Gene regulatory network inference in soybean upon infection by *Phytophthora sojae*

Brett Hale[1,2,3,¶], Sandaruwan Ratnayake[2,3,¶,#a], Ashley Flory[2,#b], Ravindu Wijeratne[4], Clarice Schmidt[5], Alison E. Robertson[5], and Asela J. Wijeratne[2,3,*]

[1]Molecular Biosciences Graduate Program, Arkansas State University, State University, AR, USA

[2]Arkansas Biosciences Institute, Arkansas State University, State University, AR, USA

[3]College of Science and Mathematics, Arkansas State University, State University, AR, USA

[4]Houston High School, Germantown, TN, USA

[5]Department of Plant Pathology and Microbiology, Iowa State University, Ames, IA, USA

[¶]These authors contributed equally to this work.

[#a]Current address: Donald Danforth Plant Science Center, St. Louis, MO, USA

[#b]Current address: St. Jude Children's Research Hospital, Memphis, TN, USA

*Corresponding author
E-mail: awijeratne@astate.edu

Short title: Plant immunity and gene regulatory networks

Abstract

Phytophthora sojae is a soil-borne oomycete and the causal agent of Phytophthora root and stem rot (PRR) in soybean (*Glycine max* [L.] Merrill). Yield losses attributed to *P. sojae* are devastating in disease-conducive environments, with global estimates surpassing 1.1 million tonnes annually. Historically, management of PRR has entailed host genetic resistance (both vertical and horizontal) complemented by disease-suppressive cultural practices (e.g., oomicide application). However, the vast expansion of complex and/or diverse *P. sojae* pathotypes necessitates developing novel technologies to attenuate PRR in field environments. Therefore, the objective of the present study was to couple high-throughput sequencing data and deep learning to elucidate molecular features in soybean following infection by *P. sojae*. In doing so, we generated transcriptomes to identify differentially expressed genes (DEGs) during compatible and incompatible interactions with *P. sojae* and a mock inoculation. The expression data were then used to select two defense-related transcription factors (TFs) belonging to WRKY and RAV families. DNA Affinity Purification and sequencing (DAP-seq) data were obtained for each TF, providing putative DNA binding sites in the soybean genome. These bound sites were used to train Deep Neural Networks with convolutional and recurrent layers to predict new target sites of WRKY and RAV family members in the DEG set. Moreover, we leveraged publicly available Arabidopsis (*Arabidopsis thaliana*) DAP-seq data for five TF families enriched in our transcriptome analysis to train similar models. These Arabidopsis data-based models were used for cross-species TF binding site prediction on soybean. Finally, we created a gene regulatory network depicting TF-target gene interactions that orchestrate an immune response against *P. sojae*. Information herein provides novel insight into molecular plant-pathogen interaction and may prove useful in developing soybean cultivars with more durable resistance to *P. sojae*.

Keywords: soybean, *P. sojae*, transcription factor, gene regulation, gene regulatory network, deep learning

Author Summary

Global food security is threatened continually by plant pathogens. One approach to circumvent these disease-causing agents entails understanding how hosts balance primary growth and defense upon pathogen perception. Molecular signatures of perception-rendered defense may be leveraged subsequently to develop resistant/tolerant crop plants. Additionally, evidence suggests that the plant immune system is characterized by tuning primary and secondary metabolic activity via transcription factor-mediated transcriptional reprogramming. Therefore, we investigated transcription factor-target gene interactions in soybean upon infection by compatible and incompatible races of *Phytophthora sojae*. Through transcriptome analysis, we found that the interactions elicited vast, overlapping transcriptional responses and identified overrepresented, defense-related transcription factor families. We then generated/acquired DNA-protein interactome data for the most represented transcription factor families in the transcriptome analysis and trained deep learning-based models to predict novel transcription factor targets. Transcription factor/target gene metrics were used to construct a gene regulatory network with prioritized components. We identified hub transcription factors belonging to WRKY and ERF families, the majority of which function in response to various biotic and abiotic stressors. These findings propose novel regulators in the soybean defense response to *Phytophthora sojae* and provide an avenue for the investigation of transcription factor-target gene interactions in plants.

Introduction

Phytophthora sojae Kaufmann and Gerdemann is a hemibiotrophic, homothallic oomycete that renders significant yield losses in soybean (*Glycine max* [L.] Merrill). The pathogen can infect host plants at any developmental stage and is denoted by damping-off in seedlings (early season) as well as root rot and subsequent chlorosis/necrosis in aboveground tissue (late season) [1]. In addition, *P. sojae* oospores can persist in a production environment for several years, limiting the efficacy of most cultural management strategies [2,3]. Thus, the most economical and environmentally benign method to manage the pathogen is the deployment of horizontal and/or vertical host genetic resistance [4]. Horizontal resistance is quantitatively inherited and provides some level of protection against all *P. sojae* pathotypes. However, early-season efficiency is reliant upon complementation with cultural practices [3] as it is only active after the first true leaf stage. Moreover, the polygenic character of horizontal resistance hinders introgression into germplasm [5,6]. Alternatively, vertical resistance (i.e., incompatibility) embodies the classic gene-for-gene concept and renders complete protection against specific pathotypes in a monogenic manner [7]. The selective pressures imposed by vertically resistant soybean have increased the virulence profile of *P. sojae* populations, restricting the use of cultivars with a specific *Resistance to P. sojae* (*Rps*) gene to 8-15 years [2,8]. Therefore, a deeper understanding of the molecular mechanisms governing soybean defense against *P. sojae* is needed to overcome pathogen evolution and ultimately attenuate disease.

During a compatible (virulent) soybean-*P. sojae* interaction, the host plant perceives microbe-associated molecular patterns (MAMPs)/pathogen-associated molecular patterns (PAMPs) and elicits PAMP-triggered immunity (PTI), a basal immune response effective against non-adapted pathogens [9]. Conversely, *P. sojae* secretes *Avirulence* (*Avr*) gene-encoded effector proteins that suppress components of PTI and promote disease. During incompatibility, a

receptor encoded by an *Rps* gene recognizes the cognate *Avr* gene product and activates effector-triggered immunity (ETI), a hypersensitive immune response that potentiates PTI and confers resistance to *P. sojae* [7,9,10]. The combined efforts of PTI and ETI to mitigate disease during incompatibility exemplify the zig-zag model of Jones and Dangl [11] and portray PTI and ETI as distinct events that occur consecutively. A growing body of evidence obscures these boundaries, particularly in plant-*Phytophthora* interactions, instead suggesting that plant defense spans a PTI:ETI continuum [12,13]. For these reasons, Wang et al. [14] proposed a three-layered model of plant immunity comprising a recognition layer, a signal-integration layer, and a defense-action layer. In the context of soybean-*P. sojae* interaction, our understanding of the signal-integration layer remains the most fragmented.

Intra- and/or extracellular pathogen perception triggers a dynamic, highly sophisticated signaling network that balances primary and secondary metabolic activity in a manner preservative of host fitness [15]. Signal integration and convergence accompanying this coordinated stress response are mediated by transcription factors (TFs) and transcriptional cofactors that comprise sensory regulatory networks embedded within phytohormone signaling pathways [16,17]. Dynamism and amplification of such networks are determined by physical interaction between TFs and nucleocytoplasmic receptors [18,19], TF phosphorylation by mitogen-activated protein kinase (MAPK) cascades [20], and feedback regulation of Ca^{2+} signaling, among other mechanisms [17]. While the abundance and diversity of TFs required for immunity vary across plant species and pathosystems [21], elucidated sensory regulatory networks tend to possess members of the bHLH, bZIP, ERF, MYB, NAC, and WRKY families [17,22,23] that collectively direct transcriptional reprogramming of downstream target genes [16]. Isolated studies have evidenced transcriptional reprogramming in soybean upon infection

by *P. sojae* [24,25,26,27] and have identified various TFs associated with defense [28,29,30,31,32,33,34,35,36,37]; yet mechanistic insight regarding TF-target gene interactions and their organization within larger hierarchical networks is lacking. This systems-level information can be unraveled using gene regulatory networks (GRNs) [38], and regulatory hubs identified subsequently through analyses of network tunability and redundancy [21].

In a simplistic model of gene regulation, TFs bind to regulatory DNA motifs in target genes to modulate transcriptional activity [39]. GRNs can be used to discern static and spatiotemporal interactions between TFs and DNA motifs as well as interaction abundance, topology, and influence on target gene expression [40,41]. Various experimental and computational methods have been developed to study gene regulation and function for a phenomenon of interest (e.g., disease resistance) [38]. As an example, TF-DNA interactome approaches such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) and DNA-affinity purification and sequencing (DAP-seq) allow the identification of many TF binding sites at once and can be used to validate interactions inferred by gene expression analysis [41,42]. However, these methodologies are technically- and economically-demanding and are thus difficult to deploy at genome-scale.

Contrarily, *in silico* exploration of TF-target interactions is easily scalable. The majority of such methods leverage guilt-by-association approaches that cannot necessarily predict causality and are thus limited in terms of elucidating regulatory pathways [43]. One can overcome this by using a bottom-up approach to identify *cis*-regulatory elements (CREs), which modulate gene expression by recruiting TFs, as a means to presume TF-target gene interactions. The most popular approach to find CREs is to employ a supervised motif method using a position-specific score matrix (i.e., position weight matrix) and map CREs to a promoter [44].

However, given that CREs are often degenerate and short, this method suffers from high false positive rates. Improvement can be made by considering the evolutionary conservation of CREs (albeit all functional CREs are not necessarily evolutionary conserved) or gene co-regulation. More recently, Deep Neural Network (DNN)-based methods were developed to detect TF binding sites (TFBS) [45,46]. The DNN-based techniques are deemed superior to others given their ability to accept minor CRE variation and sequence context surrounding TFBS and thus transfer across species [47]. For instance, leveraging Arabidopsis (*Arabidopsis thaliana*) cistrome datasets, Akagi et al. [48] constructed convolutional neural network (CNN)-based DNN models for cross-species prediction of TFBS in tomato (*Solanum lycopersicum*). Likewise, Bang et al. [49] used CNN models to predict TFBS in both maize (*Zea mays*) and soybean using maize DAP-seq data. Although high false positive rates were observed by cross-species prediction in the latter study [49], these advances demonstrate the value of DNN-based methods for TFBS prediction within and across plant species.

In the present study, we coupled transcriptomics, *in vitro* TF-DNA interaction profiling, and deep learning to construct a GRN underlying the soybean defense response to *P. sojae* infection (**Fig 1**). We first inoculated hypocotyls of soybean variety Williams 82 (possesses *Rps1k*) with mycelial slurries from *P. sojae* Races 1 or 25, rendering incompatible and compatible interactions, respectively. Transcriptomes were generated from the hypocotyls, and differential gene expression analysis was performed with expression profiles from a mock inoculation serving as a baseline. Following RNA capture-based validation of the experimental design, we assessed TF representation in the differentially expressed gene (DEG) set. The biological significance of overrepresented TF families was inferred by clustering DEGs, assigning functional annotations to each cluster, and observing TF family representation across

defense-related clusters. Next, using DAP-seq data for WRKY and RAV TFs differentially expressed in our DEG set, we obtained promoter-localized TFBS to train DNN models for the prediction of novel WRKY and RAV targets. Furthermore, cross-species target prediction was performed for MYB, WRKY, NAC, ERF, and bHLH TF families using DNN models trained with available Arabidopsis DAP-seq data. We observed the representation of predicted targets in our DEG set and used highly confident TF-target predictions to reconstruct a GRN. Findings in this study provide insight into the regulatory mechanisms governing defense against *P. sojae* and provide new/novel avenues for the molecular breeding of soybean.

Results

Transcriptome analysis and virulence screening

To identify candidate genes involved in the defense response against *P. sojae*, seeds of soybean variety Williams 82 were grown in germination paper and seedling hypocotyls inoculated with a mycelial slurry from a Race 1 isolate (R1; incompatible interaction), a Race 25 isolate (R25; compatible interaction), or sterile media (Mock) following the procedure of Dorrance et al. [50]. At 24 hrs post-infection (hpi), the hypocotyls were collected and used to generate 13 RNA-seq libraries (4 Mock, 4 R1, and 5 R25) spanning two independent inoculations, RNA isolations, and sequencing runs. Additional seedlings were maintained seven days post-infection (dpi) to compare disease development across treatments (**Fig 2a**) [50]. Collectively, the RNA-seq samples comprised over 560 million 100-bp paired-end reads with a mean mapping rate of 95% (**Table S1**). Principal Component Analysis demonstrated that samples clustered according to treatment and sequencing event. To circumvent the latter, we used ComBat-seq [51], a negative binomial regression model, to correct batch effects. We then performed differential gene expression analysis and removed any genes with expression that had been significantly changed (False Discovery Rate < 0.05) between the two batches. Furthermore,

we checked the expression of six genes that displayed stable expression in soybean upon various biotic stresses [52] and a gene possessing a *P. sojae*-inducible promoter (*GmaPPO12*) [53]. The defined reference genes showed no significant differences among treatments in the present study, while *GmaPPO12* showed strong induction due to *P. sojae* infection (**Data S1**). Cooperatively, all R1-inoculated hypocotyls displayed a localized hypersensitive response, while those inoculated with R25 were demarcated by expansive necrotic lesions (**Fig 2a**).

There were 6,042 DEGs (adjusted p-value < 0.05) between R1 and R25 treatments compared to the Mock inoculation (**Data S1**). Among them, 2,387 (39.5%) overlapped between R1 and R25, whereas 175 (2.9%) and 3,480 (57.6%) DEGs were present only in R1 or R25 treatments, respectively (**Fig 2b**). To validate and contextualize our findings, we isolated total RNA from R1-, R25-, and Mock-inoculated hypocotyls (independent experiments from those used for RNA-seq), generated adapter-ligated cDNA libraries, and performed RNA hybridization-based enrichment followed by high-throughput sequencing (i.e., Capture-seq). The RNA hybridization was performed with biotinylated RNA baits designed for seven genes that showed elevated expression upon *P. sojae* infection (i.e., pathogen-induced genes) across all NILs reported in Lin et al. [27] as well as both interaction types in our RNA-seq dataset. Baits were incorporated for the six aforementioned reference genes as an internal standard. We recovered 100% of our capture library. Further, 7/7 pathogen-induced genes showed significantly elevated expression in inoculated vs Mock samples, while 5/6 reference genes showed stable expression across treatments (**Fig S1a**,**b**).

Identification of defense-related TF families

Following transcriptome validation by Capture-seq, we used PlantTFDB [54] to identify TF-annotated genes in our DEG set (genes differentially expressed in at least one interaction).

We found 447 (7.3% of DEGs) distributed across 43 TF families. Among these families, WRKY was the most represented by total abundance (n = 65 encoding DEGs), followed by ERF (n = 56), bHLH (n = 44), MYB (n = 37), and C2H2 (n = 31) (**Fig 2c**). The RAV TF family was most represented by the percentage of genome-wide proportion (60%) (**Fig 2c**). Moreover, proportions of WRKY, ERF, HSF, CAMTA, and RAV TF-encoding genes were significantly enriched in the DEG set (hypergeometric p-value < 0.05).

