, B.; Zhai, W.-X.; Zhu, L.-H. A receptor kinase-like protein encoded by the rice disease resistance gene, *Xa21*. *Science* 1995, *270*, 1804–1806.

76. Godiard, L.; Sauviac, L.; Torii, K.U.; Grenon, O.; Mangin, B.; Grimsley, N.H.; Marco, Y. ERECTA, an LRR receptor-like kinase protein controlling development pleiotropically affects resistance to bacterial wilt. *Plant J.* 2003, *36*, 353–365.

77. Jeong, S.; Trotochaud, A.E.; Clark, S.E. The *Arabidopsis CLAVATA2* gene encodes a receptor-like protein required for the stability of the CLAVATA1 receptor-like kinase. *Plant Cell* 1999, *11*, 1925–1933.

78. Kruijt, M.; Kip, D.J.; Joosten, M.H.; Brandwagt, B.F.; de Wit, P.J. The Cf-4 and Cf-9 resistance genes against *Cladosporium fulvum* are conserved in wild tomato species. *Mol. Plant-Microbe Interact.* 2005, *18*, 1011–1021.

79. Friedman, A.R.; Baker, B.J. The evolution of resistance genes in multi-protein plant resistance systems. *Curr. Opin. Genet. Dev.* 2007, *17*, 493–499.

80. Qian, L.-H.; Zhou, G.-C.; Sun, X.-Q.; Lei, Z.; Zhang, Y.-M.; Xue, J.-Y.; Hang, Y.-Y. Distinct patterns of gene gain and loss: Diverse evolutionary modes of NBS-encoding genes in three Solanaceae crop species. *G3 Genes Genomes Genet.* 2017, *7*, 1577–1585.

81. Leister, D. Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. *Trends Genet.* 2004, *20*, 116–122.

82. Sánchez, D.; Ganfornina, M.D.; Gutiérrez, G.; Marín, A. Exon-intron structure and evolution of the Lipocalin gene family. *Mol. Biol. Evol.* 2003, *20*, 775–783.

83. Zeng, L.; Zhang, Q.; Sun, R.; Kong, H.; Zhang, N.; Ma, H. Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. *Nat. Commun.* 2014, *5*, 4956.

84. Bonardi, V.; Tang, S.; Stallmann, A.; Roberts, M.; Cherkis, K.; Dangl, J.L. Expanded functions for a family of plant intracellular immune receptors beyond specific recognition of pathogen effectors. *Proc. Natl. Acad. Sci. USA* 2011, *108*, 16463–16468.

85. Peart, J.R.; Mestre, P.; Lu, R.; Malcuit, I.; Baulcombe, D.C. NRG1, a CC-NB-LRR protein, together with N, a TIR-NB-LRR protein, mediates resistance against tobacco mosaic virus. *Curr. Biol.* 2005, *15*, 968–973.

86. Collier, S.M.; Hamel, L.-P.; Moffett, P. Cell death mediated by the N-terminal domains of a unique and highly conserved class of NB-LRR protein. *Mol. Plant-Microbe Interact.* 2011, *24*, 918–931.

87. Paal, J.; Henselewski, H.; Muth, J.; Meksem, K.; Menéndez, C.M.; Salamini, F.; Ballvora, A.; Gebhardt, C. Molecular cloning of the potato *Gro1-4* gene conferring resistance to pathotype Ro1 of the root cyst nematode *Globodera rostochiensis*, based on a candidate gene approach. *Plant J.* 2004, *38*, 285–297.

88. Vidal, S.; Cabrera, H.; Andersson, R.A.; Fredriksson, A.; Valkonen, J.P. Potato gene *Y-1* is an N gene homolog that confers cell death upon infection with potato virus Y. *Mol. Plant-Microbe Interact.* 2002, *15*, 717–727.

89. Levy, M.; Edelbaum, O.; Sela, I. Tobacco mosaic virus regulates the expression of its own resistance gene N. *Plant Physiol.* 2004, *135*, 2392–2397.

90. Katagiri, F.; Thilmony, R.; He, S.Y. The *Arabidopsis thaliana-Pseudomonas syringae* interaction. *Arabidopsis Book* 2002, *1*, e0039.

91. Pitrat, M.; Maestro, C.; Ferriere, C.; Ricard, M.; Alvarez, J. Resistance to *Aphis gossypii* in Spanish melon (*Cucumis melo*). *Cucurbit. Genet. Coop. Rep.* 1988, *11*, 50–51.

92. Liu, X.; Lin, F.; Wang, L.; Pan, Q. The in silico map-based cloning of Pi36, a rice coiled-coil–nucleotide-binding site–leucine-rich repeat gene that confers race-specific resistance to the blast fungus. *Genetics* 2007, *176*, 2541–2549.

93. Sallam, A.H.; Tyagi, P.; Brown-Guedira, G.; Muehlbauer, G.J.; Hulse, A.; Steffenson, B.J. Genome-wide association mapping of stem rust resistance in *Hordeum vulgare* subsp. *spontaneum. G3 Genes Genomes Genet.* 2017, *7*, 3491–3507.

94. Ron, M.; Avni, A. The receptor for the fungal elicitor ethylene-inducing xylanase is a member of a resistance-like gene family in tomato. *Plant Cell* 2004, *16*, 1604–1615.

95. Bai, S.; Liu, J.; Chang, C.; Zhang, L.; Maekawa, T.; Wang, Q.; Xiao, W.; Liu, Y.; Chai, J.; Takken, F.L. Structure-function analysis of barley NLR immune receptor MLA10 reveals its cell compartment specific activity in cell death and disease resistance. *PLoS Pathog.* 2012, *8*, e1002752.

96. Casey, L.W.; Lavrencic, P.; Bentham, A.R.; Cesari, S.; Ericsson, D.J.; Croll, T.; Turk, D.; Anderson, P.A.; Mark, A.E.; Dodds, P.N. The CC domain structure from the wheat stem rust resistance protein Sr33 challenges paradigms for dimerization in plant NLR proteins. *Proc. Natl. Acad. Sci. USA* 2016, *113*, 12856–12861.

97. Hammond-Kosack, K.E.; Jones, J.D. Plant disease resistance genes. *Annu. Rev. Plant Biol.* 1997, *48*, 575–607.

98. Frazier, T.P.; Palmer, N.A.; Xie, F.; Tobias, C.M.; Donze-Reiner, T.J.; Bombarely, A.; Childs, K.L.; Shu, S.; Jenkins, J.W.; Schmutz, J. Identification, characterization, and gene expression analysis of nucleotide binding site (NB)-type resistance gene homologues in switchgrass. *BMC Genom.* 2016, *17*, 892.

CHAPTER 4: IDENTIFICATION AND CHARACTERIZATION OF MITOGEN

ACTIVATED PROTEIN KINASE (MAPK) GENES IN SUNFLOWER (*HELIANTHUS*

ANNUUS L.)

This chapter has been published in the Journal *Plants*:

Neupane, S.; Schweitzer, S.E.; Neupane, A.; Andersen, E.J.; Fennell, A.; Zhou, R.;

Nepal, M.P. Identification and Characterization of Mitogen-Activated Protein Kinase

(MAPK) Genes in Sunflower (*Helianthus annuus* L.). *Plants* 2019, *8*, 28.

Abstract

Mitogen Activated Protein Kinase (MAPK) genes, known to regulate biotic and

abiotic stresses in plants, are classified into three major subfamilies: MAP Kinase (MPK),

MAPK Kinase (MKK), and MAPKK Kinase (MKKK). The main objectives of this

research were to conduct genome-wide identification of MAPK genes in *Helianthus*

annuus and examine functional divergence of these genes in relation to those in nine

other plant species (*Amborella trichopoda*, *Aquilegia coerulea*, *Arabidopsis thaliana*,

Daucus carota, *Glycine max*, *Oryza sativa*, *Solanum lycopersicum*, *Sphagnum fallax,* and

Vitis vinifera) representing diverse taxonomic groups of plant kingdom. A Hidden

Markov Model (HMM) profiling of the MAPK genes utilized reference sequences from

A. thaliana and *G. max*, yielding a total of 96 MPKs and 37 MKKs in the genomes of *A.*

trichopoda, *A. coerulea*, *C. reinhardtii*, *D. carota*, *H. annuus, S. lyccopersicum*, and *S.*

fallax species. Among them, 28 MPKs and eight MKKs were confirmed in *H. annuus*.

Phylogenetic analyses revealed four clades within each subfamily. The transcriptomic

data showed that at least 19 HaMPK and seven HaMKK genes were induced in response

to salicylic acid (SA), sodium chloride (NaCl), and polyethylene glycol (Peg) in leaves

and roots. Of the seven published sunflower microRNAs, five microRNA families are involved in targeting eight MPKs. Additionally, we discussed the need for using MAP Kinase nomenclature guidelines across plant species. Our identification and characterization of MAP Kinase genes will have implication in sunflower crop improvement, and in advancing our knowledge of the diversity and evolution of MAPK genes in the Plant Kingdom.

Keywords: Abiotic stress; evolution of gene families; homology assessment; MAP Kinase cascade genes; MAPK nomenclature; sunflower; RNA-seq

4.1. Introduction

Plant responses to abiotic and biotic stresses involve protein kinase molecules that are crucial to signal transduction pathways [1]. The protein kinase molecules are involved in phosphorylation of Serine/Threonine and Tyrosine sidechains of proteins [2]. Among these protein kinases, Mitogen-Activated Protein Kinase (MAPK) cascade genes are key components of signal transduction pathways in animals, plants, and fungi [3] that help transduce extracellular signals to intracellular responses [4]. Discovered in 1986, the MAPK gene family was originally found in animal cells as a microtubule associated protein kinase [5]. The first reports of plant MAPK gene family in 1993, identified MsERK1 in alfalfa [6] and D5 kinase in pea [7]. MsERK1 is believed to play a role as an inducer of mitosis in root nodules during symbiosis by *Rhizobium* and D5 kinase as a cell cycle regulator in pea [6, 7]. In addition to such roles in cell proliferation and cell differentiation, MAPK genes are involved in regulating various biotic (e.g. bacteria,

fungi, viruses) and abiotic stress (e.g. light, drought, UV, salinity, pH, cold) responses [8].

The stress signals trigger the MAPK cascade which is composed of reversibly phosphorylated kinases such as MAP Kinase (MAPK, MPK), MAPK Kinase (MAP2K, MAPKK, MKK), and MAPKK Kinase (MAP3K, MAPKKK, MKKK) [9, 10]. The MKKKs constitute relatively larger gene family constituting three sub-groups of genes: the MEKKs, Rafs, and ZIKs [11]. Each of these proteins in the cascade is interlinked and is activated through the recognition and phosphorylation of a specific serine/threonine amino acid motif [12]. An external or internal stimulus triggers the first step, an activation of an MKKK member, through receptor-mediated phosphorylation or intermediate bridging factors or interlinking MKKKs [10]. The phosphorylated MKKK member induces the activation of MKK through the phosphorylation of two serine or threonine amino acid residues in the conserved motif S/TxxxxxS/T [10]. The activated MKKs, which are dual specificity kinases, in turn, trigger the phosphorylation of MPKs at the Thr-Asp/Glu-Tyr [T(D/E)Y] motif located in the activation loop (T-loop) between kinase subdomains VII and VIII [3, 10, 13]. Apart from T(D/E)Y motif in many plant species, some other variants such as T(Q/V/S)Y, T(/Q/R)M, MEY, TEC in the activation loop have also been reported [1]. The MPK members phosphorylate a variety of substrates, including transcription factors, protein kinases, and cytoskeleton proteins [10, 14]. The activation of the MAPK cascade genes induces the translocation from the cytoplasm to the nucleus [15], further enacting the specific cellular response to the external stimuli through gene activation and inactivation. The detail illustration of MAP

Kinase signaling pathway in response to diverse abiotic and biotic stresses in plants is represented in Figure S1 adapted from various studies [16, 17, 18, 19, 20, 21, 22, 23, 24].

The advent of sequencing technologies and rapid progress on bioinformatics tools has assisted the sequencing of the plant genomes at a faster pace. Genome-wide identification of MPKs and MKKs has been documented in various plant species including both model and crop species [14, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39]. Previous identification and characterization of MAP Kinase cascade proteins in rice, *Arabidopsis*, and other plants [4, 39, 40] provide a wealth of information for comparative analyses of these proteins in species that have yet to be studied. The availability of the complete genome sequences from each of the major plant groups such as Asterids (*Daucus carota* [41], *Helianthus annuus* [42], *Solanum lycopersicum* [43]), Amborellales (*Amborella trichopoda* [44]), Ranunculales (*Aquilegia coerulea* [45]), Bryophyte (*Sphagnum fallax* [46]), and Algae (*Chlamydomonas reinhardtii* [47]) allowed us to identify the MPK and MKK genes of these species and assesses phylogenetic relationships. Domesticated sunflower is the fourth most important oilseed crop in the world (http://www.fao.org/) and can adapt to diverse environmental conditions such as drought and maintain the stable yields [48]. Thus, MAPK gene family might play important role in sunflower to help thrive in changing climate. The research was carried out with two major objectives: (a) detailed identification, nomenclature and functional characterization of MPK and MKK genes in *H. annuus*, (b) assess phylogenetic relationships of MPK and MKK genes of *H. annuus* with that of *A. coerulea*, *A. trichopoda*, *C. reinhardtii*, *D. carota*, *S. fallax*, and *S. lycopersicum* and including the homologs from relatively better-studied plant species from Rosids (*A. thaliana*, *G. max*, *O.sativa,* and *V. vinifera*). Findings from this

study might support further efforts in crop improvement focused on the development of

cultivars that maintain yield when challenged by biotic and abiotic stresses as well as

understand the evolution pattern of MAPK gene family in sunflower and other plant

species.

4.2. Materials and Methods

4.2.1. Retrieval and Identification of Putative MAP Kinase Cascade Genes

Genome-wide identification of MPK and MKK cascade genes were done using

protein sequences of *A. coerulea* (v 3.1), *A. trichopoda* (v 1.0), *C. reinhardtii* (v 5.5), *D.*

carota (v 2.0), *H. annuus* (r 1.2) , *S. fallax* (v 0.5), and *S. lycopersicum* (iTAG2.4)

obtained from Phytozome database [45]. The protein sequences for sunflower were

accessed from sunflower (INRA inbred genotype XRQ; *H. annuus* r1.2) whose genome

is 3.6 gigabases and encode 52,243 proteins distributed over 17 chromosomes [42]. The

20 *MPK*s and ten *MKK*s sequences of *A. thaliana* [25] and 38 MPKs and 11 MKKs

sequences of *G. max* [26] were used as reference sequences for the identification of *MPK*

and MKK proteins of sunflower. The multiple sequence alignment file of these reference

sequences was employed in HMM profiling using the program HMMER version 3.1b2

[49] at a threshold expectation value of 0.01. MPK and MKK genes were further

identified using InterProScan Version 5.27 [50], Pfam ID [51], and PROSITE ID

(http://prosite.expasy.org/). The proteins with PfamID of MAPK domain (PS01351),

ATP-binding domain (PS00107), and protein kinase domain (PS50011), serine/threonine

protein kinase active site (PS00108) were used for identification of corresponding MPK

and MKK proteins (Figure 4.1). Multiple expectation maximization for motif elicitation

(MEME) [52] and multiple sequence alignment analysis was performed to confirm the

presence of signature motifs (a) the phosphate binding P-loop, GxGxxG [1], where ATP

binds in protein kinases (b) the catalytic C-loop, D(L/I/V)K, found within the S/T PK

active site signature, and (c) the activation- or T-loop, T(D/E)Y in MPK and

GTxxYMSPER in MKK proteins. The following parameters for MEME were employed:

maxsize: 100,000, mod: zoops, nmotifs: 10, minw: 6, and maxw: 25. Further, MKK

genes were identified using BLAST [53], with an E-value cutoff of 0.01, in which *A.*

thaliana MKK sequences were used as a query, and the top ten hits for each *A. thaliana*

MKK query sequence were employed for MKK genes identification. The protein

theoretical molecular weight and isoelectric point were predicted using compute pI/Mw

tool available in ExPASy (http://au.expasy.org/tools). Subcellular localization of the

putative MPKs and MKKs genes of sunflower were analyzed using TargetP 1.1 [54].
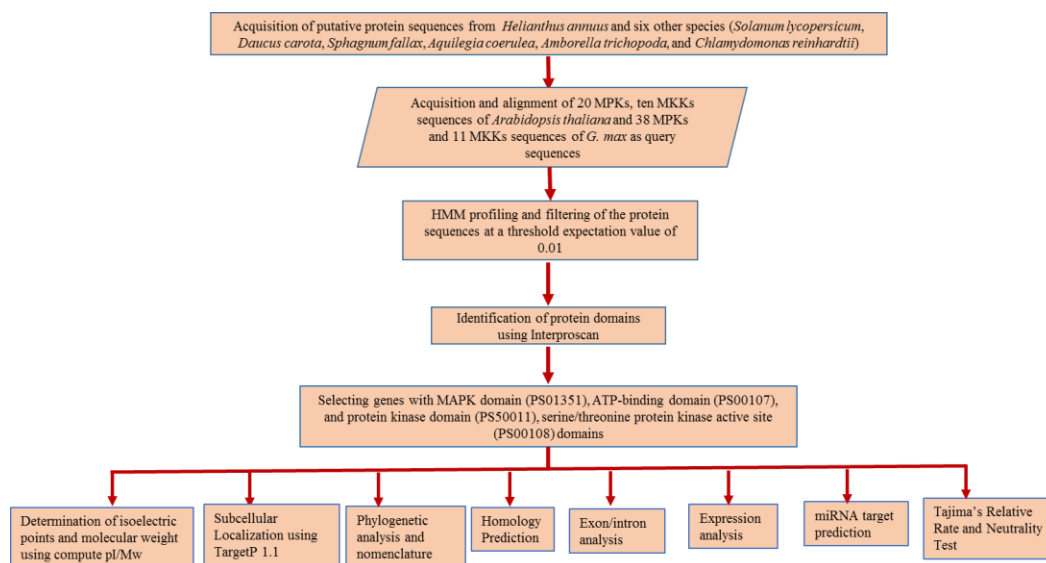
Figure 4.1. Schematic representation of *in silico* approaches used in the identification of MPK and MKK genes in seven plant species and their downstream analyses.

4.2.2. Phylogenetic Tree Construction and Homology Assessment

The multiple sequence alignment of identified MPK and MKK proteins of *H. annuus* was performed using CLUSTAL W [55] and MUSCLE [56] in Geneious [57] and subjected to phylogenetic analysis employing the maximum likelihood (ML) (100 replicates) using MEGA (version 7.0.14) [58]. The phylogenetic analyses employed the best evolutionary model (resulted from the ModelTest analysis using MEGA7) JTT+G+I (Jones–Taylor–Thornton with gamma distribution and invariant sites). Additionally, ML trees were constructed using MPK and MKK proteins of *H. annuus* with MPKs and MKKs of other plant species used in this study. The trees using MPK and MKK sequences were rooted with corresponding human MAP Kinase proteins (HsMAPK1 [GenBank: NP_002736.3] and HsMAPKK1 [GenBank: AAI37460.1]), respectively as an outgroup. Timetree was constructed using Reltime method [59] from MEGA7 to study the evolutionary divergence of MKK3 proteins belonging to all species under study. Following criteria were used for the construction of Timetree: constraints used: 3 [Divergence time: *O. sativa* and *A. trichopoda* (168-194 MYA), *G. max* and *H. annuus* (110-124 MYA), and *V. vinifera* and *A. thaliana* (105-115 MYA) obtained from http://www.timetree.org/ [60]], variance estimation method: analytical, statistical method: Maximum Likelihood, substitution model: JTT, rates among sites: 5 categories (+G, parameter = 0.6307), rate variation model allowed: ([+I], 0.00% sites), amino acid involved: 11, and total positions: 574 positions. Homology to MPKs and MKKs of other plants was assessed using BLASTp top-hit approach (https://blast.ncbi.nlm.nih.gov/Blast.cgi) with non-redundant protein sequences (nr) database.

4.2.3. Chromosomal Locations and Gene Structure

All 17 chromosome sequences of *H. annuus* accessed from the Phytozome database were uploaded into the program Geneious [57]. The chromosome locations of MPK and MKK genes were visualized using annotation file in Generic Feature Format (GFF) obtained from the annotation database of Phytozome. The exon-intron distribution pattern was obtained by the Gene Structure Display Server [61].

4.2.4. Nomenclature of MPKs and MKKs

Nomenclature of sunflower MPKs and MKKs was carried out using MAPK gene nomenclature guidelines [3, 4]. The nomenclature uses the following guidelines: a) the first letter (upper case) of the genus name followed by two to three letters of species (lower case) was used, b) a number was provided based on the homology to the *Arabidopsis* MAP Kinase cascade genes, and c) the number was followed by a hyphen and a number if paralogs are present. Such guidelines for nomenclature of MPKs and MKKs have been employed in many studies [1, 4, 26, 27, 33, 34, 35, 36, 62, 63, 64, 65]. In this study, we renamed GSVIVT01005924001 (VvMPK2) and GSVIVT0102277001 (VvMPK10) identified by Cakir and Kılıçkaya 2015 [37] as VvMPK22 and VvMPK23, respectively which were not identified in a study by Mohanta et al. 2015 [1].

4.2.5. Expression Analysis and miRNA Prediction of Sunflower MPKs and MKKs

The expression pattern of sunflower MPKs and MKKs was investigated using data accessed from NCBI SRA (Sequence Read Archive) SRP092742 [SRR4996815 (Peg treated pooled root samples), SRR4996819 (NaCl treated pooled root samples), SRR4996823 (Peg treated pooled leaf samples), SRR4996828 (Pooled Control root samples), SRR4996834 (NaCl treated pooled leaf samples), SRR4996836 (Pooled control

leaf samples), SRR4996839 (Salicylic Acid treated pooled leaf samples), and

SRR4996847 (Salicylic Acid treated pooled roots samples)]. These data are the result of

the application of one hormone treatment (0.05 µM SA), two abiotic stresses

[(polyethylene glycol 6000 (Peg) at 100g/l, which creates osmotic stress, and sodium

chloride at 100mM (NaCl)], and control [Dimethyl sulfoxide (DMSO) only] collected

from roots and leaves samples. The detailed experiment for generating transcriptomic

data is described in a study by Badouin et al. 2017 [42]. Briefly, roots and first leaves

were collected after six hours of treatments (SA, Peg, NaCl, and DMSO) applied to

sunflower (INRA inbred genotype XRQ) seedlings (two-week-old) grown in a

hydroponic system. The collection was repeated three times and was pooled after

separate RNA extractions in equimolar concentration. RNA sequencing of roots and

leaves samples done as non-oriented pair end libraries (2*76bp for roots and 2*100 for

leaves). The quality control of these reads was accessed by running the FastQC program

(version 0.11.3) [66], and trimming was done using Btrim64 (version 0.2.0) [67] to

remove low-quality bases (QC value > 20; 5-bp window size). High-quality pair-end

reads were mapped against the coding sequences of *H. annuus*

(*Hannuus: Hannuus_494_r1.2.transcript.fa.gz*) obtained from the Phytozome database

using *Salmon* tool ver.0.9.1 [68] accessed from Bioconda channel [69]. The codes that

were used for data processing are available as Appendix I; Supplementary 1. The

obtained transcript estimated quantification reads for each treatment were compared with

their respective reads from the controls to calculate the log_2Fold Change (log_2FC) and

visualized using integrated Differential Expression and Pathway analysis (iDEP 0.81

R/Bioconductor packages; http://bioinformatics.sdstate.edu/idep/) [70]. The heatmap was

generated using following criteria: Distance – correlation, Linkage –average and Cut-off

Z score – 4 to study the hierarchical clustering and expression pattern of MPK and MKK

genes in different tissues under different treatments. k –means clustering was done using

standardization normalization technique. To identify the potential miRNA targeting sites

The nucleotide sequences of the identified sunflower MPKs and MKKs were subjected to

a plant small RNA (psRNATarget) target analysis server [71] against seven published *H.*

annuus microRNAs using Schema V2 (2017 release) scoring schema.

4.2.6. Tajima's Relative Rate and Neutrality Test

Tajima's relative rate test [72] was conducted to study the statistical significance

of variations in molecular evolution in a different group of plants. The same MEGA files

used in phylogenetic tree construction were used in the program MEGA7. In this test,

three random sequences of either MPKs or MKKs of different plant species were selected

considering one of the sequences as the outgroup and χ^2 test statistic is applied. A *p*-value

of less than 0.05 was used to reject the null hypothesis of equal rates between selected

sequences of different plant groups. All positions containing gaps and missing data were

eliminated. Tajima's test of neutrality [73] was performed to understand and distinguish

the evolution pattern of randomly evolved MPKs or MKKs with non-randomly evolving

MPKs or MKKs. During the neutrality test, all positions with less than 95% site coverage

were eliminated. Therefore, fewer than 5% alignment gaps, missing data, and ambiguous

bases were allowed at any position. The grouping of A, B, C represent the statistical

groups, which should not be confused with MPKs or MKKs clades.

4.3. Results

4.3.1. Diversity of MPK and MKK Genes in Sunflower Relative to Other Species

After a careful examination of the signature motifs of the 2,419 sequences resulted from the HMM profiling using reference sequences of *A. thaliana* and *G. max* against 52,243 protein sequences of sunflower, we identified 28 MPKs (filtered from possible 244 MPKs) and 8 MKKs (filtered from possible 100 MKKs) (Table 4.1 and Table 4.2). We also used protein sequences of *A. coerulea*, *A. trichopoda*, *C. reinhardtii*, *D. carota*, *S. fallax*, and *S. lycopersicum* and identified their MPKs and MKKs, which are shown in Table S1-S2. The protein sequences identified including reference sequences used in this study and their identity in percentage are presented in Supplementary File S2. The diversity of MPK and MKK genes in the genome of *A. coerulea* (306.5Mb), *A. trichopoda* (706Mb), *C. reinhardtii* (111Mb), *D. carota* (421Mb), *H. annuus* (3600 Mb), *S. lycopersicum* (900Mb), and *S. fallax* (395 Mb) does not seem to correlate with genome size (Table 4.1).

Table 4.1 Abundance of MPK and MKK genes in the genomes of 11 species used in this study

Plant Species	Ploidy	Size of Genome (Mbs)[γ]	No. of loci [γ]	MPK	MKK
Amborella trichopoda[‡]	Diploid	706	26846	8	7
Aquilegia coerulea [‡]	Diploid	302	24823	11	5
Arabidopsis thaliana	Diploid	135	27416	20[a]	10[a]
Chlamydomonas reinhardtii[‡]	Haploid	111.1	17741	6	1
Daucus carota[‡]	Diploid	421	32,113	17	5
Glycine max	Tetraploid	975	56044	38[b]	11[b]
Helianthus annuus[‡]	Diploid	3600	52243	28	8
Oryza sativa	Diploid	372	39049	16[c]	8[c]
Solanum lycopersicum[‡]	Diploid	900	34727	15	5
Sphagnum fallax[‡]	Haploid/Diploid	395	26939	11	6
Vitis vinifera	Diploid	487	26346	14[d]	5[d]

[‡] = Plant species with MPKs and MKKs identified or revisited in this study

[γ] = References on the size of genome and number of loci *Amborella trichopoda* [44], *Arabidopsis thaliana* [74], *Aquilegia coerulea* [45], *Chlamydomonas reinhardtii* [47], *Daucus carota* [41], *Glycine max* [75], *Helianthus annuus* [42], *Oryza sativa* [76], *Solanum lycopersicum* [43], and *Sphagnum fallax* [46], and *Vitis vinifera* [77]

a= [10], b= [26], c=[4], d=[37]

Table 4.2. Sunflower MPK and MKK genes with their proposed name, GeneID, chromosomal location (Chr), strand direction (Str), start and end position of the genes on chromosome, protein length (PL), number of exon (Exo) and intron (Int), subcellular localization [Sl; M = Mitochondria and C = Chloroplast, - = Subcellular locations other than mitochondria or the chloroplast), isoelectric points (pI) and molecular weight (Mw)].

Name	Gene ID	Chr	Str	Start	End	PL	Exo	Int	Sl	pI	Mw
MPK											
HaMPK6-1	HanXRQChr01g0023391	Ha1	-	130301686	130292965	359	6	5	_	5.85	41581.61
HaMPK16-1	HanXRQChr03g0071491	Ha3	-	77378137	77372246	564	10	9	_	9.17	64059.43
HaMPK7	HanXRQChr03g0074811	Ha3	+	102410161	102406169	353	3	2	_	7.62	40274.83
HaMPK23-1	HanXRQChr03g0081221	Ha3	+	129978443	129973452	453	15	14	_	9.65	50392.28
HaMPK23-3	HanXRQChr03g0081391	Ha3	+	130506162	130500013	423	16	15	_	8.91	47648.79
HaMPK22	HanXRQChr04g0108301	Ha4	-	77321727	77315970	432	18	17	_	5.46	49633.87
HaMPK11-1	HanXRQChr04g0121371	Ha4	+	158781451	158778221	358	6	5	M	6.42	41228.21
HaMPK3-1	HanXRQChr05g0133161	Ha5	-	21064225	21061089	358	6	5	_	5.68	41323.35
HaMPK8	HanXRQChr05g0143371	Ha5	-	116774638	116767923	505	11	10	_	6.8	57051.89
HaMPK2	HanXRQChr05g0151241	Ha5	-	169574750	169571609	349	3	2	_	6.54	40295.67
HaMPK11-2	HanXRQChr06g0167011	Ha6	-	7104659	7099870	359	6	5	M	6.25	41336.17
HaMPK4	HanXRQChr06g0170261	Ha6	+	16894292	16893100	157	2	1	M	8.36	17702.54
HaMPK13-1	HanXRQChr06g0175501	Ha6	-	34635251	34631528	363	7	6	_	5.22	41353.31
HaMPK9-1	HanXRQChr06g0183531	Ha6	+	90706107	90699312	478	11	10	_	6.53	54442.91
HaMPK23-4	HanXRQChr08g0226701	Ha8	+	84318787	84308381	442	18	17	_	9.52	49480.06
HaMPK15	HanXRQChr08g0227231	Ha8	+	87599490	87591577	501	11	10	_	8.53	57073.07
HaMPK3-2	HanXRQChr08g0229941	Ha8	-	101013127	101009864	358	6	5	_	5.58	41298.31
HaMPK13-2	HanXRQChr08g0230171	Ha8	-	102808229	102804252	362	6	5	_	5.85	41552.83
HaMPK14	HanXRQChr09g0243011	Ha9	+	34673154	34669292	362	3	2	_	5.57	41423.42
HaMPK16-2	HanXRQChr09g0248301	Ha9	+	76212398	76202758	559	10	9	_	9.07	63370.4
HaMPK1	HanXRQChr09g0269211	Ha9	-	185086347	185083825	361	3	2	_	6.64	41831.44
HaMPK19-2	HanXRQChr11g0330461	Ha11	+	43791321	43784989	574	9	8	_	9.33	65344.85
HaMPK6-2	HanXRQChr11g0343001	Ha11	-	125967866	125963374	359	6	5	_	5.8	41553.72
HaMPK19-1	HanXRQChr13g0389781	Ha13	-	19048315	19044532	588	10	9	_	9.06	66613.36
HaMPK23-2	HanXRQChr13g0411961	Ha13	-	142634442	142625511	459	18	17	_	9.63	50984.95
HaMPK9-2	HanXRQChr14g0432771	Ha14	-	49683290	49679650	484	10	9	_	6.57	55530.13
HaMPK17	HanXRQChr15g0484561	Ha15	-	84424855	84420653	429	11	10	_	6.24	49909.6
HaMPK18	HanXRQChr15g0495321	Ha15	-	160155012	160149273	563	9	8	_	9.47	64374.62
MKK											
HaMKK9	HanXRQChr03g0087071	Ha3	-	148424902	148425825	308	1	0	M	6.75	34332.34
HaMKK4	HanXRQChr04g0094171	Ha4	+	471743	472816	351	1	0	C	9.04	38917.18
HaMKK6-1	HanXRQChr09g0238861	Ha9	-	9311933	9322916	357	8	7	_	6.76	39934.36
HaMKK5	HanXRQChr10g0311571	Ha10	+	219604899	219606004	355	1	0	C	9.25	39840.46
HaMKK6-2	HanXRQChr10g0318871	Ha10	+	244056044	244064185	355	8	7	_	7.13	39751.09
HaMKK2	HanXRQChr10g0319531	Ha10	-	245318274	245324118	371	9	8	_	5.43	40967.01
HaMKK1	HanXRQChr12g0354521	Ha12	-	1236278	1243005	358	10	9	_	5.77	39199.81
HaMKK3	HanXRQChr14g0450561	Ha14	-	141579116	141587170	520	12	11	M	5.79	68568.6

4.3.2. Gene Location, Subcellular Localization and Structural Variation of MPKs and MKKs in *H. annuus*

The MPK and MKK genes were distributed on all chromosomes of sunflower, with the highest of five genes in chromosome 3. The MPK genes were absent in chromosomes 2, 7, 10, 12, 10, 16 and 17 whereas, MKK genes were absent in

chromosomes 1, 2, 5, 6, 7, 8, 11, 13, 15, 16 and 17. Both MPK and MKK genes are completely absent in chromosomes 2, 7, 16 and 17. Only one HaMPK gene was found in chromosome 1 and 14 each; two HaMPKs in chromosome 4, 11, 13 and 15 each; three HaMPKs in chromosome 5 and 9 each, and four HaMPKs in chromosome 3, 6 and 8 each (Figure 4.2). Eight paralog pairs HaMPK3-1/3-2, HaMPK6-1/6-2, HaMPK9-1/9-2, HaMPK11-1/11-2, HaMPK13-1/13-2, HaMPK 16-1/16-2, HaMPK19-1/19-2 and HaMPK23-2/23-4 were located on different chromosomes. Only one paralog pair (HaMPK23-1/23-3) was present in the same chromosome (i.e. chromosome 3). Likewise, only one MKK gene was present in chromosomes 3, 4, 9, 12, and 14 while three MKKs were present in chromosome 10. The only paralog pair, HaMKK6-1/6-2 was present in different chromosomes. TargetP analysis showed that the proteins encoded by three MPKs (HaMPK11-1/11-2 and HaMPK4) and two MKKs (HaMKK9 and HaMKK3) were predicted to localize in mitochondria, two MKKs (HaMKK4 and HaMKK5) in the chloroplast, and the rest in subcellular locations other than mitochondria or the chloroplast (Table 4.2). Regarding the structural variation due to exons and introns, the number of exons in MPKs ranged from two (HaMPK4) to 18 (HaMPK22, HaMPK23-4/23-2) with an average of 8.9 exons per gene (Table 2, Figure S2). The number of exons in MKKs ranged from one (HaMKK9, HaMKK4, and HaMKK5) to 12 (HaMKK3) with an average of 6.25 exons per gene (Table 4.2, Figure S3).
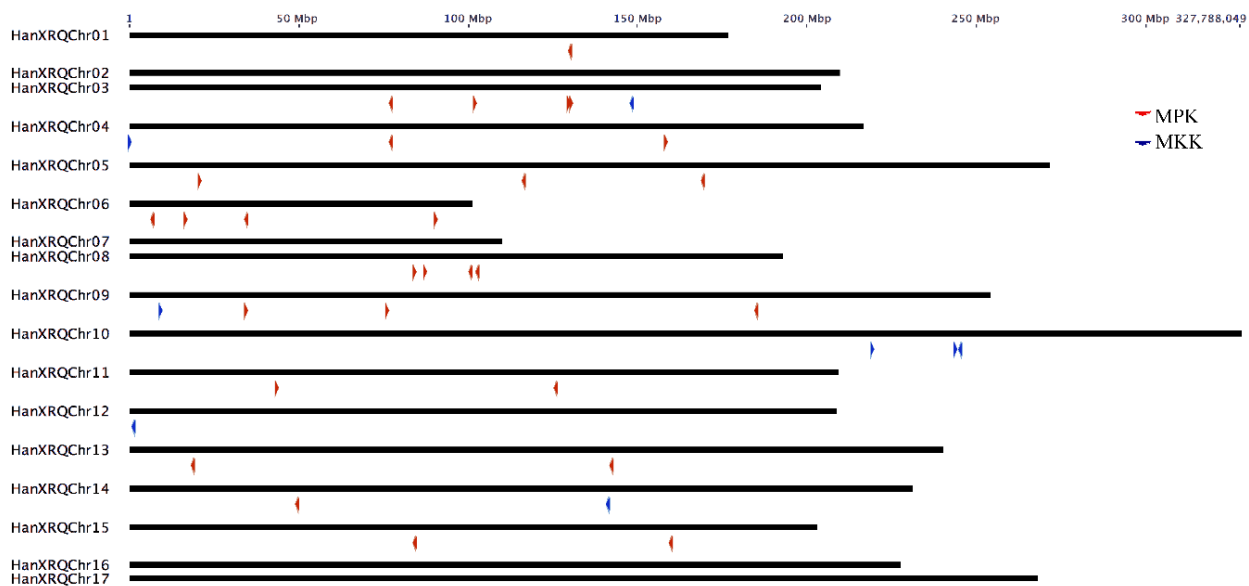
Figure 4.2. Chromosomal distribution of MPK and MKK genes in sunflower (*n* =17). Color-coded arrows represent MAP Kinase gene types and their orientation on the chromosome indicated by the black line.

4.3.3. Phylogenetic Analyses

Full-length amino acid sequences of MPKs and MKKs of sunflower, *Arabidopsis* and soybean were employed for evaluating evolutionary relationships as well as for nomenclature of the sunflower MPKs and MKKs. These sequences were subjected to multiple sequence alignment and subsequent phylogenetic analyses. Phylogenetic analyses included MPK and MKK gene sequences from, *A. coerulea*, *A. thaliana,* A. *trichopoda*, *C. reinhardtii*, *D. carota*, *G. max*, *H. annuus, O. sativa*, *S. lycopersicum,* and *V. vinifera.*

4.3.3.1. MPKs

Sunflower MPK (HaMPK) protein sequence length ranged from 349 to 588 amino acid (aa), except for HaMPK4, which was only 157 aa. The average length of

MPKs was 425 aa, with isoelectric points ranging from 5.22 (HaMPK13-1) to 9.65

(HaMPK23-1) and a predicted average molecular mass of 48523.772 Da (Table 4.1).

Twenty-eight HaMPKs identified in this study were nested into four clades (A-D; each

with bootstrap support > 70%) (Figure S4), which corresponded to their homologs in *A.*

thaliana and *G. max*, except for the Clade C MPK members of *Arabidopsis* (Table

S3). The Clade A members in this study include the previously identified group A and B

members of *A. thaliana* MPKs [3, 4]. Likewise, Clade B consists the members from

previously identified group C members of *A. thaliana* MPKs. In addition, Clade C

includes the members identified in group E of soybean MPKs [26]. The number

of HaMPKs in Clades A, B, C, and D were nine, four, five, and ten, respectively.

Sunflower MPK Clade C included five members with HaMPK22 (a homolog to

GmMPK22-1 and GmMPK22-2) and HaMPK23-1/23-2/23-3/23-4 (homologs to the

corresponding GmMPK23-1/23-2/23-4/23-4). The Clade A and B consisted members

with phosphorylation motif TEY (except for HaMPK23-1 and HaMPK23-2 that are

nested within Clade C), while those with the TDY motif were found in Clade C and D.

The sunflower MPK orthologs are shown in Table S4. The phosphate binding P-loop, the

catalytic C-loop, D(L/I/V)K, and activation- or T-loop, TxY in MPKs were defined as

(I/V/L)GxGx(S/F/G)GxV, HRD(L/I)KPxN and T(D/E)Y in sunflower, respectively.

Gene HaMAPK23-3 protein sequence had a variation in catalytic C-loop, D(L/I/V)K

motif as it possessed 'Phenylalanine (F)' instead of 'Leucine/ Isoleucine/Valine (L/I/V)'.

Other additional motifs such as VAIKKIxxxF were defined as

VA(I/V/M)KK(I/M)xxx(F/Y) in the protein sequences of MPKs. The MPKs that

belonged to clade 'C' possessed VA(I/V/M)KKMxxxY. The motifs 'DFGLAR' and

'TRWYRAPE' were found conserved in all of the MPKs of sunflower. HaMPK4 was the only member that lacked phosphate binding P-loop and VAIKKIxxxF motif. The structural analyses mapped onto phylogeny provided important insights into the duplication events. In HaMPK gene family, the number of introns ranged from one (HaMPK4) to 17 (three members from Clade C (HaMPK22, HaMPK23-4/23-2). The gene members showed a similar pattern of exon/intron structure within the clades. Majority of the HaMPKs (seven) in Clade A consist of six exons, and members, HaMPK13-1 and HaMPK4 had seven and two exons, respectively. In Clade B, all three members consisted of three exons. Three of the five members in Clade C possessed 18 exons, and HaMPK23-1 and HaMPK23-3 possessed 15 and 16 exons, respectively. Likewise, half of the gene members in Clade D (five) possessed ten exons, two (HaMPK19-2 and HaMPK18) possessed nine exons, as well as three genes (HaMPK8, HaMPK15, and HaMPK9-1), possessed 11 exons (Figure S1).

Phylogenetic analysis of full-length protein sequences was conducted to study evolutionary patterns of the MPKs in 10 plant species with sequences of *C. reinhardtii* (Figure 4.3). The MPKs were nested in four clades (Clade A-D; Table S3). Clade A is the second largest clade consisting 64 MPKs of MPK3/6/4/11/5/13/10 of all species under the study. Clade B consisted of 29 MPKs of MPK1/2/7 and 14. In cases of *S. lycopersicum* and *V. vinifera*, two species contain MPK1 and MPK7 in Clade B. Thus, MPK2 and MPK14 are absent in two species but not only MPK2. In addition, *A. trichopoda* has only AmtMPK14 in Figure 4.3. Therefore, MPK1/2/7 of *A. trichopoda* is absent. The MPK14 of *V. vinifera* and *D. carota*, MPK2 of *S. lycopersicum* and *V. vinifera*, and MPK7 of *A. trichopoda* were absent. The smallest clade, Clade C consisted

of 18 members of MPK22 and MPK23 from *H. annuus, G. max, S. lycopersicum, V. vinifera, S. fallax*, and *C. reinhardtii*. All the members of Clade A and B consisted TEY motif, whereas some members of Clade C (HaMPK23-1/23-4, GmMAPK23-1/23-2/23-3/23-4, and VvMPK22) consisted TEY motif. The largest clade, Clade D consisted 70 MPKs of MPK16/18/19/20/21/17/9/8/15, and MPK13 of *C. reinhardtii*. All clades had moderate to strong support (bootstrap values ranging from 80 to 100%). Figure 4.4a and Supplementary File S3 show the motifs related to P-loop, Catalytic C-loop, and activation or T-loop representing variations in clades A-D including other predicted conserved domains of MPK group proteins. In addition, the clade divergence was also based on the common docking site, which is important for downstream target proteins. Clade A consisted of K-M-L-V-F-D-P-N-K-R-I-V-E-E-A-L, Clade B consisted of K-M-L-V-F-D-P-S-K-R-I-S-V-T-E-A-L, Clade C consisted of S-L-C-S-W-D-P-C-K-R-P-T-A-E-E-A-L, and Clade D consisted of R-L-L-A-F-D-P-K-D-R-P-T-A-E-E-A-L consensus common docking sites (Table 4.3).

Table 4.3 Consensus Common Docking Sites in the MPK proteins belonging to clades A-D.

Clades	Consensus common docking sites
Clade A	K-M-L-V-F-D-P-N-K-R-I-V-E-E-A-L
Clade B	K-M-L-V-F-D-P-S-K-R-I-S-V-T-E-A-L
Clade C	S-L-C-S-W-D-P-C-K-R-P-T-A-E-E-A-L
Clade D	R-L-L-A-F-D-P-K-D-R-P-T-A-E-E-A-L
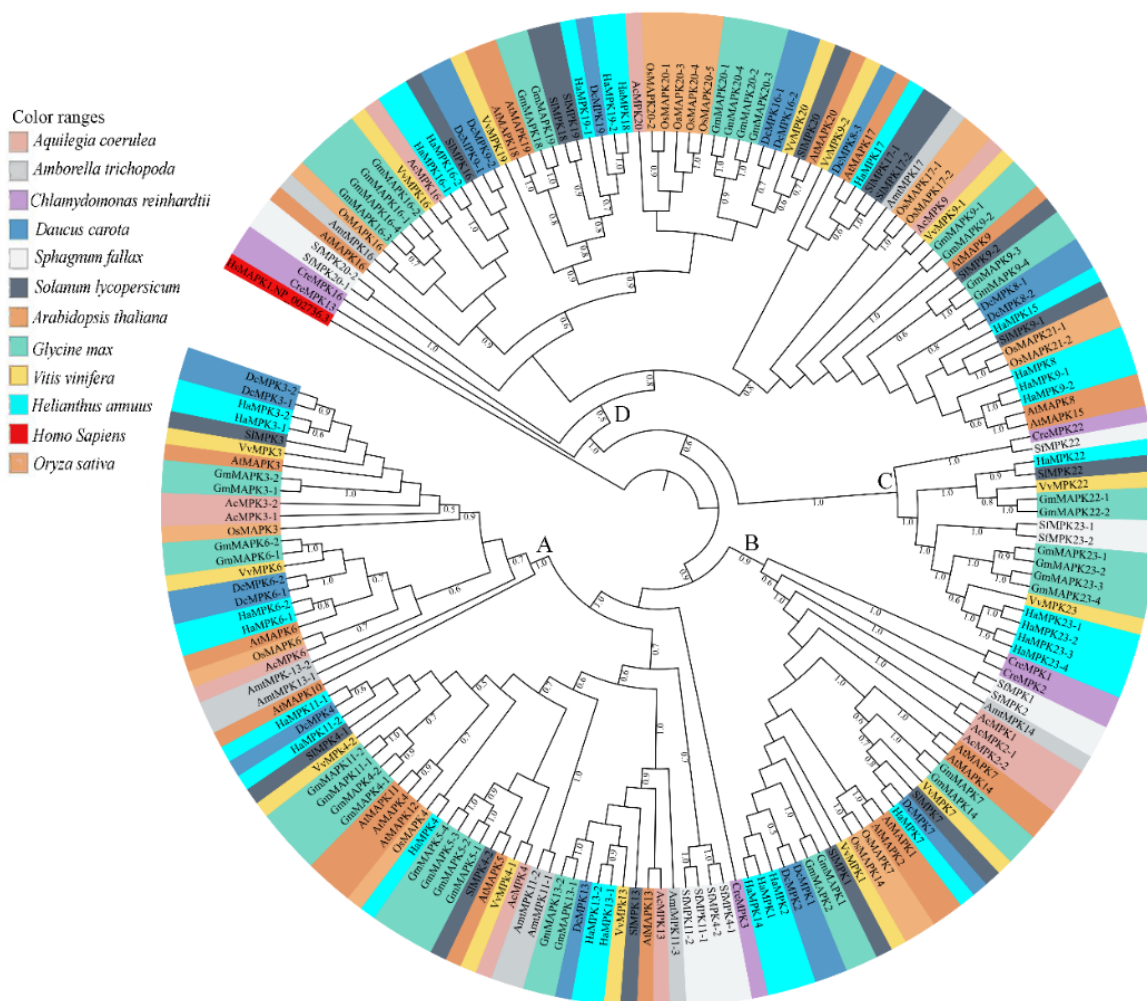
Figure 4.3 Maximum Likelihood (ML) tree constructed using full length amino acid sequences from *Amborella trichopoda* (Amt), *Arabidopsis thaliana* (At), *Aquilegia coerulea* (Ac), *Chlamydomonas reinhardtii* (Cre), *Daucus carota* (Dc), *Glycine max* (Gm), *Helianthus annuus* (Ha), *Oryza sativa* (Os), *Solanum lycopersicum* (Sl), and *Sphagnum fallax* (Sf), and *Vitis vinifera* (Vv) MPK proteins. Phylogenetic analysis with 100 bootstrap replicates was performed in the program MEGA 7. *H. sapiens*, HsMAPK1 (GenBank: NP_002736.3) was used as an outgroup. Different species are color-coded, and the MPK clades are labeled A-D.

Figure 4.4. P-loop, Catalytic C-loop, and activation or T-loop motifs representing variations in clades A-D. Panel a = MPK and Panel b = MKK

4.3.3.2. MKKs

Sunflower HaMKK protein sequence length ranged from 308 to 520 aa. The average length of proteins for MKKs was 372 aa with isoelectric points ranging from 5.43 (HaMKK2) to 9.25 (HaMKK5), and a predicted average molecular mass of 42688.86 (Table 4.1). Corresponding with their homologs in *Arabidopsis* and *G. max,* the eight identified HaMKKs are divided into four distinct clades (Figure S5). The MKK homologs of MKK1/2/6-1/6-2/3/4/5/9 were only found in sunflower. The clades divergence followed serine/threonine amino acid motif patterns in sunflower. For example, Clade A contained SxxxxxS/TxxxxxT, Clade B with SxxxxxTxxxxxT, Clade C

with SxxxxxTxxxxxS, and D with SxxxxxSxxxxxT. The HaMKKs in clade A, B, C, D were four, one, two, and one, respectively (Table S5). The orthologs of identified MKKs of sunflower in different plant species are represented in Table S6. In HaMKK gene family, the number of introns ranged from zero (HaMKK9, HaMKK4, and HaMKK5) to 11 (HaMKK3) (Table 4.2, Figure S3). Clade A members HaMKK6-1 and HaMKK6-2 consisted of eight exons, and are paralogs to each other. Remaining Clade A members, HaMKK2 and HaMKK1 consisted of nine and ten exons, respectively. The only member of Clade B, HaMKK3 consisted of twelve exons. Interestingly, without introns HaMKK9, HaMKK4 and HaMKK5 belonging to clade C and D had only one exon.

