2018-05-28

Genomewide evaluation of cis-elements and cognate transcription factors in Nicotiana attenuata predicts 27 unique transcription factor-binding site pairs

Ashton Omdahl

Honors Thesis

GENOMEWIDE EVALUATION OF CIS-ELEMENTS AND COGNATE TRANSCRIPTION FACTORS IN *NICOTIANA ATTENUATA* PREDICTS *27* UNIQUE TRANSCRIPTION FACTOR-BINDING SITE PAIRS

by Ashton R. Omdahl

Submitted to Brigham Young University in partial fulfillment of graduation requirements for University Honors

Department of Biology

Brigham Young University

June 2018

Advisor: Dr. Stephen Piccolo

Honors Coordinator: Dr. Steven Peck

ABSTRACT

GENOMEWIDE EVALUATION OF CIS-ELEMENTS AND COGNATE TRANSCRIPTION FACTORS IN *NICOTIANA ATTENUATA* PREDICTS *27* UNIQUE TRANSCRIPTION FACTOR-BINDING SITE PAIRS

Ashton R. Omdahl

Department of Biology

Bachelor of Science

Nicotiana attenuata has been widely studied for its ecological plant-herbivore relationships and response to environmental stress. The jasmonate signaling pathway regulated by jasmonate ZIM-domiain (JAZ) repressor proteins that modulate defense response levels has been of particular focus in this research. While our understanding of the genes associated with defense response and their regulation continues to expand, the transcriptional regulation of these genes is largely uncharacterized. In an effort to provide insight into these relationships, we performed genomewide analysis of transcript level data in order to predict transcription factors (TFs), their respective binding sites (TFBS), and the genes they regulate. We identified 27 unique TF-TFBS pairs and 507 genes containing cis-elements associated with these TFs. We also identified gene sets enriched for chloroplast structure and function, ribosomal structure and function, cell membrane components, and ATP binding gene ontology. Our motif enrichment and co-expression analysis results suggest that JAZb may be regulated by TFs MYC2a and MYC2b and that TF WRKY3 may be part of a self-regulation loop.

ACKNOWLEDGEMENTS

Over the course of this project, I have relied heavily on the knowledge, direction, and experience of multiple mentors who have made this project possible. I wish to acknowledge Dr. Ran Li, my Honors Thesis Reader, who provided the guiding questions in the context of molecular plant biology. I thank Dr. Shuqing Xu, who provided direction in the early stages of constructing the prediction pipeline and bioinformatic insight throughout the process. I am likewise grateful for Dr. Stephen Piccolo, who lent a listening ear and valuable perspective during the later stages of the project as I worked towards completion.

I also wish to acknowledge the Max Planck Institute for Chemical Ecology and the Director of the Department of molecular Ecology, Dr. Ian Baldwin, for providing the resources and internship opportunity in Jena, German that started me on this project.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

Introduction

Nicotiana attenuata, commonly known as the wild coyote tobacco, is a model plant that has been studied for its ecological plant-herbivore relationships. Found in the arid deserts of southern Utah, its ability to efficiently manage limited resources and maintain defensive measures against herbivores are of great interest. Many defense-related genes and transcription factors (TFs) are already characterized for *N. attenuata.*[1,2] However, the transcriptional regulation of many of these defense genes remain largely unknown, despite being well-documented in other plant species including *Arabidopsis thaliana*[3–5]. For instance, plant jasmonates (JAs), are essential to most defense responses and play important roles in various stages of development. The JA signaling pathway is regulated by jasmonate ZIM-domain (JAZ) repressor proteins, which target JA-responsive transcription factors[6]. In *A. thaliana,* a known target of these JAZ proteins is MYC2, a transcription factor known to regulate the JA-induced response[7]. Expression of JAZ genes is also directly regulated by the MYC2 transcription factor, creating a negative feedback loop that affords a fine-tuned level of regulation in this response system.[6] The exact nature of such JAZ protein transcriptional regulation in *N. attenuata* continues to be a topic of research.

In an effort to characterize the transcriptional regulation of *N. attenuata* defense response to herbivory and the signaling pathways involved (including the regulation of JA signaling), we constructed a bioinformatics pipeline to predict transcription factors (TFs), their transcription factor binding sites (TFBSs), and the associated genes they regulate. Following the model described by Yu et al.[8] for the maize genome, we performed a gene co-expression analysis on 62 sets of transcript level data and formed 1304 gene subsets, each associated with a known transcriptional regulator. We then analyzed subset gene promoters for motif enrichment and tested motifs for evolutionary conservation. Using these motifs to query online databases of known TF-TFBS pairs, we identified TFs in related plant species that bind to similar motifs. We then used these TFs to select homologous TFs in *N. attenuata* and form predicted TF-TFBS associations. To screen TF-TFBS candidates, we checked predicted TFBSs against motifs identified in a TF's subset and selected top matches for a final database of TFs-TFBS pairs (see Figure 1). Our finalized pipeline predicted 27 TF-TFBS pairs, each associated with a set of co-expressed genes from which we infer regulatory relationships for experimental testing. Our results predicted TFs involved in the regulation of JAZb and JAZd, as well as a TFBS for the JA-responsive TF WRKY3.