To predict functions of the defined TF families in the present pathosystem, K-Means clustering was used to segregate DEGs into nine co-expression clusters. In clusters 3 and 6 (hereafter “down-regulated clusters"), both compatible and incompatible interactions showed reduced expression in comparison to Mock, with mean expression higher in the incompatible interaction than in the compatible (**Fig 2d**). A reciprocal pattern was observed in clusters 1, 2, 4, 5, 7, 8, and 9 (hereafter “up-regulated clusters"), wherein the majority of genes were up-regulated compared to the Mock, and the compatible interaction displayed the highest mean expression (**Fig 2d**). We explored these trends by assigning functional annotations to each gene cluster with GO term and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses [55,56]. Interestingly, up-regulated clusters displayed enrichment for secondary metabolism and signaling-related KEGG terms, particularly those inferring MAPK signal transduction (DEGs/Total genes: 57/229), phenylpropanoid biosynthesis (48/214), flavonoid biosynthesis (23/67), and plant-pathogen interaction (64/280) (**Fig 2e**; **Data S1**). The majority of genes corresponding to these terms were differentially expressed in both compatible and incompatible interactions compared to Mock, with the exception of plant-pathogen interaction and MAPK-related pathways, where 27 and 35 genes, respectively, were differentially expressed exclusively during the compatible interaction. Down-regulated clusters

were enriched with GO and KEGG terms related to primary metabolism (e.g., photosynthesis) (**Fig 2e**; **Data S1**), which likely reflected a reallocation of cellular energy to defense [57].

Next, we examined TF representation in the co-expression clusters involved in the signaling and secondary metabolic responses to *P. sojae* (i.e., up-regulated clusters). WRKY was the most abundant TF family for four clusters (4, 7, 8, and 9), ERF for two clusters (1 and 5), and C2H2 for cluster 2 (**Fig 2e**). Furthermore, four TF families were represented in more than half of the up-regulated clusters, with ERF present in 7/7, WRKY in 6/7, C2H2 in 4/7, and MYB in 4/7.

Comprehensive identification of TFBS

Transcriptome analysis prompted the identification of TFBS (and thereby target genes) for defense-related TF families. To this end, we performed DAP-seq for GmWRKY30, whose corresponding gene (*Glyma.06G125600*) was differentially expressed by RNA-seq. In addition, *GmWRKY30* homologs promoted resistance to hemibiotrophic and necrotrophic fungi in rice (*Oryza sativa*) [58] and to *Cucumber mosaic virus* in Arabidopsis [59]. Treatments were prepared for DAP-seq by inoculating soybean hypocotyls with an R1 *P. sojae* isolate (sample hereafter referred to as "WRKY30_P1") or sterile media (hereafter "WRKY30_M1") (see "Materials and Methods" for full details). WRKY30_P1 and WRKY30_M1 displayed 6,415 and 2,083 peak regions, respectively, corresponding to various genomic features (**Fig 3**; **Data S3**; **Table S2**). Motif enrichment analysis was then performed with MEME-ChIP [60] and demonstrated that bound regions present in both samples were statistically enriched for the WRKY TF binding site, W-box (TTTGAC/T), implicating that these regions were indeed bound by a WRKY TF. To establish regulatory roles of GmWRKY30 during *P. sojae* infection, we obtained the peaks annotated as promoters (defined as 1,000 bp up- and downstream of the transcription start site [TSS]) and retained the regions shared by WRKY30_P1 and

WRKY30_M1 (235 promoters) as well as those found exclusively in WRKY30_P1 (1,110 promoters). Of these, 212 promoters corresponded to genes in our DEG set. Interestingly, 174/212 were present exclusively in the WRKY30_P1 sample (**Data S3**). Given that the only difference between WRKY30_P1 and WRKY30_M1 samples was DNA methylation marks, this observation suggested that the soybean genome undergoes differential methylation during *P. sojae* infection, perhaps as a mechanism to prevent autoimmunity [16]. Moreover, 35/212 target genes were annotated as TFs, indicating putative auto- and cross-regulatory activity of GmWRKY30 during the immune response. Thirteen of the 35 TF-encoding DEGs belonged to the WRKY TF family, and all were present in the up-regulated clusters (**Data S1**; **Data S3**). Furthermore, KEGG functional annotation revealed that eight GmWRKY30 targets were part of the MAPK or plant-pathogen interaction pathways described above (**Data S1**; **Data S3**). These data concomitantly suggest that GmWRKY30 regulates the expression of other TFs and signaling components during soybean-*P. sojae* interaction.

RAV was the most abundant TF family by a percentage of genome-wide proportion in our DEG set; therefore, we obtained DAP-seq data for a GmRAV TF from Wang et al. [61]. The corresponding gene (*Glyma.10G204400*) was significantly up-regulated in both compatible and incompatible interactions compared to Mock and displayed similar expression dynamics in Lin et al. [27] upon *P. sojae* infection. In the present analysis, GmRAV was bound to 3,409 promoters corresponding to 389 genes in our DEG set (**Data S1**; **Data S4**). Of these, 29 encoded TFs. One hundred seventy-six of the 389 targets (45%) were present in either cluster 3 or 6 (down-regulated clusters) (**Data S1**; **Data S4**). Cooperatively, functional enrichment demonstrated that GmRAV targets included genes relevant to photosynthesis and carbon metabolism (**Data S1**;

Data S4), indicating that GmRAV may repress primary metabolic activity during pathogen infection.

DNN prediction of TFBS

While TFs in a structural family have the capacity to function distinctly *in vivo*, they often share intrinsic CRE preference [62,63,64]. Therefore, we hypothesized that binding sites obtained for a single TF could be used to predict binding sites for other members of the same family. To test this hypothesis, we trained Convolutional Recurrent Neural Networks (CRNNs), which couple CNN and bi-directional long short-term memory layer architecture [65] (**Fig 4**), using peak summits of WRKY30_P1 and WRKY30_M1 samples with either 32- or 201-bp peak regions (**Fig S2**). The CRNN with a 201-bp region outperformed the CRNN with a 32-bp peak region and displayed an 89% validation accuracy, 90% test accuracy, and a false positive rate of less than 3% (**Table 1**). We trained a similar model for GmRAV, which had an 89% validation accuracy, 89% test accuracy, and a less than 3.5% false positive rate (**Table 1**). Moreover, for both GmWRKY30 and GmRAV CRNN models, the area under the receiver operating characteristic (auROC) curve was beyond 0.88, and the area under the precision-recall curve (auPRC) beyond 0.81 (**Fig S3**). To determine if a CRNN model trained for one TF could generalize to members of the same family intraspecifically, we generated AmpDAP-seq data for GmWRKY2 (homologous to AtWRKY2 and encoded by *Glyma.06G320700*). For this sample, we first observed peak distribution across genomic features (**Fig S4**; **Data S5**) and used MEME-ChIP to verify statistical enrichment of the W-box CRE within peak regions. We then used peak regions to test if the GmWRKY30 CRNN could predict GmWRKY2-bound sites. The prediction accuracy was above 82%, with a false positive rate of less than 6% (**Fig 5a**). Furthermore, we explored the interspecific generalization capacity of the GmWRKY30 CRNN by performing a

cross-species prediction on AtWRKY30 (encoded by *AT5G24110*; homologous to GmWRKY30) DAP-seq data (generated by O'Malley et al. [66] and reanalyzed by Song et al. [67]). For this analysis, the prediction accuracy was above 84% with a false positive rate of less than 7% (**Fig 5a**).

Table 1: Summary of GmWRKY and GmRAV CRNN models

Model	Train Acc. (%)	Val Acc. (%)	Test Acc. (%)	TPR (%)	TNR (%)	FPR (%)	FNR (%)
WRKY	90.97	89.77	90	33.67	56.21	2.71	7.41
RAV	89.09	89.2	89	44.88	45.33	3.49	6.3

Yet, it remained unclear whether these patterns would recur within/across TF families. Therefore, we utilized DAP-seq data for AtWRKY30, AtMYB62, AtMYB108, AtMYB119, AtNAC031, AtNAC053, and AtNAC057 [66,67] to train additional CRNNs. The AtWRKY30 model validation accuracy was 97%, test accuracy 97.33%, and false positive rate 0.13% (**Table 2**). The model was used subsequently to predict binding sites for 17 other AtWRKY TFs with available DAP-seq data [66,67] and presented a mean prediction accuracy of 89% with a mean false positive rate under 1% (**Fig 5b**; **Data S6**). The three AtMYB models had 92-98% validation accuracies, 91-98% test accuracies, and 0.8-2.76 % false positive rates (**Table 2**) and were used to predict binding sites for 5 AtMYB TFs [66,67], presenting a mean prediction accuracy of 64.92% and a mean false positive rate of 1.53% (**Fig 5b**; **Data S6**). Similarly, the three AtNAC models posed validation accuracies between 95.5-98%, test accuracies between 95-98%, and a false positive rate ranging between 0.6-1.79% (**Table 2**). These models were used for predicting

binding sites for 15 other AtNAC TFs [66,67] with a mean prediction accuracy of 79.6% and a mean false positive rate of 1.09% (**Fig 5b**; **Data S6**). These findings indicated that a model trained using the TFBS of one family member could predict the TFBS of another member with reasonable accuracy. Therefore, GmWRKY30 and GmRAV models were used to predict TFBS on promoters of the DEGs from the transcriptome analysis (**Fig 5c**). To further reduce false positives, we scanned the same promoter regions using Find Individual Motif Occurrences (FIMO) [68] with motifs obtained from the JASPAR database [69] to find CREs. The results from the FIMO scan were overlapped with our predicted sites to get a highly confident set of TFBS. From this, we obtained 3,298 GmWRKY targets with 267 corresponding to TF-encoding genes (60% of TFs in the DEG list). Similarly, GmRAV-predicted targets included 1,925 genes, 121 of which encoded TFs (27% of TFs in the DEG list) (**Data S1**).

Table 2: Summary of Arabidopsis data-based CRNN models used for generalization tests

Model	Common names	Val Acc. (%)	Test Acc. (%)	FPR (%)	TPR (%)	TNR (%)	FNR (%)
AT1G68320	AtMYB62	97.47	97	0.8	48	49.1	2.06
AT3G06490	AtMYB108	92.02	91	2.76	43.71	47.24	6.29
AT5G58850	AtMYB119	98.09	98	0.76	50.09	48.22	0.93
AT5G24110	AtWRKY30	97	97.33	0.13	47.41	49.89	2.57
AT1G76420	AtNAC031	95.52	95	1.79	47.6	47.34	3.26
AT3G17730	AtNAC057	97.33	97	0.98	48.31	49.16	1.54
AT3G10500	AtNAC053	98.15	98	0.6	48.96	48.72	1.73

Cross-species prediction of soybean TFBS

Based upon the interspecies generalization capacity of the GmWRKY30 model, we wanted to further leverage the homology between Arabidopsis and soybean to build CRNNs and conduct cross-species predictions for other defense-related TF families. In doing so, we built

CRNNs for AtERF, AtbHLH, AtC2H2, AtRAV, AtWRKY, AtMYB, and AtNAC TF families by combining available DAP-seq data for each family. With the exception of the AtRAV and AtC2H2 models, the training, validation, and testing accuracies were above 90%, and false positive rates less than 1.2% for all models (**Table 3**). Both AtRAV and AtC2H2 models had training, validation, and testing accuracies of less than 86% with increased false negative rates and were thus not used for subsequent analyses (**Table 3**). Next, we generated AmpDAP-seq data for GmMYB61 (encoded by *Glyma.10G142200*) (**Fig S4**; **Data S5**) and used AtWRKY and AtMYB CRNNs to perform cross-species predictions on GmWRKY30 DAP- and GmMYB61 AmpDAP-seq data, respectively. Both predictions demonstrated modest accuracies (approximately 61%), yet less than 1% false positive rates (**Table 4**). We posited that, while Arabidopsis-to-soybean predictions would likely miss some true TFBS, we could have confidence in those classified as bound. Therefore, we used AtMYB, AtERF, AtNAC, and AtbHLH CRNNs to predict TFBS for promoter regions of our DEGs (**Fig 5c**). Consistent with soybean CRNNs, all of the predicted TFBS were overlapped with FIMO predictions to get a highly confident set of targets.

Table 3: Summary of Arabidopsis data-based CRNN models used for cross-species predictions

Model	Train Acc. (%)	Val Acc. (%)	Test Acc. (%)	TPR (%)	TNR (%)	FPR (%)	FNR (%)
WRKY	97.79	98.12	98.00	48.94	49.83	0.63	0.60
MYB	93.23	92.22	92.00	41.93	50.16	1.04	6.88
ERF	96.10	95.85	95.76	46.14	49.62	0.43	3.80
NAC	92.34	92.11	92.00	41.69	49.68	1.20	7.40

BHLH	96.80	96.00	96.22	47.45	48.77	0.76	3.03
RAV	74.57	75.00	73.00	22.26	50.74	0.89	26.11
C2H2	85.20	85.00	84.60	36.42	48.18	1.22	14.18

Table 4: Arabidopsis-to-soybean cross-species prediction results

Model	Cross Pred. Acc. (%)	TPR (%)	TNR (%)	FPR (%)	FNR (%)
WRKY	61.16	5.81	55.35	0.52	38.32
MYB	60.81	10.92	49.89	0.90	38.29

GRN inference underpinning host defense

The combination of soybean and Arabidopsis CRNNs allowed the prediction of TFBS corresponding to 5,505 genes in the DEG set (**Fig 6a**). Global and family-level GRNs were thereby constructed with TFs represented by nodes and target genes by edges (**Fig 6b; Fig S5**). We then examined TF-level TFBS to prioritize nodes in the global GRN. To be considered, nodes had to possess statistically enriched binding motifs (q-value < 0.05) compared to a randomly shuffled input sequence (determined by the Simple Enrichment Analysis algorithm of Bailey and Grant [70]) and had to have corresponding genes expressed in the transcriptome analysis (**Fig 6d**). The 118 nodes meeting these criteria were prioritized by degree centrality, TF co-occurrence, and the expression pattern of corresponding genes. Degree centrality was determined from outdegree (number of edges directed to each node) and cumulative indegree (indegree = number of nodes to which an edge is directed; cumulative indegree = combined indegree for all edges of a node). Both measures were scale-free and displayed a power-law distribution (**Fig 6d**) as expected of GRN architecture [71]. Furthermore, TF cooperativity/co-occurrence metrics are required to effectively model causal GRNs [72]. We assessed putative TF

co-occurrence with TF-COMB (Transcription Factor Co-Occurrence using Market Basket analysis) [73] and selected cosine association score as the objective similarity measure for co-occurring TF pairs (**Fig 6c**). Cumulative cosine (total cosine association score across all co-occurrences) was determined for each node. Lastly, we calculated the mean $|\log_2 FC|$ for node-corresponding genes across both interactions (R1 vs Mock and R25 vs Mock) for further prioritization. Nodes/node-corresponding genes present in the upper quartile for all four parameters were considered hubs (**Fig 6d**). A reciprocal approach prioritized edges by indegree, cumulative outdegree (combined outdegree of all nodes to which an edge is directed), sum cumulative cosine (combined cumulative cosine of all nodes to which an edge is directed), and mean $|\log_2 FC|$ across both interactions (**Fig S6**).

Hub nodes corresponded to 14 genes encoding 13 TFs, all of which belonged to ERF and WRKY families (**Fig 6d,e**). Thirteen of these were differentially expressed in the transcriptome analysis and were present in co-expression clusters 1, 4, 5, and 8 (up-regulated clusters with defense-related functional annotations). Furthermore, we assessed KEGG annotations for putative hub node targets (n = 1,664), and 170 targets were annotated to one or more functions. Interestingly, 67% (114/170) of the targets possessed primary metabolic terms whereas 33% (56/170) possessed terms related to defense/secondary metabolism. It is therefore reasonable to hypothesize that these hub nodes are core components of the defense-growth tradeoff in soybean by regulating transcriptional reprogramming.