Phylogenetic analysis of full-length MKK amino acid sequences from the plant species with sequences of *C. reinhardtii* under this study revealed four distinct clades (Clade A- D, Figure 4.5). Figure 4.4b and Supplementary File S4 show the motifs related to P-loop, Catalytic C-loop, and activation or GTxxYMSPER representing variations in clades A-D including other predicted conserved domains of MKK group proteins. The largest clade, Clade A consisted of 26 MKKs belonging to MKK1, MKK2, and MKK6 members. While MKK3 orthologs formed Clade B consisting 12 MKKs, MKK4 and MKK5 with 16 members formed Clade C. Gene MKK4 is absent in *S. lycopersicum*, *V. vinifera*, and *D. carota*, *C. reinhardtii* species. The MKK7, MKK8, MKK9, and MKK10 formed Clade D consisting 16 of the total MKKs under study. With respect to all MKKs belonging to ten species, the phosphate-binding P-loop, the catalytic C-loop, D(L/I/V)K, and activation- or T-loop, (S/T)xxxxx(S/T) were varied according to the divergence of the. The GTxxYMSPER motif was well conserved in all species except for the OsMAPKK6 and AmtMKK6 with GTxxYMAPER in Clade A and OsMAPKK10-1 in

Clade D with GTxxYMSPEK. The ATP binding signature in MKK of sunflower

terminates with ALK except for GmMAPKK6-1 (completely absent), CrMKK3 with

AVK, VvMKK4 with ANT, OsMAPKK10-1 (completely absent), and OsMAPKK10-1

with AVK. The Timetree based on the 11 MKK3 (each MKK3 protein from all species

belonged to Clade B) sequences shows the evolutionary divergence across all species

under study. Upon use of three constraints of divergence between *O. sativa* and *A.*

trichopoda (168-194 MYA), *G. max* and *H. annuus* (110-124 MYA), and *V. vinifera* and

A. thaliana (105-115 MYA), the approximate divergence of these MKK3 proteins across

species has been found. For instance, DcMKK3 and SlMKK3 diverged 90.70 MYA from

HaMKK3 (Figure S6).
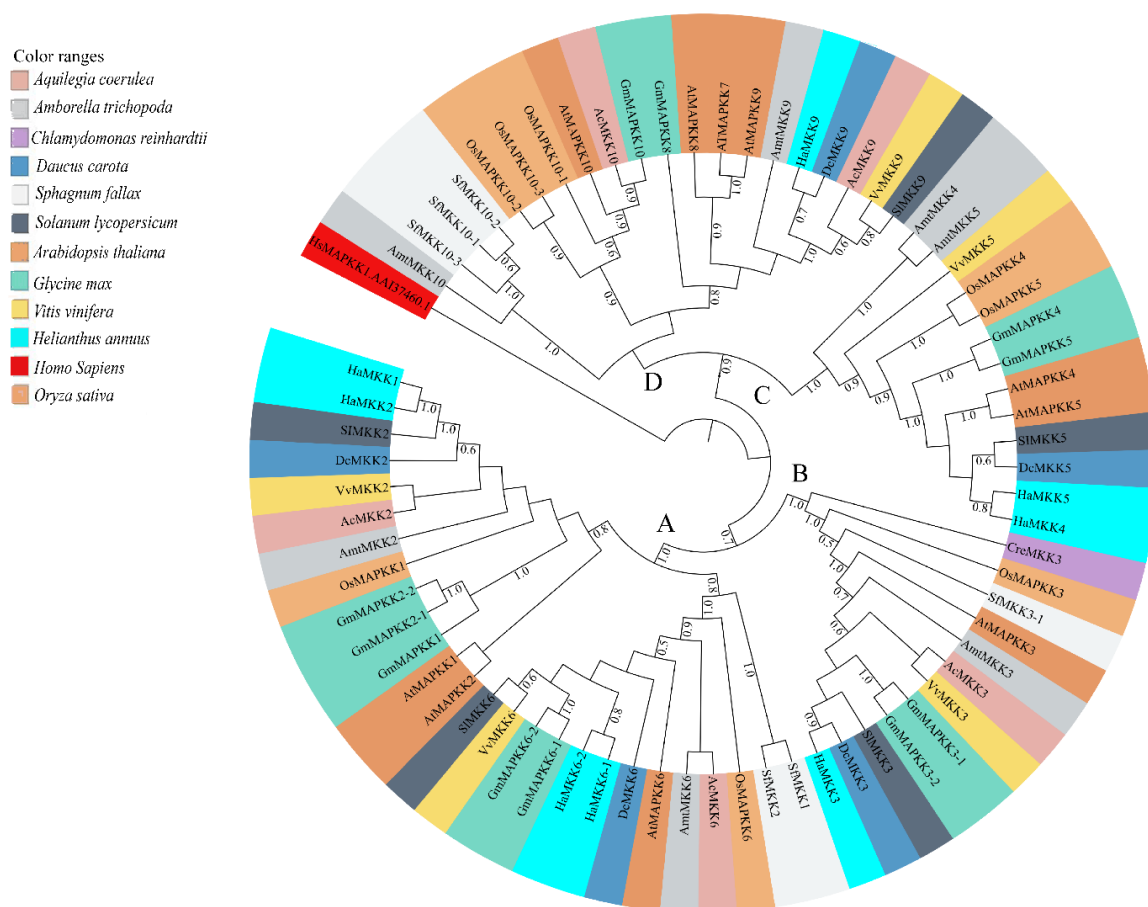
Figure 4.5. Maximum Likelihood (ML) tree constructed using full length MKK amino acid sequences from *Amborella trichopoda* (Amt), *Arabidopsis thaliana* (At), *Aquilegia coerulea* (Ac), *Chlamydomonas reinhardtii* (Cre), *Daucus carota* (Dc), *Glycine max* (Gm), *Helianthus annuus* (Ha), *Oryza sativa* (Os), *Solanum lycopersicum* (Sl), and *Sphagnum fallax* (Sf), and *Vitis vinifera* (Vv). Phylogenetic analysis with 100 bootstrap replicates was performed in the program MEGA 7. *Homo sapiens*, HsMAPKK1 (GenBank: AAI37460.1) was used as an outgroup. Different species are color-coded, and the MKK clades are labeled A-D.

4.3.4. Expression Analysis and miRNA Prediction of Sunflower MPKs and MKKs

The functional analysis of both HaMPKs and HaMKKs was studied using RNA seq data available in NCBI. Since the sunflower genome was available recently, the expression data in the public database were not found for pathogen stress. We

investigated the expression pattern of MPKs and MKKs in leaves and roots treated with one hormone treatment (SA) and two abiotic stresses (NaCl and Peg). We did observe expression patterns for all HaMPKs and HaMKKs except for HaMPK4 (Supplementary File S5). The *k-* means clustering result showed that the HaMPKs and HaMKKs were clustered into four groups (Figure S7 and Table S7). The Cluster A consisted of seven HaMPKs (from Clades A, B, and D) and four HaMKKs (from Clades A, B, and C). Cluster B consisted of three HaMKK genes (from Clades A and D) and two HaMPK genes (from Clade A). Cluster C consisted of three genes belonging to both HaMPKs (from Clades A and D) and one HaMKK (from Clade C). Cluster D consisted of 15 genes belonging to HaMPKs (belonging to clades A-D). The $\log_2$FC for each gene and hierarchical clustering of HaMPKs and HaMKKs representing the functional divergence of these genes are represented in Figure S8 and Figure 4.6, respectively. Some genes were upregulated in response to the treatments compared to the control of their respective tissues. For instance, in leaves, HaMKK5, HaMKK6-2, HaMPK3-2, HaMPK11-1, HaMPK3-2, HaMPK14, HaMPK1, HaMPK6-2, HaMPK19-1, and HaMPK18 showed $\log_2$FC > 1 in response to Peg; HaMKK5, HaMKK6-2, HaMPK11-1, HaMPK14 showed $\log_2$FC > 1 in response to NaCl; HaMPK11-1 showed $\log_2$FC > 1 in response to SA. In roots, HaMKK4, HaMKK1, HaMKK2, HaMPK3-2, HaMPK13-2, HaMPK23-2, HaMPK9-2 and HaMPK11-2 showed $\log_2$FC > 1 in response to Peg; HaMKK9, HaMPK13-2, HaMPK6-1, and HaMPK3-1 showed $\log_2$FC in range of 0.7 to 1.45 in response to SA; HaMPK6-1, HaMPK2, HaMPK23-2, and HaMPK17 showed $\log_2$FC > 0.9 in response to NaCl. In contrast, some genes were downregulated in response to the treatments compared to the control of their respective tissues. For example, in leaves,

HaMKK9, HaMKK2, and HaMPK13-2 showed $\log_2 FC$ in range of -0.6 to -0.8 in

response to Peg; HaMKK9, HaMPK7, HaMPK23-1 showed $\log_2 FC$ in range of -0.6 to -

0.8 in response to NaCl; HaMKK4, HaMPK7, and HaMPK11-2 showed log2fold change

in range of -0.58 to -2.11 in response to SA. Likewise in roots, HaMPK14 showed

$\log_2 FC$ of -0.53 in response to Peg; HaMKK6-2, HaMPK13-2, HaMPK14, and

HaMPK9-2 showed log2fold change in range of -0.62 to -1.50 in response to NaCl;

HaMPK14, HaMPK19-1, and HaMPK9-2 showed $\log_2 FC$ in range of -0.68 to -1.6 in

response to SA. In addition, the expression of HaMPKs, HaMKKs showed functional

divergence in response to stresses as the clustering of these genes in a heatmap was not

according to the clading pattern in phylogenetic trees. The potential miRNA target sites

in MPKs and MKKs identified using psRNATarget server revealed five (han-

miR156a/b/c, han-miR160a, han-miR3630-5p) of seven miRNA families that may be

involved targeting sunflower MPKs only (Table S8). HaMPK16-2, HaMPK11-1, and

HaMPK23-3 were targeted by both miRNAs (han-miR156a/b).
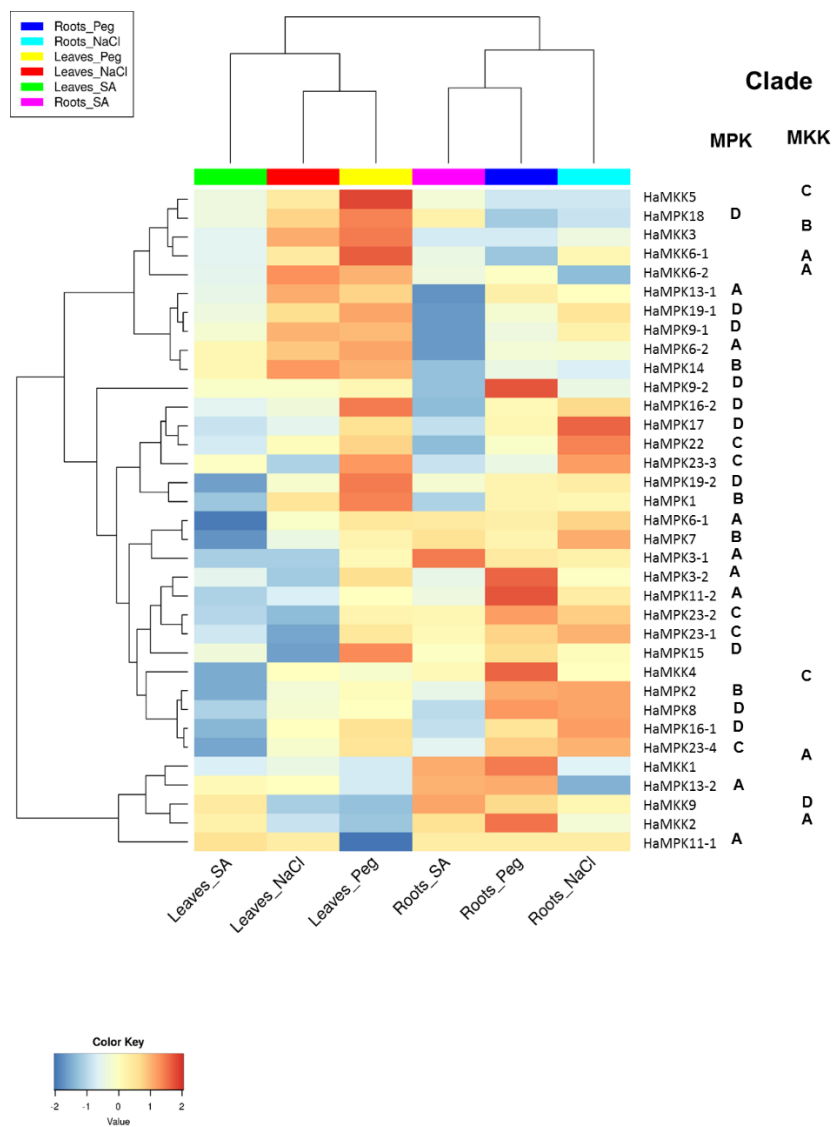
Figure 4.6. Expression profile of sunflower MPK and MKK genes visualized as a heatmap, with clade information. The heatmap was generated using $\log_2$FC and Z-score cut off of four, using iDEP [70]. The expression pattern is in response to Salicylic Acid (SA), salt (NaCl) and polyethylene glycol (Peg) in leaves and roots.

4.3.5. Tajima's Relative Rate and Neutrality Tests on MPKs and MKKs

Separate statistical analyses were performed selecting three random sequences from MPKs and MKKs group. For Tajima's relative rate test for MPKs and MKKs, the sequences were selected from the species representing a diverse taxonomic group: monocot, a dicot, basal angiosperm, bryophytes, and algae. For the analysis of MPK genes following a group of sequences were selected: a. OsMAPK4 (monocot) and HaMPK6 (dicot) with AmtMPK13-1 (basal angiosperm) b. OsMAPK4 (monocot) and HaMPK6 (dicot) with sequence SfMPK4-1 (bryophyte), and c. OsMAPK4 (monocot) and HaMPK16-1 (dicot), with sequence CreMPK2 (algae) (Table S9). The plant group combination in column 1, 2 and 3 of MPKs resulted in a p-value of 0.01, 0.0053, and 0.0007 with a χ^2 value of 6.54, 7.78 and 11.46, respectively. In MKKs, following group of sequences were selected: a. OsMAPKK5 (monocot) and HaMKK6-1 (dicot) with AmtMKK6 (basal angiosperm), b. OsMAPKK5 (monocot) and HaMKK6-1 (dicot) with sequence SfMKK3 (bryophyte) c. OsMAPKK5 (monocot) and HaMKK6-1 (dicot) with CreMKK3 (algae) (Table S10). The plant group combination in column 1, 2 and 3 of MKKs resulted in a p-value of 0, 0.04965 and 0.05687 with a χ^2 value of 100.55, 3.85 and 3.36, respectively. Tajima's Relative Rate test is commonly used to analyze variation in both DNA and amino acid sequences [78]. This test has been applied to various genes belonging to different gene families such as MAPKs and WRKY transcription factors [1, 78]. In this study, the p-value (less than 0.05) and χ^2 statistic showed randomly selected sequences of MPKs and MKKs of different plant groups to be statistically significant as we rejected the null hypothesis of equal rates between selected sequences of different plant groups. The interpretation of Tajima's D is as follows: D= 0 (observed variation is

similar to expected variation which shows the evidence of no selection), D< 0 (presence of excessive rare alleles that suggests recent selection sweep and recent population expansion) and D> 0 (lack of rare alleles that suggest balanced selection and population contraction) [72, 73]. The values in the range of greater than 2 or less than -2 are considered to be statistically significant [72, 73]. In our study, Tajima's neutrality test statistics (D) was found to be 5.391062 for MPKs and 5.928839 for MKKs (Table S11). This suggests both MPKs and MKKs have undergone a balanced selection with contraction in population size. Also, the average heterozygosity of both MAPKs and MKKs is more than that of the segregating sites suggesting a high frequency of polymorphism.

4.4. Discussion

MAPK signaling in plants plays important roles in multifaceted biological processes such as growth, development and regulation of various environmental stresses [4, 34, 36, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90]. The MPK and MKK genes have been the strong candidates for studying the evolution of gene families in plant species as well [27, 28, 39, 91]. In this study, the HMM analysis of protein sequences and examination of the signature motifs resulted in the identification of 96 MPK and 37 MKK genes in *A. coerulea*, *A. trichopoda*, *C. reinhardtii*, *D. carota*, *H. annuus*, *S. fallax*, and *S. lycopersicum*.

4.4.1. Nomenclature of MPKs and MKKs

A recent study on various Triticeae species (wheat, barley, rye, and triticale) by Goyal et al. 2018 [35] has reported numerous discrepancies in MAPK nomenclature of wheat and barley and suggested a new name based on sequence homology. A consistent nomenclature of proteins, belonging to same gene family across species based on orthology, facilitates an easy prediction and understanding the function of the particular protein [92]. Cakir and Kılıçkaya 2015 [37] reported MAP kinase cascade genes in *V. vinifera* and confirmed orthology of VvMPK14, VvMPK12, VvMPK11, VvMPK13, VvMPK7, VvMPK3, VvMKK5, VvMKK3, VvMKK2 to *Arabidopsis* AtMPK6, AtMPK3, AtMPK13, AtMPK12, AtMPK16, AtMPK9, AtMKK3, AtMKK6, and AtMKK2, respectively. Likewise, MAP Kinase cascade genes analyses in *Ziziphus jujuba* [30] provided nomenclature of MAP kinase cascade genes based on the order of appearance in different groups in the phylogenetic tree and not based on orthology (or sequence homology) to *Arabidopsis* MAP kinase cascade genes. The proper nomenclature of these MAP Kinase cascade genes should be used following an orthology or sequence homology based MAPK gene nomenclature guidelines to maintain the consistency across the plant kingdom.

4.4.2. Diversity and the Phylogenetic Relationship of MPKs

Our identification of MPKs yielded a slight variation in the number of genes from the previous studies, for example, we identified 15 MPKs in *S. lycopersicum* which is different from Kong et al. 2012 [93], who reported 16 MPKs, and Mohanta et al. 2015 [1] who found 17 MPKs in the tomato genome. The number of AcMPKs, in this study, was 11 whereas, Mohanta et al. 2015 [1] reported only 10 AcMPKs. In *C. reinhardtii*, six

CreMPKs identified in this study were consistent with Mohanta et al. 2015 [1] whereas, Dóczi et al. 2012 [39] had reported only five CreMPKs. The variation in a number of genes within the same species in different studies might come as a result of different statistical and stringency parameters employed during HMM profiling and further downstream motif analysis. The detailed study of MPKs of *D. carota, A. trichopoda, S. fallax,* and *H. annuus* has never been reported in previous studies. The number of MPK genes in sunflower is higher than that of previously identified in other numerous plant species such as *Arabidopsis* (119Mb) [3], rice (420Mb) [94] and less than soybean (1100Mb) [26]. Even the size of the sunflower genome, which is believed to have undergone the first whole genome triplication approximately 38-50 MYA, and whole genome duplication approximately 29 MYA, is about 3.5 times larger [95] than that of the soybean genome, the number of MPKs is less in sunflower than soybean. Soybean has undergone two polyploidization events approximately 59 and 13 MYA [75, 96]. Thus, recent polyploidy in plants has caused the addition of extra copies of MAPK genes to their genome [97, 98]. A slightly lower number of MPKs in sunflower might be due to past polyploidization events, recent amplification of repetitive elements causing highly similar and related sequences [99] and also sunflower genome encodes 52,243 proteins [42], which is slightly less than soybean (56,044 proteins) [75].

Phylogenetic analysis of HaMPKs revealed four distinct clades which were consistent to the MPKs previously identified in *Arabidopsis* [100], poplar [101], rice [102], *Brachypodium distachyon* [33], *Malus domestica* [32], *Ziziphus jujuba* [30], Triticeae species [35], *Brassica rapa* [28], and *Fragaria vesca* [103]. In Clade A, Sunflower has extra one copy of MPK3, MPK6, MPK11, and MPK13 genes that might

be resulted because of duplications after the divergence from *Arabidopsis*. Such extra copies of these genes were also observed in soybean [26]. The two copies of MPK3 and MPK6 was also found in *D. carota*. The clading of sunflower and other species MPK genes with the characterized *Arabidopsis* MPKs suggest their potential role in respective functions. AtMPK3 is involved in various signaling pathways related to various stresses such as wounding and hypersensitive responses elicited by Avr-R gene interaction [8, 104]. The MAP kinase genes, IbMPK3 and IbMPK6 in sweet potato (*Ipomoea batatas*) and homologs of AtMPK3 and AtMPK6 provide resistance to *Pseudomonas syringae* pv. *tabaci* (*Pta*) bacteria in tobacco leaves and induced in various abiotic stresses as well [84]. In mays, ZmMPK3, a homolog of AtMPK3 is induced upon various environmental stresses [105]. Similarly, AtMPK4 and AtMPK6 are involved in response to abiotic and biotic stress such as cold, drought, touch and wounding that result in the production of reactive oxygen species in *Arabidopsis* [106, 107]. AtMPK4 is phosphorylated and activated by the upstream components AtMEKK1 and AtMKK2 upon cold and salt stress signaling in *Arabidopsis* [107, 108]. The Clade A also consists of AtMPK5, the homolog of which in rice, OsMPK5 is well characterized to regulate stress responses [109]. All copies of MPK1/2, MPK7/14 are retained in soybean in sunflower, soybean, and *Arabidopsis*. Among them, AtMPK1, AtMPK2, AtMPK7, AtMPK14 are phosphorylated by AtMKK3 upon abscisic acid application in *A. thaliana* plantlets [110]. AtMPK1 is induced upon salt stress whereas some MPKs in rice and alfalfa such as BWMK1 and TDY1, respectively, are activated upon wounding by pathogens [111, 112, 113]. *G. max* MAP kinase 1 (GMK1), a homolog of AtMAPK1, is activated in response to salt stress in soybean [114]. Likewise, a homolog of AtMPK7

in maize, ZmMPK7 is involved in the removal of reactive oxygen species upon induction by abscisic acid and hydrogen peroxide in maize [115]. Another homolog of AtMPK1 in *Hordeum vulgare* (HvMPK4) showed enhanced resistance to *Magnaporthe grisea* and enhanced tolerance to salt stress [85]. Clade C members include the homologs to *G. max* GmMAPK22-1/22-2 and GmMAPK23-1/23-2/23-3/23-4 [26] with no MPKs in Arabidopsis. A single copy of GmMAPK22-1/GmMAPK22-2 ortholog is retained in sunflower, and hence it is named as HaMPK22. Whereas, all copies of GmMAPK23-1/23-2/23-3/23-4 are retained in sunflower and hence named as HaMPK23-1/23-2/23-2/23-3/23-4. All the members of the Clade D consist of TDY motif in T-loop and are homologs to various *Arabidopsis* and soybean MPKs belonging to MPK16/19/18/8/15/17/9.

Gene members HaMPK3-1/3-2, HaMPK6-1/6-2, HaMPK9-1/9-2, HaMPK11-1/11-2, HaMPK13-1/13-2, HaMPK 16-1/16-2, HaMPK19-1/19-2 and HaMPK23-2/23-4 are present on different chromosomes, while only paralogs HaMPK23-1/23-3 are present in the same chromosome 3. Other MPKs such as AcMPK3-1/3-2, AcMPK2-1/2-2, DcMPK3-1/3-2, DcMPK6-1/6-2, DcMPK8-1/8-2/8-3, DcMPK9-1/9-2, SfMPK4-1/4-2, SfMPK20-1/20-2, SfMPK23-1/23-2, SlMPK4-1/4-2, SlMPK17-1/17-2, SlMPK9-1/9-2, AmtMPK13-1/13-2 are present on different chromosomes. The only AmtMPK11-1/11-2 pair is present in the same scaffold (AmTr_v1.0_scaffold00001) (Table S1). This suggests the possible role of segmental duplications and transposition events that played a crucial role in the evolution of MAPKs in sunflower and other plant species except for the pair HaMPK23-1/23-3 and AmtMPK11-1/11-2 pairs in which tandem duplication might have involved. Such features of segmental and tandem duplications in MPKs are

also evidently seen in many plant species such as soybean [26], apple [32], cotton [116].

Such duplications are the major reason for the expansion of the many gene families such

as Nucleotide-binding site-leucine-rich repeat (NBS-LRR), cytochrome P450 family,

transcription factors and many more [117].

4.4.3. Diversity and Phylogenetic Relationship of MKKs

Sunflower MKKs also formed four distinct clades (A-D) with previously

identified MKKs of *Arabidopsis* and soybean. These four clades (A- D) are consistent

with the MKKs of various plant species such as *Arabidopsis* [100], rice [102], poplar

[101], *B. distachyon* [33] and apple [32]. MKK clades consist of well-characterized MKK

proteins such as AtMKK1/2/3/4/5 [118, 119, 120, 121]. Clade A consists of HaMKKs

grouped with AtMKK1/6/2, GmMAPKK6-1/6-2, GmMAPKK1, GmMAPK2-1/2-2.

Sunflower and soybean have extra one copy of MKK6 than that of *Arabidopsis* and other

plant species under study including *S. fallax*. This suggests that extra one copy of MKK6

was not seen until soybean diverged from *Arabidopsis*. Also, the retention of at least one

copy in of MKK6 in all species suggests its important role in signaling mechanism during

various stresses. We did not find a copy of MKK2-2 in sunflower as found in soybean

(GmMAPKK2-2). The characterized AtMKK1 protein (orthologue of HaMPKK1) is

induced upon the application of various stresses such as wounding, drought, cold, and

high salinity in *Arabidopsis* seedlings [118]. AtMKK2 (ortholog of HaMKK2) is

activated upon cold and salt stress signaling in *Arabidopsis* and mediate the

phosphorylation of downstream MPKs [107]. The Clade B consists of MKKs from the

MKK3 proteins across all species under study including *C. reinhardtii*. All species have a

single copy of the MKK3 proteins except *G. max* with two copies (GmMAPKK3-1/3-2).

Two copies of MKK3 proteins in soybean is expected as they went two duplication events to be a tetraploid. The time tree based on Clade B (each from all species) revealed how MKK3 proteins are conserved and retained in algae, bryophyte, Amborellales, a monocot, Ranunculales, Rosids, and Asterids. The divergence time analysis of MKK3 with CreMKK3 as the outgroup, showed bryophyte and Amborellales being sister to the land plants and other extant species which is consistent to previous studies [122, 123] and follows the evolutionary pattern as shown in Angiosperm Phylogeny Website [124]. AtMKK3 is activated on various stresses such as cold, salt, hyperosmotic and ABA treatments [120]. This suggests the potential role of HaMKK3 in such stresses. The Clade C consists of both copies of AtMKK4 and AtMKK5 in only in *A. trichopoda*, *O. sativa*, and sunflower. However, *V. vinifera*, *S. lycopersicum* and *D. carota* consist copy of MKK5 (MKK4 group absent). AtMKK4 and AtMKK5 are activated in *Arabidopsis* that mediate cell death and production of hydrogen peroxide [119]. In clade D, the orthologs for MKK9 was found in all angiosperms except in soybean and *O. sativa*. Interestingly, we found three copies of MKK10 in *S. fallax* as in *O. sativa* and one copy of MKK10 in basal angiosperm, *A. trichopoda,* and Ranunculales, *A. coerulea.* We did not find any copy of MKK10 in sunflower, *S. lycopersicum*, *D. carota*, and *V. vinifera*. We observed HaMKK4/5/9 with one exon each that correlates to the At1g51660 (AtMKK4), At3g21220 (AtMKK5), and At1g73500 (AtMKK9) consisting one exon per gene (https://www.*Arabidopsis*.org/index.jsp). Also, members belonging to the Clade C and D in *Gossypium raimondii* had one exon in each [116]. This suggests that gene members belonging to Clade C and D encode proteins that are well conserved across plant species. Altogether, the diversity in the exon-intron structures might infer duplication events

caused the evolution of these genes under different environmental conditions. Also, AtMKK1 and AtMKK2 are involved in maintaining ROS homeostasis in *Arabidopsis* [121]. Since, the paralog pairs, HaMAPKK6-1/6-2, and SfMKK10-1/10-2 are present on their different respective chromosomes, we infer the possible role of segmental duplications.

4.4.4. Expression Analysis and miRNA Prediction

In this study, we explored the expression pattern of MPKs and MKKs of sunflower under one hormone treatment, SA and two simulated abiotic stresses, NaCl for salinity, and Peg for osmotic stress in leaves and roots from the publicly available RNA seq data. The expression of all sunflower MPKs and MKKs was detected in both leaves and roots except for HaMPK4. In response to hormone SA, HaMPK11-1 was upregulated in leaves; HaMKK9, HaMPK13-2, HaMPK6-1, and HaMPK3-1 were upregulated in roots; HaMKK4, HaMPK7, and HaMPK11-2 were down regulated in leaves; HaMPK19-1, HaMPK14, and HaMPK9-2 were downregulated in roots. It has been established that SA is directly involved in MAPK phosphorylation [125]. SA- induced protein kinase (SIPK; AtMPK6) and wound-induced protein kinase (WIPK; AtMPK3) are important in balancing salicylic acid or jasmonic acid during herbivore wounding [126]. In *Arabidopsis*, AtMKK9 and AtMPK6 play important role in leaf senescence which is a complex process caused by various factors including salicylic acid [127]. Also, ZmMPK3 in *Zea mays* is activated upon the application of SA hormone [128]. Thus, HaMPK3-1, HaMKK9, and HaMPK6-1 might play important role in leaf senescence and salicylic acid pathway in sunflower. In response to NaCl, HaMKK5, HaMKK6-2, HaMPK11-1 were upregulated in leaves; HaMPK14, HaMPK6-1, HaMPK2, HaMPK23-2, and

HaMPK17 were upregulated in roots; HaMKK9, HaMPK7, HaMPK23-1 were downregulated in leaves; HaMKK6-2, HaMPK13-2, HaMPK14, and HaMPK9-2 were downregulated in roots. Among them, HaMPK17 play an important role under salinity stress as its ortholog in *Gossypium hirsutum*, GhMPK17 was induced by salt, osmosis and abscisic acid [129]. The expression pattern of some genes depended on different parts of the plant such as HaMKK6-2 was upregulated in leaves and downregulated in roots in response to NaCl. In response to Peg, HaMKK5, HaMKK6-2, HaMPK3-2, HaMPK11-1, HaMPK3-2, HaMPK14, HaMPK1, HaMPK6-2, HaMPK19-1, and HaMPK18 were upregulated in leaves; HaMKK4, HaMKK1, HaMKK2, HaMPK3-2, HaMPK13-2, HaMPK23-2, HaMPK9-2 and HaMPK11-2 were upregulated in roots; HaMKK9, HaMKK2, and HaMPK13-2 were downregulated in leaves; HaMPK14 was downregulated in roots. This reveals at least 19 HaMPK and seven MKK genes were induced upon these treatments as compared to the control. Among them, some genes are induced upon multiple treatments. For example, HaMKK4 and HaMKK6-2 were induced upon both NaCl and Peg; HaMPK6-1 was induced upon NaCl (higher expression) and Peg (lower expression); HaMPK16-2 was induced upon both SA and NaCl. The functional divergence can be observed on both HaMPKs and HaMKKs as the hierarchical clustering pattern of expression of these genes do not follow the clading pattern in the phylogenetic trees except for few genes. For example, in MPKs, HaMPK22/23-3 that belonged to Clade C, HaMPK3-1/3-2/11-2 that belonged to Clade A, HaMPK9-2/16-2/17 that belonged to clade D showed hierarchical clustering for expression of these genes. However, only HaMKK6-1/6-2 that belonged to Clade A of MKK subgroup showed hierarchical clustering for expression of these genes. This shows the functional

divergence and convergence of the HaMPk and HaMKK genes within and among the clades under different stress responses. Among seven published *H. annuus* microRNAs, five families of miRNAs are involved in possibly targeting eight MPKs. We did not find any miRNAs targeting HaMKK genes. Previous studies reported the role of miRNAs in MAPK signaling pathways of animal systems in chronic myeloid leukemia [130], papillary thyroid carcinoma [131], *Caenorhabditis elegans* [132]. Not only in animals, but studies also reported the prediction of miRNAs targeting MAPK genes of plants such as *Gossypium hirsutum* (ghr-miR5272a regulating MAPKK6) [133] and *Oryza sativa* (miR1429_5p targeting MPK17-1 and miR531 families targeting various MKKK transcripts) [134].

4.5. Conclusion

This study represents the first genome-wide identification, analysis and nomenclature of MPKs and MKKs in *H. annuus*, *D. carota* and, *S. fallax* and reassessment of these genes in *A. coerulea*, *A. trichopoda, C. reinhardtii,* and *S. lycopersicum.* We identified 28 MPKs and eight MKKs in sunflower, studied their genomic architecture, phylogenetic relationships, and functions in relation to nine other plant species (including *A. thaliana*, *G. max*, *O. sativa,* and *V. vinifera*). Though sunflower genome with 3.6 gigabases is one of the largest among plants with available complete genome sequences of species under study, the MPKs and MKKs are slightly fewer than that in soybean, which has the genome size of 975 Mbs. The phylogenetic trees and analyses of three important motifs, P-loop, Catalytic C-loop, and T-loop showed that HaMPKs and HaMKKs could be classified to four clades which are comparable to those groups identified in *A. thaliana* and *G. max.* However, clades such as

Clade A, B, and C of MPKs consisted members from different group members of *A. thaliana* and *G. max.* Among MPKs and MKK genes studied, MKK3 group of proteins are well conserved and retained in all the species under study including the outgroup, C. *reinhardtii* which warrants further exploration of these proteins across a wide array of species. The transcriptomics data analyzed under hormone and abiotic stresses treatments revealed diverse expression pattern of sunflower MPKs and MKKs exhibiting dynamic role to adapt to changing environmental conditions. We observed functional divergence of the HaMPK and HaMKK genes within the gene members of the same clade. The results from this study are generally important for understanding diversity and evolution of MAPK gene family in plants and enhancing our knowledge of MAPK signaling pathway in sunflower. These findings can help cultivar improvement in sunflower through stress-tolerance breeding.

Conflicts of Interest: The authors declare no competing interests.

Author Contributions: SN and SES performed gene identification and downstream analyses and wrote the manuscript. MPN conceived and supervised the project and guided SN draft the manuscript. SES, AN, EJA, AF, and RZ provided useful insights to data analysis, interpretation and revision of the manuscript.

Supplementary Files: Supplementary files are available at https://www.mdpi.com/2223-7747/8/2/28/s1, Table S1: MPKs identified in different species in the present study. Table S2: MKKs identified in different species in the present study. Table S3: The abundance

of the MPK gene family of sunflower and ten other species belonging to different clades

A-D. Table S4: Sunflower MPKs with their orthologs in other plant species.

Table S5: The abundance of the MKK gene family of sunflower and ten other species

belonging to different clades A-D. Table S6: Sunflower MKKs with their orthologs in

other plant species. Supplementary Table S7: *k*-means clustering of HaMPK and HaMKK

genes. Table S8: miRNAs involved in targeting MPKs and MKKs of Sunflower. Table

S9: Tajima's relative rate test of MPKs. Table S10: Tajima's relative rate test of MKKs.

Table S11: Tajima's Neutrality Test of MPKs and MKKs. Supplementary Figure S1:

Illustrating a MAP Kinase signaling pathway in response to abiotic and biotic stresses in

plants adapted from various studies. Supplementary Figure S2: Exon/intron architecture

of HaMPK genes. The blue boxes and black lines indicate the exons and introns,

respectively. Supplementary Figure S3: Exon/intron architecture of HaMKK genes. The

blue boxes and black lines indicate the exons and introns, respectively. Supplementary

Figure S4: Maximum Likelihood analysis of full-length MPK amino acid sequences from

A. thaliana (At), *G. max* (Gm) and *H. annuus* (Ha) using MEGA7 program with 100

bootstrap replicates. *Homo sapiens*, HsMAPK1 (GenBank: NP_002736.3) was used as an

outgroup. MPK clades are labeled as A, B, C and D. Supplementary Figure S5: ML tree

of full-length MKK amino acid sequences from *A. thaliana* (At), *G. max* (Gm) and *H.*

annuus (Ha) in the program MEGA7, with 100 bootstrap replicates. *Homo sapiens,*

HsMAPKK1 [GenBank: AAI37460.1] was used as an outgroup. MKK clades are labeled

as A, B, C, and D. Supplementary Figure S6: RelTime tree constructed using each

sequence of MKK3 from all species belonging to Clade B of MKK group. Supplementary

Figure S7: Clustering of sunflower MPK and MKK genes in response to diverse

conditions as simulated by Salicylic Acid (SA), salt (NaCl) and polyethylene glycol (Peg) in leaves and roots. k-means clustering method was employed for clustering of genes.

Supplementary Figure S7: The $\log_2$FC values for each HaMPK and HaMKK genes in samples treated with Salicylic Acid (SA), salt (NaCl) and polyethylene glycol (Peg) in leaves and roots Supplementary File S1: Codes used for RNA seq data processing.

Supplementary File S2: Fasta sequences of the identified MPKs and MKKs including the reference sequences used in this study. Supplementary File S3: MEME predicted conserved motifs in MPK proteins belonging to different phylogenetic clades. Supplementary File S4: MEME predicted conserved motifs in MKK proteins belonging to different phylogenetic clades. Supplementary File S5: The transcript estimated quantification reads for each treatment and their $\log_2$FC values in the additional sheet.

References

1. Mohanta, T.K.; Arora, P.K.; Mohanta, N.; Parida, P.; Bae, H. Identification of new members of the MAPK gene family in plants shows diverse conserved domains and novel activation loop variants. *BMC Genom.* 2015, *16*, 58.
2. Wang, Z.; Cole, P.A. Catalytic mechanisms and regulation of protein kinases. In *Methods Enzymolgy*; Elsevier: Amsterdam, The Netherlands, 2014; Volume 548, pp. 1–21.
3. Ichimura, K.; Shinozaki, K.; Tena, G.; Sheen, J.; Henry, Y.; Champion, A.; Kreis, M.; Zhang, S.; Hirt, H.; Wilson, C. Mitogen-activated protein kinase cascades in plants: A new nomenclature. *Trends Plant Sci.* 2002, *7*, 301–308.
4. Hamel, L.-P.; Nicole, M.-C.; Sritubtim, S.; Morency, M.-J.; Ellis, M.; Ehlting, J.; Beaudoin, N.; Barbazuk, B.; Klessig, D.; Lee, J. Ancient signals: Comparative genomics of plant MAPK and MAPKK gene families. *Trends Plant Sci.* 2006, *11*, 192–198.
5. Sturgill, T.W.; Ray, L.B. Muscle proteins related to microtubule associated protein-2 are substrates for an insulin-stimulatable kinase. *Biochem. Biophys. Res. Commun.* 1986, *134*, 565–571.
6. Duerr, B.; Gawienowski, M.; Ropp, T.; Jacobs, T. MsERK1: A mitogen-activated protein kinase from a flowering plant. *Plant Cell* 1993, *5*, 87–96.
7. Stafstrom, J.P.; Altschuler, M.; Anderson, D.H. Molecular cloning and expression of a MAP kinase homologue from pea. *Plant Mol. Biol.* 1993, *22*, 83–90.

8. Nakagami, H.; Pitzschke, A.; Hirt, H. Emerging MAP kinase pathways in plant stress signalling. *Trends Plant Sci.* 2005, *10*, 339–346.

9. Tatebayashi, K.; Takekawa, M.; Saito, H. A docking site determining specificity of Pbs2 MAPKK for Ssk2/Ssk22 MAPKKKs in the yeast HOG pathway. *EMBO J.* 2003, *22*, 3624–3634.

10. Jonak, C.; Okresz, L.; Bogre, L.; Hirt, H. Complexity, cross talk and integration of plant MAP kinase signalling. *Curr. Opin. Plant Biol.* 2002, *5*, 415–424.

11. Tena, G.; Asai, T.; Chiu, W.-L.; Sheen, J. Plant mitogen-activated protein kinase signaling cascades. *Curr. Opin. Plant Biol.* 2001, *4*, 392–400.

12. Bardwell, L. *Mechanisms of MAPK Signalling Specificity*; Portland Press Limited: London, UK, 2006.

13. Lewis, T.S.; Shapiro, P.S.; Ahn, N.G. Signal transduction through MAP kinase cascades. In *Advances in Cancer Research*; Elsevier: Amsterdam, The Netherlands, 1998; Volume 74, pp. 49–139.

14. Janitza, P.; Ullrich, K.K.; Quint, M. Toward a comprehensive phylogenetic reconstruction of the evolutionary history of mitogen-activated protein kinases in the plant kingdom. *Front. Plant Sci.* 2012, *3*, 271.

15. Marshall, C. Specificity of receptor tyrosine kinase signaling: Transient versus sustained extracellular signal-regulated kinase activation. *Cell* 1995, *80*, 179–185.

16. Kissoudis, C.; van de Wiel, C.; Visser, R.G.; van der Linden, G. Enhancing crop resilience to combined abiotic and biotic stress through the dissection of physiological and molecular crosstalk. *Front. Plant Sci.* 2014, *5*, 207.

17. Ranty, B.; Aldon, D.; Cotelle, V.; Galaud, J.-P.; Thuleau, P.; Mazars, C. Calcium sensors as key hubs in plant responses to biotic and abiotic stresses. *Front. Plant Sci.* 2016, *7*, 327.

18. Pitzschke, A.; Schikora, A.; Hirt, H. MAPK cascade signalling networks in plant defence. *Curr. Opin. Plant Biol.* 2009, *12*, 421–426.

19. Müller-Xing, R.; Xing, Q.; Goodrich, J. Footprints of the sun: Memory of UV and light stress in plants. *Front. Plant Sci.* 2014, *5*, 474.

20. Davis, P.K.; Brachmann, R.K. Chromatin remodeling and cancer. *Cancer Biol. Ther.* 2003, *2*, 23–30.

21. Rejeb, I.B.; Pastor, V.; Mauch-Mani, B. Plant responses to simultaneous biotic and abiotic stress: Molecular mechanisms. *Plants* 2014, *3*, 458–475.

22. Meyers, B.C.; Kozik, A.; Griego, A.; Kuang, H.; Michelmore, R.W. Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. *Plant Cell* 2003, *15*, 809–834.

23. Shao, Z.-Q.; Xue, J.-Y.; Wu, P.; Zhang, Y.-M.; Wu, Y.; Hang, Y.-Y.; Wang, B.; Chen, J.-Q. Large-scale analyses of angiosperm nucleotide-binding site-leucine-rich repeat (NBS-LRR) genes reveal three anciently diverged classes with distinct evolutionary patterns. *Plant Physiol.* 2016, doi:10.1104/pp.15.01487.

24. Ramirez-Prado, J.S.; Abulfaraj, A.A.; Rayapuram, N.; Benhamed, M.; Hirt, H. Plant immunity: From signaling to epigenetic control of defense. *Trends Plant Sci.* 2018, 9, 833–844.

25. Asai, T.; Tena, G.; Plotnikova, J.; Willmann, M.R.; Chiu, W.-L.; Gomez-Gomez, L.; Boller, T.; Ausubel, F.M.; Sheen, J. MAP kinase signalling cascade in *Arabidopsis* innate immunity. *Nature* 2002, *415*, 977–983.

26. Neupane, A.; Nepal, M.P.; Piya, S.; Subramanian, S.; Rohila, J.S.; Reese, R.N.; Benson, B.V. Identification, nomenclature, and evolutionary relationships of mitogen-activated protein kinase (MAPK) genes in soybean. *Evol. Bioinform.* 2013, *9*, 363.

27. Neupane, A.; Nepal, M.P.; Benson, B.V.; MacArthur, K.J.; Piya, S. Evolutionary history of mitogen-activated protein kinase (MAPK) genes in Lotus, Medicago, and Phaseolus. *Plant Signal. Behav.* 2013, *8*, e27189.

28. Wu, P.; Wang, W.; Li, Y.; Hou, X. Divergent evolutionary patterns of the MAPK cascade genes in Brassica rapa and plant phylogenetics. *Hort. Res.* 2017, *4*, 17079.

29. Wu, J.; Wang, J.; Pan, C.; Guan, X.; Wang, Y.; Liu, S.; He, Y.; Chen, J.; Chen, L.; Lu, G. Genome-wide identification of MAPKK and MAPKKK gene families in tomato and transcriptional profiling analysis during development and stress response. *PLoS ONE* 2014, *9*, e103032.

30. Liu, Z.; Zhang, L.; Xue, C.; Fang, H.; Zhao, J.; Liu, M. Genome-wide identification and analysis of MAPK and MAPKK gene family in Chinese jujube (Ziziphus jujuba Mill.). *BMC Genom.* 2017, *18*, 855.

31. Kong, X.; Pan, J.; Zhang, D.; Jiang, S.; Cai, G.; Wang, L.; Li, D. Identification of mitogen-activated protein kinase kinase gene family and MKK–MAPK interaction network in maize. *Biochem. Biophys. Res. Commun.* 2013, *441*, 964–969.

32. Zhang, S.; Xu, R.; Luo, X.; Jiang, Z.; Shu, H. Genome-wide identification and expression analysis of MAPK and MAPKK gene family in *Malus domestica*. *Gene* 2013, *531*, 377–387.

33. Chen, L.; Hu, W.; Tan, S.; Wang, M.; Ma, Z.; Zhou, S.; Deng, X.; Zhang, Y.; Huang, C.; Yang, G. Genome-wide identification and analysis of MAPK and MAPKK gene families in *Brachypodium distachyon*. *PLoS ONE* 2012, *7*, e46744.

34. Piao, Y.; Jin, K.; He, Y.; Liu, J.; Liu, S.; Li, X.; Piao, Z. Genome-wide identification and role of MKK and MPK gene families in clubroot resistance of *Brassica rapa*. *PLoS ONE* 2018, *13*, e0191015.

35. Goyal, R.K.; Tulpan, D.; Chomistek, N.; Fundora, D.G.-P.; West, C.; Ellis, B.E.; Frick, M.; Laroche, A.; Foroud, N.A. Analysis of MAPK and MAPKK gene families in wheat and related Triticeae species. *BMC Genom.* 2018, *19*, 178.

36. Wang, L.; Hu, W.; Tie, W.; Ding, Z.; Ding, X.; Liu, Y.; Yan, Y.; Wu, C.; Peng, M.; Xu, B. The MAPKKK and MAPKK gene families in banana: Identification, phylogeny and expression during development, ripening and abiotic stress. *Sci. Rep.* 2017, *7*, 1159.

37. Cakir, B.; Kılıçkaya, O. Mitogen-activated protein kinase cascades in *Vitis vinifera*. *Front. Plant Sci.* 2015, *6*, 556.

38. Zhang, X.; Mi, X.; Chen, C.; Wang, H.; Guo, W. Identification on mitogen-activated protein kinase signaling cascades by integrating protein interaction with transcriptional profiling analysis in cotton. *Sci. Rep.* 2018, *8*, 8178.

39. Dóczi, R.; Ökrész, L.; Romero, A.E.; Paccanaro, A.; Bögre, L. Exploring the evolutionary path of plant MAPK networks. *Trends Plant Sci.* 2012, *17*, 518–525.

40. Widmann, C.; Gibson, S.; Jarpe, M.B.; Johnson, G.L. Mitogen-activated protein kinase: Conservation of a three-kinase module from yeast to human. *Physiol. Rev.* 1999, *79*, 143–180.

41. Iorizzo, M.; Ellison, S.; Senalik, D.; Zeng, P.; Satapoomin, P.; Huang, J.; Bowman, M.; Iovene, M.; Sanseverino, W.; Cavagnaro, P. A high-quality carrot genome

assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nat. Genet.* 2016, *48*, 657.

42. Badouin, H.; Gouzy, J.; Grassa, C.J.; Murat, F.; Staton, S.E.; Cottret, L.; Lelandais-Brière, C.; Owens, G.L.; Carrère, S.; Mayjonade, B. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* 2017, *546*, 148.

43. Consortium, T.G. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 2012, *485*, 635.

44. Albert, V.A.; Barbazuk, W.B.; Der, J.P.; Leebens-Mack, J.; Ma, H.; Palmer, J.D.; Rounsley, S.; Sankoff, D.; Schuster, S.C.; Soltis, D.E. The Amborella genome and the evolution of flowering plants. *Science* 2013, *342*, 1241089.

45. Goodstein, D.M.; Shu, S.; Howson, R.; Neupane, R.; Hayes, R.D.; Fazo, J. Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res.* 2012, *40*, D1178–D1186.

46. Shaw, A.; Schmutz, J.; Devos, N.; Shu, S.; Carrell, A.; Weston, D. The Sphagnum Genome Project: A new model for ecological and evolutionary genomics. In *Advances in Botanical Research*; Elsevier: Amsterdam, The Netherlands, 2016; Volume 78, pp. 167-187.

47. Merchant, S.S.; Prochnik, S.E.; Vallon, O.; Harris, E.H.; Karpowicz, S.J.; Witman, G.B.; Terry, A.; Salamov, A.; Fritz-Laylin, L.K.; Maréchal-Drouard, L. The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science* 2007, *318*, 245–250.

48. Kane, N.C.; Rieseberg, L.H.J.G. Selective sweeps reveal candidate genes for adaptation to drought and salt tolerance in common sunflower, *Helianthus annuus*. *Genetics* 2007, *175*, 1823–1834.

49. Finn, R.D.; Clements, J.; Arndt, W.; Miller, B.L.; Wheeler, T.J.; Schreiber, F.; Bateman, A.; Eddy, S.R. HMMER web server: 2015 update. *Nucleic Acids Res.* 2015, *43*, W30–W38.

50. Jones, P.; Binns, D.; Chang, H.-Y.; Fraser, M.; Li, W.; McAnulla, C.; McWilliam, H.; Maslen, J.; Mitchell, A.; Nuka, G. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* 2014, *30*, 1236–1240.

