QUANTIFYING THE *GLYCINE MAX*
PROXIMAL *CIS*–REGULOME DURING PATHOGENESIS

by

Parsa Hosseini
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
In Partial fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Bioinformatics and Computational Biology

Committee:

_____	Dr. Mohsin S. Jafri, Dissertation Chair
_____	Dr. Benjamin F. Matthews, Dissertation Director
_____	Dr. Patrick Gillevet, Committee Member
_____	Dr. James D. Willett, Committee Member
_____	Dr. Ivan Ovcharenko, Committee Member
_____	Dr. James D. Willett, Director, School of Sytems Biology
_____	Dr. Rick Diecchio, Associate Dean, Student and Academic Affairs, College of Science
_____	Dr. Peggy Agouris, Interim Dean, College of Science
Date: _____	Fall Semester 2013 George Mason University Fairfax, VA

Quantifying the *Glycine max* Proximal *Cis*–regulome during Pathogenesis

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Parsa Hosseini
Master of Science
Towson University, 2010
Bachelor of Science
University of Maryland, 2007

Director: Dr. Benjamin F. Matthews

Fall Semester 2013
George Mason University
Fairfax, VA

Dedication

To Heather, thank you for everything you have done to make this aspiration a reality. Through thick and thin, you have stood beside me flawlessly and supported my every academic venture without hesitation. You put your dreams on hold to help me see this journey to its end, and for that alone, I am eternally grateful. I may never be able to thank you for all your contributions and support, but hopefully this dedication is a good place to start.

Acknowledgments

I wish to extend my appreciations to my dissertation research advisor, Dr. Benjamin F. Matthews. The years I spent in his lab at the United States Department of Agriculture – Soybean Genomics and Improvement Laboratory (USDA–SGIL) were by–far the best years of my academic career and am beyond grateful for his teachings and guidance.

My gratitude also goes out to Dr. Ivan Ovcharenko and his laboratory at the National Center for Biotechnology Information (NCBI), National Institutes of Health. Dr. Ovcharenko taught me principles of transcription regulation and am deeply appreciative to have worked under his guidance.

I am indebted to Dr. Patrick Gillevet for countless hours of insightful discussions. Thank you for advising me throughout this academic journey and pushing me to grow as a scientist. Thank you also to Dr. James Willett for teaching me the elegant world of systems biology which I have come to cherish. Much gratitude goes to my dissertation chair, Dr. Saleet Jafri. Thank you wholeheartedly for your guidance and helping me keep on–track with this pursuit. Without question, you were there when I had concerns and you made sure I saw this aspiration to its very end.

My gratitude goes out to all USDA–SGIL personnel, namely Dr. Perry Cregan, Dr. Brett Cooper, Dr. Mark Tucker, Dr. Marcial Pastor–Corrales, Dr. Savithiry Natarajan, Rita Walker, Margaret MacDonald, Eric Brewer, Reham Youssef, and Hunter Beard. Thank you for welcoming me to SGIL and embracing me like one of your own. Much acknowledgement is owed to former SGIL members whom I had the pleasure of frequently working with, namely Dr. Arianne Tremblay and Dr. Vincent P. Klink.

My gratitude also goes out to the Computational Biology Branch (CBB) of NCBI. CBB provided software and hardware resources to facilitate completion of numerous portions of this dissertation. For such assistance as well as the many CBB friends made along the way, thank you wholeheartedly.

Last but certainly not least, this dissertation would not be possible without the unconditional love and support from family, friends, co–workers, or anyone who touched my life in any way. Thank you for shaping my life, supporting my aspirations, and fostering an academically stimulating environment.

As silly as this may sound, my gratitude also goes out to our pug puppy, Chumpy. Thank you for your unconditional love and for always sitting by the door for me day after day. To be honest, I have always wondered what he would say if he was able to talk.

Much gratitude to the faculty and staff at George Mason University School of Systems Biology and the George Mason University Academic Common Market. Funding and support was provided by George Mason University and the USDA–SGIL.

Table of Contents

List of Tables

List of Figures

Abstract

QUANTIFYING THE *GLYCINE MAX* PROXIMAL *CIS*–REGULOME DURING PATHOGENESIS

Parsa Hosseini, PhD

George Mason University, 2013

Dissertation Director: Dr. Benjamin F. Matthews

Transcription regulation is a highly orchestrated dynamic which mediates every aspect of organismal development. Following host perception of positive or negative stress, hormone–driven signaling amplifies extracellular cues and triggers a multifaceted, exquisite array of downstream signaling cascades. These cascades go on to drive synthesis of small metabolites and regulatory proteins known as transcription factors which mediate transcription regulation. Transcription factors drive transcription expression by binding onto short non–coding genomic regions known as transcription factor binding sites. Additional regulatory proteins are recruited to collectively bind to a regulatory element, bringing about tightly regulated, tissue–specific gene expression.

With transcriptomic assays capable of quantifying cDNA at unprecedented levels of granularity and resolution, we can now quantify not only these regulatory elements but entire transcriptomes in a matter of hours. Novel questions can now be proposed, questions which necessitate utilization of these high–throughput platforms. Investigators can now build novel isoform models and ultimately get one step closer to filling in gaps sprinkled throughout the organismal systematic landscape.

These technologies, known by many as next-generation sequencing (NGS), encompass ultrafast, parallel sequencing assays. Their name is fitting for the intent: unprecedented nucleotide resolution with a dynamic range of expression. A popular NGS assay, RNA sequencing, or RNA–Seq for short, provides the ability to sequence an entire cDNA library at high–resolution. Unlike traditional microarrays, probe-sets are not required, making it possible to quantify novel transcript isoforms, identify structural variants, and SNPs.

Quantification of the soybean (*Glycine max*) root and leaf transcriptome is an active area of agriculture research, however, quantification of the soybean proximal regulome is still in its infancy. In this dissertation, we investigate the transcriptomic and regulatory interplay between both soybean roots and leaves and two major compatible pathogens. We present numerous potentially novel defense–response candidate genes as well as a regulatory signature that captures and models host–pathogen interplay in a biologically–sound manner.

Epilogue

"Most of an organism, most of the time, is developing from one pattern into another, rather than from homogeneity into a pattern. One would like to be able to follow this more general process mathematically also. The difficulties are, however, such that one cannot hope to have any very embracing theory of such processes, beyond the statement of the equations. It might be possible, however, to treat a few particular cases in detail with the aid of a digital computer."

– Alan Turing (1952)

"The chemical basis of morphogenesis."
Phil. Trans. R. Soc. Lond. B., **237**(641)

Chapter 1: Introduction

1.1 Motivation

With next–generation sequencing (NGS) technologies now a staple assay in today's molecular biology, we can now investigate biological phenomena and quantify systematic interplay with relative ease. NGS is not a singular technology but rather an umbrella–term to encapsulate parallel, dynamic, and high–resolution biological sequencing. Two popular NGS technologies include RNA sequencing (RNA–Seq) and Chromatin Immunoprecipitation sequencing (ChIP–Seq). As implied by the former assay, entire transcriptomes are sequenced, producing billions of short reads. These reads are subsequently mapped onto a reference index such as a transcriptome. Each feature in this index is then quantified, shedding light onto magnitude of differential expression. ChIP–Seq on the other hand quantifies protein–DNA binding rather than RNA transcription. Similar to RNA–Seq, ChIP–Seq also generates billions of short reads which get mapped to a reference index. Resultant mappings reveal insight into transcription factor (TF) binding affinity and likely TF binding sites (TFBSs). Advances in NGS chemistry and parallelism have democratized the financial barrier of entry; seemed like yesterday when NGS was exclusive solely to large, well–funded institutes. Even still, NGS costs continue to drop with no end in sight.

Virtually all scientific disciplines from plant biology to personalized medicine have benefited from NGS assays such as RNA–Seq or ChIP–Seq. One such field is soybean (*Glycine max*) genomics which utilizes NGS assays to decipher interplay between the host plant and a compatible pathogen. By quantifying the transcriptome of soybean leaves and roots upon pathogenesis, the commandeering nature of the pathogen can be quantified. Supplementing this transcriptional profile with quantification of regulatory elements therefore reveals systematic interplay between soybean and the pest.

In–closing, the motivation of this dissertation is to investigate the soybean transcriptome and proximal regulome during pathogenesis. The intent of such investigation is to answer biologically–relevant questions regarding plant–pathogen interplay and ultimately reveal novel insight into host transcription regulation in the face of a pest.

1.2 Plant hormone–driven signaling

Plant hormones, otherwise known as phytohormones, drive virtually all aspects of plant development, from seed germination to floral development. Due to the sessile and immobile nature of plants, plants must effectively adapt to an ever–changing range of external stimuli encompassing numerous biotic and abiotic stressors. To mediate such adaptations, plants utilize an interconnected, exquisite collection of phytohormones. Among the first identified phytohormones were ethylene (ETH), auxin (indole-3-acetic acid; IAA), gibberellin (GA) and abscisic acid (ABA). As of lately, numerous phytohormones have joined these ranks, namely salicylic acid (SA) and jasmonic acid (JA). All together, phytohormone cross–talk mediates systematic interplay so as to enable plant development in the face of positive or negative stress.

1.3 Core pathways involved in defense–response

Pathway databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [1] and MetaCyc [2] contain well annotated signaling pathways, allowing for effortless tracking of desired metabolites and enzymes from one pathway to the next. Well–studied pathways such as phenylalanine biosynthesis, flavonoid biosynthesis, isoflavonoid biosynthesis, and phenylpropanoid biosynthesis have been shown to orchestrate synthesis of numerous phytohormones and just as many secondary metabolites. Utilizing KEGG and MetaCyc tool–sets can therefore reveal insight into the roles defense response metabolites play throughout the stress response landscape.

1.3.1 Phenylalanine, tyrosine and tryptophan biosynthesis

Beginning with the glycolysis derivative, phosphoenolpyruvate (PEP), this very molecule ultimately serves as the precursor to phenylalanine, tyrosine, and tryptophan (Figure 1.1). A critical aspect of this pathway is the synthesis of shikimate, a seven–carbon carboxylic acid (KEGG: C00493) from 3–Dehydroshikimate. A phosphate group is subsequently attached by shikimate kinase (EC: 2.7.1.71) and catalyzed by chorismate synthase (EC: 4.2.3.5).

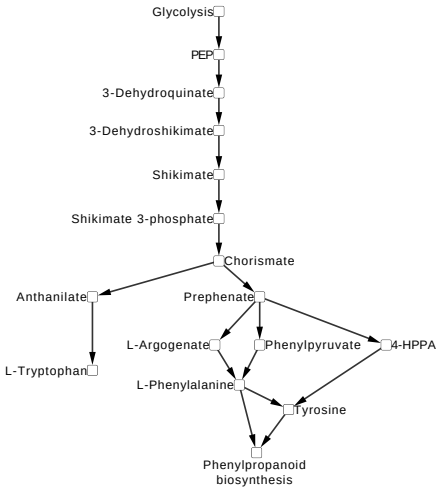


Figure 1.1: Phenylalanine and tyrosine biosynthesis.

Chorismate, a ten–carbon aromatic is subsequently synthesized and forks into two independent branch-points which ultimately generate three amino acids [3]. The first branch–point leads to synthesis of L–tryptophan, while the second branch–point leads to synthesis of tyrosine and phenylalanine. What makes chorismate such a critical metabolite is its ability to undergo modification of its hydroxyl group and become isochorismate. With the help of isochorismate pyruvate lyase (EC: 4.2.99.21), isochorismate becomes salicylic acid (SA), a classical phytohormone known for its defense–response signaling capabilities [4]. Besides orchestrating synthesis of SA, chorismate regulates synthesis of indoleacetate (IAA; auxin), a well–studied phytohormone involved in defense response [5].

1.3.2 Phenylpropanoid biosynthesis

Metabolites generated from phenylalanine, tyrosine and tryptophan biosynthesis serve as the entry–point for synthesis of phenylpropanoids. These metabolites are synthesized from phenylalanine and go on to assist in regulation of many aspects of plant development [6–8]. Phenylpropanoid biosynthesis begins with phenylalanine ammonia lyase (PAL; EC: 4.3.1.24) removing an NH_2 group from L–phenylalanine (Figure 1.2a). The resultant compound, cinnamic acid (cinnamate) is subsequently processed by 4-coumarate–CoA ligase (EC: 6.2.1.12) which converts this nine–carbon phenylalanine derivative into cinnamoyl–CoA.

(a) Flavonoid biosynthesis precursors.

(b) Biosynthesis of major flavonoids.

Figure 1.2: Phenylpropanoid and flavonoid biosynthesis.

1.3.3 Flavonoid biosynthesis

Similar to phenylpropanoids, flavonoids are secondary metabolites which execute a diverse set of biological roles ranging from defense response to floral pigmentation [9, 10]. Their antioxidative properties thus make this metabolite a beneficial human supplement. The entry–point to flavonoid biosynthesis is p–Coumaroyl–CoA, a metabolite generated from phenylpropanoid biosynthesis (Figure 1.2b). Chalcone synthase (CHS; EC: 2.3.1.74) converts p–Coumaroyl–CoA to naringenin chalcone. CHS provides the ability to generate

chalcones: metabolites well known for their defense–response roles [11]. From here, chalcone isomerase (CHI; EC: 5.5.1.6) converts naringenin chalcone to naringenin which then goes on to synthesize the isoflavonoid metabolite, genistein.

1.3.4 Aspartate and methionine biosynthesis

The amino acid aspartate is synthesized by the Krebs cycle by–product oxaloacetate [12]. What makes aspartate an interesting compound is its ability to contribute to the synthesis of fellow amino acids threonine, lysine and methionine [13]. With respect to methionine biosynthesis, the first synthesis step involves conversion of aspartate to yield the methionine precursor, *O*–phosphohomoserine (OPS). OPS is converted into cystathionine, a seven–carbon compound which is subsequently converted to yield homocysteine [14, 15]. With the help of S-methyltransferase (EC: 2.1.1.14), a CH_3 methyl group is added to homocysteine resulting in methionine (Figure 1.3). From here on, methionine can serve as the precursor for ETH biosynthesis, a critical phytohormone responsible for a diverse range of physiological processes [16, 17]. To begin this conversion, S–adenosylmethionine synthetase (EC: 2.5.1.6) converts methionine to S–adenosylmethionine (SAM) which is subsequently converted to the ethylene precursor 1–aminocyclopropane–1–carboxylate (ACC) [18–20]. ACC is subsequently converted to ethylene using ACC oxidase (EC: 1.14.17.4).

1.3.5 α–linolenic acid and jasmonic acid biosynthesis

Fatty acids (FAs) such as α–linolenic acid are hydrocarbon compounds with a carboxylic acid tail. Such compounds play fundamental roles in not just plant defense but all throughout the organismal landscape, from signaling to defense response. The shorthand FA naming convention entails counting from the carboxyl terminal carbon the number of carbon atoms as well as the number of double bonds. Counts are then delimited with a colon. For instance, α–linolenic acid contains 18 carbon atoms and 3 carbon double bonds, hence the derived shorthand name of 18:3 (Figure 1.4). Often, double bond indices may also be written in conjunction to the shorthand convention. In the case of α–linolenic acid, the revised

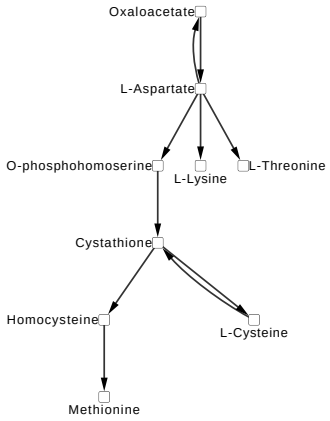


Figure 1.3: Biosynthesis of methionine from aspartate.

convention would be 9,12,15–18:3. Polyunsaturated FAs can be divided into classes known as omega (ω). Beginning at the opposite end of the FA, omega pertains to the carbon at which the first double bond occurs. As is the case of α–linolenic acid, the first double bond is found at carbon 3, thus classifying this FA as an ω–3 FA.

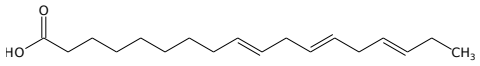


Figure 1.4: α–linolenic acid ($C_{18}H_{30}O_2$).

The FA α–linolenic acid is particularly important within the systematic defense response context. As a matter of fact, this FA is a precursor for numerous jasmonates, secondary metabolites inextricably linked with plant defense response. The first phase in jasmonate synthesis utilizes lipoxygenase (EC: 1.13.11.12) to convert α–linolenic acid to 13(S)–hydroperoxy linolenic acid (Figure 1.5). Allene oxidase synthase (AOS) slightly modifies the hydroxyl group of this product to yield 12,13(S)–epoxylinolenic acid [21]. Allene oxidase cyclase (AOC) converts this 18–carbon product to generate 12–OPDA, subsequently followed by OPC–8. OPC–8 then undergoes β–oxidation, producing JA–CoA. Amazingly,

JA–CoA signals biosynthesis of fellow JA metabolites, namely jasmonate, methyl–jasmonate (Me–JA), and jasmonate–isoleucine (JA–Ile) [22]. Jasmonates are tried–and–true secondary metabolites actively involved in biotic stress response and capable of rapid induction following herbivory or wound perception [23–25].

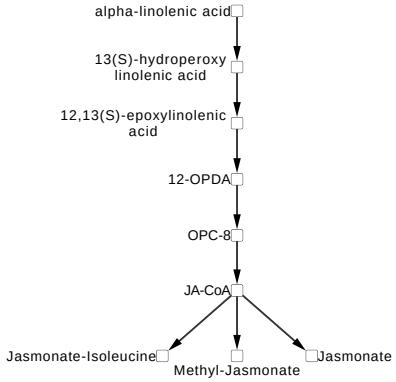


Figure 1.5: Biosynthesis of jasmonates.

1.4 Organization of the Dissertation

This dissertation is solely comprised of manuscripts which are either accepted, in–review or submitted for publication in peer–reviewed journals.

At the time of this writing, chapter 2 has appeared in print, while chapters 3, 4, 5, 6 have been submitted or are in–review. Chapters 2, 3, 4, 5, and 6 have been reformatted in this dissertation. All authors affiliated with their respective manuscripts have read and approved such works.

Chapter 2

P. Hosseini, I. Ovcharenko, and B. F. Matthews. Using an ensemble of statistical metrics to quantify large sets of plant transcription factor binding sites. *Plant Methods*, **9**(12), 2013.

Chapter 3

P. Hosseini and B. F. Matthews. Regulatory interplay between soybean root and soybean cyst nematode during a resistant and susceptible reaction. *Submitted.*

Chapter 4

P. Hosseini, A. Tremblay, D. Huang, I. Ovcharenko, and B. F. Matthews. Quantifying *Glycine max* proximal regulatory elements during a *Phakopsora pachyrhizi* reaction. *In–review.*

Chapter 5

P. Hosseini, R. Youssef, and B. F. Matthews. Phytohormone treated *Glycine max* roots capture host defense response and proximal binding site landscape. *Submitted.*

Chapter 6

P. Hosseini and B. F. Matthews. The *Glycine max* proximal regulatory element landscape sheds light on host–pathogen defense response interplay. *In–review.*

Chapter 2: Over–represented TFBSs during SR pathogenesis

2.1 Background

2.1.1 Abstract

Background

From initial seed germination through reproduction, plants continuously reprogram their transcriptional repertoire to facilitate growth and development. This dynamic is mediated by a diverse but inextricably-linked catalog of regulatory proteins called transcription factors (TFs). Statistically quantifying TF binding site (TFBS) abundance in promoters of differentially expressed genes can be used to identify binding site patterns in promoters that are closely related to stress-response. Output from today's transcriptomic assays necessitates statistically-oriented software to handle large promoter-sequence sets in a computationally tractable fashion.

Results

We present Marina, an open-source software for identifying over-represented TFBSs from amongst large sets of promoter sequences, using an ensemble of 7 statistical metrics and binding-site profiles. Through software comparison, we show that Marina can identify considerably more over-represented plant TFBSs compared to a popular software alternative.

Conclusions

Marina was used to identify over-represented TFBSs in a two time–point RNA-Seq study exploring the transcriptomic interplay between soybean (*Glycine max*) and soybean rust

(*Phakopsora pachyrhizi*). Marina identified numerous abundant TFBSs recognized by transcription factors that are associated with defense-response such as WRKY, HY5 and MYB2. Comparing results from Marina to that of a popular software alternative suggests that regardless of the number of promoter-sequences, Marina is able to identify significantly more over-represented TFBSs.

2.1.2 Definitions and Presumptions

We define a list of transcription factor binding sites (TFBSs), $t_1, t_2, \ldots, t_N$, where t_i is either a DNA motif, m_i or position weight matrix (PWM), w_i. The former is a variable-length character-string from the four-nucleotide DNA alphabet, while the latter is a two-dimensional matrix of preset weights.

A group, G_a, is a FASTA file populated with user-provided promoter sequences. Let $G_a, G_{a+1}, \ldots, G_N$ represent a list of N groups such that $N \geq 2$. We define a contingency matrix, c_i, as a 2×2 matrix, used to model t_i over-representation across G_a and G_{a+1}. A set of statistical metrics, S, quantify degree of t_i over-representation given c_i.

2.1.3 Transcription factors and binding site representation

Plants are constantly surrounded by stimulus, be-they deleterious pathogens or positive stimuli such as light and nutrients. In order for the plant to respond to these signals, plants must utilize regulatory proteins known as transcription factors (TFs) to facilitate transcriptional reprogramming in a dynamic, tissue-dependent manner. These proteins bind to enhancer or promoter *cis*-elements and facilitate the recruitment of RNA polymerase II. This combinatorial binding of TFs facilitates downstream execution of adaptative signals in the face of drought, herbivory, and high salinity. By quantifying binding–sites for these regulatory proteins, inherent transcriptional dynamics and magnitude of over-representation can be inferred.

TFs are classified into families by inherent DNA-binding signatures otherwise known as protein domains. In *Arabidopsis thaliana*, for instance, there are 64 known TF families[26],

and it is not uncommon for different TF family members to exhibit relatively similar functionality. This redundancy is especially true when it comes to stress-response [27–29].

DNA motifs and PWMs are two models frequently used to define a TFBS. The former is a short *cis*–element region presumed to be a TFBS, while the latter models nucleotide propensities of a TFBS in the form of a matrix[30, 31]. PWMs have been used across a broad spectrum of plant investigations such as identification of conserved exonic splice-site enhancers in *Arabidopsis thaliana* [32], prediction of potential seed-storage regulatory elements in mustards, grasses and legumes [33], and identification of novel regulatory elements in *Arabidopsis thaliana* [34]. With assays such as ChIP-ChIP and ChIP-Seq, novel regulatory elements can be identified and modeled as a PWM[35].

2.2 Implementation

Marina is an operating-system independent GUI software tool built using the Java programming language. This manuscript builds on the works of Chekmenev et al. [36], Loots et al.[37], and Kel et al.[38], by implementing multiple statistical metrics to identify the maximum number of biologically-sound TFBSs, while accounting for cases when large promoter sets are provided.

To begin analysis with Marina, at least two FASTA files populated with user-provided promoter sequences are required. Each FASTA file is known as a group. A group, for instance, could represent promoter sequences of interest for a particular condition or time–point. The Marina workflow (Figure 2.1) is partitioned into three distinct phases. The first phase performs abundance-estimation given a catalog of known TFBS models (Figure 2.1a). Initial abundance derivation is performed by mapping TFBS models onto user-provided promoter sequences. Subsequently, low-quality TFBSs are removed (Figure 2.1b). Finally, statistical metrics quantify and rank TFBS over-representation (Figure 2.1c).

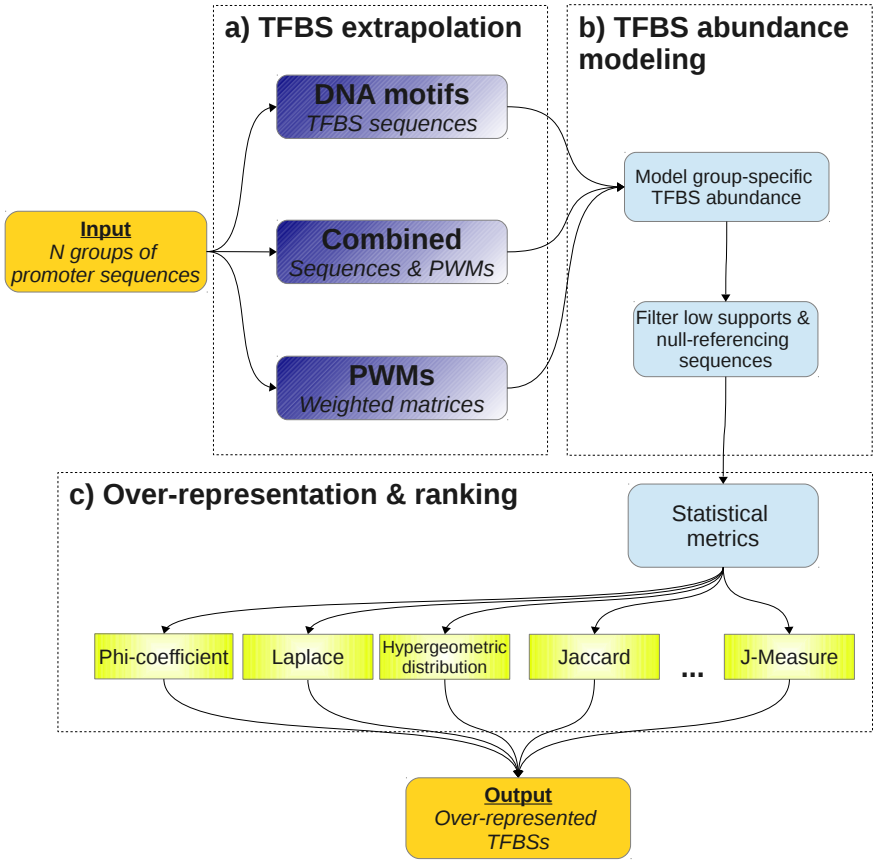


Figure 2.1: High–level overview and flow of the Marina algorithm.

2.2.1 Binding site mapping

In order to effectively quantify TFBS abundance using this tool, TFBS models must be provided. These models are in the form of either DNA motifs or PWMs. Cumulatively, 1,240 TFBS models were mined and utilized throughout this study. Of these models, 1,160 were DNA motifs with the remaining 80 being PWMs; motif-to-PWM ratio of 13:1. Plant DNA motif and PWM models originated from AthaMap[39], AGRIS[40], PlantCARE[41], TRANSFAC[42], and JASPAR[43]. DNA motifs and PWMs were stored in either a tab-delimited or FASTA file format, respectively. Due to licensing restrictions, Marina does not come pre-packaged with a catalog of TFBS models, however several PWMs are provided,

built from known PDB structures using the 3DTF web-server[44]. Be it PWMs or DNA motifs, a user-friendly schema is provided for importing custom TFBS profiles.

DNA motif and PWM mapping

To efficiently derive over-representation using DNA motifs, Marina scans promoter sequences for any occurrence of this motif using the Boyer-Moore-Horspool algorithm[45]. Due to the short length of many DNA motifs, elements such as ARF1 (TGTCTC)[46] may ubiquitously map throughout a promoter sequence with many mappings having little biological significance. Though this tool provides the option to filter short-length models be it PWMs or DNA motifs, resultant abundance estimations may seldom be biologically significant simply due to the likelihood of spurious mappings.

Marina maps each PWM onto promoter sequences using a concurrent implementation of the P–MATCH algorithm[36]. P–MATCH calculates a likelihood that a particular candidate promoter region contains a TFBS. By default, Marina uses a probability-cutoff of 0.80; any sub-sequence with a score greater than this cutoff is rendered a potential TFBS. Alongside DNA motif and PWM extrapolations is a third pseudo-extrapolation known as combined mode. This mode simply performs the two former extrapolations back-to-back, merging their results into a singular data-structure. Combined mode capitalizes on the abundance of DNA motifs and probabilistic power of PWMs.

2.2.2 Modeling TFBS over-representation

TFBS abundances across all promoter sequences are modeled using a group-specific acyclic graph. Each graph is organized such that group name is the root-node and each TFBS is a child leaf node. Every TFBS node references a list of promoter sequences containing this TFBS. Per graph child node, two measures are used to model initial TFBS abundance: raw counts and support[47]. The former is simply defined as the number of promoter sequences which contain this particular TFBS. Raw counts are a useful, comparable metric if all groups have approximately the same number of promoter sequences. Unfortunately

some groups may be larger than others, resulting in skewed and uncontrastable counts. To circumvent this possibility, the latter probabilistic measure, support, comes in helpful. Support, $P(t_i, G_a)$, is a data-mining metric for representing abundance of a TFBS within a particular group. A collection of statistical metrics continue where support leaves off, providing a means of deducing TFBS abundance. Both raw-counts and support serve as viable metrics to initially model TFBS abundance, however there may be cases were a rift between the two measures can appear. For example, suppose a single TFBS mapped only once to a group. Due to such minimal mapping, raw-count will be small but support would be large. Both low-support and low-count thresholds exist so as to filter corresponding graph nodes. Such graph trimming ensures that high-support and/or high-count TFBS nodes remain as they are more likely of having correlations to a particular group[48]. A caveat with threshold cutoffs is that low-abundance TFBSs will get discarded.