Discussion

P. sojae is a yield-devastating soybean pathogen subject to rapid genetic diversification and expansion within and across production environments. To unravel regulatory signatures of host defense during *P. sojae* infection, we coupled multi-omic and computational analyses to identify TF-target gene interactions at 24 hpi. In doing so, we conducted the first comparative

transcriptomic study for compatible and incompatible soybean-*P. sojae* interactions within a single host genotype. Similar gene expression profiles were observed across the interaction types, implicating significant overlap between PTI- and PTI + ETI-mediated defense at 24 hpi. Enkerli et al. [74] revealed ultrastructural differences between compatible and incompatible soybean-*P. sojae* interactions at 4 hpi, with programmed cell death and impedance of hyphal growth evident in the incompatible interaction by 15 hpi. Thus, it is likely that maximal expression of PTI-potentiating transcripts during incompatibility precludes or at least precedes the *P. sojae* transition from biotrophy to necrotrophy (12-24 hpi [75,76]), and succedent activity reflects a reduction in hypersensitivity required to offset fitness costs of induced resistance. This concept draws parallels to other pathosystems in which compatible and incompatible interactions displayed consistent trends in gene expression, with the incompatible interaction eliciting a heightened, more immediate immune response [77,78,79]. A complementary explanation is that PTI/ETI convergence is attributed to the *P. sojae* arsenal present during both compatible and incompatible interactions, including MAMPs and a conserved suite of effectors [80]. Moreover, overlapping expression profiles may correspond primarily to PTI signatures not targeted by *Avr*-encoded effectors. Nevertheless, findings herein necessitate the investigation of compatible and incompatible interactions in tandem to elucidate *Rps* gene-exclusive defense mechanisms.

Plant immune signaling is remarkably tunable yet robust, allowing the coordination of defense and growth in a manner that maximizes host fitness [21,81,82]. Prior studies suggest signal integration underpinning the defense-growth trade-off is imposed by TF regulatory networks that modulate immune responses through transcriptional reprogramming [17]. In the present study, K-means clustering of DEGs rendered nine gene co-expression clusters, seven of which were up-regulated and corresponded to defense-related functional annotations (e.g.,

MAPK signaling and secondary metabolism). Cooperatively, these clusters were enriched with statistically overrepresented TF families (i.e., MYB, WRKY, NAC, ERF, and C2H2) known to regulate plant secondary metabolism [22,23] and reported in prior soybean-*P. sojae* studies [29,30,32,33,34,35,37,83]. Furthermore, WRKY and ERF were the most abundant TF families across the seven clusters. This is consistent with findings in Arabidopsis where MAPK-WRKY and MAPK-ERF complexes regulated core immune signaling through transcriptional reprogramming [17,20,84]. The remaining gene co-expression clusters were down-regulated and demarcated by growth and reproductive functional terms. Interestingly, DAP-seq suggested that gene targets of RAV (the most represented TF family in the DEG set by a percentage of genome-wide proportion) were enriched in these two clusters. This finding is consistent with prior studies wherein GmRAV has been reported to play a role in photosynthesis, senescence, abiotic stress tolerance, and phytohormone-mediated signaling [85,86,87] and act as a transcriptional repressor to delay flowering [61]. To our knowledge, the present study is the first to propose a function for GmRAV during immunity, where it acts as a repressor of primary metabolism. Thus, our transcriptome analysis evidences transcriptional reprogramming governing the defense-growth trade-off in soybean upon *P. sojae* infection.

DAP-seq data were generated/obtained for the most represented TF families by total abundance and percentage of genome-wide proportion in the transcriptome analysis, and promoter-localized DAP-seq peaks were used to train CRNNs (DNNs composed of convolutional and recurrent layers) for the prediction of novel TFBS. We leveraged CRNNs for their capacity to learn randomly composite, predictive sequence patterns [47]. Previous studies suggest that binding site motifs, along with nearby sequence features and their organization in the genome, play a vital role in TF binding. The convolutional filters in CRNNs can capture and

train these binding site motifs and nearby sequence features, while recurrent layers can learn their multidimensional organization [47]. In addition, such hybrid model architecture has been used successfully to predict TF-target interactions with human data [65,88,89]. Here, our CRNN models were capable of predicting TFBS for the selected TF families in soybean and Arabidopsis with ~90% accuracy. The exclusive use of DEG promoters for binding site prediction increased the likelihood that targets were biologically valid, as the correlation between stable TF binding and TF regulation is vastly inconsistent and oftentimes poor [72,90,91]. Moreover, CRNNs trained for one TF could find TFBS for other members of the same family. We supported this notion in soybean by generating binding site data for a second WRKY TF and using the pre-existing WRKY CRNN to predict its targets. We validated our findings in another species by training additional CRNNs with Arabidopsis DAP-seq data for WRKY, MYB, and NAC TFs and predicting TFBS for various members of the respective families. In every instance, CRNNs were capable of generalizing with acceptable accuracy, posing significant potential as an alternative to wet lab-based TFBS assays. Altogether, findings herein reflect the ability to train highly accurate CRNNs for the prediction of TFBS in plants.

The DNA sequence preference of TFs is largely conserved across phylogenetically-related species, leading to the advent of deep learning-based approaches for cross-species TFBS prediction [47,64,91]. However, recent attempts at mouse-to-human/human-to-mouse and maize-to-soybean cross-species predictions suffered from high false positive rates [47,49]. We hypothesized that we could overcome such limitations due to the evolutionary proximity of soybean and Arabidopsis (two diploid, dicotyledonous species). In the present study, Arabidopsis-to-soybean predictions had moderate accuracy (approximately 60%) with low false positive rates (less than 1%). Interestingly, soybean-to-Arabidopsis predictions displayed a

higher accuracy than the Arabidopsis-to-soybean. Nitta et al. [92] investigated TFBS conservation between *Drosophila* and mammals, finding that novel binding site specificities could arise via gene duplication and subsequent divergence. Perhaps the lower Arabidopsis-to-soybean prediction accuracies reflect the expansion of WRKY and MYB families in soybean, rendering soybean-specific CRE preferences. Furthermore, the lower prediction accuracies may indicate contributions to TF genomic occupancy beyond DNA sequence affinity (e.g., chromatin state; presence of cofactors), which have been demonstrated to significantly influence immunity-related transcriptional dynamics in plants [16]. Thus, the integration of complementary information (e.g., ATAC-seq data) into existing CRNN frameworks will allow for more accurate model training in the future.

We predicted TFBS for WRKY, RAV, NAC, ERF, bHLH, and MYB TF families with soybean- and Arabidopsis-based CRNNs, overlaid CRNN predictions with FIMO scans, and constructed global and family-level GRNs. Within a GRN, some TFs act as hubs to regulate many genes, while target genes are typically regulated by multiple TFs [72]. Therefore, we identified hub nodes, which presumably have an inordinate effect on phenotype [93], by integrating motif enrichment analysis, degree centrality, TF co-occurrence, and gene expression metrics. Interestingly, all hub-corresponding genes encoded WRKY or ERF TFs, suggesting these families are central components of the host immune response at 24 hpi. One could attribute this in part to the use of intraspecies prediction for WRKY and RAV and cross-species prediction for the other TF families, the latter of which was prone to high false negative rates that likely influenced degree centrality. Yet, WRKY and ERF were the most represented TF families in defense-related DEG clusters derived from the transcriptome analysis, which, when coupled with functional annotations of hub node targets, reinforced the pertinence of the two families for

host defense. Moreover, the majority of hub genes identified here demonstrated differential expression in soybean upon various biotic and abiotic stresses (**Fig 6d**) [27, 94,95,96,97,98,99,100,101,102] . Additional efforts must be used to functionally validate TF-target predictions, which remains a bottleneck in GRN research [72,103].

Nevertheless, this research poses limitations that must be considered. First, the inoculation procedure of Dorrance et al. [50] was used in the present study: hypocotyl wounding occurred prior to the placement of a mycelial slurry (sterile media for Mock tissues). It is possible that mechanical tissue disruption increased damage-associated molecular pattern (DAMP)-related signaling, and that DAMP + MAMP perception triggered PTI beyond what occurs *in situ* [104]. Second, the existing network lacks the temporal resolution required to fully elucidate dynamic phenomena such as plant-pathogen interactions [72]. Thus, future experiments will include more natural inoculation procedures, the integration of time-series expression and epigenomic data into CRNN frameworks, the evaluation of additional defense-relevant TF families, and the molecular investigation of *P. sojae* effector targets.

Conclusions

In summary, TF families emphasized in this study (particularly ERF, WRKY, and RAV) are likely core components of sensory regulatory networks required for the balance of primary and secondary metabolic responses to *P. sojae* infection. Interactions predicted here shed light on convergent and discrete transcriptional complexes during host immunity and provide a framework for data integration, functional validation, and the prediction of novel regulatory components for disease resistance [21]. We also provide a framework for improved TFBS prediction by coupling high-throughput sequencing data and CRNNs. Consequently, the information herein may prove useful for the circumvention of *P. sojae* pathogenicity through the

modulation of defense-related pathways and the resultant derivation of disease-resistant soybean genotypes.

Materials and Methods

Biological materials, pathogenicity testing, and RNA isolation

For all experiments, soybean and *P. sojae* materials were generated at Iowa State University in the lab of Dr. Alison Robertson. Axenic-grown *P. sojae* isolates were transferred from a 20% clarified vegetable juice (V8) medium onto a soft-diluted V8 medium and incubated at 25°C in the dark as defined by Dorrance et al. [50]. Concomitantly, seeds of soybean varieties Williams 82 (*Rps1k*) and Williams (*Rps*) were grown on moistened germination paper at 25°C, 16 h d^{-1} light, and 90% relative humidity. On day 7 of plant growth, seedling hypocotyls were incised 1 cm below the cotyledonary node with a sterile razor blade and inoculated with 0.1 mL mycelial slurry of a *P. sojae* Race 1 isolate (avirulent on *Rps1k*), a Race 25 isolate (virulent on *Rps1k*), or sterile media. At 24 hpi, the mycelial slurries were washed off with deionized water, and a 2 cm fragment of each hypocotyl was cut and immediately frozen in liquid nitrogen.

For each inoculation, 10 or more seedlings per treatment were kept 7 d to monitor disease development. Seedlings were scored as asymptomatic or symptomatic depending on the absence/presence of lesions and necrotic tissue in at least 90% of replicates. For an inoculation to be successful, Williams seedlings displayed disease symptoms when infected with either *P. sojae* pathotype. Williams 82 seedlings displayed hypersensitivity upon inoculation with Race 1 and were symptomatic upon infection with Race 25. Moreover, mock inoculations rendered asymptomatic, clean wounds for both varieties.

Total RNA was isolated from frozen Williams 82 hypocotyls with the NEB Monarch Total RNA Miniprep Kit (Cat #T2010S) and quantified using a Qubit fluorometer paired with the RNA high-sensitivity assay kit (Cat #Q32852). RNA purity was estimated from A260/A280 and A260/A230 ratios using a NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific) and further assessed by gel electrophoresis (1% agarose gel at 120V for 40 min). Samples were stored at −80 °C until use.

RNA-seq

Total RNA was sent to Novogene Corporation (Sacramento, CA, USA) for library preparation and sequencing. RNA purity was assessed using a NanoPhotometer® spectrophotometer (Implen, Westlake Village, CA, USA). RNA integrity and quantitation were monitored with the RNA Nano 6000 Assay Kit (Cat #5067-1511) of the Agilent Bioanalyzer 2100 system (Agilent Technologies, CA, USA). Following quality control, cDNA libraries were prepared from 1 µg total RNA using the NEBNext Ultra™ II RNA Library Prep Kit for Illumina (Cat #E7770S) paired with the NEBNext Poly(A) mRNA Magnetic Isolation Module (Cat #E7490) following manufacturer's instructions. Library quality was assessed on the Agilent Bioanalyzer 2100 system, and libraries were sequenced on an Illumina platform.

Resultant short reads were processed using fastp software (v0.20.1) [105] for the removal of adapter sequences and low-quality reads (Phred <33). Clean reads were then mapped to the soybean reference genome (Gmax_508_v4.0.softmasked) with HISAT2 (v2.0.5) [106]. FeatureCounts (v1.5.0-p3) [107] was used to summarize read counts per gene. Only the genes that had a mean count > 20 across all samples were considered for further analyses to improve the sensitivity for differential gene expression analysis. The batch-level bias was removed using Combat-Seq in the Bioconductor R package sva (v3.44.0) [51], and differential gene expression

analysis was performed using the DESeq2 package (v1.34.0) [108]. Genes with expression significantly changed in the R1 and R25 treatments compared to the Mock treatment (adjusted *p*-value □≤0.05) were deemed DEGs. Furthermore, we used two experimental batches as a factor for the DESeq2 model, and any genes with an adjusted p-value □≤0.05 between the two batches were removed from further analyses.

The 6,042 genes in our DEG set were clustered using the Bioconductor R package coseq (v1.17.2) [109,110]. The batch effect-corrected count data were used as the input for the coseqR and centered log-ratio-transformation and trimmed means of M values normalization were performed to normalize the counts. Genes assigned to each cluster were used for visualization and GO and KEGG enrichment analyses. For GO enrichment analysis, GO terms were downloaded from the Gene Ontology Meta Annotator for Plants (GOMAP) database [111], and the enrichment analysis was performed for each co-expression cluster using the TopGO Bioconductor package (v2.48.0) [112]. GO terms with a *p*-value □≤0.05 were considered enriched terms. Similarly, for the KEGG enrichment analysis, terms were downloaded from the KEGG database, and enrichment analysis was performed using the clusterProfiler Bioconductor package (v4.4.4) [113]. KEGG pathways with an adjusted *p*-value □≤0.05 were considered overrepresented. Genes encoding putative TFs were annotated using PlantTFDB [54]. To calculate the statistical significance between observed TF abundance in our DEG set and their genome-wide proportions, we conducted a proportion test using the prop.test function in RStudio (v3.6.3) [114]. Data visualization was performed using ggplot2 (v3.3.6) [115].

<u>Capture-seq</u>

Capture-seq was used to validate the expression of pathogen-induced genes and internal standards (*n* = 13 genes). Hypocotyl inoculation and RNA isolation were performed as described above. Adapter-ligated cDNA libraries were then prepared from total RNA using the NEBNext Ultra™ II RNA Library Prep Kit for Illumina (Cat #E7770S) following appendix modifications for size selection of 300 nt inserts (420 nt final library size). In doing so, mRNA was fragmented using First-Strand Synthesis Reaction Buffer and Random Primer Mix (2X) at 94°C for 10 min (compared to 15 min in the protocol). Moreover, the incubation time during first-strand cDNA synthesis was increased from 15 to 50 min at 42°C. Size selection of libraries was performed using 25 and then 10 µl of Agencourt AMPure XP beads (Beckman Coulter, Brea, CA, USA, Cat #A63880). All libraries were quantified with a Qubit fluorometer using the dsDNA high-sensitivity assay kit (Cat #Q32851) and visualized by gel electrophoresis (2% agarose gel at 100V for 60 min).