51. Finn, R.D.; Bateman, A.; Clements, J.; Coggill, P.; Eberhardt, R.Y.; Eddy, S.R.; Heger, A.; Hetherington, K.; Holm, L.; Mistry, J. Pfam: The protein families database. *Nucleic Acids Res.* 2013, *42*, W30–W38.

52. Bailey, T.L.; Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 1994, *2*, 28–36.

53. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* 1990, *215*, 403–410.

54. Emanuelsson, O.; Brunak, S.; von Heijne, G.; Nielsen, H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Prot.* 2007, *2*, 953–971.

55. Thompson, J.D.; Higgins, D.G.; Gibson, T.J. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 1994, *22*, 4673–4680.

56. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004, *32*, 1792–1797.

57. Kearse, M.; Moir, R.; Wilson, A.; Stones-Havas, S.; Cheung, M.; Sturrock, S.; Buxton, S.; Cooper, A.; Markowitz, S.; Duran, C. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 2012, *28*, 1647–1649.

58. Kumar, S.; Stecher, G.; Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 2016, 33, 1870–1874.

59. Tamura, K.; Battistuzzi, F.U.; Billing-Ross, P.; Murillo, O.; Filipski, A.; Kumar, S. Estimating divergence times in large molecular phylogenies. *Proc. Natl. Acad. Sci. USA* 2012, *109*, 19333–19338.

60. Kumar, S.; Stecher, G.; Suleski, M.; Hedges, S.B. Evolution TimeTree: A resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* 2017, *34*, 1812–1819.

61. Hu, B.; Jin, J.; Guo, A.-Y.; Zhang, H.; Luo, J.; Gao, G. GSDS 2.0: An upgraded gene feature visualization server. *Bioinformatics* 2015, *31*, 1296–1297.

62. He, S. Genome-wide identification and transcriptional expression analysis of mitogen-activated protein kinase and mitogen-activated protein kinase kinase genes in Capsicum annuum. *Front. Plant Sci.* 2015, *6*, 780.

63. Liang, W.; Yang, B.; Yu, B.-J.; Zhou, Z.; Li, C.; Jia, M.; Sun, Y.; Zhang, Y.; Wu, F.; Zhang, H. Identification and analysis of MKK and MPK gene families in canola (*Brassica napus* L.). *BMC Genom.* 2013, *14*, 392.

64. Wang, J.; Pan, C.; Wang, Y.; Ye, L.; Wu, J.; Chen, L.; Zou, T.; Lu, G. Genome-wide identification of MAPK, MAPKK, and MAPKKK gene families and transcriptional profiling analysis during development and stress response in cucumber. *BMC Genom.* 2015, *16*, 386.

65. Liu, Y.; Zhang, D.; Wang, L.; Li, D. Genome-Wide Analysis of Mitogen-Activated Protein Kinase Gene Family in Maize. *Plant Mol. Biol. Rep.* 2013, *31*, 1446–1460.

66. Andrews, S. *FastQC: A Quality Control Tool for High Throughput Sequence Data*; Babraham Institute: Cambridge, UK, 2010.

67. Kong, Y.J.G. Btrim: A fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. *Genomics* 2011, *98*, 152–153.

68. Patro, R.; Duggal, G.; Love, M.I.; Irizarry, R.A.; Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 2017, *14*, 417–419.

69. Grüning, B.; Dale, R.; Sjödin, A.; Chapman, B.A.; Rowe, J.; Tomkins-Tinch, C.H.; Valieris, R.; Köster, J.; Bioconda, T. Bioconda: Sustainable and comprehensive software distribution for the life sciences. *Nat. Methods* 2018, *15*, 475–476.

70. Ge, S.X. iDEP: An integrated web application for differential expression and pathway analysis. *bioRxiv* 2017, doi:10.1101/148411.

71. Dai, X.; Zhao, P.X. psRNATarget: A plant small RNA target analysis server. *Nucleic Acids Res.* 2011, *39*, W155–W159.

72. Tajima, F. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* 1993, *135*, 599–607.

73. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 1989, *123*, 585–595.

74. Lamesch, P.; Berardini, T.Z.; Li, D.; Swarbreck, D.; Wilks, C.; Sasidharan, R.; Muller, R.; Dreher, K.; Alexander, D.L.; Garcia-Hernandez, M. The *Arabidopsis* Information

Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Res.* 2011, *40*, D1202–D1210.

75. Schmutz, J.; Cannon, S.B.; Schlueter, J.; Ma, J.; Mitros, T.; Nelson, W.; Hyten, D.L.; Song, Q.; Thelen, J.J.; Cheng, J. Genome sequence of the palaeopolyploid soybean. *Nature* 2010, *463*, 178–183.

76. Ouyang, S.; Zhu, W.; Hamilton, J.; Lin, H.; Campbell, M.; Childs, K.; Thibaud-Nissen, F.; Malek, R.L.; Lee, Y.; Zheng, L. The TIGR rice genome annotation resource: Improvements and new features. *Nucleic Acids Res.* 2006, *35*, D883–D887.

77. Jaillon, O.; Aury, J.-M.; Noel, B.; Policriti, A.; Clepet, C.; Casagrande, A.; Choisne, N.; Aubourg, S.; Vitulo, N.; Jubin, C. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 2007, *449*, 463.

78. Mohanta, T.K.; Park, Y.-H.; Bae, H. Novel genomic and evolutionary insight of WRKY transcription factors in plant lineage. *Sci. Rep.* 2016, *6*, 37309.

79. Xu, J.; Zhang, S. Mitogen-activated protein kinase cascades in signaling plant growth and development. *Trends Plant Sci.* 2015, *20*, 56–64.

80. De Smet, I.; Voß, U.; Jürgens, G.; Beeckman, T. Receptor-like kinases shape the plant. *Nat. Cell Biol.* 2009, *11*, 1166.

81. Li, J.; Tax, F.E. Receptor-like kinases: Key regulators of plant development and defense. *J. Integr. Plant Biol.* 2013, *55*, 1184–1187.

82. Lee, J.S.; Wang, S.; Sritubtim, S.; Chen, J.G.; Ellis, B.E. *Arabidopsis* mitogen-activated protein kinase MPK12 interacts with the MAPK phosphatase IBR5 and regulates auxin signaling. *Plant J.* 2009, *57*, 975–985.

83. Dai, Y.; Wang, H.; Li, B.; Huang, J.; Liu, X.; Zhou, Y.; Mou, Z.; Li, J. Increased expression of MAP KINASE KINASE7 causes deficiency in polar auxin transport and leads to plant architectural abnormality in *Arabidopsis*. *Plant Cell* 2006, *18*, 308–320.

84. Kim, H.S.; Park, S.-C.; Ji, C.Y.; Park, S.; Jeong, J.C.; Lee, H.-S.; Kwak, S.-S. Molecular characterization of biotic and abiotic stress-responsive MAP kinase genes, IbMPK3 and IbMPK6, in sweetpotato. *Plant Physiol. Biochem.* 2016, *108*, 37–48.

85. Abass, M.; Morris, P.C. The Hordeum vulgare signalling protein MAP kinase 4 is a regulator of biotic and abiotic stress responses. *J. Plant Physiol.* 2013, *170*, 1353–1359.

86. Jalmi, S.K.; Sinha, A.K. ROS mediated MAPK signaling in abiotic and biotic stress-striking similarities and differences. *Front. Plant Sci.* 2015, *6*, 769.

87. Jaggi, M.; Gupta, M.; Tuteja, N.; Sinha, A.K. Mitogen-Activated Protein Kinases in Abiotic Stress Tolerance in Crop Plants:"-Omics"Approaches. *Improv. Crop Prod. Sustain. Agric.* 2013, 107-132.

88. Taj, G.; Agarwal, P.; Grant, M.; Kumar, A. MAPK machinery in plants: Recognition and response to different stresses through multiple signal transduction pathways. *Plant Signal. Behav.* 2010, *5*, 1370–1378.

89. Mishra, N.S.; Tuteja, R.; Tuteja, N. Signaling through MAP kinase networks in plants. *Arch. Biochem. Biophys.* 2006, *452*, 55–68.

90. Rasmussen, M.W.; Roux, M.; Petersen, M.; Mundy, J. MAP kinase cascades in *Arabidopsis* innate immunity. *Front. Plant Sci.* 2012, *3*, 169.

91. Kültz, D. Evolution of osmosensory MAP kinase signaling pathways. *Am. Zool.* 2001, *41*, 743–757.

92. Lee, D.; Redfern, O.; Orengo, C. Predicting protein function from sequence and structure. *Nat. Rev. Mol. Cell Biol.* 2007, *8*, 995.

93. Kong, F.; Wang, J.; Cheng, L.; Liu, S.; Wu, J.; Peng, Z.; Lu, G. Genome-wide analysis of the mitogen-activated protein kinase gene family in *Solanum lycopersicum*. *Gene* 2012, *499*, 108–120.

94. Liu, Q.; Xue, Q. Computational identification and phylogenetic analysis of the MAPK gene family in *Oryza sativa*. *Plant Physiol. Biochem.* 2007, *45*, 6–14.

95. Barker, M.S.; Li, Z.; Kidder, T.I.; Reardon, C.R.; Lai, Z.; Oliveira, L.O.; Scascitelli, M.; Rieseberg, L.H. Most Compositae (Asteraceae) are descendants of a paleohexaploid and all share a paleotetraploid ancestor with the Calyceraceae. *Am. J. Bot.* 2016, *103*, 1203–1211.

96. Gill, N.; Findley, S.; Walling, J.G.; Hans, C.; Ma, J.; Doyle, J.; Stacey, G.; Jackson, S.A. Molecular and chromosomal evidence for allopolyploidy in soybean. *Plant Physiol.* 2009, *151*, 1167–1174.

97. Soltis, P.S.; Soltis, D.E. The role of genetic and genomic attributes in the success of polyploids. *Proc. Natl. Acad. Sci. USA* 2000, *97*, 7051–7057.

98. Gottlieb, L. Conservation and duplication of isozymes in plants. *Science* 1982, *216*, 373–380.

99. Seiler, G. Wild annual Helianthus anomalus and H. deserticola for improving oil content and quality in sunflower. *Ind. Crops Prod.* 2007, *25*, 95–100.

100. Colcombet, J.; Hirt, H. *Arabidopsis* MAPKs: A complex signalling network involved in multiple biological processes. *Biochem. J.* 2008, *413*, 217–226.

101. Nicole, M.-C.; Hamel, L.-P.; Morency, M.-J.; Beaudoin, N.; Ellis, B.E.; Séguin, A. MAP-ping genomic organization and organ-specific expression profiles of poplar MAP kinases and MAP kinase kinases. *BMC Genom.* 2006, *7*, 223.

102. Rao, K.P.; Richa, T.; Kumar, K.; Raghuram, B.; Sinha, A.K. In silico analysis reveals 75 members of mitogen-activated protein kinase kinase kinase gene family in rice. *DNA Res.* 2010, *17*, 139–153.

103. Wei, W.; Chai, Z.; Xie, Y.; Gao, K.; Cui, M.; Jiang, Y.; Feng, J.J.P.o. Bioinformatics identification and transcript profile analysis of the mitogen-activated protein kinase gene family in the diploid woodland strawberry *Fragaria vesca*. *PLoS ONE* 2017, *12*, e0178596.

104. Zhang, S.; Liu, Y.; Klessig, D.F. Multiple levels of tobacco WIPK activation during the induction of cell death by fungal elicitins. *Plant J.* 2000, *23*, 339–347.

105. Wang, J.; Ding, H.; Zhang, A.; Ma, F.; Cao, J.; Jiang, M. A novel mitogen-activated protein kinase gene in maize (*Zea mays*), ZmMPK3, is involved in response to diverse environmental cues. *J. Integr. Plant Biol.* 2010, *52*, 442–452.

106. Yuasa, T.; Ichimura, K.; Mizoguchi, T.; Shinozaki, K. Oxidative stress activates ATMPK6, an *Arabidopsis* homologue of MAP kinase. *Plant Cell Physiol.* 2001, *42*, 1012–1016.

107. Ichimura, K.; Mizoguchi, T.; Yoshida, R.; Yuasa, T.; Shinozaki, K. Various abiotic stresses rapidly activate *Arabidopsis* MAP kinases ATMPK4 and ATMPK6. *Plant J.* 2000, *24*, 655–665.

108. Teige, M.; Scheikl, E.; Eulgem, T.; Dóczi, R.; Ichimura, K.; Shinozaki, K.; Dangl, J.L.; Hirt, H. The MKK2 pathway mediates cold and salt stress signaling in *Arabidopsis*. *Mol. Cell* 2004, *15*, 141–152.

109. Rohila, J.S.; Yang, Y. Rice mitogen-activated protein kinase gene family and its role in biotic and abiotic stress response. *J. Integr. Plant Biol.* 2007, *49*, 751–759.

110. Danquah, A.; Zélicourt, A.; Boudsocq, M.; Neubauer, J.; Frei dit Frey, N.; Leonhardt, N.; Pateyron, S.; Gwinner, F.; Tamby, J.P.; Ortiz-Masia, D. Identification and characterization of an ABA-activated MAP kinase cascade in *Arabidopsis thaliana*. *Plant J.* 2015, *82*, 232–244.

111. Cheong, Y.H.; Moon, B.C.; Kim, J.K.; Kim, C.Y.; Kim, M.C.; Kim, I.H.; Park, C.Y.; Kim, J.C.; Park, B.O.; Koo, S.C. BWMK1, a rice mitogen-activated protein kinase, locates in the nucleus and mediates pathogenesis-related gene expression by activation of a transcription factor. *Plant Physiol.* 2003, *132*, 1961–1972.

112. Schoenbeck, M.A.; Samac, D.A.; Fedorova, M.; Gregerson, R.G.; Gantt, J.S.; Vance, C.P. The alfalfa (Medicago sativa) TDY1 gene encodes a mitogen-activated protein kinase homolog. *Mol. Plant-Microbe Interact.* 1999, *12*, 882–893.

113. Mizoguchi, T.; Irie, K.; Hirayama, T.; Hayashida, N.; Yamaguchi-Shinozaki, K.; Matsumoto, K.; Shinozaki, K. A gene encoding a mitogen-activated protein kinase kinase kinase is induced simultaneously with genes for a mitogen-activated protein kinase and an S6 ribosomal protein kinase by touch, cold, and water stress in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* 1996, *93*, 765–769.

114. Im, J.H.; Lee, H.; Kim, J.; Kim, H.B.; Seyoung, K.; Kim, B.M.; An, C.S. A salt stress-activated mitogen-activated protein kinase in soybean is regulated by phosphatidic acid in early stages of the stress response. *J. Plant Biol.* 2012, *55*, 303–309.

115. Zong, X.-j.; Li, D.-p.; Gu, L.-k.; Li, D.-q.; Liu, L.-x.; Hu, X.-l. Abscisic acid and hydrogen peroxide induce a novel maize group C MAP kinase gene, ZmMPK7, which is responsible for the removal of reactive oxygen species. *Planta* 2009, *229*, 485.

116. Zhang, X.; Xu, X.; Yu, Y.; Chen, C.; Wang, J.; Cai, C.; Guo, W.J.S.r. Integration analysis of MKK and MAPK family members highlights potential MAPK signaling modules in cotton. *Sci. Rep.* 2016, *6*, 29781.

117. Lespinet, O.; Wolf, Y.I.; Koonin, E.V.; Aravind, L. The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.* 2002, *12*, 1048–1059.

118. Matsuoka, D.; Nanmori, T.; Sato, K.i.; Fukami, Y.; Kikkawa, U.; Yasuda, T. Activation of AtMEK1, an *Arabidopsis* mitogen-activated protein kinase kinase, in vitro and in vivo: Analysis of active mutants expressed in E. coli and generation of the active form in stress response in seedlings. *Plant J.* 2002, *29*, 637–647.

119. Ren, D.; Yang, H.; Zhang, S. Cell death mediated by MAPK is associated with hydrogen peroxide production in *Arabidopsis*. *J. Biol. Chem.* 2002, *277*, 559–565.

120. Hwa, C.-M.; Yang, X.-C. The AtMKK3 pathway mediates ABA and salt signaling in *Arabidopsis*. *Acta Physiol. Planta* 2008, *30*, 277–286.

121. Liu, Y. Roles of mitogen-activated protein kinase cascades in ABA signaling. *Plant Cell Rep.* 2012, *31*, 1–12.

122. Zhong, B.; Xi, Z.; Goremykin, V.V.; Fong, R.; Mclenachan, P.A.; Novis, P.M.; Davis, C.C.; Penny, D. Streptophyte algae and the origin of land plants revisited using heterogeneous models with three new algal chloroplast genomes. *Mol. Biol. Evol.* 2013, *31*, 177–183.

123. Jansen, R.K.; Cai, Z.; Raubeson, L.A.; Daniell, H.; Leebens-Mack, J.; Müller, K.F.; Guisinger-Bellian, M.; Haberle, R.C.; Hansen, A.K.; Chumley, T.W.; et al. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and

identifies genome-scale evolutionary patterns. *Proc. Natl. Acad. Sci. USA* 2007, *104*, 19369–19374.

124. Stevens, P.F.; Davis, H. *Angiosperm Phylogeny Website*; Missouri Botanical Garden: St. Louis, MO, USA, 2001.

125. Mockaitis, K.; Howell, S.H.J.T.P.J. Auxin induces mitogenic activated protein kinase (MAPK) activation in roots of *Arabidopsis* seedlings. *Plant J.* 2000, *24*, 785–796.

126. Seo, S.; Katou, S.; Seto, H.; Gomi, K.; Ohashi, Y. The mitogen-activated protein kinases WIPK and SIPK regulate the levels of jasmonic and salicylic acids in wounded tobacco plants. *Plant J.* 2007, *49*, 899–909.

127. Zhou, C.; Cai, Z.; Guo, Y.; Gan, S. An *Arabidopsis* mitogen-activated protein kinase cascade, MKK9-MPK6, plays a role in leaf senescence. *Plant Physiol.* 2009, *150*, 167–177.

128. Wu, T.; Kong, X.P.; Zong, X.J.; Li, D.P.; Li, D.Q. Expression analysis of five maize MAP kinase genes in response to various abiotic stresses and signal molecules. *Mol. Biol. Rep.* 2011, *38*, 3967–3975.

129. Zhang, J.; Zou, D.; Li, Y.; Sun, X.; Wang, N.-N.; Gong, S.-Y.; Zheng, Y.; Li, Xue B.L. GhMPK17, a cotton mitogen-activated protein kinase, is involved in plant response to high salinity and osmotic stresses and ABA signaling. *PLoS ONE* 2014, *9*, e95642.

130. Chakraborty, C.; Sharma, A.R.; Patra, B.C.; Bhattacharya, M.; Sharma, G.; Lee, S.-S. MicroRNAs mediated regulation of MAPK signaling pathways in chronic myeloid leukemia. *Oncotarget* 2016, *7*, 42683.

131. Liu, Z.; Zhang, J.; Gao, J.; Li, Y. MicroRNA-4728 mediated regulation of MAPK oncogenic signaling in papillary thyroid carcinoma. *Saudi J. Biol. Sci.* 2018, *25*, 986–990.

132. Johnson, S.M.; Grosshans, H.; Shingara, J.; Byrom, M.; Jarvis, R.; Cheng, A.; Labourier, E.; Reinert, K.L.; Brown, D.; Slack, F.J. RAS is regulated by the let-7 microRNA family. *Cell* 2005, *120*, 635–647.

133. Li, S.; Yu, B. miRNA limits MAP kinase-mediated immunity: Optimization of plant fitness. *Journal of Experimental Botany* 2017, *68*, 5685.

134. Raghuram, B.; Sheikh, A.H.; Sinha, A.K. Regulation of MAP kinase signaling cascade by microRNAs in *Oryza sativa*. *Plant Signal. Behav.* 2014, *9*, e97213

CHAPTER 5: CHARACTERIZATION OF INDUCED SUSCEPTIBILITY EFFECTS

ON SOYBEAN-SOYBEAN APHID INTERACTION

The data portion of this chapter is in review for publication in *BMC Data Note.*

Abstract

Soybean aphid (*Aphis glycines* Matsumura; SBA) is one of the major pests of

soybean (*Glycine max*) in the United States of America. Four biotypes of soybean aphids

have been confirmed in the United States suggesting the insect's ability to adapt to the

host resistance. One previous study on soybean, soybean-aphid interactions showed that

avirulent (biotype 1) and virulent (biotype 2) biotypes could co-occur, and potentially

interact on resistant and susceptible soybean resulting in an induced susceptibility. The

main objective of this research was to employ RNA sequencing approach to characterize

the induced susceptibility effect in which initial feeding by virulent aphids can increase

the suitability of avirulent aphids in resistant soybean cultivars. The interactions were

evaluated using SBA-resistant (*Rag1*) and SBA-susceptible soybean cultivars with

biotype 1 and biotype 2 soybean aphids. Demographic and transcriptomic responses of

susceptible and resistant (*Rag1*) soybean cultivars to aphid feeding were investigated in

soybean plants colonized by aphids (biotype 1) in presence or absence of inducer

population (biotype 2) at day 1 and day 11. WGCNA analysis revealed that 11 and 15

KEGG pathway modules were enriched for day 1 and day 11 samples, respectively. In

addition, enriched transcription factor (TF) binding motifs were identified in time course

and resistant and susceptible reactions. In the presence of inducer population, we found

746 and 243 DEGs in susceptible and resistant cultivars, respectively at day 1, whereas,

981 and 377 DEGs were found in susceptible and resistant cultivars, respectively at day 11. Enrichment analysis showed a response to chitin, lignin catabolic and metabolic process, asparagine metabolic process, response to chemical unique to treatment with no inducer population, whereas, response to reactive oxygen species, photosynthesis, regulation of endopeptidase activity unique to treatment with inducer population. Furthermore, 14 DEGs were observed in *Rag* QTLs regions, particularly six DEGs in *Rag1* containing QTL. The identified DEGs in the experiment in both resistant and susceptible cultivars during the interaction of soybean and soybean-aphids are potential candidates for furthering investigation into induced susceptibility.

5.1. Introduction

The invasive species have severely affected the agriculture system in numerous ways such as reducing yields and increasing costs of managing them affecting integrated pest management (IPM) [1, 2]. *Aphis glycines* Matsumura (Hemiptera: Aphididae), the soybean aphid (SBA), a common invasive pest of soybean [*Glycine max* (L.) Merr] was first reported in North America in 2000 [3]. It is regarded as a common insect pest in China and many Asian countries [4]. In 2003, soybean aphid spread over 21 states of the U. S. and three Canadian provinces [5]. By the season of 2009, soybean aphid developed in the eastern region (New York and Ontario, Canada) beginning from July through August, as well as in Midwestern region (Minnesota, Wisconsin, and Iowa) spreading to 30 different states of the U.S. [6, 7]. The main reasons for over spreading are its alate (winged) form and occurrence of its overwintering host, *Rhamnus cathartica* (common buckthorn) [8]. The eggs laid on the common buckthorn hatch during March–April into

parthenogenically dividing (for 2 to 3 generations) apteran (wingless) aphids. These

wingless aphids turn in to winged form and travel to their secondary host, soybean during

May–June [9]. They prefer to feed on the ventral side of the leaves mainly in young

soybean trifoliate leaves [10] and feed on the phloem sap and draw assimilates from

soybean plants [11, 12]. The chief symptoms of disease caused by aphid infestation are

plant stunting, leaf yellowing and wrinkling with a reduced photosynthesis, poor pod fill,

and reduced yield, seed size, and seed quality than of healthy soybeans [13]. SBA

deposits honeydew on soybean leaves that aids as a vector for various viruses such as

Soybean mosaic virus, Alfalfa mosaic virus, and Bean yellow mosaic virus [9]. The SBA

population can double in 1.5 days under favorable conditions whereas doubling time in

fields is up to 6.8 days. Because of this high reproducing capability, they can undergo

multiplication up to 15 generations in the growing season of soybean [11]. The economic

loss due to SBA was approximately $4 billion annually [9].

For an effective management approach, soybean lines that are naturally resistant

to the aphids can be used to control SBA. Many researchers surveyed soybean germplasm

collection and have identified soybean lines that have shown resistance to *A. glycines*.

The resistance mechanism of the plant can be implemented in controlling pests without

disturbing the environment [14]. Various dominant and recessive resistance to *A. glycines*

(*Rag*) loci have been identified in soybean lines through various genetic analysis. Up to

now, 16 *Rag* QTLs [*Rag1*[15], *Rag1b* [16], *Rag1c* [17], *Rag2* [18], *Rag3* [19, 20] , *Rag4*

[19], *Rag3* [16] , *Rag3b* [21], *Rag3c* [22], *Rag4* [17], *Rag5* [23] , *Rag6* [22];

qChrom.07.1, *qChrom.16.1*, *qChrom.13.1*, *qChrom.17.1* [24] in various soybean plant

introductions (PI).

Despite the identification of many monogenic and oligogenic genes for host plant resistance, the discovery of virulent biotypes of *A. glycines* that can survive on resistant varieties has been a serious threat. It has been estimated that the soybean cultivar with alone *Rag* and combined *Rag1* and *Rag2* can diminish the *A. glycines* growth by 34% and 49% respectively [25]. Up to now, four biotypes of aphid (biotype 1, biotype 2, biotype 3, and biotype 4) have been prevalent in the U.S having capability to reproduce in susceptible as well as resistant cultivars (with single or multiple *Rag* genes) [26, 27, 28]. Hence, the diverse population of both virulent and avirulent that appear phenotypically similar can engender induced susceptibility on the resistant plants [29]. The interaction between insect herbivores with their own host creates the condition called induced susceptibility that assists other consequent herbivores [30]. This type of susceptibility takes place between conspecifics on susceptible as well as resistant plants [30, 31]. The phenotype of conspecific can be both virulent and avirulent biotype. Few studies have been done to understand induced susceptibility in *A. glycines* to answer the reason for a high number of *A. glycines* population in resistant soybean cultivars in North America. Varenhorst et al. 2015, [32] showed that virulent *A. glycines* increase the suitability of resistant soybean for avirulent conspecifics.

RNA sequencing (RNA-Seq) has been a standard tool for studying qualitative and quantitative gene expression assay providing information on transcript abundance with their variation [33, 34]. A comprehensive understanding of the transcriptomes would help in understanding the molecular interactions between soybean and *A. glycines*. A number of studies have been carried out using RNA-seq to unravel the molecular interactions for soybean-*A. glycines* herbivory with different objectives [35, 36, 37]. Brechenmacher et

al., 2015 [37] identified 396 differentially regulated proteins and 2361 significantly

regulated genes in different time response (up to 48 hours) soybean aphid infestation

using two *Rag*2 and/or *Rag*2 near-isogenic lines of soybean. Among them, a gene of

unknown function, a mitochondrial protease, and NBS-LRR gene those map within *Rag*2

locus are significantly upregulated in the presence of aphids. Prochaska et al., 2015 [35]

identified 3 and 36 differentially expressed genes (DEGs) at day 5 and day 15 in resistant

cultivar (KS4202), respectively whereas 0 and 11 DEGs at day 5 and day 15 in

susceptible cultivar (K-03-4686), respectively. Most of the DEGs were related to WRKY

transcription factors, peroxidases, and cytochrome p450s. Previously, Li et al. 2008 [38]

studied soybean defense response to *A. glycines* generating transcript profiles using

cDNA microarrays. In this study, they identified 140 genes related to the cell wall,

transcription factors, signaling and secondary metabolism in response to resistance using

resistant (cv. Dowling) and susceptible (cv. Williams 82) soybean cultivars. Studham and

MacIntosh 2013 [39] used oligonucleotide microarrays to study soybean- *A. glycines*

interaction using aphid-resistant LD16060 with *Rag1* gene and aphid-susceptible SD01-

76R. They identified 49 and 284 DEGs in 1 day of infestation (doi) and 7 doi in

susceptible cultivar, respectively whereas 0 and 1 DEGs in 1 doi and 7 doi in resistant

cultivar respectively studying transcript profiles determined after 1 and 7 days of aphid

infestation. They suggested that the response of defense genes in the resistant plants are

in constitutive in nature whereas, in susceptible plants, the defense genes are elicited only

upon aphid infestation.

This study is aimed to characterize induced susceptibility in soybean through the

analysis of the transcriptional response of soybean in the presence of biotype 1 and

biotype 2 soybean-aphids. Results of the study would have implication in soybean-aphid management and developing soybean cultivar with durable resistance to *A. glycines*.

5.2. Materials and Methods

5.2.1. Plant Material and Aphid Colonies

Two genotypes of soybean were used: susceptible soybean cultivar was LD12-15838R and the resistant cultivar was LD12-15813Ra. The resistant cultivar contains *Rag1* QTL. These genotypes were infested with two aphid populations defined namely biotype 1 (avirulent) and biotype 2 (virulent biotype [26]). The biotypes are defined by the response to *Rag1* genes and were identified in Illinois [26]. The biotype 1 and biotype 2 populations originated from a colony maintained by Iowa State University (Ames, IA). Among them, colonies of biotype 1 originated from field populations in Ohio and were maintained in a colony at the Ohio State University biotype. At South Dakota State University, aphid colonies were maintained using susceptible cultivar SD01-76R for biotype 1 and resistant cultivar LD12-15813Ra for biotype 2. The aphid populations used in this study were randomly selected removing the leaves from the soybean plants used for maintaining the colonies.

5.2.2. Induced Susceptibility Experiment

To characterize induced susceptibility effects, randomized complete block design (RCBD) greenhouse experiment was conducted using twelve treatments, three replications (plants) in three blocks (nine experimental units per treatment). We followed the treatments as explained by the procedure by [32]. The initial feeding population of *A. glycines* was termed as an inducer population and the subsequent feeding population of *A. glycines* was termed as a response population. Three seeds of LD12-15838R and LD12-

15813Ra were planted into damp soil (Professional Growing Mix, Sun Gro Horticulture, MA, USA) in each pot of dimension of 10.1 cm by 8.89 cm (500 ml; Belden Plastics, MN, USA). Pots were placed onto plastic flats ($87 \times 15 \times 5$ cm). The soybean plants were watered filling the flats when top soil began to dry. The plants were thinned down to one plant per pot upon reaching the V1 developmental growth stage. V2 staged soybean plants (Day 0) were infested with avirulent inducer populations using with a combination of zero inducer (none), 50 *A. glycines* (50 avirulent), or 50 *A. glycines* (50 virulent) onto a ventral side of a middle leaf of first trifoliate except the control plants. The infested trifoliate was covered with a small no-see-um mesh net (Quest Outfitters, Sarasota, FL) and secured with the paper clip and tangle trap to confine within the first trifoliate of the plants. After 24 hrs. (Day 1), one-day leaves from second trifoliate were collected from one replication set of each block and snap frozen in the liquid nitrogen. After sample collection from one replication, response population of 15 *A. glycines* (15 avirulent), or 15 *A. glycines* (50 virulent) were added upon the middle leaf of second trifoliate (except on sampled and control plants). The whole plants were covered with the large no-see-um mesh net (Quest Outfitters, Sarasota, FL) to confine movement of aphids between the plants. The response population was allowed to move freely about the plant with the exception of first trifoliate. This ensures the spatial isolation of inducer and response populations. The response populations were counted on each plant to confirm the colonization by the response populations on day 5. On day 11, the response population of aphids was counted and the day 11 leaf samples from the one replication sets of each block were collected and snap frozen in the liquid nitrogen. The samples were kept at -80ºC for further analysis. The greenhouse conditions were maintained

approximately 24-25°C and a 16-hour photo period (16 light: 8 dark). An overview

representing experimental methods used for the experiment is shown in Figure 5.1.

The aphid counts (response population) collected at 11th day after the inducer

infestations were analyzed using R statistical software version 3.2.4 (https://www.r-

project.org/). The main effects of the inducer population, soybean cultivar, and the

interaction of inducer population by soybean cultivar were analyzed using the model

Response Counts ~ Inducer + Inducer: Cultivar. We checked the effect of both treatment

and block for susceptible and resistant cultivars separately. The model Aphid Counts ~

Treatment + Block was applied in the analysis of variance (ANOVA). The treatment

means based on *A. glycines* numbers were separated using Fisher- least significant

difference (LSD) test at $P < 0.05$ using agricolae package [40] in R. The average SBA

counts were plotted (Figure 5.3) using GraphPad Prism 8.0.2 (San Diego, California

USA, www.graphpad.com).

5.2.3. RNA Extraction, Library Construction, and RNA-sequencing

RNA was extracted from the leave samples from resistant and susceptible

cultivars treated with no aphids, biotype 2: biotype1 collected at day 1 and no aphids,

biotype 2: biotype1 and no aphids: biotype1 at day 11. Briefly, leaf samples from each

treatment were grounded in liquid nitrogen with pestle and mortar to a fine powder

followed by their processing for total RNA extraction using PureLink RNA mini kit

(Invitrogen, USA). RNA samples were treated with TURBO[TM] DNase (Invitrogen, USA)

to remove any DNA contamination following the manufacturer's instructions.

Assessment of the isolated RNA integrity was performed by 1% agarose gel

electrophoresis, and RNA concentration was measured by Nanodrop 2000 (Thermo Fisher Scientific, USA). Three replicates from these treatments in resistant and susceptible cultivars were pooled in equimolar concentration. The cDNA libraries were constructed and sequenced at South Dakota State University Sequencing Facilities. RNAseq library construction was prepared using Illumina's TruSeq Stranded mRNA Kit v1 (San Diego, CA). The libraries were quantified by QuBit dsDNA HS Assay (Life Technologies, Carlsbad, CA). The libraries were sequenced on an Illumina NextSeq 500 using a NextSeq 500/550 High Output Reagent Cartridge v2 (San Diego, CA) at 75 cycles. Fastq files were generated and demultiplexed on Illumina's BaseSpace cloud network (San Diego, CA).

5.2.4. RNA-seq Analysis

Quality control of reads was assessed using FastQC program (version 0.11.3) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) [41]. The FastQC results were visualized using MultiQC v1.3 [42]. Low quality bases (QC value < 20) and adapters were removed by trimming using the program Trimmomatic (version 0.36) [43] (options: PE -phred33 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 HEADCROP:8 MINLEN:30). High-quality single-end reads were mapped against the primary coding sequences of *G. max*. The coding sequences (*Gmax:Gmax_275_Wm82.a2.v1.transcript_primaryTranscriptOnly.fa.gz*) were obtained from the Phytozome database and aligned using Salmon ver.0.9.1 [44] accessed from Bioconda [45]. The codes that were used for data processing are available in Appendix II. A flow chart showing the RNA-seq data analysis pipeline is shown in Figure 5.2. The read quants were filtered with 0.5 counts per million (CPM) in at least one sample. The

quantified raw reads were transformed using regularized log (rlog) which is implemented in the DESeq2 package. The transformed data were subjected to exploratory data analysis such as hierarchical clustering, K-means clustering, principal component analysis (PCA), and visualization of clusters using the t-SNE map. Gene co-expression networks were constructed for divided datasets with the weighted gene co-expression network analysis (WGCNA) package [46] using following parameters: most variable genes to include-3000 genes, soft threshold- 4, minimum module size- 20. The quantified transcript reads obtained from Salmon were employed in CLC Genomics Workbench 9.5 (https://www.qiagenbioinformatics.com/) to obtain the differentially expressed genes (DEGs) using Karl's z-test with false discovery rate (FDR) <0.01 and log2fold change more than a 2-fold. The annotations of the DEGs were obtained from Soybase [47] (www.soybase.org). To understand the molecular pathways enriched GO Biological processes, GO Cellular, GO molecular function, and KEGG pathways for DEGs were analyzed using a graphical enrichment tool REVIGO [48], ShinyGO [49] and integrated Differential Expression and Pathway analysis (iDEP 0.81, R/Bioconductor packages) [50]. The enriched transcription factor binding motifs in promoters in different comparisons were identified in 300bp upstream of DEGs using both iDEP and ShinyGO [49]. The biological relevance of DEGs was visualized using MapMan [51]. The total transcripts of soybean were first converted to bins using the Mercator tool [52] and uploaded to MapMan to assign bins to each differentially expressed transcript.
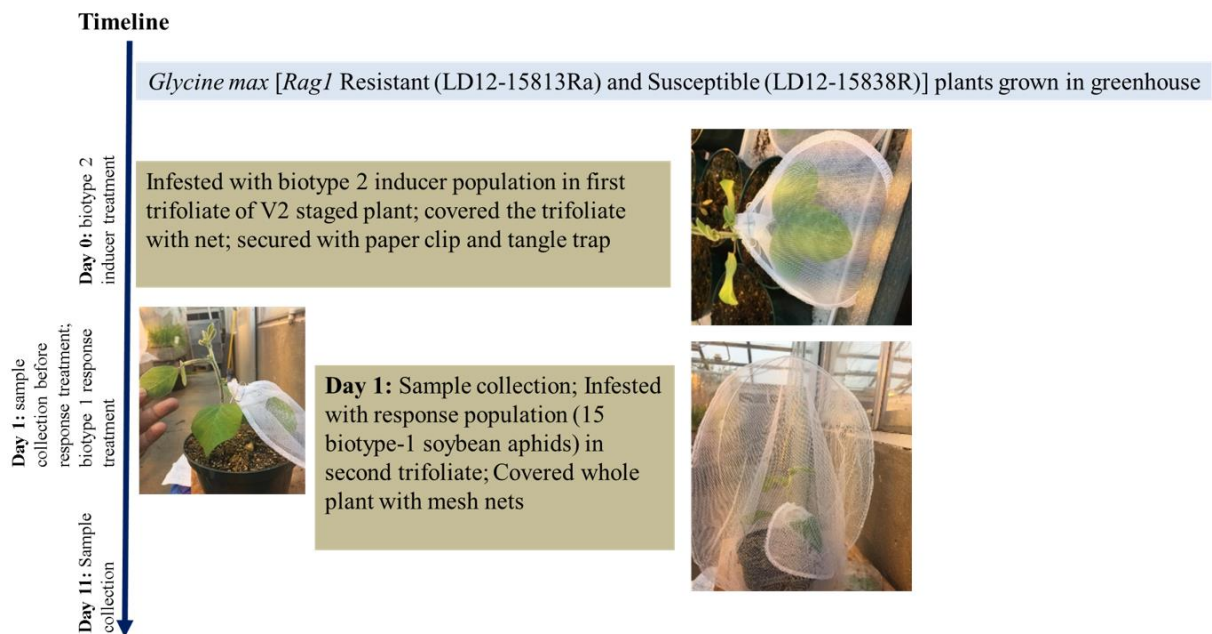
Figure 5.1. An overview of the greenhouse experiment on induced susceptibility effects of soybean-aphids on two cultivars of soybean.

Extraction of RNA

Pool RNA in equimolar concentration and construction of RNA-seq libraries

Sequencing libraries using an Illumina NextSeq 500

Submission to NCBI SRA ← Raw datasets

Assessment of quality of data using FastQC and visualization using multiQC

Trimming using Trimmomatic 0.36

Mapping and quantification using Salmon tool ← Clean reads

Coding sequences of *G. max*

Filtering, assessment, and visualization of transcriptomic data using iDEP 0.81 and CLC Genomics Workbench

Figure 5.2. An overview of RNA-seq data analysis pipeline for the characterization of induced susceptibility effects of soybean-aphids on two cultivars of soybean.

5.3. Results

    5.3.1. Greenhouse Experiment

The hypothesis of response population being positively affected by the presence

of inducer population or conspecifics was tested considering the main effects of the

inducer population, soybean cultivar, and the interaction of inducer population by

soybean cultivar. The response population was significantly affected by the main effects

of inducer population ($F = 15.821$, df = 1, $P = 0.000130$) and cultivar ($F = 11.642$, df =

1, $P = 0.000926$). Induced susceptibility effect on both susceptible and resistant soybean

cultivars, as we observed increased response population densities in the 50 virulent

inducer population treatments compared to the none inducer population treatment (Figure

5.3a and b). Also, the interaction of inducer population on soybean cultivar was

significant ($F = 3.956$, df = 1, $P = 0.049386$) as the response population in the resistant

cultivar was lower than that of a susceptible cultivar.

Upon application of model Response counts ~ Treatment + Block was applied in

analysis of variance (ANOVA), both treatment ($F = 10.950$, df = 5, $P = 6.92e\text{-}07$) and

block (F= 4.497, df = 2, $P = 0.0167$) effect were significant in susceptible cultivars.

Whereas, block effect was insignificant ($F = 0.588$, df = 2, $P = 0.56$) in resistant

cultivars. Thus, we applied a reduced model Response counts ~ Treatment in resistant

cultivars. One way ANOVA was applied to observe the significance of treatment ($F =

7.601$, df = 5, $P = 2.52e\text{-}05$) in resistant cultivars. Fisher- least significant difference

(LSD) test at $P < 0.05$ was applied to see the separation of treatment means based on *A.*

glycines numbers. In susceptible cultivars, we observed the separation of means of

avirulent (response) population between the treatments with none, biotype 2 as an inducer

with biotype 1 as an inducer. Response populations for the biotype 2 as inducer population treatments were 84.4% greater than the response population that received the "none inducer" treatment in the susceptible cultivar. In resistant cultivars, we did not observe the separation of means of avirulent (response) population between the treatments with zero, biotype 2 as an inducer with biotype 1 as an inducer. However, response populations for the biotype 2 as inducer population treatments were 228% greater than the response population that received the "none inducer" treatment in the resistant cultivar.
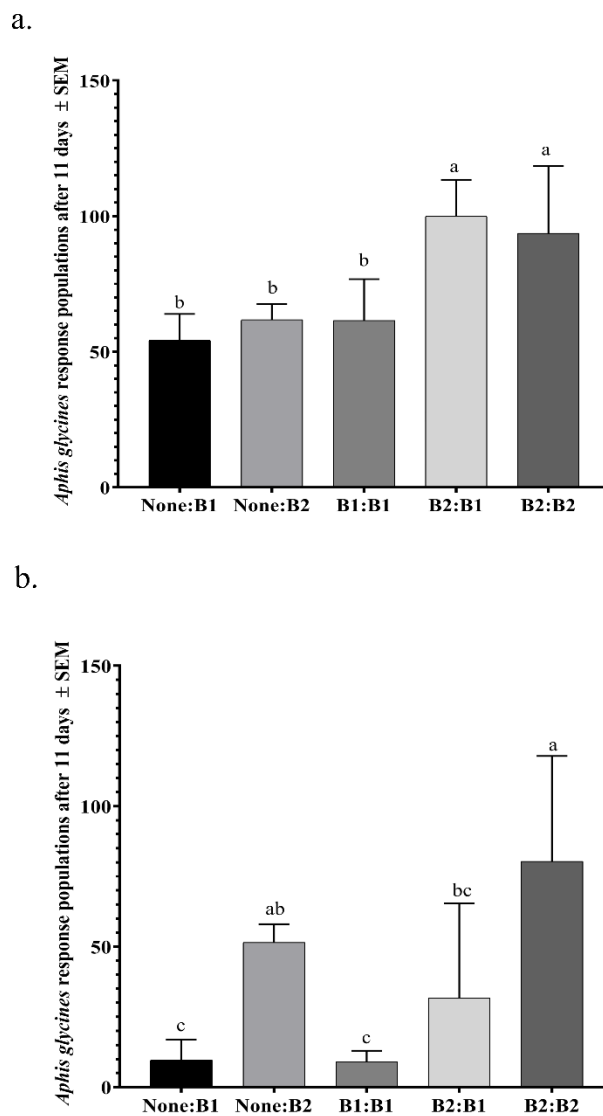
Figure 5.3. Effect of avirulent (B1) and virulent (B2) inducer populations on avirulent (B1) and virulent (B2) response populations on both (a) susceptible and (b) resistant soybean. For this experiment, the susceptible soybean cultivar was LD12-15838R and the resistant cultivar was LD12-15813Ra. Lowercase letters indicate significance among treatments ($P < 0.05$). Plotted values represent the means of the avirulent response population.

5.3.2. RNA-seq Analysis

A total of 10 RNA libraries were prepared and sequenced with the sequencing depth ranging from 24,779,816 to 29,72,4913. Total reads of 266,535,654 were subjected to FastQC analysis to determine the data quality using various quality metrics such as mean quality scores, per sequence quality scores, per sequence GC content, and sequence length distribution (Figure 5.4, Table 5.1). The Phred quality scores per base for all the samples were higher than 30. The GC content ranged from 45 to 46% and followed the normal distribution. After trimming, more than 99% of the reads were retained as the clean and good quality reads. Upon mapping these reads, we obtained high mapping rate ranging from 90.4% to 92.9%. Among the mapped reads, 85.8% to 91.9% reads were uniquely mapped.

Table 5.1. Statistics of the transcriptomic data using RNA-seq pipeline used in this study.

Sample Name	Raw Reads	% GC	Read Length	Trimmed reads	Percentage of clean reads	Mapped Reads	Percentage of mapped reads	Number of Uniquely mapped reads	Percent uniquely mapped	Accession
Control: No aphids; Susceptible soybean; Day 1	25252863	46%	75	25092599	99.37	23394131	93.23119937	20651941	88.27829937	SRR8848027
Control: No aphids; Susceptible soybean; Day 11	27576285	45%	75	27428725	99.46	25212419	91.91976295	22903908	90.84375442	SRR8848028
Control: No aphids; Resistant soybean; Day 1	26009250	45%	75	25842889	99.36	23766133	91.96391704	21001237	88.36623526	SRR8848025
Control: No aphids; Resistant soybean; Day 11	27852647	44%	75	27688740	99.41	25665595	92.69325726	23579647	91.87259052	SRR8848026
Inducer: None; Response: 15 biotype 1; Susceptible soybean; Day 11	26191613	45%	75	26048380	99.45	23554300	90.42520111	21541233	91.4535053	SRR8848031
Inducer: 50 biotype 2; Response:15 biotype 1; Susceptible soybean; Day 1	26008870	46%	75	25862409	99.4	23894333	92.39020619	20704930	86.6520526	SRR8848032
Inducer: 50 biotype 2; Response: 15 biotype 1; Susceptible soybean; Day 11	27213494	46%	75	27046904	99.39	24598524	90.94765153	21099681	85.77620755	SRR8848029
Inducer: None; Response: 15 biotype 1; Resistant soybean; Day 11	26274980	45%	75	26116361	99.4	24249196	92.85059278	22100532	91.13923612	SRR8848030
Inducer: 50 biotype 2; Response: 15 biotype 1; Resistant soybean; Day 1	26424818	45%	75	26275488	99.43	24065562	91.58940074	21158309	87.91944688	SRR8848023
Inducer: 50 biotype 2; Response: 15 biotype 1; Resistant soybean; Day 11	27730834	45%	75	27562105	99.39	25387621	92.11060258	23022198	90.68277016	SRR8848024
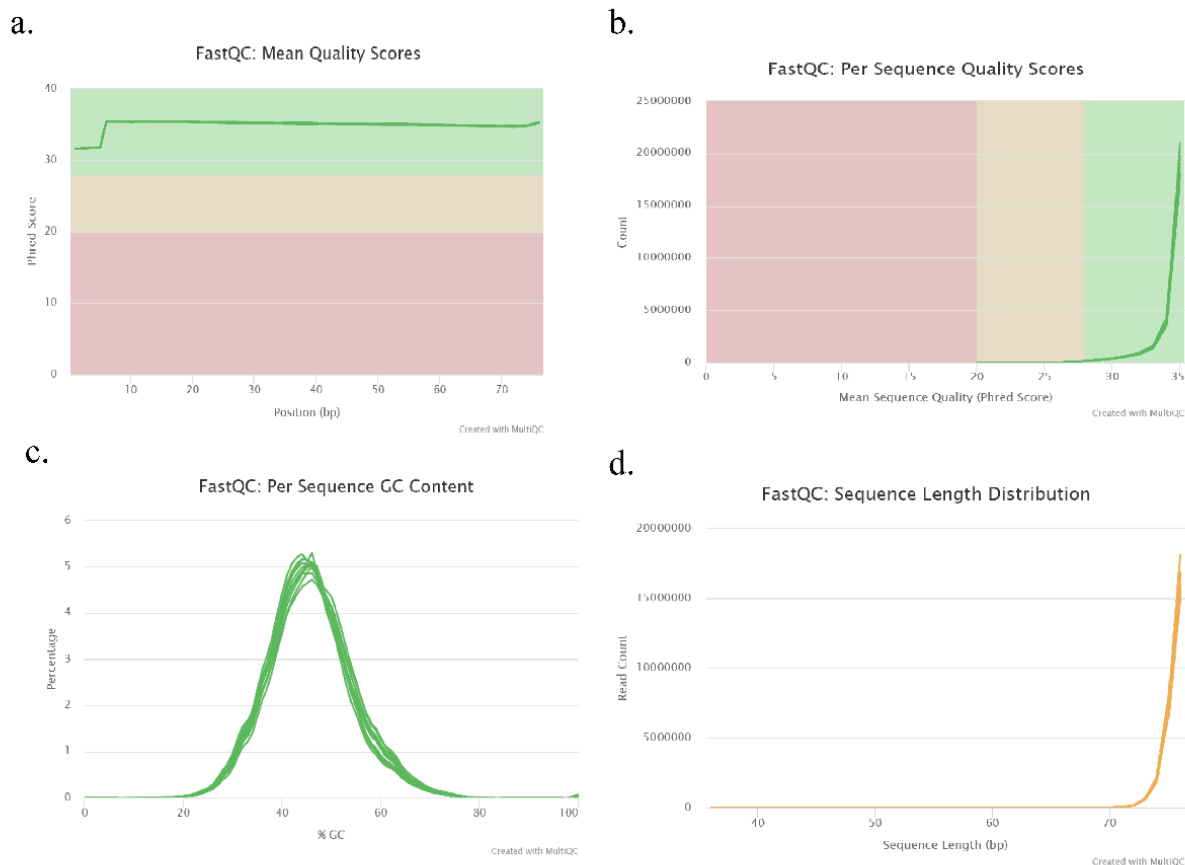
Figure 5.4. Quality metrics of *G. max* sequencing data. (a) Mean quality scores per position, (b) Per sequence quality scores, (c) GC content distribution, and (d) Read length distribution.