2.2.3 Derivation of TFBS over–representation

Given remaining TFBSs nodes, Marina aims to deduce magnitude of over-representation per TFBS, t_i, by contrasting its abundance across groups G_a and G_{a+1}. To facilitate this objective, a collection of 7 knowledge discovery metrics, S, are implemented (Table 2.1). Though a single metric can theoretically suffice, employing the entire set provides a means to appreciate unique features per measure and avoid individual bias. This table is by no means exhaustive as there are well over 20 frequently used metrics [49, 50]. The metrics in this table were selected so that there exists a sound mixture of both well-studied association and correlation measures.

In order to utilize such measures, TFBS abundances must be modeled in a suitable data-structure. A contingency matrix, c_i, is an ideal data-structure candidate as it models TFBS distributions throughout multiple, independent groups (Table 2.2). Each metric within S processes frequencies within a contingency matrix, c_i, so as to quantitatively deduce over-representation of TFBS, t_i. Certainly not all metrics deduce magnitude of TFBS over-representation the same, resulting in difficulties as to which TFBSs are unanimously

Table 2.1: Statistical metrics utilized throughout Marina.

Metric	Equation	Output range	Ref.
Confidence (CF)	$max(P(G_a\|t_i), P(t_i\|G_a))$	$0 \dots 1$	[51]
Cosine (CO)	$\frac{P(t_i,G_a)}{\sqrt{P(t_i)P(G_a)}}$	$0 \dots \sqrt{P(t_i,G_a)} \dots 1$	[52]
Jaccard (JAC)	$\frac{P(t_i,G_a)}{P(t_i)+P(G_a)-P(t_i,G_a)}$	$0 \dots 1$	[53]
Kappa coefficient (K)	$\frac{P(t_i,G_a)+P(\overline{t_i},\overline{G_a})-P(t_i)P(G_a)-P(\overline{t_i})P(\overline{G_a}))}{1-P(t_i)P(G_a)-P(\overline{t_i})P(\overline{G_a})}$	$-1 \dots 1$	[54]
Laplace (LP)	$max\left(\frac{NP(t_i,G_a)+1}{NP(t_i)+2}, \frac{NP(t_i,G_a)+1}{NP(G_a)+2}\right)$	$0 \dots 1$	[55]
Lift (LI)	$\frac{P(t_i,G_a)}{P(t_i)P(G_a)}$	$0 \dots \infty$	[56]
Phi coefficient (PHI)	$\frac{P(t_i,G_a)-P(t_i)P(G_a)}{\sqrt{P(t_i)P(G_a)(1-P(t_i))(1-P(G_a))}}$	$-1 \dots 1$	[57]

most over-represented by all metrics. A solution to bringing uniform over-representation agreement across all metrics is to standardize contingency matrix counts using Iterative Proportional Fitting (IPF) [58].

Table 2.2: A 2×2 contingency matrix.

	G_a	$\overline{G_a}$	
t_i	$c_i(0,0)$	$c_i(1,0)$	$n(t_i)$
$\overline{t_i}$	$c_i(0,1)$	$c_i(1,1)$	$n(\overline{t_i})$
	$n(G_a)$	$n(\overline{G_a})$	N

Iterative Proportional Fitting (IPF)

IPF is an algorithm for standardizing counts in a two-dimensional contingency matrix such that matrix row and column marginals are equal to one another (Table 2.3). Through such adjustment, inherent associations and correlations can be discovered [59]. By performing IPF-standardization, output for all 7 metrics become normalized so as to agree which TFBSs are the most over-represented.

Table 2.3: IPF–normalization adjusts counts in a contingency matrix.

	G_a	$\overline{G_a}$	
t_i	x	$N/2 - x$	$N/2$
$\overline{t_i}$	$N/2 - x$	x	$N/2$
	$N/2$	$N/2$	N

Equations 2.1 and 2.2 present an implementation of the IPF algorithm originally outlined by Tan et al.[60]. The former equation adjusts counts, a, such that they are equal on the diagonal axis. The latter equation then subtracts the remainder of the counts from that of the entire matrix sum, N.

$$c_{i1,1} = c_{i0,0} = a = \frac{N\sqrt{c_{i1,1}c_{i0,0}}}{2\left(\sqrt{c_{i1,1}c_{i0,0}} + \sqrt{c_{i1,0}c_{i0,1}}\right)} \tag{2.1}$$

$$c_{i0,1} = c_{i1,0} = \frac{N}{2} - a \tag{2.2}$$

2.3 Results and Discussion

2.3.1 Case study: Over-represented TFBSs during SR inoculation

To evaluate the functionality of this software tool, we utilized a two time-course RNA-Seq study that investigates soybean (*Glycine max*) transcriptional dynamics upon pathogenesis with soybean rust (SR; *Phakopsora pachyrhizi*). As outlined in our previous study, susceptible Williams 82 soybean leaves were inoculated with SR and assayed using RNA-Seq 10 days after inoculation (dai) [61]. An accompanying uninoculated control was also assayed to serve as a baseline condition. In both the control and 10 dai samples, a total of 5,940,995 70bp reads and 5,574,892 40bp reads were respectively sequenced using the Illumina platform (GenomeAnalyzer IIx). Sequenced reads were deposited in NCBI SRA under accessions SRX100854, SRX129967 and SRX100853, SRX129959, respectively. Per

run, quality assessment and control (QA/QC) entailed removal of low quality reads and trimming of low-quality 3' ends should its quality score be less than 22. Reads were also discarded if they had at most one nucleotide mismatch to either the human genome (Hg19) or the JCVI Microbial Resource [62]. Upon QA/QC completion, a total of 5,015,459 control reads and 5,420,745 10 dai reads passed filtering; quality-scores of 27 and 30, respectively. For each time–point, reads were mapped with at-most 3 nucleotide mismatches onto the soybean transcriptome (Glyma 1.0) using BWA[63]. Custom Python scripts inferred differential expression by deriving RPKM[64] and $log_2\left(\frac{RPKM_{10dai}}{RPKM_{0dai}}\right)$ per transcript.

Two gene-sets were then declared to contain the top 600 induced and 600 suppressed differentially expressed genes (DEGs), respectively. Per gene set, the promoter sequence 2.5kb upstream from each genes transcription start site (TSS) was retrieved and appended to a FASTA file. Both FASTA files in-conjunction with 80 plant PWMs and 1,160 plant-specific DNA motifs served as input into Marina. Using default settings, Marina identified 71 over-represented TFBSs given the control group (promoters of suppressed transcripts, G_S) and the query group (promoters of induced transcripts, G_I) (Table 2.4).

Table 2.4: Over–represented TFBSs without IPF.

	Metrics								Count	
TFBS	LP	CO	JAC	LI	CF	K	PHI	p–value	G_S	G_I
ABF1	20	39	39	20	20	3	2	8.21e-274	130	169
ABFS	9	9	10	9	9	16	12	2.38e-31	10	20
ABI3/FUS3	67	19	17	67	67	41	58	3.03e-47	14	7
ABI4(2)	64	34	33	64	64	64	67	4.46e-172	66	43
AG	14	20	21	14	14	13	18	4.61e-82	30	42
AGP1	48	57	56	48	48	58	49	2.41e-720	427	398
ALFIN1	34	58	57	34	34	34	34	1.58e-731	440	426
ARF1	65	29	24	65	65	57	62	1.24e-113	40	25

ARR10	39	65	65	39	39	43	39	1.83e-895	579	552
ARR2	69	27	22	69	69	60	69	4.02e-99	33	15
ATHB-5	43	68	68	43	43	49	43	1.54e-901	584	555
ATHB1	40	67	67	40	40	45	40	3.16e-901	584	556
ATHB5-1	63	21	20	63	63	44	55	3.20e-78	26	18
ATHB5-2	37	60	60	37	37	37	37	9.77e-769	470	452
ATHB6	27	23	25	27	27	29	32	3.06e-109	41	46
ATHB9	53	38	36	53	53	55	52	2.10e-225	95	81
AtLEC2	55	51	51	55	55	68	61	1.06e-611	336	284
ATML1/PDF2	71	18	11	71	71	54	71	8.73e-38	10	1
AtMYB2	29	33	34	29	29	23	31	1.60e-170	70	76
AtMYB77	60	32	31	60	60	56	57	2.95e-141	53	40
AtMYC2	2	2	2	2	2	30	8	0.00027	1	7
AtSPL3	30	45	46	30	30	8	26	7.99e-426	220	236
BLR/RPL/PNY	35	61	61	35	35	35	35	1.44e-777	478	462
bZIP910(2)	10	12	16	10	10	14	11	6.06e-42	14	26
bZIP911	12	11	13	12	12	19	14	4.35e-37	12	21
bZIP911(1)	11	10	12	11	11	20	13	2.52e-34	11	20
bZIP911(2)	18	13	14	16	16	32	29	3.73e-38	12	16
CBF	43	68	68	43	43	49	43	1.54e-901	584	555
CDC5	4	4	4	4	4	18	3	1.34e-10	3	13
DOF2	42	71	71	42	42	48	42	1.25e-902	585	556
DPBF1/2	51	55	55	51	51	66	54	1.85e-712	418	379
E2Fa	70	13	9	70	70	38	64	8.05e-24	6	1
E2Fc/d	1	1	1	1	1	26	5	0.0003	1	8
EmBP-1	25	43	43	25	25	5	17	3.31e-397	203	228
GAMYB	47	59	59	47	47	53	47	2.04e-743	447	422
Gamyb	58	28	26	58	58	40	50	6.18e-120	44	36

GATA-1	17	24	28	18	18	12	16	5.92e-120	47	62
GATA-1/2/3/4	16	15	18	17	17	28	27	9.29e-54	18	24
GT-3b	13	25	29	13	13	7	7	1.24e-128	52	76
HAHB4	46	64	64	46	46	52	46	1.03e-891	575	546
HAT5	43	68	68	43	43	49	43	1.54e-901	584	555
HSE	19	26	30	19	19	11	15	1.64e-130	52	68
HVH21	41	66	66	41	41	46	41	3.88e-900	583	555
HY5	6	8	8	6	6	21	10	6.24e-20	6	15
ID1	28	31	32	28	28	27	33	4.01e-146	58	63
MYB.PH3(1)	56	41	41	56	56	61	56	4.15e-333	154	130
MYB.PH3(2)	52	49	49	52	52	62	53	6.93e-564	306	276
MYB98	62	36	35	62	62	65	65	2.64e-210	85	60
O2	33	56	58	33	33	6	28	2.96e-731	446	457
OsbHLH66	26	40	40	26	26	9	20	1.72e-308	147	165
OsCBT	3	3	3	3	3	24	6	1.54e-7	2	10
P	57	52	52	57	57	71	66	2.57e-629	347	286
PCF2	61	47	44	61	61	70	70	3.56e-441	215	160
PCF5	59	48	48	59	59	69	68	2.61e-498	254	201
PEND	31	35	37	31	31	15	30	3.82e-230	101	108
PIF3(2)	21	22	23	21	21	17	25	4.17e-99	37	46
RAP2.2	66	30	27	66	66	59	63	1.61e-125	45	28
RAV1(1)	49	54	53	49	49	63	51	1.95e-688	400	366
RAV1(2)	38	62	62	38	38	39	38	1.07e-854	543	519
STF1	24	37	38	24	24	10	22	1.24e-242	109	124
TAC1	68	17	15	68	68	42	59	1.47e-44	13	6
TaMYB80	54	50	50	54	54	67	60	2.7e-594	324	276
TBP	36	63	63	36	36	36	36	1.54e-881	568	547
TEIL	50	42	42	50	50	47	48	8.45e-340	160	146

TGA1	23	46	47	23	23	2	9	3.41e-468	253	293
TGA1a	32	53	54	32	32	4	23	3.32e-688	413	433
WRKY11	7	7	7	8	8	31	19	2.34e-14	4	9
WRKY18/40/62	7	6	6	7	7	33	21	2.87e-11	3	7
WRKY26/38/43	15	16	19	15	15	25	24	4.16e-56	19	26
WRKY6	5	5	5	5	5	22	4	1.09e-10	3	12
ZAP1	22	44	45	22	22	1	1	2.46e-415	219	268

As shown in Table 2.4, there exists no consensus amongst the various metrics as to which TFBS is truly the most over-represented. There are however some TFBSs that are ranked by all metrics in a relatively uniform manner: AG, ATHB6, and ABFS. For all other TFBSs, it is difficult to deduce magnitude of over-representation. Such a scenario warrants IPF-standardization as it normalizes metric-ranks to agree in-concert which TFBSs are the most over-represented (Table 2.5). By visually contrasting this table with that of Table 2.4, it is clear that unstandardized ranks from Laplace Correction (LP), Confidence (CF) and Lift (LI) perfectly equal their IPF-standardized counterpart.

Table 2.5: IPF identifies over–represented TFBSs.

	Metrics						
TF	LP	CO	JAC	LI	CF	K	PHI
ABF1	20	20	20	20	20	20	20
ABFS	9	9	9	9	9	9	9
ABI3/FUS3	67	67	67	67	67	67	67
ABI4(2)	64	64	64	64	64	64	64
AG	14	14	14	14	14	14	14
AGP1	48	48	48	48	48	48	48
ALFIN1	34	34	34	34	34	34	34
ARF1	65	65	65	65	65	65	65

ARR10	39	39	39	39	39	39	39
ARR2	69	69	69	69	69	69	69
ATHB-5	43	43	43	43	43	43	43
ATHB1	40	40	40	40	40	40	40
ATHB5-1	63	63	63	63	63	63	63
ATHB5-2	37	37	37	37	37	37	37
ATHB6	27	27	27	27	27	27	27
ATHB9	53	53	53	53	53	53	53
AtLEC2	56	56	56	56	56	56	56
ATML1/PDF2	71	71	71	71	71	71	71
AtMYB2	29	29	29	29	29	29	29
AtMYB77	60	60	60	60	60	60	60
AtMYC2	2	2	2	2	2	2	2
AtSPL3	30	30	30	30	30	30	30
BLR/RPL/PNY	35	35	35	35	35	35	35
bZIP910(2)	10	10	10	10	10	10	10
bZIP911	12	12	12	12	12	12	12
bZIP911(1)	11	11	11	11	11	11	11
bZIP911(2)	17	17	17	17	17	17	17
CBF	43	43	43	43	43	43	43
CDC5	4	4	4	4	4	4	4
DOF2	42	42	42	42	42	42	42
DPBF1/2	51	51	51	51	51	51	51
E2Fa	70	70	70	70	70	70	70
E2Fc/d	1	1	1	1	1	1	1
EmBP-1	25	25	25	25	25	25	25
GAMYB	47	47	47	47	47	47	47
Gamyb	58	58	58	58	58	58	58

GATA-1	18	18	18	18	18	18	18
GATA-1/2/3/4	16	16	16	16	16	16	16
GT-3b	13	13	13	13	13	13	13
HAHB4	46	46	46	46	46	46	46
HAT5	43	43	43	43	43	43	43
HSE	19	19	19	19	19	19	19
HVH21	41	41	41	41	41	41	41
HY5	6	6	6	6	6	6	6
ID1	28	28	28	28	28	28	28
MYB.PH3(1)	55	55	55	55	55	55	55
MYB.PH3(2)	52	52	52	52	52	52	52
MYB98	62	62	62	62	62	62	62
O2	33	33	33	33	33	33	33
OsbHLH66	26	26	26	26	26	26	26
OsCBT	3	3	3	3	3	3	3
P	57	57	57	57	57	57	57
PCF2	61	61	61	61	61	61	61
PCF5	59	59	59	59	59	59	59
PEND	31	31	31	31	31	31	31
PIF3(2)	21	21	21	21	21	21	21
RAP2.2	66	66	66	66	66	66	66
RAV1(1)	49	49	49	49	49	49	49
RAV1(2)	38	38	38	38	38	38	38
STF1	24	24	24	24	24	24	24
TAC1	68	68	68	68	68	68	68
TaMYB80	54	54	54	54	54	54	54
TBP	36	36	36	36	36	36	36
TEIL	50	50	50	50	50	50	50

TGA1	23	23	23	23	23	23	23
TGA1a	32	32	32	32	32	32	32
WRKY11	8	8	8	8	8	8	8
WRKY18/40/62	7	7	7	7	7	7	7
WRKY26/38/43	15	15	15	15	15	15	15
WRKY6	5	5	5	5	5	5	5
ZAP1	22	22	22	22	22	22	22

Many over-represented TFBSs have defense or stress-response functions

Given the list of IPF-standardized TFBSs (Table 2.5), all 4 WRKY TFBSs were over-represented at 10 dai. These abundances are supported by numerous studies which show that WRKY genes are perceived upon PAMP signals or abiotic stressors. [65–68]. WRKY genes drive defense-response by regulating NONEXPRESSOR OF PR1 (NPR1) expression by binding to W-box motifs in the NPR1 promoter. NPR1 protein binds with TGA TFs which regulate pathogenesis-response (PR) expression, hence providing a means of positively regulating SA-defense response [69–71].

Similar to WRKY, a bZIP family TFBS, HY5, was also over-represented 10 dai. Inextricably linked to photomorphogenesis, this TF is also known for its positive regulation of auxin signaling; a phytohormone which regulates defense response[72, 73]. Through interactions with HY1 and MYC2, HY5 is able to regulate photomorphogenesis, ABA and JA signaling[74, 75]. Much like MYC2, AtMYB2 is not only over-represented at 10 dai but also plays a role in ABA-signaling. Microarray analyses on *Arabidopsis* plants with 35S:AtMYC2/AtMYB2 over-expression constructs revealed induced expression of several ABA-regulated genes[76].

The GT (Trihelix) TF family member, GT-3b, was over-represented at 10 dai. Much is unknown about this TF family let alone GT-3b, however what is known is that many

GT members, like HY5, regulate photomorphogenic signaling[77]. A recent study showed how GT-2a and GT-2b over-expression positively-regulates ABA-signaling[78]. Though an over-expressed GT-3b construct was not part of this recent study, translating findings from GT-2a and GT-2b over to GT-3b could reveal potentially novel insights into whether GT-3b plays a part in ABA and defense-signaling roles.

Strong relationship between degree of TFBS over-representation and IPF-rank

Due to the multi-dimensional nature of unstandardized TFBS ranks (Table 2.4), dimensionality reduction was performed to facilitate rank visualization on a 2D coordinate plane. To carry-out such analysis, Principal Component Analysis (PCA) followed by bi-variate clustering was executed using the R library clusplot[79]. All 71 TFBSs were partitioned into 6 discrete clusters and labeled based on their respective IPF-standardized rank (Figure 2.2). Interestingly, there appears to be a strong relationship between the magnitude of TFBS over-representation and IPF-standardized rank. This suggests that IPF-standardization is suitable for deducing magnitude of over-represented TFBSs.

Comparative software analysis

Several actively-used software tools and web-interfaces are available to quantify TFBS over-representation [39,40,80–82]. We classified such tools into two classes: software that deduce TFBS over-representation given either 1) one promoter-sequence set or 2) at least two promoter-sequence sets. Marina falls into this latter class and as does a popular software tool, F-MATCH [38]. Both tools require two FASTA files as input such that one file serves as a query sequence-set while the other a baseline control. Degree of over-representation is therefore deduced by statistically contrasting TFBS over-representation across these two groups. Both software tools were compared using three independent sets of promoter-sequences of varying sizes. Each of these three analyses encompassed promoter-sequences of DEGs 10 dai from our prior soybean – soybean rust RNA-Seq study [61]. F-MATCH

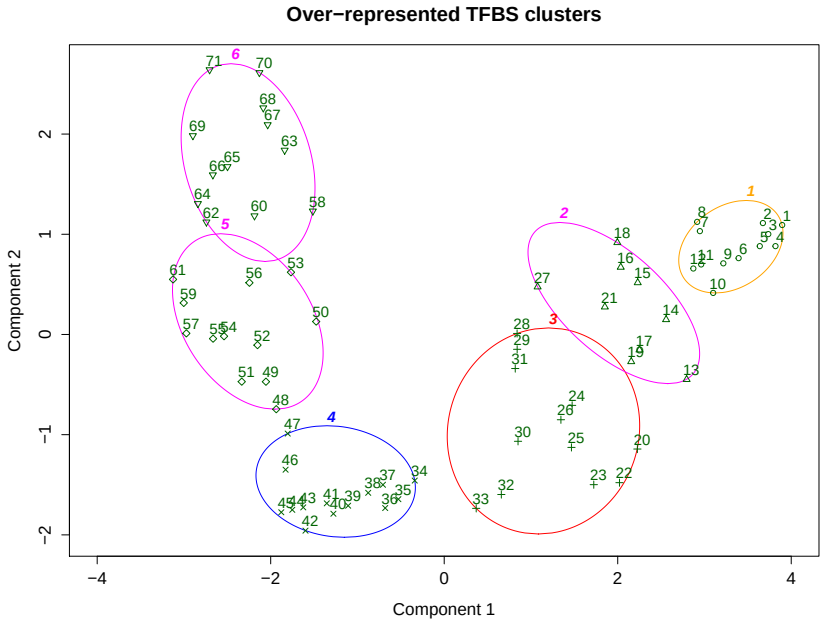


Figure 2.2: Dimensionality reduction of Marina–derived TFBSs.

and Marina identify relatively the same number of over-represented TFBSs when promoter-sequence sets are 600 sequences in size (Table 2.6). As these promoter sets increase in size, Marina maintains steady consistency as to identification of over-represented TFBSs, while F-MATCH failed to detect any over-represented TFBSs. We believe the reasoning behind why F-MATCH yields 0 over-represented TFBSs while Marina identified almost 50 TFBSs to be attributed towards usage of the binomial distribution by F-MATCH, which is known to be sensitive to large test sets. As far as output consistency between the two tools, our only comparison pertains to results obtained with 600 sequences sets. Given the 44 F-MATCH and 47 Marina over-represented TFBSs, 21 TFBSs were shared between the two result-sets. Unlike F-MATCH, we did not include TRANSFAC Professional PWMs in our analysis. We believe by introducing such PWMs, the number of shared TFBSs would certainly increase.

Table 2.6: Comparing Marina to F–MATCH.

Group size (# sequences)	PWMs (x 80)		DNA motifs (x 1,160)	
	F-MATCH	Marina	F-MATCH	Marina
600	44	47	N/A	24
1500	0	50	N/A	41
2500	0	53	N/A	44

2.4 Conclusions

Marina is an operating-system independent software tool to identify over-represented TFBSs across different groups of promoter sequences. It is freely available under the BSD license and can be downloaded at `http://mason.gmu.edu/~phossein/marina/`. The Java 7 virtual machine is required.

We illustrate its usage using an RNA-Seq plant-pathogen study, however promoter sequences from any organism can be analyzed using Marina as long as compatible TFBS models are provided. We also show its capability to identify over-represented TFBSs regardless of input size. Given large sets of DEGs, our results show that by contrasting their promoter sequences, TFBSs perceived during defense and stress response were significantly over-represented. Other lesser-known TFBSs joined these ranks, raising questions as to potential candidate TFs affiliated with defense-response.

The underlying algorithms within this tool are guided by a catalog of user-provided TFBSs models be-it DNA motifs or PWMs. Thankfully, many regulatory element resources and databases exist. By contrasting this software tool to a popular alternative, we show that Marina is built for large promoter-sequence sets while being able to identify biologically sound over-representative TFBSs.

Chapter 3: Soybean root and soybean cyst nematode interplay

3.1 Background

3.1.1 Abstract

Background

Plant–parasitic nematodes (PPNs) are obligate parasites that feed on the roots of living host plants. Often, these nematodes can lay hundreds of eggs, each capable of surviving without a host for as long as 12 years. When it comes to wreaking havoc on agricultural yield, few nematodes can compare to the soybean cyst nematode (SCN). Quantifying soybean (*Glycine max*) transcription factor binding sites (TFBSs) during a late–stage SCN resistant and susceptible reaction can shed light onto the systematic interplay between host and pathogen, thereby elucidating underlying *cis*–regulatory mechanisms.

Results

We sequenced the soybean root transcriptome at 6 and 8 days upon independent inoculation with a resistant and susceptible SCN population. Genes such as β–1,4 glucanase, chalcone synthase, superoxide dismutase and various heat shock proteins (HSPs) exhibited reaction–specific expression profiles. Several likely defense–response genes candidates were also identified which are believed to confer SCN resistance. To explore magnitude of TFBS representation during SCN pathogenesis, a multivariate statistical software identified 46 over–represented TFBSs which capture soybean regulatory dynamics across both reactions.

Conclusions

Our results reveal a set of soybean TFBSs which are over–represented solely throughout a resistant and susceptible SCN reaction. This set furthers our understanding of soybean *cis*–regulatory dynamics by providing reaction–specific levels of over–representation at 6 and 8 days after inoculation (dai) with SCN.

3.2 Introduction

Obligate parasites, such as plant–parasitic nematodes (PPNs), are infamously known for their ability to suppress host defense mechanisms and cripple yield of many agricultural crops. Such devastation is tightly orchestrated by nematode effector proteins that commandeer host–plant metabolic machinery. One of the most destructive PPNs to soybean yield is the soybean cyst nematode (SCN; *Heterodera glycines*). Worldwide, approximately 1.5 billion dollars in soybean yield is lost annually due to SCN infestations[83, 84]. In SCN susceptible soybeans, this devastation begins when the female juvenile–stage 2 (J2) nematode penetrates the host root. J2 effector proteins are injected into the root, dissolving plant cell walls and driving formation of a metabolically–active, multinucleated feeding site known as a syncytium[85]. Newly–molted J3 males and females feed from this nutrient–rich syncytium, subsequently molt into J4 larvae and copulate[86]. After approximately 30 days post–copulation, a hardened sac of SCN eggs known as a cyst becomes visible to the naked–eye. In the resistant reaction however, cysts are not visible since nematodes can neither form a nutrient–rich syncytium nor copulate. Thus, these nematodes starve to death in the resistant reaction.

With next–generation sequencing (NGS) now becoming a central assay in transcriptomics, entire transcriptomes can now be sequenced at unprecedented resolution. Fueled by the economic impact of SCN infestations, numerous studies have utilized NGS assays to sequence and quantify the soybean transcriptome [87–90]. In this study, we extend such works by conducting transcriptomic and regulatory analyses on soybean roots (Peking cv.)

inoculated with SCN. We sequence the soybean root transcriptome and contrast resistant and susceptible SCN reactions at 6 and 8 days after inoculation (dai). Our findings reveal likely defense–response gene candidates and a potential regulatory "signature" that captures TFBS over–representation throughout both resistant and susceptible reactions.

3.3 Results and Discussion

3.3.1 Illumina sequencing and read alignment

cDNA libraries from soybean roots were generated after independently inoculating roots for both 6 and 8 dai in two SCN populations, NH1RHg (confers resistant reaction in Peking; Race 3) and TN8 (confers susceptible reaction in Peking; Race 14). A baseline control cDNA library was also created from roots uninoculated with SCN. RNA was prepared using the Illumina TruSeq sample preparation kit. Single–end RNA–sequencing (RNA–Seq) was performed on the Illumina GAIIx, producing a total of 30 million reads 80 bp in length. Across all sequenced libraries, quality assessment subtracted between 10% – 19% of reads for being either a contaminant sequence or of low quality (Table 3.1).

Table 3.1: RNA–Seq summary upon SCN inoculation.

SCN population	Time–point	SRA	Reads	Filtered	Hits soybean
Uninoculated	Control	SRR849499	2,141,303	401,913	1,201,664
Race 3	6 dai	SRR847313	8,069,844	1,130,372	4,640,251
	8 dai	SRR848922	7,319,342	745,019	4,135,793
Race 14	6 dai	SRR848921	9,160,690	1,624,774	4,486,182
	8 dai	SRR849498	4,078,344	637,475	2,193,208
Total	–	–	30,769,523	4,539,553	16,657,098

Using the BWA aligner [63], quality reads were mapped against the soybean transcriptome build version 1.1 [91]. Reads aligning to multiple transcripts were identified and assigned to the transcript with the highest alignment score. In total, 59% to 67% of quality–assessed reads mapped to the soybean transcriptome.

3.3.2 Many isoforms are involved in defense response

Differential expression tests were performed using the R package DESeq[92]. Soybean transcripts were functionally annotated using both Gene Ontology (GO)[93] and PFAM[94]. Both an RPKM [64] and log_2 RPKM were computed against a baseline uninoculated sample. To render a soybean transcript differentially expressed (DE), the transcript had to have a log_2 RPKM beyond ± 0.5 and have at least 5 mapped reads. A total of 25,245 soybean transcripts were identified to be DE in at least one of the samples. Over 270 DE transcripts perceived during pathogenesis were mined (Table 3.2).

Table 3.2: Numerous DE genes involved in defense–response.