120-nt biotinylated RNA baits were designed by Integrated DNA Technologies (IDT, Coralville, IA, USA) and encoded the sense strand of the genes of interest. Sequence capture was then performed as described in the xGen hybridization protocol from IDT (http://sfvideo.blob.core.windows.net/sitefinity/docs/default-source/protocol/xgen-hybridization-capture-of-dna-libraries.pdf?sfvrsn=ab880a07_6) with modifications. 500 ng of each barcoded Illumina library was pooled into a single 1.5 mL low-bind microcentrifuge tube and combined with 5 µg of salmon sperm DNA (Cat #15632011) and 2 µl of xGen Universal BlockersTS Mix (Cat #1075474) to prevent bait hybridization with repetitive elements/adapter sequences. Samples were dried for ~90 min in an ISS110 SpeedVac System (Thermo Fisher Scientific). Pelleted material was resuspended in 8.5 µl of 2X hybridization buffer, 2.7 µl of Hybridization Buffer Enhancer, 4 µl of a working bait stock (100 attomoles/bait/µl), and 1.8 µl of nuclease-free

water to a final volume of 17 µl. A hybridization reaction was then performed with an initial denaturation at 95°C for 30 s followed by a 16-hr incubation at 65°C. M-270 Streptavidin beads were equilibrated, pelleted by a magnet, and mixed with the hybridization product for another 45-min incubation at 65°C. Heated and room temperature washes were performed as recommended and 20 µl of nuclease-free water was added to the beads. Post-capture PCR was then performed by adding the following components to the capture product: 1.25 µl of xGen library amplification primer, 10 µl of 5X Phusion HF Buffer, 1 µl of 10 mM dNTP, 0.5 µl Phusion High-Fidelity DNA Polymerase (Cat #M0530S), and 17.25 µl of nuclease-free water to a final volume of 50 µl. Amplification settings included polymerase activation at 98°C for 45 s followed by 10 cycles of denaturation at 98°C for 15 s, annealing at 60°C for 30 s, extension at 72°C for 30 s for each cycle, and a final extension at 72°C for 1 min. PCR fragments were purified using Agencourt AMPure XP beads (Cat #A63880) and eluted using 0.1X Tris-Ethylenediamine Tetraacetic Acid. Captured libraries were quantified with a Qubit fluorometer using the dsDNA high-sensitivity assay kit (Cat #Q32851). Library quality was assessed on the Agilent Bioanalyzer 2100 system (Agilent Technologies, Santa Clara, CA, USA) at Novogene Corporation. Samples were then sequenced on an Illumina platform with 150-bp paired-end reads.

The resultant short reads were preprocessed to remove poor-quality reads/sequencing artifacts using Cutadapt (3.0) [116] and BBDuk (https://sourceforge.net/projects/bbmap/) (Phred <30). The preprocessed short reads were aligned to soybean primary transcripts (Gmax_275_Wm82.a2.v1.transcript_primaryTranscriptOnly.fa) obtained from Phytozome [117] using Kallisto pseudoaligner (v0.46.1) [118] with default parameters. Differential gene

expression analysis and further data processing were performed in RStudio (v3.6.3) [114] as described for the RNA-seq analysis.

DAP- and AmpDAP-seq

DAP- and AmpDAP-seq experiments were conducted as described previously [42]. The open reading frame of each TF was cloned independently into a Gateway-compatible pIX-HALO expression vector containing an N-terminal HaloTag (Arabidopsis Biological Resource Center, stock #CD3-1742). Protein complexes were then expressed *in vitro* using the TNT SP6 Coupled Wheat Germ Extract System (Promega, Madison, WI, USA, Cat #L4130) and purified using Magne HaloTag Beads (Cat #G7282). To prepare DAP-seq samples, hypocotyls were inoculated with a mycelial slurry from a *P. sojae* Race 1 isolate or sterile media as described above. Tissues were collected at 24 hpi, and genomic DNA was isolated using the Zymo Research *Quick*-DNA Plant/Seed Miniprep kit (Cat #D6020, Irvine, CA, USA) with the addition of 2-mercaptoethanol. Three replicates per treatment were pooled to minimize biological variation, and ~5 µg DNA per pool was fragmented and ligated with modified Illumina adapters. Adapter-ligated fragments were then incubated with an immobilized HALO-tagged TF protein. For AmpDAP-seq, fragments were PCR-amplified prior to incubation with a TF complex, permitting binding in the absence of secondary modifications. In both instances, bound DNA was eluted and indexed using unique barcoded primers during PCR enrichment. Indexed libraries were quantified with a Qubit fluorometer paired with a dsDNA high-sensitivity assay kit (Cat #Q32851). Following quantification, libraries were sequenced on an Illumina platform with 150-bp paired-end reads at the University of Arkansas for Medical Sciences (Little Rock, AR, USA).

Short reads were aligned to the soybean reference genome Williams 82 Assembly 4 Annotation 1 (Gmax_508_v4.0.softmasked) using Burrows-Wheeler Alignment tool (v0.7.17-r1188) [119]

and duplicated reads were removed using sambamba (v0.6.8) [120]. DAP-seq peaks were called using the Model-based Analysis of ChIP-Seq peak caller (v2.2.7.1) [121] with empty vector samples as the background control. Bioconductor packages ChIPQC (v1.8.2) [122] and ChIPseeker (v1.32.0) [123] were used to assess the DAP-seq data quality and peak annotation, respectively. The peak annotation was performed using Williams 82 Assembly 4 Annotation 1 (Gmax_508_v4.0.softmasked) genome annotation.

Data preprocessing for CRNNs

To predict TFBS, CRNNs were trained with the aforementioned GmWRKY DAP-seq data. We combined the peak regions from both samples using pandas (v1.4.0) [124] and pybedtools (v0.9.0) [125] to obtain a non-redundant set of peak regions. Similarly, we obtained DAP-seq data for GmRAV from Wang et al. [61] and performed peak calling as mentioned above. The two biological replicates from the latter study were pooled together to get a non-redundant list of peaks. Sequences corresponding to 201-bp peak summit regions were obtained using the soybean reference genome with soft masked (Gmax_508_v4.0.softmasked obtained from Phytozome) and the same genome assembly version used for all other processing and annotations. A negative data set was created from the shuffle tool from BEDTools (v2.30.0) [126], excluding positive binding site regions equal to an exact number of positive binding sites with an identical bp length. The masked regions and low-intensity sequences were removed from the FASTA files. **Fig S2** illustrates the selection of 201-bp bound and unbound sites. The bound sites were given the label “1” (Positive), and unbound sites were labeled “0” (Negative). Both positive and negative data sets were combined to obtain the complete data set for each TF. Next, the sequences were converted into one-hot encoded binary information. In the one-hot encoding process, we consider the DNA sequence as a one-dimensional sequence represented by 4 binary

channels. The encoding was conducted as follows: A = (1 0 0 0), C = (0 1 0 0), G = (0 0 10), and T= (0 0 0 1). The input for each TF was (*n*, 201, 4) three-dimensional array where *n* is the number of sequences for each data set.

For Arabidopsis, Song et al. [67] reanalyzed the DAP-seq data from O'Malley et al. [66], and we obtained the 32-bp peak summit regions from the reanalyzed dataset. Based on Akagi et al. [48], 32-bp summit peak regions perform the best for Arabidopsis DNN models. Therefore, similar to soybean, 32-bp sequences corresponding to positive and negative sets were obtained from the Arabidopsis genome sequences (genome version: TAIR 10; downloaded from The Arabidopsis Information Resource; TAIR) [127].

Model architecture, training, and testing

We used CRNN architecture for both soybean and Arabidopsis models. The input to the CRNN network was the one-hot encoded 201-bp window of DNA sequence, which was passed through a convolutional layer with 256 20-bp filters with a Rectified Linear Unit (ReLU) activation. The next layer was a convolutional layer with 64 8-bp filters with ReLU activation. This was followed by a 128-node time-distributed layer with ReLU activation. Next was a bi-directional long short-term memory network with 64 internal nodes, followed by a 50% Dropout layer. The final layer was a single sigmoid-activated neuron. See **Fig 4** for full architecture.

For the model training, data sets were split into 70% training, 15% validation, and 15% testing. The training was conducted with Keras v2.8 [128] with backend TensorFlow (v2.8) [129] using the Adam optimizer with a 0.001 learning rate. The training ran for 100 epochs. Early stopping with a patience of 20 was used to prevent overfitting. The models were trained with a batch size of 256.

Testing the capacity of models to generalize across TF families

We wanted to test the ability of a model trained using binding site data for one TF to predict binding sites across the TF family. First, we obtained AmpDAP-seq data for GmWRKY2 and obtained 201-bp peak summit regions and their corresponding DNA sequences as described above. Next, we used these as input for the model trained using GmWRKY30. Model accuracy was calculated based on the number of correct predictions/number of bound sites X 100.

Given the limited amount of soybean binding site data, we used Arabidopsis DAP-seq data to further verify model generalization capacity. We trained models using Arabidopsis DAP-seq data for AtWRKY30 (*AT5G24110*), AtMYB62 (*AT1G68320*), AtMYB108 (*AT5G58850*), AtMYB119 (*AT3G06490*), AtNAC031 (*AT1G76420*), AtNAC053 (*AT3G10500*), and AtNAC057 (*AT3G17730*) [66,67] using the methods described above. Similar to soybean, 201-bp peak summit regions for 17 WRKY family members, 15 NAC family members, and five MYB family members were used as input for each model. Model predictions were recorded for each family member and model accuracy was calculated as described above.

Cross-species predictions using the Arabidopsis DAP-seq data

We built models for AtWRKY, AtMYB, AtERF, AtNAC, AtBHLH, and AtC2H2 TF families using Arabidopsis DAP-seq peaks. Akagai et al. [48] used 32-bp peak summit regions to train models with combined DAP-seq data to conduct cross-species prediction. Therefore, we adopted a similar approach and created a combined data set for each TF family to train models using 32-bp peak summit regions (with the exception of AtRAV, for which we obtained data from the ReMap database [130]). In total, we combined DAP-seq data from 24 WRKY, 19 MYB, 17 NAC, 13 ERF, 4 C2H2, and 3 bHLH to create family-level datasets. Further, we trained models using 201- and 32-bp peak regions. Both model types showed high accuracies and low false

positive rates. However, since Akagai et al. [48] used 32-bp peak summit regions to train models to conduct cross-species predictions with success, we opted for this window size for our cross-species prediction. Nevertheless, our model architecture remained the same as the 201-bp models that were used for soybean DAP-seq data, while the convolution layer filter lengths were different (in the first CONV layer, filter length [Kernel] was 20 for soybean models and ten Arabidopsis models).

To test the ability of Arabidopsis models to successfully predict soybean binding targets, we used AtWRKY- and AtMYB-trained models and 201-bp summit peak sequences obtained from soybean DAP-seq (GmWRKY30) or AmpDAP data (GmMYB2) as the input for each model. Model prediction was compared with true TFBS to get cross-species prediction accuracies.

Predicting new gene targets

For the target prediction, we selected 1,000 bp regions on both sides of the TSS (**Fig 4**), as it has been shown that binding sites can reside on either side of the TSS [67]. Next, we obtained predictions on *n* bp (201- or 32-bp) sliding windows throughout the 2-kb sequence. Then, based on widow predictions, we determined if the gene is a potential target gene for a specific TF. If at least one window had a positive prediction, we considered that a potential target.

When obtaining predictions on the target gene, the biggest challenge is finding the optimal stride length for the sliding window. For soybean models, we tested three stride lengths: 10, 100, and 150 bp. To evaluate the best stride, we obtained predictions on 201-bp sliding windows throughout the 2-kb sequence using each stride for genes with GmWRKY30 and GmRAV binding sites on the promoter regions. We then calculated the model prediction with true bindings to get the best stride (**Fig 4**). In both models, the 10-bp stride has the highest true

positive counts. Therefore, we selected the 10 bp as the optimal stride with a 201-bp window region. Leveraging the optimal stride length, we predicted binding sites using GmWRKY30 and GmRAV models for all 6,042 DEGs. If at least one window had a positive prediction, we considered the gene a potential target of the TF. Similarly, we used AtMYB, AtNAC, AtERF, and AtbHLH models to predict binding sites for their respective families in soybean.

The window regions that were predicted to contain TF binding sites were scanned using FIMO software [68] from the MEME suite (v5.0.5) [131] using the default parameters. To retrieve matches, we obtained MEME core plants position frequency matrix files from the JASPAR database [69], and any motif with a *q*-value (Benjamini–Hochberg corrected *p*-value) <0.01 was used for the overlap analyses. We used pybedtools (v0.9.0) (with options: a_and_b = a.intersect(b,wb=True)) [125] to get the overlaps between FIMO-predicted and CRNN-predicted binding sites. Then, we annotated the overlapped binding sites using their Arabidopsis trans-acting factor information obtained from the JASPAR database. Global and family-level GRNs were visualized using Cytoscape (v3.9.1) [132].

Motif enrichment analysis

Sequences corresponding to TFBS were used as input for Simple Enrichment Analysis (v5.4.1) [70] to find statistically enriched motifs with a randomly shuffled input sequence used as the background. Soybean homologs for the Arabidopsis TFs were derived from the gene families of the PANTHER classification system [133].

TF co-occurrence

Motif co-occurrence analysis was performed with Transcription Factor Co-Occurrence using Market Basket analysis (TF-COMB) Python module [73] with default parameters (except

for the count_within() function where default options were changed to max_dist=50, binarize=True, max_overlap=1).

Accession numbers

Sequence data generated in this study will be submitted to the NCBI Sequence Read Archive upon manuscript acceptance. In addition, any file may be obtained from the corresponding author.

Author contributions

BH, AF, and CS conducted wet lab experiments. SR, RW, and AW performed DL. BH, SR, and AW contributed to the data analysis. AR and AW designed the experiments and acquired funding. All authors contributed to the writing and editing of the manuscript and approved the final version prior to publication.

Acknowledgments

The authors acknowledge and are appreciative of the intellectual contributions of Shyaron Poudel, Jeff Gaither, and the Arkansas High-Performance Computing Center.