5.3.3. WGCNA Analysis

The co-expression networks were used to detect correlated networks of genes and their enrichment in the divided datasets to compare difference on the day 1 and day 11 treatments. Weighted gene co-expression network analysis identified a network of 3,000 genes divided into 11 co-expression modules in four day 1 samples, and a network of 2,999 genes divided into 15 co-expression modules in six day 11 samples (Figure 5.5,

Supplementary File 1). In entire modules, the enrichment analysis found several highly

enriched KEGG pathways for day 1 and day 11 samples. The only KEGG pathways

enriched in day 1 samples, but not in day 11 samples were Phenylpropanoid biosynthesis,

Glycine, serine and threonine metabolism, Alpha-Linolenic acid metabolism, Glycolysis /

Gluconeogenesis, and Cysteine and methionine metabolism. Whereas, the only KEGG

pathways enriched in day 11 samples, but not in day 1 samples were Plant-pathogen

Interaction, Flavonoid biosynthesis, MAPK signaling pathway, Glucosinolate

biosynthesis, and Alanine, aspartate and glutamate metabolism, Cutin, suberine, and wax

biosynthesis. (Table S1). The common pathways for both time points included

Biosynthesis of secondary metabolites, Metabolic pathways, Ribosome, Porphyrin, and

chlorophyll metabolism.

a.

Modules (N=11)

b.

Modules (N=15)
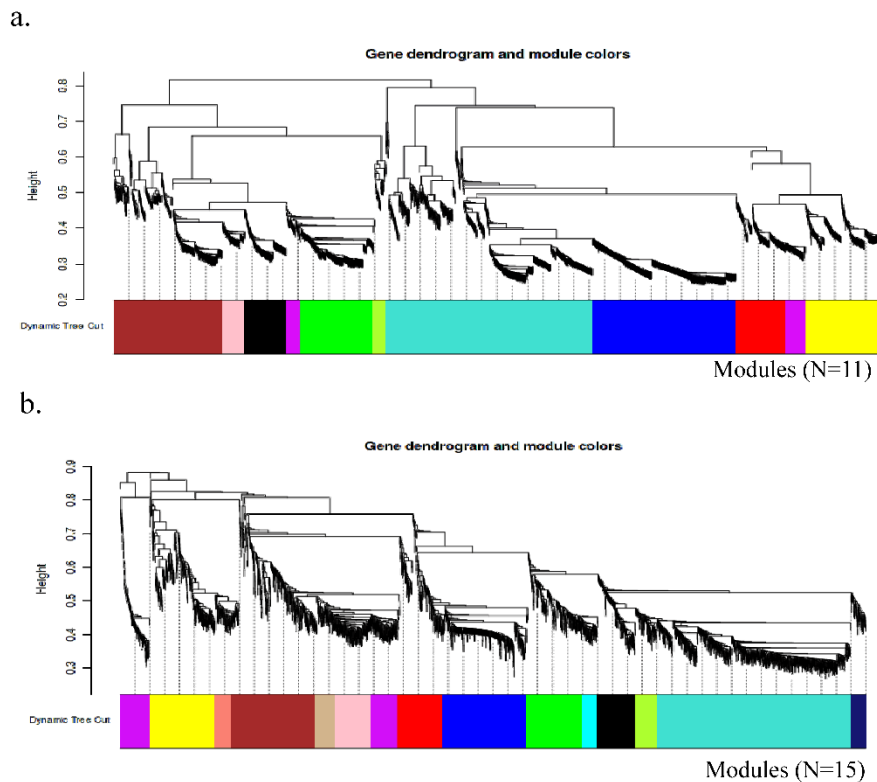
Figure 5.5. Weighted gene co-expression network analysis identified a network of 3,000 genes divided into 10 co-expression modules in day-1 samples (a), and a network of 2,999 genes divided into 15 co-expression modules in day-11 samples (b).

5.3.4. Hierarchical and K-means Clustering

After filtering with 0.5 counts per million (CPM) in at least one sample and rlog transformation, a total of 37,468 genes (66.9% of original 55,983) were retained for clustering and visualization. We used a hierarchical clustering method to determine if individual gene expression patterns clustered according to the time period. The hierarchical clustering based on 3,000 most variable genes, sample distances (Figure 5.6a) indicated that samples clustered on the basis of time points of sample collection (Day 1 and Day 11). Figure 5.6b represents the correlation between the samples using the top 75% genes in a range of 0.96-1. Figure 5.6c represents the standard deviation (SD)

distribution of the top variable 3,000 genes. Regarding the PCA, PC1 is correlated with time with 55% variance, and PC2 is correlated with Treatment with 18% variance (Figure 5.6d).

We used gene clustering to assess if day 11 had more gene clusters enriched for defense-related pathways than day 1 samples. K-means clustering identified five clusters of correlated genes in day 11 samples (Figure 5.7), and Cluster A was enriched primarily with the biosynthesis of secondary metabolites, Cluster B was enriched with various plant defense-related pathways such as MAPK signaling pathway, Plant-pathogen interaction, and plant hormone signal transduction (Supplemental File 2). Four clusters were identified (A-D) in day 1 samples, of which Cluster A was enriched with photosynthesis, carbohydrate metabolism, Cluster B was enriched with fatty acid metabolism, glucosinolate biosynthesis, Cluster C and D were enriched with biosynthesis of secondary metabolites, and plant defense-related pathways such as MAPK signaling pathway (Supplemental File 2). Promoter analysis of clusters in day 1 samples found 80 enriched transcription factor binding motifs in four clusters (A, B, C, and D). Enriched binding motifs consisted of twelve transcription factor families: AP2, AT hook, bHLH, bZIP, CG-1, CxC, Homeodomain, Myb/SANT, NAC/NAM, TBP, TCP, and WRKY. Promoter analysis in day 11 samples found 100 enriched transcription factor binding motifs, consisting of eight transcription factor families: AP2, bHLH, bZIP, CG-1, E2F, LOB, Myb/SANT, and TCP (Supplemental File 2). Six transcription factor families (AP2, bHLH, bZIP, CG-1, Myb/SANT, TCP) were found in both time periods. Four transcription factor families were unique to day 1 samples (AT hook, CxC,

Homeodomain, and NAC/NAM), whereas, two transcription factor families were unique

to day 11samples (E2F and LOB).
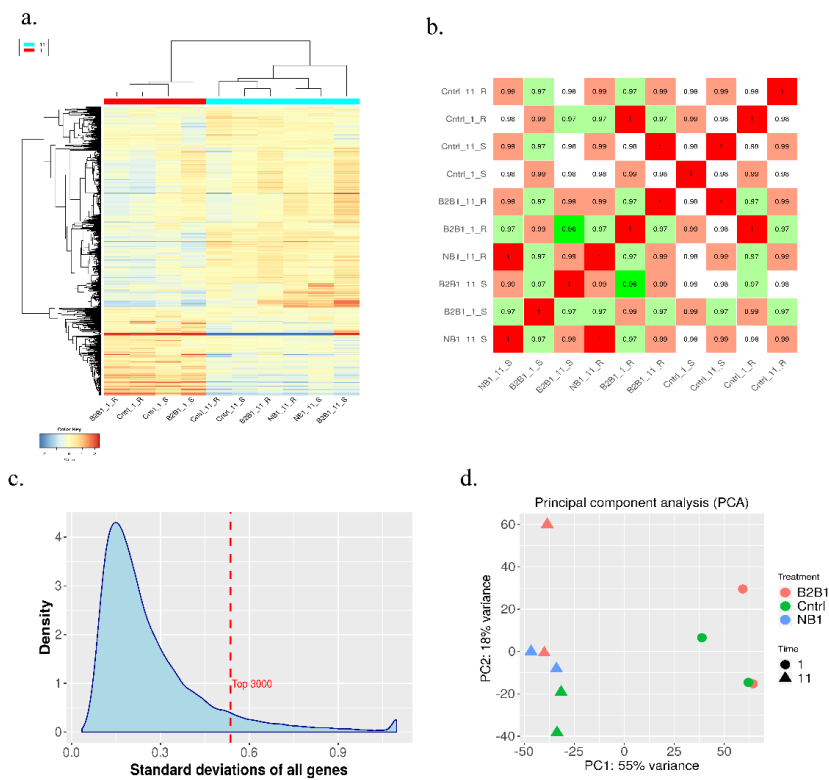
Figure 5.6. Assessment of transcriptomic data. (a) Heatmap of top 3,000 variable genes, (b) Correlation matrix, (c) Gene SD distribution, and (d) A PCA plot.
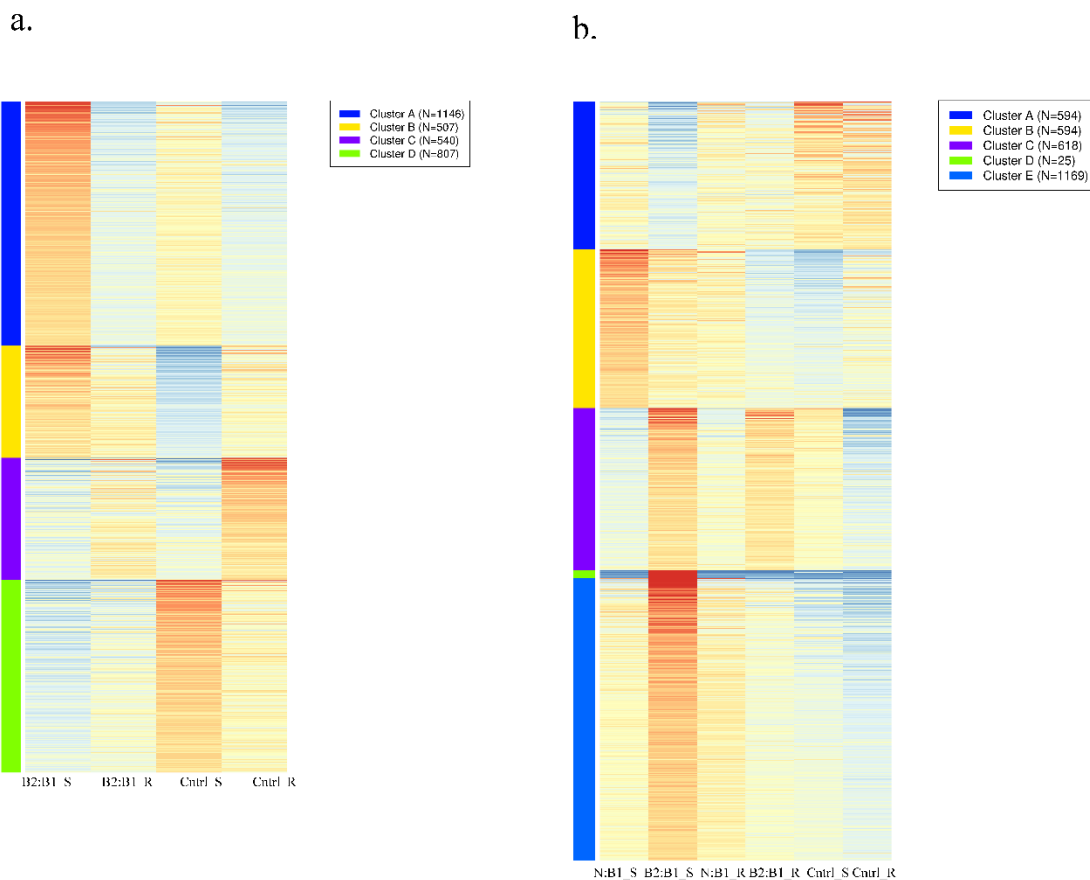
Figure 5.7. K-means clustering of top 3,000 most highly variable genes in Day-1 samples (a) and Day-11 samples (b).

5.3.5. Gene Expression Analysis

The pair wise comparisons between treatments in two different treatments (none: B1 and B2: B1) at day 1 and day 11 with FDR < 0.01 and log2fold-change > 2 as cutoffs resulted differentially expressed genes (DEGs) shown in Table 5.2. We further investigated these genes using Venn diagrams (Figure 5.8). At day 1, we found 746 and 243 DEGs in susceptible and resistant cultivars, respectively treated with biotype 2 as inducer and biotype 1 as response population (B2: B1). Whereas, 981 and 407 DEGs

were found in susceptible and resistant cultivars, respectively at day 11 treated with

biotype 2 as inducer and biotype 1 as response population (B2: B1). At day 11 we found

520 and 377 DEGs in susceptible and resistant cultivars treated with no inducer and

biotype 1 as a response population (none: B1). In total, at day 11, we found 1,274 and

638 DEGs in susceptible and resistant cultivars, respectively upon comparing treatments

with none: B1 and B2: B1.

Table 5.2. A number of up-regulated and down regulated DEGs in different comparisons.

Time	Comparisons	Cultivar	Up regulated	Down regulated
Day 1	B2:B1 vs Control	Susceptible	364	382
	B2:B1 vs Control	Resistant	239	4
Day 11	none:B1 vs Control	Susceptible	196	324
	B2:B1 vs Control	Susceptible	660	321
	none:B1 vs Control	Resistant	154	223
	B2:B1 vs Control	Resistant	214	196

a.

N:B1(R)

228   152   258

B2:B1(R)

b.

N:B1(S)

290   233   751

B2:B1(S)
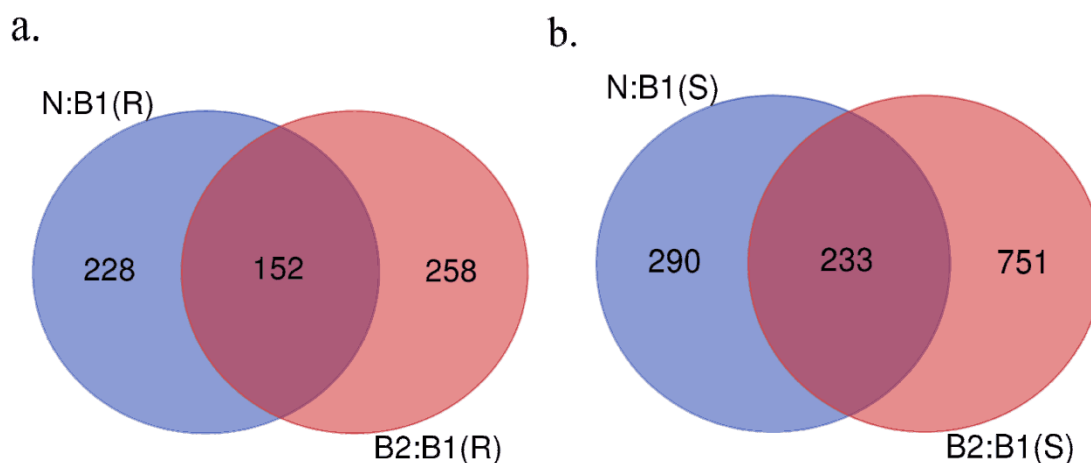
Figure 5.8. Venn diagram showing DEGs for two treatments N: B1 (none: B1) and B2:
B1 in resistant (a) and susceptible cultivars (b).

5.3.6. GO, KEGG Enrichment and MapMan Analysis

The 1,274 and 638 DEGs in susceptible and resistant cultivars, respectively were subjected to GO enrichment analysis for biological process, molecular function, and KEGG pathways. In susceptible cultivar, the DEGs were enriched for various biological processes including Jasmonic acid mediated signaling pathway, Response to chitin, Phenylpropanoid metabolic process, Regulation of defense response, Response to chemical or organic substance, Response to wounding, Hormone metabolic process, Reactive oxygen species metabolic process, Regulation of macromolecule biosynthetic and metabolic processes. Among them, Jasmonic acid mediated signaling pathway was unique to none: B1 treatment and Phenylpropanoid biosynthetic process and Glucosinolate metabolic process were unique to B2: B1 treatment (Figure 5.9). In terms of KEGG pathways, these genes were enriched for Cutin, suberine and wax biosynthesis (FDR=5.36E-07), Biosynthesis of secondary metabolites (FDR=5.36E-07), Glucosinolate biosynthesis (FDR=1.04E-05), Phenylpropanoid biosynthesis (FDR=4.89E-05), MAPK signaling pathway (FDR=8.84E-05), Plant hormone signal transduction (FDR=0.047596525) and others represented in Figure S1 and Supplementary File 3. Promoter analysis of 1,274 DEGs showed 30 enriched transcription factor binding motifs. Enriched binding motifs consisted of seven transcription factor families: AP2, bHLH, bZIP, CG-1, LOB, SBP, and TCP (Supplemental File 3).

a.
- 1.1e-07 Response to stress
- 4.7e-10 Defense response
- 4.6e-07 Response to organic substance
- 1.6e-10 Cellular response to organic substance
- 1.1e-10 Cellular response to chemical stimulus
- 1.4e-06 Response to endogenous stimulus
- 1.1e-06 Response to hormone
- 2.4e-10 Response to acid chemical
- 5.5e-10 Cellular response to oxygen-containing compound
- 4.9e-12 Cellular response to acid chemical
- 8.6e-09 Response to oxygen-containing compound
- 5.2e-08 Cellular response to hormone stimulus
- 1.6e-08 Hormone-mediated signaling pathway
- 7.1e-08 Cellular response to endogenous stimulus
- 6.3e-16 Cellular response to jasmonic acid stimulus
- 6.3e-16 Jasmonic acid mediated signaling pathway
- 3.5e-15 Response to jasmonic acid
- 6.8e-08 Regulation of signal transduction
- 4.1e-16 Regulation of jasmonic acid mediated signaling pathway
- 6.8e-08 Regulation of cell communication
- 6.8e-08 Regulation of signaling
- 1.1e-11 Response to wounding
- 1.8e-09 Regulation of response to stress
- 1.1e-11 Regulation of defense response
- 1.6e-07 Regulation of response to stimulus
- 1.6e-07 Cellular response to stimulus
- 5.9e-09 Response to stimulus
- 7.8e-10 Response to chemical
- 2.4e-06 Hormone metabolic process
- 5.1e-08 Hormone biosynthetic process

b.
- 4.6e-02 Cellular response to toxic substance
- 4.2e-02 Cellular oxidant detoxification
- 4.6e-02 Detoxification
- 4.6e-02 Cellular detoxification
- 3.3e-02 Hydrogen peroxide catabolic process
- 3.3e-02 Hydrogen peroxide metabolic process
- 4.2e-02 Antibiotic catabolic process
- 4.2e-02 Cofactor catabolic process
- 4.6e-02 Reactive oxygen species metabolic process
- 2.6e-02 Response to oxidative stress
- 1.8e-02 Oxidation-reduction process
- 3.3e-02 Lignin catabolic process
- 3.3e-02 Phenylpropanoid catabolic process
- 3.6e-02 Lignin metabolic process
- 2.6e-02 Phenylpropanoid metabolic process
- 2.6e-02 Lipid transport
- 2.6e-02 Lipid localization
- 4.6e-02 Cutin biosynthetic process

c.
- 8.2e-03 Cellular response to toxic substance
- 8.0e-03 Cellular detoxification
- 8.2e-03 Detoxification
- 1.0e-02 Cellular oxidant detoxification
- 3.0e-03 Response to toxic substance
- 2.4e-05 Oxidation-reduction process
- 1.9e-02 Response to organonitrogen compound
- 2.5e-06 Response to chitin
- 5.3e-03 Response to nitrogen compound
- 1.2e-05 Response to drug
- 1.4e-05 Response to oxygen-containing compound
- 2.2e-06 Response to chemical
- 3.3e-04 Phenylpropanoid metabolic process
- 5.7e-06 Secondary metabolic process
- 1.9e-02 Secondary metabolite biosynthetic process
- 4.4e-03 Phenylpropanoid biosynthetic process
- 2.0e-02 Lignin metabolic process
- 5.3e-03 Proanthocyanidin biosynthetic process
- 3.4e-03 Response to wounding
- 4.4e-03 Glycosinolate metabolic process
- 4.4e-03 S-glycoside metabolic process
- 4.4e-03 Glucosinolate metabolic process
- 2.0e-02 Indole glucosinolate metabolic process
- 1.4e-03 Flavone biosynthetic process
- 1.4e-03 Flavone metabolic process
- 1.2e-04 Flavonoid metabolic process
- 8.4e-05 Flavonoid biosynthetic process
- 5.3e-03 Pigment metabolic process
- 3.0e-03 Pigment biosynthetic process
- 1.9e-02 Asparagine metabolic process

Figure 5.9 Enriched GO biological processes specific to treatments in susceptible cultivar at day 11. None: B1 (a), common (b), and B2: B1(c).

Differentially expressed genes visualized using biotic stress pathway integrated into MapMan showed distinct expression patterns in none: B1 and B2: B1 treatments in both susceptible and resistant cultivars. The biotic stress overview pathway generated by MapMan demonstrated the involvement of multifaceted defense related genes in the presence of inducer and no inducer population in both susceptible and resistant plants. In susceptible reaction, 280 (of 523) with 26 bins and 362 (of 984) DEGs with 25 bins were associated with the biotic stress pathway for none: B1 and B2: B1 treatments, respectively. As compared to treatment none: B1, upregulated genes related to abiotic stress (bin 20.2), peroxidases (bin 26.12), abscisic acid hormone pathway (bin 17.1), respiratory burst (bin 20.1.1), glutathione S transferase (bin 26.9), pathogenesis related (PR)-proteins (bin 20.1.7), and secondary metabolism (bin 16), and heat shock proteins (HSPs) (bin 20.2.1).
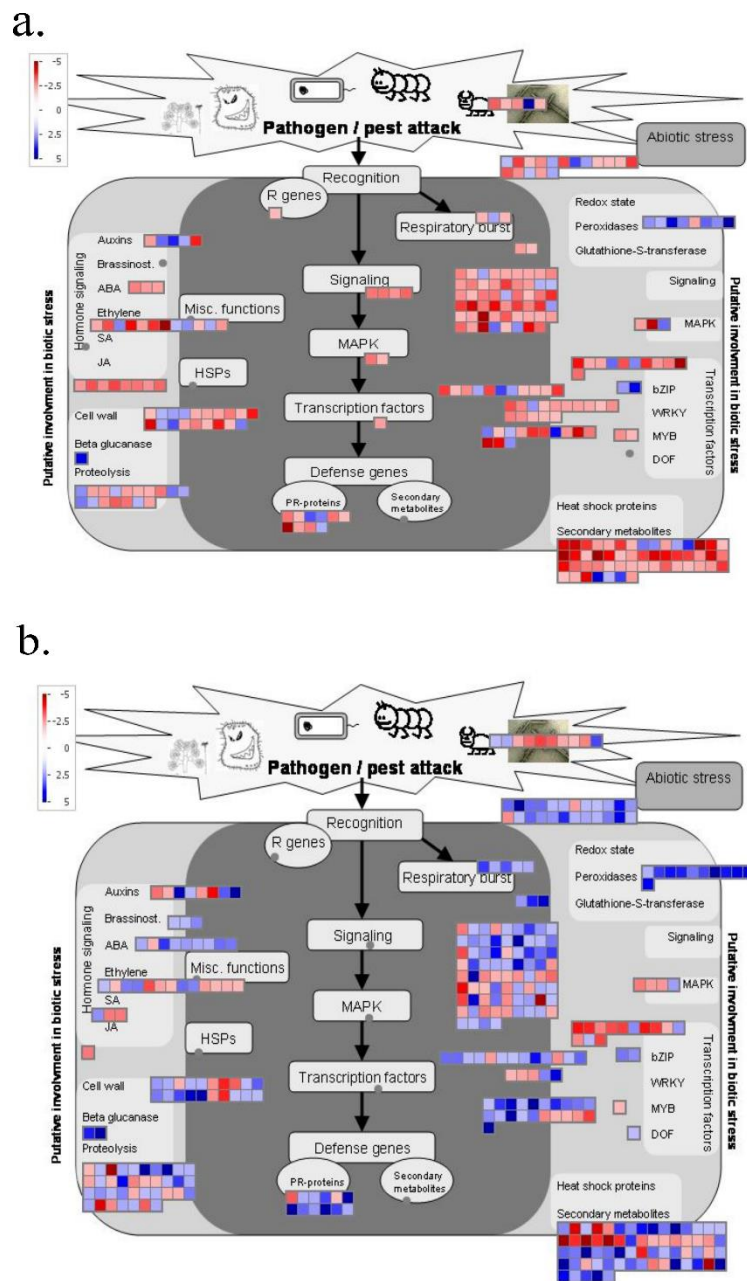
Figure 5.10. Biotic stress pathway overview of differentially expressed genes in susceptible cultivar at day 11. None: B1 (a) and B2: B1 (b). Blue color indicates the up-regulated and red color indicates the down regulated genes. False discovery rate (FDR) *p* < 0.01 and log2fold change ≥ 2 or ≤ -2 were used to identify the differentially expressed genes.

Whereas, 638 DEGs in resistant cultivar were particularly enriched for

Photosynthesis, Lignin metabolic process, negative regulation of endopeptidase activity,

response to cytokinin, Inositol catabolic process which were different from the

susceptible reaction (Figure S2). Among them, response to chitin, lignin catabolic and

metabolic process, asparagine metabolic process, response to chemical were unique to

none: B1 treatment, whereas, response to reactive oxygen species, photosynthesis, and

regulation of endopeptidase activity were unique to B2: B1 treatment (Figure 5.10).

These genes were enriched for KEGG pathways such as Photosynthesis (FDR=

0.005883), Glutathione metabolism (FDR=0.009895), Cutin, suberine and wax

biosynthesis (FDR=0.012764), Cysteine and methionine metabolism (FDR=0.046797),

and Flavonoid biosynthesis (FDR=0.046797) (Supplementary File 4). Promoter analysis

of 638 DEGs showed 30 enriched transcription factor binding motifs. Enriched binding

motifs consisted of four transcription factor families: bHLH, bZIP, CG-1, and TCP

(Supplemental File 4).
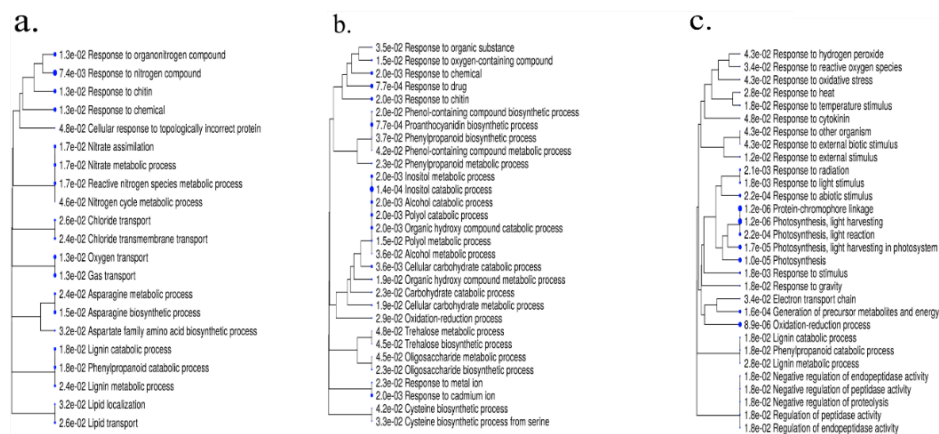
Figure 5.11. Enriched GO biological processes specific to treatments in resistant cultivar at day 11. None: B1 (a), common (b), and B2: B1 (c).

In resistant reaction, MapMan biostress pathway revealed 154 (of 380) with 21 bins and 176 (of 410) DEGs with 23 bins associated with the biotic stress pathway for none: B1 and B2: B1 treatments, respectively. As compared to treatment none: B1, upregulated genes related to transcription factors [WRKY (bin 27.3.32), MYB (27.3.25)], peroxidases (bin 26.12), abscisic acid hormone pathway (bin 17.1), respiratory burst (bin 20.1.1), glutathione S transferase (bin 26.9), salicylic acid hormone pathway (bin 17.8), jasmonic acid hormone pathway (bin 17.7), pathogenesis related (PR)-proteins (bin 20.1.7), and secondary metabolism (bin 16), and heat shock proteins (HSPs) (bin 20.2.1).
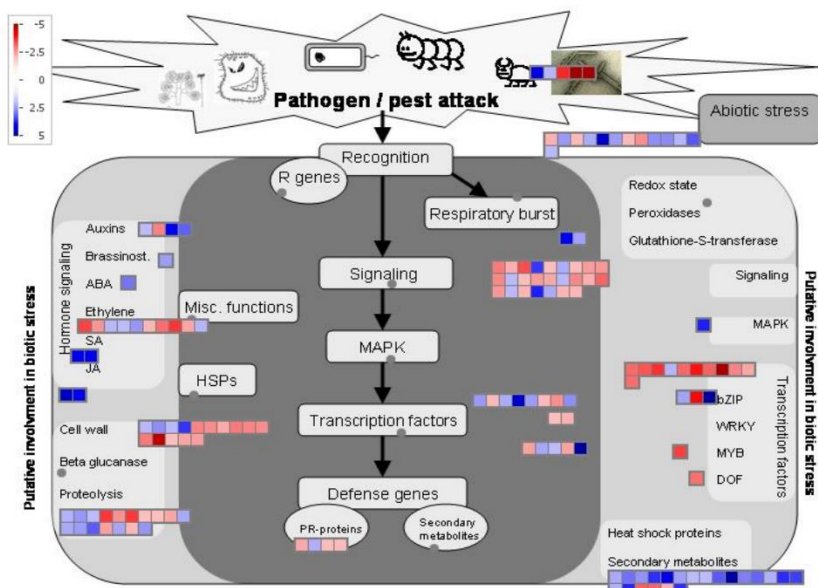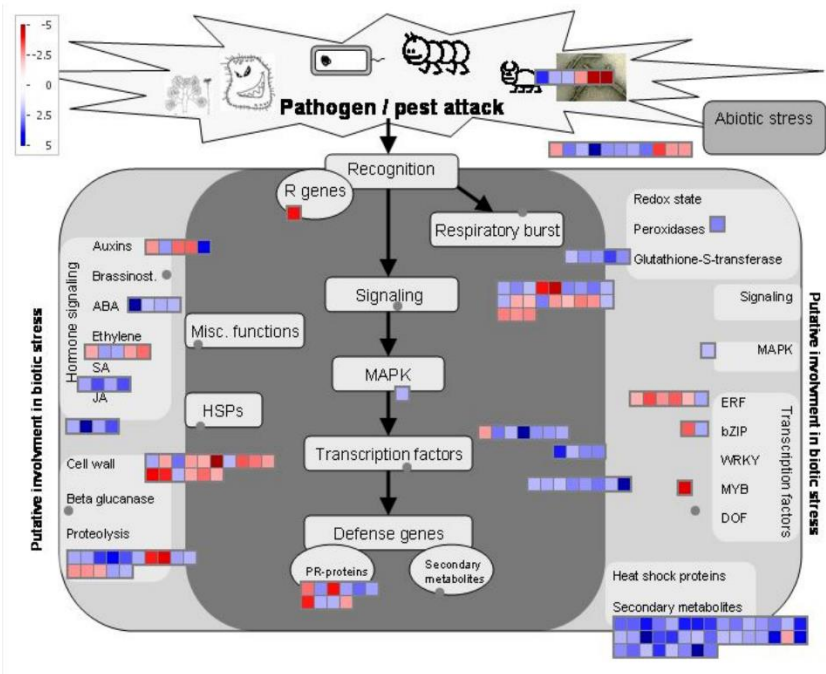
Figure 5.12. Biotic stress pathway overview of differentially expressed genes in resistant cultivar at day 11. None: B1 (a) and B2:B1 (b). Blue color indicates the up-regulated and red color indicates the down regulated genes. False discovery rate (FDR) $p < 0.01$ and log2fold change ≥ 2 or $\leq$ -2 were used to identify the differentially expressed genes.

5.3.7. Comparison of the DEGs between Two-time Points

Further, we compared DEGs for samples treated with biotype 2 as inducer and biotype 1 as response population (B2: B1) on day1 and day 11. Of 626 DEGs in resistant cultivar, 216 were unique to day 1 samples, 383 were unique to day 11 samples and 27 were expressed at both time points (Figure 5.11a). Likewise, of 1,621 DEGs in susceptible cultivar, 637 were unique to day 1 samples, 872 were unique to day 11 samples and 112 were expressed at both time points (Figure 5.11b).

a.

Day1(R)

216   27   383

Day11(R)

b.

Day1(S)

637   112   872
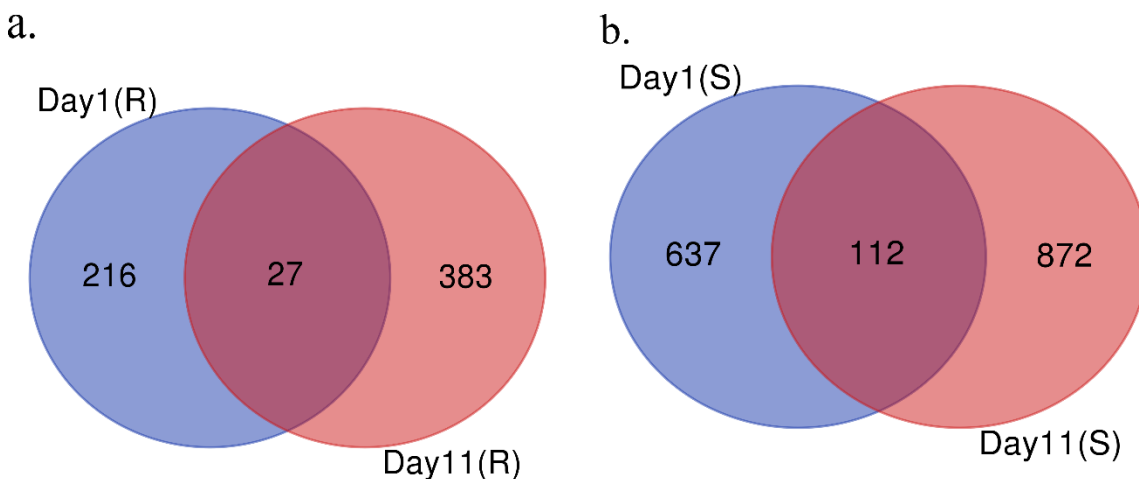
Day11(S)

Figure 5.13. Venn diagram showing DEGs for treatment B2: B1 at day 1 and day 11 in resistant (a) and susceptible (b) cultivars.

At day 1, MapMan biostress pathway revealed 284 (of 749) with 24 bins and 90 (of 243) DEGs with 16 bins associated with the biotic stress pathway in susceptible and resistant cultivars with B2: B1 treatment, respectively. As compared to a susceptible reaction, the resistant reaction showed fewer bins associated with the biostress pathway with almost all upregulated genes.
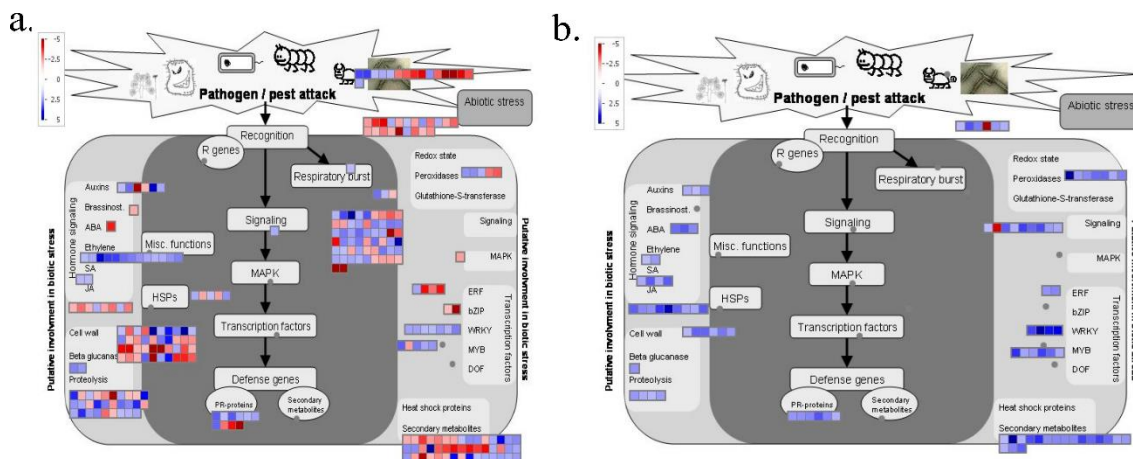
Figure 5.14. Biotic stress pathway overview of differentially expressed genes at day 1 with B2: B1 treatment. Susceptible (a) and resistant (b). Blue color indicates the up-regulated and red color indicates the down regulated genes. False discovery rate (FDR) p < 0.01 and log2fold change ≥ 2 or ≤ -2 were used to identify the differentially expressed genes.

Table 5.3. List of 27 common DEGs for treatment B2:B1 at day 1 and day 11 in a resistant cultivar.

Feature ID	Day 1 log2fc	Day 11 log2fc	Top Arabidopsis Hit	Symbols	Gene Description
Glyma.01G021000	3.35	3.84	AT4G37990.1	ELI3-2, ELI3, ATCAD8, CAD-B2	elicitor-activated gene 3-2
Glyma.02G054200	2.32	2.38	AT1G19640.1	JMT	jasmonic acid carboxyl methyltransferase
Glyma.02G108700	2.47	2.87	AT3G29000.1		Calcium-binding EF-hand family protein
Glyma.03G068200	2.46	-6.07	AT1G73330.1	ATDR4, DR4	drought-repressed 4
Glyma.03G222600	2.72	2.62	AT2G47140.1		NAD(P)-binding Rossmann-fold superfamily protein
Glyma.06G004400	2.42	2.54	AT4G38650.1		Glycosyl hydrolase family 10 protein
Glyma.06G145300	2.30	2.80	AT5G05340.1		Peroxidase superfamily protein
Glyma.06G182700	2.48	2.14	AT5G14740.2	CA2, CA18, BETA CA2	carbonic anhydrase 2
Glyma.06G299900	3.11	2.17	AT2G31180.1	ATMYB14, MYB14AT, MYB14	myb domain protein 14
Glyma.07G034900	2.73	2.29	AT1G55020.1	LOX1, ATLOX1	lipoxygenase 1
Glyma.08G199300	2.22	3.42	AT1G14520.2	MIOX1	myo-inositol oxygenase 1
Glyma.11G051800	2.24	2.60	AT4G37340.1	CYP81D3	cytochrome P450, family 81, subfamily D, polypeptide 3
Glyma.12G092600	2.74	5.48	AT3G51680.1		NAD(P)-binding Rossmann-fold superfamily protein
Glyma.12G199200	3.49	2.56	AT3G23250.1	MYB15, ATY19, ATMYB15	myb domain protein 15
Glyma.13G084000	2.11	-5.01	AT1G77760.1	NIA1, GNR1, NR1	nitrate reductase 1
Glyma.13G349300	2.30	2.26	AT3G13790.1	ATCWINV1, ATBFRUCT1	Glycosyl hydrolases family 32 protein
Glyma.14G102900	3.91	2.80	AT1G80840.1	WRKY40, ATWRKY40	WRKY DNA-binding protein 40
Glyma.15G009500	2.73	3.27	AT1G80160.1		Lactoylglutathione lyase / glyoxalase I family protein
Glyma.15G072400	2.05	4.26	AT4G27450.1		Aluminium induced protein with YGL and LRDR motifs
Glyma.17G079000	2.12	2.01	AT3G29575.4	AFP3	ABI five binding protein 3
Glyma.17G156100	2.61	4.78	AT4G37850.1		basic helix-loop-helix (bHLH) DNA-binding superfamily protein
Glyma.17G222500	3.47	3.84	AT1G80840.1	WRKY40, ATWRKY40	WRKY DNA-binding protein 40
Glyma.17G242400	2.06	2.21	AT4G25000.1	ATAMY1, AMY1	alpha-amylase-like
Glyma.18G246100	3.62	3.78	AT4G37850.1		basic helix-loop-helix (bHLH) DNA-binding superfamily protein
Glyma.18G273200	2.38	2.42	AT5G24530.1	DMR6	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein
Glyma.19G245400	2.42	-3.05	AT3G04720.1	PR4, HEL, PR-4	pathogenesis-related 4
Glyma.20G036100	3.70	7.00	AT2G02990.1	RNS1, ATRNS1	ribonuclease 1

The 27 overlapped DEGs in resistant cultivar represented the variation in the level of expression at day 1 and day 11 (Table 5.3). In both cultivars, we observed no difference, increased and decreased expression pattern of the genes (Table 5.3 and Table S2). Particularly in resistant cultivar, elicitor-activated gene 3-2 (Glyma.01G021000), jasmonic acid carboxyl methyltransferase (Glyma.02G054200), Calcium-binding EF-hand family protein (Glyma.02G108700), 2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein (Glyma.18G273200) showed static level or no difference of expression at day 1 and day 11. Whereas, the expression of NAD(P)-binding Rossmann-fold superfamily protein (Glyma.03G222600) increased from 2.74 to 5.48 log2fc, myo-inositol oxygenase 1 (Glyma.08G199300) from 2.22 to 3.42 log2fc, Aluminium induced protein with YGL and LRDR motifs (Glyma.15G072400) from 2.05 to 4.26 log2fc, basic helix-loop-helix (bHLH) DNA-binding superfamily protein from 2.61 to 4.68 log2fc, and ribonuclease 1 (Glyma.20G036100) from 3.70 to 7.00 log2fc. In contrast, the expression of drought-repressed 4 (Glyma.03G068200) was decreased from 2.46 to -6.07 log2fc, nitrate reductase 1 (Glyma.13G084000) from 2.11 to -5.01, pathogenesis-related 4 (Glyma.19G245400) from 2.42 to -3.05 log2fc. At day 1, particularly, seven genes belonging to peroxidases and six cytochrome P450s were highly upregulated. In addition, disease resistance-responsive (dirigent-like protein) family protein (Glyma.03G045600, Glyma.08G019900, Glyma.12G030300) were expressed in the range of 2.6 to 3.3 log2foldchange, Kunitz family trypsin and protease inhibitor protein (Glyma.08G235300, Glyma.16G212500, Glyma.08G235400) were expressed in the range of 2.2 to 4.7 log2foldchange, laccase 3 (Glyma.02G231600, Glyma.14G198900) were expressed in the range of 2.8 to 3.1, TRICHOME

BIREFRINGENCE-LIKE 27 (Glyma.01G034600, Glyma.02G031400) and TRICHOME BIREFRINGENCE-LIKE 42 (Glyma.12G233500) were expressed in the range of 2.0 to 2.1 log2foldchange, Ferritin/ribonucleotide reductase-like family protein (Glyma.16G056300) was expressed by 3.129 log2foldchange, WRKY24 (Glyma.18G238600) was expressed by 4.2 log2foldchange (Supplementary File 5). At day 11, particularly, wall associated kinase 5 (Glyma.13G035900) was expressed by 2.78 log2fold change, glutathione S-transferases (Glyma.18G043700, Glyma.11G212900, Glyma.07G139600, Glyma.07G139900, Glyma.07G139700) were expressed in the range of 2.0 to 3.5 log2foldchange, Toll interleukin-1 receptor-like Nucleotide-binding site Leucine-rich repeat (TNL) genes (Glyma.03G048600, Glyma.03G052800, Glyma.03G048700, Glyma.03G047700) expressed in the range of 2.4 to 3.1 log2foldchange, senescence-related genes (Glyma.06G273600, Glyma.13G222100, Glyma.15G090100, Glyma.16G052000) were expressed in the range of 2.1 to 4.1 log2foldchange, UDP-glucosyltransferases (Glyma.08G109100, Glyma.02G105000, Glyma.10G062600, Glyma.20G196000) were expressed in the range of 2.1 to 3.65 log2foldchange, myo-inositol oxygenases (Glyma.07G013900, Glyma.05G224500) were expressed in range of 3.5 to 3.9 log2foldchange, ferritin 4 (Glyma.02G262500) was expressed by 2.26 log2foldchange, WRKY40 (Glyma.17G222300) and WRKY67 (Glyma.03G002300) were expressed in the range of 2.1 to 2.7 log2foldchange (Supplementary File 5).

5.3.8. DEGs Coincident with *Rag* QTL Genes

We identified 1,691 non-redundant genes in the *Rag* QTLs: *Rag1* [15], *Rag1b* [16], *Rag1c* [17], *Rag2* [18], *Rag3* [19, 20] , *Rag4* [19], *Rag3* [16] , *Rag3b* [21], *Rag3c* [22], *Rag4* [17], *Rag5* [23] , *Rag6* [22]; *qChrom.07.1*, *qChrom.16.1*, *qChrom.13.1*, *qChrom.17.1* [24] and compared with the DEGs found in the resistant cultivar, LD12-15813Ra. We found 14 DEGs that were coincident with the *Rag* QTL genes with lipoxygenase 1 (Glyma.07G034900) being up-regulated at both day 1 (2.73 log2foldchange) and day 11 (2.29 log2foldchnage) treated with B2 as an inducer population and B1 as a response population and Gibberellin-regulated family protein (Glyma.17G237100) downregulated at day 11 in both treatment conditions none: B1 and B2: B1. Protein kinase family proteins with leucine-rich repeat domain (Glyma.16G065600) and Gibberellin-regulated family protein (Glyma.17G237100) were downregulated at day 1 and day 11, respectively treated with B2 as an inducer population and B1 as a response population. Likewise, arabinogalactan protein 22 (Glyma.07G087200) and Gibberellin-regulated family protein (Glyma.17G237100) were downregulated treated with no inducer population and B1 as a response population at day 11.

Table 5.4. List of DEGs coincident with *Rag* QTL genes

Time	Treatment	GeneID	Log2foldchange	E-value	Top Arabidopsis Hit	Symbols	Gene Description
Day 1	B2:B1	Glyma.07G034800	3.181	0	AT1G55020.1	LOX1, ATLOX1	lipoxygenase 1
		Glyma.13G183000	5.276	0	NA		
		Glyma.13G183500	2.729	9.57E-10	NA		
		Glyma.16G053300	3.141	0	AT5G41040.1		HXXXD-type acyl-transferase family protein
		Glyma.16G056300	3.129	8.33E-08	AT3G27060.1	TSO2, ATTSO2	Ferritin/ribonucleotide reductase-like family protein
		Glyma.16G065600	-4.430	3.91E-03	AT1G35710.1		Protein kinase family protein with leucine-rich repeat domain
		Glyma.07G034900	2.73	0.00E+00	AT1G55020.1	LOX1, ATLOX1	lipoxygenase 1
Day 11	B2:B1	Glyma.17G237100	-3.45066	0	AT5G59845.1		Gibberellin-regulated family protein
		Glyma.16G052000	2.146059	0	AT3G02040.1	SRG3	senescence-related gene 3
		Glyma.07G051500	2.670101	0	AT4G00870.1		basic helix-loop-helix (bHLH) DNA-binding superfamily protein
		Glyma.13G035900	2.780425	0	AT1G21230.1	WAK5	wall associated kinase 5
		Glyma.07G061500	2.96304	0	AT2G46240.1	BAG6, ATBAG6	BCL-2-associated athanogene 6
		Glyma.07G034900	2.29	0	AT1G55020.1	LOX1, ATLOX1	lipoxygenase 1
	none:B1	Glyma.07G051500	2.86115	0	AT4G00870.1		basic helix-loop-helix (bHLH) DNA-binding superfamily protein
		Glyma.07G087200	-2.05063	0	AT5G53250.1	AGP22, ATAGP22	arabinogalactan protein 22
		Glyma.17G237100	-2.03562	1.99E-10	AT5G59845.1		Gibberellin-regulated family protein

5.4. Discussion

This experiment is the first attempt to characterize the induced susceptibility effect that promotes the avirulent *A. glycines* populations in both resistant and susceptible cultivar treated with virulent inducer populations. Previously, this effect was initially tested with *Rag1* + *Rag*2 (IA3027RA12) cultivar and subsequent tests in near-isogenic soybean cultivars containing no *Rag* genes (IA3027), *Rag1* (IA3027RA1), using biotype 1 and biotype 2 soybean aphids in a growth chamber and semi-field settings [32]. We first validated this effect using susceptible soybean cultivar (LD12-15838R) with no *Rag* gene and the resistant cultivar (LD12-15813Ra) with *Rag1* gene in the greenhouse settings with slight modifications on response population density (15 instead of five). In

the meantime, we collected leaves samples for the transcriptomic study. We observed both 'feeding facilitation' [53] and 'obviation of resistance' [31] which are the two subcategories of induced susceptibility. Feeding facilitation refers to the condition where conspecifics are favored on either susceptible or resistant host plants in presence of herbivore, irrespective of its genotype. Whereas, obviation of resistance refers to the condition where avirulent conspecifics on the resistant plant are favored in the resistant host plant in the presence of virulent herbivore. We chose treatments with no aphids (control), biotype 2: biotype1 (B2: B1) and no aphids: biotype1 (none: B1) collected at day 1 and day 11 for the transcriptomic study. These treatments were chosen as we expect some insights on gene expression pattern in resistant and susceptible cultivars in time course response in presence or absence of virulent soybean aphids as an inducer population and avirulent soybean aphids as a response population. The day 1 samples were selected expecting some response to the host by the inducer population. The day 11 samples were selected as we expected both physical and metabolic changes caused by both inducer and response populations.