		Median RPKM			
		Race3		Race14	
Function	n	6 dai	8 dai	6 dai	8 dai
β–1,4–G	19	1.05	0.85	-0.53	-0.78
4CL	37	-0.84	-0.55	-0.93	-0.64
A–8 LOX	22	0.81	0.70	-1.89	-1.24
ChR	5	1.04	1.01	-3.65	-2.8
ChI	11	0.64	0.61	-0.74	-1.01
ChS	19	0.82	1.41	-0.82	-0.83
GST	33	0.65	0.56	-0.61	-0.72
GLY I	11	0.55	0.58	-0.66	-1.06
L–13S LOX	23	0.60	0.61	-1.96	-1.65
PCS	9	0.10	0.80	-0.78	-0.68
PR5	18	0.76	0.62	-2.77	-0.84
PR10	22	0.67	0.07	-0.75	-0.77
PDI	20	0.69	0.83	-0.67	-0.88
RnDR	6	1.39	0.93	-0.6	-0.66
SOD	17	0.70	0.65	0.70	-0.15

DE transcripts were subsequently binned based on annotated function, yielding bins of differing size, n. To estimate a bin–specific median RPKM, a 95% bootstrap confidence interval (CI) with 50,000 replicates was predicted (Table 3.3).

Virtually all annotation classifications exhibited induced expression profiles exclusive to the resistant reaction. For instance, all 19 transcripts of β–1,4–glucanase (β–1,4–G) were generally induced throughout the resistant but suppressed in the susceptible reaction. Numerous studies reveal how a pathogenic nematode can commandeer not only β–1,4–glucanase but other cellulases to drive formation of a nematode feeding site [95–97]. Critical genes encoding isoflavonoid and flavonoid biosynthesis such as chalcone synthase (ChS), chalcone reductase (ChR), and chalcone isomerase (ChI) also exhibited similar induced expression profiles. Glutathione S-transferase (GST) genes were also induced in the resistant reaction. GST is a class of enzymes involved in reactions leading to xenobiotic degradation [98], and has been shown to be induced during an SCN resistant reaction [99–101].

Table 3.3: Confidence intervals of genes involved in defense–response.

	95% CI			
	Race3		Race14	
Function	6 dai	8 dai	6 dai	8 dai
β–1,4–G	(0.64, 1.36)	(0.60, 1.16)	(-2.38, 0.94)	(-2.76, -0.09)
4CL	(-1.71, -0.44)	(-1.77, -0.06)	(-1.13, -0.6)	(-1.93, -0.29)
A–8 LOX	(-0.10, 2.31)	(0.24, 2.90)	(-2.66, -1.47)	(-3.09, -0.67)
ChR	(0.68, 5.13)	(0.91, 4.51)	(-4.07, -3.23)	(-5.07, -2.59)
ChI	(-0.04, 2.95)	(-0.03, 4.49)	(-0.84, 2.43)	(-1.3, -0.56)
ChS	(0.28, 2.17)	(0.87, 2.2)	(-1.07, -0.31)	(-1.66, -0.44)
GST	(0.48, 2.13)	(0.17, 1.9)	(-1.93, -0.24)	(-2.24, -0.04)
GLY I	(0.43, 2.72)	(0.05, 2.38)	(-0.80, 0.16)	(-2.67, -0.71)
L–13S LOX	(0.36, 2.50)	(-0.38, 2.66)	(-2.34, -1.57)	(-3.35, -1.07)
PCS	(-0.70, 2.03)	(-1.05, 2.51)	(-2.28, -0.03)	(-2.19, -0.23)
PR5	(0.02, 4.10)	(0.06, 1.86)	(-6.26, 0.37)	(-2.25, 2.10)
PR10	(0.53, 2.85)	(-0.87, 2.01)	(-2.32, 0.04)	(-2.33, 0.16)
PDI	(0.35, 1.59)	(0.11, 3.71)	(-0.84, 0.75)	(-1.88, 0.80)
RnDR	(1.06, 1.97)	(0.89, 0.96)	(-1.90, 0.64)	(-0.71, 0.36)
SOD	(0.29, 0.91)	(0.51, 0.75)	(0.34, 4.95)	(-1.01, 4.30)

Transcripts of genes encoding two lipoxygenase (LOX) gene family members, arachidonate 8-lipoxygenase (A–8 LOX; EC: 1.13.11.40) and linoleate 13S-lipoxygenase (L–13S LOX (LOX2); EC: 1.13.11.12) were also induced throughout the resistant reaction. The role A–8 LOX plays during a nematode reaction has yet to be elucidated, however lipoxygenases in–general are consistently induced throughout a resistant SCN reaction [102–105]. This raises speculation that A–8 LOX may be perceived during SCN pathogenesis.

Ribonucleoside–diphosphate reductase (RnDR; EC: 1.17.4.1) as well as protein disulfide–isomerase (PDI; EC: 5.3.4.1) were induced in the resistant reaction. Both RnDR and PDI are thioredoxins, a family of reductases known to play defense–response roles upon perception of a pathogen [106–108]. Little is known about the role RnDR plays in SCN pathogenesis, however an earlier microarray study examined abaxial and adaxial soybean embryo expression profiles upon exposure to auxin 2,4-dichlorophenoxyacetic acid (2,4–D). Microarray results revealed differentially expressed levels of expressed transcripts of RnDR 21 days after auxin inoculation[109]. PDI on the other hand, is a well–studied thioreductase expressed during plant defense[110, 111], especially in soybean roots undergoing a resistant SCN reaction[112].

Pathogenesis–related (PR) transcripts were induced in the resistant reaction. PR genes are expressed not just during pathogen perception [87, 113–118] but also following abiotic stress[119], phytohormone signaling[120] and drought[121]. Two such PR genes, PR–5 and PR–10, were both not only induced in the resistant reaction but had similar expression profiles even in the susceptible reaction. Glyoxalase I (GLY I; lactoylglutathione lyase, EC: 4.4.1.5), was also induced throughout the resistant reaction. Though not as induced compared to PR genes, GLY I has been shown to exhibit an induced expression profile in pumpkin seeds exposed to numerous abiotic stresses[122]. Little is known about the role phytochelatin synthetase (PCS) plays throughout SCN pathogenesis, however PCS has been shown to be induced during aphid herbivory[123].

3.3.3 Reaction–dependent Gene Ontology enrichments

To identify statistically significant Gene Ontology (GO) annotations, the top 750 induced and 750 suppressed genes across all SCN samples each independently underwent GO Process enrichment using the AgriGO web–server[124]. Numerous GO Processes were statistically significant across resistant and susceptible reactions (Table 3.4). GO Process p–values were adjusted using Bonferroni False Discovery Rate (FDR) and all GO Processes with adjusted p–values less than 0.05 were selected.

Table 3.4: Distribution of GO enrichments during SCN inoculation.

		Race3		Race14	
		6 dai	8 dai	6 dai	8 dai
Count	Induced	53	48	25	19
	Suppressed	73	104	113	86

The top 10 most statistically significant GO Processes within induced genes were subsequently identified (Table 3.5). Processes such as "defense response", "response to hormone stimulus", and "response to stress", were revealed to be statistically significant mainly in the resistant reaction when compared to the susceptible. Similarly, the top 10 most statistically significant GO Processes within suppressed genes were also identified (Table 3.6). Contrasting GO Processes in suppressed genes to that of induced genes reveals an entirely different catalog of annotations. For instance, 9 of the 10 GO Processes in suppressed genes are statistically significant across both resistant and susceptible reactions. This indicates that nematode effectors are generally operable in a race–independent manner and capable of effortlessly suppressing a majority of crucial basal processes. The most suppressed GO Processes were "photosynthesis", "photosynthesis, light harvesting", "photosynthesis, light reaction", and "generation of precursor metabolites and energy". Interestingly, it has been shown in prior studies that PPNs can suppress photosynthesis in tomato plants by disrupting cytokinin and gibberellin signaling[125,126]. Aside from photosynthetic processes, those

associated with metabolism and biosynthesis were highly suppressed across both reactions. This suggests that both resistant and susceptible SCN populations share a common goal of crippling basal metabolic machinery and suppressing the host machinery responsible for photosynthesis.

Table 3.5: GO Processes within induced transcripts during SCN pathogenesis.

Term	Description	$-log_{10}FDR$	
		Race 3	Race 14
GO:0006325	chromatin organization	7.18	0
GO:0051276	chromosome organization	6.11	0
GO:0006952	defense response	6.69	1.45
GO:0006323	DNA packaging	11.55	0
GO:0051704	multi-organism process	3.69	0
GO:0009825	multidimensional cell growth	6.08	1.79
GO:0034728	nucleosome organization	11.85	0
GO:0006996	organelle organization	3.48	0
GO:0009725	response to hormone stimulus	2.50	5.95
GO:0006950	response to stress	5.35	0

Table 3.6: GO Processes within suppressed transcripts during SCN pathogenesis.

Term	Description	$-log_{10}FDR$	
		Race 3	Race 14
GO:0006091	generation of precursor metabolites and energy	83.18	87.31
GO:0006096	glycolysis	1.48	3.95
GO:0009853	photorespiration	6.48	9.04
GO:0015979	photosynthesis	215.70	211.61
GO:0019684	photosynthesis, light reaction	132.78	130.48
GO:0009767	photosynthetic electron transport chain	43.33	47.11
GO:0009773	photosynthetic electron transport in photosystem I	23.73	28.55
GO:0009416	response to light stimulus	11.30	13.85
GO:0009314	response to radiation	10.71	13.19
GO:0000302	response to reactive oxygen species	0	3.73

3.3.4 Many over–represented TFBSs during SCN pathogenesis

The 1,000 most induced and 1,000 most suppressed genes were identified for each sample and the promoter sequence 2kb upstream from each genes transcription start site was retrieved and appended to a FASTA file. To quantify abundance of *cis*–regulatory TFBSs within promoter sequences, we used a collection of 68 plant Position Weight Matrices (PWMs) from AthaMap[39] and JASPAR[43]. PWMs are multi–dimensional matrices frequently used to model regulatory elements, namely TFBSs. Each cell in a PWM represents a weight as to the likelihood a particular base at a specific index is a regulatory element. Thus, mapping PWMs onto promoter sequences and statistically quantifying its abundance reveals insight into the magnitude of TFBS over–representation. To efficiently execute such mapping, we had developed a multivariate statistical software named Marina[127]. Marina maps TFBS models such as PWMs onto promoter sequences and infers magnitude of TFBS over–representation using 7 knowledge–discovery metrics. The Iterative Proportional Fitting (IPF) algorithm [58] normalizes output produced from each of the 7 metrics, enabling unanimous agreement across the metrics as to the magnitude of TFBS over–representation. IPF scores range from 1 to N whereby N is the total number of over–represented TFBSs. Scores in the range of 1 represent over–represented TFBSs while scores in the range of N represent highly under–represented TFBSs.

For all SCN samples, Marina mapped all 68 plant PWMs onto promoter sequences of both induced and suppressed genes and contrasted TFBS abundance between these sequence sets. In total, 46 TFBSs were over–represented in at least one of the four samples (Figure 3.1). There were 29 TFBSs over–represented across all four samples. If a TFBS was not over–represented in a specific sample, that TFBS was assigned an IPF score of $N+1$ so as to serve as a proxy for being highly under–represented. Contrasting TFBS IPF scores across samples reveals that 30 of the 46 TFBSs either increase or decrease in IPF score regardless of the reaction (Figure 3.1). For instance, the TFBS for STF1 exhibits a relatively modest increase in its IPF score across both reactions. Interestingly, STF1 IPF score increases from 11th to 1st from 6 dai to 8 dai respectively in the resistant reaction.

IPF score for the HAHB4 TFBS greatly increased in the resistant and susceptible reaction. A prior study found HAHB4 to contribute to jasmonic acid and ethylene signaling crosstalk [128]. Similarly, TFBSs for DOF2 and DOF3 exhibited relatively weak increases in IPF scores across resistant and susceptible samples. DOF transcripts have not been explicitly quantified as–far as their gene expression during SCN pathogenesis, however such proteins have been detected during auxin signaling [129]. In contrast to DOF2 and DOF3, the TFBS for TEIL had a near–50% jump in IPF scores across both reactions. Being the tobacco homolog of ethylene insensitive (EIN3), TEIL gene products have been shown to bind directly to the promoter sequence of PR1a[130]. Interestingly, across both resistant and susceptible reactions, TEIL scores appear to be relatively equal to one another.

The *A. thaliana* MYB77 homolog, AtMYB77, exhibits a mild change in IPF score across both resistant and susceptible reactions. Across both reactions, AtMYB77 IPF scores were generally under–represented at 6 dai but become slightly over–represented at 8 dai. The TFBS for OsCBT exhibited pronounced IPF scores. In both the resistant and susceptible reaction, OsCBT was highly over–represented only at 6 dai. It was shown that OsCBT mutants conferred increased pathogen resistance upon inoculation with *Magnaporthe grisea*, revealing that OsCBT suppresses defense response[131].

3.3.5 TFBSs are over–represented in a reaction–dependent manner

The remaining 16 TFBSs from the total set of 46 TFBSs were over–represented in one reaction compared to the other. Such TFBSs can expose novel insight into TFBSs over–representation patterns respective to a specific reaction.

ZAP1, a TFBS for a WRKY TF [67], appears to be highly over–represented during the resistant reaction but slightly under–represented in the susceptible reaction. Similarly, PIF3–1 and PIF3–2 were both under–represented during the susceptible reaction, however slightly over–represented in the resistant reaction. It has been shown that PIF plays roles in phytochrome signaling[132]. Since photosynthetic processes are heavily suppressed within resistant and susceptible reactions (Table 3.6), such suppression explains why PIF3–1 and

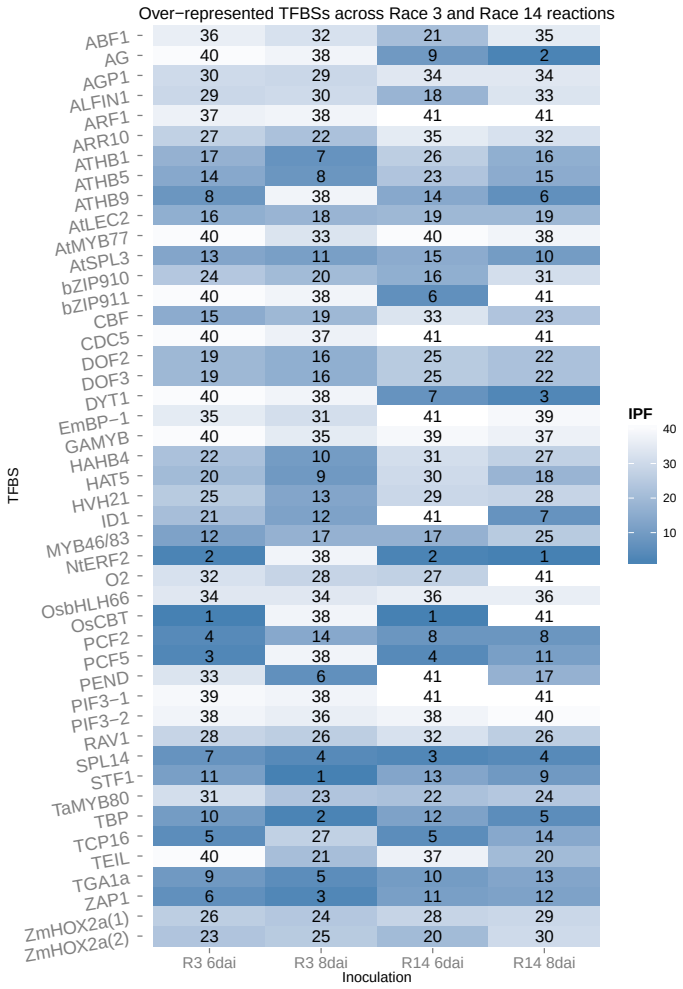


Figure 3.1: Heatmap of TFBSs over–represented during SCN pathogenesis.

PIF3–2 have severely under–represented IPF scores. Indeed SCN pathogenesis does not only disrupt the photosynthetic machinery but also the plants ability to execute sound phytochrome signaling. Numerous TFBSs of TFs perceived during defense response such as bZIP910, bZIP911, TaMYB80, and TGA1a were all over–represented during the resistant reaction but under–represented in the susceptible.

3.4 Conclusions

We used RNA–Seq to sequence soybean whole–root (Peking cv.) at both 6 and 8 dai upon inoculation with a resistant (NH1–RHg; Race 3) and susceptible (TN8; Race 14) SCN population. Contrasting TFBSs over–represented in promoter sequences of DE soybean genes after 6 and 8 dai revealed underlying transcriptomic and *cis*–regulatory dynamics within the soybean root during pathogenesis. In–total, over 30 million reads from soybean whole–root was sequenced and differential expression analysis revealed over 270 transcripts to be statistically and biologically significant during defense–response. Several viable defense–response gene candidates joined these ranks, including glyoxalase I, arachidonate–8 lipoxygenase, phytochelatin synthetase, and ribonucleoside-diphosphate reductase. A set of 46 TFBSs were rendered over/under–represented across all resistant and susceptible samples. Interestingly, 30 of these TFBSs were either over or under–represented across both reactions. Thus, our results reveal presence of a biologically–sound regulatory signature that identifies reaction–specific soybean regulatory patterns during both resistant and susceptible SCN reactions.

3.5 Methods

3.5.1 Plant procurement and SCN inoculation

Glycine max cv. Peking seeds were surface–sterilized by treating the seeds with 10% bleach (0.6% sodium hypochlorite) for 10 minutes, followed by several washes with distilled water. Seeds were planted in sterile sand in 20x20 cm flats. Eight days later, seedlings were gently lifted out of the sand and rinsed clean with water. Five seedlings for each time–point were placed on moistened germination paper in 8x12x3.5 cm plastic trays. The two SCN populations NH1–RHg and TN8 were independently harvested from stock plants [133]. Females were crushed with a rubber stopper and eggs were washed through a 250 micron screen and collected on a 25 micron screen. Eggs were left to hatch on a rotary shaker for 3 days. J2 stage nematodes were further purified by passing them through a 30 micron cloth

into deionized, distilled water and gently centrifuged at 250 relative centrifugal force (RCF) for 1 minute to concentrate to 2,000 J2/ml. Roots from each plant were inoculated with 1 ml of inoculum. Roots were covered with a second piece of moistened germination paper and the trays were placed in a larger tray with 0.5 cm water and wrapped in a semi-clear plastic bag for the duration of the time–points. Uninoculated control plants were also placed in trays and collected separately. Four of the plant roots following 6 and 8 dai with SCN were harvested, immediately frozen in liquid nitrogen, ground to a fine powder in a mortar and pestle, and stored in microfuge tubes at –80°C until RNA extraction. A fifth root was stained for visualization of nematode infection with acid fuchsin [134]. RNA was extracted after 6 and 8 days after inoculation by phenol/chloroform and lithium chloride precipitation [135]. RNA was treated with DNase to remove any genomic DNA remaining in the samples. RNA integrity was checked by visualizing the intact 18S and 28S ribosomal bands on an agarose gel and concentrations were measured on a NanoDrop spectrophotometer (Thermo Scientific; Waltham, MA).

3.5.2 RNA extraction and cDNA isolation

cDNA libraries were prepared using the TruSeq RNA Prep Kit according to the manufacturer instruction (Illumina). Briefly, mRNA was purified from four micrograms of total RNA diluted in fifty microliters of nuclease–free ultra pure water using magnetic beads. Resulting mRNA was fragmented at 94°C for eight minutes. Seventeen microliters of fragmented mRNA was used as template for cDNA synthesis performed by a Superscript II Reverse Transcriptase. Second–strand synthesis was immediately performed and fifty microliters of double stranded DNA was transferred to a new tube and submitted to end repair followed by adenylation of 3' ends. Once adenylation of 3' ends reached completion, adapters containing different indexes were ligated to each library. DNA fragments having adapter molecules on both ends were amplified and enriched. Quantification and quality control were performed by loading one microliter of cDNA libraries on an Agilent DNA–1000 chip and running it on an Agilent Technologies 2100 Bioanalyzer.

3.5.3 Deep–sequencing and transcriptome quantification

For both NH1–RHg (Race 3) and TN8 (Race 14) reactions, cDNA libraries were sequenced from 8 day old soybean whole–root independently inoculated with SCN at 6 dai and 8 dai. Single–end RNA–sequencing was performed on the Illumina GAIIx at the United States Department of Agriculture (USDA), Beltsville, MD. An uninoculated whole–root control was also sequenced using the same sequencing protocol. To remove low quality reads across all sequencing runs, custom bash scripts filtered all reads should its 3' tail have a quality score of less than 22. To remove contaminant reads, sequences were subtracted if they mapped at least once to both the Ensembl human genome (Hg19) or the JCVI Microbial Resource[62]. Remaining sequences were mapped to the soybean transcriptome (build 1.1) using BWA. Across all SCN inoculated samples, transcript counts underwent normalization and variance estimation using the DESeq R package. To infer magnitude of differential expression, RPKM was computed for all inoculated and uninoculated samples and $log_2(\frac{RPKM_{inoculated}}{RPKM_{uninoculated}})$ was subsequently derived. A transcript was rendered non–differentially expressed if it failed any of the following conditions: log_2 RPKM not beyond ± 0.5, an adjusted p–value greater than or equal to 0.05, or fewer than 5 mapped reads.

3.5.4 Functional annotation & Gene Ontology (GO) enrichment

Functional annotation comprised of homology–based analysis of all sequences in the Phytozome soybean transcriptome. Of these 73,320 soybean transcriptomic sequences, 7,810 sequences were subtracted for being either a scaffold or duplicate sequence. BLASTX[136] aligned the remaining 65,510 query sequences onto all UniProt plant proteins[137]. The top–scoring UniProt function annotation was assigned to the query if it did not contain ambiguous keywords, namely "Hypothetical", "Uncharacterized" or "Unknown".

For all samples, soybean Phytozome accessions for the top 750 induced and top 750 suppressed transcripts were identified. Gene Ontology (GO) enrichment on each accession–set

was performed using the AgriGO web–server. AgriGO settings were modified to quantify GO annotations using the hypergeometric distribution and Bonferroni p–value false–discovery rate (FDR) correction. To measure GO Process statistical significance in both resistant and susceptible reactions, the $-log_{10}FDR$ per GO Process was summed across both 6 and 8 dai time–points. Subsequently, the top 30 most statistically significant GO Processes from the top 750 induced and suppressed transcript sets were identified.

Chapter 4: Interplay between soybean rust and susceptible soybean plants

4.1 Abstract

In virtually all soybean–growing regions of the world, one of the most devastating pathogens is *Phakopsora pachyrhizi*, an obligate biotroph that causes Soybean Rust (SR). At least 10% of annual soybean yield in the United States is lost due to this pathogen. In spite of this, soybean remains amongst one of the top agricultural exports of the United States. To better understand interaction of SR with soybean at the molecular level and how aspects may be regulated, we examined changes in expression of soybean genes and quantified proximal *cis*–regulatory elements within promoter sequences of differentially expressed soybean genes. Understanding soybean *cis*–regulatory dynamics during an SR time–course contributes to our understanding of plant–pathogen interplay by revealing numerous transcription factor binding sites rendered over–represented and statistically significant during infection.

4.2 Introduction

Phytopathogenic fungi are plant pathogens which feed and live off nutrients within host tissues. The invasive mechanism of these pathogens is made possible through a cocktail of effector proteins that cripple plant defenses and commandeer host metabolic machinery [138, 139]. Since many agricultural crops are hosts to phytopathogenic fungi, agricultural yield–loss at the hand of these pathogens is inevitable. In the case of soybean (*Glycine max*), one phytopathogenic fungus, *Phakopsora pachyrhizi*, has gained notoriety for devastating entire soybean yields. As a result, quantifying soybean proximal *cis*–regulatory elements during fungal pathogenesis could shed light on soybean regulatory dynamics during soybean

rust pathogenesis.

Soybean is a major agricultural export by the United States and is a legume rich in antioxidants, oils, and protein [140]. Unfortunately, *Phakopsora pachyrhizi* and its causative foliar disease, soybean rust (SR), have been shown to sever yield with losses as much as 10% to 50% in the United States [141,142]. In Brazil however, SR can be found in approximately 80% of soybean fields, resulting in millions of dollars in lost yield annually [143]. Unfortunately, no soybean cultivar is yet available that is resistant to all SR strains. Certainly alternative measures such as fungicide treatment and fungitoxic essential oils have shown promising results when it comes to reducing the extent of SR infection [144–147].

The SR life cycle begins with the germination of asexual urediniospores on host leaves. Following post–germination is the elongation of a thin extremity known as the germ tube. The germ tube grows on the host leaf surface and eventually forms a specialized structure known as an appresorium which punctures leaf epidermis, thereby allowing the fungus to penetrate the leaf. Such penetrance is made possible by an appresorial cone, a cone–shaped extremity grown out of the appresorium. Once the fungus has entered plant mesophyll, primary hyphae are produced which lead to development of a specialized structure known as haustorium. The haustorium is dedicated to absorbing host nutrients, enabling further pathogenesis. Over the course of fungal differentiation, haustoria will be formed throughout the host tissue. On the leaf surface, red–brown spots representing localized necrosis become visible to the naked eye. Within leaf mesophyll however, hyphae aggregate together, forming uredinia primordia. Such structures then differentiate to form asexual urediniospores. Overall, the entire SR lifecycle can take between 7 to 10 days [148–150].

In two prior studies [61, 151], we sequenced the soybean leaf transcriptome at 7 hours after inoculation (hai), 48 hai and 10 days after inoculation (dai) with SR. In this study, we extend both works by examining and quantifying soybean proximal *cis*–regulatory elements within differentially expressed soybean transcripts.

4.3 Materials and Methods

4.3.1 Plant procurement and RNA sequencing

In an earlier study [61], we used RNA–Sequencing (RNA–Seq) to quantify the soybean leaf transcriptome at 10 dai with SR. A control sample 15 seconds after inoculation was also sequenced as part of this prior investigation. In a follow–up study [151], we utilized RNA–Seq to further quantify the soybean leaf transcriptome at two additional time–points: 7 hai and 48 hai with SR (Table 4.1).

Custom Python scripts parsed FASTQ files produced from both prior studies and removed low quality reads if its 3' tail had a quality score less than 25, or at least 35% of the entire read had a quality score less than 22.

Table 4.1: RNA–Seq runs investigating soybean–SR interplay.

Time–point	SRA	# reads	Reference
Control	SRR352328	4,467,871	[61]
7 hai	SRR863029	7,543,421	[151]
48 hai	SRR863032	9,082,363	[151]
10 dai	SRR352327	3,510,311	[61]

4.3.2 Differential expression analysis and functional annotation

Reads passing quality filters were subsequently mapped onto both the human genome (Hg19) and all microbial genomes from the JCVI Microbial Resource [62] using the BWA aligner [63] to remove potentially contaminant reads. Reads across all four RNA–Seq time–points that passed quality and contaminant checks were subsequently mapped onto the soybean transcriptome build 1.1 [91] using the TopHat splice–aware aligner [152]. The Cuffdiff software [153] inferred differential expression between each SR inoculated and control sample. A transcript was rendered differentially expressed if its Benjamini–Hochberg p–value was below 0.05 and had at least a log_2 Fragments Per Kilobase per Million fragments

mapped reads (FPKM) of at least ±1.5.

For all inoculated RNA–Seq runs, Phytozome accession numbers for the top 750 induced and suppressed soybean transcript isoforms were subsequently identified and termed an accession–set. The AgriGO web–server [124] analyzed each accession–set so as to identify statistically over–represented Gene Ontology (GO) terms; a process known as GO enrichment. A term was rendered enriched if its Hochberg false discovery rate (FDR) adjusted hypergeometric p–value was below 0.05. For each enriched GO Process, its $-log_{10}FDR$ was subsequently summed across all time–points and the resultant summation was ranked in descending order, producing a list of enriched GO Processes.

4.3.3 Derivation of over–represented soybean binding sites

For 7 hai, 48 hai, and 10 dai samples, the top 1,000 induced and suppressed soybean transcripts were identified and their promoter sequences 2kb upstream of their transcription start site were appended to two FASTA files respectively. A total of six FASTA files were therefore created, each termed a positive set. To effectively control for length distribution and GC content within of the six positive sets, a matching control set of 1,000 promoter sequences were randomly generated from non–differential soybean transcripts. Each control set differed in GC content by at most ± 7% when compared to its respective positive set; on average, GC content differed by ± 4%. Upon generation of a control set, the positive set was paired with its respective control set and termed a positive–control pair.

To model *cis*–regulatory transcription factor binding sites (TFBSs), a set of 66 plant Position Weight Matrices (PWMs) were mined from AthaMap [39], JASPAR [43], and TRANSFAC [42]. PWMs are multi–dimensional matrices frequently used to model regulatory elements, namely TFBSs. Each cell in a PWM represents a weight as to the likelihood a particular base at a specific index is a regulatory element. Thus, mapping PWMs onto promoter sequences and statistically quantifying relative abundance reveals insight into the magnitude of TFBS over–representation.