Supporting information

All code/additional information needed for the full reproduction of this study will be made available on GitHub upon acceptance of this manuscript. In addition, all files may be obtained from the corresponding author.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Tyler BM. Phytophthora sojae: root rot pathogen of soybean and model oomycete. Mol Plant Path. 2007;8(1):1–8. doi: 10.1111/j.1364-3703.2006.00373.x

2. Schmitthenner AF. Problems and Progress in Control of Phytophthora Root Rot of Soybean. Plant Dis. 1985; 69(4):362. doi: 10.1094/PD-69-362

3. Dorrance AE. Management of Phytophthora sojae of soybean: a review and future perspectives. Can J Plant Path. 2018;40(2):210–9. doi: 10.1080/07060661.2018.1445127

4. Sahoo DK, Abeysekara NS, Cianzio SR, Robertson AE, Bhattacharyya MK. A Novel Phytophthora sojae Resistance Rps12 Gene Mapped to a Genomic Region That Contains Several Rps Genes. PLOS ONE. 2017;12(1):e0169950. doi: 10.1371/journal.pone.0169950

5. Kou Y, Wang S. Broad-spectrum and durability: understanding of quantitative disease resistance. Curr Opin Plant Biol. 2010;13(2):181–5. doi: 10.1016/j.pbi.2009.12.010

6. Sugimoto T, Kato M, Yoshida S, Matsumoto I, Kobayashi T, Kaga A, et al. Pathogenic diversity of Phytophthora sojae and breeding strategies to develop Phytophthora-resistant soybeans. Breed Sci. 2012;61(5):511–22. doi: 10.1270/jsbbs.61.511

7. Zhong C, Sun S, Yao L, Ding J, Duan C, Zhu Z. Fine Mapping and Identification of a Novel Phytophthora Root Rot Resistance Locus RpsZS18 on Chromosome 2 in Soybean. Front Plant Sci. 2018;9. Available from: https://doi.org/10.3389/fpls.2018.00044

8. Dorrance AE, Jia H, Abney TS. Evaluation of Soybean Differentials for Their Interaction with Phytophthora sojae. Plant Health Prog. 2004;5(1):9. doi: 10.1094/PHP-2004-0309-01-RS

9. Dodds PN, Rathjen JP. Plant immunity: towards an integrated view of plant–pathogen interactions. Nat Rev Genet. 201;11(8):539–48. doi: 10.1038/nrg2812

10. Ngou BPM, Ahn HK, Ding P, Jones JDG. Mutual potentiation of plant immunity by cell-surface and intracellular receptors. Nature. 2021;592(7852):110–5. doi: 10.1038/s41586-021-03315-7

11. Jones JDG, Dangl JL. The plant immune system. Nature. 2006;444(7117):323–9. doi: 10.1038/nature05286

12. Thomma BPHJ, Nürnberger T, Joosten MHAJ. Of PAMPs and Effectors: The Blurred PTI-ETI Dichotomy. Plant Cell. 2011;23(1):4–15. doi: 10.1105/tpc.110.082602

13. Naveed ZA, Wei X, Chen J, Mubeen H, Ali GS. The PTI to ETI Continuum in Phytophthora-Plant Interactions. Front Plant Sci. 2020. Available from: https://doi.org/10.3389/fpls.2020.593905

14. Wang Y, Tyler BM, Wang Y. Defense and Counterdefense During Plant-Pathogenic Oomycete Infection. Ann Rev Microbiol. 2019;73(1):667–96. doi: 10.1146/annurev-micro-020518-120022

15. Lu Y, Tsuda K. Intimate Association of PRR- and NLR-Mediated Signaling in Plant Immunity. MPMI. 2021;34(1):3–14. doi: 10.1094/MPMI-08-20-0239-IA

16. Moore JW, Loake GJ, Spoel SH. Transcription Dynamics in Plant Immunity. Plant Cell. 2011 Aug;23(8):2809–20. doi: 10.1105/tpc.111.087346

17. Tsuda K, Somssich IE. Transcriptional networks in plant immunity. New Phytol. 2015;206(3):932–47. doi: 10.1111/nph.13286

18. Bai S, Liu J, Chang C, Zhang L, Maekawa T, Wang Q, et al. Structure-Function Analysis of Barley NLR Immune Receptor MLA10 Reveals Its Cell Compartment Specific Activity in Cell Death and Disease Resistance. PLOS Path. 2012;8(6):e1002752. doi: 10.1371/journal.ppat.1002752

19. Bhattacharjee S, Garner C, Gassmann W. New clues in the nucleus: transcriptional reprogramming in effector-triggered immunity. Front Plant Sci. 2013;4. Available from: https://doi.org/10.3389/fpls.2013.00364

20. Meng X, Zhang S. MAPK Cascades in Plant Disease Resistance Signaling. Annual Review of Phytopathology. 2013;51(1):245–66. doi: 10.1146/annurev-phyto-082712-102314

21. Delplace F, Huard-Chauveau C, Berthomé R, Roby D. Network organization of the plant immune system: from pathogen perception to robust defense induction. Plant J. 2022;109(2):447–70. doi: 10.1111/tpj.15462

22. Yang CQ, Fang X, Wu XM, Mao YB, Wang LJ, Chen XY. Transcriptional Regulation of Plant Secondary Metabolism. J Integr Plant Biol. 2012;54(10):703–12. doi: 10.1111/j.1744-7909.2012.01161.x

23. Ng DWK, Abeysinghe JK, Kamali M. Regulating the Regulators: The Control of Transcription Factors in Plant Defense Signaling. Int J Mol Sci. 2018;19(12):3737. doi: 10.3390/ijms19123737

24. Tyler BM, Jiang RHY, Zhou L, Tripathy S, Dou D, Torto-Alalibo T, et al. Functional Genomics and Bioinformatics of the Phytophthora sojae Soybean Interaction. In: Gustafson JP, Taylor J, Stacey G, editors. Genomics of Disease. New York, NY: Springer; 2008. p. 67–78. (Stadler Genetics Symposia Series). doi: 10.1007/978-0-387-76723-9_6

25. Zhou L, Mideros SX, Bao L, Hanlon R, Arredondo FD, Tripathy S, et al. Infection and genotype remodel the entire soybean transcriptome. BMC Genom. 2009;10(1):49. doi: 10.1186/1471-2164-10-49

26. Wang H, Waller L, Tripathy S, St. Martin SK, Zhou L, Krampis K, et al. Analysis of Genes Underlying Soybean Quantitative Trait Loci Conferring Partial Resistance to Phytophthora sojae. Plant Genome. 2010;3(1). doi: 10.3835/plantgenome2009.12.0029

27. Lin F, Zhao M, Baumann DD, Ping J, Sun L, Liu Y, et al. Molecular response to the pathogen Phytophthora sojae among ten soybean near isogenic lines revealed by comparative transcriptomics. BMC Genom. 2014;15(1):18. doi: 10.1186/1471-2164-15-18

28. Dong L, Cheng Y, Wu J, Cheng Q, Li W, Fan S, et al. Overexpression of GmERF5, a new member of the soybean EAR motif-containing ERF transcription factor, enhances resistance to Phytophthora sojae in soybean. J Exp Bot. 2015;66(9):2635–47. doi: 10.1093/jxb/erv078

29. Fan S, Dong L, Han D, Zhang F, Wu J, Jiang L, et al. GmWRKY31 and GmHDL56 Enhances Resistance to Phytophthora sojae by Regulating Defense-Related Gene Expression in Soybean. Front Plant Sci. 2017;8. Available from: https://doi.org/10.3389/fpls.2017.00781

30. Zhao Y, Chang X, Qi D, Dong L, Wang G, Fan S, et al. A Novel Soybean ERF Transcription Factor, GmERF113, Increases Resistance to Phytophthora sojae Infection in Soybean. Front Plant Sci. 2017;8. Available from: https://doi.org/10.3389/fpls.2017.00299

31. Cheng Q, Dong L, Gao T, Liu T, Li N, Wang L, et al. The bHLH transcription factor GmPIB1 facilitates resistance to Phytophthora sojae in Glycine max. J Exp Bot. 2018;69(10):2527–41. doi: 10.1093/jxb/ery103

32. Cui X, Yan Q, Gan S, Xue D, Wang H, Xing H, et al. GmWRKY40, a member of the WRKY transcription factor genes identified from Glycine max L., enhanced the resistance to Phytophthora sojae. BMC Plant Biol. 2019;19(1):598. doi: 10.1186/s12870-019-2132-0

33. Jahan MA, Harris B, Lowery M, Coburn K, Infante AM, Percifield RJ, et al. The NAC family transcription factor GmNAC42–1 regulates biosynthesis of the anticancer and neuroprotective glyceollins in soybean. BMC Genom. 2019;20(1):149. doi: 10.1186/s12864-019-5524-5

34. Jahan MA, Harris B, Lowery M, Infante AM, Percifield RJ, Kovinich N. Glyceollin Transcription Factor GmMYB29A2 Regulates Soybean Resistance to Phytophthora sojae. Plant Physiol. 2020;183(2):530–46. doi: 10.1104/pp.19.01293

35. Liu T, Wang H, Liu Z, Pang Z, Zhang C, Zhao M, et al. The 26S Proteasome Regulatory Subunit GmPSMD Promotes Resistance to Phytophthora sojae in Soybean. Front Plant Sci. 2021;12. Available from: https://doi.org/10.3389/fpls.2021.513388

36. Yu G, Zou J, Wang J, Zhu R, Qi Z, Jiang H, et al. A soybean NAC homolog contributes to resistance to Phytophthora sojae mediated by dirigent proteins. Crop J. 2022;10(2):332–41. doi: 10.1016/j.cj.2021.08.009

37. Gao H, Jiang L, Du B, Ning B, Ding X, Zhang C, et al. GmMKK4-activated GmMPK6 stimulates GmERF113 to trigger resistance to Phytophthora sojae in soybean. Plant J. 2022;111(2):473–95. doi: 10.1111/tpj.15809

38. Ko DK, Brandizzi F. Network-based approaches for understanding gene regulation and function in plants. Plant J. 2020;104(2):302–17. doi: 10.1111/tpj.14940

39. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 2012;13(9):613–26. doi: 10.1038/nrg3207

40. Krouk G, Lingeman J, Colon AM, Coruzzi G, Shasha D. Gene regulatory networks in plants: learning causality from time and perturbation. Genome Biol. 2013;14(6):123. doi: 10.1186/gb-2013-14-6-123

41. Springer N, de León N, Grotewold E. Challenges of Translating Gene Regulatory Information into Agronomic Improvements. Trends Plant Sci. 2019;24(12):1075–82. doi: 10.1016/j.tplants.2019.07.004

42. Bartlett A, O'Malley RC, Huang S shan C, Galli M, Nery JR, Gallavotti A, et al. Mapping genome-wide transcription-factor binding sites using DAP-seq. Nat Protoc. 2017;12(8):1659–72. doi: 10.1038/nprot.2017.055

43. Windram O, Penfold CA, Denby KJ. Network Modeling to Understand Plant Immunity. Ann Rev Phytopath. 2014;52(1):93–111. doi: 10.1146/annurev-phyto-102313-050103

44. Stormo GD, Schneider TD, Gold L, Ehrenfeucht A. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. Nucleic Acids Res. 1982;10(9):2997–3011. doi: 10.1093/nar/10.9.2997

45. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33(8):831–8. doi: 10.1038/nbt.3300

46. Chen C, Hou J, Shi X, Yang H, Birchler JA, Cheng J. DeepGRN: prediction of transcription factor binding site across cell-types using attention-based deep neural networks. BMC Bioinformatics. 2021;22(1):38. doi: 10.1186/s12859-020-03952-1

47. Cochran K, Srivastava D, Shrikumar A, Balsubramani A, Hardison RC, Kundaje A, et al. Domain-adaptive neural networks improve cross-species prediction of transcription factor binding. Genome Res. 2022;32(3):512–23. doi: 10.1101/gr.275394.121

48. Akagi T, Masuda K, Kuwada E, Takeshita K, Kawakatsu T, Ariizumi T, et al. Genome-wide cis-decoding for expression design in tomato using cistrome data and explainable deep learning. Plant Cell. 2022;34(6):2174–87. doi: 10.1093/plcell/koac079

49. Bang S, Galli M, Crisp PA, Gallavotti A, Schmitz RJ. Identifying transcription factor-DNA interactions using machine learning. in silico Plants. 2022;diac014. doi: 10.1093/insilicoplants/diac014

50. Dorrance AE, Berry SA, Anderson TR, Meharg C. Isolation, Storage, Pathotype Characterization, and Evaluation of Resistance for Phytophthora sojae in Soybean. Plant Health Prog. 2008;9(1):35. doi: 10.1094/PHP-2008-0118-01-DG

51. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genom Bioinform. 2020;2(3):lqaa078. doi: 10.1093/nargab/lqaa078

52. Bansal R, Mittapelly P, Cassone BJ, Mamidala P, Redinbaugh MG, Michel A. Recommended Reference Genes for Quantitative PCR Analysis in Soybean Have Variable Stabilities during Diverse Biotic Stresses. PLOS ONE. 2015;10(8):e0134890. doi: 10.1371/journal.pone.0134890

53. Chai C, Lin Y, Shen D, Wu Y, Li H, Dou D. Identification and Functional Characterization of the Soybean GmaPPO12 Promoter Conferring Phytophthora sojae Induced Expression. PLOS ONE. 2013;8(6):e67670. doi: 10.1371/journal.pone.0067670

54. Tian F, Yang DC, Meng YQ, Jin J, Gao G. PlantRegMap: charting functional regulatory maps in plants. Nucleic Acids Res. 2020;48(D1):D1104–13. doi: 10.1093/nar/gkz1020

55. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25–9. doi: 10.1038/75556

56. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000;28(1):27–30. doi: 10.1093/nar/28.1.27

57. Rojas C, Senthil-Kumar M, Tzin V, Mysore K. Regulation of primary plant metabolism during plant-pathogen interactions and its contribution to plant defense. Front Plant Sci. 2014;5. Available from: https://doi.org/10.3389/fpls.2014.00017

58. Peng X, Hu Y, Tang X, Zhou P, Deng X, Wang H, et al. Constitutive expression of rice WRKY30 gene increases the endogenous jasmonic acid accumulation, PR gene expression and resistance to fungal pathogens in rice. Planta. 2012;236(5):1485–98. doi: 10.1007/s00425-012-1698-7

59. Zou L, Yang F, Ma Y, Wu Q, Yi K, Zhang D. Transcription factor WRKY30 mediates resistance to Cucumber mosaic virus in Arabidopsis. Biochem Biophys Res Comm. 2019;517(1):118–24. doi: 10.1016/j.bbrc.2019.07.030

60. Machanick P, Bailey TL. MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics. 2011;27(12):1696–7. doi: 10.1093/bioinformatics/btr189

61. Wang Y, Xu C, Sun J, Dong L, Li M, Liu Y, et al. GmRAV confers ecological adaptation through photoperiod control of flowering time and maturity in soybean. Plant Physiol. 2021;187(1):361–77. doi: 10.1093/plphys/kiab255

62. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, et al. DNA-Binding Specificities of Human Transcription Factors. Cell. 2013;152(1):327–39. doi: 10.1016/j.cell.2012.12.009

63. Kribelbauer JF, Rastogi C, Bussemaker HJ, Mann RS. Low-Affinity Binding Sites and the Transcription Factor Specificity Paradox in Eukaryotes. Annu Rev Cell Dev Biol. 2019;35:357–79. doi: 10.1146/annurev-cellbio-100617-062719

64. Kelley DR. Cross-species regulatory sequence activity prediction. PLOS Comp Biol. 2020;16(7):e1008050. doi: 10.1371/journal.pcbi.1008050

65. Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res. 2016;44(11):e107. doi: 10.1093/nar/gkw226

66. O'Malley RC, Huang S shan C, Song L, Lewsey MG, Bartlett A, Nery JR, et al. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell. 2016;165(5):1280–92. doi: 10.1016/j.cell.2016.04.038

67. Song Q, Lee J, Akter S, Rogers M, Grene R, Li S. Prediction of condition-specific regulatory genes using machine learning. Nucleic Acids Res. 2020;48(11):e62. doi: 10.1093/nar/gkaa264

68. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27(7):1017–8. doi: 10.1093/bioinformatics/btr064

69. Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Berhanu Lemma R, Turchi L, Blanc-Mathieu R, et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2022;50(D1):D165–73. doi: 10.1093/nar/gkab1113

70. Bailey TL, Grant CE. SEA: Simple Enrichment Analysis of motifs. bioRxiv. 2021. doi: 10.1101/2021.08.23.457422

71. Albert R. Scale-free networks in cell biology. J Cell Sci. 2005;118(21):4947–57. doi: 10.1242/jcs.02714

72. Alvarez JM, Brooks MD, Swift J, Coruzzi GM. Time-Based Systems Biology Approaches to Capture and Model Dynamic Gene Regulatory Networks. Ann Rev Plant Biol. 2021;72(1):105–31. doi: 10.1146/annurev-arplant-081320-090914

73. Bentsen M, Heger V, Schultheis H, Kuenne C, Looso M. TF-COMB–Discovering grammar of transcription factor binding sites. Comp Struct Biotech J. 2022;20:4040-51. doi: 10.1016/j.csbj.2022.07.025

74. Enkerli K, Mims CW, Hahn MG. Ultrastructure of compatible and incompatible interactions of soybean roots infected with the plant pathogenic oomycete Phytophthora sojae. Can J Bot. 1997;75(9):1493–508. doi: 10.1139/b97-864

75. Ward EWB. The interaction of soya beans with Phytophthora megasperma f.sp. glycinea: pathogenicity. Biological control of soil-borne plant pathogens. 1990;311–27.