The initial WGCNA analysis revealed 11 and 15 co-expression modules on day 1 and day 11, respectively enriched for various pathways in both resistant and susceptible cultivars. At day 1 or 24 hours, we found an enriched pathway for cysteine and methionine metabolism which was also enriched in the DEGs in resistant cultivar discussed below. Many plant species utilize S-methylmethionine and glutathione to transport sulphur molecules in the phloem [54]. Aphids might have an efficient mechanism for the production of methionine and cysteine from the phloem metabolites [55]. It has been shown that peach aphid and pea aphid in symbiosis with the

endosymbiont, *Buchnera aphidicola* incorporate sulphur from inorganic sulphate transported to the phloem sap [55, 56]. The presence of aphid endosymbiotic bacteria [57] in the aphids might be one of the causes for feeding facilitation and obviation of resistance by soybean aphids. The possibility of the role of endosymbionts including plant viruses and aphid effector molecules causing induced susceptibility was discussed by Varenhorst et al. 2015 [32]. Other enriched pathways at day 1 were Phenylpropanoid biosynthesis, Alpha-Linolenic acid metabolism, Fatty acid biosynthesis. Previous studies have shown that the phenylpropanoid pathway was induced in the resistant (*Rag1*) Dowling cultivar at 6 and 12 h after aphid feeding [58]. The pathways related to α-linolenic acid metabolism and fatty acid biosynthesis corresponds to the precursor for jasmonic acid pathway via the oxylipin pathway [59]. This shows that soybean aphids can induce hormone response inducing changes in fatty acid metabolism within 24 hours. The production of various phytohormones such as JA including SA and ET upon aphid infestation on the response of resistant (*Rag1*) and susceptible near-isogenic soybean lines [39]. Such effect was also seen in two soybean varieties (DK 27–52 and DK 28–52) when infested with soybean aphid in the field environment [60]. At day 11, we found an enriched pathway for Flavonoid biosynthesis, Plant pathogen interaction, MAPK signaling pathway, and Glucosinolate biosynthesis. The interaction of plant and pathogen involves pathogen-associated molecular patterns (PAMPs) of pathogens by pattern recognition receptors (PRRS) of the host [61]. These plant-parasite interactions have caused a battle in the molecular avenue where evolutionary arms race takes place [62]. There are various models that describe plant-pathogen interactions such as the gene for gene model, guard model, decoy model, bait and switch model, and zig-zag model [63,

64, 65, 66]. Zig-zag model depicts the interaction between plant and parasites [64]. It is still unknown if aphid and other insects interaction follow the particular model [62]. As reviewed by Wu and Baldwin, 2010 [67] early defense signaling events take place in a cell of insect attacked leaf. Briefly, aphid elicitors are perceived by the receptors on plasma membrane trigger Ca^{2+} channels and produce Ca^{2+}. Ca^{2+} binds with NADPH oxidase which gets enhanced through phosphorylation by CDPKs eventually producing reactive oxygen species (ROS). MAPK pathways are activated quickly among which SIPK and WIPK trigger the synthesis of Jasmonic acid (JA) and JA-Ile (JA-isoleucine) which is a central regulator of plant innate immunity [68]. Another enriched pathway at day 11 was Glucosinolate biosynthesis. The involvement of secondary metabolites such as glucosinolates have been documented in two separate studies as a defensive compound when *Myzus persicae* infested *Arabidopsis* for three [69] and seven days [70].

The K-means clustering revealed five and four clusters for day 11 and day 1 samples. The pathway enrichment analysis of the clusters supported the enrichment of entire modules obtained from WGCNA analysis. Enriched binding motifs of these clusters revealed AT hook [71], CxC [72], Homeodomain [73], and NAC/NAM [74] transcription factor families unique to day 1 samples. whereas, two transcription factor families were unique to day 11samples (E2F [75], and LOB [76]). Six transcription factor families (AP2 [77], bHLH [78], bZIP [79] , CG-1[80], Myb/SANT [81], TCP [82]) were found in both time periods. Among them, AP2, bHLH, bZIP, CG-1, LOB, SBP, and TCP were particularly enriched in susceptible reaction whereas, bHLH, bZIP, CG-1, and TCP were enriched in resistant reaction.

At day 11, upon analyzing DEGs between treatments none: B1 and B2: B1 we observed a significant number of DEGs in susceptible cultivar than resistant cultivar (1,274 vs 638). The DEGs in susceptible cultivar were enriched for many biological processes related to defense programs such as MAPK signaling pathway, Plant hormone signal transduction, and Plant-pathogen interaction. These are the major components of the PTI program of defense mechanism. Such an effect in which significant induction of defense programs in susceptible cultivar aphid-susceptible SD01-76R when infested with soybean aphid for 21 days [83]. The DEGs in resistant cultivar were enriched for Photosynthesis, Glutathione metabolism, Cutin, suberine and wax biosynthesis, Cysteine and methionine metabolism, and Flavonoid biosynthesis. Particularly, glutathione metabolism was enriched in which one gene, glutathione peroxidase 6 (Glyma.01G219400) was upregulated by 2.12 log2foldchange in the none: B1 treatments. Whereas, three genes belonging to glutathione S-transferases (Glyma.07G139700, Glyma.07G139900, Glyma.14G067200) were upregulated by 2.04 to 2.5 log2foldchange in the B2: B1 treatments. The structural damage on the host upon aphid feeding may be linked to the accumulation of excessive reactive oxygen species (ROS) in the attacked organs [84]. Plant glutathione S-transferases (GSTs) make such endogenous substrates and xenobiotics (e.g., ROS) less toxic upon adding glutathione molecule via nucleophilic or addition reactions [85]. We observed glutathione peroxidase6 being upregulated in none: B1 treatments. Sometimes, GSTs exhibit glutathione-peroxidase activity for the reduction of hydroperoxides [86]. The enrichment of Cysteine and methionine metabolism which was also observed in initial WGCNA analysis at day 1 was observed in DEGs in resistant cultivar at day 11. This shows that cysteine and methionine

metabolism pathway is active from initial aphid feeding to the 11th day. Another enriched pathway in the resistant cultivar was photosynthesis. Previous transcriptomic study on soybean near-isogenic lines differing in alleles for an aphid resistance gene, *Rag5* following infestation by soybean aphid biotype 2 has shown DEGs enriched for photosynthesis [87]. Physiologically, photosynthesis rates have been reduced up to 50% on soybean aphid infested leaflets [88].

The comparison of DEGs was further expanded to see a pattern of the expression of DEGs especially focusing to common and unique genes at day 1 and day 11 in the resistant cultivar when treated with biotype 2 as an inducer population. Particularly on day 1, we observed upregulation TRICHOME BIREFRINGENCE-LIKE 27 and -42 proteins and laccase 3. TRICHOME BIREFRINGENCE-LIKE genes contribute to the synthesis and deposition of the secondary wall [89]. Likewise, laccase genes also play a role in cell wall lignification [90]. The cell wall modification and deposition of callose are considered as the chemical defense responses during PAMP-triggered immunity (PTI) response after recognition of components from the aphid saliva [91]. The role of laccase in tolerance to the insect pests cotton bollworm (*Helicoverpa armigera*) and cotton aphid (*Aphis gosypii*) has been shown in cotton [92]. Upregulation of other genes at day 1 involved peroxidases, cytochrome P450s. The role of peroxidases in scavenging ROS during the defense mechanism has been clearly documented in plant-aphid interactions [93, 94] including soybean-soybean aphid interaction [35]. Plant cytochrome P450s are importantly involved in jasmonic acid mediated plant defense in response to wound and insect attack [95]. Other DEGs belonged to disease resistance-responsive (dirigent-like protein) family protein, Kunitz family trypsin, protease inhibitor protein, and

Ferritin/ribonucleotide reductase-like family protein. Previously, DIR-like proteins were also upregulated during feeding of spruce (*Picea* spp.) stem-by boring insects (i.e., white pine weevil, *Pissodes strobi*) in bark tissue and defoliating insects (i.e., western spruce budworm, *Choristoneura occidentalis*) in green apical shoots [96]. Kunitz family trypsin and protease inhibitor protein target various proteases of phytophagous pests and pathogen as a resistance response [97]. Previously, Kunitz family trypsin and protease inhibitor genes were reported as a differentially expressed gene in tolerant soybean cultivar upon soybean aphid feeding [35]. Another gene that encodes Ferritin/ribonucleotide reductase-like family protein was upregulated at day 1 response. The differential expression of ferritin as a resistance response has been shown in previous studies as a part of constitutive resistance mechanism in soybean-soybean aphid interactions [35, 38, 83]. Upregulation of ferritins in resistant plants can limit the availability of iron to the insect [83]. At day 11, four TNL genes, four homologs of WRKY40, one homolog of WRKY67, four senescence-related genes, four UDP-glucosyltransferases, two myo-inositol oxygenases, five glutathione S-transferases were uniquely upregulated in the resistant cultivar when treated with biotype 2 as an inducer. The expression of four TNL genes at day 11 shows the involvement of canonical resistance genes. Numerous plant resistance genes involved in defense mechanism encode proteins containing nucleotide- binding site (NBS) and a leucine-rich repeat (LRR) motifs [98]. For example, *Vat* gene, which confers resistance to *Aphis gossypii* in melon (*Cucumis melo*) is also an NBS-LRR gene [99].

We examined DEGs that are coincident with the 1,691 genes that were assessed from the *Rag* QTLs. The cultivar used in this experiment is LD12-15813Ra with *Rag1*

gene. The mapping and inheritance mechanism of the *Rag1* gene has been well studied in various soybean cultivars [58, 100, 101, 102]. *Rag1* loci were finely mapped as a 115 kb interval on chromosome 7 through genetic mapping using cultivar Dowling (PI 548663; donor parent of *Rag1*) and Dwight (PI587386; aphid-susceptible parent) [15]. We found 14 DEGs that were coincident with the *Rag* QTL genes. Among them, six genes belonged to *Rag1* QTLs. These genes belonged to lipoxygenase 1 (Glyma.07G034800, Glyma.07G034900) basic helix-loop-helix (Glyma.07G051500, Glyma.07G051500), BCL-2-associated athanogene (Glyma.07G061500) were upregulated while arabinogalactan protein 22 (Glyma.07G087200) was downregulated.

The present study is an effort to characterize the interactions between two different biotypes of soybean aphids in susceptible and resistant soybean cultivars. We validated induced susceptibility effects using demographic datasets obtained from the greenhouse experiment. Further, this effect was characterized using genetic datasets obtained from RNA-seq technique. The characterization was limited to two treatments: one with no inducer population and biotype 1 as a response population and another with biotype 2 as an inducer population and biotype 1 as a response population in both resistant and susceptible cultivars. Many DEGs were common and unique in two cultivars and treatments that were enriched for various biological processes and pathways and were functionally related to known defense mechanisms reported in various host-aphid systems. The responses to aphid biotype 1 infestation in presence or absence of inducer population at day1 and 11 revealed significant differences on the gene enrichment and regulation in resistant and susceptible cultivars. The assessment of DEGs in *Rag* genes QTLs, particularly in *Rag1* containing QTL on chromosome 7, six non-

NBS-LRR genes – Glyma.07G034800, Glyma.07G034900, Glyma.07G051500,

Glyma.07G051500, Glyma.07G061500, Glyma.07G087200 revealed distinct expression

in treatments with absence or presence of inducer population at day 1 and day 11.

However, four TNL genes – Glyma.03G048600, Glyma.03G052800, Glyma.03G048700,

Glyma.03G047700 were upregulated in resistant cultivar treated with biotype 2 as an

inducer population and biotype 1 as response population at day 11 which suggest their

crucial role in the interaction effects. Further experiments based upon metabolomics,

proteomics, and validation of the candidate genes will be needed to understand the

mechanism underlying induced susceptibility effects.

Supplementary Files

https://figshare.com/s/ef9f55016e4d594111a3

Data Record

The raw fastq files were submitted to the National Center for Biotechnology

Information (NCBI, https://www.ncbi.nlm.nih.gov/) and are available with accession

numbers accession (SRR8848023- SRR8848032) under Bioproject PRJNA530958. The

data could be retrieved using fastq-dump tool SRA toolkit

(http://www.ncbi.nlm.nih.gov/sra). The raw transcript abundance counts for all the

samples was deposited at the Gene Expression Omnibus (GEO) database, GSE129626.

Please see Table 1 and reference list for details and links to the data.

Acknowledgments

References

1. Desneux, N.; Wajnberg, E.; Wyckhuys, K.A.; Burgio, G.; Arpaia, S.; Narváez-Vasquez, C.A.; González-Cabrera, J.; Ruescas, D.C.; Tabone, E.; Frandon, J. Biological invasion of European tomato crops by *Tuta absoluta*: ecology, geographic expansion and prospects for biological control. *Journal of Pest Science* 2010, *83*, 197-215.
2. Pimentel, D.; Lach, L.; Zuniga, R.; Morrison, D. Environmental and economic costs of nonindigenous species in the United States. *Bioscience* 2000, *50*, 53-65.
3. Hartman, G.; Domier, L.; Wax, L.; Helm, C.; Onstad, D.; Shaw, J.; Solter, L.; Voegtlin, D.; d'Arcy, C.; Gray, M. Occurrence and distribution of *Aphis glycines* on soybeans in Illinois in 2000 and its potential control. *Plant Health Progress* 2001, *10*.
4. Wu, Z.; Schenk-Hamlin, D.; Zhan, W.; Ragsdale, D.W.; Heimpel, G.E. The soybean aphid in China: a historical review. *Annals of the Entomological Society of America* 2004, *97*, 209-218.
5. Venette, R.; Ragsdale, D. Assessing the invasion by soybean aphid (Homoptera: Aphididae): where will it end? *Annals of the Entomological Society of America* 2004, *97*, 219-226.
6. Meng, J. Genetic analysis of soybean aphid resistance gene in soybean K1621. *Dissertation*. Kansas State University 2010. http://hdl.handle.net/2097/4599
7. Ragsdale, D.W.; Landis, D.A.; Brodeur, J.; Heimpel, G.E.; Desneux, N. Ecology and management of the soybean aphid in North America. *Annual Review of Entomology* 2011, *56*, 375-399.
8. Voegtlin, D.J.; O'neil, R.J.; Graves, W.R. Tests of suitability of overwintering hosts of *Aphis glycines*: identification of a new host association with *Rhamnus*

alnifolia L'Héritier. *Annals of the Entomological Society of America* 2004, *97*, 233-234.

9. Hill, C.; Chirumamilla, A.; Hartman, G. Resistance and virulence in the soybean-*Aphis glycines* interaction. *Euphytica* 2012, *186*, 635-646.

10. McCornack, B.P.; Costamagna, A.; Ragsdale, D. Within-plant distribution of soybean aphid (Hemiptera: Aphididae) and development of node-based sample units for estimating whole-plant densities in soybean. *Journal of Economic Entomology* 2008, *101*, 1488-1500.

11. Ragsdale, D.W.; McCornack, B.; Venette, R.; Potter, B.; MacRae, I.V.; Hodgson, E.W.; O'Neal, M.E.; Johnson, K.D.; O'neil, R.; DiFonzo, C. Economic threshold for soybean aphid (Hemiptera: Aphididae). *Journal of Economic Entomology* 2007, *100*, 1258-1267.

12. Chandran, P.; Reese, J.C.; Khan, S.A.; Wang, D.; Schapaugh, W.; Campbell, L.R. Feeding behavior comparison of soybean aphid (Hemiptera: Aphididae) biotypes on different soybean genotypes. *Journal of Economic Entomology* 2013, *106*, 2234-2240.

13. Fox, C.M.; Kim, K.-S.; Cregan, P.B.; Hill, C.B.; Hartman, G.L.; Diers, B.W. Inheritance of soybean aphid resistance in 21 soybean plant introductions. *Theoretical and Applied Genetics* 2014, *127*, 43-50.

14. Smith, C.M. *Plant resistance to insects. A fundamental approach*; John Wiley and Sons Ltd.: 1989; p.^pp.

15. Kim, K.-S.; Bellendir, S.; Hudson, K.A.; Hill, C.B.; Hartman, G.L.; Hyten, D.L.; Hudson, M.E.; Diers, B.W. Fine mapping the soybean aphid resistance gene *Rag1* in soybean. *Theoretical and Applied Genetics* 2010, *120*, 1063-1071.

16. Bales, C.; Zhang, G.; Liu, M.; Mensah, C.; Gu, C.; Song, Q.; Hyten, D.; Cregan, P.; Wang, D. Mapping soybean aphid resistance genes in PI 567598B. *Theoretical and Applied Genetics* 2013, *126*, 2081-2091.

17. Zhang, G.; Gu, C.; Wang, D. Molecular mapping of soybean aphid resistance genes in PI 567541B. *Theoretical and Applied Genetics* 2009, *118*, 473-482.

18. Mian, M.R.; Kang, S.-T.; Beil, S.E.; Hammond, R.B. Genetic linkage mapping of the soybean aphid resistance gene in PI 243540. *Theoretical and Applied Genetics* 2008, *117*, 955-962.

19. Wang, D.; Bales, C.; Yuan, J.; Zhang, Z., Aphid resistant soybean plants. In Google Patents: 2015.

20. Zhang, G.; Gu, C.; Wang, D. A novel locus for soybean aphid resistance. *Theoretical and Applied Genetics* 2010, *120*, 1183-1191.

21. Zhang, G.; Gu, C.; Wang, D. Mapping and validation of a gene for soybean aphid resistance in PI 567537. *Molecular Breeding* 2013, *32*, 131-138.

22. Zhang, S.; Zhang, Z.; Wen, Z.; Gu, C.; An, Y.-Q.C.; Bales, C.; DiFonzo, C.; Song, Q.; Wang, D. Fine mapping of the soybean aphid-resistance genes *Rag6* and *Rag3c* from *Glycine soja* 85-32. *Theoretical and Applied Genetics* 2017, *130*, 2601-2615.

23. Jun, T.-H.; Mian, M.R.; Michel, A.P. Genetic mapping revealed two loci for soybean aphid resistance in PI 567301B. *Theoretical and Applied Genetics* 2012, *124*, 13-22.

24. Bhusal, S.J.; Jiang, G.-L.; Song, Q.; Cregan, P.B.; Wright, D.; Gonzalez-Hernandez, J.L. Genome-wide detection of genetic loci associated with soybean aphid resistance in soybean germplasm PI 603712. *Euphytica* 2017, *213*, 144.

25. McCarville, M.T.; Hodgson, E.W.; O'Neal, M.E. Soybean aphid-resistant soybean varieties for Iowa. 2012.

26. Kim, K.-S.; Hill, C.B.; Hartman, G.L.; Mian, M.; Diers, B.W. Discovery of soybean aphid biotypes. *Crop Science* 2008, *48*, 923-928.

27. Hill, C.B.; Crull, L.; Herman, T.K.; Voegtlin, D.J.; Hartman, G.L. A new soybean aphid (Hemiptera: Aphididae) biotype identified. *Journal of Economic Entomology* 2010, *103*, 509-515.

28. Alt, J.; Ryan-Mahmutagic, M. Soybean aphid biotype 4 identified. *Crop Science* 2013, *53*, 1491-1495.

29. Claridge, M.; Den Hollander, J. The biotype concept and its application to insect pests of agriculture. *Crop Protection* 1983, *2*, 85-95.

30. Robert, C.A.; Erb, M.; Hibbard, B.E.; Wade French, B.; Zwahlen, C.; Turlings, T.C. A specialist root herbivore reduces plant resistance and uses an induced plant volatile to aggregate in a density-dependent manner. *Functional Ecology* 2012, *26*, 1429-1440.

31. Baluch, S.D.; Ohm, H.W.; Shukle, J.T.; Williams, C.E. Obviation of wheat resistance to the hessian Fly through systemic induced susceptibility. *Journal of Economic Entomology* 2012, *105*, 642-650.

32. Varenhorst, A.; McCarville, M.; O'Neal, M. An Induced Susceptibility Response in Soybean Promotes Avirulent *Aphis glycines* (Hemiptera: Aphididae) Populations on Resistant Soybean. *Environmental Entomology* 2015, nvv051.

33. Griffith, M.; Walker, J.R.; Spies, N.C.; Ainscough, B.J.; Griffith, O.L. Informatics for RNA sequencing: a web resource for analysis on the cloud. *PLoS Computational Biology* 2015, *11*, e1004393.

34. Shan, X.; Li, Y.; Jiang, Y.; Jiang, Z.; Hao, W.; Yuan, Y. Transcriptome profile analysis of maize seedlings in response to high-salinity, drought and cold stresses by deep sequencing. *Plant Molecular Biology Reporter* 2013, *31*, 1485-1491.

35. Prochaska, T.J.; Donze-Reiner, T.; Marchi-Werle, L.; Palmer, N.; Hunt, T.E.; Sarath, G.; Heng-Moss, T. Transcriptional responses of tolerant and susceptible soybeans to soybean aphid (*Aphis glycines* Matsumura) herbivory. *Arthropod-Plant Interactions* 2015, *9*, 347-359.

36. Bansal, R.; Mian, M.; Mittapalli, O.; Michel, A.P. RNA-Seq reveals a xenobiotic stress response in the soybean aphid, *Aphis glycines*, when fed aphid-resistant soybean. *BMC Genomics* 2014, *15*, 972.

37. Brechenmacher, L.; Nguyen, T.H.N.; Zhang, N.; Jun, T.-H.; Xu, D.; Mian, M.R.; Stacey, G. Identification of soybean proteins and genes differentially regulated in near isogenic lines differing in resistance to aphid infestation. *Journal of Proteome Research* 2015, *14*, 4137-4146.

38. Li, Y.; Zou, J.; Li, M.; Bilgin, D.D.; Vodkin, L.O.; Hartman, G.L.; Clough, S.J. Soybean defense responses to the soybean aphid. *New Phytologist* 2008, *179*, 185-195.

39. Studham, M.E.; MacIntosh, G.C. Multiple phytohormone signals control the transcriptional response to soybean aphid infestation in susceptible and resistant soybean plants. *Molecular Plant-Microbe Interactions* 2013, *26*, 116-129.

40. De Mendiburu, F. Agricolae: statistical procedures for agricultural research. *R package version 1.2* 2014, *1*.

41. Andrews, S. FastQC: a quality control tool for high throughput sequence data. 2010.

42. Ewels, P.; Magnusson, M.; Lundin, S.; Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016, *32*, 3047-3048.

43. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014, *30*, 2114-2120.

44. Patro, R.; Duggal, G.; Love, M.I.; Irizarry, R.A.; Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* 2017, *14*, 417.

45. Grüning, B.; Dale, R.; Sjödin, A.; Chapman, B.A.; Rowe, J.; Tomkins-Tinch, C.H.; Valieris, R.; Köster, J.; Bioconda, T. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature methods* 2018, *15*, 475.

46. Langfelder, P.; Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008, *9*, 559.

47. Grant, D.; Nelson, R.T.; Cannon, S.B.; Shoemaker, R.C. SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Research* 2009, gkp798.

48. Supek, F.; Bošnjak, M.; Škunca, N.; Šmuc, T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLOS ONE* 2011, *6*, e21800.

49. Ge, S.; Jung, D. ShinyGO: a graphical enrichment tool for animals and plants. *bioRxiv* 2018, 315150.

50. Ge, S.X.; Son, E.W.; Yao, R. iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data. *BMC Bioinformatics* 2018, 19, 534.

51. Thimm, O.; Bläsing, O.; Gibon, Y.; Nagel, A.; Meyer, S.; Krüger, P.; Selbig, J.; Müller, L.A.; Rhee, S.Y.; Stitt, M. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal* 2004, *37*, 914-939.

52. Lohse, M.; Nagel, A.; Herter, T.; May, P.; Schroda, M.; Zrenner, R.; Tohge, T.; Fernie, A.R.; Stitt, M.; Usadel, B. Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant Cell Environment* 2014, *37*, 1250-1258.

53. Price, P.W.; Denno, R.F.; Eubanks, M.D.; Finke, D.L.; Kaplan, I. *Insect ecology: behavior, populations and communities*; Cambridge University Press: 2011; p.^pp.

54. Bourgis, F.; Roje, S.; Nuccio, M.L.; Fisher, D.B.; Tarczynski, M.C.; Li, C.; Herschbach, C.; Rennenberg, H.; Pimenta, M.J.; Shen, T.-L. S-methylmethionine plays a major role in phloem sulfur transport and is synthesized by a novel type of methyltransferase. *The Plant Cell* 1999, *11*, 1485-1497.

55. Wilson, A.C.; Ashton, P.D.; Calevro, F.; Charles, H.; Colella, S.; Febvay, G.; Jander, G.; Kushlan, P.F.; Macdonald, S.J.; Schwartz, J.F. Genomic insight into the amino acid relations of the pea aphid, *Acyrthosiphon pisum*, with its symbiotic bacterium *Buchnera aphidicola*. *Insect Molecular Biology* 2010, *19*, 249-258.

56. Douglas, A. Sulphate utilization in an aphid symbiosis. *Insect Biochemistry and Molecular Biology* 1988, *18*, 599-605.

57. Oliver, K.M.; Degnan, P.H.; Burke, G.R.; Moran, N.A. Facultative symbionts in aphids and the horizontal transfer of ecologically important traits. *Annual Review of Entomology* 2010, *55*, 247-266.

58. Li, Y.; Hill, C.B.; Carlson, S.R.; Diers, B.W.; Hartman, G.L. Soybean aphid resistance genes in the soybean cultivars Dowling and Jackson map to linkage group M. *Molecular Breeding* 2007, *19*, 25-34.

59. Howe, G.A. Jasmonates as signals in the wound response. *Journal of Plant Growth Regulation* 2004, *23*, 223-237.

60. Kanobe, C.; McCarville, M.T.; O'Neal, M.E.; Tylka, G.L.; MacIntosh, G.C. Soybean aphid infestation induces changes in fatty acid metabolism in soybean. *PloS one* 2015, *10*, e0145660.

61. Boller, T.; Felix, G. A renaissance of elicitors: perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors. *Annual Review of Plant Biology* 2009, *60*, 379-406.

62. Jaouannet, M.; Rodriguez, P.A.; Thorpe, P.; Lenoir, C.J.; MacLeod, R.; Escudero-Martinez, C.; Bos, J.I. Plant immunity in plant–aphid interactions. *Frontiers in Plant Science* 2014, *5*.

63. Flor, H.H. Current status of the gene-for-gene concept. *Annual Review of Phytopathology* 1971, *9*, 275-296.

64. Jones, J.D.; Dangl, J.L. The plant immune system. *Nature* 2006, *444*, 323-329.

65. van der Hoorn, R.A.; Kamoun, S. From Guard to Decoy: a new model for perception of plant pathogen effectors. *Plant Cell* 2008, *20*, 2009-2017.

66. Collier, S.M.; Moffett, P. NB-LRRs work a "bait and switch" on pathogens. *Trends in Plant Science* 2009, *14*, 521-529.

67. Wu, J.; Baldwin, I.T. New insights into plant responses to the attack from insect herbivores. *Annual Review of Genetics* 2010, *44*, 1-24.

68. Bigeard, J.; Colcombet, J.; Hirt, H. Signaling mechanisms in pattern-triggered immunity (PTI). *Molecular Plant* 2015, *8*, 521-539.

69. Kim, J.H.; Jander, G. *Myzus persicae* (green peach aphid) feeding on Arabidopsis induces the formation of a deterrent indole glucosinolate. *The Plant Journal* 2007, *49*, 1008-1019.

70. Hillwig, M.S.; Chiozza, M.; Casteel, C.L.; Lau, S.T.; Hohenstein, J.; Hernández, E.; Jander, G.; MacIntosh, G.C. Abscisic acid deficiency increases defence responses against *Myzus persicae* in *Arabidopsis*. *Molecular Plant Pathology* 2016, *17*, 225-235.

71. Fujimoto, S.; Matsunaga, S.; Yonemura, M.; Uchiyama, S.; Azuma, T.; Fukui, K. Identification of a novel plant MAR DNA binding protein localized on chromosomal surfaces. *Plant Molecular Biology* 2004, *56*, 225-239.

72. Hobert, O.; Jallal, B.; Ullrich, A. Interaction of Vav with ENX-1, a putative transcriptional regulator of homeobox gene expression. *Molecular and Cellular Biology* 1996, *16*, 3066-3073.

73. Scott, M.P.; Tamkun, J.W.; Hartzell, G.W. The structure and function of the homeodomain. *Biochimica et Biophysica Acta* 1989, *989*, 25-48.

74. Ernst, H.A.; Olsen, A.N.; Larsen, S.; Lo Leggio, L. Structure of the conserved domain of ANAC, a member of the NAC family of transcription factors. *EMBO Reports* 2004, *5*, 297-303.

75. Zheng, N.; Fraenkel, E.; Pabo, C.O.; Pavletich, N.P. Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F-DP. *Genes & Development* 1999, *13*, 666-674.

76. Thatcher, L.F.; Kazan, K.; Manners, J.M. Lateral organ boundaries domain transcription factors: new roles in plant defense. *Plant Signaling & Behavior* 2012, *7*, 1702-1704.

77. Ohme-Takagi, M.; Shinshi, H. Ethylene-inducible DNA binding proteins that interact with an ethylene-responsive element. *The Plant cell* 1995, *7*, 173-182.

78. Massari, M.E.; Murre, C. Helix-Loop-Helix Proteins: Regulators of Transcription in Eucaryotic Organisms. *Molecular and Cellular Biology* 2000, *20*, 429-440.

79. Landschulz, W.; Johnson, P.; McKnight, S. The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. *Science* 1988, *240*, 1759-1764.

80. Silva, O.d.C.e. CG-1, a parsley light-induced DNA-binding protein. *Plant Molecular Biology* 1994, *25*, 921-924.

81. Klempnauer, K.H.; Sippel, A.E. The highly conserved amino-terminal region of the protein encoded by the v-myb oncogene functions as a DNA-binding domain. *The EMBO Journal* 1987, *6*, 2719-2725.

82. Cubas, P.; Lauter, N.; Doebley, J.; Coen, E. The TCP domain: a motif found in proteins regulating plant growth and development. *The Plant Journal* 1999, *18*, 215-222.

83. Hohenstein, J.D.; Studham, M.E.; Klein, A.; Kovinich, N.; Barry, K.; Lee, Y.-J.; MacIntosh, G.C. Transcriptional and Chemical Changes in Soybean Leaves in Response to Long-Term Aphid Colonization. *Frontiers in Plant Science* 2019, *10*.

84. Sytykiewicz, H.; Chrzanowski, G.; Czerniewicz, P.; Sprawka, I.; Łukasik, I.; Goławska, S.; Sempruch, C. Expression Profiling of Selected Glutathione Transferase Genes in *Zea mays* (L.) Seedlings Infested with Cereal Aphids. *PLOS ONE* 2014, *9*, e111863.

85. Dixon, D.P.; Hussey, P.J.; Hawkins, T.; Edwards, R. Enzyme activities and subcellular localization of members of the Arabidopsis glutathione transferase superfamily. *Journal of Experimental Botany* 2009, *60*, 1207-1218.

86. Wagner, U.; Edwards, R.; Dixon, D.P.; Mauch, F. Probing the Diversity of the Arabidopsis glutathione S-Transferase Gene Family. *Plant Molecular Biology* 2002, *49*, 515-532.

87. Lee, S.; Cassone, B.J.; Wijeratne, A.; Jun, T.-H.; Michel, A.P.; Mian, M.R. Transcriptomic dynamics in soybean near-isogenic lines differing in alleles for an aphid resistance gene, following infestation by soybean aphid biotype 2. *BMC Genomics* 2017, *18*, 472.

88. Macedo, T.B.; Bastos, C.S.; Higley, L.G.; Ostlie, K.R.; Madhavan, S. *Photosynthetic Responses of Soybean to Soybean Aphid (Homoptera: Aphididae) Injury*; BIOONE: 2003; Vol. 96, p.^pp.188-193, 186.

89. Bischoff, V.; Nita, S.; Neumetzler, L.; Schindelasch, D.; Urbain, A.; Eshed, R.; Persson, S.; Delmer, D.; Scheible, W.-R. TRICHOME BIREFRINGENCE and Its Homolog AT5G01360 Encode Plant-Specific DUF231 Proteins Required for Cellulose Biosynthesis in *Arabidopsis*. *Plant Physiology* 2010, *153*, 590-602.

90. Schuetz, M.; Benske, A.; Smith, R.A.; Watanabe, Y.; Tobimatsu, Y.; Ralph, J.; Demura, T.; Ellis, B.; Samuels, A.L. Laccases direct lignification in the discrete secondary cell wall domains of protoxylem. *Plant Physiology* 2014, *166*, 798-807.

91. Tjallingii, W.F. Salivary secretions by aphids interacting with proteins of phloem wound responses. *Journal of Experimental Botany* 2006, *57*, 739-745.

92. Hu, Q.; Min, L.; Yang, X.; Jin, S.; Zhang, L.; Li, Y.; Ma, Y.; Qi, X.; Li, D.; Liu, H. Laccase GhLac1 modulates broad-spectrum biotic stress tolerance via manipulating phenylpropanoid pathway and jasmonic acid synthesis. *Plant Physiology* 2018, *176*, 1808-1823.

93. Argandona, V.; Chaman, M.; Cardemil, L.; Munoz, O.; Zuniga, G.; Corcuera, L. Ethylene production and peroxidase activity in aphid-infested barley. *Journal of Chemical Ecology* 2001, *27*, 53-68.

94. Soffan, A.; Aldawood, A.S.; Alghamdi, S.S. Peroxidase and Polyphenol Oxidase Activity in Moderate Resistant and Susceptible *Vicia faba* Induced by *Aphis craccivora* (Hemiptera: Aphididae) Infestation. *Journal of Insect Science* 2014, *14*.

95. Park, J.H.; Halitschke, R.; Kim, H.B.; Baldwin, I.T.; Feldmann, K.A.; Feyereisen, R. A knock-out mutation in allene oxide synthase results in male sterility and defective wound signal transduction in *Arabidopsis* due to a block in jasmonic acid biosynthesis. *The Plant Journal* 2002, *31*, 1-12.

96. Ralph, S.G.; Jancsik, S.; Bohlmann, J. Dirigent proteins in conifer defense II: Extended gene discovery, phylogeny, and constitutive and stress-induced gene expression in spruce (*Picea* spp.). *Phytochemistry* 2007, *68*, 1975-1991.

97. Major, I.T.; Constabel, C.P. Functional Analysis of the Kunitz Trypsin Inhibitor Family in Poplar Reveals Biochemical Diversity and Multiplicity in Defense against Herbivores. *Plant Physiology* 2008, *146*, 888-903.

98. Dangl, J.L.; Jones, J.D. Plant pathogens and integrated defence responses to infection. *Nature* 2001, *411*, 826-833.

99. Dogimont, C.; Bendahmane, A.; Pitrat, M.; Burget-Bigeard, E.; Hagen, L.; Le Menn, A.; Pauquet, J.; Rousselle, P.; Caboche, M.; Chovelon, V., Transfer the resistance-promoting Vat allele, by transgenesis, to *Aphis gossypii*-sensitive melon varieties; aphicides. In Google Patents: 2009.

100. Hill, C.B.; Li, Y.; Hartman, G.L. A single dominant gene for resistance to the soybean aphid in the soybean cultivar Dowling. *Crop Science* 2006, *46*, 1601-1605.

101. Hill, C.B.; Li, Y.; Hartman, G.L. Soybean aphid resistance in soybean Jackson is controlled by a single dominant gene. *Crop Science* 2006, *46*, 1606-1608.

102.	Van Nurden, A.; Scott, R.; Hesler, L.; Tilmon, K.; Glover, K.; Carter, C. Inheritance of soybean aphid resistance from PI 71506. *Journal of Crop Improvement* 2010, *24*, 400-416.

CHAPTER 6: TRANSCRIPTOME PROFILING OF INTERACTION EFFECTS OF

SOYBEAN CYST NEMATODES AND SOYBEAN APHIDS ON SOYBEAN

The data portion of this chapter is under review for publication in Nature's

Scientific Data Journal.

Abstract

Soybean aphid (*Aphis glycines*; SBA) and soybean cyst nematode (*Heterodera glycines*; SCN) are two economically important pests of soybean (*Glycine max*) in the United States. The main objective of this research was to use demographic and genetic data sets to characterize three-way interactions among soybean, soybean-aphid and soybean-cyst-nematode. The interactions were evaluated using SCN-resistant and SCN-susceptible soybean cultivars with three treatments/cultivar: SBA (biotype 1), SCN (HG type 0), or SCN: SBA in randomized complete block design in six blocks. The experiment was conducted in a greenhouse water bath using cone-tainers. Treatments receiving SCN were infested at planting with 2000 nematode eggs. Treatments with soybean-aphids were infested at second trifoliate growth stage (V2) with 15 biotype 1 SBA. SBA populations were counted at 5, 15, and 30 days post infestation (dpi). SCN eggs were sampled at 30 dpi. The number of SCN eggs was significantly greater on the susceptible cultivar and no effect in resistant cultivar in the presence of SBA. The SBA population density was negatively affected by SCN populations. RNA-seq analysis revealed 4, 637 DEGs at 5 dpi and 19,032 DEGs at 30 dpi samples treated with SCN, SBA and both when compared between resistant and susceptible cultivars. Further

analysis was narrowed to DEGs that are common in all treatments and discordant in resistant cultivar focused on treatment with SBA and SCN. WGCNA analysis revealed seven and nine modules at 5 dpi and 30 dpi, respectively. PGSEA analysis revealed several pathways enriched: 'Plant pathogen interaction' and 'cutine, suberine, and wax biosynthesis' pathways at 5 dpi and 'isoflavonoid biosynthesis' and 'one carbon pool by folate' pathways enriched at 30 dpi. In addition, enriched transcription factor (TF) binding motifs were identified in up-regulated and down-regulated DEGs in different comparisons. The identified DEGs in this experiment, particularly in resistant cultivar during SBA and SCN interactions are potential candidates for dual and durable pest resistance warranting further validation.

6.1. Introduction

Soybean [*Glycine max* (L.) Merr.], considered as the source of high-quality sugar, protein, and oil, is one of the most important crops worldwide [1]. Soybean aphid, *Aphis glycines* Matsumura (Hemiptera: Aphididae) and soybean cyst nematode (SCN), *Heterodera glycines* Ichinohe are the two most economically important pests of soybean in the Midwestern United States [2, 3]. Soybean aphid, an aboveground herbivore (pest), feeds on phloem sap whereas SCN, a belowground pest, infests the soybean roots. These infestations can co-occur and amplify further reduction in soybean yield [4, 5]. In the United States, annual economic losses due to the SBA and SCN have been estimated to be approximately $4 billion and $1.3 billion, respectively [6, 7, 8]. To counteract these devastating pests, farmers rely on various management strategies that include host plant resistance and chemical measures [9, 10, 11]. For SBA, dependency on the use of

chemical management has resulted in pyrethroid resistance in SBA populations in Iowa, Minnesota, North Dakota and South Dakota as well as the impacts on non-target beneficial organisms[12, 13]. In addition, the long-term use of SCN resistance has resulted in SCN populations that are capable of overcoming the resistance genes (i.e., HG types)[14]. Although host plant resistance has not been implemented on a large scale for SBA management, multiple virulent SBA biotypes have been discovered in the U.S. Virulent SBA biotypes and SCN races threaten the sustainability of host plant resistance for these two pests[14, 15, 16, 17]. Thus, genetic data generated from greenhouse experiments on the effects of SBA and SCN on soybean cultivars are of tremendous importance for unraveling resistance genes and regulatory networks that can potentially be used for developing durable resistance in soybean to both pests.

Although above and belowground herbivores are spatially segregated, they both share the host plant through systemic tissues and are able to influence each other[18]. Previously, the influence of SCN on soybean aphid infestation or *vice versa* has been studied on soybean using demographic datasets [4, 5, 19, 20, 21]. McCarville et al. 2012 [4] conducted experiments on various soybean cultivars [SCN susceptible (DK 28-52, IA 3018, IA 3041) and SCN resistant (DK 27-52, AG 2821 V, IA 3028)] to understand the effect of SBA, SCN, and fungus *Cadophora gregata* on soybean[16]. Their study showed 5.24 times increase in SCN reproduction in the presence of soybean aphid and the fungus. In contrast, the aphid population decreased by 26.4% in the presence of SCN and *C. gregata* and the aphid exposure reduced by 19.8% in SCN resistant cultivars. Later, McCarville et al. [5] demonstrated the relationship between the aboveground feeding of soybean aphid and belowground reproduction of SCN in the SCN resistant Dekalb 27-52

(PI 88788 derived) cultivar, and SCN susceptible Kenwood 94 cultivar. In 30 days, both

SCN eggs and the number of females increased by 33% in SCN-resistant cultivar and

reduced by 50% in the SCN-susceptible cultivar. In 60 days, the number of SCN eggs

and female count remained unaffected in the resistant cultivar but decreased in the

susceptible cultivar. The authors concluded that soybean aphid feeding improved the

quality of soybean as a host for SCN, but this result was varied significantly with the

cultivar and length of the experiment. Apart from these demographic studies, molecular

characterization of SBA-SCN-soybean interaction has not been reported previously.

RNA-Sequencing (RNA-Seq) has been a standard tool for studying qualitative

and quantitative gene expression assays that provide information on transcript abundance

with their variation [22, 23]. The major objective of this study was to evaluate differential

gene expression of soybean plants that are infested with SCN in the presence or absence

of SBA. To achieve the objective, we conducted experiments on two genotypes of

G. max [*H. glycines* susceptible Williams 82 (PI518671), and *H. glycines* resistant

MN1806CN] that were infested with biotype 1 SBA and HG Type 0 SCN for RNA-

sequencing. More than 1.1 billion reads (61.4 GB) of transcriptomic data were obtained

from 47 samples derived from the experiment using whole roots of *G. max*. An overview

of the experimental design and transcriptome analysis pipeline is shown in Figure 6.1. A

comprehensive understanding of these transcriptome data will enhance our understanding

of interactions among soybean, SBA, and SCN at the molecular level. The rapid

advancement of bioinformatics tools is facilitating the search of candidate genes and their

function that might play a crucial role in various pathways for host resistance against both herbivores.

6.2. Materials and Methods

6.2.1. Plant Material, Aphid, and SCN

Two cultivars of soybean were used – Williams 82 and MN1806CN. Williams 82 is susceptible to both HG Type 0 (race 3) of the SCN and SBA. MN1806CN is resistant to HG Type 0 (race 3) of the SCN. Soybean aphid biotype 1 populations were obtained from Ohio State University and were reared on susceptible cultivar LD12-15838R. This biotype is defined by an avirulent response to all known SBA resistance (*Rag*) genes and was first identified in Illinois [24]. The SCN population used was HG type 0, which is defined by having less than 10% reproduction documented by studies of SCN resistance and is avirulent to all SCN resistance genes in soybean.

6.2.2. Experimental Design and Sample Collection

A greenhouse experiment was designed using a randomized complete block design (RCBD) with eight treatments (four treatments per cultivars) with eight experimental units (plants) in six blocks. The treatments were factors of soybean genotype, SBA infestation, and SCN infestation. For examples, each of the soybean genotypes received one of the following combinations: SCN:no SBA, no SCN:SBA, SCN:SBA, or no SBA:no SCN (control). For this experiment, the soil-sand mixture was prepared by adding construction sand and clay soil including SCN (HG type 0) infested clay soil in the ratio of 3:1. The 125 cc of the mixture was distributed in cone-tainers (diameter of 3.8 cm, a depth of 21 cm and a volume of 164 cc; Greenhouse Megastore, USA). For SCN included treatments, each cone-tainer received approximately 2,000 SCN

eggs. The cone-tainers with three soybean seeds were arranged in a 2.0 U.S. gallon (7.57 liter) plastic buckets (Leaktite, USA) filled with construction sand (Quikrete, GA). These buckets were kept in a water bath for maintaining soil temperature between 26.7 ºC and 28.9 ºC to ensure the reproduction of SCN (i.e.~ 30 days) [5]. The plants were grown under 16 hour cycle of light and dark. The plants were thinned down to one plant per cone-tainer upon reaching the second vegetative growth stage (V2). The V2-staged plants with the SBA included treatments were infested with 15 mixed age biotype 1 SBA using a 000 fine tip paintbrush (Winsor & Newton, England). The SBA were applied on the abaxial surface of the first trifoliate of V2-staged plants. All plants in each bucket were covered with a large no-see-um mesh net (Quest Outfitters, Sarasota, FL) to prevent inter-bucket movement of aphids. After SBA infestation, soybean plants were regularly checked to confirm the successful establishment of soybean aphids. Soybean aphid populations were counted at 5, 15, and 30 days post-infestation (dpi). SCN eggs were sampled at 30 dpi. The whole roots were collected on 5 and 30 dpi by snap freezing in liquid Nitrogen and stored at -80 ºC for further analysis. The 5 dpi and 30 dpi root samples treated with each treatment were collected from Water bath I and Water bath II, respectively, representing each plant from three blocks (three biological replicates). The SCN soil and SCN infested roots were used for SCN cysts collection (except root samples collected for transcriptomic study) and the soil was examined for SCN counts. An overview of the experimental design and transcriptome analysis pipeline is shown in Figure 6.1.
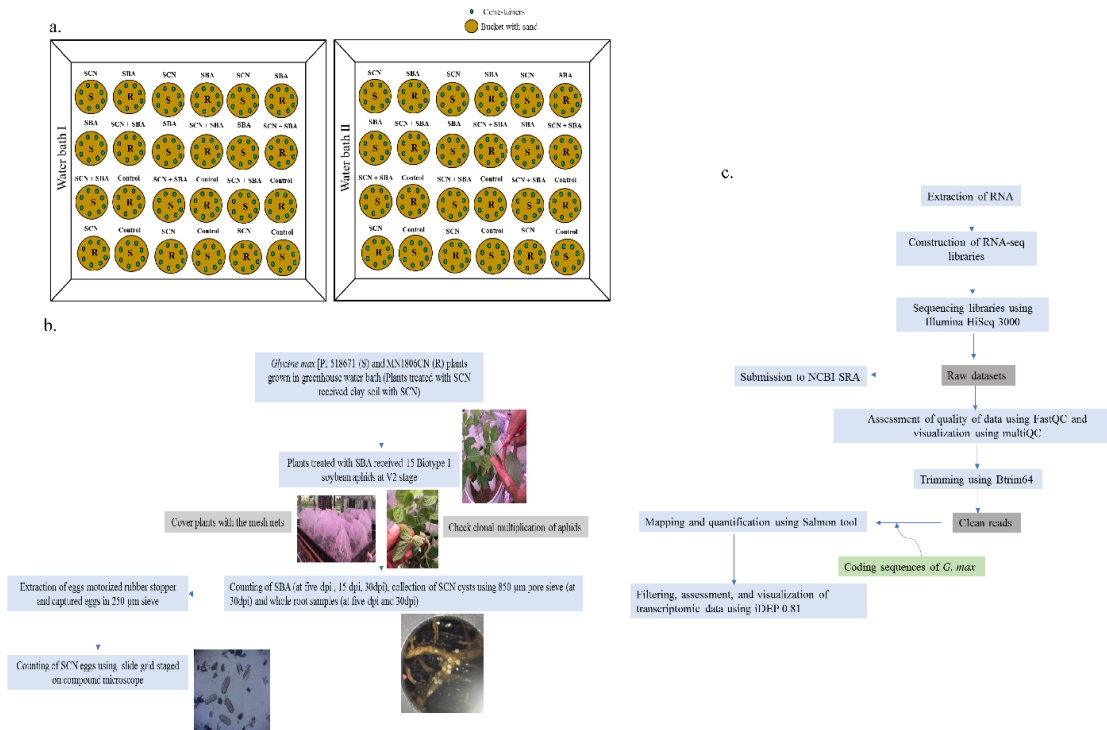
Figure 6.1. An overview of greenhouse experiments and transcriptomic data analysis pipeline. (a) A randomized complete block design (RCBD) using two water baths (Water bath I and Water bath II), (b) A flow chart representing experimental methods used for soybean-cyst-nematode and soybean-aphid interaction using two cultivars of soybean, and (c) A flow chart showing RNA-seq data analysis pipeline.

6.2.3. RNA Extraction, Library Construction, and RNA-sequencing

RNA was extracted from all samples representing three biological replicates of each treatment that constituted 24 samples collected at 5 and 30 dpi each. Frozen root samples from each treatment were grounded in liquid nitrogen with a mortar and pestle to a fine powder followed by total RNA extraction using PureLink RNA mini kit (Invitrogen, USA). RNA samples were treated with TURBO™ DNase (Invitrogen, USA) to remove any DNA contamination following the manufacturer's instructions. Assessment of the isolated RNA integrity was performed by 1% agarose gel

electrophoresis, and RNA concentration was measured by Nanodrop 2000 (Thermo Fisher Scientific, USA). The cDNA libraries were constructed using NEBNext Ultra II RNA library 96 single index kit prep kit and sequenced using Illumina HiSeq 3000 (single read end utilizing a 100-bp read length) at Iowa State University Sequencing Facilities.

6.2.4. Aphid and SCN Egg Counts

Soybean aphid populations were counted at 5, 15 and 30 dpi. SCN eggs were sampled at 30 days. The SCN soil and roots were washed with water in the bucket and mixed properly using a hand. After, mixing the solution was passed through the 850 µm pore sieve and captured in 250 µm pore sieve. The females and cysts were then ground on 250 µm pore sieve using a motorized rubber stopper and eggs were released and recovered in 25 µm pore sieve passing through 75 µm pore sieve. Eggs were suspended in 50ml of water and the number of *H. glycines* were counted under the compound microscope using 1ml as the representative sample of the solution.

6.2.5. SCN and SBA Count Data Analysis

The SBA and SCN counts data were analyzed using GraphPad Prism 8.0.2. The 30 dpi SCN counts and SBA counts collected at 5, 15, and 30 dpi were analyzed separately. One-way ANOVA was employed for 30 dpi SCN counts and statistical significance between the treatments was calculated using Tukey's multiple comparisons. The 5, 15, and 30 dpi SBA counts were analyzed using two- way ANOVA with Geisser-Greenhouse correction and statistical significance between the treatments were calculated

using Tukey's multiple comparisons. The linear regression between SBA and SCN was based on counts obtained at 30 dpi.