To quantify TFBS PWM over–representation given a positive and control set, we used

the Marina software [127]. Marina performs TFBS over–representation analysis using a set of 7 statistical metrics to compute magnitude of TFBS over–representation. To yield a collective measure of TFBS over–representation across all 7 metrics, Marina utilizes the Iterative Proportional Fitting (IPF) algorithm [58]. Doing so enables the investigator to identify the most over–represented TFBSs with ease. IPF ranks range from 1 to N whereby a rank of 1 represents a highly over–represented TFBS while N represented a highly under–represented TFBS. If a TFBS was over– or under–represented in one SR inoculation but not in another, that TFBS was set an IPF rank of $N + 1$ so as to serve as a proxy for a highly under–represented TFBS. For each TFBS, the $\text{IPF}_{induced}$ / $\text{IPF}_{suppressed}$ ratio was derived per time–point to provide a time–point measure of TFBS over–representation. Resultant IPF ratios were visualized using the gplots R package [154].

4.3.4 Building a soybean TFBS classifier

To assess the quality of TFBS hierarchical clustering (Figure 4.1), we built LASSO regression classifiers [155]. TFBS PWMs fitted by a LASSO model receive real–number weights which represent the magnitude of TFBS classification to a particular sequence set.

For all positive–control pairs, Marina generated an abundance matrix to model TFBS counts within individual promoter sequences. Using the glmnet R package [156], a LASSO classifier with 10–fold cross–validation (CV) was subsequently built for each abundance matrix. CV is a popular statistical procedure for testing predictive power of a classifier. In the case of 10–fold CV used in our models, the classifier is trained on nine–tenths of the input matrix and tested on the remaining one–tenth. Thus, classifier testing can shed light into how statistically accurate hierarchical clustering of TFBS over–representation truly is.

4.4 Results and Discussion

4.4.1 GO Processes capture soybean–SR pathogenesis interplay

Numerous enriched GO Processes within induced soybean transcripts were identified across all three time–points (Table 4.2). Not surprising, "defense response", "response to stimuli", "response to stress", "flavonoid biosynthesis process", and "phenylpropanoid biosynthetic process" were enriched across all three time–points. These five GO Processes alone appear to be moderately enriched early in SR pathogenesis, but increase several fold during the 10 dai time–point. An earlier microarray study quantified leaves of both resistant and susceptible soybean plants after SR inoculation and found GO Processes "defense response" and "flavonoid biosynthetic process" to be enriched during a 7 dai reaction [157]. When comparing enriched GO Processes within induced transcripts to that of suppressed transcripts, there appears to be an entirely different catalog of enriched GO Processes (Table 4.3). Despite this fact, there is some enrichment overlap. GO Processes such as "aromatic compound biosynthetic process", "cellular amino acid derivative biosynthetic process", "cellular amino acid derivative metabolic process", "flavonoid biosynthetic process", "lignin biosynthetic process", "phenylpropanoid biosynthetic process", and "phenylpropanoid metabolic process", are enriched across both induced and suppressed transcripts.

Within suppressed transcripts, 12 of the 15 total enrichments were either associated with biosynthesis or metabolism. This majority alone is testament to the commandeering nature of SR, bent on crippling host machinery and leeching nutrients from its host. A further 12 of the 15 GO Processes were enriched in either 7 hai or 48 hai but not at 10 dai. These 12 processes were generally associated with amino acid biosynthesis, fatty acid biosynthesis, and lignin biosynthesis. A plausible explanation why a bulk of the biosynthesis–related processes were not enriched at 10 dai is that by this time–point, uredinial primordia and resultant urediniospores are released, rupturing the host cell epidermis [158], collapsing the cell wall, lignin, and membranes formed from fatty acids.

Table 4.2: GO Processes within induced transcripts during SR pathogenesis.

GO Process	$-log_{10}FDR$		
	7 hai	48 hai	10 dai
aromatic compound biosynthetic process	1.31	2.96	6.82
cellular amino acid derivative biosynthetic process	1.60	3.34	6.82
cellular amino acid derivative metabolic process	1.62	3.09	5.80
cellular aromatic compound metabolic process	0	1.92	4.96
defense response	0	5.26	5.80
flavonoid biosynthetic process	0	1.77	4.32
lignin biosynthetic process	3.68	5.26	5.08
lignin metabolic process	3.68	5.26	5.08
petal formation	0	8.27	0
petal morphogenesis	0	6.25	0
phenylpropanoid biosynthetic process	1.77	3.85	8.15
phenylpropanoid metabolic process	1.34	3.42	8.15
response to inorganic substance	3.13	3.14	1.77
response to stimulus	1.68	2.68	2.49
response to stress	3.01	3.34	4.13

4.4.2 Over–represented soybean TFBSs capture SR dynamics

Of the 66 TFBS PWMs used in this study, 25 TFBSs were over–represented in at least one of the three time–points. To model inoculation–specific magnitude of TFBS over–representation, a TFBS IPF ratio was computed given $\text{IPF}_{induced}$ / $\text{IPF}_{suppressed}$. Thus, ratios closest to 0 are over–represented within promoter sequences of induced transcripts whereas ratios much larger than 0 are over–represented within promoter sequences of suppressed transcripts. To aide visualization of TFBS over–representation, hierarchical clustering was performed on TFBS ratios, creating 4 fixed clusters each designated by their respective color and integer (Figure 4.1).

TFBSs within cluster 1 ($n = 15$) were predominantly over–represented within promoter sequences of induced transcripts across 7 hai, 48 hai, and 10 dai. At 10 dai however, ratios of some TFBSs changed entirely and generally became more abundant within promoter

Table 4.3: GO Processes within suppressed transcripts during SR pathogenesis.

GO Process	$-log_{10}FDR$		
	7 hai	48 hai	10 dai
aromatic compound biosynthetic process	1.62	1.41	0
cellular amino acid derivative biosynthetic process	1.38	2.15	0
cellular amino acid derivative metabolic process	0	1.47	0
fatty acid biosynthetic process	0	2.80	0
fatty acid metabolic process	0	2.22	0
flavonoid biosynthetic process	1.49	0	0
lignin biosynthetic process	1.43	0	0
lipid localization	8.74	7.80	19.96
lipid metabolic process	0	2.80	0
lipid transport	2.00	2.17	6.82
multidimensional cell growth	0	0	1.77
phenylpropanoid biosynthetic process	2.00	1.60	0
phenylpropanoid metabolic process	1.66	1.47	0
positive regulation of biosynthetic process	0	1.64	0
positive regulation of cellular biosynthetic process	0	1.64	0

sequences of suppressed sequences. The TFBS for ATHB1, for instance, exhibited approximately equal IPF ratios throughout the three time–points. When contrasting against fellow homeodomain TFBS, ATHB5, both TFBSs remain relatively equal in IPF ratios, but vary considerably at 10 dai. ATHB5 has been shown to play roles in regulating abscisic acid (ABA) accumulation [159,160]. The TaMYB80 TFBS shared a similar IPF ratio to that of the ATHB5 TFBS. TaMYB80 has been shown to play active roles in meristem formation [28] and pollen development regulation [161], respectively. As a protein family however, MYB proteins have been shown to play a number of diverse roles in plants, ranging from pathogen defense [27,162] to metabolism [163–165].

TFBSs within cluster 2 ($n = 2$) are generally over–represented within promoter sequences of induced transcripts and have an IPF ratio far more closer to 0 when compared to cluster 1. ABI4, for instance, is known for its involvement in both abscisic acid (ABA) signaling and stress response regulation [166–168]. Similar to the TFBS of ABI4, the TFBS of ZAP1 exhibited similar magnitudes of over–representation. Being a member of the WRKY

TF family, ZAP1 has shown to regulate salicylic acid and stress response signaling [69,169].

TFBSs comprising cluster 3 ($n = 6$) are heavily over–represented within promoter sequences of suppressed transcripts at 7 hai. However at 48 hai, IPF ratios do not follow this consistency. Only at 10 dai do IPF ratios appear to be generally found within promoter sequences of induced transcripts. For instance, TGA1A has been shown to bind to promoter sequences of xenobiotic regulators [170]. The magnitude of TGA1A TFBS over–representation at 7 hai could therefore reveal insight into how TGA1A contributes to regulating temporal defense response. Similarly, the TFBS of PCF5 is highly over–represented at 7 hai within promoter sequences of suppressed transcripts, but becomes over–represented within induced transcripts at both 48 hai and 10 dai. Belonging to the TCP protein family, this family is known for its ability to regulate biosynthesis of jasmonic acid [171]. Thus, suppressive efforts by SR effector proteins to undermine host defense mechanisms such as jasmonic acid signaling may explain the decreases in PCF5 TFBS representation at 7 hai.

TFBSs comprising cluster 4 ($n = 2$) were over–represented within promoter sequences of induced transcripts at both 7 hai and 10 dai. The TFBS for DYT1, for instance, has been shown to regulate pollen development [172]. It comes to no surprise that manipulation of host photosynthesis machinery by SR disrupts regulatory factors associated with this very process, hence the severe under–represented nature of the DYT1 TFBS.

4.4.3 Accurate classification of soybean TFBSs

A cross–validated LASSO model was generated for all six positive–control pairs (Figure 4.2). The efficiency of LASSO classification, or area under the curve (AUC), maximizes at 1.0. AUC curves approaching this vicinity therefore represent excellent classification performance. Our LASSO classifiers collectively reached AUC ranges of 0.82–0.87, indicative of very good classification performance. Thus, our LASSO models can confidently classify TFBSs given a positive set and a corresponding control.

To quantify extent of TFBS classification within each positive set, a numerical LASSO weight per TFBS was generated from each model. Positive weights signified classification of

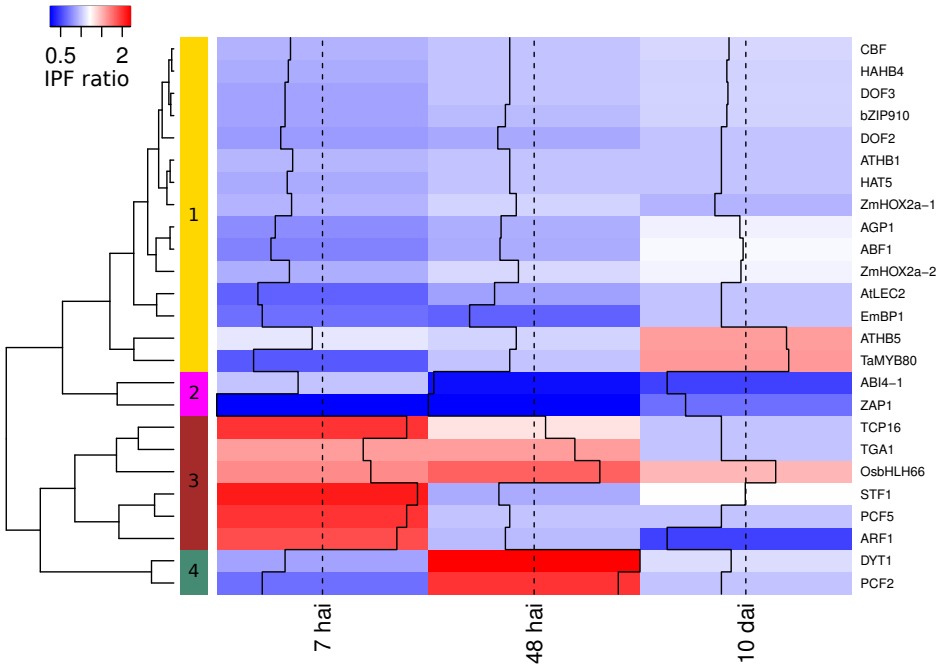


Figure 4.1: Over–represented TFBSs during a time–course SR inoculation.

a TFBS to the positive set be–it promoter sequences from induced or suppressed transcripts. Negative weights however meant that the TFBS was classified predominately in the control set. Weights equal to zero signified no inherent classification to neither the positive nor control set. Of the 25 over–represented TFBSs, all but 1 TFBS had non–zero weights across all time–points. In each of the three time–points, the number of TFBSs were enumerated if its weight was positive and was over–represented in promoters of induced or suppressed transcripts. Subsequent enumeration was then visualized to illustrate over–represented TFBSs classified within the positive sets of promoter sequences (Figure 4.3).

Of the 24 TFBSs, 9 (ZAP1, AtLEC2, ABI4–1, EmBP1, HAT5, DOF3, ZmHOX2a-2, HAHB4, CBF) were collectively weighted in all inoculations and over–represented within promoters of induced transcripts (Figure 4.3a). This should come as no surprise since many of these regulatory elements are TFBSs of TFs perceived in defense response. For instance, ZAP1, otherwise known as WRKY1, exhibited high levels of TFBS over–representation

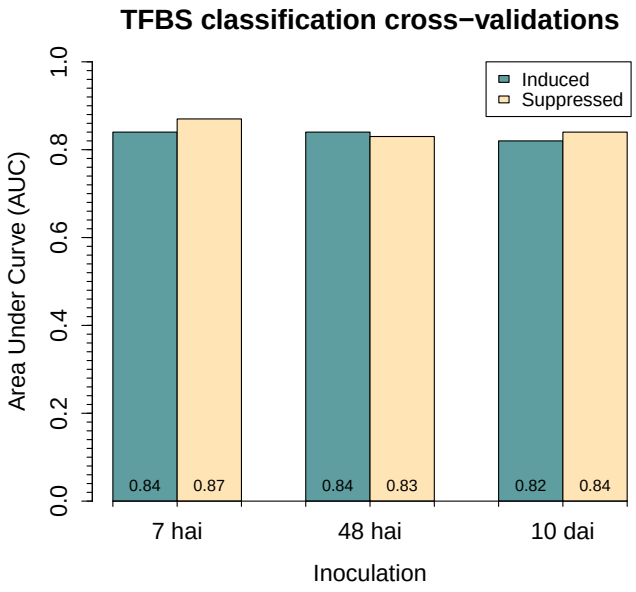
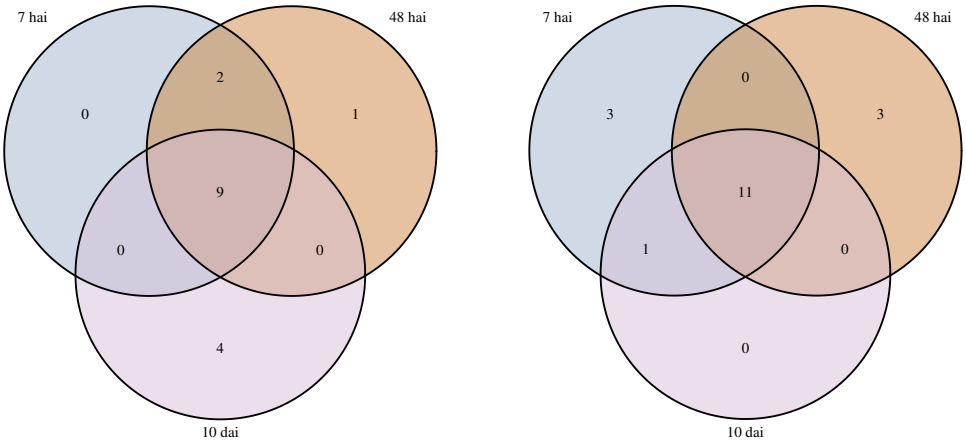


Figure 4.2: TFBS classification within promoter sequences.

within promoters of induced transcripts. WRKY proteins have been well–studied and are cataloged extensively as to their roles in biotic stress response [68, 173]. The TFBS for ABI4–1 also exhibited significantly high levels of over–representation, with numerous studies implicating its involvement in regulating abscisic acid biosynthesis and jasmonic acid signaling [174, 175]. Aside from weighted TFBSs within promoters of induced transcripts, 11 TFBSs were both weighted and over–represented within suppressed loci (Figure 4.3b). Interestingly, numerous TFBSs of TFs induced during stress response, such as ZAP1, OsbHLH66, and TGA1, were present within this set, however their weights were trivial when compared to promoters of induced transcripts. Nonetheless, this set similarity should come as no surprise since well studied TF families (WRKY, bZIP, and MYB) have multi–purpose functionalities and are found in promoters of induced and suppressed loci.



(a) Promoters of induced transcripts. (b) Promoters of suppressed transcripts.

Figure 4.3: Distribution of classified over–represented TFBSs.

4.5 Final remarks

Much progress has been made by the soybean community to advance our understanding of soybean–pathogen transcriptomic and regulatory interplay. With next–generation sequencing assays now used extensively in molecular biology, host dynamics can be examined at unprecedented nucleotide resolution. These assays have proven invaluable towards quantifying soybean tissues during a pathogen time–course as they can identify potentially novel insight into host *cis*–regulatory dynamics during defense response.

In this study, we build on prior analyses by quantifying soybean *cis*–regulatory elements during a 7 hai, 48 hai, and 10 dai time–course treatment with soybean rust. We show there to be 25 TFBSs over–represented across the three time–points, with many serving as TFBSs of TFs perceived in defense or stress response. To gauge whether these TFBSs were predominately over–represented within promoter sequences of induced or suppressed transcripts, linear regression models quantified such classification. Indeed numerous TFBSs were found within promoters of induced transcripts, serving as a glimpse into soybean regulatory dynamics during defense response.

Chapter 5: Soybean root transcriptomes following phytohormone treatment

5.1 Abstract

Jasmonic acid, ethylene, and auxin are amongst a collection of empirically studied plant hormones induced during plant defense response. Otherwise known as phytohormones, these signaling molecules are regulated through numerous elegant and precise biochemical cascades in a tissue–specific manner. In the face of commandeering stimuli such as pathogens, phytohormones drive synthesis of defense response elicitors. Thus, quantifying host cDNA and upstream transcription factor binding sites following phytohormone treatment could reveal potentially novel insight into defense response signaling and co–regulatory interplay. In this study, we quantified soybean (*Glycine max*) root cDNA following treatment with jasmonic acid, ethylene, and auxin. Statistical analysis of promoter sequences of differentially expressed transcripts revealed numerous over–represented soybean binding sites of transcription factors induced during stress response. Our results provide a biologically–sound catalog of differential transcripts and corresponding binding sites over–represented following soybean root treatment with jasmonic acid, ethylene, and auxin.

5.2 Introduction

Phytohormones are plant hormones which regulate virtually every aspect of plant development. Systematic interplay amongst such molecules, even at very low concentrations, drives biochemical signaling cascades which govern processes such as seed germination and leaf senescence to flower differentiation. In the face of herbivory and pathogenesis, this tightly

regulated interplay, be–it synergistic or antagonistic, is necessary for orchestrated and perfectly timed signaling cross–talk. Throughout the past several decades, our understanding of plant hormones and their corresponding receptors has grown orders of magnitude, with a bulk of such advances discovered in the *Arabidopsis thaliana* model. Amongst the first empirically studied phytohormones were auxin (IAA), abscisic acid (ABA), gibberellin (GA), and ethylene (ETH). These phytohormones were termed "classical" phytohormones due to their core ability to regulate plant organ differentiation [176,177]. More recently discovered signaling molecules such as jasmonic acid (JA), salicylic acid (SA), and brassinosteriods (BS) have also joined the phytohormone catalog. Upon perception of external stimuli such as pests or pathogens, virtually all phytohormones are involved in systematic cross–talk [178–181]. Thus, treating plant tissues with various phytohormones and quantifying cDNA abundance of such samples could provide a transcriptomic and regulatory snapshot into both systematic cross–talk and defense–response signaling.

In this study, we independently treated soybean (*Glycine max*) roots with three phytohormones: JA, IAA, and ETH. Treated roots were sequenced using RNA–Sequencing (RNA–Seq), generating almost 300 million high–quality reads. Functional analysis of differential transcripts revealed numerous distinct transcriptomic profiles following either individual or pairs of treatments. In addition, statistical analysis of soybean transcription factor binding sites identified numerous proximal regulatory elements of genes perceived in defense–response signaling. Our results provide a potentially novel catalog of biologically–sound and over–represented soybean *cis*–regulatory proximal binding sites following IAA, ETH, and JA treatment.

5.3 Materials and Methods

5.3.1 Plant procurement and phytohormone treatment

Soybean plants (*Glycine max* cv. Williams 82) were grown in a greenhouse under natural light at 25°C. Two weeks after sowing plants, roots from twelve plants were treated, three

with JA, ETH, and IAA (Invitrogen, USA), respectively. JA and IAA were dissolved in 1 mL ethanol and dispersed in water to produce 5 mM and 0.5 mM concentrations solutions, respectively. Roots of each of the plants were submerged into 250 mL containing JA or IAA for 8 hrs in the dark. Plants treated with ETH were placed in an airtight cabinet in the dark. ETH was injected to achieve a final concentration of 1 μl/L. To serve as a baseline, control plants were also grown and maintained in the dark for 8 hrs but treated with neither JA, IAA nor ETH.

5.3.2 RNA isolation and cDNA sequencing

Total RNA was extracted from roots of each plant using the RNeasy Mini Kit (Qiagen, USA). Two RNA libraries were generated from untreated roots as well as roots treated with JA and IAA. For roots treated with ETH, only one RNA library was generated. cDNA libraries were generated for all 7 RNA libraries and sequenced using the Illumina paired–end sequencing recipe by Expression Analysis (Durham, NC).

5.3.3 Transcriptome assembly and quantification

Paired–end FASTQ files from each of the sequenced lanes were mapped to the Phytozome *Glycine max* transcriptome build 1.1 [91] using the BWA short–read aligner [63]. Reads mapping to more than one transcript were assigned to the transcript with the highest alignment score. Following read alignment, custom bash scripts enumerated read–counts per transcript. The DESeq R package [92] performed read–count normalization, variance estimation across replicates, and differential expression hypothesis testing. A transcript was rendered differentially expressed (DE) if it had at least 5 mapped reads in all replicates, a $\log_2$ fold–change greater than ± 1.5 and an adjusted p–value less than 0.05.

5.3.4 Gene Ontology analysis

Across each treatment, accession identifiers for the top 200 differentially induced and suppressed transcripts were selected and termed an accession–set. The AgriGO web–server [124]

identified statistically significant Gene Ontology (GO) terms within each accession–set, a process known as GO enrichment. AgriGO settings were modified so as to execute statistical testing using the hypergeometric distribution and p–value correction using Hochberg false discovery rate (FDR).

5.3.5 Identification of outlier differential transcripts

In the context of this study, outlier transcripts were defined as transcripts with statistically significant log_2 fold–change variance following two treatments. To investigate transcripts with outlier expression profiles, all transcripts DE following pairs of treatments were identified. Expression profiles were subsequently represented as an $i \times j$ matrix. Each matrix column, j, referenced transcript log_2 fold–change expression, while each row i referenced a unique DE transcript accession identifier. For each matrix row, i, the Mahalanobis distance, m_i, was computed. The 97.5% quantile with j degrees of freedom from the chi–squared distribution, q, was subsequently derived. A transcript was rendered an outlier if $m_i \geq q$.

5.3.6 Identification of over–represented soybean binding sites

Following each of the three phytohormone treatments, Phytozome accession identifiers for the top 400 induced soybean transcripts were retrieved. For all accession identifiers, its promoter sequence 2kb upstream from its transcription start site was subsequently identified. Thus, three sets of promoter sequences were produced, each termed a query set. To effectively identify over–represented transcription factor binding sites (TFBSs) within each query set, a matching control set was generated from promoter sequences of non–differential transcripts following all treatments. Each control set was generated to match length distribution of its respective query set and differ in GC content by at–most 2%. To model TFBSs, 71 publicly available position weight matrices (PWMs) were retrieved from AthaMap [39] and JASPAR [43]. The Marina software [127] was used to statistically contrast abundance of all 71 PWMs within each query and control set. Following abundance analyses, a PWM abundance matrix was generated by Marina to represent frequency of individual PWMs

within all promoter sequences. To effectively quantify magnitude of TFBS classification to either the query or control set, a LASSO classifier was built for each matrix using the glmnet R package [156]. Each classifier was built using 10–fold cross–validation whereby the model was built using nine–tenths of the matrix and tested on the remaining one–tenth. Magnitude of TFBS classification was modeled as real–number weights, with positive and negative weights indicative of classification towards the query or control set, respectively. All TFBSs with negative weights were assigned a weight of 0 to indicate no classification towards the query set.

5.4 Results and Discussion

5.4.1 The phytohormone–treated soybean root transcriptome

Approximately 290 million 50bp paired–end reads were generated upon sequencing all 7 cDNA libraries. Phred quality (q) scores throughout all lanes were either 38 or 39, indicating base calling accuracy of approximately 99.99% (Table 5.1). Similarly, percentage of reads mapping to the soybean transcriptome consistently ranged between 89% to 92%, averaging a respectable alignment percentage of 90%.

To determine linearity between each replicate, Pearson correlation coefficients (ρ) were derived given read counts following JA, IAA, and control replicates (Figure 5.1a). The high Pearson coefficients associated with these treatments ($\rho \geq 0.98$) thus indicates presence of robust, biologically–sound replicates. Of the 73,026 transcripts and 53,917 genes making up the Phytozome *Glycine max* 1.1 transcriptome build, 6,632 transcripts and 5,498 genes were rendered DE upon treatment with at least one phytohormone (Figure 5.1b). DE transcripts were subsequently stratified based on their log_2 fold–change, f, and classified into one of six categories: heavily induced ($f \geq 5$), moderately induced ($2 \leq f < 5$), lightly induced ($1.5 \leq f < 2$), lightly suppressed ($-2 < f \leq -1.5$), moderately suppressed ($-5 < f \leq -2$), and heavily suppressed ($f \leq -5$). Across all treatments, heavily induced and heavily suppressed transcripts were the least abundant transcript group (Figure 5.1c). On the contrary,

Table 5.1: Sequencing phytohormone–treated root transcriptomes.

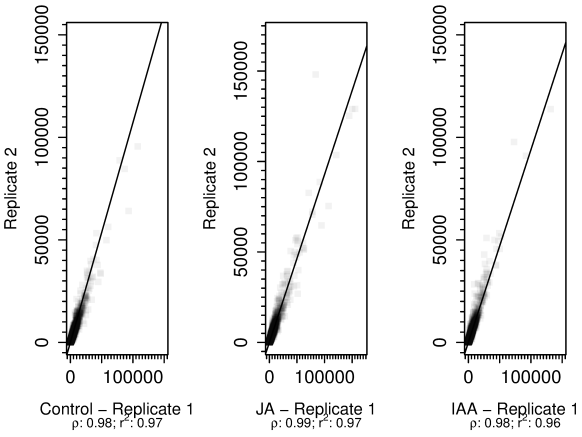
Treatment	Replicate	# reads	Map (%)	SRA	q–score
Untreated	1	42,820,946	89	SRR976388	38
	2	40,705,166	89	SRR976389	38
JA	1	40,631,916	90	SRR976390	39
	2	43,081,146	89	SRR976392	38
IAA	1	40,906,124	91	SRR976393	38
	2	39,896,498	92	SRR976395	38
ETH	1	39,837,842	90	SRR976396	38
Total	–	287,879,638	μ =90	–	μ =38

moderately induced and lightly induced transcripts were the most abundant category, but only after JA and IAA treatment. ETH treatment however resulted in increased abundance of lightly suppressed and moderately suppressed transcripts.

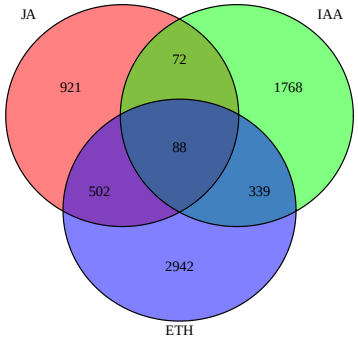
Of the 6,632 DE transcripts, 88 transcripts were DE following all treatments (Figure 5.1d). Almost one–third of these transcripts (n=34) were collectively induced. Of these transcripts, 15 transcripts were identified to functionally encode numerous enzymes central to metabolism and biosynthesis, such as phosphoethanolamine (EC: 3.1.3.75), glucosucrase (EC: 3.2.1.26), glutathione S-transferase (EC: 2.5.1.18), and flavonol synthase (EC: 1.14.11.23). Thus, there exists a core set of induced transcripts following IAA, ETH, and JA treatment which collectively contribute towards the regulation of critical basal operations.