76. Moy P, Qutob D, Chapman BP, Atkinson I, Gijzen M. Patterns of Gene Expression Upon Infection of Soybean Plants by Phytophthora sojae. MPMI. 2004;17(10):1051–62. doi: 10.1094/MPMI.2004.17.10.1051

77. Rinaldi C, Kohler A, Frey P, Duchaussoy F, Ningre N, Couloux A, et al. Transcript Profiling of Poplar Leaves upon Infection with Compatible and Incompatible Strains of the Foliar Rust Melampsora larici-populina. Plant Physiol. 2007;144(1):347–66. doi: 10.1104/pp.106.094987

78. Mine A, Seyfferth C, Kracher B, Berens ML, Becker D, Tsuda K. The Defense Phytohormone Signaling Network Enables Rapid, High-Amplitude Transcriptional Reprogramming during Effector-Triggered Immunity. Plant Cell. 2018;30(6):1199–219. doi: 10.1105/tpc.17.00970

79. Duan Y, Duan S, Armstrong MR, Xu J, Zheng J, Hu J, et al. Comparative Transcriptome Profiling Reveals Compatible and Incompatible Patterns of Potato Toward Phytophthora infestans. G3-Genes Genom Genet. 2020;10(2):623–34. doi: 10.1534/g3.119.400818

80. Yuan M, Ngou BPM, Ding P, Xin XF. PTI-ETI crosstalk: an integrative view of plant immunity. Curr Opin Plant Biol. 2021;62:102030. doi: 10.1016/j.pbi.2021.102030

81. Denancé N, Sánchez-Vallet A, Goffner D, Molina A. Disease resistance or growth: the role of plant hormones in balancing immune responses and fitness costs. Front Plant Sci. 2013;4. Available from: https://doi.org/10.3389/fpls.2013.00155

82. Kim Y, Tsuda K, Igarashi D, Hillmer RA, Sakakibara H, Myers CL, et al. Mechanisms Underlying Robustness and Tunability in a Plant Immune Signaling Network. Cell Host Microbe. 2014;15(1):84–94. doi: 10.1016/j.chom.2013.12.002

83. Zhang C, Cheng Q, Wang H, Gao H, Fang X, Chen X, et al. GmBTB/POZ promotes the ubiquitination and degradation of LHP1 to regulate the response of soybean to Phytophthora sojae. Commun Biol. 2021;4(1):1–15. doi: 10.1038/s42003-021-01907-7

84. Mao G, Meng X, Liu Y, Zheng Z, Chen Z, Zhang S. Phosphorylation of a WRKY Transcription Factor by Two Pathogen-Responsive MAPKs Drives Phytoalexin Biosynthesis in Arabidopsis. Plant Cell. 2011;23(4):1639–53. doi: 10.1105/tpc.111.084996

85. Zhao L, Luo Q, Yang C, Han Y, Li W. A RAV-like transcription factor controls photosynthesis and senescence in soybean. Planta. 2008;227(6):1389–99. doi: 10.1007/s00425-008-0711-7

86. Zhao L, Hao D, Chen L, Lu Q, Zhang Y, Li Y, et al. Roles for a soybean RAV-like orthologue in shoot regeneration and photoperiodicity inferred from transgenic plants. J Exp Bot. 2012;63(8):3257–70. doi: 10.1093/jxb/ers056

87. Zhao SP, Xu ZS, Zheng WJ, Zhao W, Wang YX, Yu TF, et al. Genome-Wide Analysis of the RAV Family in Soybean and Functional Identification of GmRAV-03 Involvement in Salt and Drought Stresses and Exogenous ABA Treatment. Front Plant Sci. 2017;8. Available from: https://doi.org/10.3389/fpls.2017.00905

88. Quang D, Xie X. FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. Methods. 2019;166:40–7. doi: 10.1016/j.ymeth.2019.03.020

89. Srivastava D, Aydin B, Mazzoni EO, Mahony S. An interpretable bimodal neural network characterizes the sequence and preexisting chromatin predictors of induced transcription factor binding. Genome Biol. 2021;22(1):20. doi: 10.1186/s13059-020-02218-6

90. Swift J, Coruzzi GM. A matter of time — How transient transcription factor interactions create dynamic gene regulatory networks. Biochim Biophys Acta Gene Regul Mech. 2017;1860(1):75–83. doi: 10.1016/j.bbagrm.2016.08.007

91. Brooks MD, Juang CL, Katari MS, Alvarez JM, Pasquino A, Shih HJ, et al. ConnecTF: A platform to integrate transcription factor–gene interactions and validate regulatory networks. Plant Physiol. 2021;185(1):49–66. doi: 10.1093/plphys/kiaa012

92. Nitta KR, Jolma A, Yin Y, Morgunova E, Kivioja T, Akhtar J, et al. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. Ren B, editor. eLife. 2015;4:e04837. doi: 10.7554/eLife.04837

93. Varala K, Marshall-Colón A, Cirrone J, Brooks MD, Pasquino AV, Léran S, et al. Temporal transcriptional logic of dynamic regulatory networks underlying nitrogen signaling and use in plants. Proc. Natl. Acad. Sci. U.S.A. 2018;115(25):6494–9. doi: 10.1073/pnas.1721487115

94. Atencio L, Salazar J, Lauter AN, Gonzales MD, O'Rourke JA, Graham MA. Characterizing short and long term iron stress responses in iron deficiency tolerant and susceptible soybean (Glycine max L. Merr.). Plant Stress. 2021;2:100012. doi: 10.1016/j.stress.2021.100012

95. Liu Y, Xue Y, Xie B, Zhu S, Lu X, Liang C, Tian J. Complex gene regulation between young and old soybean leaves in responses to manganese toxicity. Plant Phys and Biochem. 2020;155:231-42. doi: 10.1016/j.plaphy.2020.07.002

96. Chen LM, Fang YS, Zhang CJ, Hao QN, Cao D, Yuan SL, Chen HF, Yang ZL, Chen SL, Shan ZH, Liu BH. GmSYP24, a putative syntaxin gene, confers osmotic/drought, salt stress tolerances and ABA signal pathway. Sci Rep. 2019;9(1):1-2. doi: 10.1038/s41598-019-42332-5

97. Zhang H, Yang Y, Sun C, Liu X, Lv L, Hu Z, Yu D, Zhang D. Up□regulating GmETO1 improves phosphorus uptake and use efficiency by promoting root growth in soybean. Plant Cell Env. 2020;43(9):2080-94. doi: 10.1111/pce.13816

98. Chang C, Tian L, Ma L, Li W, Nasir F, Li X, Tran LS, Tian C. Differential responses of molecular mechanisms and physiochemical characters in wild and cultivated soybeans against invasion by the pathogenic Fusarium oxysporum Schltdl. Physiol Plant. 2019;166(4):1008-25. doi: 10.1111/ppl.12870

99. McCabe CE, Cianzio SR, O'Rourke JA, Graham MA. Leveraging RNA-Seq to characterize resistance to Brown stem rot and the Rbs3 locus in soybean. Mol Plant-Microbe Interact. 2018;31(10):1083-94. doi: 10.1094/MPMI-01-18-0009-R

100. Liu X, Chu S, Sun C, Xu H, Zhang J, Jiao Y, Zhang D. Genome-wide identification of low phosphorus responsive microRNAs in two soybean genotypes by high-throughput sequencing. Funct Integr Genomics. 2020; 20(6):825-38. doi: 10.1007/s10142-020-00754-9

101. Zhao J, Zheng L, Wei J, Wang Y, Chen J, Zhou Y, Chen M, Wang F, Ma Y, Xu ZS. The soybean PLATZ transcription factor GmPLATZ17 suppresses drought tolerance by interfering with stress-associated gene regulation of GmDREB5. Crop J. 2022. doi: 10.1016/j.cj.2022.03.009

102. Han X, Wang J, Zhang Y, Kong Y, Dong H, Feng X, Li T, Zhou C, Yu J, Xin D, Chen Q. Changes in the m6A RNA methylome accompany the promotion of soybean root growth by rhizobia under cadmium stress. J Haz Mat. 2023;441:129843. doi: 10.1016/j.jhazmat.2022.129843

103. Schrynemackers M, Kueffner R, Geurts P. On protocols and measures for the validation of supervised methods for the inference of biological networks. Front Genet. 2013;4. Available from: https://doi.org/10.3389/fgene.2013.00262

104. Zhou F, Emonet A, Dénervaud Tendon V, Marhavy P, Wu D, Lahaye T, et al. Co-incidence of Damage and Microbial Patterns Controls Localized Immune Responses in Roots. Cell. 2020;180(3):440-453.e18. doi: 10.1016/j.cell.2020.01.013

105. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90. doi: 10.1093/bioinformatics/bty560

106. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907–15. doi: 10.1038/s41587-019-0201-4

107. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923–30. doi: 10.1093/bioinformatics/btt656

108. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. doi: 10.1186/s13059-014-0550-8

109. Rau A, Maugis-Rabusseau C. Transformation and model choice for RNA-seq co-expression analysis. Brief Bioinform. 2018;19(3):425–36. doi: 10.1093/bib/bbw128

110. Godichon-Baggioni A, Maugis-Rabusseau C, Rau A. Clustering transformed compositional data using K-means, with applications in gene expression and bicycle sharing system data. J Appl Stat. 2019;46(1):47–65. doi: 10.1080/02664763.2018.1454894

111. Wimalanathan K, Lawrence-Dill CJ. Gene Ontology Meta Annotator for Plants (GOMAP). Plant Methods. 2021;17(1):54. doi: 10.1186/s13007-021-00754-1

112. Alexa A, Rahnenführer J. Gene set enrichment analysis with topGO. Bioconductor Improv. 2009;27:1-26.

113. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS J Integr Biol. 2012;16(5):284–7. doi: 10.1089/omi.2011.0118

114. RStudio T. RStudio: integrated development for R. Rstudio Team, PBC, Boston, MA URL http://www. rstudio. com. 2020.

115. Wickham H, Chang W, Wickham MH. Package ‘ggplot2’. Create elegant data visualizations using the grammar of graphics. Version. 2016;2(1):1-89.

116. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet j.2011;17(1):10–2. doi: 10.14806/ej.17.1.200

117. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40(D1):D1178–86. doi: 10.1093/nar/gkr944

118. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34(5):525–7. doi: 10.1038/nbt.3519

119. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25(14):1754–60. doi: 10.1093/bioinformatics/btp324

120. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015;31(12):2032–4. doi: 10.1093/bioinformatics/btv098

121. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137. doi: 10.1186/gb-2008-9-9-r137

122. Carroll TS, Liang Z, Salama R, Stark R, de Santiago I. Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. Front Genet. 2014;5. Available from: https://doi.org/10.3389/fgene.2014.00075

123. Yu G, Wang LG, He QY. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics. 2015;31(14):2382–3. doi: 10.1093/bioinformatics/btv145

124. McKinney W. pandas: a Foundational Python Library for Data Analysis and Statistics. 2011;9.

125. Dale RK, Pedersen BS, Quinlan AR. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics. 2011;27(24):3423–4. doi: 10.1093/bioinformatics/btr539

126. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics. 2014;47(1):11.12.1-11.12.34. doi: 10.1002/0471250953.bi1112s47

127. Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, et al. The arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. genesis. 2015;53(8):474–85. doi: 10.1002/dvg.22877

128. Chollet F. keras. 2015.

129. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M. {TensorFlow}: a system for {Large-Scale} machine learning. In12th USENIX symposium on operating systems design and implementation (OSDI 16) 2016 (pp. 265-283).

130. Hammal F, de Langen P, Bergon A, Lopez F, Ballester B. ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. Nucleic Acids Res. 2022;50(D1):D316–25. doi: 10.1093/nar/gkab996

131. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. Nucleic Acids Res. 2015 Jul 1;43(W1):W39–49. doi: 10.1093/nar/gkv416

132. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. 2003;13(11):2498–504. doi: 10.1101/gr.1239303

133. Mi H, Muruganujan A, Huang X, Ebert D, Mills C, Guo X, Thomas PD. Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v. 14.0). Nat Protoc. 2019;14(3):703-21. doi: 10.1038/s41596-019-0128-8

Supporting information captions

Fig S1 Capture-seq validation of RNA-seq data. (**a**) Heatmap of reference gene expression in RNA- and Capture-seq. (**b**) Heatmap of pathogen-induced gene expression.

Fig S2 Selection of 201-bp bound (peak regions) and unbound sites (negative dataset) during model training.

Fig S3 auROC and auPRC curves for soybean data-trained models. (**a**) auROC curve for GmWRKY30 CRNN. (**b**) auROC curve for GmRAV CRNN. (**c**) auPRC curve for GmWRKY30 CRNN. (**d**) auPRC curve for GmRAV CRNN.

Fig S4 AmpDAP-seq data for GmMYB61 and GmWRKY2. (**a**) Heatmap of DAP peak binding within 1,000 bp of the TSS region. (**b**) Distribution of DAP peaks across genomic features.

Fig S5 GRNs for defense-related TF families. (**a**) bHLH GRN. (**b**) ERF GRN. (**c**) MYB GRN. (**d**) NAC GRN. (**e**) RAV GRN. (**f**) WRKY GRN.

Fig S6 Prioritization of target genes. (**a**) Density plots of indegree, cumulative outdegree, sum cumulative cosine, and mean $|\log_2 FC|$ (Mean $|FC|$) for target genes. Red polygons represent the upper quartile for each parameter. Furthermore, 254 targets were in the upper quarter for all four parameters. (**b**) Heatmap depicting the $\log_2 FC$ (FC) of prioritized target genes across both interactions.

Table S1 RNA-seq mapping statistics.

Table S2 Mapping statistics for DAP- and AmpDAP-seq.

Data S1 DEGs from RNA-seq, corresponding co-expression cluster assignments, functional annotations, and target gene assignments.

Data S2 Capture-seq gene list and expression data.

Data S3 GmWRKY30 DAP-seq binding site data.

Data S4 GmRAV DAP-seq binding site data.

Data S5 GmMYB61 and GmWRKY2 AmpDAP-seq binding site data.

Data S6 Arabidopsis generalization prediction accuracies.

Figure and table captions

Fig 1 Schematic overview of the study design. (**a**) Soybean plants harboring *Rps1k* were inoculated with a Race 1 *P. sojae* isolate, Race 25 isolate, or sterile media. Inoculated hypocotyls were used for RNA-seq. Capture-seq was performed subsequently to validate the RNA-seq data. (**b**) Overrepresented TF families were identified from the RNA-seq analysis. DAP-seq data was

generated/obtained for the families most represented by total abundance and percentage of genome-wide proportion. (**c**) DL models were trained using DAP-seq binding site data. The capacity of some models to generalize across a given TF family was performed intra- and interspecifically. For several TF families of interest, soybean- or Arabidopsis-based DNNs were trained and used to predict TFBS. (**d**) DNN predictions were overlapped with FIMO motif scans, and the highly confident targets were used to construct a GRN.