6.2.6. Pre-processing of Sequencing Data

Quality control of reads was assessed using FastQC program (version 0.11.3) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) [25]. The FastQC results were visualized using MultiQC v1.3[26], and low quality bases (QC value > 20; 5-bp window size) were removed by trimming in the program Btrim64 (version 0.2.0) [27]. High-quality single-end reads were mapped against the primary coding sequences of *G. max*. The coding sequences

(*Gmax: Gmax_275_Wm82.a2.v1.transcript_primaryTranscriptOnly.fa.gz*) were obtained from the Phytozome database and aligned using Salmon ver.0.9.1[28] accessed from Bioconda [29]. Downstream analyses of the quantified transcript reads were performed using integrated Differential Expression and Pathway analysis (iDEP 0.81, R/Bioconductor packages) [30]. The missing data of one of the replicates of control at 30d in the resistant cultivar, MN1806CN were imputed averaging the counts from the other two replicates of cultivar at the same time point. The downstream analyses for obtained transcript estimated quantification reads were performed using integrated Differential Expression and Pathway analysis (iDEP 0.81, R/Bioconductor packages) [30]. The read quants were filtered with 0.5 counts per million (CPM) in at least one sample. The quantified raw reads were transformed using regularized log (rlog) which is implemented in the DESeq2 package. The project was deposited into the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) accession (SRR8427366-SRR8427408) under Bioproject PRJNA514200 (Table 6.1). The raw

transcript abundance counts for all the samples was deposited at the Gene Expression

Omnibus (GEO) database, GSE125103 (Supplementary File 2). The transformed

transcript abundance counts, hierarchical clustering, correlation matrices, and clusters are

represented by Supplementary Files 3, 4, 5, and 6, respectively.

Table 1: Statistics of the transcriptomic data using RNA-seq pipeline used in this study

Sample	Number of raw reads	GC %	Read Length	Trimmed reads	Percentage of clean reads	Mapped Reads	Percentage of mapped reads	Number of Uniquely mapped reads	Percent uniquely mapped	Accession
PI518671_treatment_SCN_30d_R1	29,875,777	44	100	29,868,305	99.97%	26,306,640	88.1	24,916,413	83.4	SRR8427366
PI518671_treatment_SCN_30d_R2	20,569,129	45	100	20,564,513	99.98%	18,327,957	89.1	17,356,148	84.4	SRR8427367
PI518671_treatment_SCN_30d_R3	23,663,582	44	100	23,657,909	99.98%	20,899,976	88.3	19,646,683	83.0	SRR8427368
PI518671_treatment_Aphid_30d_R1	24,553,476	45	100	24,546,368	99.97%	21,032,002	85.7	19,429,157	79.2	SRR8427369
PI518671_treatment_Aphid_30d_R2	25,372,180	45	100	25,364,647	99.97%	22,011,320	86.8	19,706,012	77.7	SRR8427362
PI518671_treatment_Aphid_30d_R3	37,691,731	44	100	37,682,590	99.98%	31,646,750	84.0	29,865,320	79.3	SRR8427363
PI518671_treatment_SCNAphid_30d_R1	23,727,017	45	100	23,721,761	99.98%	21,457,335	90.5	20,276,187	85.5	SRR8427364
PI518671_treatment_SCNAphid_30d_R2	22,378,982	44	100	22,373,777	99.98%	19,622,486	87.7	18,602,604	83.1	SRR8427365
PI518671_treatment_SCNAphid_30d_R3	27,673,846	44	100	27,668,291	99.98%	23,304,305	84.2	22,080,120	79.8	SRR8427370
MN1806CN_treatment_SCN_30d_R1	25,200,882	43	100	25,192,664	99.97%	18,589,872	73.8	17,402,401	69.1	SRR8427371
MN1806CN_treatment_SCN_30d_R2	22,192,100	43	100	22,186,459	99.97%	18,350,922	82.7	17,417,979	78.5	SRR8427383
MN1806CN_treatment_SCN_30d_R3	20,653,286	43	100	20,648,111	99.97%	15,975,636	77.4	15,083,771	73.1	SRR8427384
MN1806CN_treatment_Aphid_30d_R1	20,903,446	44	100	20,896,290	99.97%	17,025,027	81.5	15,982,207	76.5	SRR8427385
MN1806CN_treatment_Aphid_30d_R2	21,708,115	44	100	21,701,712	99.97%	16,458,081	75.8	15,472,937	71.3	SRR8427386
MN1806CN_treatment_Aphid_30d_R3	26,617,069	44	100	26,610,582	99.98%	22,222,510	83.5	21,021,087	79.0	SRR8427387
MN1806CN_treatment_SCNAphid_30d_R1	19,498,275	43	100	19,491,491	99.97%	15,139,964	77.7	14,203,387	72.9	SRR8427388
MN1806CN_treatment_SCNAphid_30d_R2	27,765,044	44	100	27,759,095	99.98%	22,021,174	79.3	20,747,251	74.7	SRR8427389
MN1806CN_treatment_SCNAphid_30d_R3	43,325,617	44	100	43,312,161	99.97%	33,076,203	76.4	29,935,328	69.1	SRR8427390
MN1806CN_treatment_control_30d_R1	24,104,763	45	100	24,099,789	99.98%	18,112,259	75.2	17,132,109	71.1	SRR8427391
MN1806CN_treatment_control_30d_R2	32,183,362	44	100	32,174,938	99.97%	26,274,456	81.7	24,162,028	75.1	SRR8427392
PI518671_treatment_control_30d_R1	20,522,473	44	100	20,518,044	99.98%	17,937,163	87.4	17,022,590	83.0	SRR8427405
PI518671_treatment_control_30d_R2	28,600,503	44	100	28,593,731	99.98%	25,409,842	88.9	24,045,140	84.1	SRR8427404
PI518671_treatment_control_30d_R3	20,577,190	44	100	20,570,977	99.97%	17,574,516	85.4	16,585,012	80.6	SRR8427407
PI518671_treatment_SCN_5d_R1	20,389,378	44	100	20,383,629	99.97%	17,826,706	87.5	16,736,123	82.1	SRR8427406
PI518671_treatment_SCN_5d_R2	10,518,888	44	100	10,516,365	99.98%	9,444,170	89.8	8,950,048	85.1	SRR8427401
PI518671_treatment_SCN_5d_R3	21,303,947	44	100	21,298,111	99.97%	18,909,955	88.8	17,897,118	84.0	SRR8427400
PI518671_treatment_Aphid_5d_R1	20,262,293	45	100	20,256,610	99.97%	18,157,064	89.6	16,851,551	83.2	SRR8427403
PI518671_treatment_Aphid_5d_R2	51,680,716	44	100	51,666,055	99.97%	45,293,720	87.7	42,794,964	82.8	SRR8427402
PI518671_treatment_Aphid_5d_R3	20,328,355	44	100	20,322,387	99.97%	18,171,819	89.4	17,083,986	84.1	SRR8427399
PI518671_treatment_SCNAphid_5d_R1	21,569,888	44	100	21,563,432	99.97%	18,502,664	85.8	17,044,428	79.0	SRR8427398
PI518671_treatment_SCNAphid_5d_R2	57,520,568	44	100	57,503,170	99.97%	47,902,174	83.3	45,268,224	78.7	SRR8427381
PI518671_treatment_SCNAphid_5d_R3	16,889,301	45	100	16,883,954	99.97%	14,700,125	87.1	13,744,624	81.4	SRR8427382
MN1806CN_treatment_SCN_5d_R1	25,443,012	44	100	25,435,147	99.97%	21,929,527	86.2	20,483,059	80.5	SRR8427379
MN1806CN_treatment_SCN_5d_R2	20,043,049	45	100	20,037,212	99.97%	17,551,266	87.6	16,336,263	81.5	SRR8427380
MN1806CN_treatment_SCN_5d_R3	9,847,269	45	100	9,844,767	99.97%	8,472,717	86.1	7,992,925	81.2	SRR8427377
MN1806CN_treatment_Aphid_5d_R1	20,503,738	45	100	20,497,489	99.97%	16,815,160	82.0	15,666,380	76.4	SRR8427378
MN1806CN_treatment_Aphid_5d_R2	14,359,303	45	100	14,355,678	99.97%	12,268,563	85.5	11,559,112	80.5	SRR8427375
MN1806CN_treatment_Aphid_5d_R3	19,094,540	45	100	19,088,178	99.97%	16,590,158	86.9	15,245,807	79.9	SRR8427376
MN1806CN_treatment_SCNAphid_5d_R1	20,636,498	44	100	20,630,026	99.97%	16,806,607	81.5	15,865,622	76.9	SRR8427373
MN1806CN_treatment_SCNAphid_5d_R2	22,488,050	44	100	22,482,625	99.98%	19,286,899	85.8	18,060,389	80.3	SRR8427374
MN1806CN_treatment_SCNAphid_5d_R3	22,033,213	45	100	22,028,303	99.98%	16,862,396	76.5	15,964,103	72.5	SRR8427408
MN1806CN_treatment_control_5d_R1	18,937,367	46	100	18,932,017	99.97%	14,805,819	78.2	12,707,453	67.1	SRR8427396
MN1806CN_treatment_control_5d_R2	26,710,585	43	100	26,702,238	99.97%	20,226,195	75.7	18,092,239	67.8	SRR8427394
MN1806CN_treatment_control_5d_R3	21,327,385	46	100	21,320,799	99.97%	16,776,843	78.7	14,820,338	69.5	SRR8427372
PI518671_treatment_control_5d_R1	17,242,793	45	100	17,239,066	99.98%	16,044,618	93.1	14,976,834	86.9	SRR8427397
PI518671_treatment_control_5d_R2	22,062,929	46	100	22,055,685	99.97%	20,094,996	91.1	17,347,038	78.7	SRR8427395
PI518671_treatment_control_5d_R3	21,220,300	44	100	21,213,623	99.97%	19,994,447	94.3	18,592,042	87.6	SRR8427393

6.2.7. Analysis of RNA-seq Data

To reduce the mean dependent variance, the quantified reads were transformed as shown in Figure 6.3b-d. The transformed data were subjected to exploratory data analysis such as hierarchical clustering (Figure 6.4a), principal component analysis (PCA) (Figure 6.4b), and visualized using t-SNE map (Figure 6.4c) [31] to assess the global transcriptomic data. To reduce the complexity in RNA seq analysis we divided the counts data sets to two different subsets of samples belonging to different time point. Gene co-expression networks were constructed for divided datasets with the weighted gene co-expression network analysis (WGCNA) package [32] using following parameters: most variable genes to include- 2000 genes, soft threshold- 5, minimum module size- 20. DESeq2 package [33] was used to identify differentially expressed genes (DEGs) with more than a 2-fold change and with a false discovery rate (FDR) ≤ 0.01. We tested the effects of gene expression using different factors such as cultivar and treatment using the model (Expression ~ cultivar + treatment + cultivar: treatment) in different comparisons. The term cultivar: treatment refers to the interaction between cultivar and treatment. The annotations of the DEGs were obtained from Soybase [34] (www.soybase.org). The enriched up regulated and down regulated transcription factor binding motifs in promoters in different comparisons were identified using 300bp upstream of DEGs using ShinyGO [35] and iDEP 0.81. To understand the molecular pathways enriched GO Biological processes, GO Cellular, GO molecular function, and KEGG pathway for DEGs were identified using a graphical enrichment tool REVIGO [36], ShinyGO [35] and iDEP 0.81. The biological relevance of DEGs were visualized using MapMan [37].

The total transcripts of soybean were first converted to bins using the Mercator tool [38]

and uploaded to MapMan to assign bins to each differentially expressed transcript.

Parametric Analysis of Gene Set Enrichment (PGSEA) method (with all samples) [39]

with pathway significant cutoff (FDR) of 0.2 using fold change values of DEGs applying

in divided datasets. The codes used for RNA-seq data processing in the current study are

available as Appendix II; Supplementary 1.

6.3. Results
6.3.1. Greenhouse Experiment

The SCN egg counts assessed from the resistant and susceptible cultivars were

analyzed using one way ANOVA ($F = 87.44$, df $= 3$, $P < 0.001$). The SCN eggs number

in treatments with SCN only, and with Aphid and SCN together did not show a

significant difference between treatments in the resistant plants, whereas the SCN eggs

number was significantly higher in the susceptible plants at 30 d after aphid infestation

(Figure 6.2a.). To observe the relationship between the aphid counts and SCN counts,

linear regression analysis was performed. In resistant and susceptible cultivars, aphid and

SCN counts showed positive and negative relationships, respectively. However, these

relationships were insignificant [resistant ($F = 0.7$, $P = 0.43$, $R^2 = 0.1$), susceptible ($F =$

0.2, $P = 0.65$, $R^2 = 0.03$)]. However, the result showed a significant negative relationship

between population density of SCN and aphids (F=143.5, $P < 0.0001$, $R^2 = 0.91$) when

SCN egg and aphid counts considered in both cultivars (Figure 6.2b). The aphid counts

for different treatments varied across the trials. The 30 d aphid counts showed a

significant difference between all types of treatments in which 25% decrease in aphid

counts receiving SCN in resistant plants and 90% decrease in susceptible plants (Figure

6.2c). At 5 dpi and 15 dpi, aphid counts did not show a significant difference between all types of treatments.
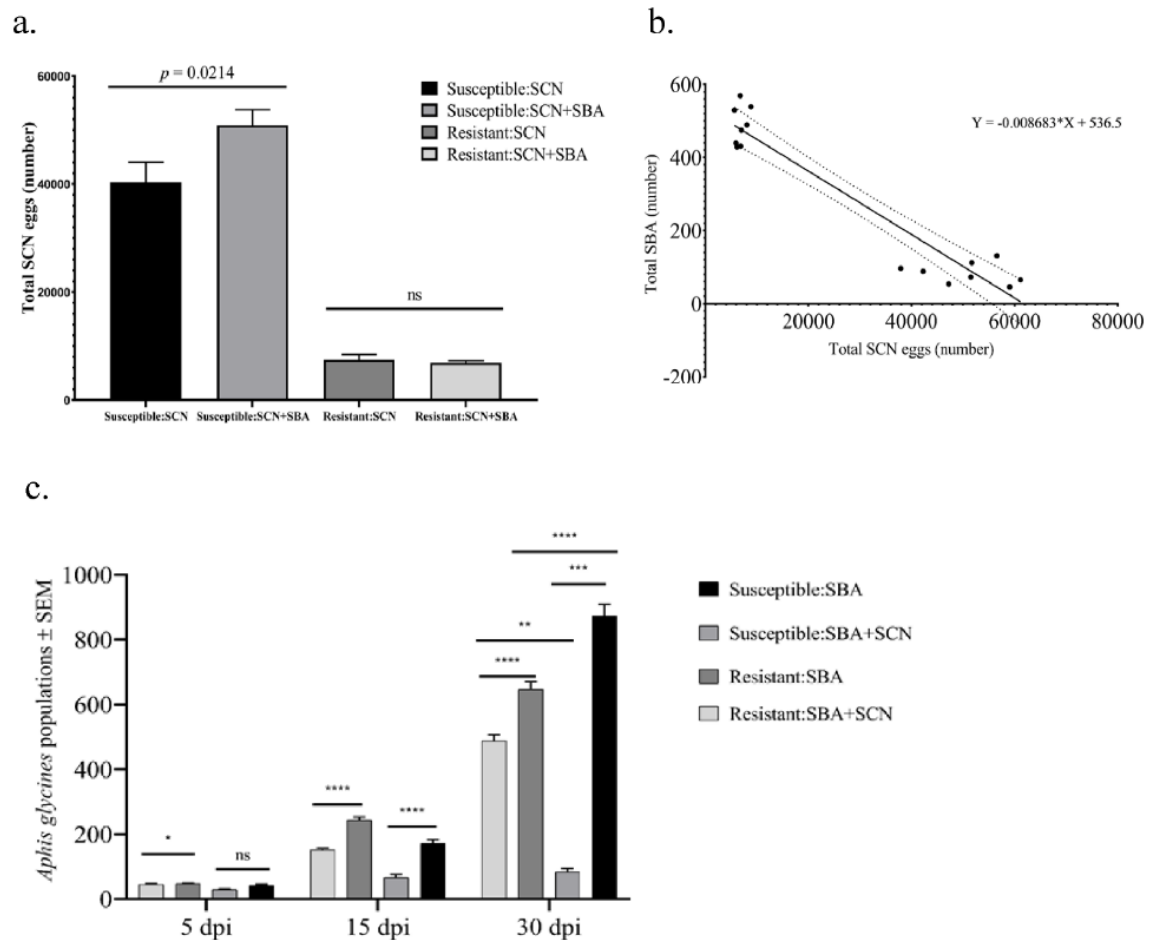
Figure 6.2. SBA and SCN counts analysis. a. The 30 dpi count of SCN eggs after infected with approximately 2000 initial SCN eggs populations on both susceptible (Williams 82 PI 518671) and resistant (MN1806CN) soybean. Error bar represents standard error mean. b. Relationship between total SCN eggs and total aphid number on 30 dpi sampling after aphid infestation. c. A number of avirulent soybean aphid (B1) populations after infested with 15 initial populations on both susceptible (Williams 82 PI 518671) and resistant (MN1806CN) soybean. Error bar represents standard error mean. [ns = P > 0.05, *=P ≤ 0.05, **= P ≤ 0.01, ***= P ≤ 0.001, **** = P ≤ 0.0001 (For the last two choices only)]

6.3.2. Transcriptomic Analysis and Assessment of Transcriptomic Data

A total of 48 RNA libraries were prepared and sequenced with the sequencing depth ranging from 9,847,269 to 57,520,568 except for the control sample in the resistant cultivar, MN1806CN collected at 30d. Total reads of more than 1.1 billion were subjected to FastQC analysis to determine the data quality using various quality metrics such as mean quality scores, per sequence quality scores, per sequence GC content, and sequence length distribution (Figure 6.3, Table 6.1). The Phred quality scores per base for all the samples were higher than 30. The GC content ranged from 43 to 45% and followed the normal distribution. After trimming, more than 99% of the reads were retained as the clean and good quality reads. Upon mapping these reads, we obtained high mapping rate ranging from 73.8% to 94.3%. Among the mapped reads, 67.1% to 87.6% reads were uniquely mapped. The 43,122 genes passed the filter upon filtering with 0.5 CPM in at least one sample. To reduce the mean dependent variance, the quantified transcript reads were transformed as shown in Figure 6.4a-c. The transformed data were subjected to hierarchical clustering and principal component analysis (PCA) followed by visualization using t-SNE map [31] in order to assess the global transcriptomic data. The hierarchical clustering of top 6000 variable genes based on two time points (5 dpi and 30 dpi) showed distinct clustering except for some samples (Figure 6.5a; Supplementary File 4). Figure 6.5b represents the standard deviation (SD) distribution of the top variable 6,000 genes. Figure 6.5c represents the correlation between the samples using the top 75% genes. The t-SNE map revealed four clusters (A, B, C, and D) for 6,000 variable genes (Figure 6.4d; Supplementary File 6). Regarding the PCA, PC1 is correlated with

time (P =1.16e-06) with 28% variance, and PC2 is correlated with Treatment (P =2.02e-08) with 15% variance (Figure 6.4e).
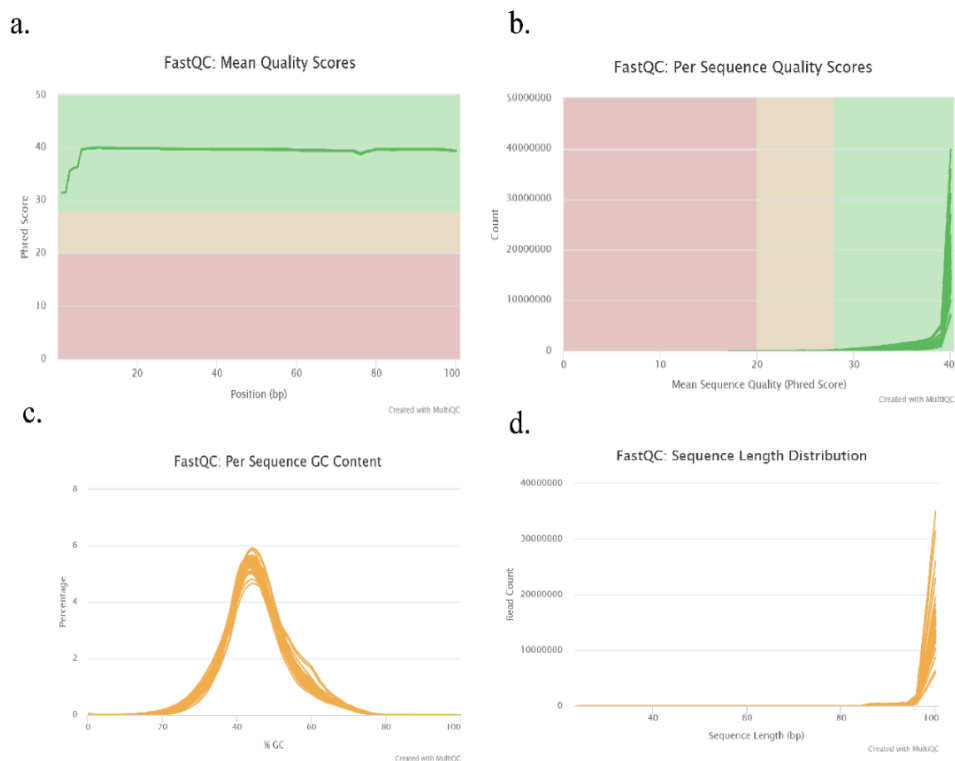
Figure 6.3. Quality metrics of *G. max* sequencing data. (a) Mean quality scores per position. (b) Per sequence quality scores. (c) GC content distribution. (d) Read length distribution
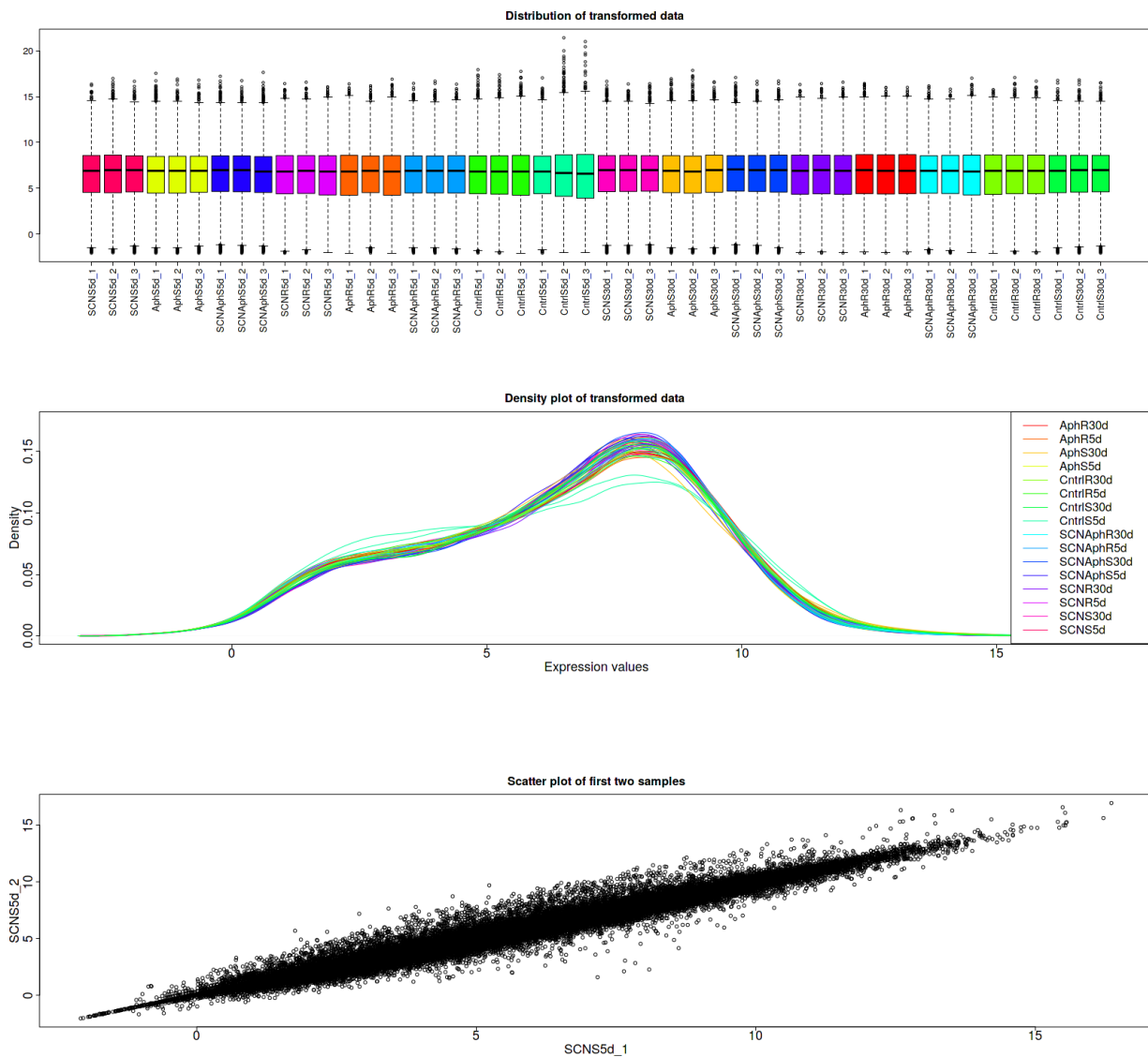
Figure 6.4. Pre-processing of transcriptomic data. (a) Distribution of transformed data. (b) Density plot of transformed data. (c) Scatter plot of the first two samples (SCNS5d_1 vs SCNS5d_2).

Figure 6.5. Assessment of transcriptomic data. (a) Heatmap of top 6,000 variable genes, (b) Gene SD distribution, (c) Correlation matrix, (d) Visualization of top 6,000 genes shown in the t-SNE map, and (e) A PCA plot.
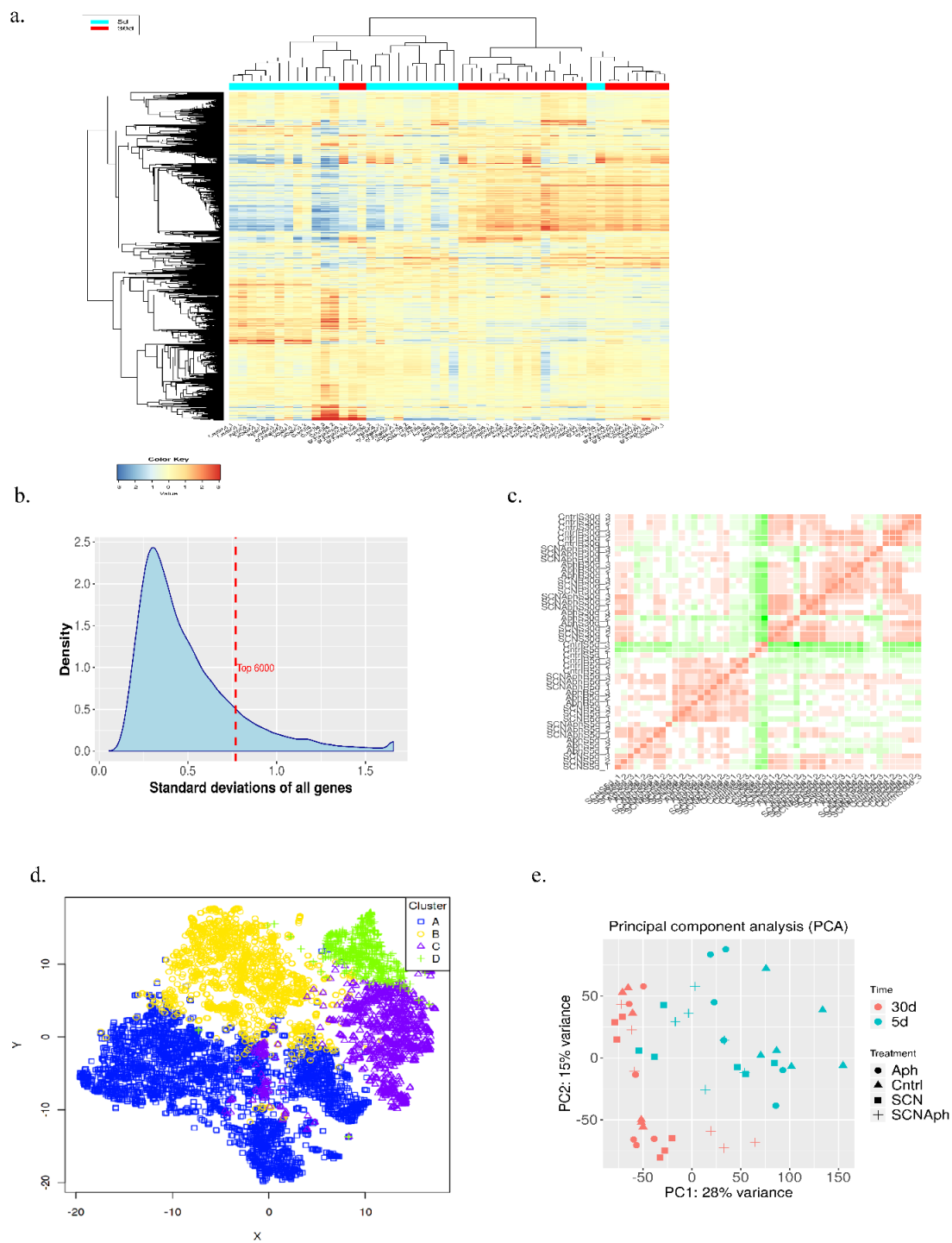
6.3.3. WGCNA Analysis Revealed Oxidative Stress at 30 dpi

The co-expression networks were used to detect correlated networks of genes and

their enrichment in the divided datasets to compare difference on the 5 dpi and 30 dpi

treatments. Weighted gene co-expression network analysis identified a network of 2,000

genes divided into seven co-expression modules 5 dpi samples, and a network of 1,994

genes divided into nine co-expression modules in 30 dpi samples (Supplementary File 7).

GO (Gene Ontology) Biological Process enrichment analysis found several highly

enriched pathways for both 5 dpi and 30 dpi samples, including nodulation, defense

response, cell wall organization, oxidation reduction process, interspecies interaction

between organisms. The only GO pathways enriched in 30 dpi samples, but not in 5 dpi

samples, were hydrogen peroxide metabolic process and reactive oxygen species

metabolic process (Table S2).
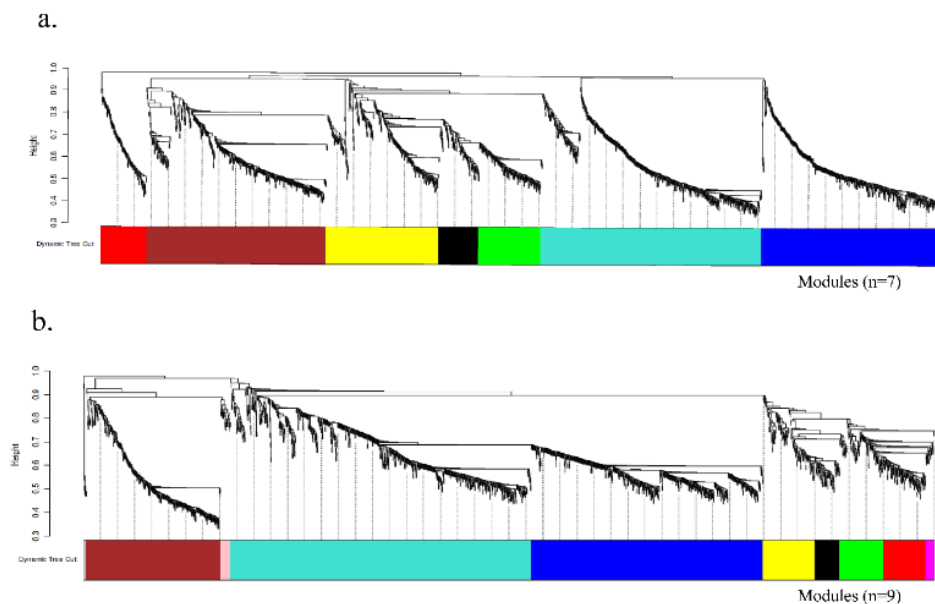
Figure 6.6. Weighted gene co-expression network analysis identified a network of 2,000 genes divided into seven co-expression modules in (a) 5 dpi samples, and a network of 1,994 genes divided into nine co-expression modules in (b) 30 dpi samples.

6.3.4. Comparison of the DEGs between Susceptible and Resistant Cultivars

The pair wise comparisons between treatments in two different cultivars with FDR < 0.01 and fold-change > 2 as cutoffs resulted in a total of 4, 637 DEGs in 5 dpi and 19,032 DEGs in 30 dpi samples treatment with SCN, SBA and both SBA and SCN (Supplementary Files 8 and 9). The MA plots were used to visualize up regulated and down regulated DEGs for each comparison as shown in Figure 6.7. We further investigated these genes using Venn diagrams (Figure 6.7). Among these comparisons 242 and 1535 DEGs overlapped in all treatments 5dpi and 30 dpi samples, respectively. These overlapped genes in all treatments are important for understanding the role of these genes in a common pathway for the interactions of these pests. The expression pattern of these genes visualized using heatmap and their biological functions visualized using GO annotations and KEGG pathway are shown in Figure 6.8 and 6.9. At 5 dpi, 242 genes were enriched for GO molecular functions of transferase activity, transferring acyl groups (GO: 0016746), ADP binding (GO: 0043531), and adenyl ribonucleotide binding (GO: 0032559). These genes are enriched for various KEGG pathways of circadian rhythm (Enrichment FDR= 0.028680109; Glyma.08G110900, Glyma.08G109200, Glyma.08G109400), Flavonoid biosynthesis (Enrichment FDR = 0.028680109; Glyma.08G110900, Glyma.08G109200, Glyma.08G109400), Isoquinoline alkaloid biosynthesis (Enrichment FDR = 0.035955005; Glyma.18G143600, Glyma.15G071200). The enriched genes in the flavonoid biosynthesis pathway are represented by Figure 6.8d. Overrepresented TF binding motifs in the promoters of these 242 genes revealed

Homeodomain, Myb/SANT, and CG-1 as the enriched transcription family

(Supplementary File 10).

At 30 dpi, 1535 DEGs were enriched for GO biological processes of oxidation-

reduction process (GO: 0055114; 15.06%), carbohydrate metabolic process (GO:

0005975; 5.26%), lipid metabolic process (GO: 0006629; 3.52%), extracellular

polysaccharide biosynthetic process (GO: 0045226; 0.06%). These genes were enriched

for various KEGG pathways of Amino sugar and nucleotide sugar metabolism

(Enrichment FDR = 1.10E-05; 21 genes), Phenylpropanoid biosynthesis (Enrichment

FDR = 1.12E-05; 23 genes), Biosynthesis of secondary metabolites (Enrichment FDR =

0.018866; 66 genes), Metabolic pathways (Enrichment FDR = 0.033888; 102 genes).

Overrepresented TF binding motifs in the promoters of these 1535 genes has revealed

AP2, B3, bHLH, bZIP, Myb/SANT, SBP, TCP as the enriched transcription family

(Supplementary File 11).
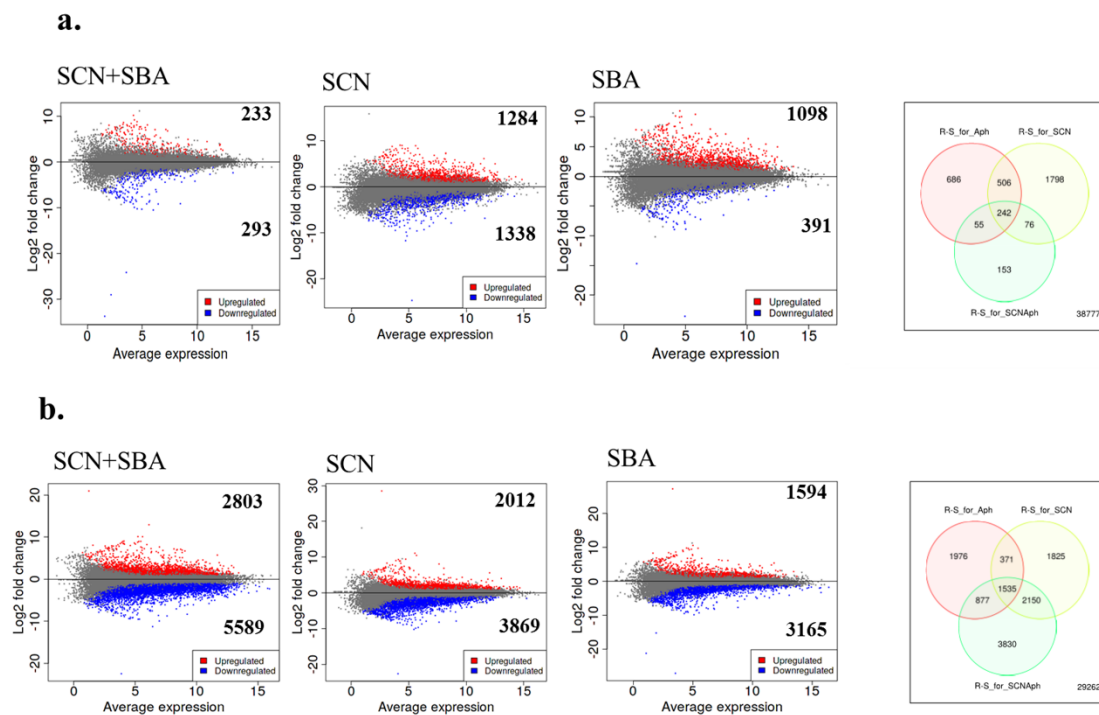
Figure 6.7. Visualization of DEGs using MA plots and Venn diagrams obtained from the comparison of the DEGs between susceptible and resistant cultivars. (a) 5 dpi, (b) 30 dpi.

a.

b.

c.

d.
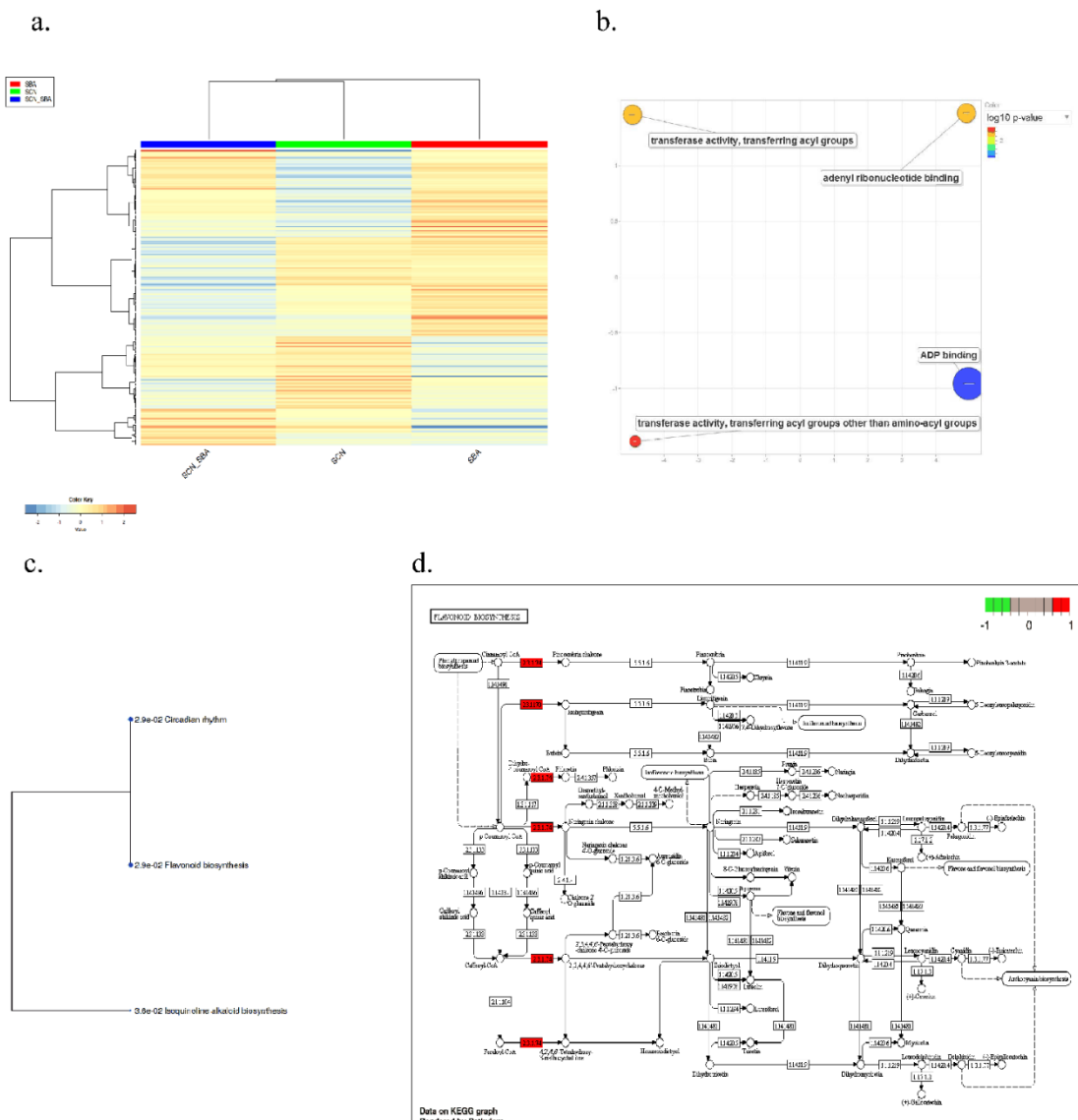
Figure 6.8. Assessment of 242 genes overlapped in treatments with SCN, SBA, and both SBA and SCN at 5 dpi in comparison of the DEGs between susceptible and resistant cultivars. (a) Heatmap based on log2foldchange (b) Enriched GO molecular functions (c) A hierarchical tree representing enriched KEGG pathways (d) A KEGG pathway representing Flavonoid Biosynthesis pathway with genes overrepresented.

Figure 6.9. Assessment of 1535 DEGs overlapped in treatments with SCN, SBA, and both SBA and SCN at 30 dpi in comparison of the DEGs between susceptible and resistant cultivars. (a) Heatmap based on log2foldchange (b) Enriched GO biological processes (c) A hierarchical tree representing enriched KEGG pathways (d) A KEGG pathway representing Phenylpropanoid Biosynthesis pathway with genes overrepresented.
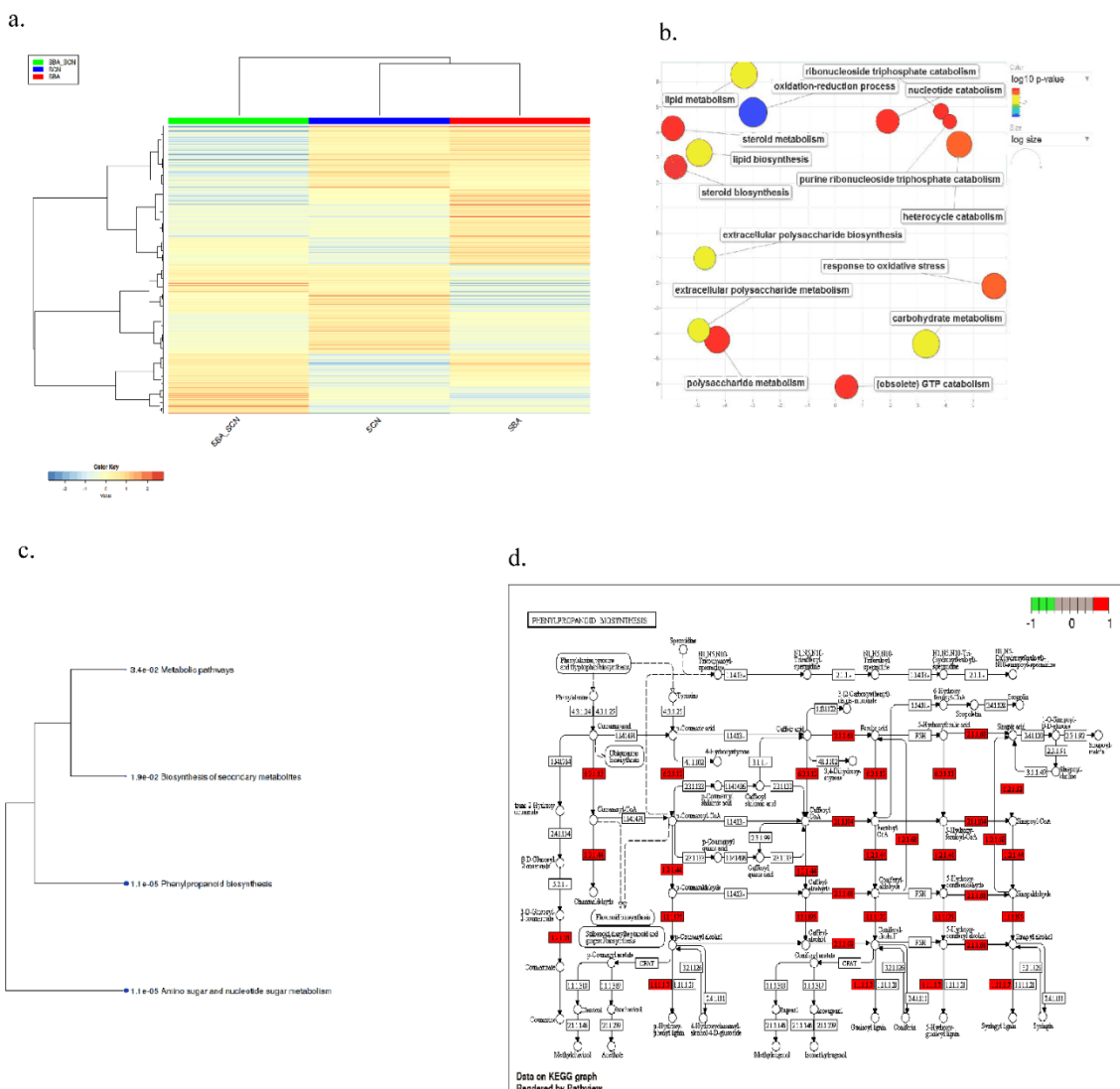
6.3.5. DEGs Coincident with SCN QTLs

The non-redundant 251 genes were assessed from the SCN QTLs [40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50]. In this study, we found three genes (of 242 DEGs overlapped at all treatments) at 5 dpi and ten genes (of 1535 DEGs overlapped at all treatments) at 30 dpi located in SCN QTLs (Supplementary File 12; Figure 6.10). Among them, Glyma.18G022400 (Transmembrane amino acid transporter protein), Glyma.18G022500 [soluble N-ethylmelaimide sensitive factor (NSF) attachment protein (*GmSNAP18*)], and Glyma.18G022700 (Wound-induced protein WI12) were up regulated from 2.53 log2fold change to 5.01 log2fold change in resistant cultivar as compared to susceptible cultivar in both time periods. These genes are present in a 31-kilobase (kb) segment at *rhg1-b* loci in Peking (PI548402) that play a significant role in SCN resistance [48, 49]. A recent study by Liu et al. 2017 [51] narrowed down the interval to ~14.3 kb in the recombinant lines of Forrest cultivar that contained three genes in three tandem repeats with in *rhg1-a* locus. These genes encode armadillo/β-catenin-like repeat (Glyma.18G022300), amino acid transporter (AAT), and soluble N-ethylmelaimide sensitive factor (NSF) attachment protein (*GmSNAP18*). However, Glyma.08G108900 [*Rhg4* (*GmSHMT08*)] gene, SCN-resistant allele [50], was downregulated at 30 dpi which was not found as DEG at 5 dpi. Other down regulated genes at 30 dpi involved Glyma.01G186900, Glyma.11G233500, and Glyma.18G023500 which belonged to protein kinases. The Glyma.14G043300 gene that belongs to the receptor like protein (RLP) was upregulated by 5.16 to 10.60 log2 fold change in all treatments at 30 dpi.

a.

b.
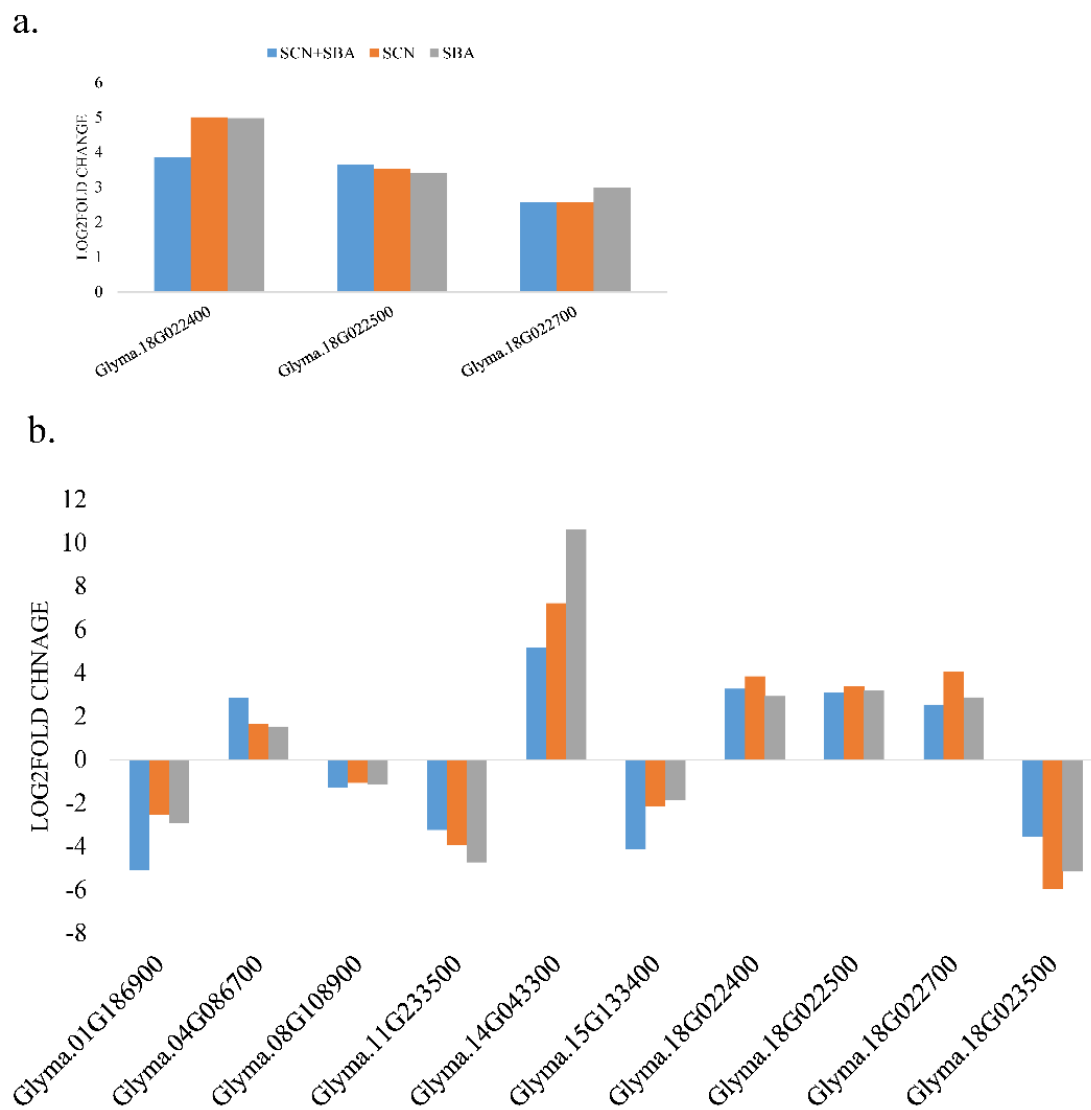
Figure 6.10. Log2fold change of the DEGs coincident with SCN QTLs upon a comparison of the DEGs between susceptible and resistant cultivars. (a) 5dpi (b) 30 dpi.