5.4.2 Ontology analysis captures systematic phytohormone interplay

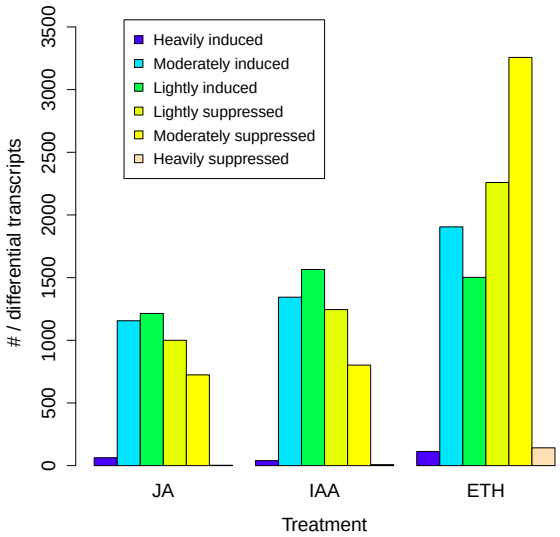
Following all treatments, the top 200 most induced transcripts DE solely after at least one treatment were identified and termed a transcript set. For each transcript set, the top 20 enriched Gene Ontology (GO) Processes were identified within induced transcripts (Figure 5.2a). Such results could therefore provide a glimpse into systematic interplay orchestrated solely by each individual phytohormone. Of the 20 enriched GO Processes within induced transcripts, 10 were exclusively enriched following IAA treatment. Such GO enrichments were exclusively associated with classical IAA regulatory roles such as root development



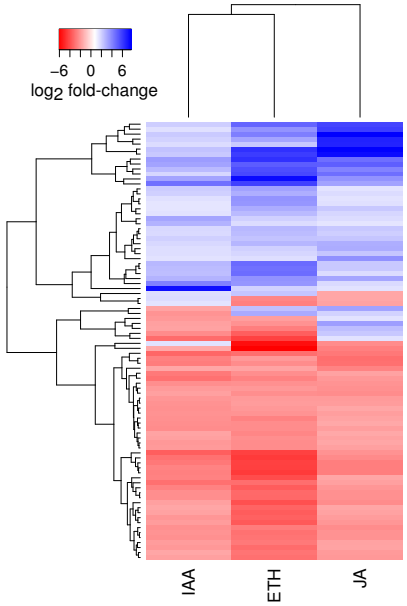
(a) Linearity between replicate treatments.



(b) Differential transcript distribution.



(c) Classifying differential transcripts.

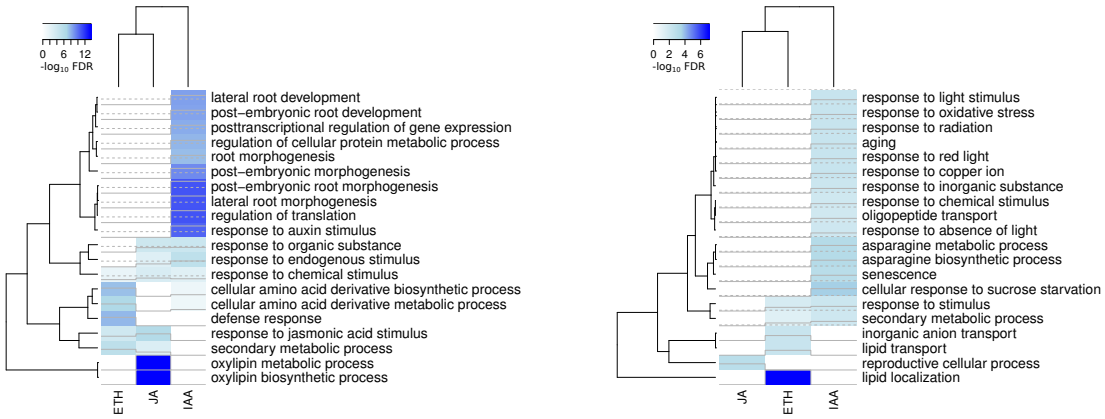


(d) Treatment–exclusive expression profiles.

Figure 5.1: RNA–Seq analysis of phytohormone–treated soybean roots.

and post–embryonic differentiation. An additional 2 enrichments ("response to jasmonic acid stimulus" and "secondary metabolic process") were statistically significant following both JA and ETH treatments. This mutually–shared set of ontologies should come as no surprise considering the well studied co–regulation between JA and ETH [182–184]. Two

of the most enriched GO Processes within induced transcripts were "oxylipin metabolic process" and "oxylipin biosynthetic process", both exclusive to transcripts following JA treatment. This should come as no surprise since oxylipins represent a family of fatty acids such as JA, signaling molecules known for their ability to regulate stress response signaling [185–187].



(a) Induced transcripts.

(b) Suppressed transcripts.

Figure 5.2: GO Processes within transcripts following at least one treatment.

Contrasting GO enrichments within induced transcripts to those from suppressed transcripts yields pronounced differences between the two enrichment sets (Figure 5.2b). To begin with, almost all (n=14) of the 20 enriched GO Processes appear within suppressed transcripts solely following IAA treatment. Of the remaining 6 Processes, 3 were enriched exclusively following ETH treatment.

5.4.3 Outlier transcripts following treatment–pairs

Of the 1,001 transcripts DE following any two treatments, A and B, outlier analysis revealed numerous transcripts which exhibited pronounced expression variances (Figure 5.3). Such results shed light on systematic interplay by revealing expression profiles that could potentially capture individual hormone–driven signaling dynamics.

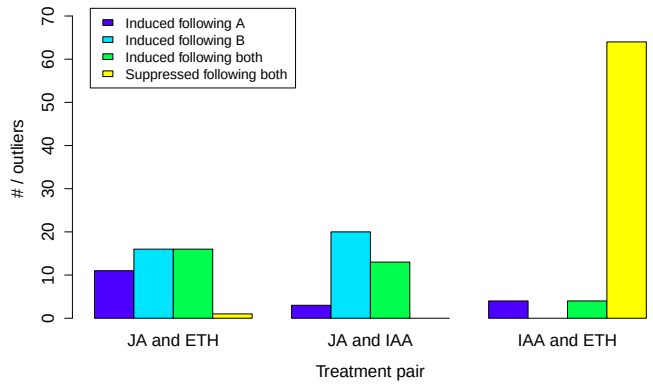


Figure 5.3: Distribution of outlier transcripts following treatments.

Outlier transcripts following JA and ETH treatment

A total of 44 transcripts were rendered outliers following JA and ETH treatment. Of this set, 16 transcripts were induced following both treatments (Figure 5.4a). An additional 11 transcripts were identified to be induced following JA treatment but suppressed following ETH treatment. One such loci, Glyma02g09130.2, encodes alliin lyase (EC: 4.4.1.4), an enzyme shown to play critical roles in herbivory stress response [188,189]. An additional subset of 16 transcripts were induced following ETH treatment and suppressed following JA treatment. Such transcripts encoded numerous transcription factors (TFs) critical to ethylene interplay. Four genomic loci from this set (Glyma13g18330.1, Glyma13g18370.2, Glyma19g34670.2, Glyma10g04170.2) encoded an AP2/ERF TF, a regulatory protein critical to ethylene signaling [190, 191]. Two additional soybean loci (Glyma14g06710.1, Glyma01g32310.1) encoded shikimate O–hydroxycinnamoyltransferase (EC: 2.3.1.133) and peroxidase (EC: 1.11.1.7), respectively. Both enzymes are well–known for their roles in flavonoid and phenylpropanoid biosynthesis, two key defense signaling pathways [7].

Outlier transcripts following JA and IAA treatment

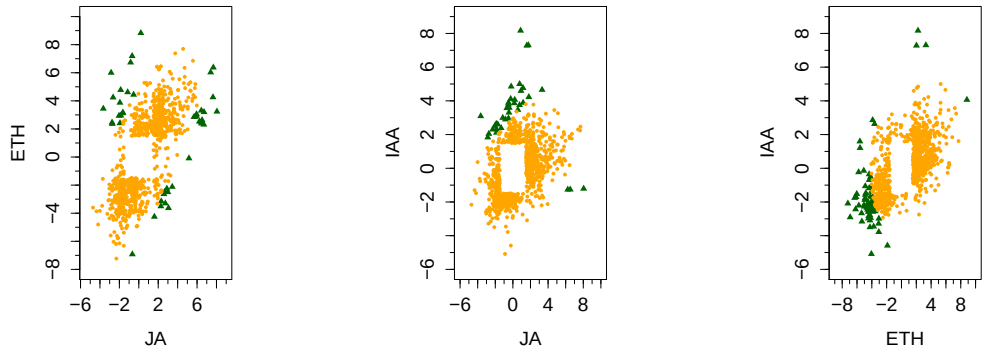
A set of 36 transcripts were identified as outliers following both JA and IAA treatment. Interestingly, only 3 transcripts (Glyma13g06880.1, Glyma07g05740.1, Glyma03g27120.1) were induced after JA treatment but suppressed after IAA treatment (Figure 5.4b). Of the remaining 33 transcripts, 20 were induced following IAA treatment but suppressed following JA treatment. Surprisingly, five of these 20 transcripts (Glyma14g06710.1, Glyma13g18330.1, Glyma13g18370.2, Glyma10g04170.2, Glyma19g34670.2) were also outlier transcripts following JA and ETH treatment. Of these 20 outliers, the most induced following only IAA treatment was Glyma16g03600.1, a vital gene encoding 1–aminocyclopropane–1–carboxylate synthase (EC: 4.4.1.14), a key regulator of ETH biosynthesis [19, 192]. Lastly, 13 of the 36 total outlier transcripts were induced following both IAA and JA treatment. Four of these transcripts (Glyma17g18040.1, Glyma05g21680.1, Glyma02g13910.1, Glyma17g18080.1) encoded GH3 proteins which have been shown in soybean plants to undergo rapid induction following IAA treatment [193].

Outlier transcripts following ETH and IAA treatment

A set of 72 outliers were identified following both ETH and IAA treatment. Within this set, four soybean transcripts (Glyma02g09130.2, Glyma18g06230.1, Glyma13g34221.1, Glyma06g14640.1) were induced solely after IAA treatment but suppressed after ETH treatment (Figure 5.4c). A further four outlier transcripts (Glyma05g21680.1, Glyma17g18080.1, Glyma17g18040.1, Glyma10g04160.1) were induced following both IAA and ETH treatment. Surprisingly, the last transcript in this set encodes an AP2/ERF TF, a family of proteins with well–studied roles in host stress and defense response.

5.4.4 Proximal binding sites shed light on defense–response signaling

For each treatment, a LASSO regression model was built to classify over–represented TFBSs to either promoter sequences of the query set or control set. TFBSs classified to the former set were assigned a positive LASSO weight whereas TFBSs classified to the latter



(a) JA and ETH. (b) JA and IAA. (c) ETH and IAA.

Figure 5.4: Outlier soybean transcripts following pairs of treatments.

set were assigned a negative LASSO weight. Thus, magnitude of the LASSO weight is indicative of the degree of classification to one set over the other. To quantify robustness of the LASSO classifier, area under the curve (AUC) of the receiver operating characteristic (ROC) curve was computed. An AUC score in the range of 1.0 therefore indicates excellent classification. AUC scores given the ETH, IAA and JA LASSO models were 0.87, 0.86, and 0.82, respectively. Of the 71 TFBS PWMs used in this study, less than half (n=33) were over–represented in promoter sequences of transcripts induced after at least one treatment (Figure 5.5).

AP2 / EREBP binding site over–representation

The APETALA2 / ethylene responsive element binding protein (AP2 / EREBP) TF family has been shown to be play critical roles in plant defense signaling [194–197]. One such TFBS, ABI4–1, exhibited over–representation within promoters of transcripts following both JA and ETH treatment. Fellow AP2 / EREBP TFBS, RAV1–2, exhibited similar over–representation but was also over–represented in promoters following IAA treatment. Functional annotation of 142 soybean loci with at least 3 occurrences of the RAV1–2 TFBS in their promoter sequence revealed these loci encode gene products associated with defense response signaling (Table 5.2). Such results were supported by an earlier study which

probed the role of RAV1–2 TFs in stress response following infection of pepper leaves with *Xanthomonas campestris* [198].

Table 5.2: Differential genes with RAV1–2 binding sites.

Phytozome ID	EC	EC Description	Function
Glyma20g30835.3 Glyma03g31912.1 Glyma10g04160.1 Glyma13g18330.1	–	–	AP2/ERF TF
Glyma15g05520.1 Glyma08g19500.1 Glyma06g00880.1	–	–	IAA–inducible protein
Glyma08g19980.1 Glyma13g18340.1	–	–	ETH response factor
Glyma03g33340.2 Glyma10g05480.3	2.3.1.74	Chalcone synthase	Chalcone synthase
Glyma13g34420.1 Glyma07g37280.1 Glyma17g03330.1 Glyma03g05530.1 Glyma15g06830.1	–	–	Pathogenesis–related

In contrast, ABI4–1 TFBSs were less over–represented. Numerous transcripts containing this TFBS in their promoter encoded gene products involved in defense response, namely 1–aminocyclopropane–1–carboxylate (ACC) synthase, chalcone synthase, glutathione S-transferase (GST), and auxin–inducible proteins (Table 5.3).

bZIP binding site over–representation

Much like AP2/EREBP TFs, bZIP proteins have also been shown to regulate aspects of defense signaling [199,200]. bZIP TFBSs bZIP911, O2, and TGA1 were collectively classified within induced promoters following all three treatments. TGA1 TFBSs were however found at least twice in 323 soybean promoters following all treatments. Sequence analysis of these loci indicated presence of statistically significant annotations associated with ethylene signaling, defense response, and xenobiotic perception (Table 5.4).

Fellow bZIP TFBSs ABF1 and PEND were over–represented exclusively following ETH

Table 5.3: Differential genes with ABI4–1 binding sites.

Phytozome ID	EC	EC Description	Function
Glyma05g23020.1 Glyma17g16990.1	4.4.1.14	ACC synthase	ACC synthase
Glyma02g16090.1 Glyma10g03710.2	–	–	IAA–inducible protein
Glyma08g11630.2	2.3.1.74	Naringenin–chalcone synthase	Chalcone synthase
Glyma03g33340.2	2.5.1.18	Glutathione S–transferase	Glutathione transferase
Glyma08g15600.1	2.7.11.1	Serine/threonine protein kinase	WD repeat-containing protein

Table 5.4: Differential genes with TGA1 binding sites.

Phytozome ID	EC	EC Description	Function
Glyma01g44270.1	6.2.1.12	4-coumarate–CoA ligase	4-coumarate–CoA ligase
Glyma14g05350.1 Glyma14g05355.1 Glyma08g05500.1	1.14.11.23	Flavonol synthase	ACC–oxidase
Glyma05g21680.1 Glyma17g18080.1 Glyma17g18040.1	–	–	Auxin–responsive GH3
Glyma08g11630.2 Glyma08g11530.1 Glyma05g28610.2	2.3.1.74	Chalcone synthase	Chalcone synthase
Glyma18g49761.1 Glyma13g28810.2 Glyma19g34650.1	–	–	ETH–responsive TF
Glyma07g03920.2 Glyma08g20210.1 Glyma08g20220.1 Glyma08g20190.1 Glyma04g11640.1 Glyma04g11870.1	1.13.11.40	Arachidonate 8-LOX	Lipoxygenase
Glyma07g00920.1 Glyma08g20220.2 Glyma08g20200.2 Glyma07g00886.1 Glyma14g31400.1	1.13.11.12	Linoleate 13S–LOX	Lipoxygenase

and JA treatment, respectively. More specifically, a total of 128 soybean loci had at least one occurrence of the PEND TFBS in their promoter sequence. Aside from the role PEND plays in DNA binding within chloroplast [201], little is known about its regulatory contributions towards defense signaling. Surprisingly, functional annotation of these 128 PEND–containing loci revealed strikingly similar GO Processes to that of TGA1–containing loci (Table 5.5). Thus, further investigation examining PEND defense response cross-talk could provide novel insight into JA–driven systematic regulation. As far as ABF1 TFBSs are concerned, these TFBSs were less represented than PEND, being found at least twice in 54 soybean promoters. Annotating such transcripts revealed generally similar annotations to that of PEND and TGA1–containing loci.

Table 5.5: Differential genes with PEND binding sites.

Phytozome ID	EC	EC Description	Function
Glyma05g36310.1	1.14.11.23	Flavonol synthase	ACC–oxidase
Glyma14g08560.1	4.2.1.92	Hydroperoxide dehydratase	Allene–oxidase synthase
Glyma04g42990.1 Glyma06g11760.1 Glyma08g15440.1 Glyma06g00880.1	–	–	IAA–inducible protein
Glyma08g20190.1 Glyma13g42340.1 Glyma07g03910.1	1.13.11.40	Arachidonate 8-LOX	Lipoxygenase
Glyma07g00920.1 Glyma08g20200.2 Glyma07g00886.1 Glyma08g20230.2	1.13.11.12	Linoleate 13S–LOX	Lipoxygenase
Glyma16g24831.2	2.1.1.141	Jasmonate O–methyltransferase	Salicylic acid methyl–transferase–like
Glyma08g01430.1	–	–	WRKY TF

MYB binding site over–representation

MYB TFs are key regulators of biotic stress response. In total, 7 MYB TFBSs were represented from which over–representation of 2 TFBSs (GAMYB, AtMYB61) were statistically

significant. The remaining 5 TFBSs (TaMYB80, ARR10, MYB83, MYB46, AtMYB77) exhibited nominal over–representation following all treatments.

A total of 31 and 32 soybean loci had at least once occurrence of the GAMYB TFBS within their promoter sequence following IAA and ETH treatment, respectively. AtMYB61 TFBSs were found at least three times in 106 and 132 soybean loci promoter sequences following JA and IAA treatment, respectively. Functional annotation of the former 106 loci revealed enriched levels of jasmonic acid signaling processes as well as gene products classically involved in defense response (Table 5.6). On the other hand, the latter set of 132 soybean loci revealed a pronounced catalog of processes, with almost all associated with regulation of lateral shoot differentiation and development.

Table 5.6: Differential genes with AtMYB61 binding sites.

Phytozome ID	EC	EC Description	Function
Glyma08g46610.1 Glyma08g46610.2 Glyma14g05350.1	1.14.11.23	Flavonol synthase	ACC-oxidase
Glyma08g15440.1	–	–	IAA–induced protein
Glyma08g11630.2 Glyma08g11530.1	2.3.1.74	Chalcone synthase	Chalcone synthase
Glyma07g04630.1	–	–	JAZ2
Glyma01g41290.1 Glyma04g01531.1 Glyma13g17640.1 Glyma17g04850.1 Glyma17g05540.1	–	–	TIFY

WRKY binding site over–representation

Otherwise known as WRKY1, ZAP1 is a member of the WRKY protein family, a TF family known for its ability to drive defense–response signaling [67, 202]. The ZAP1 TFBS was found at least three times in 79 soybean loci promoters following all treatments. Collectively, these transcripts encoded genes involved in phenylpropanoid and flavonoid biosynthesis, vital pathways central to biotic stress signaling (Table 5.7).

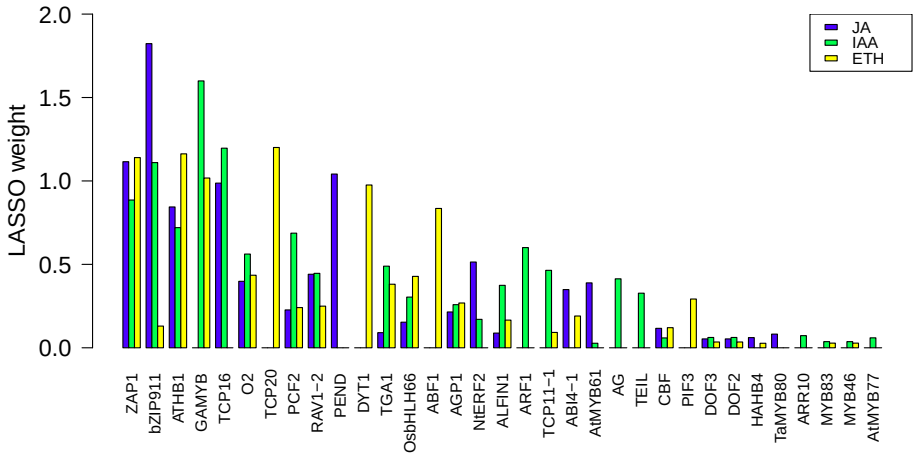


Figure 5.5: Over–represented binding sites following phytohormone treatments.

Table 5.7: Differential genes with ZAP1 binding sites.

Phytozome ID	EC	EC Description	Function
Glyma07g05420.1	1.14.11.23	Flavonol synthase	ACC–oxidase
Glyma15g01560.1	–	–	IAA–induced protein
Glyma08g11630.2	2.3.1.74	Naringenin–chalcone synthase	Chalcone synthase
Glyma13g28810.2 Glyma17g02711.1	–	–	ETH–responsive TF
Glyma05g29400.1	2.5.1.18	Glutathione S–transferase	Glutathione transferase
Glyma13g34420.1	–	–	Pathogen–related
Glyma10g27860.5	–	–	WRKY TF

5.5 Discussion

Hormone–driven signaling within plants is an exquisite, tightly regulated orchestra of small metabolites working in–concert to trigger transcription of downstream gene products. Numerous plant hormones, namely phytohormones, officiate such interplay. Well–studied phytohormones include ethylene, auxin, salicylic acid, jasmonic acid, gibberellin, and brassinosteriods. Each phytohormone is armed with a unique transcriptional repertoire designed to regulate crucial aspects of plant development, from seed germination to defense response

and cell death. The plant signaling community has made ever so increasing strides when it comes to deciphering phytohormone signaling mechanisms. Coupled with advances in next–generation sequencing assays, individual actors involved in such signaling can now be quantified, providing potentially novel transcriptomic snapshots.

In this study, soybean (*Glycine max*) roots were treated with three well–studied phytohormones: ethylene (ETH), auxin (IAA), and jasmonic acid (JA). Sequencing the transcriptome of treated roots produced almost 300 million paired–end reads. Expression analysis of transcripts following each treatment revealed presence of 88 transcripts differentially expressed following all three treatments. Functional analysis of these transcripts revealed a majority of transcripts encode critical gene products vital to defense response and secondary metabolite biosynthesis. Further analysis of transcripts exclusively expressed after each treatment revealed distinct transcriptional profiles. For instance, IAA–treated roots contained a significant number of differential transcripts encoding gene products involved in root development and embryonic differentiation. On the other hand, JA–treated roots exhibited a significant number of transcripts encoding oxylipin biosynthesis, a family of fatty–acids critical to defense signaling.

Statistical examination of proximal binding site abundance within promoters of induced transcripts reveals over–represented levels of numerous binding sites of transcription factors involved in defense signaling. Amongst the most represented binding sites were ZAP1, bZIP911, GAMYB, O2, and TGA1; all serving as binding sites for genes involved in various defense signaling processes. Interestingly, annotation of loci containing PEND binding sites yielded statistically significant annotations associated with jasmonic acid signaling, even though little is known of the role PEND plays in defense response. Further examination of how PEND co–regulates defense response signaling would therefore be an investigative avenue with potential novelty.

In closing, our results provide a high–coverage snapshot of the plant signaling and regulatory defense–response landscape by sequencing the phytohormone–treated soybean root transcriptome and quantifying proximal regulatory elements.

Chapter 6: Soybean promoter sequences and the defense–response landscape

6.1 Abstract

Plant transcription regulation is a vital biological process that governs every aspect of growth and development, from seed germination, flower differentiation, to senescence and pathogen perception. Regulatory proteins known as transcription factors drive transcription regulation by binding onto non–coding genomic regions within the target gene promoter sequence and recruit additional regulatory factors which collectively transcribe the target gene. The sessile nature of plants therefore necessitates fine–tuned regulatory machinery in the face of stimuli be–it positive or negative stress. In this study, we perform genome–wide computational analysis on soybean (*Glycine max*) promoter sequences with intent of cataloging binding sites present within host defense response elicitors. We reveal binding profiles of soybean genes involved in defense response, such as glutathione S–transferase, flavonol reductase, hexokinase, and allene oxide cyclase.

6.2 Introduction

Regulating transcription of plant genes is a dynamic and non–linear biological process orchestrated by proteins known as transcription factors (TFs). These regulatory proteins bind to non–coding genomic regions adjacent to the target gene known as transcription factor binding sites (TFBSs). This delicate dynamic is of utmost importance throughout the life of a plant, driving virtually every aspect of its development. During defense response, the plants regulatory machinery is put on high–alert as it goes into overdrive regulating synthesis of hundreds of small metabolites that collectively meet the pathogen head–on.

In a stunning show of events, a pathogen can even commandeer and manipulate such dynamics, fooling the plant into transcribing a weaker set of defense elicitors. For instance, the whitefly parasitoid, *Encarsia formosa*, has been shown to suppress *Arabidopsis thaliana* jasmonic acid defense signaling but induce salicylic acid signaling which is a less effective *Encarsia formosa* countermeasure [203]. Therefore, it goes without saying that transcription regulation is a delicate network comprised of many TF families, TFBSs, and signaling cascades.

The promoter sequence of a transcript is a non–coding stretch of genomic sequence that serves as the starting point for transcription. Numerous software tools have been developed to analyze promoter sequences and derive statistically significant TFBSs [31, 37, 80, 204]. By utilizing such analytical software tools, promoter sequences of entire genomes could be analyzed, leading to insight as to which TFBSs are present within particular loci. In this study, we performed genome–wide analysis of soybean promoter sequences to identify abundant patterns of TFBSs with respect to soybean loci functionally involved in defense response. Numerous TFBSs of TFs empirically associated with defense response were analyzed, identifying numerous functional transcripts which could possibly serve as potentially novel gene candidates likely involved in defense elicitation. Our results therefore aim to provide a genomic–scale investigation probing the interplay between soybean transcripts involved in defense and associated binding sites.

6.3 Results

6.3.1 Identification of binding sites in soybean promoter sequences

For each of the 73,320 transcript isoforms making up the *Glycine max* transcriptome build, the promoter sequence 2kb upstream from its transcription start site was identified. A collection of 71 position weight matrices (PWMs) were retrieved to statistically model TFBS abundance in promoter sequences. The Marina software [127] mapped all PWMs onto each soybean promoter sequence and produced an abundance matrix that enumerates TFBS

counts in each promoter sequence. Of the 71 TFBS PWMs, 64 passed default filtering criterion set by Marina. Of these, 40 (62%) were present in at least one soybean promoter sequence. This low percentage could be attributed to the fact that numerous PWMs were from plants distantly related to soybean, namely rice (*Oryza sativa*) and corn (*Zea mays*).

To identify abundant TFBSs within promoter sequences, TFBSs found at least 3 times in any promoter of soybean loci were enumerated. TFBS counts, c, were subsequently partitioned into three bins: low abundance ($c \leq 95$), medium abundance ($95 < c \leq 2,452$), and high abundance ($c \geq 2,453$). The most abundant TFBSs appear to be ATHB5, HAHB4, ATHB1, and HAT5, each found three or more times in promoters of at least 58,450 loci (Figure 6.1). A plausible explanation for such pronounced abundance could be attributed to the homogeneous nature of their PWMs. For instance, the ATHB5 PWM (Figure 6.2a) has weights heavily skewed to a particular nucleotide. A similar skew is present with the HAT5 PWM (Figure 6.2b). Thus, explicitly weighted bases often yield many ubiquitous false–positive mappings since the resulting short, consensus motif may occur spuriously throughout a promoter sequence.

6.3.2 Proximal binding sites and the defense–response landscape

Analysis of soybean promoter sequences revealed numerous TFBSs serving as docking sites of proteins induced during pathogen response. Such TFBSs were classified into distinct groupings based on the TF the binding site interacts with. Functional annotation of transcripts containing various TFBSs in their promoter sequence revealed biologically–sound insight into not only the transcriptomic defense response landscape, but also *cis*–regulatory dynamics involving nearby regulatory elements.

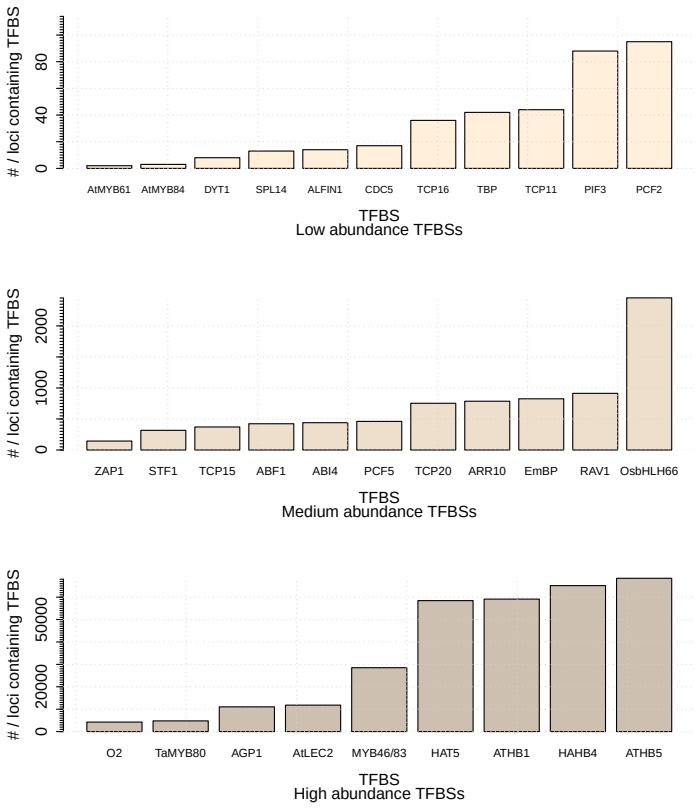


Figure 6.1: Enumerating TFBSs found at least 3 times in soybean promoters.