Fig 2 Pathogenicity testing and transcriptome analysis. (**a**) Disease development in Race 25- (top) and Race 1-treated (bottom) hypocotyls at seven days post-infection. (**b**) Venn diagram of DEGs between different treatments. (**c**) TF representation among DEGs from RNA-seq. WRKY was the most represented TF family by total abundance and RAV by the percentage of genome-wide proportion. (**d**) K-means clustering of DEGs. DEGs were assigned to nine co-expression clusters. Of these, seven displayed increased expression ($\log_2$FC [FC] >0) in infected vs Mock treatments, while two demonstrated decreased expression (FC <0). (**e**) Functional enrichment and TF representation for gene co-expression clusters. (left panel) Top five GO categories by *p*-value. (middle panel) Top five KEGG terms by adjusted *p*-value (□≤0.05). (right panel) top 3 TF families (abundance) for each cluster.

Fig 3 DAP-seq identification of GmWRKY30 and GmRAV TFBS. (**a**) Distribution of DAP-seq peaks across genomic features. (**b**) Distance of peaks from the TSS.

Fig 4 Schematic illustration of CRNN architecture and TFBS prediction.

Fig 5 Generalization testing for soybean and Arabidopsis CRNNs and schematic illustration of TFBS prediction for defense-related TF families. (**a**) The GmWRKY30 CRNN was used to predict TFBS interspecifically with AtWRKY30 DAP-seq data (left barplot) and intraspecifically with GmWRKY2 AmpDAP-seq data (right barplot). (**b**) AtWRKY, AtMYB, and AtNAC CRNNs were trained with available DAP-seq data and used to predict binding sites for other members of their respective families. (**c**) The Arabidopsis-based models, along with the GmWRKY30 and GmRAV models, were used to predict TFBS in our DEG set. These predictions were overlaid with FIMO scans to elucidate TF-target gene interactions.

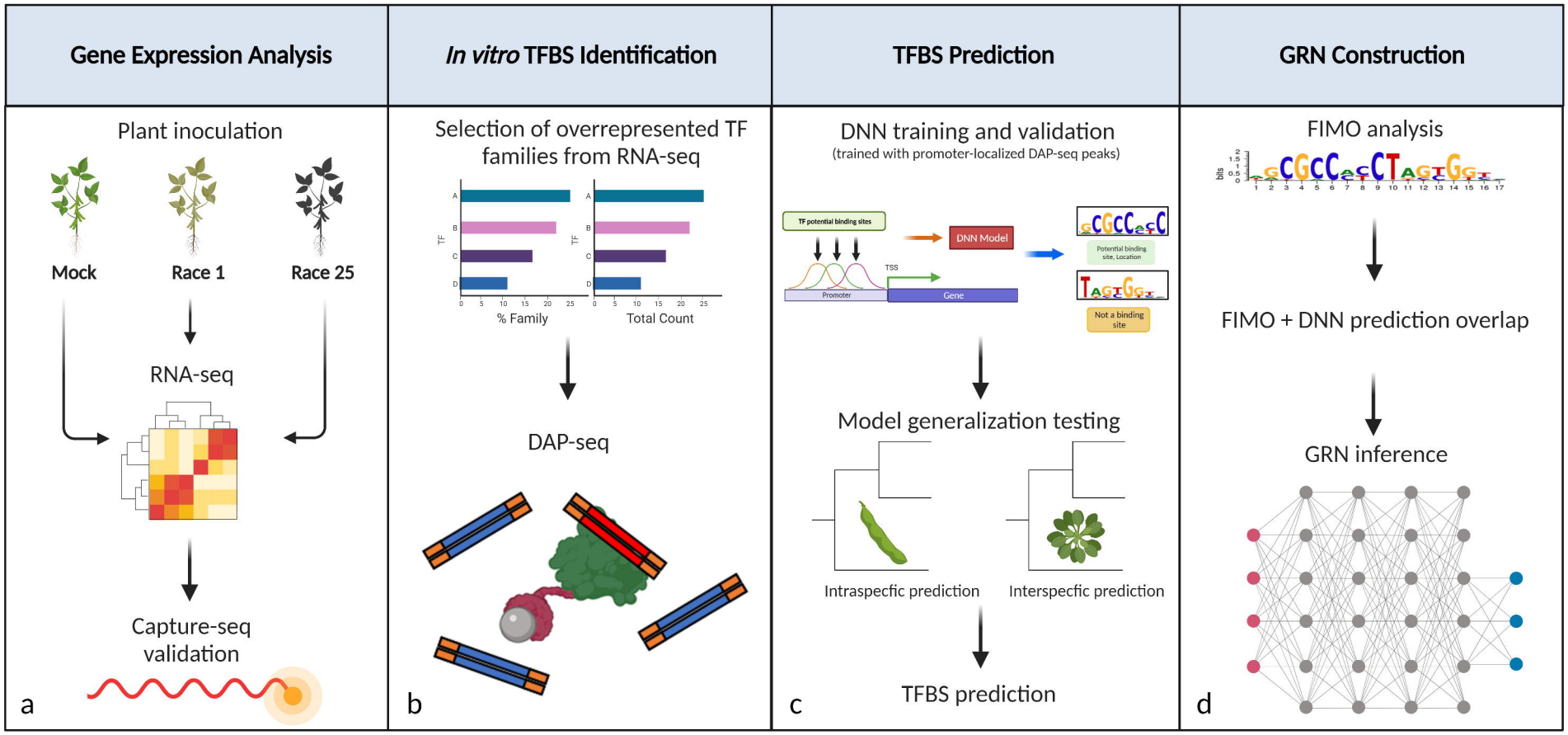
Fig 6 GRN inference at 24 hpi. (**a**) (left) $\log_2$FC (FC) of DEGs across interaction types, (middle) WRKY and RAV binding site representation in the DEG set derived from DAP-seq, and (right) binding site representation for each TF family in the DEG set derived from CRNN + FIMO prediction. The bar plot shows the total number of target genes for each family, as well as the number of TF-encoding target genes (light green). (**b**) Hairball of the global GRN. Nodes and edges represent TFs and target genes, respectively. Node size corresponds to outdegree. (**c**) Scatterplot of the top co-occurring TF pairs by cosine association score identified with TF-COMB. The datapoint color reflects the total number of shared targets for a given TF pair. (**d**) Prioritization of nodes. Nodes that were statistically enriched by Simple Enrichment Analysis and were represented in the transcriptome analysis (n = 118) were prioritized by outdegree, cumulative indegree, cumulative cosine, and mean |log2FC| (Mean |FC|). Blue polygons represent the upper quartile for each parameter. Fourteen genes/13 TFs were in the upper quarter for all four parameters. (**e**) Hairball of the hub nodes. Node size corresponds to outdegree.

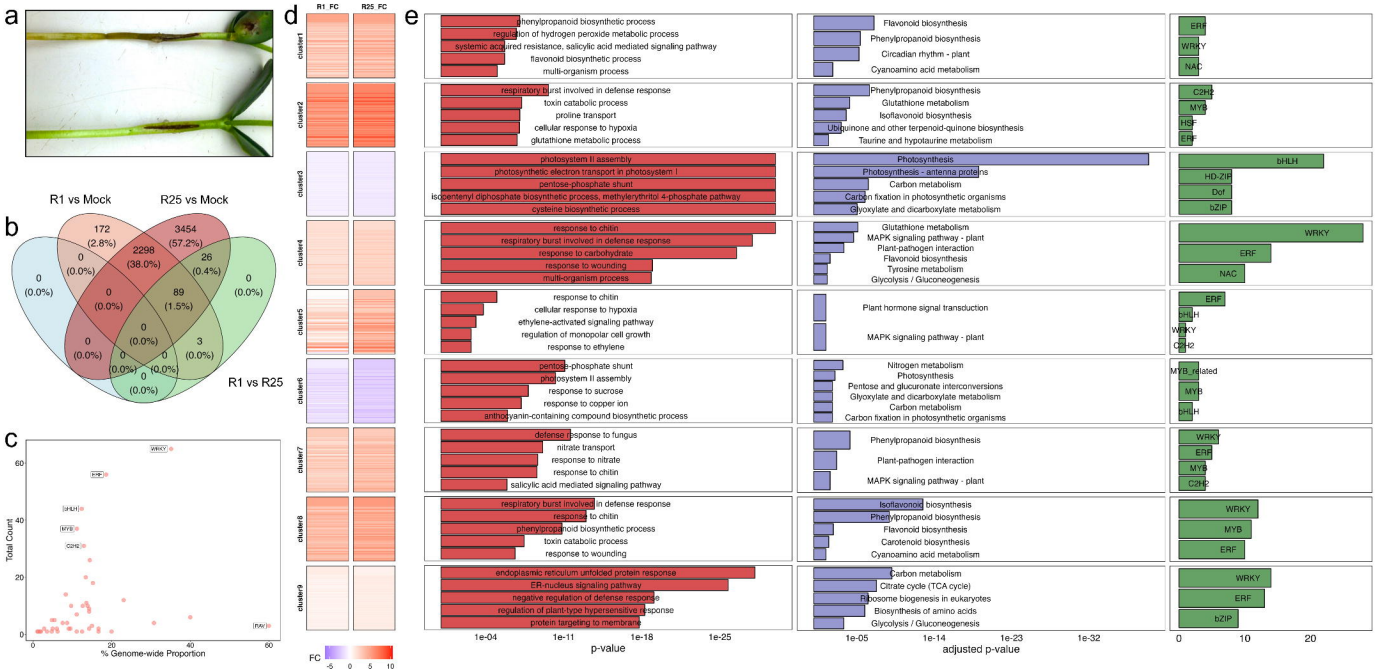
Table 1 Summary of GmWRKY and GmRAV CRNN models. TPR: True Positive Rate; TNR: True Negative Rate; FPR: False Positive Rate; FNR: False Negative Rate

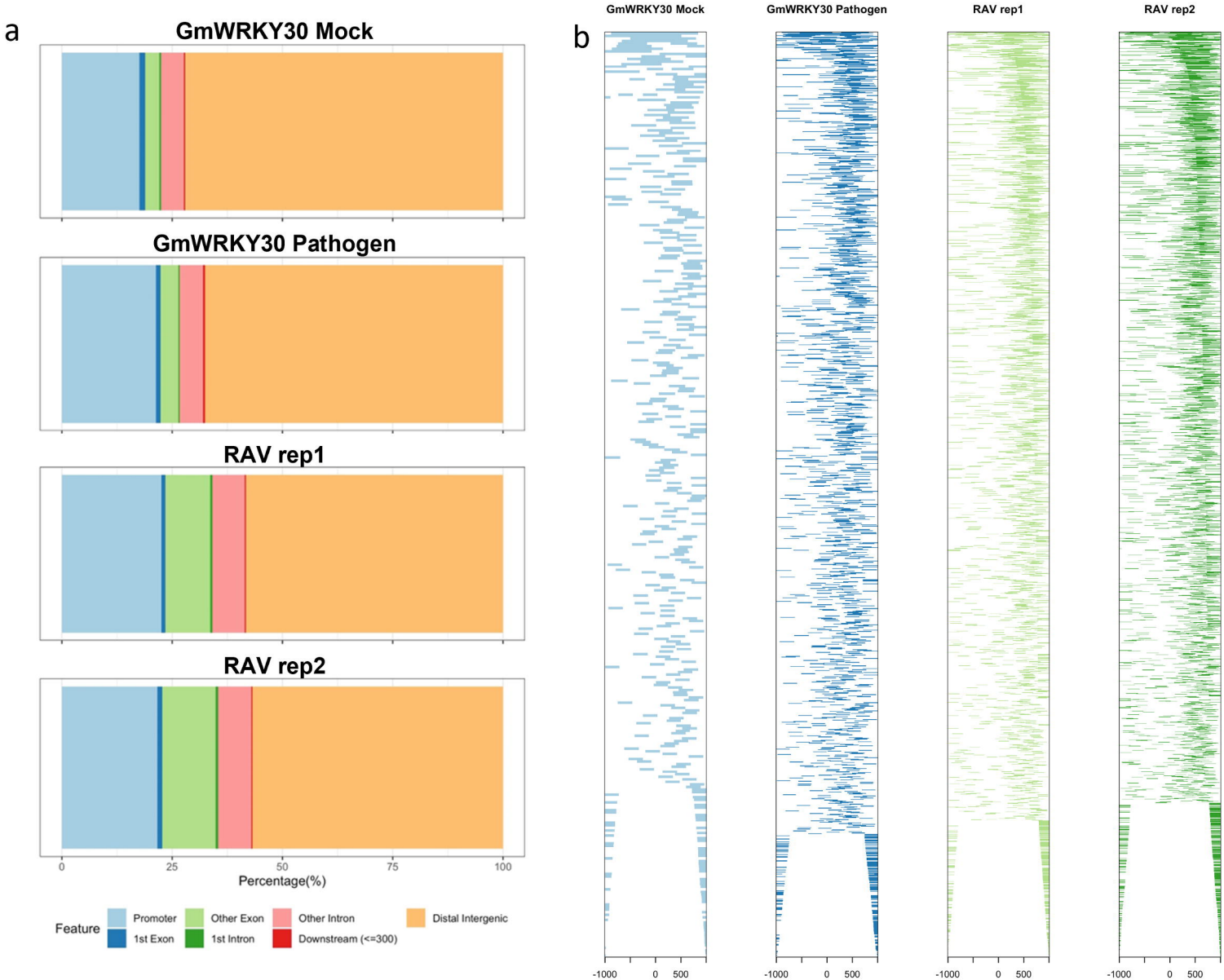
Table 2 Summary of Arabidopsis data-based CRNN models used for generalization tests. TPR: True Positive Rate; TNR: True Negative Rate; FPR: False Positive Rate; FNR: False Negative Rate

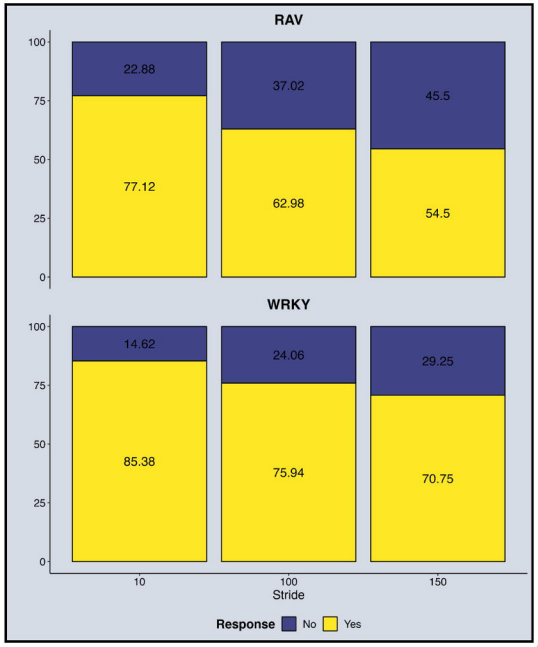
Table 3 Summary of Arabidopsis data-based CRNN models used for cross-species predictions. TPR: True Positive Rate; TNR: True Negative Rate; FPR: False Positive Rate; FNR: False Negative Rate

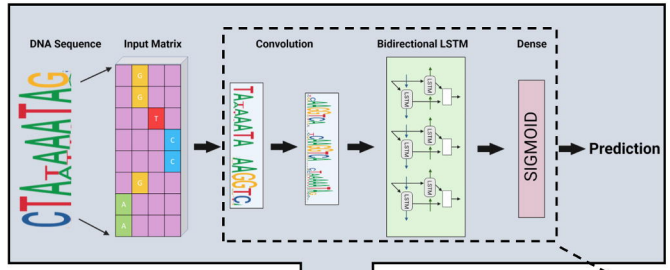
Table 4 Arabidopsis-to-soybean cross-species prediction results. TPR: True Positive Rate; TNR: True Negative Rate; FPR: False Positive Rate; FNR: False Negative Rate