6.3.6. Comparison of DEGs within Susceptible and Resistant Cultivars

The purpose of these comparisons was to find uniquely expressed genes in the resistant cultivar at 5 dpi and 30 dpi. Overall, in all comparisons, we found fewer DEGs in resistant cultivar as compared to the susceptible cultivar. At 5dpi, 44 genes were differentially expressed in samples treated with both SCN and SBA in which 30 genes were up regulated and 14 genes were down regulated in the resistant cultivar. Whereas, at 30 dpi 578 genes were differentially expressed in which 214 genes were up regulated and 364 genes were down regulated in the resistant cultivar (Figure 6.11 and 6.12). At 5 dpi, we did not find any genes shared by all the treatments in the resistant cultivar, whereas, 40 genes were shared in the samples treated with SCN and the samples treated with both SCN and SBA. At 30 dpi, 139 genes were shared by all the treatments in the resistant cultivar. The transcriptome changes in these genes in treatments with SBA, SCN, and both SBA and SCN is shown in the heatmap (Figure 6.11). These genes were enriched for various pathways such as nine genes in MAPK signaling pathway, seven genes in plant-pathogen interaction, three genes in fatty acid metabolism, five genes in plant hormone signal transduction, 15 genes in metabolic pathways, two genes in alpha-linolenic acid metabolism (Table 6.1). The overrepresentation of genes for KEGG pathway of Plant hormone signal transduction and α-Linolenic acid metabolism is shown in Figure 6.13c and 6.13d, respectively. Overrepresented TF binding motifs in the promoters of the 139 genes have revealed AP2, bZIP, C2H2 ZF, GRAS, Myb/SANT, NAC/NAM, SBP as the enriched transcription family (Supplementary File 13).
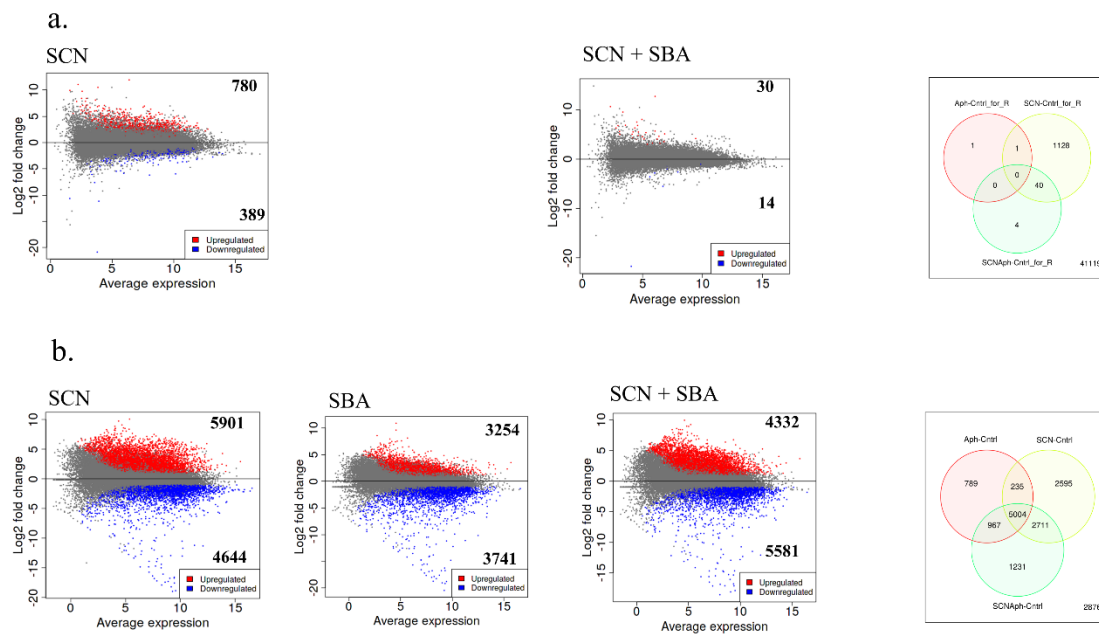
Figure 6.11. Visualization of DEGs using MA plots and Venn diagrams obtained from the comparison of the DEGs within susceptible and resistant cultivars at 5 dpi. (a) The resistant cultivar, (b) susceptible cultivar.
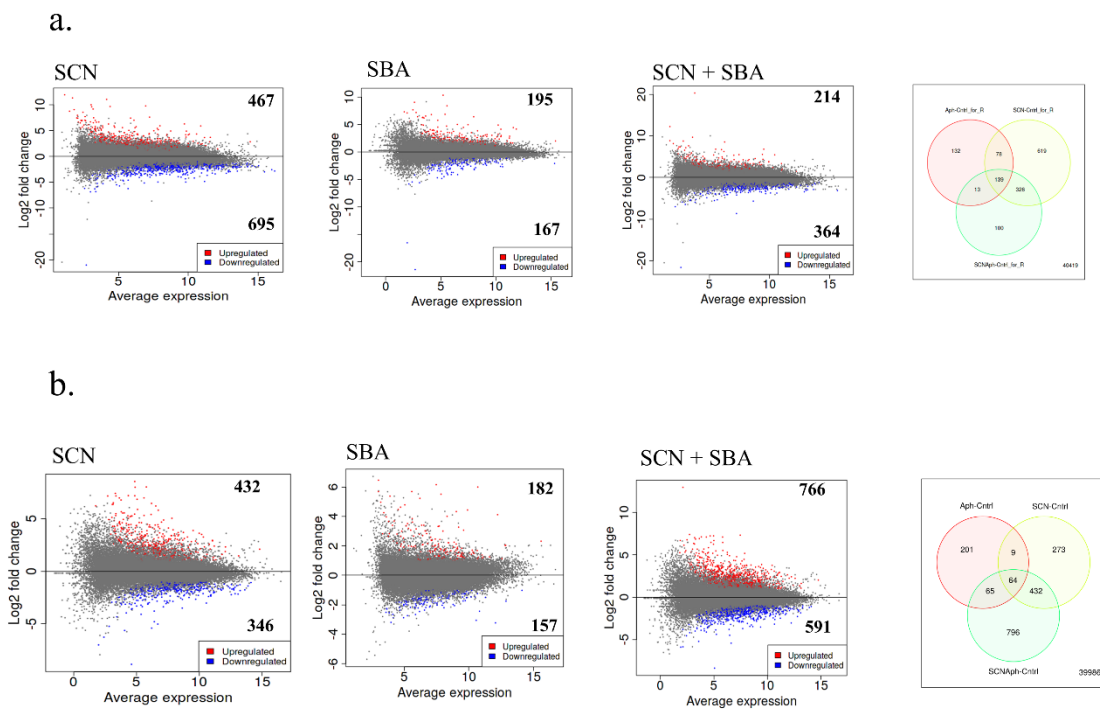
Figure 6.12. Visualization of DEGs using MA plots and Venn diagrams obtained from the comparison of the DEGs within susceptible and resistant cultivars at 30 dpi. (a) The resistant cultivar, (b) susceptible cultivar.

Table 6.1. Enriched KEGG pathways in 139 DEGs overlapped in treatments with SCN, SBA, and both SBA and SCN at 30 dpi in a resistant cultivar.

Enrichment FDR	Genes in list	Total Genes	Functional Category	Genes
1.31E-07	9	241	MAPK signaling pathway	Glyma.05G021100 Glyma.15G182000 Glyma.17G078300 Glyma.01G160100 Glyma.15G062400 Glyma.09G073200 Glyma.15G062500 Glyma.15G062700 Glyma.02G042500
2.88E-05	7	262	Plant-pathogen interaction	Glyma.05G021100 Glyma.15G182000 Glyma.17G078300 Glyma.15G062400 Glyma.09G073200 Glyma.15G062500 Glyma.15G062700
0.002244	3	61	Fatty acid biosynthesis	Glyma.20G007900 Glyma.04G197400 Glyma.07G161900
0.005913	3	94	Fatty acid metabolism	Glyma.20G007900 Glyma.04G197400 Glyma.07G161900
0.008062	3	134	Glycerolipid metabolism	Glyma.10G011000 Glyma.02G286500 Glyma.01G102900
0.008062	2	37	Linoleic acid metabolism	Glyma.13G030300 Glyma.20G054000
0.008062	15	2853	Metabolic pathways	Glyma.20G007900 Glyma.10G011000 Glyma.12G156600 Glyma.04G197400 Glyma.02G286500 Glyma.03G085500 Glyma.13G030300 Glyma.07G161900 Glyma.01G160100 Glyma.05G180600 Glyma.01G102900 Glyma.04G220600 Glyma.07G100500 Glyma.20G054000 Glyma.02G042500
0.008062	3	116	Peroxisome	Glyma.20G007900 Glyma.07G161900 Glyma.20G196900
0.013871	5	512	Plant hormone signal transduction	Glyma.07G015200 Glyma.13G354700 Glyma.15G062400 Glyma.15G062500 Glyma.15G062700
0.018845	2	77	Fatty acid degradation	Glyma.20G007900 Glyma.07G161900
0.018845	3	204	Amino sugar and nucleotide sugar metabolism	Glyma.12G156600 Glyma.01G160100 Glyma.02G042500
0.018845	2	75	α-Linolenic acid metabolism	Glyma.13G030300 Glyma.20G054000
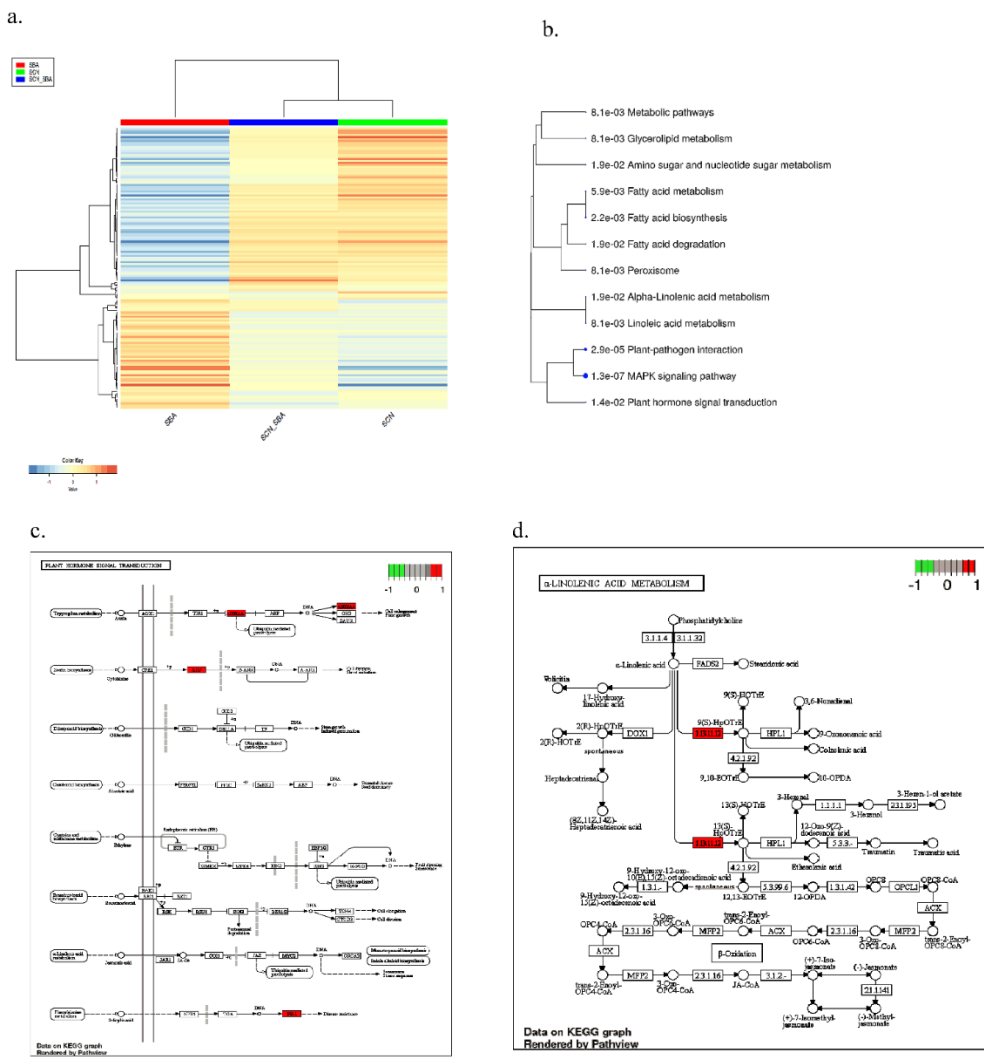
Figure 6.13. Assessment of 139 DEGs overlapped in treatments with SCN, SBA, and both SBA and SCN at 30 dpi in resistant cultivar. (a) Heatmap based on log2foldchange (b) A hierarchical tree representing enriched KEGG pathways (c) A KEGG pathway representing Plant Hormone Signal Transduction pathway with genes overrepresented. (d) A KEGG pathway representing α-Linolenic acid metabolism pathway with genes overrepresented.

6.3.7. MapMan Analysis of DEGs

Differentially expressed genes visualized using biotic stress pathway integrated into MapMan showed distinct expression patterns in SBA, SCN, and SCN + SBA

treatments in both susceptible and resistant cultivars. The biotic stress overview pathway generated by MapMan demonstrated the involvement of multifaceted defense related genes in presence either SBA or SCN or both in susceptible and resistant plants at 30 dpi. In resistant reaction, 164 (of 362) with 22 bins, 398 (of 1162) DEGs with 25 bins, 215 (of 578) DEGs with 24 bins were associated with the biotic stress pathway for treatments with SBA, SCN, SBA + SCN, respectively (Figure 6.12). Likewise, in susceptible reaction, 118 (of 339) with 22 bins, 350 (of 778) DEGs with 25 bins, 561 (of 1357) DEGs with 26 bins were associated with the biotic stress pathway for treatments with SBA, SCN, SBA + SCN, respectively (Figure 6.13). In a resistant reaction, there was consistent up-regulation of five genes (Glyma.17g078300, Glyma.09g073200, Glyma.15g182000, Glyma.05g021100, Glyma.08g005900) related to respiratory burst (bin 20.1.1) in treatments with SBA, SCN, SBA + SCN treatments. The expression of genes encoding signaling compounds such as calcium, receptor like Kinases, MAP kinases, proteolysis, heat shock proteins, and ethylene were varied across the treatments. In the context of hormone metabolism, a diverse number of genes were associated with ethylene, auxins, abscisic acid, salicylic acid, and jasmonic acid biosynthesis. In the presence of SCN only, 16 of 19 genes that were associated with cell wall metabolism were upregulated. However, in the presence of both SBA and SCN, 6 of 10 genes associated with cell wall metabolism were upregulated mostly related to pectate lyases, polygalactouronases, and esterases. In susceptible reaction, as compared to other treatments, 14 genes, most of them upregulated, were associated with redox reaction (bin 21) mainly related to thioredoxin, ascorbate and glutathione metabolism in presence of both SBA and SCN.

The genes related to transcription factors AP2/ERF (bin 27.3.3) and MYB (27.3.25) showed consistent upregulation in all treatments in susceptible reaction.
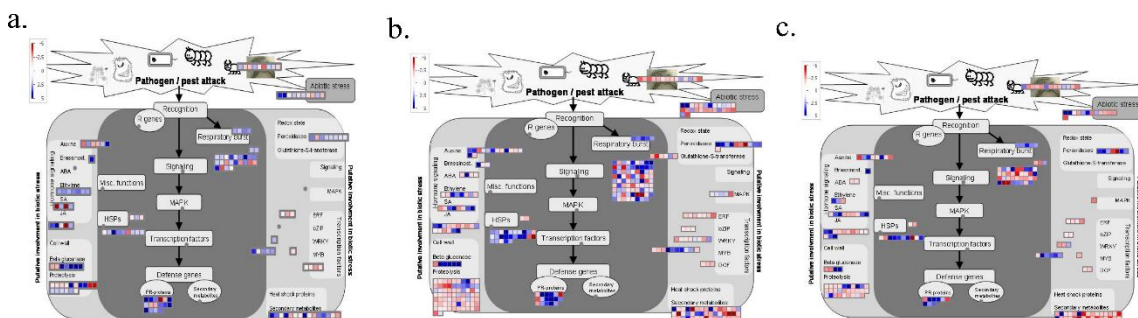
Figure 6.14. Biotic stress pathway overview of differentially expressed genes in resistant cultivar at day 30. (a) SBA, (b) SCN, and (c) SBA + SCN. Blue color indicates the up-regulated and red color indicates the down regulated genes. False discovery rate (FDR) *p* < 0.01 and logfold change ≥ 2 or ≤ -2 were used to identify the differentially expressed genes.
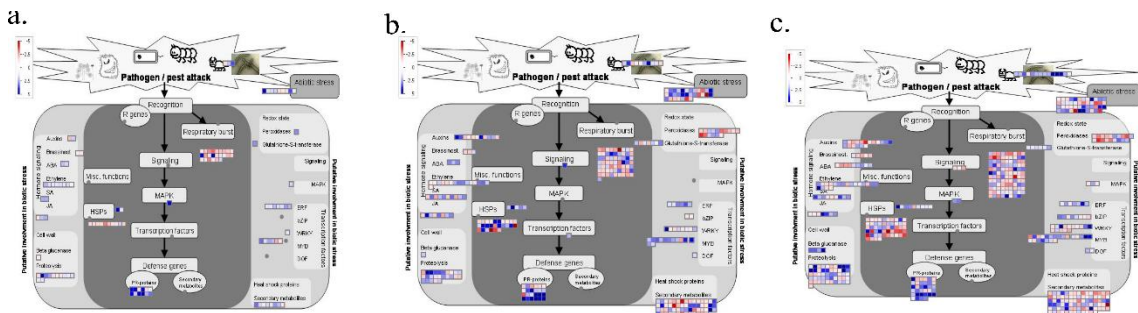
Figure 6.15. Biotic stress pathway overview of differentially expressed genes in susceptible cultivar at day 30. (a) SBA, (b) SCN, (c) SBA + SCN. Blue color indicates the up-regulated and red color indicates the down regulated genes. False discovery rate (FDR) *p* < 0.01 and logfold change ≥ 2 or ≤ -2 were used to identify the differentially expressed genes.

6.3.8. DEGs Unique to Resistant Cultivar

To identify the important genes responsible for HG type 0 and biotype 1 soybean aphid resistance in resistant cultivar, we specifically focused on the DEGs that are uniquely expressed in the resistant cultivar. For instance, only four and 100 genes were unique to the samples treated with both SCN and SBA at 5 dpi and 30 dpi, respectively in the resistant cultivar. The four genes at 5 dpi were Glyma.03G044900 (Dirigent-like protein), Glyma.13G147600 (2OG-Fe(II) oxygenase superfamily), Glyma.16G214400 (Exo70 exocyst complex subunit), and Glyma.20G089400 (Proteasome component domain protein) (Table 6.2).

Table 6.2. List of four DEGs uniquely expressed in the resistant cultivar treated with SBA and SCN at 5 dpi.

Gene ID	log2foldchange	p-value	Top Arabidopsis Hit	Gene Description	Gene Ontology Biological Process
Glyma.03g044900	8.04	7.16E-03	AT5G49040.1	Disease resistance-responsive (dirigent-like protein) family protein	GO:0006952 GO:0009807
Glyma.13g147600	-3.59	6.27E-03	AT2G36690.1	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein	GO:0009058 GO:0055114
Glyma.16g214400	7.50	4.90E-03	AT5G58430.1; ATEXO70B1, EXO70B1	exocyst subunit exo70 family protein B1	GO:0006887 GO:0006904 GO:0009738 GO:0035556
Glyma.20g089400	-1.04	2.81E-04	AT5G15610.2	Proteasome component (PCI) domain protein	GO:0006302 GO:0006312 GO:0007062 GO:0007129 GO:0007131 GO:0008150 GO:0009560 GO:0009909 GO:0034968 GO:0042138 GO:0045132

Among 100 genes, uniquely expressed at 30 dpi samples, three genes belonged to cytochrome P450s in which Glyma.03G160100 (CYP94B1), Glyma.10G115900 (CYP71B34) were down regulated and Glyma.12G239100 (CYP712A1) was up

regulated. Another group of genes belonged to UDP-glucosyltransferase activity, response to auxin stimulus, thioredoxin metabolism and many more (Table 6.3). The gene Glyma.04G096400 showed high expression (20-fold change) which belonged to cysteine-type endopeptidase inhibitor activity (GO: 0004870). The GO molecular function enrichment on these 100 genes showed most of the genes were enriched in glucosyltransferase activity, transferase activity, Thioredoxin-disulfide reductase activity, Oxidoreductase activity, and Calcium ion binding. Overrepresented TF binding motifs in the promoters of the 100 genes have revealed WRKY, TCP, SBP, GRAS, and bZIP as the enriched transcription family (Supplementary File 14).

Table 6.3. Enriched KEGG pathways in 100 DEGs uniquely expressed at 30 dpi samples treated with both SBA and SCN at 30 dpi in a resistant cultivar.

Enrichment FDR	Genes in list	Total genes	Functional Category	Genes
0.000104666	6	236	Quercetin 3-O-glucosyltransferase activity	Glyma.01G046300 Glyma.09G128300 Glyma.09G162400 Glyma.11G000500 Glyma.14G175400 Glyma.16G158100
0.000104666	6	236	Quercetin 7-O-glucosyltransferase activity	Glyma.01G046300 Glyma.09G128300 Glyma.09G162400 Glyma.11G000500 Glyma.14G175400 Glyma.16G158100
0.000894105	6	371	UDP-glucosyltransferase activity	Glyma.01G046300 Glyma.09G128300 Glyma.09G162400 Glyma.11G000500 Glyma.14G175400 Glyma.16G158100
0.002120752	6	458	Glucosyltransferase activity	Glyma.01G046300 Glyma.09G128300 Glyma.09G162400 Glyma.11G000500 Glyma.14G175400 Glyma.16G158100
0.004252586	6	544	UDP-glycosyltransferase activity	Glyma.01G046300 Glyma.09G128300 Glyma.09G162400 Glyma.11G000500 Glyma.14G175400 Glyma.16G158100
0.027351491	6	812	Transferase activity, transferring hexosyl groups	Glyma.01G046300 Glyma.09G128300 Glyma.09G162400 Glyma.11G000500 Glyma.14G175400 Glyma.16G158100
0.029538121	7	1139	Transferase activity, transferring glycosyl groups	Glyma.01G046300 Glyma.09G128300 Glyma.09G162400 Glyma.11G000500 Glyma.14G175400 Glyma.16G158100 Glyma.20G004900
0.037769863	2	63	Protein-disulfide reductase activity	Glyma.08G295600 Glyma.18G127400
0.049127578	2	78	Thioredoxin-disulfide reductase activity	Glyma.08G295600 Glyma.18G127400
0.049127578	2	81	Oxidoreductase activity, acting on a sulfur group of donors, disulfide as acceptor	Glyma.08G295600 Glyma.18G127400
0.049282369	4	478	Calcium ion binding	Glyma.06G079900 Glyma.12G089800 Glyma.03G157800 Glyma.13G191200

6.3.9. Enriched Transcription Factor (TF) Binding Motifs

TF motifs enriched in gene promoters (300 bp) of up- or down-regulated DEGs were analyzed to reveal gene regulatory mechanisms. Overrepresented promoters of DEGs for different comparisons were analyzed for multiple comparisons using the

transcription factor (TF) target gene sets in enrichment analyses. For instance, in resistant

cultivar treated with both SCN and SBA at 5 dpi (44 DEGs), TF family of homeodomain,

WRKY, NAC/NAM, bZIP, and WRKY were overrepresented in up-regulated genes and

none in down-regulated genes (Table 6.4). Likewise, in resistant cultivar treated with

both SCN and SBA at 30 dpi (578 DEGs), WRKY, TBP, bHLH were overrepresented in

downregulated genes and TBP (ATA), bZIP, Myb/SANT, AT hook, bHLH, CG-1 were

overrepresented in up regulated genes (Table 6.5). For other comparisons, enriched TF

motifs at 5 dpi and 30 dpi are presented in Supplementary File 15 and 16, respectively.

Table 6.4. Enriched transcription factor (TF) binding motifs in 44 DEGs in resistant
cultivar treated with both SCN and SBA at 5 dpi.

List	Motif	TF	TF family	FDR
	GCTGTCA	Glyma0041s00360.1	Homeodomain	2.50E-02
	GTCA	Glyma01g42410.1	Homeodomain	2.50E-02
	GTCA	Glyma01g03450.1	Homeodomain	2.50E-02
	TGACGGC	Glyma03g39040.1	Homeodomain	2.50E-02
	GTCAAC	Glyma01g43420.1	WRKY	3.00E-02
	GTCAA	Glyma01g43130.1	WRKY	3.00E-02
	GTCAA	Glyma07g36640.1	WRKY	3.10E-02
	GTCAA	Glyma15g37120.1	WRKY	3.20E-02
	GTCAA	Glyma02g45530.1	WRKY	3.20E-02
	GGTCAA	Glyma10g13720.1	WRKY	3.20E-02
Up regulated	GTCAAC	Glyma03g37870.1	WRKY	3.20E-02
	GTCAAC	Glyma02g15920.1	WRKY	3.20E-02
	GTCAAC	Glyma01g39600.1	WRKY	3.80E-02
	TTACGTAA	Glyma07g05660.1	NAC/NAM	3.80E-02
	TGTCGG	Glyma01g00510.1	B3	4.40E-02
	GTCAAC	Glyma09g06980.1	WRKY	5.60E-02
	GTCAA	Glyma01g06150.1	NAC/NAM	5.60E-02
	TGACGTCA	Glyma01g21020.1	bZIP	5.60E-02
	GTCAAC	Glyma06g17690.1	WRKY	6.00E-02
	GTCAA	Glyma04g06470.1	WRKY	6.70E-02

Table 6.5. Enriched transcription factor (TF) binding motifs in 44 DEGs in resistant cultivar treated with both SCN and SBA at 30 dpi.

List	Motif	TF	TF family	FDR
Down regulated	GTCAAC	Glyma09g06980.1	WRKY	6.30E-08
	GTCAA	Glyma04g08060.1	WRKY	6.30E-08
	GTCAAC	Glyma01g06870.1	WRKY	6.30E-08
	ATA	Glyma03g04500.1	TBP	1.20E-07
	GTCAAC	Glyma02g15920.1	WRKY	1.20E-07
	GTCAAC	Glyma01g43420.1	WRKY	2.00E-07
	GTCAA	Glyma09g39040.1	WRKY	2.00E-07
	CACGTG	Glyma01g09010.1	bHLH	2.60E-07
	GTCAAC	Glyma03g37870.1	WRKY	2.60E-07
	GACGTG	Glyma19g30680.1	bZIP	3.00E-07
	CACGTG	Glyma05g07490.1	bHLH	3.00E-07
	GTCAA	Glyma01g43130.1	WRKY	3.10E-07
	AGTCAACG	Glyma02g01420.1	WRKY	3.20E-07
	AGTCAA	Glyma09g39000.1	WRKY	3.20E-07
	GTCAAC	Glyma01g39600.1	WRKY	3.70E-07
	CACGTG	Glyma03g32740.1	bHLH	4.50E-07
	GTCAA	Glyma02g45530.1	WRKY	4.50E-07
	CACGTG	Glyma02g00980.1	bHLH	4.50E-07
	GTCAACG	Glyma01g05050.1	WRKY	4.70E-07
	GTCAAC	Glyma06g17690.1	WRKY	4.80E-07
Up regulated	ATA	Glyma03g04500.1	TBP	1.50E-08
	ACACGTG	Glyma08g08220.1	bZIP	2.80E-06
	GACGTG	Glyma19g30680.1	bZIP	4.30E-06
	ACGTGG	Glyma01g01740.1	bZIP	4.30E-06
	GGATAA	Glyma01g00600.1	Myb/SANT	1.30E-05
	ACGTGGC	Glyma01g38380.1	bZIP	1.30E-05
	GGATAA	Glyma13g43120.1	Myb/SANT	1.30E-05
	ACACGTG	Glyma04g04170.1	bZIP	1.30E-05
	ATATAATT	Glyma06g01700.1	AT hook	1.90E-05
	CACGTGT	Glyma09g06770.1	bHLH	1.90E-05
	CACGTG	Glyma02g00980.1	bHLH	1.90E-05
	CACGTG	Glyma01g04610.1	bHLH	2.60E-05
	CCACGTG	Glyma01g39450.1	bHLH	4.30E-05
	CACGTG	Glyma01g39360.1	bHLH	4.80E-05
	GCCACGTG	Glyma08g41620.1	bHLH	4.80E-05
	CACGTG	Glyma06g41620.1	bHLH	6.20E-05
	CACGTG	Glyma01g02250.1	bHLH	6.50E-05
	GGAT	Glyma05g36290.1	Myb/SANT	6.50E-05
	CACGTG	Glyma03g32740.1	bHLH	8.80E-05
	CGCGT	Glyma05g31190.1	CG-1	8.80E-05

6.3.10. PGSEA and KEGG Pathway Analysis

Pathway analysis of genes expressed at the 5 dpi *versus* 30 dpi was carried out using Parametric Analysis of Gene Set Enrichment (PGSEA). The analysis was conducted on the 2000 most differentially expressed genes (DEGs) using a false discovery rate cutoff of 0.1 (Supplementary File 17 and 18). GO molecular function

annotations confirmed the observed differential patterns in 5 dpi and 30 dpi treatments.

At 5 dpi, treatments involving SCN in both resistant and susceptible cultivar revealed

transcription factor activity (FDR= 1.36E-03) and modulation of various binding

functions (Figure 6.16a). At 30 dpi, treatments involving SCN in both resistant and

susceptible cultivar revealed the higher activity of ubiquitin-protein transferase activity

(FDR= 1.45E-05) with modulation of binding activity related to carbohydrate metabolism

(Figure 6.16b). The analysis revealed differential patterns in KEGG metabolic pathways.

At 5 dpi, Plant-pathogen Interaction (FDR= 3.98E-03), Ubiquitin mediated proteolysis

(FDR= 8.46E-03), Phenylalanine, tyrosine and tryptophan biosynthesis (FDR= 3.81E-

03), Cutin, suberin and wax biosynthesis (9.01E-03), Alpha-Linolenic acid metabolism

(6.57E-04), fatty acid degradation (FDR= 1.59E-E-03) pathways were enriched (Figure

6.15a). Whereas, at 30 dpi, most of the pathways were related to carbohydrate

metabolism [starch and sugar metabolism (FDR= 6.32E-04) , Pentose and glucuronate

interconversions (FDR= 6.32E-04), fructose and mannose metabolism (FDR=2.09E-04),

galactose metabolism (FDR= 4.81E-04)], fatty acid metabolism (FDR= 1.50E-04)

including fatty acid biosynthesis (FDR= 3.47E-04), fatty acid elongation (FDR= 7.17E-

04), Phenylpropanoid biosynthesis (3.80E-05), isoflavonoid biosynthesis (FDR= 1.68E-

04), one carbon pool by folate (FDR= 2.89E-03) (Figure 6.17b). The KEGG pathways

(Plant Pathogen Interaction and cutin, suberin and wax biosynthesis at 5 dpi and

isoflavonoid biosynthesis and one carbon pool by folate at 30 dpi) for enriched DEGs

with both SCN and SBA in resistant cultivar are represented in Figure 6.18 and Figure

6.19, respectively. The KEGG pathways for other comparisons at 5 dpi and 30 dpi are

represented in Supplementary Files 19 and 20, respectively.
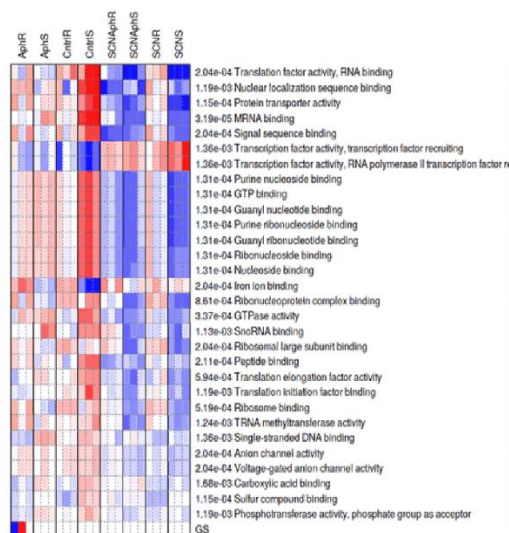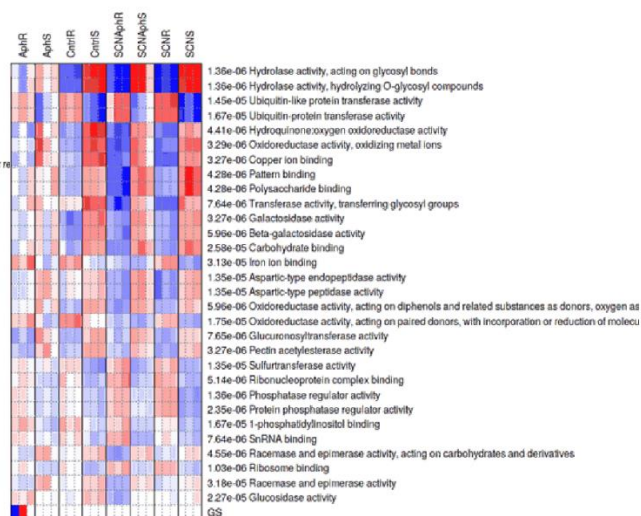
Figure 6.16. Gene Ontology (GO) molecular annotations overrepresented at (a) 5 dpi (b) 30 dpi upon PGSEA analysis. Red and blue indicate higher and lower pathway activities, respectively.
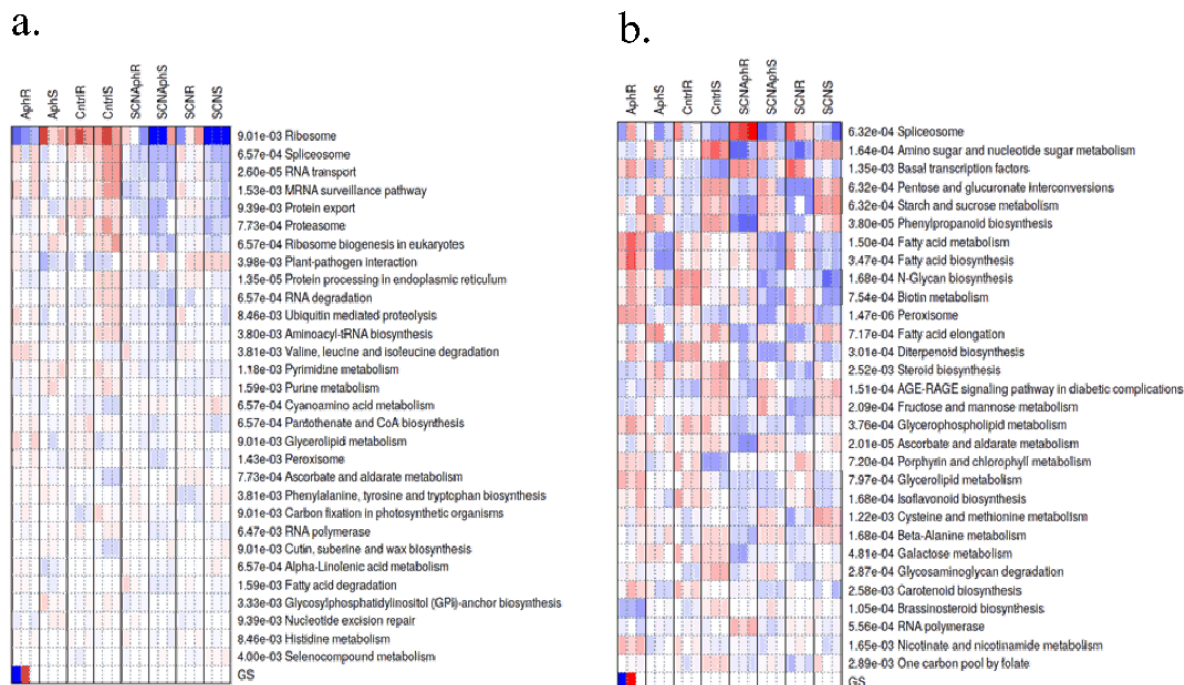
Figure 6.17. KEGG pathways overrepresented at (a) 5 dpi (b) 30 dpi upon PGSEA analysis. Red and blue indicate higher and lower pathway activities, respectively.
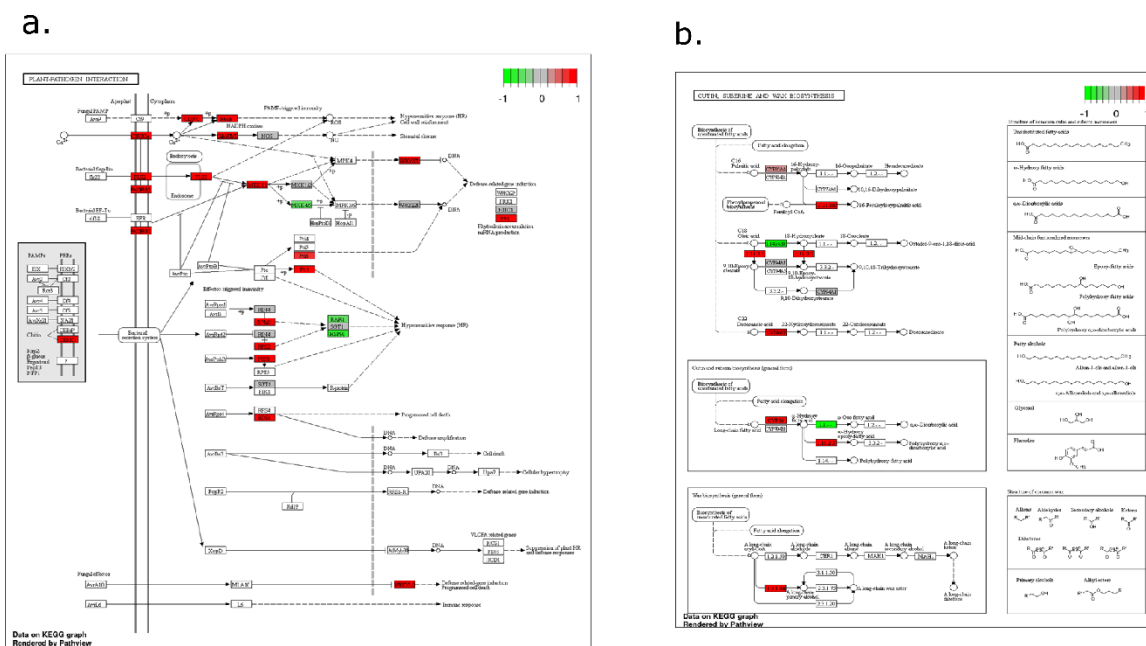
Figure 6.18. Expression profiles of (a) 'Plant Pathogen Interaction' and (b) 'Cutine, Suberine, and Wax Biosynthesis Pathway' visualized on a KEGG diagram for SCN + SBA in resistant cultivar at 5 dpi. Red and green indicate genes induced and suppressed, respectively.
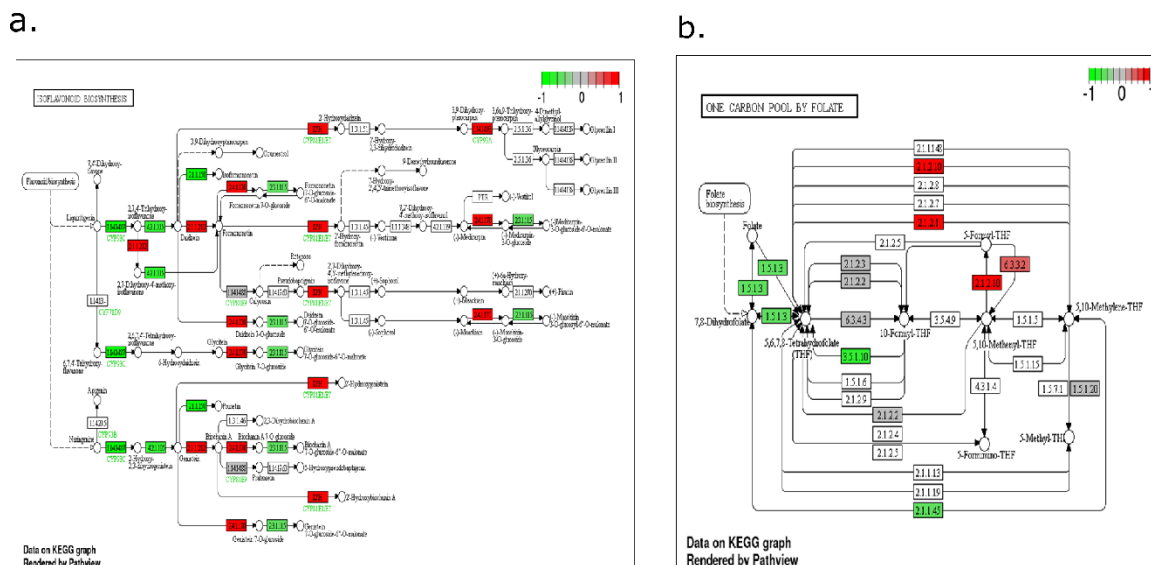
Figure 6.19. Expression profiles of (a) 'Isoflavonoid biosynthesis' (b) 'One carbon pool by folate' pathway visualized on a KEGG diagram for SCN + SBA in resistant cultivar at 30 dpi. Red and green indicate genes induced and suppressed, respectively.

6.4. Discussion

This study is the first to develop and use RNA sequences to study interaction effects of soybean-cyst-nematodes and soybean-aphids on soybean using SCN resistant, MN1806CN and SCN susceptible, Williams 82 (PI 518671) cultivars. MN1806CN cultivar carries the *Rps1k* gene for resistance to Phytophthora root rot making it resistant to races 1, 4, and 17, and is susceptible to the soybean-aphid. We used a similar greenhouse set up employed by McCarville et al. 2014 [5]. However, we utilized SCN resistant cultivar MN1806CN and SCN susceptible Williams 82 (PI 518671) instead of SCN resistant (Dekalb 27-52, PI 88788 derived) and SCN susceptible (Kenwood 94) soybean cultivars, 15 biotype 1 SBA instead of either zero, five, or ten biotype 1 SBA, data collected after 5 and 30 dpi by SBA instead of collecting data at 30 day after planting. McCarville et al. 2014 [5] showed both the SCN eggs and the number of

females increased by 33% in SCN-resistant cultivar and reduced by 50% in the SCN-susceptible cultivar in 30 days after planting. In our greenhouse experiment, at 30 dpi, SBA feeding significantly affected the reproduction of SCN depending on the cultivars. We observed a significant difference in SCN egg counts in susceptible cultivars and no effect in resistant cultivar in the presence of SBA. This suggests that the application of aphid increases the reproduction of the belowground SCN in the susceptible plants and no effects in the resistant plants. However, if we consider for each treatment, the final SCN egg counts have increased from the approximate initial counts of 2,000 eggs in both resistant and susceptible cultivars. Previously, Heeren et al. 2012 [21] utilized resistant and susceptible lines with respect to both soybean aphid and SCN in order to study the interaction effect of soybean aphid and SCN in the field conditions. The study showed that the effect of soybean aphid feeding on soybean on the SCN reproduction was not observed in any of the soybean cultivars as the SCN eggs and aphid densities, less than 100 SCN eggs per 100 cc of soil and less than 10 aphids per plant for less than 10 days, respectively, were too low in some of the cultivars.

We analyzed our data with respect to the aphid counts at 5 dpi, 15 dpi, and 30 dpi to see the trend on the SBA populations in the presence or absence of SCN. At 5 dpi and 15 dpi, we could not observe a significant difference on the SBA counts between all types of treatments. However, at 30 dpi we observed a significant difference on the SBA counts between all types of treatments. The facilitation at lower herbivore densities and competition at higher herbivore densities might be the reason for differences on the population densities of aphids depending on the length of the experiment [52]. Particularly, we observed 90% decrease in susceptible plants and 25% decrease in aphid

counts receiving SCN in resistant plants. The decline in the SBA populations compared to the plants that did not receive SCN might be due to the competition for food resources as both herbivores absorb nutrient assimilates via phloem and affect each other [52]. A similar pattern was also observed in a study by McCarville et al. 2014 [5] as SBA populations declined at 30 d and 60 d experiments when infested with five and ten SBA. Further, this relationship was shown by the regression analysis that showed a negative relationship between population density of SCN and aphids at 30 dpi.

The negative relationship was also observed in a various host-aphid-nematode interaction studies. The nematode, *Pratylenchus penetrans* infection on *Brassica nigra* caused a decreased infestation of shoot herbivore, *Pieris rapae* [53]. Bezemer et al. 2005 [54] reported decreased fertility of aphids *Rhopalosiphum padi* infesting *Agrostis capillaris* and *Anthoxanthum odoratum* because of decreased amino acid in the phloem sap of nematode infected plants. A similar type of effect was seen in the offspring of aphid *Myzus persicae* on *Plantago lanceolata* infected with nematode *P. penetrans* [55]. Hol et al. 2010 [56] reported a detrimental effect on aphids, *Brevicoryne brassicae* in the presence of nematodes (*H. schachtii*) in *B. oleracea*. Also, the reproduction of aphid (*Schizaphis rufula*) was reduced in the presence of three nematodes (*Pratylenchus*, *Meloidogyne*, and *Heterodera spp.*) in the plant, *Ammophila arenaria* in laboratory conditions [57]. The possible reason might be mechanical factors such as changes in waxes of the cuticle, leaf toughness or water content in the presence of nematodes [58]. The water stress in the aerial part of the host plant might affect the insects that rely on phloem feeding [59]. Also, decreased shoot herbivory could be because of the accumulation of phenolics and glucosinolates [53, 60].

We utilized RNA-sequencing approach to investigate the interaction effects of soybean cyst nematodes and soybean aphids on soybean at two time points after SBA infestation. RNA-seq produces data of transcripts with higher sensitivity, reproducibility and comprehensive dynamic ranges than conventional methods [61]. The RNA-seq data generated at two time points (5 dpi and 30 dpi) were analyzed separately to remove the complexity in the analyses.

In this study, identification of DEGs in resistant and susceptible cultivars was of particular interest, with the treatment that received both SCN and SBA to study the genes that are differentially expressed during the interaction. We did a comparison between and within the cultivars. Upon comparison, the discordant expression of genes particularly in resistant cultivar was considered important. In total, we found 4 and 100 discordantly expressed DEGs in resistant cultivar at 5dpi and 30 dpi, respectively. At 5dpi, Dirigent-like protein, 2OG-Fe (II) oxygenase superfamily), genes encoding Exo70 exocyst complex subunit, and Proteasome component domain proteins were differentially expressed in the resistant cultivar that received both SCN and SBA. Dirigent (DIR) -like protein are particularly induced in different kinds of biotic such as wounding and abiotic stresses ranging from drought, cold, abscisic acid (ABA), H_2O_2, salinity, and osmotic stress [62, 63, 64]. These proteins play a crucial role in plant defenses against pathogens and lignin and lignan formation [65]. In the present study, one DIR-like protein (Glyma.03G044900) was upregulated by 8.04 log2foldchange at 5dpi. DIR-like proteins were also upregulated during feeding of spruce (*Picea* spp.) stem-by boring insects (i.e., white pine weevil, *Pissodes strobi*) in bark tissue and defoliating insects (i.e., western spruce budworm, *Choristoneura occidentalis*) in green apical shoots [62]. In soybean,

GmDIR22 conferred resistance to *Phytophthora sojae* regulating lignan biosynthesis [66]. Another gene, Glyma.16g214400 that belonged to exocyst subunit exo70 family protein B1 was upregulated by 7.50 log2fold change. The exocyst subunit Exo70B1 interacts with soluble *N*-ethylmaleimide–sensitive factor attachment protein receptor (SNARE) complex protein SNAP in the process of vesicular trafficking, which mediates the exocytosis [67]. Previously, the role of α-SNAP, which is one of the important genes in *Rhg1* mediated SCN resistance, has been unraveled in SCN resistance in many studies [48, 49, 68, 69]. SNAP protein, involves in vesicle trafficking that affects the exocytosis of food in syncytium which in turn affecting the nematode physiology [48]. Another important DEG is Glyma.13G147600 (2OG-Fe (II) oxygenase superfamily), which is down regulated by 3.59 log2fold change. The 2OG-Fe(II) oxygenase superfamily that constitutes flavone synthase I (FNS I), flavonol synthase (FLS), anthocyanidin synthase (ANS) and flavanone 3β-hydroxylase (FHT), play important role in flavonoid biosynthesis [70]. The remaining DEG, Glyma.20G089400 (Proteasome component domain protein) was also downregulated by 1.04 log2fold change. The plant proteasomes play an important role in an auxin signaling pathway, oxidative stress and hyper sensitive responses, which are an important component of plant defenses [71].