AP2/EREBP binding sites

Like many regulatory proteins, APETALA 2 / ETHYLENE RESPONSE ELEMENT BINDING PROTEIN (AP2 / EREBP) TFs play key roles in regulating key aspects of plant development, ranging from flower differentiation to biotic response [205–207]. Of the 4 AP2/EREBP TFBS PWMs present in this study (ANT, ABI4, RAV1, EmBP1), the ANT TFBS was the least abundant, found in promoters of four loci (Glyma14g33170.1, Glyma07g20411.1, Glyma15g02350.1, Glyma15g22975.1). The latter three TFBSs, ABI4, RAV1, and EmBP1, were found at least twice in 440, 913, and 825 soybean loci, respectively. Gene Ontology (GO) enrichments "sulfur amino acid biosynthetic process" and "protein import" were exclusively associated with loci containing ABI4. On the other hand, annotating

(a) ATHB5. (b) HAT5.

Figure 6.2: Sequence logos of abundant TFBSs.

transcripts with RAV1 in their promoter sequences revealed 23 significant GO Processes such as "auxin mediated signaling pathway", "cellular response to chemical stimulus", and "response to auxin stimulus". A total of 26 transcripts made up these 3 enrichments, encoding various defense response elicitors such as glutathione S–transferase (GST) and six auxin–responsive proteins (Table 6.1). Thus, such results may further our understanding of the role transcripts with RAV1 TFBSs play in auxin–driven defense response [198].

Table 6.1: Soybean genes with AP2/EREBP binding sites.

Phytozome ID	EC	EC Description	PFAM Description
Glyma14g39090.1	2.5.1.18	GST	GST
Glyma08g21450.1	2.7.1.1	hexokinase	Hexokinase
Glyma12g14900.1 Glyma09g35580.1 Glyma08g16490.1 Glyma06g43270.1 Glyma06g43220.1 Glyma05g27580.1	–	–	Auxin responsive protein

ABI3/VP1 binding sites in loci encoding auxin signaling

ABI3/VP1 TFs represent proteins which mediate auxin signaling [208]. The ABI3/VP1 TF, ARF1, is of particular interest in plant defense response as it has been shown to bind auxin responsive elements [46, 209]. ARF1 TFBSs were found in 1,826 soybean promoters, but

were not found occurring more than 3 times in any one promoter sequence. GO analysis of loci containing this binding site revealed most transcripts being involved in processes such as "auxin mediated signaling pathway", "phenylpropanoid metabolic process", "phenylpropanoid biosynthetic process", "hormone-mediated signaling", and "pathway response to auxin stimulus". Transcripts mapping to such processes encoded cinnamyl–alcohol dehydrogenase (CAD), glutathione–S transferase, as well as various proteins induced in the presence of auxin (Table 6.2).

Table 6.2: Soybean genes with ABI3/VP1 binding sites.

Phytozome ID	EC	EC Description	PFAM Description
Glyma10g40870.1	1.1.1.195	CAD	Zinc–binding dehydrogenase
Glyma15g40220.1	2.5.1.18	GST	GST
Glyma06g43310.1 Glyma06g43400.1 Glyma06g43470.1 Glyma08g16490.1 Glyma09g35310.1 Glyma09g35370.1 Glyma09g35510.1 Glyma12g03870.1	–	–	Auxin responsive protein

bZIP binding sites in loci vital to stimulus perception

Basic Leucine Zipper (bZIP) TFs, such as O2, bZIP911, ABF1, and STF1, encompass proteins which regulate a diverse set of basal plant operations, from seed maturation, to light perception, and pathogen response [199, 210]. GO enrichments within transcripts containing O2 and ABF1 TFBSs generally exhibited similar GO enrichment levels, sharing 71 GO Processes between each other. Of these, the most significant ontologies ($p < 0.001$) were "response to abiotic stimulus", "inositol phosphate-mediated signaling", "response to cytokinin stimulus", "response to endogenous stimulus", "response to jasmonic acid stimulus", and "jasmonic acid mediated signaling pathway". A set of 50 soybean transcripts mapping to the latter two enrichments were identified and functionally annotated. From

this subset, 8 transcripts contained the No Apical Meristem (NAM) domain (Table 6.3). NAM domains have been shown to contribute towards actively regulating plant development and stress [211]. An additional two transcripts encoded glutaredoxins (EC: 1.20.4.1), small enzymes with noted roles in the mediation of SA–JA cross–talk [212, 213]. Further examination of such soybean transcripts could therefore provide potentially novel insight into glutaredoxin–driven host–pathogen interplay.

Table 6.3: Soybean genes with bZIP binding sites.

Phytozome ID	EC	EC Description	PFAM Description
Glyma12g33510.1 Glyma15g08520.1	1.20.4.1	glutaredoxin	Glutaredoxin
Glyma05g32850.1 Glyma06g11970.1 Glyma06g16440.1 Glyma06g38410.1 Glyma12g22880.1 Glyma12g35000.1 Glyma13g35550.1 Glyma14g24220.1	–	–	No Apical Meristem (NAM)

Diverse functionalities in loci containing MYB binding sites

A set of 8 MYB TFBSs (MYB46/83, TaMYB80, ARR10, CDC5, AtMYB84, AtMYB15, AtMYB61, AtMYB77) were identified at least once within all soybean promoters. MYB46/83 was the most abundant MYB TFBS, found at least three times in 28,516 soybean promoters. On the other hand, AtMYB77 was the least abundant, found only in 139 promoter sequences. MYB proteins have been well–studied and their defense response and regulatory roles have been cataloged [162, 214, 215]. GO enrichment on transcripts containing any of these 8 TFBS reveals pronounced variances in transcription functionality (Figure 6.3). TFBSs namely ARR10 and AtMYB84 were found in transcripts predominately associated with amino acid metabolism. Transcripts solely containing ARR10 in their promoters revealed such transcripts to be enriched for processes associated with protein ubiquitination

and modification. Virtually all enrichments were exclusively present within transcripts containing a specific TFBS. This alone captures the diverse systematic and independent roles regulated by MYB proteins. For instance: transcripts containing MYB46/83 were exclusively associated with defense response signaling. An earlier study investigating *Arabidopsis thaliana* interplay with the pathogen *Botrytis cinerea* revealed *myb46* mutants confer increased susceptibility upon perception of this fungal pathogen [216, 217].

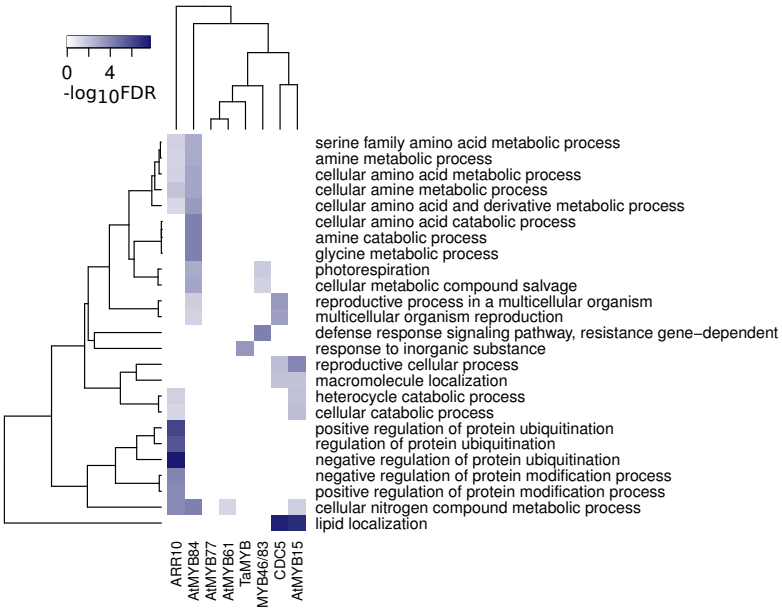


Figure 6.3: Loci with MYB TFBSs captures numerous basal processes.

bHLH binding sites in loci critical to biotic stress response

Basic Helix–Loop–Helix (bHLH) TFs have been shown to be involved in numerous signaling pathways ultimately leading to the plant defense phytohormone, jasmonic acid [218, 219]. Only 3 bHLH PWMs were available: OsbHLH66, PIF3, DYT1. Of these, the most abundant TFBS was OsbHLH66, being found in 15,757 promoters or 15.5% of all promoter sequences. DYT1 however was the least abundant, being found in 913 promoters. GO analysis of transcripts with at least one occurrence of these TFBSs in their promoter sequence revealed

stark differences in transcript functionality (Figure 6.4). Generally, transcripts containing PIF3 and OsbHLH66 TFBSs in their promoters tend to execute approximately similar systematic functionalities, ranging from mediating cytokinin perception, to regulating jasmonic acid signaling. Filtering soybean transcripts containing both PIF3 and OsbHLH TFBSs and which mapped to "response to wounding", "response to jasmonic acid stimulus", "jasmonic acid mediated signaling pathway", and "response to abiotic stimulus" produced a set of 217 transcripts. Several transcripts within this set encoded enzymes involved in stress signaling and synthesis of metabolites such as allene oxide cyclase (AOC), glutathione S-transferase, anthocyanidin synthase, and flavonoid 3'—hydroxylase (Table 6.4). Besides functional analysis of transcripts containing PIF3 and OsbHLH66 TFBSs, transcripts containing DYT1 appear to be exclusively associated with maintaining ion homeostasis. Little information is known of the role DYT1 proteins play in defense response, however bHLH TFs are a diverse family of regulatory proteins controlling virtually every aspect of the plant developmental process [220, 221].

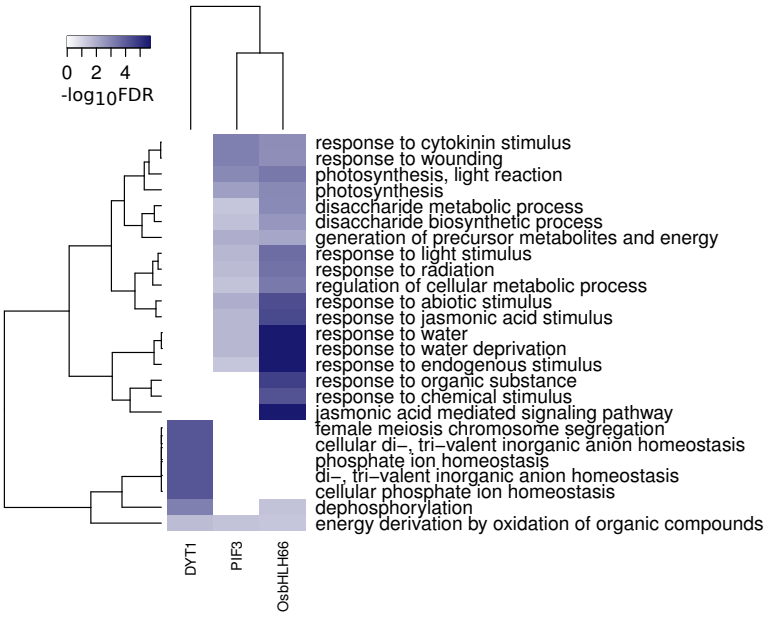


Figure 6.4: Annotation of transcripts containing various bHLH binding sites.

Table 6.4: Soybean genes with bHLH binding sites.

Phytozome ID	EC	EC Description	PFAM Description
Glyma01g22160.1	5.3.99.6	AOC	AOC
Glyma15g40200.1	2.5.1.18	GST	GST
Glyma01g42350.1	1.14.11.19	anthocyanidin synthase	2OG–Fe(II) oxygenase
Glyma17g08550.1	1.14.13.21	flavonoid–3'–hydroxylase	Cytochrome P450

TCP binding sites in loci regulating defense response

TEOSINTE BRANCHED1, CYCLOIDEA, and PCF (TCP) TFs regulate numerous developmental processes in plants, from shoot meristem differentiation [222], to circadian regulation [223], and jasmonic acid signaling [224, 225]. A total of 6 TCP–family TFBSs (PCF2, PCF5, TCP11, TCP15, TCP16, TCP20) were identified within soybean promoters. The least abundant of these TFBSs was TCP16, found in 3,851 or 5.25% of promoter sequences. In contrast, PCF5 was the most abundant TCP TFBS, being found in 12,772, or 17.4% of promoter sequences. A total of 1,040 soybean promoters had at least one occurrence of all 6 TCP TFBSs in their promoters. From within these 1,040 transcripts, GO analysis revealed enrichments associated with defense response, transcription regulation and biosynthesis (Figure 6.5). TCP11, TCP15, TCP20, and PCF2 all appear to be enriched for "response to endogenous stimulus" and "response to jasmonic acid stimulus". Interestingly, 4 transcripts (Glyma03g41700.1, Glyma04g00960.1, Glyma12g35550.1, Glyma19g34370.1) were mapped to the latter enrichment. The first transcript in this set, Glyma03g41700.1, encoded an auxin–inducible GH3 protein family [226], while the latter two, Glyma12g35550.1, Glyma19g34370.1 contained AUX/IAA protein domains (Table 6.5). Thus, transcripts containing TCP binding sites may possibly interplay with auxin and jasmonic acid to enable auxin–driven or jasmonic acid–driven defense signaling. Prior studies have examined TCP–auxin cross–talk during plant development [227, 228], however examining such interplay strictly with respect to defense response would certainly be a potentially novel area of

plant–pathogen inquiry. Surprisingly, many enrichments appear to be lower in magnitude within transcripts containing either TCP16 or PCF5. Nonetheless, the vast number of GO enrichments associated with biosynthetic regulation captures the systematic and diverse role TCP TFs play in plant development and defense.

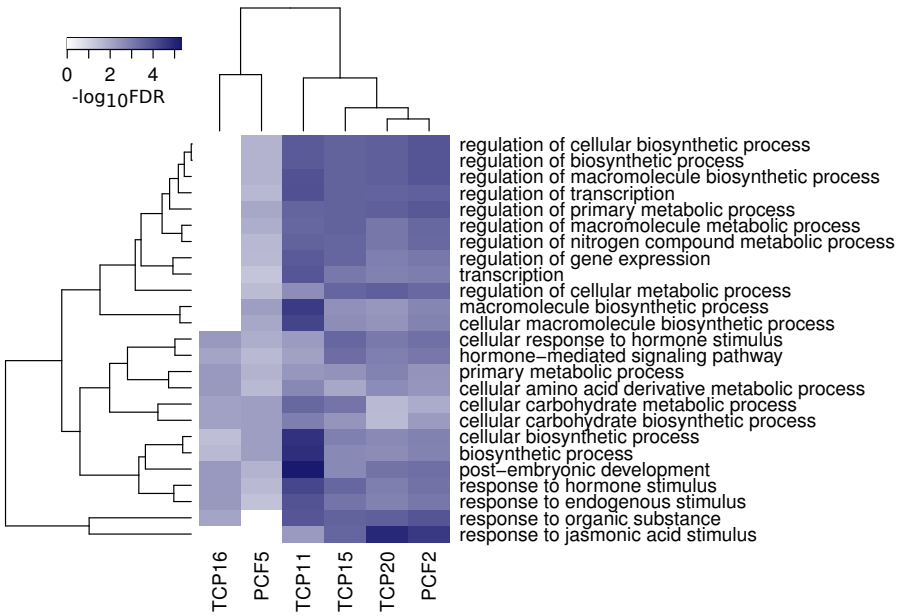


Figure 6.5: Annotation of transcripts containing TCP and PCF binding sites.

Table 6.5: Soybean genes with TCP binding sites.

Phytozome ID	EC	EC Description	PFAM Description
Glyma03g41700.1	–	–	GH3
Glyma12g35550.1 Glyma19g34370.1	–	–	AUX/IAA domain

TBP binding sites in loci encoding phenylpropanoid biosynthesis

The TATA–box binding protein (TBP) is central to regulating assembly of the transcription initiation complex. Numerous studies have examined systematic interplay involving TBP

and its ability to regulate defense response [229–231]. GO enrichment on soybean transcripts with at least one occurrence of the TBP TFBS in its promoter sequence revealed statistically significant annotations involved in multicellular differentiation and metabolism. Making up such processes included "phenylpropanoid biosynthetic process" and "flavonoid biosynthetic process". A set of 24 transcripts mapped to both these processes, encoding enzymes such as 4–coumarate–CoA ligase (4CL), hydroxyindole methyltransferase, and flavonol reductase (Table 6.6).

Table 6.6: Soybean genes with TBP binding sites.

Phytozome ID	EC	EC Description	PFAM Description
Glyma18g08550.1	6.2.1.12	4CL	AMP–binding enzyme
Glyma19g45000.2	2.1.1.4	hydroxyindole methyltransferase	O–methyltransferase
Glyma11g29460.1	1.2.1.44	flavonol reductase	NAD–dependent epimerase
Glyma15g16490.1 Glyma13g21120.1 Glyma06g11590.1 Glyma09g05170.1 Glyma02g15390.1	1.14.11.23	flavonol synthase	2OG-Fe(II) oxygenase

WRKY1 and PR–1 interplay

ZAP1, also known as WRKY1, was the only WRKY TFBS with an available PWM. In terms of stress response, WRKY TFs play vital roles in regulating synthesis of defense–response elicitors such as pathogenesis–related 1 (PR–1) genes [67,169]. Within all promoter sequences, ZAP1 TFBSs occurred at least twice in 1,197 promoters. GO enrichment on this set revealed highly significant processes involved in regulation of both transcription and protein kinase cascades. Processes such as "phenylpropanoid biosynthesis" and "aromatic compound biosynthesis" were amongst such enrichments. A subset of 22 transcripts mapped to lesser–enriched processes ($p < 0.013$), namely "phenylpropanoid metabolic process", "positive regulation of response to stimulus", and "regulation of immune response". Such transcripts encoded anthocyanin reductase and a MAP2K protein (Table 6.7).

Table 6.7: Soybean genes with WRKY/ZAP1 binding sites.

Phytozome ID	EC	EC Description	PFAM Description
Glyma07g19180.1	2.7.11.1	serine/threonine protein kinase	Leucine Rich Repeat
Glyma08g06630.1	1.3.1.77	anthocyanin reductase	NAD–dependent epimerase
Glyma19g00220.1	–	MAP2K	Protein kinase domain

6.4 Discussion

In this study, promoter sequences throughout the soybean (*Glycine max*) genome were analyzed to investigate links between binding site specificities and transcript functionality. A set of 40 TFBS PWMs were found to occur in at least one soybean promoter. Filtering this set revealed a subset of 31 TFBS PWMs with three or more occurrences in any given promoter. Stratifying such PWMs into their respective TF family produced 8 distinct categories (AP2/EREBP, ABI3/VP1 bZIP, MYB, bHLH, TCP, TBP, WRKY). GO analysis on transcripts within each of these categories revealed statistically significant annotations involved predominantly in defense–response and stress signaling. Soybean loci mapping to such functions were identified, with most encoding enzymes associated with phenylpropanoid biosynthesis, flavonoid biosynthesis, and jasmonic acid and auxin signaling. Such findings therefore conclude that frequency of certain TFBSs can indeed help decipher functionality of transcripts involved in defense–response.

6.5 Materials and Methods

6.5.1 Acquisition of soybean promoter sequences

A total of 73,320 soybean transcript identifiers were mined from the Phytozome plant genomics resource [91]. Per identifier, the promoter sequence 2kb upstream from its transcription start site was retrieved and appended to a FASTA file.

6.5.2 Identification of over–represented soybean binding sites

TFBSs are frequently modeled as two–dimensional matrices known as position weight matrices (PWMs). Each cell in this data–structure therefore references the likelihood of a nucleotide being part of a regulatory element. A set of 21 PWMs were mined from JASPAR[43] while another 50 were mined from AthaMap[39]. PWMs ≤ 7 columns wide were filtered and not utilized in analysis. The Marina software mapped all valid TFBS PWMs onto all promoter sequences. A PWM alignment was rendered successful if at least 85% of the PWM aligned against the promoter sequence. Custom Python scripts parsed PWM mappings along all promoter sequences and enumerated PWM frequency in the form of an abundance matrix.

6.5.3 Functional annotation of soybean transcripts

Functional annotation comprised of identifying statistically significant GO Processes within soybean transcripts, a process known as GO enrichment. Accession identifiers for such transcripts underwent GO enrichment using the AgriGO web–server [124]. A GO Process was termed enriched if its Hochberg–adjusted p–value was less than 0.05 and at least 5 transcripts were mapped to the respective annotated ontology.

Chapter 7: Conclusions and Future Work

7.1 Research Implications

This dissertation is comprised of five original first–author manuscripts. Each study investigates a specific segment of the soybean proximal *cis*–regulome and transcriptome during biotic stress perception. Collectively, such studies aspire to contribute novel and biologically interesting findings to the plant–pathogen research community.

Chapter 2 proposes a novel software tool named `Marina`. This software identifies statistically over–represented transcription factor binding sites (TFBSs) given a set of 7 statistical metrics. Section 2.2 discusses such metrics in–depth as well as its ability to be used in inferring magnitude of TFBS over–representation. Section 2.3 examines how `Marina` fares against a leading TFBS analysis tool when it comes to identifying over–represented TFBSs. Results indicate that indeed `Marina` identified more over–represented TFBSs even as the number of input promoter sequences increased. Thus, chapter 2 proposes a scalable software tool which yields biologically–sound and over–represented TFBSs.

Chapter 3 examines the soybean transcriptome after inoculation with resistant and susceptible soybean cyst nematode (SCN) populations. Section 3.2 discusses RNA sequencing of the soybean transcriptome post–inoculation. Differential transcripts in each reaction reveal a biologically–sound set of transcripts which capture host defense response dynamics across an inoculation time–course. Section 3.3.2 reveals reaction–specific Gene Ontology (GO) annotations across differential transcripts. Such results reveal the commandeering nature of SCN through its ability to suppress host metabolism and biochemical synthesis processes. Section 3.3.3 revealed presence of numerous over–represented TFBSs within promoter sequences of differentially expressed transcripts. Such over–represented TFBSs could be used to compile a *cis*–regulatory signature which captures proximal soybean regulatory

element dynamics during SCN pathogenesis.

Chapter 4 quantifies the soybean leaf transcriptome following inoculation with soybean rust (SR). Such an investigation builds on prior peer–reviewed manuscripts by computationally analyzing promoter sequences of differential transcripts following an SR time–course. Section 4.4 reveals GO enrichments given soybean transcripts induced and suppressed following SR inoculation. Promoter sequences of such transcripts were analyzed in section 4.4.1, revealing a set of 25 soybean binding sites over–represented following pathogenesis. Thus, the ultimate goal of this study is to be used as a platform for the development of synthetic promoters designed to increase transcription of genes associated with pathogen and stress response.

Chapter 5 quantifies the soybean root transcriptome following inoculation with phytohormones jasmonic acid (JA), auxin (IAA), and ethylene (ETH). Unlike traditional experimentation in plant–pathogen studies in–which a tissue is infected with a deleterious pest, this chapter is dedicated towards understanding the plant defense–response signaling landscape. Section 5.4 reveals a high–coverage, high–quality root transcriptome build following high–throughput sequencing. Section 5.4.2 statistically identifies transcripts with significant changes in fold expression following two independent phytohormone treatments. These differential transcripts were termed "outliers". Across all phytohormone treatments pairs, outlier transcripts capture defense–response interplay and cross–talk. Section 5.4.3 reveals numerous statistically–sound TFBSs over–represented following individual phytohormone treatments. Such TFBSs were found to be not only over–represented but also present in promoter sequences of genes commonly involved in defense response. Thus, chapter 5 provides a high–coverage map of the soybean root transcriptome following numerous phytohormone treatments. Analysis of GO enrichments reveal functions which codify hormone–driven defense signaling. Further analysis of promoter sequences from differential transcripts reveal over–represented levels of TFBSs of TFs induced during defense response. Our findings therefore aim to contribute potentially novel insight into phytohormone–specific expression profiles of numerous transcripts.

Chapter 6 provides a detailed genomic–scale investigation into the binding site profiles of soybean genes involved in defense response. Unlike prior chapters which were built around quantifying host–pathogen interplay, this study is dedicated entirely to the analysis and quantification of promoter sequences within the soybean genome. Section 6.3.1 reveals numerous TFBSs that were found in promoter sequences of soybean genes which play roles in plant defense. Soybean accession numbers mapping to such annotations could therefore be used to drive functional genomic assays be–they over–expression or knockout.

7.2 Conclusions

This dissertation is a collective series of manuscripts exploring the soybean transcriptome and proximal regulome upon inoculation with two major host pathogens: soybean rust (SR) and soybean cyst nematode (SCN). Results reveal numerous biologically–sound transcripts involved in defense response such as glutathione S–transferase, lipoxygenase, hexokinase, and flavonoid synthase, amongst others. Several promising stress response gene candidates such as arachidonate–8 lipoxygenase, phytochelatin synthetase, and ribonucleoside-diphosphate reductase may also exhibit defensive properties and would be excellent candidates for empirical validation. Quantifying a complex organism is an incredibly multi-dimensional, non–linear process, however advances in high–throughput sequencing allow high–resolution quantification of transcriptomes in a matter of hours. Manuscripts in this dissertation utilize such assays to probe the soybean transcriptome and gauge its defensive countermeasures.

Alongside exploring the transcriptomic landscape of infected soybean tissues, this dissertation also gauges the proximal binding site landscape without inoculation with a pest. Manuscripts comprising this dissertation collectively reveal the presence of a signature that captures binding site profiles of genes involved in stress response. In general, this signature traditionally encompasses TFBSs of TFs such as WRKY, AP2/EREBP, MYB, and bZIP. We extend this TFBS set by showing that TFBSs of TFs such as ABI3/VP1, GT–3b, TCP, and TBP, are over–represented in promoters of soybean genes involved in defense response.

In conclusion, this dissertation explores the systematic interplay between the soybean host and its two major pathogens: soybean rust and soybean cyst nematode. Utilizing high–throughput sequencing assays reveals high–coverage transcriptomic maps and analysis of promoter sequences reveals biologically–sound binding site profiles capturing defense response dynamics.

7.3 Future Work

This dissertation lays the foundations for further analyses involving soybean–pathogen interplay. The entirety of this dissertation was comprised of RNA–Seq studies investigating the soybean transcriptome and analysis of upstream promoter sequences. Indeed such analyses provide a glimpse into regulatory element dynamics, however transcription regulation is a process often spanning the reaches of promoter sequences.

A bulk of the organismal transcriptional landscape is mediated many kilobases from the target gene. Assays such as Hi–C, ChIP–Seq and resultant histone marks can be employed to measure such dynamics by quantifying *cis*– or *trans*– distal enhancer activity. Thus, superimposing ChIP–Seq assays with corresponding RNA–Seq runs provides a truly systematic snapshot.

Analysis of proximal regulatory elements is an excellent first–step to quantifying the organismal transcriptional landscape. Such findings could be translated towards building of synthetic promoter sequences by knocking–out or over–representing certain TFBSs. Systematically however, these proximal elements often collaborate with distal enhancers, silencers, and insulators to paint a global transcriptional landscape. Thus, introducing assays designed for quantification of distal elements will reveal novel insight into soybean regulatory element dynamics during pathogenesis.

Appendix A: Abbreviations

Abbreviations

4CL:	4–Coumarate–CoA ligase	**L–13S LOX:**	Linoleate 13S–LOX
AOC:	Allene oxidase cyclase	**LP:**	Laplace correction
A–8 LOX:	Arachidonate 8–LOX	**LI:**	Lift metric
ABA:	Abscisic acid	**LOX:**	Lipoxygenase
ChI:	Chalcone isomerase	**NPR1:**	Non-expressor of PR1
ChR:	Chalcone reductase	**PCS:**	Phytochelatin synthetase
CO:	Cosine metric	**PDI:**	Protein disulfide–isomerase
CF:	Confidence metric	**PHI:**	Phi-coefficient
DE:	Differentially expressed	**PPN:**	Plant–parasitic nematode
DEG:	Differentially expressed gene	**PR:**	Pathogen–related
ETH:	Ethylene	**PWM:**	Position Weight Matrix
GLY I:	Glyoxalase I	**RnDR:**	Ribonucleotide reductase
GO:	Gene Ontology	**SA:**	Salicylic acid
GST:	Glutathione S–transferase	**SCN:**	Soybean cyst nematode
IAA:	indole-3-acetic acid; IAA; auxin	**SOD:**	Super–oxide dismutase
IPF:	Iterative Proportional Fitting	**SR:**	Soybean rust
JA:	Jasmonic acid	**TF:**	Transcription Factor
JAC:	Jaccard index	**TFBS:**	TF binding site
K:	Cohen's Kappa	**TSS:**	Transcription Start Site

Bibliography

Bibliography

[1] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[2] R. Caspi, H. Foerster, C. A. Fulcher, P. Kaipa, M. Krummenacker, M. Latendresse, S. Paley, S. Y. Rhee, A. G. Shearer, C. Tissier, T. C. Walk, P. Zhang, and P. D. Karp, "The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases," *Nucleic Acids Research*, vol. 36, no. suppl 1, pp. D623–D631, 2008.

[3] D. L. Siehl, "The biosynthesis of tryptophan, tyrosine, and phenylalanine from chorismate," *Plant amino acids: Biochemistry and biotechnology*, pp. 171–204, 1999.