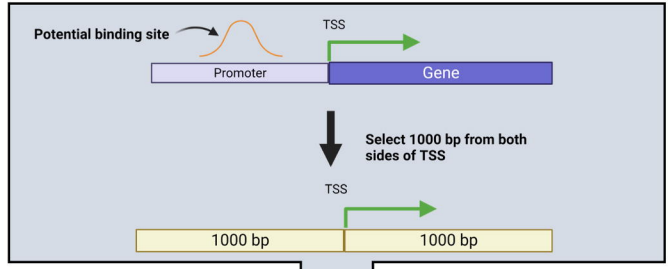

Gene Expression Analysis
Plant inoculation
Mock
Race 1
Race 25
RNA-seq
Capture-seq validation
a
In vitro TFBS Identification
Selection of overrepresented TF families from RNA-seq
TF
% Family
Total Count
DAP-seq
b
TFBS Prediction
DNN training and validation
(trained with promoter-localized DAP-seq peaks)
TF potential binding sites
DNN Model
TSS
Promoter
Gene
Potential binding site, Location
Not a binding site
Model generalization testing
Intraspecfic prediction
Interspecfic prediction
TFBS prediction
c
GRN Construction
FIMO analysis
bits
FIMO + DNN prediction overlap
GRN inference
d

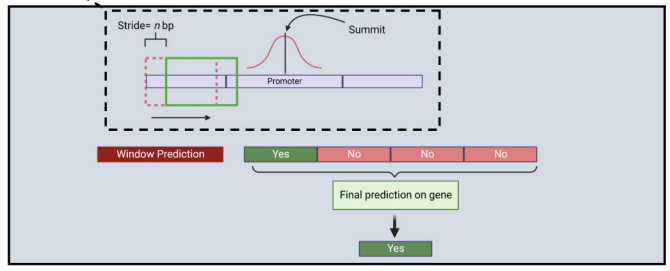

a
b
R1 vs Mock
R25 vs Mock
R1 vs R25
172 (2.8%)
3454 (57.2%)
2298 (38.0%)
0 (0.0%)
26 (0.4%)
0 (0.0%)
0 (0.0%)
89 (1.5%)
0 (0.0%)
0 (0.0%)
0 (0.0%)
3 (0.0%)
0 (0.0%)
0 (0.0%)
0 (0.0%)
c
WRKY
ERF
bHLH
MYB
C2H2
NAC
Total Count
% Genome-wide Proportion
d
R1_FC
R25_FC
FC
e
cluster1
phenylpropanoid biosynthetic process
regulation of hydrogen peroxide metabolic process
systemic acquired resistance, salicylic acid mediated signaling pathway
flavonoid biosynthetic process
multi-organism process
Flavonoid biosynthesis
Phenylpropanoid biosynthesis
Circadian rhythm - plant
Cyanoamino acid metabolism
ERF
WRKY
NAC
cluster2
respiratory burst involved in defense response
toxin catabolic process
proline transport
cellular response to hypoxia
glutathione metabolic process
Phenylpropanoid biosynthesis
Glutathione metabolism
Isoflavonoid biosynthesis
Ubiquinone and other terpenoid-quinone biosynthesis
Taurine and hypotaurine metabolism
C2H2
MYB
HSF
ERF
cluster3
photosystem II assembly
photosynthetic electron transport in photosystem I
pentose-phosphate shunt
isopentenyl diphosphate biosynthetic process, methylerythritol 4-phosphate pathway
cysteine biosynthetic process
Photosynthesis
Photosynthesis - antenna proteins
Carbon metabolism
Carbon fixation in photosynthetic organisms
Glyoxylate and dicarboxylate metabolism
bHLH
HD-ZIP
Dof
bZIP
cluster4
response to chitin
respiratory burst involved in defense response
response to carbohydrate
response to wounding
multi-organism process
Glutathione metabolism
MAPK signaling pathway - plant
Plant-pathogen interaction
Flavonoid biosynthesis
Tyrosine metabolism
Glycolysis / Gluconeogenesis
WRKY
ERF
NAC
cluster5
response to chitin
cellular response to hypoxia
ethylene-activated signaling pathway
regulation of monopolar cell growth
response to ethylene
Plant hormone signal transduction
MAPK signaling pathway - plant
ERF
bHLH
WRKY
C2H2
cluster6
pentose-phosphate shunt
photosystem II assembly
response to sucrose
response to copper ion
anthocyanin-containing compound biosynthetic process
Nitrogen metabolism
Photosynthesis
Pentose and glucuronate interconversions
Glyoxylate and dicarboxylate metabolism
Carbon metabolism
Carbon fixation in photosynthetic organisms
MYB_related
MYB
bHLH
cluster7
defense response to fungus
nitrate transport
response to nitrate
response to chitin
salicylic acid mediated signaling pathway
Phenylpropanoid biosynthesis
Plant-pathogen interaction
MAPK signaling pathway - plant
WRKY
ERF
MYB
C2H2
cluster8
respiratory burst involved in defense response
response to chitin
phenylpropanoid biosynthetic process
toxin catabolic process
response to wounding
Isoflavonoid biosynthesis
Phenylpropanoid biosynthesis
Flavonoid biosynthesis
Carotenoid biosynthesis
Cyanoamino acid metabolism
WRKY
MYB
ERF
cluster9
endoplasmic reticulum unfolded protein response
ER-nucleus signaling pathway
negative regulation of defense response
regulation of plant-type hypersensitive response
protein targeting to membrane
Carbon metabolism
Citrate cycle (TCA cycle)
Ribosome biogenesis in eukaryotes
Biosynthesis of amino acids
Glycolysis / Gluconeogenesis
WRKY
ERF
bZIP
1e-04
1e-11
1e-18
1e-25
p-value
1e-05
1e-14
1e-23
1e-32
adjusted p-value
0
10
20


a
GmWRKY30 Mock
GmWRKY30 Pathogen
RAV rep1
RAV rep2
0
25
50
75
100
Percentage(%)
Feature
Promoter
1st Exon
Other Exon
1st Intron
Other Intron
Downstream (<=300)
Distal Intergenic
b
GmWRKY30 Mock
GmWRKY30 Pathogen
RAV rep1
RAV rep2
-1000
0
500

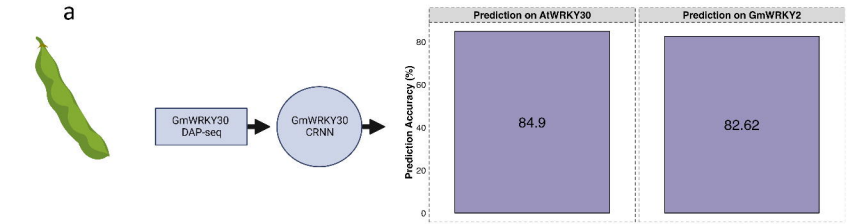


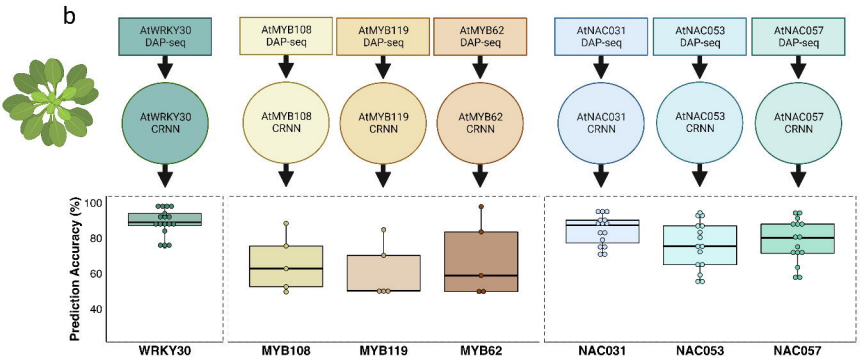


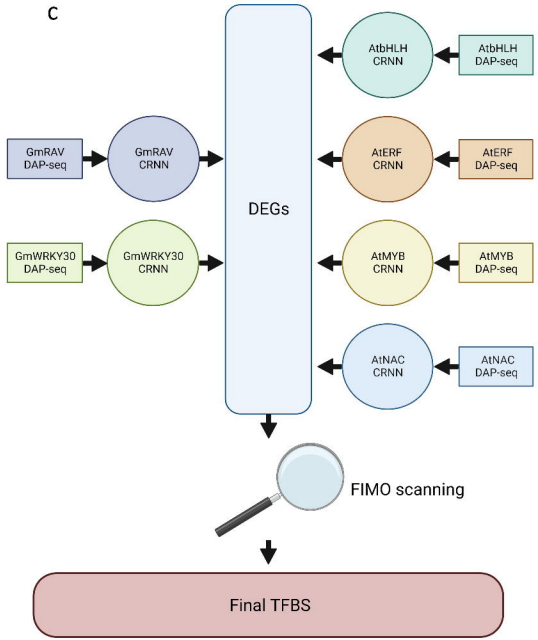


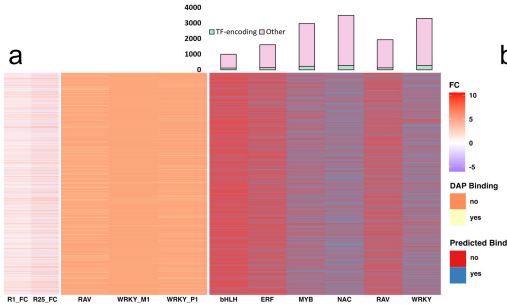


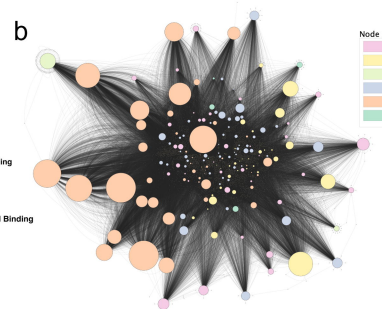
conv1d_input InputLayer	input: output:	[(None, 201, 4)]	[(None, 201, 4)]
conv1d Conv1D	input: output:	(None, 201, 4)	(None, 201, 512)
dropout Dropout	input: output:	(None, 201, 512)	(None, 201, 512)
conv1d_1 Conv1D	input: output:	(None, 201, 512)	(None, 201, 64)
dropout_1 Dropout	input: output:	(None, 201, 64)	(None, 201, 64)
time_distributed(dense) TimeDistributed(Dense)	input: output:	(None, 201, 64)	(None, 201, 128)
bidirectional(lstm) Bidirectional(LSTM)	input: output:	(None, 201, 128)	(None, 201, 128)
dropout_2 Dropout	input: output:	(None, 201, 128)	(None, 201, 128)
flatten Flatten	input: output:	(None, 201, 128)	(None, 25728)
dense_1 Dense	input: output:	(None, 25728)	(None, 1)

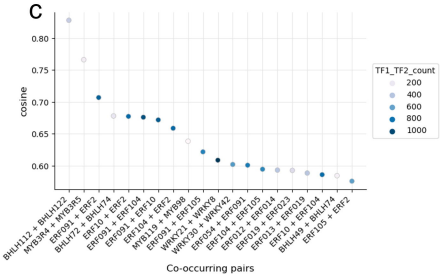

a
GmWRKY30 DAP-seq
GmWRKY30 CRNN
Prediction on AtWRKY30
Prediction on GmWRKY2
Prediction Accuracy (%)
0
20
40
60
80
84.9
82.62

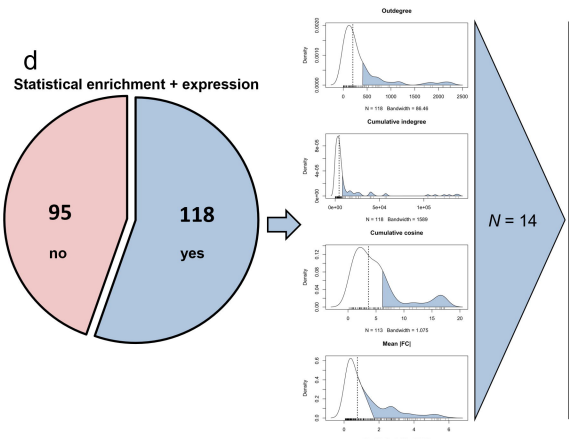
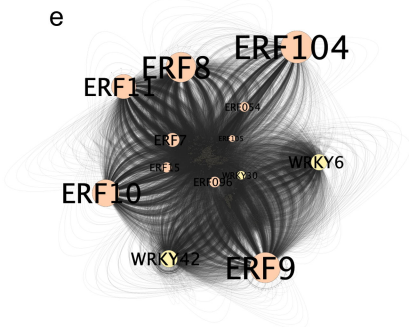

b
AtWRKY30 DAP-seq
AtMYB108 DAP-seq
AtMYB119 DAP-seq
AtMYB62 DAP-seq
AtNAC031 DAP-seq
AtNAC053 DAP-seq
AtNAC057 DAP-seq
AtWRKY30 CRNN
AtMYB108 CRNN
AtMYB119 CRNN
AtMYB62 CRNN
AtNAC031 CRNN
AtNAC053 CRNN
AtNAC057 CRNN
Prediction Accuracy (%)
40
60
80
100
WRKY30
MYB108
MYB119
MYB62
NAC031
NAC053
NAC057


c
GmRAV DAP-seq
GmRAV CRNN
GmWRKY30 DAP-seq
GmWRKY30 CRNN
DEGs
AtbHLH CRNN
AtbHLH DAP-seq
AtERF CRNN
AtERF DAP-seq
AtMYB CRNN
AtMYB DAP-seq
AtNAC CRNN
AtNAC DAP-seq
FIMO scanning
Final TFBS









TF ID	At ID	Gm ID	Abiotic Stress	Biotic Stress	Literature
ERF105	AT5G51190	Glyma.01G206700	—	—	—
WRKY30	AT5G24110	Glyma.06G125600	Fe deficiency, Mn toxicity	*P. sojae*	27,94,95
WRKY30	AT5G24110	Glyma.07G057400	Mn toxicity	—	95
WRKY42	AT4G04450	Glyma.07G262700	Mn toxicity	—	95
WRKY6	AT1G62300				
WRKY42	AT4G04450	Glyma.09G005700	Fe deficiency, Mn toxicity, ABA stress	—	94,95,96
WRKY6	AT1G62300				
ERF15	AT2G31230	Glyma.10G036700	P deficiency	*Fusarium oxysporum*, *Phialophara gregata*	97,98,99
ERF104	AT5G61600	Glyma.10G194200	—	—	—
ERF15	AT2G31230	Glyma.13G123100	P deficiency	*Phialophara gregata*	99,100
ERF054	AT4G28140	Glyma.13G298600	Drought	—	101
ERF10	AT1G03800	Glyma.15G077000	—	—	—
ERF11	AT1G28370				
ERF7	AT3G20310				
ERF8	AT1G53170				
ERF9	AT5G44210				
ERF10	AT1G03800	Glyma.15G077100	Cd toxicity	—	102
ERF11	AT1G28370				
ERF7	AT3G20310				
ERF8	AT1G53170				
ERF9	AT5G44210				
WRKY42	AT4G04450	Glyma.15G110300	ABA stress	—	96
WRKY6	AT1G62300				
WRKY42	AT4G04450	Glyma.17G011400	Fe deficiency, Cd toxicity	—	94,102
WRKY6	AT1G62300				
ERF096	AT5G43410	Glyma.20G070100	P deficiency	—	100