At 30 dpi, we found 100 DEGs that were uniquely expressed in the resistant cultivar and 21 of them were upregulated. Particularly, Glyma.04G096400 with high expression (20 log2fold change) shows cysteine-type endopeptidase inhibitor activity, and possess cystatin domain. The cystatins are basically low molecular weight proteins that inhibit various exogenous proteases or digestive enzymes of invasive pests and pathogens [72]. It has been demonstrated that the serine protease activity of *H. glycines*

has been inhibited by cowpea trypsin inhibitor (CpTI) [73]. Numerous studies on the

expression of both native and transgenic cystatins have shown resistance to

phytonematodes in a wide range of hosts (see review in [74]). The transgenic expression

of rice cystatin in eggplant improved nematode resistance in eggplant against root knot

nematode, *Meloidogyne incognita* [74]. Three DEGs belonged to cytochrome P450s

[Glyma.03G160100 (CYP94B1), Glyma.10G115900 (CYP71B34), and

Glyma.12G239100 (CYP712A1)]. *CYP94*-genes play important role in Jasmonic Acid

signaling pathway via catalyzing the sequential ω-oxidation of JA-Ile [75]. Whereas,

CYP71 is involved in flavonoid biosynthesis in producing isoflavone and pterocarpan

derivatives such as glyceollin in soybean in pathogen-infected tissues [76]. RNA-seq

analysis of two *Glycine soja* genotypes, PI 424093 and PI 468396B, upon infestation

by HG type 2.5.7 revealed upregulation of JA, including SA, and ET pathways [77].

Upon pathway enrichment of 100 DEGs, six genes (Glyma.01G046300,

Glyma.09G128300, Glyma.09G162400, Glyma.11G000500 Glyma.14G175400, and

Glyma.16G158100) were enriched for glucosyltransferase activity and four genes

(Glyma.06G079900, Glyma.12G089800, Glyma.03G157800, and Glyma.13G191200)

were enriched for calcium ion binding activity. Previously, the role of glucosyltransferase

has been shown in *Mi*-mediated nematode resistance in tomato [78], which plays an

important role in carbohydrate and cell-wall biosynthesis [79]. Calcium/calmodulin-

mediated signaling has been shown to be involved in responses to *H. glycines* infection

in *G. soja* [77].

PSGEA analysis, performed for understanding biological function, showed

distinct enriched pathways at 5 dpi and 30 dpi. Plant-pathogen interaction; ubiquitin-

mediated proteolysis; phenylalanine, tyrosine and tryptophan biosynthesis; cutin, suberin and wax biosynthesis; alpha-linolenic acid metabolism; and fatty acid degradation pathways were enriched at 5 dpi. Plant-pathogen interaction and ubiquitin-mediated proteolysis play important role in plant immunity [80]. The interaction of plant and pathogen involves pathogen-associated molecular patterns (PAMPs) of pathogens by pattern-recognition receptors (PRRs) of host and are regulated by E3 ubiquitin ligase [80]. E3 ubiquitin ligase has been previously reported in involvement in phytonematodes such as *Heterodera schachtii* [81] and *Globodera rostochiensis* [82]. Phenylalanine, tyrosine, and tryptophan biosynthesis pathway is related to the shikimate pathway. It is shown that the chorismite mutase enzyme in root-knot nematode and potato cyst nematode alters the shikimate pathway of the host plant [83]. Other pathways such as cutin, suberin and wax biosynthesis, alpha-linolenic acid metabolism, and fatty acid degradation pathways, are related to plant lipid metabolisms, which are important for the production of JA, cutins, and suberins in plant defense via wounding [84, 85].

At 30 dpi, most of the pathways were related to carbohydrate metabolism, fatty acid metabolism including fatty acid biosynthesis, fatty acid elongation, phenylpropanoid biosynthesis, isofalvonoid biosynthesis, and one carbon pool by folate. Phenylpropanoid biosynthesis and isoflavonoid biosynthesis pathways are particularly related to the metabolism that produces compounds such as flavonoids, anthocynanins, lignin, suberin, salicylic acid, coumarins and furanocoumarins [86]. It has been shown that phloem-feeding-insect, whitefly *Bemisia tabaci*, when infested in *Nicotiana tabacum* activates the phenylpropanoid pathway [87]. We expect the carbohydrate metabolism pathway to be enriched at 30 dpi, as SBA and SCN might be competing for the limited food. The "one

carbon pool by folate" pathway is related to folate-mediated one-carbon metabolism, and

Rhg4 resistance gene *SHMT* (*GmSHMT08*) catalyzes methylene carbon of glycine to

tetrahydrofolate (THF) [88, 89]. Also, *GmSHMT08* changes its enzymatic properties in

the resistant allele negatively affecting the folate homeostasis in the syncytium that

causes hypersensitive responses (HR) leading to programmed cell death (PCD) [89, 90].

We examined DEGs that are coincident with the 251 genes assessed from the SCN QTLs.

Remarkably, we found three genes at 5 dpi and ten genes at 30 dpi, located in SCN QTLs

among common genes in all treatments in comparisons of resistant versus susceptible

cultivar. We were interested in finding out if the resistant cultivar, MN1806CN provided

rhg1 and *Rhg4* mediated resistance. Three genes Glyma.18G022400 (Transmembrane

amino acid transporter protein), Glyma.18G022500 [soluble N-ethylmelaimide sensitive

factor (NSF) attachment protein (*GmSNAP18*)], and Glyma.18G022700 (Wound-induced

protein WI12), upregulated at both 5 and 30 dpi in all treatments are important for *rhg1*-

mediated SCN resistance [48, 49, 51]. We could not find *Rhg4* gene as DEG at 5 dpi but

was downregulated at 30 dpi in all treatments. This indicates that the resistant cultivar,

MN1806CN might possess *rhg1*mediated SCN resistance.

In summary, the expressions of DEGs were changed after SCN and SBA infection

during SCN susceptible and resistant soybean interactions. Many genes revealed various

pathways and networks involved in the interaction effects of SCN and SBA on soybean.

Although a huge number of genes were differentially expressed, when compared between

resistant and susceptible cultivars, the comparison within the cultivars exhibited fewer

DEGs conferring resistance against both SBA and SCN in the resistant cultivar. One

limitation was that the cultivar that was resistant to SCN was susceptible to SBA. Various

GO enrichment and KEGG pathways identified several molecular mechanisms involved in SCN-SBA interaction. Identified role of transcription factors in the SBA-SCN interaction in this study can be used for future research and breeding for SCN and SBA resistance in soybean.

Supplementary Files

https://figshare.com/s/b96d56d94a0cd9230f96

Acknowledgments

References

1. Yu, X.; Yuan, F.; Fu, X.; Zhu, D. Profiling and relationship of water-soluble sugar and protein compositions in soybean seeds. *Food Chemistry* 2016, *196*, 776-782.
2. Hartman, G.; Domier, L.; Wax, L.; Helm, C.; Onstad, D.; Shaw, J.; Solter, L.; Voegtlin, D.; d'Arcy, C.; Gray, M. Occurrence and distribution of *Aphis glycines* on soybeans in Illinois in 2000 and its potential control. *Plant Health Progress* 2001, *10*.
3. Wrather, J.A.; Koenning, S.R. Estimates of disease effects on soybean yields in the United States 2003 to 2005. *Journal of Nematology* 2006, *38*, 173.
4. McCarville, M.; O'Neal, M.; Tylka, G.; Kanobe, C.; MacIntosh, G. A nematode, fungus, and aphid interact via a shared host plant: implications for soybean management. *Entomologia Experimentalis et Applicata* 2012, *143*, 55-66.

5.   McCarville, M.T.; Soh, D.H.; Tylka, G.L.; O'Neal, M.E. Aboveground feeding by soybean aphid, *Aphis glycines*, affects soybean cyst nematode, *Heterodera glycines*, reproduction belowground. *PloS one* 2014, *9*.

6.   Hill, C.; Chirumamilla, A.; Hartman, G. Resistance and virulence in the soybean-*Aphis glycines* interaction. *Euphytica* 2012, *186*, 635-646.

7.   Koenning, S.R.; Wrather, J.A. Suppression of soybean yield potential in the continental United States by plant diseases from 2006 to 2009. *Plant Health Progress* 2010, *10*.

8.   Niblack, T.; Lambert, K.; Tylka, G. A model plant pathogen from the kingdom animalia: *Heterodera glycines*, the soybean cyst nematode. *Annu. Rev. Phytopathol.* 2006, *44*, 283-303.

9.   Olson, K.D.; Badibanga, T.M.; DiFonzo, C. Farmers' awareness and use of IPM for soybean aphid control: report of survey results for the 2004, 2005, 2006, and 2007 crop years; 2008.

10.   Niblack, T. Soybean cyst nematode management reconsidered. *Plant disease* 2005, *89*, 1020-1026.

11.   Ragsdale, D.W.; Landis, D.A.; Brodeur, J.; Heimpel, G.E.; Desneux, N. Ecology and management of the soybean aphid in North America. *Annual review of entomology* 2011, *56*, 375-399.

12.   Koch, R.L.; Potter, B.D.; Glogoza, P.A.; Hodgson, E.W.; Krupke, C.H.; Tooker, J.F.; DiFonzo, C.D.; Michel, A.P.; Tilmon, K.J.; Prochaska, T.J. Biology and economics of recommendations for insecticide-based management of soybean aphid. *Plant Health Progress* 2016, *17*, 265-269.

13.   Varenhorst, A.J.; O'Neal, M.E. The response of natural enemies to selective insecticides applied to soybean. *Environmental Entomology* 2012, *41*, 1565-1574.

14.   Tylka, G.L. Understanding soybean cyst nematode HG types and races. *Plant Health Progress* 2016, *17*, 149-151.

15.   Mitchum, M.G.; Wrather, J.A.; Heinz, R.D.; Shannon, J.G.; Danekas, G. Variability in distribution and virulence phenotypes of *Heterodera glycines* in Missouri during 2005. *Plant Disease* 2007, *91*.

16.   Hesler, L.S.; Chiozza, M.V.; O'neal, M.E.; MacIntosh, G.C.; Tilmon, K.J.; Chandrasena, D.I.; Tinsley, N.A.; Cianzio, S.R.; Costamagna, A.C.; Cullen, E.M. Performance and prospects of R ag genes for management of soybean aphid. *Entomologia Experimentalis et Applicata* 2013, *147*, 201-216.

17.   Mitchum, M.G. Soybean resistance to the soybean cyst nematode *Heterodera glycines*: an update. *Phytopathology* 2016, *106*, 1444-1450.

18.   Megías, A.G.; Müller, C. Root herbivores and detritivores shape above-ground multitrophic assemblage through plant-mediated effects. *Journal of Animal Ecology* 2010, *79*, 923-931.

19.   Hong, S.; Donaldson, J.; Gratton, C. Soybean cyst nematode effects on soybean aphid preference and performance in the laboratory. *Environ. Entomol.* 2010, *39*, 1561-1569.

20.   Hong, S.; MacGuidwin, A.; Gratton, C. Soybean aphid and soybean cyst nematode interactions in the field and effects on soybean yield. *Journal of Economic Entomology* 2011, *104*, 1568-1574.

21. Heeren, J.; Steffey, K.; Tinsley, N.; Estes, R.; Niblack, T.; Gray, M. The interaction of soybean aphids and soybean cyst nematodes on selected resistant and susceptible soybean lines. *J. Appl. Entomol.* 2012, *136*, 646-655.

22. Griffith, M.; Walker, J.R.; Spies, N.C.; Ainscough, B.J.; Griffith, O.L. Informatics for RNA sequencing: a web resource for analysis on the cloud. *PLoS Comp. Biol.* 2015, *11*, e1004393.

23. Shan, X.; Li, Y.; Jiang, Y.; Jiang, Z.; Hao, W.; Yuan, Y. Transcriptome profile analysis of maize seedlings in response to high-salinity, drought and cold stresses by deep sequencing. *Plant molecular biology reporter* 2013, *31*, 1485-1491.

24. Kim, K.-S.; Hill, C.B.; Hartman, G.L.; Mian, M.; Diers, B.W. Discovery of soybean aphid biotypes. *Crop Science* 2008, *48*, 923-928.

25. Andrews, S. FastQC: a quality control tool for high throughput sequence data. 2010.

26. Ewels, P.; Magnusson, M.; Lundin, S.; Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016, *32*, 3047-3048.

27. Kong, Y.J.G. Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. *Genomics* 2011, *98*, 152-153.

28. Patro, R.; Duggal, G.; Love, M.I.; Irizarry, R.A.; Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 2017, *14*, 417.

29. Grüning, B.; Dale, R.; Sjödin, A.; Chapman, B.A.; Rowe, J.; Tomkins-Tinch, C.H.; Valieris, R.; Köster, J.; Bioconda, T. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature methods* 2018, *15*, 475.

30. Ge, S.X.; Son, E.W.; Yao, R. iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data. BMC Bioinformatics 2018, 19, 534.

31. Maaten, L.v.d.; Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research* 2008, *9*, 2579-2605.

32. Langfelder, P.; Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008, *9*, 559.

33. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology and Evolution* 2014, *15*, 550.

34. Grant, D.; Nelson, R.T.; Cannon, S.B.; Shoemaker, R.C. SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Research* 2009, gkp798.

35. Ge, S.; Jung, D. ShinyGO: a graphical enrichment tool for animals and plants. *bioRxiv* 2018, 315150.

36. Supek, F.; Bošnjak, M.; Škunca, N.; Šmuc, T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLOS ONE* 2011, *6*, e21800.

37. Thimm, O.; Bläsing, O.; Gibon, Y.; Nagel, A.; Meyer, S.; Krüger, P.; Selbig, J.; Müller, L.A.; Rhee, S.Y.; Stitt, M. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal* 2004, *37*, 914-939.

38. Lohse, M.; Nagel, A.; Herter, T.; May, P.; Schroda, M.; Zrenner, R.; Tohge, T.; Fernie, A.R.; Stitt, M.; Usadel, B. Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant Cell Environment* 2014, *37*, 1250-1258.

39. Furge, K.; Dykema, K. PGSEA: parametric gene set enrichment analysis. *R package version 1.2* 2006.

40. Vuong, T.D.; Sonah, H.; Meinhardt, C.G.; Deshmukh, R.; Kadam, S.; Nelson, R.L.; Shannon, J.G.; Nguyen, H.T. Genetic architecture of cyst nematode resistance revealed by genome-wide association study in soybean. *BMC Genomics* 2015, *16*, 593.

41. Zhang, H.; Li, C.; Davis, E.L.; Wang, J.; Griffin, J.D.; Kofsky, J.; Song, B.-H. Genome-Wide Association Study of Resistance to Soybean Cyst Nematode (*Heterodera glycines*) HG Type 2.5. 7 in Wild Soybean (Glycine soja). *Frontiers in Plant Science* 2016, *7*.

42. Zhang, H.; Song, Q.; Griffin, J.D.; Song, B.-H. Genetic architecture of wild soybean (Glycine soja) response to soybean cyst nematode (*Heterodera glycines*). *Molecular Genetics Genomics* 2017, *292*, 1257-1265.

43. Chang, H.-X.; Lipka, A.E.; Domier, L.L.; Hartman, G.L. Characterization of disease resistance loci in the USDA soybean germplasm collection using genome-wide association studies. *Phytopathology* 2016, *106*, 1139-1151.

44. Li, Y.-h.; Shi, X.-h.; Li, H.-h.; Reif, J.C.; Wang, J.-j.; Liu, Z.-x.; He, S.; Yu, B.-s.; Qiu, L.-j. Dissecting the genetic basis of resistance to soybean cyst nematode combining linkage and association mapping. *The Plant Genome* 2016, *9*.

45. Wen, Z.; Tan, R.; Yuan, J.; Bales, C.; Du, W.; Zhang, S.; Chilvers, M.I.; Schmidt, C.; Song, Q.; Cregan, P.B. Genome-wide association mapping of quantitative resistance to sudden death syndrome in soybean. *BMC Genomics* 2014, *15*, 809.

46. Han, Y.; Zhao, X.; Cao, G.; Wang, Y.; Li, Y.; Liu, D.; Teng, W.; Zhang, Z.; Li, D.; Qiu, L. Genetic characteristics of soybean resistance to HG type 0 and HG type 1.2. 3.5. 7 of the cyst nematode analyzed by genome-wide association mapping. *BMC Genomics* 2015, *16*, 598.

47. Kim, M.; Hyten, D.L.; Bent, A.F.; Diers, B.W. Fine mapping of the SCN resistance locus *rhg1-b* from PI88788. *The Plant Genome* 2010, *3*, 81-89.

48. Cook, D.E.; Lee, T.G.; Guo, X.; Melito, S.; Wang, K.; Bayless, A.M. Copy number variation of multiple genes at *Rhg1* mediates nematode resistance in soybean. *Science* 2012, *338*.

49. Bayless, A.M.; Smith, J.M.; Song, J.; McMinn, P.H.; Teillet, A.; August, B.K.; Bent, A.F. Disease resistance through impairment of α-SNAP–NSF interaction and vesicular trafficking by soybean Rhg1. *Proceedings of the National Academy of Sciences* 2016, *113*, E7375-E7382.

50. Kandoth, P.K.; Liu, S.; Prenger, E.; Ludwig, A.; Lakhssassi, N.; Heinz, R.; Zhou, Z.; Howland, A.; Gunther, J.; Eidson, S. Systematic mutagenesis of serine hydroxymethyltransferase reveals essential role in nematode resistance. *Plant Physiology* 2017, pp. 00553.02017.

51. Liu, S.; Kandoth, P.K.; Lakhssassi, N.; Kang, J.; Colantonio, V.; Heinz, R.; Yeckel, G.; Zhou, Z.; Bekal, S.; Dapprich, J. The soybean GmSNAP18 gene

underlies two types of resistance to soybean cyst nematode. *Nature Communications* 2017, *8*, 14822.

52. Soler, R.; Erb, M.; Kaplan, I. Long distance root–shoot signalling in plant–insect community interactions. *Trends in Plant Science* 2013, *18*, 149-156.

53. Van Dam, N.M.; Raaijmakers, C.E.; Van Der Putten, W.H. Root herbivory reduces growth and survival of the shoot feeding specialist Pieris rapae on Brassica nigra. *Entomologia Experimentalis et Applicata* 2005, *115*, 161-170.

54. Bezemer, T.; De Deyn, G.; Bossinga, T.; Van Dam, N.; Harvey, J.; Van der Putten, W. Soil community composition drives aboveground plant–herbivore–parasitoid interactions. *Ecology Letters* 2005, *8*, 652-661.

55. Wurst, S.; van der Putten, W.H. Root herbivore identity matters in plant-mediated interactions between root and shoot herbivores. *Basic and Applied Ecology* 2007, *8*, 491-499.

56. Hol, W.G.; De Boer, W.; Termorshuizen, A.J.; Meyer, K.M.; Schneider, J.H.; Van Dam, N.M.; Van Veen, J.A.; Van Der Putten, W.H. Reduction of rare soil microbes modifies plant–herbivore interactions. *Ecology Letters* 2010, *13*, 292-301.

57. Vandegehuchte, M.L.; De La Peña, E.; Bonte, D. Interactions between root and shoot herbivores of Ammophila arenaria in the laboratory do not translate into correlated abundances in the field. *Oikos* 2010, *119*, 1011-1019.

58. Kuhlmann, F.; Müller, C. UV-B impact on aphid performance mediated by plant quality and plant changes induced by aphids. *Plant Biology* 2010, *12*, 676-684.

59. Huberty, A.F.; Denno, R.F. Plant water stress and its consequences for herbivorous insects: a new synthesis. *Ecology* 2004, *85*, 1383-1398.

60. Kaplan, I.; Halitschke, R.; Kessler, A.; Rehill, B.J.; Sardanelli, S.; Denno, R.F. Physiological integration of roots and shoots in plant defense strategies links above-and belowground herbivory. *Ecology Letters* 2008, *11*, 841-851.

61. Jiang, L.; Romero-Carvajal, A.; Haug, J.S.; Seidel, C.W.; Piotrowski, T. Gene-expression analysis of hair cell regeneration in the zebrafish lateral line. *Proceedings of the National Academy of Sciences of the United States of America* 2014, *111*, E1383-E1392.

62. Ralph, S.G.; Jancsik, S.; Bohlmann, J. Dirigent proteins in conifer defense II: Extended gene discovery, phylogeny, and constitutive and stress-induced gene expression in spruce (Picea spp.). *Phytochemistry* 2007, *68*, 1975-1991.

63. Wu, R.; Wang, L.; Wang, Z.; Shang, H.; Liu, X.; Zhu, Y.; Qi, D.; Deng, X. Cloning and expression analysis of a dirigent protein gene from the resurrection plant Boea hygrometrica. *Progress in Natural Science* 2009, *19*, 347-352.

64. Moura, J.C.M.S.; Bonine, C.A.V.; de Oliveira Fernandes Viana, J.; Dornelas, M.C.; Mazzafera, P. Abiotic and biotic stresses and changes in the lignin content and composition in plants. *Journal of Integrative Plant Biology* 2010, *52*, 360-376.

65. Effenberger, I.; Zhang, B.; Li, L.; Wang, Q.; Liu, Y.; Klaiber, I.; Pfannstiel, J.; Wang, Q.; Schaller, A. Dirigent proteins from cotton (Gossypium sp.) for the atropselective synthesis of gossypol. *Angewandte Chemie International Edition* 2015, *54*, 14660-14663.

66. Li, N.; Zhao, M.; Liu, T.; Dong, L.; Cheng, Q.; Wu, J.; Wang, L.; Chen, X.; Zhang, C.; Lu, W. A novel soybean dirigent gene GmDIR22 contributes to promotion of lignan biosynthesis and enhances resistance to *Phytophthora sojae*. *Frontiers in plant science* 2017, *8*, 1185.

67. He, B.; Guo, W. The exocyst complex in polarized exocytosis. *Current Opinion in Cell Biology* 2009, *21*, 537-542.

68. Matsye, P.D.; Lawrence, G.W.; Youssef, R.M.; Kim, K.-H.; Lawrence, K.S.; Matthews, B.F.; Klink, V.P. The expression of a naturally occurring, truncated allele of an α-SNAP gene suppresses plant parasitic nematode infection. *Plant Molecular Biology* 2012, *80*, 131-155.

69. Cook, D.; Bayless, A.; Wang, K.; Guo, X.; Song, Q.; Jiang, J.; Bent, A. Distinct copy number, coding sequence and locus methylation patterns underlie Rhg1-mediated soybean resistance to soybean cyst nematode. *Plant Physiology* 2014, pp. 114.235952.

70. Cheng, A.-X.; Han, X.-J.; Wu, Y.-F.; Lou, H.-X. The function and catalysis of 2-oxoglutarate-dependent oxygenases involved in plant flavonoid biosynthesis. *International Journal of Molecular Sciences* 2014, *15*, 1080-1095.

71. Suty, L.; Lequeu, J.; Lançon, A.; Etienne, P.; Petitot, A.-S.; Blein, J.-P. Preferential induction of 20S proteasome subunits during elicitation of plant defense reactions: towards the characterization of "plant defense proteasomes". *The International Journal of Biochemistry and Cell Biology* 2003, *35*, 637-650.

72. Martínez, M.; Cambra, I.; González-Melendi, P.; Santamaría, M.E.; Díaz, I. C1A cysteine-proteases and their inhibitors in plants. *Physiol. Plant.* 2012, *145*, 85-94.

73. Lilley, C.; Urwin, P.; McPherson, M.; Atkinson, H. Characterization of intestinally active proteinases of cystnematodes. *Parasitology* 1996, *113*, 415-424.

74. Papolu, P.K.; Dutta, T.K.; Tyagi, N.; Urwin, P.E.; Lilley, C.J.; Rao, U. Expression of a Cystatin Transgene in Eggplant Provides Resistance to Root-knot Nematode, Meloidogyne incognita. *Frontiers in Plant Science* 2016, *7*.

75. Bruckhoff, V.; Haroth, S.; Feussner, K.; König, S.; Brodhun, F.; Feussner, I. Functional Characterization of CYP94-Genes and Identification of a Novel Jasmonate Catabolite in Flowers. *PLOS ONE* 2016, *11*, e0159875.

76. Latunde-Dada, A.O.; Cabello-Hurtado, F.; Czittrich, N.; Didierjean, L.; Schopfer, C.; Hertkorn, N.; Werck-Reichhart, D.; Ebel, J. Flavonoid 6-Hydroxylase from Soybean (Glycine maxL.), a Novel Plant P-450 Monooxygenase. *The Journal of Biological Chemistry* 2001, *276*, 1688-1695.

77. Zhang, H.; Kjemtrup-Lovelace, S.; Li, C.; Luo, Y.; Chen, L.P.; Song, B.-H. Comparative RNA-Seq Analysis Uncovers a Complex Regulatory Network for Soybean Cyst Nematode Resistance in Wild Soybean (*Glycine soja*). *Scientific Reports* 2017, *7*, 9699.

78. Schaff, J.E.; Nielsen, D.M.; Smith, C.P.; Scholl, E.H.; Bird, D.M. Comprehensive Transcriptome Profiling in Tomato Reveals a Role for Glycosyltransferase in Mi-Mediated Nematode Resistance. *Plant Physiology* 2007, *144*, 1079-1092.

79. Egelund, J.; Skjøt, M.; Geshi, N.; Ulvskov, P.; Petersen, B.L. A Complementary Bioinformatics Approach to Identify Potential Plant Cell Wall Glycosyltransferase-Encoding Genes. *Plant Physiology* 2004, *136*, 2609-2620.

80. Macho, A.P.; Zipfel, C. Plant PRRs and the activation of innate immune signaling. *Molecular Cell* 2014, *54*, 263-272.

81. Hewezi, T.; Piya, S.; Qi, M.; Balasubramaniam, M.; Rice, J.H.; Baum, T.J. Arabidopsis miR827 mediates post-transcriptional gene silencing of its ubiquitin E3 ligase target gene in the syncytium of the cyst nematode Heterodera schachtii to enhance susceptibility. *The Plant Journal* 2016, *88*, 179-192.

82. Walter, A.J.; Willforss, J.; Lenman, M.; Alexandersson, E.; Andreasson, E. RNA seq analysis of potato cyst nematode interactions with resistant and susceptible potato roots. *European Journal of Plant Pathology* 2018, *152*, 531-539.

83. Davis, E.L.; Hussey, R.S.; Mitchum, M.G.; Baum, T.J. Parasitism proteins in nematode–plant interactions. *Current Opinion in Plant Biology* 2008, *11*, 360-366.

84. Howe, G.A. Jasmonates as signals in the wound response. *Journal of Plant Growth Regulation* 2004, *23*, 223-237.

85. Li, Y.; Beisson, F.; Koo, A.J.; Molina, I.; Pollard, M.; Ohlrogge, J. Identification of acyltransferases required for cutin biosynthesis and production of cutin with suberin-like monomers. *Proceedings of the National Academy of Sciences* 2007, *104*, 18339-18344.

86. Dixon, R.A.; Achnine, L.; Kota, P.; Liu, C.-J.; Reddy, M.S.S.; Wang, L. The phenylpropanoid pathway and plant defence—a genomics perspective. *Molecular Plant Pathology* 2002, *3*, 371-390.

87. Alon, M.; Malka, O.; Eakteiman, G.; Elbaz, M.; Moyal Ben Zvi, M.; Vainstein, A.; Morin, S. Activation of the Phenylpropanoid Pathway in *Nicotiana tabacum* Improves the Performance of the Whitefly *Bemisia tabaci* via Reduced Jasmonate Signaling. *PLOS ONE* 2013, *8*, e76619.

88. Ros, R.; Muñoz-Bertomeu, J.; Krueger, S. Serine in plants: biosynthesis, metabolism, and functions. *Trends in Plant Science* 2014, *19*, 564-569.

89. Liu, S.; Kandoth, P.K.; Warren, S.D.; Yeckel, G.; Heinz, R.; Alden, J. A soybean cyst nematode resistance gene points to a new mechanism of plant resistance to pathogens. *Nature* 2012, *492*.

90. Wu, X.-Y.; Zhou, G.-C.; Chen, Y.-X.; Wu, P.; Liu, L.-W.; Ma, F.-F.; Wu, M.; Liu, C.-C.; Zeng, Y.-J.; Chu, A.E. Soybean cyst nematode resistance emerged via artificial selection of duplicated serine hydroxymethyltransferase genes. *Frontiers in Plant Science* 2016, *7*, 998.

CHAPTER 7: CONCLUSIONS

Current study utilized both *in silico* as well as greenhouse experiments to study stress responsive genes in the genomes of soybean and sunflower. This study has identified 117 TNL R genes in soybean, where most of the genes were under purifying selection except for a few accessions under positive selection. Approximately 63% of the regular TNL genes were found in clusters in soybean, which signifies their origin is primarily by tandem duplications. Characterization of these TNL genes is warranted for understanding resistance pathways paving avenues toward crop improvement. This study has also confirmed 352 NBS encoding genes in sunflower genome, and reported their phylogenetic relationships and functional divergence. These genes also formed clusters and showed structural conservation in signature domains and exon/intron architecture in CNL, TNL and RNL types of NBS genes. Interestingly, the RNLs were nested within the CNL-A clade, making the CNL clade paraphyletic, which warrants further analysis in future. All of the NBS-encoding genes have undergone purifying selection and available expression data have revealed their functional divergence. Further characterization of the NBS genes will help us understand resistance pathways as well as develop durable resistance necessary for crop improvement in sunflower and soybean.

This study has become the first to report 28 MPKs and eight MKKs in the genome of sunflower and to conduct comparative analyses of the genomic architecture and phylogenetic relationships with nine other plant species representing diverse taxonomic groups of plant kingdom. Though sunflower genome with 3.6 gigabases is one

of the largest among the plants under this study, the MPKs and MKKs are slightly fewer than that in soybean, which has the genome size of 975 Mbs. Among MPKs and MKK genes studied, MKK3 group of proteins were highly conserved and retained in all the species under study, including outgroup, *C. reinhardtii*. This result warrants further investigation through an exploration of a wide array of species. Transcriptomic data analyzed under hormone and abiotic stresses treatments revealed a diverse expression pattern of sunflower MPKs and MKKs, exhibiting a dynamic role to adapt to changing environmental conditions. The results advance our understanding of the diversity and evolution of MAPK genes and their signaling pathways in sunflower, and are expected to help in cultivar improvement through stress-tolerance breeding.

Present study characterized induced susceptibility effects of soybean and soybean aphid interaction using demographic data from greenhouse experiments and genetic data based on RNA-sequencing. The characterization was limited to two treatments: one with no inducer population and biotype 1 as a response population, and another with biotype 2 as an inducer population and biotype 1 as a response population, in both resistant and susceptible cultivars. Kal's z-test integrated with CLC Genomics Workbench (https://www.qiagenbioinformatics.com/) was used to study the differential gene expression for pooled samples with no replications. Many DEGs were common and unique in two cultivars and treatments that were enriched for various biological processes and pathways and were functionally related to known defense mechanisms reported in various host-aphid systems. The responses to aphid biotype 1 infestation in presence or absence of inducer population at day1 and day11 revealed significant differences on the gene enrichment and regulation in resistant and susceptible cultivars. Assessment of

DEGs in *Rag* genes QTLs, particularly in *Rag1* containing QTL on chromosome 7, six

non-NBS-LRR genes – Glyma.07G034800, Glyma.07G034900, Glyma.07G051500,

Glyma.07G051500, Glyma.07G061500, Glyma.07G087200 revealed distinct expression

in treatments with absence or presence of inducer population at day 1 and day 11.

However, four TNL genes – Glyma.03G048600, Glyma.03G052800, Glyma.03G048700,

and Glyma.03G047700 (identified in the study in chapter 2) were upregulated in resistant

cultivar treated with biotype 2 as an inducer population and biotype 1 as response

population at day 11 which suggest their crucial role in the interaction effects. Further

experiments based upon metabolomics, proteomics, and validation of the candidate genes

will be needed to understand the mechanism underlying induced susceptibility effects.

In the last project, a three-way interaction among soybean, SBA and SCN was

characterized. Various DEGs whose expressions were changed in the days after the SCN

and SBA infection during SCN susceptible and resistant soybean interactions are

reported. Many genes revealed various pathways and networks involved in the interaction

effects of SCN and SBA on soybean. Although a huge number of genes were found

differentially expressed between resistant and susceptible cultivars, fewer DEGs

conferring resistance against both SBA and SCN were found in the resistant cultivar. In

total, four and 100 DEGs were found in resistant cultivar at 5dpi and 30 dpi, respectively.

In the present study, these genes are inferred to play important roles during SBA-SCN

interaction on soybean. One limitation was that the cultivar resistant to SCN was

susceptible to SBA. Various GO enrichment and KEGG pathways identified several

molecular mechanisms involved in three-way interaction, and transcription factors

identified for the interaction can be used for future research and breeding programs. Further work will be needed for the functional validation of identified DEGs.

Although significant progress has been made on pinpointing specific genes in the *Rhg* QTLs, narrowing down to very specific genes responsible for soybean aphid resistance warrants further investigations. The advent of sequencing technologies has made now the availability of soybean, soybean aphid, and SCN (de novo assembly) genomes. This should speed the discovery of particular molecular cues in terms of effector and host resistance components. With the development of various gene editing tools such as CRISPR/Cas9 system and advancement in producing various mutant hosts could help on understanding the function of these genes. In addition, since soybean aphid and SCN have been co-existed in many soybean fields, integrative system biology approaches might yield results useful for the plant-pest management.

APPENDIX

APPENDIX I: Codes Used for RNA-seq Analyses as Described in Chapter 4

```
# Codes used for data processing
#--------------------------------------------In Unix--------------------------------------------------------

#Logged in to SOUTH DAKOTA STATE UNIVERSITY High Performance Cluster
ssh username@blackjack

###Download SRR files
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR499/005/SRR4996815/SRR4996815_1.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR499/005/SRR4996815/SRR4996815_2.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR499/009/SRR4996819/SRR4996819_1.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR499/009/SRR4996819/SRR4996819_2.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR499/003/SRR4996823/SRR4996823_1.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR499/003/SRR4996823/SRR4996823_2.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR499/008/SRR4996828/SRR4996828_1.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR499/008/SRR4996828/SRR4996828_2.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR499/004/SRR4996834/SRR4996834_1.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR499/004/SRR4996834/SRR4996834_2.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR499/006/SRR4996836/SRR4996836_1.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR499/006/SRR4996836/SRR4996836_2.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR499/009/SRR4996839/SRR4996839_1.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR499/009/SRR4996839/SRR4996839_2.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR499/007/SRR4996847/SRR4996847_1.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR499/007/SRR4996847/SRR4996847_2.fastq.gz

###gunzip files
##for one pair samples, run
gunzip SRR4996815_1.fastq.gz
gunzip SRR4996815_2.fastq.gz

###Qulity control of raw data using FastQC
#For one pair raw read files,run
fastqc SRR4996815_1.fastq
fastqc SRR4996815_2.fastq

###Btrim64 to trim low-quality bases
#for one pair raw read files, run
btrim64-static -q -t /path/to/file/SRR4996815_1.fastq -o /path/for/output/file/SRR4996815_1_trimmed.fastq
btrim64-static -q -t /path/to/file/SRR4996815_2.fastq -o /path/for/output/file/SRR4996815_2_trimmed.fastq

###Activate Biconda Channel
conda activate environment_name

##Run Salmon tool

#Buid Index
```

salmon index -t /path/to/file/Hannuus_494_r1.2.transcript.fa.gz -i /path/for/output/file/Ha_transcripts_index --
    type quasi -k 31

##Quantify reads
#for one one pair trimmed files
salmon quant -i /path/to/file/Ha_transcripts_index -l A -1 /path/to/file/SRR4996815_1_trimmed.fastq -2
    /path/to/file/SRR4996815_2_trimmed -o SRR4996815_count

#---Customized R codes via iDEP 0.81--------------------------------
    ------
# hierarchical clustering tree
 x <- readData.out$data
 maxGene <- apply(x,1,max)

# Parameters for heatmap
 input_geneCentering <- TRUE   #centering genes ?
 input_sampleCentering <- FALSE   #Center by sample?
 input_geneNormalize <- TRUE   #Normalize by gene?
 input_sampleNormalize <- FALSE   #Normalize by sample?
 input_noSampleClustering <- FALSE   #Use original sample order
 input_heatmapCutoff <- 4   #Remove outliers beyond number of SDs
 input_distFunctions <- 1   #which distant funciton to use (Correlation)
 input_hclustFunctions <- 1   #Linkage type (Average)
 input_heatColors1 <- 5   #Colors (Blue-white-brown)

APPENDIX II: Codes Used for RNA-seq Analyses as Described in Chapter 5 and 6

---In Unix--

#Logged in to SOUTH DAKOTA STATE UNIVERSITY High Performance Cluster
ssh username@blackjack
password

###Upload fastq.gz file in the cluster

###gunzip files
##for one sample, run
gunzip 3-1a_S25_L005_R1_001.fastq.gz

###Qulity control of raw data using FastQC
##For one pair raw read files,run
fastqc 3-1a_S25_L005_R1_001.fastq

##MultiQC of all samples
mutiqc .

##Btrim64 to trim low-quality bases
#for one raw read file, run
btrim64-static -q -t /path/to/file/3-1a_S25_L005_R1_001.fastq -o /path/for/output/file/3-1a_S25_L005_R1_001_trimmed

##Trimmomatic to trim low-quality bases (Chapter 5)
java -jar trimmomatic-0.36.jar SE -phred33 /stor2/neupanex/SoybeanAphid/2017_07_28_Madav-44509465/5A_Madav_07_2017-53233373/5A.fastq /stor2/neupanex/SoybeanAphid/2017_07_28_Madav-44509465/5A_trimmomatric.fastq ILLUMINACLIP:TruSeq3-SE:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

###Activate Biconda Channel
conda activate environment_name

##Run Salmon tool

#Buid Index
salmon index -t /path/to/file/Gmax_275_Wm82.a2.v1.transcript_primaryTranscriptOnly.fa.gz -i /path/for/output/file/Gm_transcripts_index --type quasi -k 31

##Quantify reads
#for one trimmed file
salmon quant -i /path/to/file/Gm_transcripts_index -l A -1 /path/to/file/3-1a_S25_L005_R1_001.fastq -2 /path/to/file/3-1a_S25_L005_R1_001_trimmed -o 3-1a_S25_L005_R1_001_count

```
#----------------------------------------------------------Customized R codes via iDEP 0.81------------------------

###########################
# 1. Read data
###########################
 setwd('C:/Users/owner/Downloads')
 source('iDEP_core_functions.R')

 # Input files
 inputFile <- 'Downloaded_Converted_Data.csv'  # Expression matrix
 # Experiment design file
 sampleInfoFile <- 'Downloaded_sampleInfoFile.csv'
  #Gene symbols, location etc.
 geneInfoFile <- 'Glycine_max__gmax_eg_gene_GeneInfo.csv'
 # pathway database in SQL; can be GMT format
 geneSetFile <- 'Glycine_max__gmax_eg_gene.db'
 STRING10_speciesFile <- 'https://raw.githubusercontent.com/iDEP-
SDSU/idep/master/shinyapps/idep/STRING10_species.csv'

 # Parameters
 input_missingValue <- 'geneMedian'          #Missing values imputation method
 input_dataFileFormat <- 1           #1- read counts, 2 FKPM/RPKM or DNA microarray
 input_minCounts <- 0.5    #Min counts
 input_NminSamples <- 1  #Minimum number of samples
 input_countsLogStart <- 4          #Pseudo count for log CPM
 input_CountsTransform <- 3        #Methods for data transformation of counts. 1-EdgeR's logCPM; 2-
VST; 3-rlog

 #Read data files
 readData.out <- readData(inputFile)
 readSampleInfo.out <- readSampleInfo(sampleInfoFile)
 input_selectOrg ="NEW"
 input_noIDConversion = TRUE
 allGeneInfo.out <- geneInfo(geneInfoFile)
 converted.out = NULL
 convertedData.out <- convertedData()
 nGenesFilter()
 convertedCounts.out <- convertedCounts()  # converted counts, just for compatibility
 readCountsBias()  # detecting bias in sequencing depth

###########################
# 2. Pre-Process
###########################

 # Box plot
 x = readData.out$data
 boxplot(x, las = 2, col=col1,
   ylab='Transformed expression levels',
   main='Distribution of transformed data')

 # Density plot
 par(parDefault)
 densityPlot()
```

```
# Scatter plot of the first two samples
plot(x[,1:2],xlab=colnames(x)[1],ylab=colnames(x)[2],
  main='Scatter plot of first two samples')




###########################
# 3. Heatmap
###########################
# hierarchical clustering tree
x <- readData.out$data
maxGene <- apply(x,1,max)
# remove bottom 25% lowly expressed genes, which inflate the PPC
x <- x[which(maxGene > quantile(maxGene)[1] ) ,]
plot(as.dendrogram(hclust2( dist2(t(x)))), ylab="1 - Pearson C.C.", type = "rectangle")
#Correlation matrix
input_labelPCC <- TRUE #Show correlation coefficient?
correlationMatrix()

# Parameters for heatmap
input_geneCentering <- TRUE      #centering genes ?
input_sampleCentering <- FALSE #Center by sample?
input_geneNormalize <- TRUE      #Normalize by gene?
input_sampleNormalize <- FALSE#Normalize by sample?
input_noSampleClustering <- FALSE        #Use original sample order
input_heatmapCutoff <- 4#Remove outliers beyond number of SDs
input_distFunctions <- 1   #which distant funciton to use (Correlation)
input_hclustFunctions <- 1            #Linkage type (Average)
input_heatColors1 <- 5     #Colors (Blue-white-brown)




###########################
# 4. k-Means clustering
###########################
input_nGenesKNN <- 2000          #Number of genes fro k-Means
input_nClusters <- 4        #Number of clusters
maxGeneClustering = 12000
input_kmeansNormalization <- 'geneStandardization'#Normalization
input_KmeansReRun <- 0#Random seed

distributionSD()  #Distribution of standard deviations
KmeansNclusters()  #Number of clusters

Kmeans.out = Kmeans()   #Running K-means
KmeansHeatmap()   #Heatmap for k-Means


#Read gene sets for enrichment analysis
sqlite  <- dbDriver('SQLite')
input_selectGO3 <- 'KEGG'          #Gene set category
input_minSetSize <- 15    #Min gene set size
input_maxSetSize <- 6000          #Max gene set size
GeneSets.out <-readGeneSets( geneSetFile,
  convertedData.out, input_selectGO3,input_selectOrg,
```

```
   c(input_minSetSize, input_maxSetSize)  )
# Alternatively, users can use their own GMT files by
#GeneSets.out <- readGMTRobust('somefile.GMT')
KmeansGO()  #Enrichment analysis for k-Means clusters


input_seedTSNE <- 0       #Random seed for t-SNE
input_colorGenes <- TRUE          #Color genes in t-SNE plot?
tSNEgenePlot()  #Plot genes using t-SNE


##########################
# 5. PCA
##########################
input_selectFactors <- 'Treatment'  #Factor coded by color
input_selectFactors2 <- 'Cultivar'   #Factor coded by shape
input_tsneSeed2 <- 0       #Random seed for t-SNE
#PCA, MDS and t-SNE plots
PCAplot()


##########################
# 6. DEG1
##########################
input_CountsDEGMethod <- 3       #DESeq2= 3,limma-voom=2,limma-trend=1
input_limmaPval <- 0.01   #FDR cutoff
input_limmaFC <- 2       #Fold-change cutoff
input_selectModelComprions <- c('Cultivar: R vs. S','Treatment: Aph vs. Cntrl','Treatment: SCN vs.
Aph','Treatment: SCN vs. Cntrl','Treatment: SCNAph vs. Aph','Treatment: SCNAph vs. Cntrl','Treatment:
SCNAph vs. SCN')          #Selected comparisons
input_selectFactorsModel <- c('Cultivar','Treatment')#Selected comparisons
input_selectInteractions <- 'Cultivar:Treatment'       #Selected comparisons
input_selectBlockFactorsModel <- NULL   #Selected comparisons
factorReferenceLevels.out <- c('Cultivar:R','Treatment:Cntrl')


limma.out <- limma()
limma.out$comparisons
DEG.data.out <- DEG.data()
input_selectComparisonsVenn <- c('SCN-Aph','SCNAph-Aph','SCNAph-SCN')#Selected comparisons for
Venn diagram
input_UpDownRegulated <- FALSE          #Split up and down regulated genes
vennPlot() # Venn diagram
sigGeneStats() # number of DEGs as figure
sigGeneStatsTable() # number of DEGs as table


##########################
# 7. DEG2
##########################
input_selectContrast <- 'SCN-Aph' #Selected comparisons
selectedHeatmap.data.out <- selectedHeatmap.data()
selectedHeatmap()   # heatmap for DEGs in selected comparison

# Save gene lists and data into files
write.csv( selectedHeatmap.data()$genes, 'heatmap.data.csv')
write.csv(DEG.data(),'DEG.data.csv' )
write(AllGeneListsGMT() ,'AllGeneListsGMT.gmt')


input_selectGO2 <- 'KEGG'          #Gene set category
```

```
geneListData.out <- geneListData()
volcanoPlot()
scatterPlot()
MAplot()
geneListGOTable.out <- geneListGOTable()
# Read pathway data again
GeneSets.out <-readGeneSets( geneSetFile,
  convertedData.out, input_selectGO2,input_selectOrg,
  c(input_minSetSize, input_maxSetSize)  )
input_removeRedudantSets <- TRUE        #Remove highly redundant gene sets?
geneListGO()

# STRING-db API access
STRING10_species = read.csv(STRING10_speciesFile)
ix = grep('Mus musculus', STRING10_species$official_name )
findTaxonomyID.out <- STRING10_species[ix,1] # find taxonomyID
findTaxonomyID.out
# users can also skip the above and assign NCBI taxonomy id directly by
# findTaxonomyID.out = 10090 # mouse 10090, human 9606 etc.
STRINGdb_geneList.out <- STRINGdb_geneList() #convert gene lists
input_STRINGdbGO <- 'Process'   #'Process', 'Component', 'Function', 'KEGG', 'Pfam', 'InterPro'
stringDB_GO_enrichmentData()

# PPI network retrieval and analysis
input_nGenesPPI <- 100   #Number of top genes for PPI retrieval and analysis
stringDB_network1(1) #Show PPI network
write(stringDB_network_link(), 'PPI_results.html') # write results to html file
browseURL('PPI_results.html') # open in browser

###########################
# 8. Pathway analysis
###########################
input_selectContrast1 <- 'SCNAph-SCN'     #select Comparison
#input_selectContrast1 = limma.out$comparisons[3] # manually set
input_selectGO <- 'KEGG'         #Gene set category
#input_selectGO='custom' # if custom gmt file
input_minSetSize <- 15    #Min size for gene set
input_maxSetSize <- 2000          #Max size for gene set
# Read pathway data again
GeneSets.out <-readGeneSets( geneSetFile,
  convertedData.out, input_selectGO,input_selectOrg,
  c(input_minSetSize, input_maxSetSize)  )
input_pathwayPvalCutoff <- 0.2    #FDR cutoff
input_nPathwayShow <- 30          #Top pathways to show
input_absoluteFold <- FALSE       #Use absolute values of fold-change?
input_GenePvalCutoff <- 1         #FDR to remove genes

input_pathwayMethod = 1  # 1  GAGE
gagePathwayData.out <- gagePathwayData()  # pathway analysis using GAGE
gagePathwayData.out
 pathwayListData.out = pathwayListData()
enrichmentPlot(pathwayListData.out, 25  )
enrichmentNetwork(pathwayListData.out )
enrichmentNetworkPlotly(pathwayListData.out)
```

```
input_pathwayMethod = 3  # 1  fgsea
fgseaPathwayData.out <- fgseaPathwayData() #Pathway analysis using fgsea
fgseaPathwayData.out
pathwayListData.out = pathwayListData()
enrichmentPlot(pathwayListData.out, 25  )
enrichmentNetwork(pathwayListData.out )
enrichmentNetworkPlotly(pathwayListData.out)

PGSEAplot() # pathway analysis using PGSEA


###########################
# 9. Co-expression network
###########################
 input_mySoftPower <- 5  #SoftPower to cutoff
 input_nGenesNetwork <- 1000      #Number of top genes
 input_minModuleSize <- 20        #Module size minimum
 wgcna.out = wgcna()   # run WGCNA
 softPower() # soft power curve
 modulePlot() # plot modules
 listWGCNA.Modules.out = listWGCNA.Modules() #modules

 input_selectGO5 <- 'GOBP'          #Gene set
 # Read pathway data again
 GeneSets.out <-readGeneSets( geneSetFile,
   convertedData.out, input_selectGO5,input_selectOrg,
   c(input_minSetSize, input_maxSetSize)  )
 input_selectWGCNA.Module <- 'Entire network'      #Select a module
 input_topGenesNetwork <- 15       #SoftPower to cutoff
 input_edgeThreshold <- 0.4          #Number of top genes
 moduleNetwork()           # show network of top genes in selected module

 input_removeRedudantSets <- TRUE         #Remove redundant gene sets
 networkModuleGO()        # Enrichment analysis of selected module
```