[4] M. C. Wildermuth, J. Dewdney, G. Wu, and F. M. Ausubel, "Isochorismate synthase is required to synthesize salicylic acid for plant defence," *Nature*, vol. 414, no. 6863, pp. 562–565, 2001.

[5] B. Bartel, "Auxin biosynthesis," *Annual Review of Plant Biology*, vol. 48, no. 1, pp. 51–66, 1997.

[6] B. Weisshaar and G. I. Jenkins, "Phenylpropanoid biosynthesis and its regulation," *Current Opinion in Plant Biology*, vol. 1, no. 3, pp. 251–257, 1998.

[7] T. Vogt, "Phenylpropanoid Biosynthesis," *Molecular Plant*, vol. 3, no. 1, pp. 2–20, 2010.

[8] R. A. Dixon and N. L. Paiva, "Stress–induced phenylpropanoid metabolism," *The Plant Cell*, vol. 7, no. 7, pp. 1085–1097, 1995.

[9] B. Winkel-Shirley, "Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology," *Plant Physiology*, vol. 126, no. 2, pp. 485–493, 2001.

[10] T. Vogt, P. Pollak, N. Tarlyn, and L. P. Taylor, "Pollination–or wound–induced kaempferol accumulation in petunia stigmas enhances seed production," *The Plant Cell Online*, vol. 6, no. 1, pp. 11–23, 1994.

[11] L. Pourcel, J. M. Routaboul, V. Cheynier, L. Lepiniec, and I. Debeaujon, "Flavonoid oxidation in plants: from biochemical properties to physiological functions," *Trends in Plant Science*, vol. 12, no. 1, pp. 29–36, 2007.

[12] Sweetlove, L. J. and Beard, K. F. M. and Nunes–Nesi, A. and Fernie, A. R. and Ratcliffe, R. G., "Not just a circle: flux modes in the plant TCA cycle," *Trends in Plant Science*, vol. 15, no. 8, pp. 462–470, 2010.

[13] R. A. Azevedo, P. Arruda, W. L. Turner, and P. J. Lea, "The biosynthesis and metabolism of the aspartate derived amino acids in higher plants," *Phytochemistry*, vol. 46, no. 3, pp. 395–419, 1997.

[14] S. Ravanel, B. Gakière, D. Job, and R. Douce, "The specific features of methionine biosynthesis and metabolism in plants," *Proceedings of the National Academy of Sciences*, vol. 95, no. 13, pp. 7805–7812, 1998.

[15] H. Hesse and R. Hoefgen, "Molecular aspects of methionine biosynthesis," *Trends in Plant Science*, vol. 8, no. 6, pp. 259–262, 2003.

[16] K. L. C. Wang, H. Li, and J. R. Ecker, "Ethylene biosynthesis and signaling networks," *The Plant Cell Online*, vol. 14, no. suppl 1, pp. S131–S151, 2002.

[17] H. Kende, "Ethylene Biosynthesis," *Annual Review of Plant Biology*, vol. 44, no. 1, pp. 283–307, 1993.

[18] H. Kende, "Enzymes of ethylene biosynthesis," *Plant Physiology*, vol. 91, no. 1, pp. 1–4, 1989.

[19] D. O. Adams and S. F. Yang, "Ethylene biosynthesis: identification of 1–aminocyclopropane–1–carboxylic acid as an intermediate in the conversion of methionine to ethylene," *Proceedings of the National Academy of Sciences*, vol. 76, no. 1, pp. 170–174, 1979.

[20] K. J. Bradford and S. F. Yang, "Xylem transport of 1–aminocyclopropane–1–carboxylic acid, an ethylene precursor, in waterlogged tomato plants," *Plant Physiology*, vol. 65, no. 2, pp. 322–326, 1980.

[21] J. Liu and X. J. Wang, "An integrative analysis of the effects of auxin on jasmonic acid biosynthesis in *Arabidopsis thaliana*," *Journal of Integrative Plant Biology*, vol. 48, no. 1, pp. 99–103, 2006.

[22] J. G. Turner, C. Ellis, and A. Devoto, "The jasmonate signal pathway," *The Plant Cell Online*, vol. 14, no. suppl 1, pp. S153–S164, 2002.

[23] A. Santner, L. I. A. Calderon-Villalobos, and M. Estelle, "Plant hormones are versatile chemical regulators of plant growth," *Nature Chemical Biology*, vol. 5, no. 5, pp. 301–307, 2009.

[24] R. A. Creelman and J. E. Mullet, "Jasmonic acid distribution and action in plants: regulation during development and response to biotic and abiotic stress," *Proceedings of the National Academy of Sciences*, vol. 92, no. 10, pp. 4114–4119, 1995.

[25] V. A. Halim, A. Vess, D. Scheel, and S. Rosahl, "The role of salicylic acid and jasmonic acid in pathogen defence," *Plant Biology*, vol. 8, no. 3, pp. 307–313, 2006.

[26] D. M. Riao-Pachn, S. Ruzicic, I. Dreyer, and B. Mueller-Roeber, "PlnTFDB: an integrative plant transcription factor database," *BMC Bioinformatics*, vol. 8, no. 42, 2007.

[27] K. Singh, R. C. Foley, and L. Oate-Sánchez, "Transcription factors in plant defense and stress responses.," *Current Opinion in Plant Biology*, vol. 5, pp. 430–436, October 2002.

[28] C. Dubos, R. Stracke, E. Grotewold, B. Weisshaar, C. Martin, and L. Lepiniec, "MYB transcription factors in *Arabidopsis*," *Trends in Plant Science*, vol. 15, pp. 573–581, October 2010.

[29] J. L. Carrasco, G. Ancillo, E. Mayda, and P. Vera, "A novel transcription factor involved in plant defense endowed with protein phosphatase activity," *The EMBO Journal*, vol. 22, pp. 3376–3384, May 2003.

[30] M. L. Bulyk, "Computational prediction of transcription–factor binding site locations," *Genome Biology*, vol. 5, December 2003.

[31] K. Cartharius, K. Frech, K. Grote, B. Klocke, M. Haltmeier, A. Klingenhoff, M. Frisch, M. Bayerlein, and T. Werner, "MatInspector and beyond: promoter analysis based on transcription factor binding sites," *Bioinformatics*, vol. 21, pp. 2933–2942, July 2005.

[32] M. Pertea, S. M. Mount, and S. L. Salzberg, "A computational survey of candidate exonic splicing enhancer motifs in the model plant *Arabidopsis thaliana*," *BMC Bioinformatics*, vol. 8, pp. 159+, May 2007.

[33] F. Fauteux and M. Stromvik, "Seed storage protein gene promoters contain conserved DNA motifs in *Brassicaceae*, *Fabaceae* and *Poaceae*," *BMC Plant Biology*, vol. 9, no. 1, pp. 126+, 2009.

[34] K. Vandepoele, T. Casneuf, and Y. van de Peer, "Identification of novel regulatory modules in dicotyledonous plants using expression data and comparative genomics," *Genome Biology*, vol. 7, November 2007.

[35] N. Leelavati and I. Ovcharenko, "Identifying regulatory elements in eukaryotic genomes," *Briefings in Functional Genomics and Proteomics*, vol. 8, pp. 215–230, June 2009.

[36] D. S. Chekmenev, C. Haid, and A. E. Kel, "P–Match: transcription factor binding site search by combining patterns and weight matrices," *Nucleic Acids Research*, vol. 33, pp. W432–W437, July 2005.

[37] G. G. Loots, I. Ovcharenko, L. Pachter, I. Dubchak, and E. M. Rubin, "rVista for comparative sequence-based discovery of functional transcription factor binding sites," *Genome Research*, vol. 12, pp. 832–839, May 2002.

[38] A. E. Kel, N. Voss, R. Jauregui, O. V. Kel-Margoulis, and E. Wingender, "Beyond microarrays: Finding key transcription factors controlling signal transduction pathways," *BMC Bioinformatics*, vol. 7, no. S2, 2006.

[39] L. Blow, S. Engelmann, M. Schindler, and R. Hehl, "AthaMap, integrating transcriptional and post-transcriptional data," *Nucleic Acids Research*, vol. 37, pp. D983–D986, October 2009.

[40] S. K. Palaniswamy, S. James, H. Sun, R. S. Lamb, R. V. Davuluri, and E. Grotewold, "AGRIS and AtRegNet: A platform to link *cis*–regulatory elements and transcription factors into regulatory networks," *Plant Physiology*, vol. 140, pp. 818–829, March 2006.

[41] S. Rombauts, K. Florquin, M. Lescot, K. Marchal, P. Rouz, and Y. van de Peer, "Computational approaches to identify promoters and cis-regulatory elements in plant genomes," *Plant Physiology*, vol. 132, pp. 1162–76, June 2003.

[42] E. Wingender, X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, H. Michael, R. Ohnhauser, M. Pruss, F. Schacherer, S. Thiele, and S. Urbach, "The TRANSFAC system on gene expression regulation," *Nucleic Acids Research*, vol. 29, no. 1, pp. 281–283, 2001.

[43] A. Sandelin, W. Alkema, P. Engstrom, W. W. Wasserman, and B. Lenhard, "JASPAR: an open-access database for eukaryotic transcription factor binding profiles," *Nucleic Acids Research*, vol. 32, pp. D91–D94, January 2007.

[44] R. R. Gabdoulline, D. Eckweiler, A. E. Kel, and P. Stegmaier, "3DTF: a web server for predicting transcription factor PWMs using 3D structure–based energy calculations," *Nucleic Acids Research*, vol. 40, no. Web–Server Issue, pp. 180–185, 2012.

[45] N. R. Horspool, "Practical fast searching in strings," *Software Practice and Experience*, vol. 10, no. 6, pp. 501–506, 1980.

[46] T. Ulmasov, G. Hagen, and T. J. Guilfoyle, "ARF1, a transcription factor that binds to auxin response elements," *Science*, vol. 276, pp. 1865–1868, June 1997.

[47] S. Ramakrishnan and A. Rakesh, "Mining sequential patterns: generalizations and performance improvements," in *Proceedings of the 5th International Conference on Extending Database Technology, EDBT*, pp. 3–17, 1996.

[48] P. N. Tan, V. Kumar, and J. Srivastava, "Selecting the right interestingness measure for association patterns," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pp. 32–41, 2002.

[49] L. Geng and H. J. Hamilton, "Interestingness measures for data mining: a survey," *ACM Computing Surveys*, vol. 38, September 2006.

[50] M. Steinbach, P. N. Tan, H. Xiong, and V. Kumar, "Objective measures for association pattern analysis," *Contemporary Mathematics*, no. 443, pp. 205–226, 2007.

[51] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in *ACM SIGMOD – International Conference on Management of Data*, May 1993.

[52] A. Merceron and K. Yacef, "Interestingness measures for association rules in educational data," in *Proceedings of Educational Data Mining 2008: 1st International Conference on Educational Data Mining*, pp. 57–66, 2008.

[53] P. Jaccard, "Ètude comparative de la distribution florale dans une portion des Alpes et des Jura," *Bulletin de la Société Vaudoise des Sciences Naturelles*, vol. 37, pp. 547–579, 1901.

[54] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, pp. 37–46, April 1960.

[55] I. J. Good, *The estimation of probabilities: an essay on modern Bayesian methods*, vol. 30 of *Research Monograph*. M.I.T. Press, 1965.

[56] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, "Dynamic itemset counting and implication rules for market basket data," in *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data*, pp. 255–264, May 1997.

[57] H. Cramér, *Mathematical methods of statistics*. Princeton mathematical series, Princeton University Press, 1946.

[58] W. E. Deming and F. F. Stephan, "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known," *Annals of Mathematical Statistics*, vol. 11, no. 4, pp. 427–444, 1940.

[59] F. Mosteller, "Association and estimation in contingency tables," *Journal of the American Statistical Association*, vol. 63, pp. 1–28, March 1968.

[60] P. N. Tan, V. Kumar, and J. Srivastava, "Selecting the right objective measure for association analysis," *Information Systems*, vol. 29, pp. 293–313, 2004.

[61] A. Tremblay, P. Hosseini, N. W. Alkharouf, S. Li, and B. F. Matthews, "Gene expression in leaves of susceptible *Glycine max* during infection with *Phakopsora pachyrhizi* using next generation sequencing," *Sequencing*, p. 14, 2011.

[62] J. D. Peterson, L. A. Umayam, T. M. Dickinson, E. K. Hickey, and O. White, "The comprehensive microbial resource," *Nucleic Acids Research*, vol. 29, pp. 123–125, January 2001.

[63] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows–Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[64] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA–Seq," *Nature Methods*, vol. 5, pp. 621–628, July 2008.

[65] P. J. Rushton, I. E. Somssich, P. Ringler, and Q. J. Shen, "WRKY transcription factors," *Trends in Plant Science*, vol. 15, pp. 247–258, March 2010.

[66] T. Eulgem, "Dissecting the WRKY web of plant defense regulators," *PLoS Pathogens*, vol. 2, November 2006.

[67] T. Eulgem, P. J. Rushton, S. Robatzek, and I. E. Somssich, "The WRKY superfamily of plant transcription factors," *Trends in Plant Science*, vol. 5, pp. 199–206, May 2000.

[68] T. Eulgem and I. E. Somssich, "Networks of WRKY transcription factors in defense signaling," *Current Opinion in Plant Biology*, vol. 10, pp. 366–371, August 2007.

[69] S. P. Pandey and I. E. Somssich, "The role of WRKY transcription factors in plant immunity," *Plant Physiology*, vol. 150, no. 4, pp. 1648–1655, 2009.

[70] D. Yu, C. Chen, and Z. Chen, "Evidence for an important role of WRKY DNA binding proteins in the regulation of NPR1 gene expression," *Plant Cell*, vol. 13, pp. 1527–1540, July 2001.

[71] X. Dong, "NPR1, all things considered," *Current Opinions in Plant Biology*, vol. 7, pp. 547–552, October 2004.

[72] R. Sibout, P. Sukumar, C. Hettiarachchi, M. Holm, G. K. Muday, and C. S. Hardtke, "Opposite root growth phenotypes of hy5 versus hy5 hyh mutants correlate with increased constitutive auxin signaling," *PLoS Genetics*, vol. 2, November 2004.

[73] C. P. Cluis, C. F. Mouchel, and C. S. Hardtke, "The *Arabidopsis* transcription factor HY5 integrates light and hormone signaling pathways," *Plant Journal*, vol. 38, pp. 332–347, April 2004.

[74] B. R. V. P. Prasad, S. V. Kumar, A. Nandi, and S. Chattopadhyay, "Functional interconnections of HY1 with MYC2 and HY5 in *Arabidopsis* seedling development," *BMC Plant Biology*, vol. 12, March 2012.

[75] M. Boter, O. Ruíz-Rivero, A. Abdeen, and S. Prat, "Conserved MYC transcription factors play a key role in jasmonate signaling both in tomato and *Arabidopsis*," *Genes & Development*, vol. 18, pp. 1577–1591, July 2004.

[76] H. Abe, T. Urao, T. Itom, M. Seki, K. Shinozaki, and K. Yamaguchi-Shinozaki, "*Arabidopsis* AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling," *Plant Cell*, vol. 15, pp. 63–78, January 2003.

[77] R. N. Kaplan-Levy, P. B. Brewer, T. Quon, and D. R. Smyth, "The trihelix family of transcription factors – light, stress and development," *Trends in Plant Science*, vol. 17, pp. 163–171, March 2012.

[78] Z. M. Xie, H. F. Zou, G. Lei, W. Wei, Q. Y. Zhou, C. F. Niu, Y. Liao, A. G. Tian, B. Ma, W. K. Zhang, J. S. Zhang, and S. Y. Chen, "Soybean trihelix transcription factors GmGT-2A and GmGT-2B improve plant tolerance to abiotic stresses in transgenic *Arabidopsis*," *PLoS ONE*, vol. 4, p. e6898, September 2009.

[79] G. Pison, A. Struyf, and P. J. Rousseeuw, "Displaying a clustering with CLUSPLOT," *Computational Statistics and Data Analysis*, vol. 30, pp. 381–392, June 1999.

[80] J. Keilwagen, J. Grau, I. A. Paponov, S. Posch, M. Strickert, and I. Grosse, "*De–novo* discovery of differentially abundant transcription factor binding sites including their positional preference," *PLoS Computational Biology*, vol. 7, pp. e1001070+, February 2011.

[81] Y. Y. Yamamoto and J. Obokata, "ppdb: a plant promoter database," *Nucleic Acids Research*, vol. 36, pp. D977 – D981, January 2008.

[82] W. C. Chang, T. Y. Lee, H. D. Huang, H. Y. Huang, and R. L. Pan, "PlantPAN: Plant promoter analysis navigator for identifying combinatorial *cis*–regulatory elements with distance constraint in plant gene groups," *BMC Genomics*, vol. 9, pp. 561+, November 2008.

[83] J. A. Wrather, T. R. Anderson, D. M. Arsyad, Y. Tan, L. D. Ploper, A. Porta-Puglia, R. H. H., and J. T. Yorinori, "Soybean disease loss estimates for the top ten soybean-producing counries in 1998," *Canadian Journal of Plant Pathology*, vol. 23, no. 2, pp. 115–121, 2001.

[84] P. D. Matsye, G. W. Lawrence, R. Youssef, K. H. Kim, K. S. Lawrence, B. F. Matthews, and V. P. Klink, "The expression of a naturally occurring, truncated allele of an α–SNAP gene suppresses plant parasitic nematode infection," *Plant Molecular Biology*, vol. 80, pp. 131–155, September 2012.

[85] B. Y. Endo, "Penetration and development of *Heterodera glycines* in soybean roots and related anatomical changes," *Phytopathology*, vol. 54, pp. 79–88, 1964.

[86] V. P. Klink, P. Hosseini, M. H. MacDonald, N. W. Alkharouf, and B. F. Matthews, "Population–specific gene expression in the plant pathogenic nematode *Heterodera glycines* exists prior to infection and during the onset of a resistant or susceptible reaction in the roots of the *Glycine max* genotype Peking," *BMC Genomics*, vol. 10, March 2009.

[87] X. Li, X. Wang, S. Zhang, D. Liu, Y. Duan, and W. Dong, "Comparative profiling of the transcriptional response to soybean cyst nematode infection of soybean roots by deep sequencing," *Chinese Science Bulletin*, vol. 56, pp. 1904–1911, June 2011.

[88] X. Li, X. Wang, S. Zhang, D. Liu, Y. Duan, and W. Dong, "Identification of Soybean MicroRNAs Involved in Soybean Cyst Nematode Infection by Deep Sequencing," *PLoS ONE*, vol. 7, p. e39650, June 2012.

[89] N. Hamamouch, C. Li, T. Hewezi, T. J. Baum, M. G. Mitchum, R. S. Hussey, L. O. Vodkin, and E. L. Davis, "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*, vol. 63, pp. 3683–3695, June 2012.

[90] S. K. Guttikonda, N. C. Trupti, J. Bisht, H. Chen, Y. C. An, D. Pandey, S. Xu, and O. Yu, "Whole genome co–expression analysis of soybean cytochrome P450 genes identifies nodulation-specific P450 monooxygenases," *BMC Plant Biology*, vol. 10, no. 243, 2010.

[91] D. M. Goodstein, S. Shu, R. Howson, R. Neupane, R. D. Hayes, J. Fazo, T. Mitros, W. Dirks, U. Hellsten, N. Putnam, and D. S. Rokhsar, "Phytozome: a comparative platform for green plant genomics," *Nucleic Acids Research*, pp. 1178–1186, 2012.

[92] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biology*, vol. 11, no. 10, p. R106, 2010.

[93] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene Ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[94] A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, and E. L. L. Sonnhammer, "The Pfam Protein Families Database," *Nucleic Acids Research*, vol. 30, no. 1, pp. 276–280, 2002.

[95] H. M. Ibrahim, P. Hosseini, N. W. Alkharouf, E. H. Hussein, A. E. K. E. Gamal E. D., M. A. Aly, and B. F. Matthews, "Analysis of gene expression in soybean (*Glycine max*) roots in response to the root–knot nematode *Meloidogyne incognita* using microarrays and KEGG pathways," *BMC Genomics*, vol. 12, May 2011.

[96] M. Goellner, X. Wang, and E. L. Davis, "Endo-β-1,4-glucanase expression in compatible plant–nematode interactions," *Plant Cell*, vol. 13, no. 10, pp. 2241–2255, 2001.

[97] M. L. Tucker, A. Burke, C. A. Murphy, V. K. Thai, and M. L. Ehrenfried, "Gene expression profiles for cell wall–modifying proteins associated with soybean cyst nematode infection, petiole abscission, root tips, flowers, apical buds, and leaves," *Journal of Experimental Botany*, vol. 58, no. 12, pp. 3395–3406, 2007.

[98] D. A. Dalton, C. Boniface, Z. Turner, A. Lindahl, H. J. Kim, L. Jelinek, M. Govindarajulu, R. E. Finger, and C. G. Taylor, "Physiological roles of glutathione s–transferases in soybean root nodules," *Plant Physiology*, vol. 150, pp. 521–530, May 2009.

[99] M. Mazarei, W. Liu, H. Al-Ahmad, P. R. Arelli, V. R. Pantalone, and C. N. Stewart, "Gene expression profiling of resistant and susceptible soybean lines infected with soybean cyst nematode," *Theoretical Applied Genetics*, vol. 123, pp. 1193–1206, November 2011.

[100] N. Alkharouf, R. Khan, and B. F. Matthews, "Analysis of expressed sequence tags from roots of resistant soybean infected by the soybean cyst nematode," *Genome*, vol. 47, pp. 380–388, April 2004.

[101] P. K. Kandoth, N. Ithal, J. Recknor, T. Maier, D. Nettleton, T. J. Baum, and M. G. Mitchum, "The Soybean Rhg1 locus for resistance to the soybean cyst nematode *Heterodera glycines* regulates the expression of a large number of stress– and defense–related genes in degenerating feeding cells," *Plant Physiology*, vol. 155, pp. 1960–1975, April 2011.

[102] V. P. Klink and B. F. Matthews, "Emerging approaches to broaden resistance of soybean to soybean cyst nematode as supported by gene expression studies," *Plant Physiology*, vol. 151, pp. 1017–1022, November 2009.

[103] V. P. Klink, P. D. Matsye, K. S. Lawrence, and G. W. Lawrence, *Soybean – Pest Resistance*, ch. Engineered Soybean Cyst Nematode Resistance, pp. 147–180. InTech, February 2013.

[104] P. Veronico, D. Giannino, M. T. Melillo, A. Leone, A. Reyes, M. W. Kennedy, and T. Bleve-Zacheo, "A novel lipoxygenase in pea roots: its function in wounding and biotic stress," *Plant Physiology*, vol. 141, pp. 1045–1055, July 2006.

[105] N. Ithal, J. Recknor, D. Nettleton, T. Maier, T. J. Baum, and M. G. Mitchum, "Developmental transcript profiling of cyst nematode feeding cells in soybean roots," *Molecular Plant–Microbe Interactions*, vol. 20, no. 5, pp. 510–525, 2007.

[106] C. Vieira Dos Santos and P. Rey, "Plant thioredoxins are key actors in the oxidative stress response," *Trends in Plant Science*, vol. 11, no. 7, pp. 329–334, 2006.

[107] C. Laloi, D. Mestres-Ortega, Y. Marco, Y. Meyer, and J. Reichheld, "The *Arabidopsis* cytosolic thioredoxin h5 gene induction by oxidative stress and its W-box-mediated response to pathogen elicitor," *Plant Physiology*, vol. 134, no. 3, pp. 1006–16, 2004.

[108] D. Wang, N. D. Weaver, M. Kesarwani, and X. Dong, "Induction of protein secretory pathway is required for systemic acquired resistance," *Science*, vol. 308, pp. 1036–1040, May 2005.

[109] F. Thibaud-Nissen, R. T. Shealy, A. Khanna, and L. O. Vodkin, "Clustering of microarray data reveals transcript patterns associated with somatic embryogenesis in soybean," *Plant Physiology*, vol. 132, no. 1, pp. 118–136, 2003.

[110] S. Ray, J. M. Anderson, F. I. Urmeev, and S. B. Goodwin, "Rapid induction of a protein disulfide isomerase and defense-related genes in wheat in response to the hemibiotrophic fungal pathogen *Mycosphaerella graminicola*," *Plant Molecular Biology*, vol. 53, pp. 701–714, November 2003.

[111] C. W. Gruber, M. Cemazar, R. J. Clark, T. Horibe, R. F. Renda, M. A. Anderson, and D. J. Craik, "A novel plant protein–disulfide isomerase involved in the oxidative folding of cystine knot defense proteins," *Journal of Biological Chemistry*, vol. 282, pp. 20435–20446, July 2007.

[112] V. P. Klink, P. Hosseini, P. D. Matsye, N. W. Alkharouf, and B. F. Matthews, "Syncytium gene expression in *Glycine max* (pi 88788) roots undergoing a resistant reaction to the parasitic nematode *Heterodera glycines*," *Plant Physiology and Biochemistry*, vol. 48, no. 2–3, pp. 176–93, 2010.

[113] A. J. Afzal, A. Natarajan, N. Saini, M. J. Iqbal, M. Geisler, H. A. El Shemy, R. Mungur, L. Willmitzer, and D. A. Lightfoot, "The Nematode Resistance Allele at the rhg1 Locus Alters the Proteome and Primary Metabolism of Soybean Roots," *Plant Physiology*, vol. 151, pp. 1264–80, November 2009.

[114] *Soybean – Genetics and Novel Techniques for Yield Enhancement*, ch. Changes in the Expression of Genes in Soybean Roots Infected by Nematodes, pp. 87–106. InTech, November 2011.

[115] N. W. Alkharouf, V. P. Klink, I. B. Chouikha, H. Beard, M. H. MacDonald, S. Meyer, H. T. Knap, R. Khan, and B. Matthews, "Timecourse microarray analyses reveal global changes in gene expression of susceptible *Glycine max* (soybean) roots during

infection by *Heterodera glycines* (soybean cyst nematode)," *Planta*, vol. 224, pp. 838–852, September 2006.

[116] N. Ithal, J. Recknor, D. Nettleton, L. Hearne, T. Maier, T. J. Baum, and M. G. Mitchum, "Parallel genome–wide expression profiling of host and pathogen during soybean cyst nematode infection of soybean," *Molecular Plant–Microbe Interactions*, vol. 20, pp. 293–305, March 2007.

[117] M. G. Mitchum and T. J. Baum, "Genomics of the Soybean Cyst Nematode–Soybean Interaction," in *Genetics and Genomics of Soybean* (G. Stacey, ed.), vol. 2 of *Plant Genetics and Genomics: Crops and Models*, pp. 321–341, Springer New York, 2008.

[118] B. F. Matthews, H. Beard, M. H. MacDonald, S. Kabir, R. Youssef, P. Hosseini, and E. Brewer, "Engineered resistance and hypersusceptibility through functional metabolic studies of 100 genes in soybean to its major pathogen, the soybean cyst nematode," *Planta*, vol. 237, pp. 1337–1357, February 2013.

[119] M. Hashimoto, L. Kisseleva, S. Sawa, T. Furukawa, S. Komatsu, and T. Koshiba, "A novel rice pr10 protein, rsospr10, specifically induced in roots by biotic and abiotic stresses, possibly via the jasmonic acid signaling pathway," *Plant Cell Physiology*, vol. 45, no. 5, pp. 550–559, 2004.

[120] S. Kitajima and F. Sato, "Plant Pathogenesis–Related Proteins: Molecular Mechanisms of Gene Expression and Protein Function," *Journal of Biochemistry*, vol. 125, pp. 1–8, January 1999.

[121] C. Dubos and C. Plomion, "Drought differentially affects expression of a PR–10 protein, in needles of maritime pine (*Pinus pinaster* Ait.) seedlings," *Journal of Experimental Botany*, vol. 52, pp. 1143–1144, May 2001.

[122] M. A. Hossain, M. Z. Hossain, and M. Fujita, "Stress–induced changes of methylglyoxal level and glyoxalase I activity in pumpkin seedlings and cDNA cloning of glyoxalase I gene," *Australian Journal of Crop Science*, vol. 3, no. 2, pp. 53–64, 2009.

[123] C. Michael Smith and E. V. Boyko, "The molecular bases of plant resistance and defense responses to aphid feeding: current status," *Entomologia Experimentalis et Applicata*, vol. 122, pp. 1–16, January 2007.

[124] Z. Du, X. Zhou, Y. Ling, Z. Zhang, and Z. Su, "agriGO: a GO analysis toolkit for the agricultural community," *Nucleic Acids Research*, vol. 38, no. Web-Server Issue, pp. 64–70, 2010.

[125] C. H. Brueske and G. B. Bergeson, "Investigation of growth hormones in xylem exudate and root tissue of tomato infected with root–knot nematode," *Journal of Experimental Botany*, vol. 23, pp. 14–22, February 1972.

[126] B. R. Loveys and A. F. Bird, "The influence of nematodes on photosynthesis in tomato plants," *Physiological Plant Pathology*, vol. 3, pp. 525–529, July 1973.

[127] P. Hosseini, I. Ovcharenko, and B. F. Matthews, "Using an ensemble of statistical metrics to quantify large sets of plant transcription factor binding sites," *Plant Methods*, vol. 9, April 2013.

[128] P. A. Manavella, C. A. Dezar, G. Bonaventure, I. T. Baldwin, and R. L. Chan, "HAHB4, a sunflower HD–Zip protein, integrates signals from the jasmonic acid and ethylene pathways during wounding and biotic stress responses," *The Plant Journal*, vol. 56, pp. 376–388, November 2008.

[129] S. Yanagisawa, "Dof domain proteins: plant-specific transcription factors associated with diverse phenomena unique to plants," *Plant Cell Physiology*, vol. 45, no. 4, pp. 386–391, 2004.

[130] T. Hibi, S. Kosugi, T. Iwai, M. Kawata, S. Seo, I. Mitsuhara, and Y. Ohashi, "Involvement of ein3 homologues in basic pr gene expression and flower development in tobacco plants.," *Journal of Experimental Botany*, vol. 58, pp. 3671–3678, October 2007.

[131] S. C. Koo, M. S. Choi, H. J. Chun, D. B. Shin, B. S. Park, Y. H. Kim, H. Park, H. S. Seo, J. T. Song, K. Y. Kang, D. Yun, W. S. Chung, M. J. Cho, and M. C. Kim, "The calmodulin–binding transcription factor OsCBT suppresses defense responses to pathogens in rice," *Molecules and Cells*, vol. 27, no. 5, pp. 563–570, 2009.

[132] J. Kim, H. Yi, G. Choi, B. Shin, P. Song, and G. Choi, "Functional characterization of phytochrome interacting factor 3 in phytochrome-mediated light signal transduction," *Plant Cell*, vol. 15, no. 10, pp. 2399–2407, 2003.

[133] S. Sardanelli and W. J. Kenworthy, "Soil moisture control and direct seeding for bioassay of *Heterodera glycines* on soybean," *The Journal of Nematology*, vol. 29, no. 4S, pp. 625–634, 1997.

[134] D. W. Byrd Jr., T. Kirkpatrick, and K. R. Barker, "An Improved Technique for Clearing and Staining Plant Tissues for Detection of Nematodes," *The Journal of Nematology*, vol. 15, no. 1, pp. 142–143, 1983.

[135] C. V. Mujer, D. L. Andrews, J. R. Manhart, S. K. Pierce, and M. E. Rumpho, "Chloroplast genes are expressed during intracellular symbiotic association of *Vaucheria litorea* plastids with the sea slug *Elysia chlorotica*," *Proceedings of the National Academy of Sciences*, vol. 93, no. 22, pp. 12333–12338, 1996.

[136] S. F. Altschul, T. L. Madden, A. A. Schffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped Blast and PsiBlast: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[137] The UniProt Consortium, "Reorganizing the protein space at the Universal Protein Resource (UniProt)," *Nucleic Acids Research*, vol. 40, no. D1, pp. D71–D75, 2012.

[138] P. M. Houterman, B. J. C. Cornelissen, and M. Rep, "Suppression of plant resistance gene-based immunity by a fungal effector," *PLoS Pathogens*, vol. 4, no. 5, p. e1000061, 2008.

[139] I. Stergiopoulos and P. J. G. M. de Wit, "Fungal effector proteins," *Annual Review of Phytopathology*, vol. 47, pp. 233–263, 2009.

[140] R. E. Soria-Guerra, S. Rosales-Mendoza, S. Chang, J. S. Haudenshield, A. Padmanaban, S. Rodriguez-Zas, G. L. Hartman, S. A. Ghabrial, and S. S. Korban, "Transcriptome analysis of resistant and susceptible genotypes of *Glycine tomentella* during *Phakopsora pachyrhizi* infection reveals novel rust resistance genes," *Theoretical and Applied Genetics*, vol. 120, no. 7, pp. 1315–1333, 2010.

[141] X. B. Yang, "Assessment and management of the risk of soybean rust," in *Proceedings of the Soybean Rust Workshop*, pp. 9–11, National Soybean Research Laboratory, Urbana, Illinois, USA, 1995.

[142] M. L. Posada-Buitrago and R. D. Frederick, "Expressed sequence tag analysis of the soybean rust pathogen *Phakopsora pachyrhizi*," *Fungal Genetics and Biology*, vol. 42, no. 12, pp. 949–962, 2005.

[143] J. T. Yorinori, W. M. Paiva, R. D. Frederick, L. M. Costamilan, P. F. Bertagnolli, G. E. Hartman, C. V. Godoy, and J. Nunes Jr, "Epidemics of soybean rust *(Phakopsora pachyrhizi)* in Brazil and Paraguay from 2001 to 2003," *Plant Disease*, vol. 89, no. 6, pp. 675–677, 2005.

[144] A. C. da Silva, P. E. de Souza, J. E. B. P. Pinto, B. M. da Silva, D. C. Amaral, and E. de Arruda Carvalho, "Essential oils for preventative treatment and control of Asian soybean rust," *European Journal of Plant Pathology*, vol. 134, no. 4, pp. 865–871, 2012.

[145] G. L. Hartman, M. R. Miles, and R. D. Frederick, "Breeding for resistance to soybean rust," *Plant Disease*, vol. 89, no. 6, pp. 664–666, 2005.

[146] M. R. Miles, C. Levy, W. Morel, T. Mueller, T. Steinlage, N. Van Rij, R. D. Frederick, and G. L. Hartman, "International fungicide efficacy trials for the management of soybean rust," *Plant Disease*, vol. 91, no. 11, pp. 1450–1458, 2007.

[147] H. Scherm, R. S. C. Christiano, P. D. Esker, E. M. Del Ponte, and C. V. Godoy, "Quantitative review of fungicide efficacy trials for managing soybean rust in Brazil," *Crop Protection*, vol. 28, no. 9, pp. 774–782, 2009.

[148] A. Tremblay, *Soybean Rust: Five Years of Research*, ch. 14. InTech, 2011.

[149] M. A. Marchetti, F. A. Uecker, and K. R. Bromfield, "Uredial development of *Phakopsora pachyrhizi* in soybeans," *Phytopathology*, vol. 65, no. 7, pp. 822–823, 1975.

[150] A. M. A. P. Morales, A. Borém, M. A. Graham, and R. V. Abdelnoor, "Advances on molecular studies of the interaction soybean–Asian rust," *Crop Breeding and Applied Biotechnology*, vol. 12, no. 1, pp. 1–7, 2012.

[151] A. Tremblay, P. Hosseini, S. Li, N. W. Alkharouf, and B. F. Matthews, "Analysis of *Phakopsora pachyrhizi* gene expression in critical pathways at four time–points during infection of a susceptible soybean cultivar using deep sequencing," *BMC Genomics*, vol. 14, no. 614, 2013.

[152] C. Trapnell, L. Pachter, and S. L. Salzberg, "TopHat: discovering splice junctions with RNA–Seq," *Bioinformatics*, vol. 25, no. 9, pp. 1105–1111, 2009.

[153] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. Van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nature Biotechnology*, vol. 28, no. 5, pp. 511–515, 2010.

[154] G. R. Warnes, B. Bolker, and T. Lumley, "gplots: Various R programming tools for plotting data," *R package version*, vol. 2, no. 4, 2009.

[155] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B*, pp. 267–288, 1996.

[156] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–33, 2010.

[157] M. van de Mortel, J. C. Recknor, M. A. Graham, D. Nettleton, J. D. Dittman, R. T. Nelson, C. V. Godoy, R. V. Abdelnoor, A. M. Almeida, T. J. Baum, and S. A. Whitham, "Distinct biphasic mRNA changes in response to Asian soybean rust infection," *Molecular Plant–Microbe Interactions*, vol. 20, no. 8, pp. 887–899, 2007.

[158] E. Koch, F. Ebrahim-Nesbat, and H. H. Hoppe, "Light and electron microscopic studies on the development of soybean rust (*Phakopsora pachyrhizi* Syd.) in susceptible soybean leaves," *Phytopathol Z.*, vol. 106, no. 4, pp. 302–320, 1983.

[159] H. Johannesson, Y. Wang, and P. Engström, "DNA–binding and dimerization preferences of *Arabidopsis* homeodomain-leucine zipper transcription factors in vitro," *Plant Molecular Biology*, vol. 45, no. 1, pp. 63–73, 2001.

[160] A. Himmelbach, T. Hoffmann, M. Leube, B. Höhener, and E. Grill, "Homeodomain protein ATHB6 is a target of the protein phosphatase ABI1 and regulates hormone responses in *Arabidopsis*," *The EMBO Journal*, vol. 21, no. 12, pp. 3029–3038, 2002.

[161] H. A. Phan, S. F. Li, and R. W. Parish, "MYB80, a regulator of tapetal and pollen development, is functionally conserved in crops," *Plant Molecular Biology*, vol. 78, no. 1–2, pp. 171–183, 2012.

[162] P. J. Rushton and I. E. Somssich, "Transcriptional control of plant genes responsive to pathogens," *Current Opinion in Plant Biology*, vol. 1, no. 4, pp. 311–315, 1998.

[163] A. Mahjoub, M. Hernould, J. Joubès, A. Decendit, M. Mars, F. Barrieu, S. Hamdi, and S. Delrot, "Overexpression of a grapevine R2R3–MYB factor in tomato affects vegetative development, flower morphology and flavonoid and terpenoid metabolism," *Plant Physiology and Biochemistry*, vol. 47, no. 7, pp. 551–561, 2009.

[164] A. Gonzalez, M. Zhao, J. M. Leavitt, and A. M. Lloyd, "Regulation of the anthocyanin biosynthetic pathway by the TTG1/bHLH/Myb transcriptional complex in *Arabidopsis* seedlings," *The Plant Journal*, vol. 53, no. 5, pp. 814–827, 2008.

[165] J. Zhou, C. Lee, R. Zhong, and Z. H. Ye, "MYB58 and MYB63 are transcriptional activators of the lignin biosynthetic pathway during secondary cell wall formation in *Arabidopsis*," *The Plant Cell*, vol. 21, no. 1, pp. 248–266, 2009.

[166] M. A. Hossain, Y. Lee, J. I. Cho, C. H. Ahn, S. K. Lee, J. S. Jeon, H. Kang, C. H. Lee, G. An, and P. B. Park, "The bZIP transcription factor OsABF1 is an ABA responsive element binding factor that enhances abiotic stress signaling in rice," *Plant Molecular Biology*, vol. 72, no. 4-5, pp. 557–566, 2010.

[167] Y. Fujita, M. Fujita, R. Satoh, K. Maruyama, M. M. Parvez, M. Seki, K. Hiratsu, M. Ohme-Takagi, K. Shinozaki, and K. Yamaguchi-Shinozaki, "AREB1 is a transcription activator of novel ABRE–dependent ABA signaling that enhances drought stress tolerance in *Arabidopsis*," *The Plant Cell*, vol. 17, no. 12, pp. 3470–3488, 2005.

[168] B. A. T. Adie, J. Pérez-Pérez, M. M. Pérez-Pérez, M. Godoy, J. J. Sánchez-Serrano, E. A. Schmelz, and R. Solano, "Aba is an essential signal for plant resistance to pathogens affecting ja biosynthesis and the activation of defenses in *Arabidopsis*," *The Plant Cell*, vol. 19, no. 5, pp. 1665–1681, 2007.

[169] J. Dong, C. Chen, and Z. Chen, "Expression profiles of the *Arabidopsis* WRKY gene superfamily during plant defense response," *Plant Molecular Biology*, vol. 51, no. 1, pp. 21–37, 2003.

[170] C. Johnson, E. Boden, M. Desai, P. Pascuzzi, and J. Arias, "In vivo target promoter-binding activities of a xenobiotic stress–activated TGA factor," *The Plant Journal*, vol. 28, no. 2, pp. 237–243, 2001.

[171] C. Schommer, J. F. Palatnik, P. Aggarwal, A. Chételat, P. Cubas, E. E. Farmer, U. Nath, and D. Weigel, "Control of jasmonate biosynthesis and senescence by miR319 targets," *PLoS Biology*, vol. 6, no. 9, p. e230, 2008.

[172] Y. Song, H. Li, Q. L. Shi, H. Jiang, H. Chen, X. L. Zhong, J. F. Gao, Y. L. Cui, and Z. N. Yang, "The Arabidopsis bHLH transcription factor DYT1 is essential for anther development by regulating callose dissolution," *Journal of Shanghai Normal University*, vol. 38, pp. 174–182, 2009.

[173] X. Liu, X. Bai, X. Wang, and C. Chu, "OsWRKY71, a rice transcription factor, is involved in rice defense response," *Journal of Plant Physiology*, vol. 164, no. 8, pp. 969–979, 2007.

[174] R. Bari and J. D. G. Jones, "Role of plant hormones in plant defence responses," *Plant Molecular Biology*, vol. 69, no. 4, pp. 473–488, 2009.

[175] P. I. Kerchev, T. K. Pellny, P. D. Vivancos, G. Kiddle, P. Hedden, S. Driscoll, H. Vanacker, P. Verrier, R. D. Hancock, and C. H. Foyer, "The transcription factor ABI4 is required for the ascorbic acid–dependent regulation of growth and regulation of jasmonate–dependent defense signaling pathways in *Arabidopsis*," *The Plant Cell Online*, vol. 23, no. 9, pp. 3319–3334, 2011.

[176] A. Santner and M. Estelle, "Recent advances and emerging trends in plant hormone signalling," *Nature*, vol. 459, no. 7250, pp. 1071–1078, 2009.

[177] V. N. Matiru and F. D. Dakora, "The rhizosphere signal molecule lumichrome alters seedling development in both legumes and cereals," *New Phytologist*, vol. 166, no. 2, pp. 439–444, 2005.

[178] G. Loake and M. Grant, "Salicylic acid in plant defence–the players and protagonists," *Current Opinion in Plant Biology*, vol. 10, no. 5, pp. 466–472, 2007.

[179] H. Gundlach, M. J. Muller, T. M. Kutchan, and M. H. Zenk, "Jasmonic acid is a signal transducer in elicitor–induced plant cell cultures," *Proceedings of the National Academy of Sciences*, vol. 89, no. 6, pp. 2389–2393, 1992.

[180] O. Lorenzo, R. Piqueras, J. J. Sánchez-Serrano, and R. Solano, "ETHYLENE RESPONSE FACTOR1 integrates signals from ethylene and jasmonate pathways in plant defense," *The Plant Cell Online*, vol. 15, no. 1, pp. 165–178, 2003.

[181] C. M. J. Pieterse, A. Leon-Reyes, S. Van der Ent, and S. C. M. Van Wees, "Networking by small–molecule hormones in plant immunity," *Nature Chemical Biology*, vol. 5, no. 5, pp. 308–316, 2009.

[182] J. Sun, V. Cardoza, D. M. Mitchell, L. Bright, G. Oldroyd, and J. M. Harris, "Crosstalk between jasmonic acid, ethylene and Nod factor signaling allows integration of diverse inputs for regulation of nodulation," *The Plant Journal*, vol. 46, no. 6, pp. 961–970, 2006.

[183] L. C. van Loon, B. P. J. Geraats, and H. J. M. Linthorst, "Ethylene as a modulator of disease resistance in plants," *Trends in Plant Science*, vol. 11, no. 4, pp. 184–191, 2006.

[184] Thomma, B. P. H. J. and Penninckx, I. A. M. A. and Cammue, B. and Broekaert, W. F., "The complexity of disease signaling in *Arabidopsis*," *Current Opinion in Immunology*, vol. 13, no. 1, pp. 63–68, 2001.

[185] N. A. Eckardt, "Oxylipin signaling in plant stress responses," *The Plant Cell Online*, vol. 20, no. 3, pp. 495–497, 2008.

[186] P. E. Staswick and I. Tiryaki, "The oxylipin signal jasmonic acid is activated by an enzyme that conjugates it to isoleucine in *Arabidopsis*," *The Plant Cell Online*, vol. 16, no. 8, pp. 2117–2127, 2004.

[187] G. A. Howe and A. L. Schilmiller, "Oxylipin metabolism in response to stress," *Current Opinion in Plant Biology*, vol. 5, no. 3, pp. 230–236, 2002.

[188] E. Bartholomeus Kuettner, R. Hilgenfeld, and M. S. Weiss, "Purification, characterization, and crystallization of alliinase from garlic," *Archives of Biochemistry and Biophysics*, vol. 402, no. 2, pp. 192–200, 2002.

[189] M. Jia, H. Wu, K. L. Clay, R. Jung, B. A. Larkins, and B. C. Gibbon, "Identification and characterization of lysine-rich proteins and starch biosynthesis genes in the opaque2 mutant by transcriptional and proteomic analysis," *BMC Plant Biology*, vol. 13, no. 1, p. 60, 2013.

[190] C. Y. Yang, F. C. Hsu, J. P. Li, N. N. Wang, and M. C. Shih, "The AP2/ERF transcription factor AtERF73/HRE1 modulates ethylene responses during hypoxia in *Arabidopsis*," *Plant Physiology*, vol. 156, no. 1, pp. 202–212, 2011.

[191] M. Pré, M. Atallah, A. Champion, M. De Vos, C. M. J. Pieterse, and J. Memelink, "The AP2/ERF domain transcription factor ORA59 integrates jasmonic acid and ethylene signals in plant defense," *Plant Physiology*, vol. 147, no. 3, pp. 1347–1357, 2008.

[192] A. B. Bleecker and H. Kende, "Ethylene: a gaseous signal molecule in plants," *Annual Review of Cell and Developmental Biology*, vol. 16, no. 1, pp. 1–18, 2000.

[193] Z. B. Liu, T. Ulmasov, X. Shi, G. Hagen, and T. J. Guilfoyle, "Soybean GH3 promoter contains multiple auxin-inducible elements.," *The Plant Cell Online*, vol. 6, no. 5, pp. 645–657, 1994.

[194] L. Van Der Fits and J. Memelink, "The jasmonate–inducible AP2/ERF–domain transcription factor ORCA3 activates gene expression via interaction with a jasmonate–responsive promoter element," *The Plant Journal*, vol. 25, no. 1, pp. 43–53, 2001.

[195] N. Gutterson and T. L. Reuber, "Regulation of disease resistance pathways by AP2/ERF transcription factors," *Current Opinion in Plant Biology*, vol. 7, no. 4, pp. 465–471, 2004.

[196] Z. S. Xu, M. Chen, L. C. Li, and Y. Z. Ma, "Functions and Application of the AP2/ERF Transcription Factor Family in Crop Improvement," *Journal of Integrative Plant Biology*, vol. 53, no. 7, pp. 570–585, 2011.

[197] J. Mizoi, K. Shinozaki, and K. Yamaguchi-Shinozaki, "AP2/ERF family transcription factors in plant abiotic stress responses," *Biochimica et Biophysica Acta*, vol. 1819, no. 2, pp. 86–96, 2012.

[198] K. H. Sohn, S. C. Lee, H. W. Jung, J. K. Hong, and B. K. Hwang, "Expression and functional roles of the pepper pathogen-induced transcription factor RAV1 in bacterial disease resistance, and drought and salt stress tolerance," *Plant Molecular Biology*, vol. 61, no. 6, pp. 897–915, 2006.

[199] M. Jakoby, B. Weisshaar, W. Droge-Laser, J. Vicente-Carbajosa, J. Tiedemann, T. Kroj, and F. Parcy, "bzip transcription factors in *Arabidopsis*," *Trends in Plant Science*, vol. 7, no. 3, pp. 106–111, 2002.

[200] S. C. Lee, H. W. Choi, I. S. Hwang, and B. K. Hwang, "Functional roles of the pepper pathogen-induced bZIP transcription factor, CAbZIP1, in enhanced resistance to pathogen infection and environmental stresses," *Planta*, vol. 224, no. 5, pp. 1209–1225, 2006.

[201] N. Sato and N. Ohta, "DNA–binding specificity and dimerization of the DNA-binding domain of the PEND protein in the chloroplast envelope membrane," *Nucleic Acids Research*, vol. 29, no. 11, pp. 2244–2250, 2001.

[202] K. Hara, M. Yagi, T. Kusano, and H. Sano, "Rapid systemic accumulation of transcripts encoding a tobacco WRKY transcription factor upon wounding," *Molecular and General Genetics*, vol. 263, no. 1, pp. 30–37, 2000.

[203] P. J. Zhang, C. X. Xu, J. M. Zhang, Y. B. Lu, J. N. Wei, Y. Q. Liu, A. David, W. Boland, and T. C. J. Turlings, "Phloem–feeding whiteflies can fool their host plants, but not their parasitoids," *Functional Ecology*, 2013.

[204] T. L. Bailey, N. Williams, C. Misleh, and W. W. Li, "MEME: discovering and analyzing DNA and protein sequence motifs," *Nucleic Acids Research*, vol. 34, no. suppl 2, pp. W369–W373, 2006.

[205] J. L. Riechmann and E. M. Meyerowitz, "The AP2/EREBP family of plant transcription factors," *Biological Chemistry*, vol. 379, pp. 633–646, 1998.

[206] C. P. Song, M. Agarwal, M. Ohta, Y. Guo, U. Halfter, P. Wang, and J. K. Zhu, "Role of an *Arabidopsis* AP2/EREBP-type transcriptional repressor in abscisic acid and drought stress responses," *The Plant Cell Online*, vol. 17, no. 8, pp. 2384–2396, 2005.

[207] D. Kizis, V. Lumbreras, and M. Pages, "Role of AP2/EREBP transcription factors in gene regulation during abiotic stress," *FEBS Letters*, vol. 498, no. 2, pp. 187–189, 2001.

[208] G. Lazarova, Y. Zeng, and A. R. Kermode, "Cloning and expression of an ABSCISIC ACID-INSENSITIVE 3 (ABI3) gene homologue of yellow-cedar (*Chamaecyparis nootkatensis*)," *Journal of Experimental Botany*, vol. 53, no. 371, pp. 1219–1221, 2002.

[209] G. Hagen and T. Guilfoyle, "Auxin–responsive gene expression: genes, promoters and regulatory factors," *Plant Molecular Biology*, vol. 49, no. 3-4, pp. 373–385, 2002.

[210] T. Oyama, Y. Shimura, and K. Okada, "The *Arabidopsis* HY5 gene encodes a bZIP protein that regulates stimulus–induced development of root and hypocotyl," *Genes & Development*, vol. 11, no. 22, pp. 2983–2995, 1997.

[211] M. C. Jewell, B. C. Campbell, and I. D. Godwin, "Transgenic plants for abiotic stress resistance," in *Transgenic Crop Plants*, pp. 67–132, Springer–Verlag, 2010.

[212] N. Rouhier, J. Couturier, and J. P. Jacquot, "Genome–wide analysis of plant glutaredoxin systems," *Journal of Experimental Botany*, vol. 57, no. 8, pp. 1685–1696, 2006.

[213] C. H. Foyer and G. Noctor, "Redox homeostasis and antioxidant signaling: a metabolic interface between stress perception and physiological responses," *The Plant Cell Online*, vol. 17, no. 7, pp. 1866–1875, 2005.

[214] X. Dai, Y. Xu, Q. Ma, W. Xu, T. Wang, Y. Xue, and K. Chong, "Overexpression of an R1R2R3 MYB gene, OsMYB3R–2, increases tolerance to freezing, drought, and salt stress in transgenic *Arabidopsis*," *Plant Physiology*, vol. 143, no. 4, pp. 1739–1751, 2007.

[215] H. Abe, K. Yamaguchi-Shinozaki, T. Urao, T. Iwasaki, D. Hosokawa, and K. Shinozaki, "Role of *Arabidopsis* MYC and MYB homologs in drought–and abscisic acid–regulated gene expression," *The Plant Cell Online*, vol. 9, no. 10, pp. 1859–1868, 1997.

[216] V. Ramirez, A. Agorio, A. Coego, J. Garcia-Andrade, M. J. Hernandez, B. Balaguer, P. B. F. Ouwerkerk, I. Zarra, and P. Vera, "MYB46 modulates disease susceptibility to *Botrytis cinerea* in *Arabidopsis*," *Plant Physiology*, vol. 155, no. 4, pp. 1920–1935, 2011.

[217] V. Ramirez, J. Garcia-Andrade, and P. Vera, "Enhanced disease resistance to *Botrytis cinerea* in myb46 *Arabidopsis* plants is associated to an early down–regulation of *CesA* genes," *Plant Signaling & Behavior*, vol. 6, no. 6, pp. 911–913, 2011.

[218] T. Qi, S. Song, Q. Ren, D. Wu, H. Huang, Y. Chen, M. Fan, W. Peng, C. Ren, and D. Xie, "The jasmonate–ZIM–domain proteins interact with the WD–repeat/bHLH/MYB complexes to regulate jasmonate–mediated anthocyanin accumulation and trichome initiation in *Arabidopsis thaliana*," *The Plant Cell Online*, vol. 23, no. 5, pp. 1795–1814, 2011.

[219] Z. Cheng, L. Sun, T. Qi, B. Zhang, W. Peng, Y. Liu, and D. Xie, "The bHLH transcription factor MYC3 interacts with the jasmonate ZIM–domain proteins to mediate jasmonate response in *Arabidopsis*," *Molecular Plant*, vol. 4, no. 2, pp. 279–288, 2011.

[220] P. D. Duek and C. Fankhauser, "bhlh class transcription factors take centre stage in phytochrome signalling," *Trends in Plant Science*, vol. 10, no. 2, pp. 51–54, 2005.

[221] A. Feller, K. Machemer, E. L. Braun, and E. Grotewold, "Evolutionary and comparative analysis of MYB and bHLH plant transcription factors," *The Plant Journal*, vol. 66, no. 1, pp. 94–116, 2011.

[222] T. Koyama, M. Furutani, M. Tasaka, and M. Ohme-Takagi, "TCP transcription factors control the morphology of shoot lateral organs via negative regulation of the expression of boundary–specific genes in *Arabidopsis*," *The Plant Cell Online*, vol. 19, no. 2, pp. 473–484, 2007.

[223] E. Giraud, S. Ng, C. Carrie, O. Duncan, J. Low, C. P. Lee, O. Van Aken, A. H. Millar, M. Murcha, and J. Whelan, "TCP transcription factors link the regulation of genes encoding mitochondrial proteins with the circadian clock in *Arabidopsis thaliana*," *The Plant Cell Online*, vol. 22, no. 12, pp. 3921–3934, 2010.

[224] Y. I. Zhang and J. G. Turner, "Wound–induced endogenous jasmonates stunt plant growth by inhibiting mitosis," *PLoS One*, vol. 3, no. 11, p. e3699, 2008.

[225] C. Reinbothe, A. Springer, I. Samol, and S. Reinbothe, "Plant oxylipins: role of jasmonic acid during programmed cell death, defence and leaf senescence," *The FEBs Journal*, vol. 276, no. 17, pp. 4666–4681, 2009.

[226] J. E. Park, J. Y. Park, Y. S. Kim, P. E. Staswick, J. Jeon, J. Yun, S. Y. Kim, J. Kim, Y. H. Lee, and C. M. Park, "GH3–mediated auxin homeostasis links growth regulation with stress adaptation response in *Arabidopsis*," *Journal of Biological Chemistry*, vol. 282, no. 13, pp. 10036–10046, 2007.

[227] T. Koyama, N. Mitsuda, M. Seki, K. Shinozaki, and M. Ohme-Takagi, "TCP transcription factors regulate the activities of ASYMMETRIC LEAVES1 and miR164, as well as the auxin response, during differentiation of leaves in *Arabidopsis*," *The Plant Cell Online*, vol. 22, no. 11, pp. 3574–3588, 2010.

[228] J. Hao, L. Tu, H. Hu, J. Tan, F. Deng, W. Tang, Y. Nie, and X. Zhang, "GbTCP, a cotton TCP transcription factor, confers fibre elongation and root hair development by a complex regulating system," *Journal of Experimental Botany*, vol. 63, no. 17, pp. 6267–6281, 2012.

[229] Q. Zhu, M. I. Ordiz, T. Dabi, R. N. Beachy, and C. Lamb, "Rice TATA binding protein interacts functionally with transcription factor IIB and the RF2a bZIP transcriptional activator in an enhanced plant in vitro transcription system," *The Plant Cell Online*, vol. 14, no. 4, pp. 795–803, 2002.

[230] W. Mahomed and N. van den Berg, "EST sequencing and gene expression profiling of defence–related genes from *Persea americana* infected with *Phytophthora cinnamomi*," *BMC Plant Biology*, vol. 11, no. 1, p. 167, 2011.

[231] J. D. Jones and J. L. Dangl, "The plant immune system," *Nature*, vol. 444, no. 7117, pp. 323–329, 2006.

Curriculum Vitae

Parsa Hosseini graduated from St. Joseph's College, Melbourne, Australia, in 2001. He received his Bachelor of Science in Cell Biology and Molecular Genetics from the University of Maryland in 2007 and his Master of Science in Computer Science from Towson University in 2010.