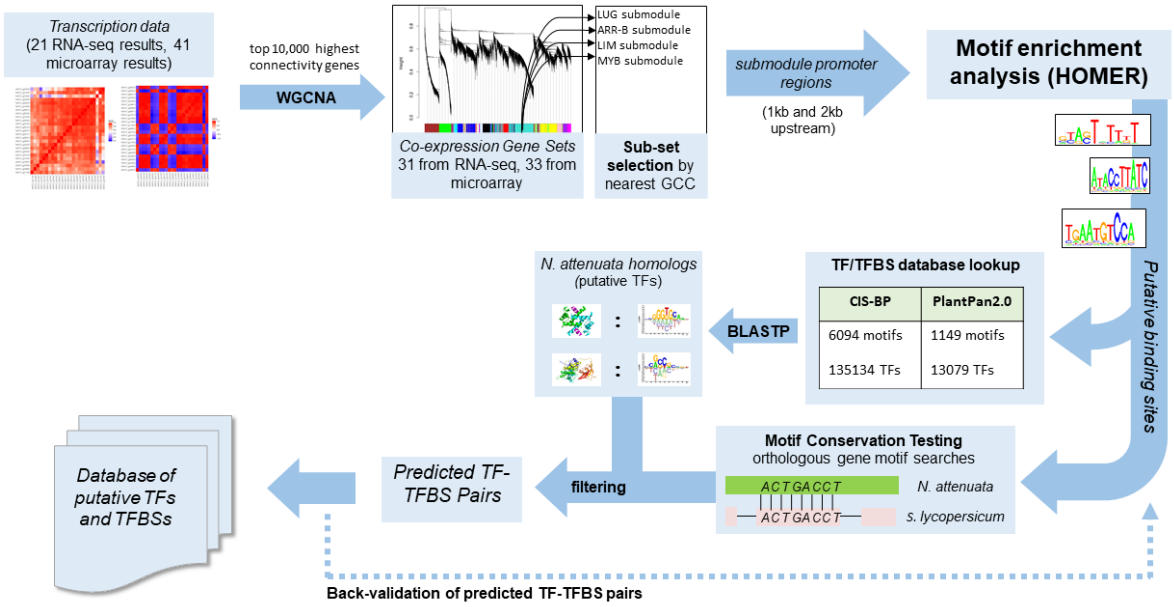

Figure 1: Summary of putative transcription factor and transcription factor binding site prediction pipeline. Analysis begins with transcript measurement data--21 samples of RNA-seq data and 41 samples of microarray data from various tissue and treatment samples--to generate a co-expression network based on highest connectivity genes via the Weighted Correlation Network Analysis (WGCNA) package[9]. This generated 31 co-expression gene sets based on the RNA-seq data and 33 from the microarray datasets; these we further clustered into subsets centered on transcriptionally relevant genes. We then identified overrepresented motifs in the 1kb and 2kb regions upstream of subset genes transcription start sites using the HOMER[10] software suite. Motifs appearing in at least 20% of gene subset promoter regions were checked for conservation against orthologous genes in *S. lycopersicum* and searched in plant transcription factor binding site (TFBS) databases. We used transcription factors (TFs) associated with the searched motif database hits as search queries in a BLASTP search to find homologous proteins in *N. attenuata.* We screened predicted TF-TFBS pairs by comparing motifs found in gene subsets associated with predicted TFs with the predicted TFBS and kept only TFBSs that best matched TF subset motifs. Data sets or outputs are italicized, while procedures are written in bold.

Results

Our pipeline identified 27 TF-TFBS pairs, each associated with a set of co-expressed genes as potential regulated candidates for a total of 1164 unique genes with predicted transcription factor binding on 507 of these genes. Among these, we identified gene subsets associated with chloroplast structure and function, ribosomal structure and function, cell membrane components, and ATP binding. While experimental validation of our immediate results has yet to be performed, we observe that several of our TF-TFBS and regulatory predictions align with results from past experiments and regulatory patterns observed in related species.

Transcript level results

As described in the Methods section, we based our analysis on 21 samples of RNA-seq data and 41 microarray experiment datasets from various experimental conditions and tissue types (see Methods or the *Nicotiana Attenuata* Data Hub[11] for a

more complete description of the experimental conditions). Using the Weighted Correlation Network Analysis (WGCNA) R package,[9] we selected the 10,000 genes with the highest network connectivity separately for the RNA-seq data and microarray data results. Visual plots of Gini correlation for the 25 highest connectivity genes from both datasets are given in Figure 2.

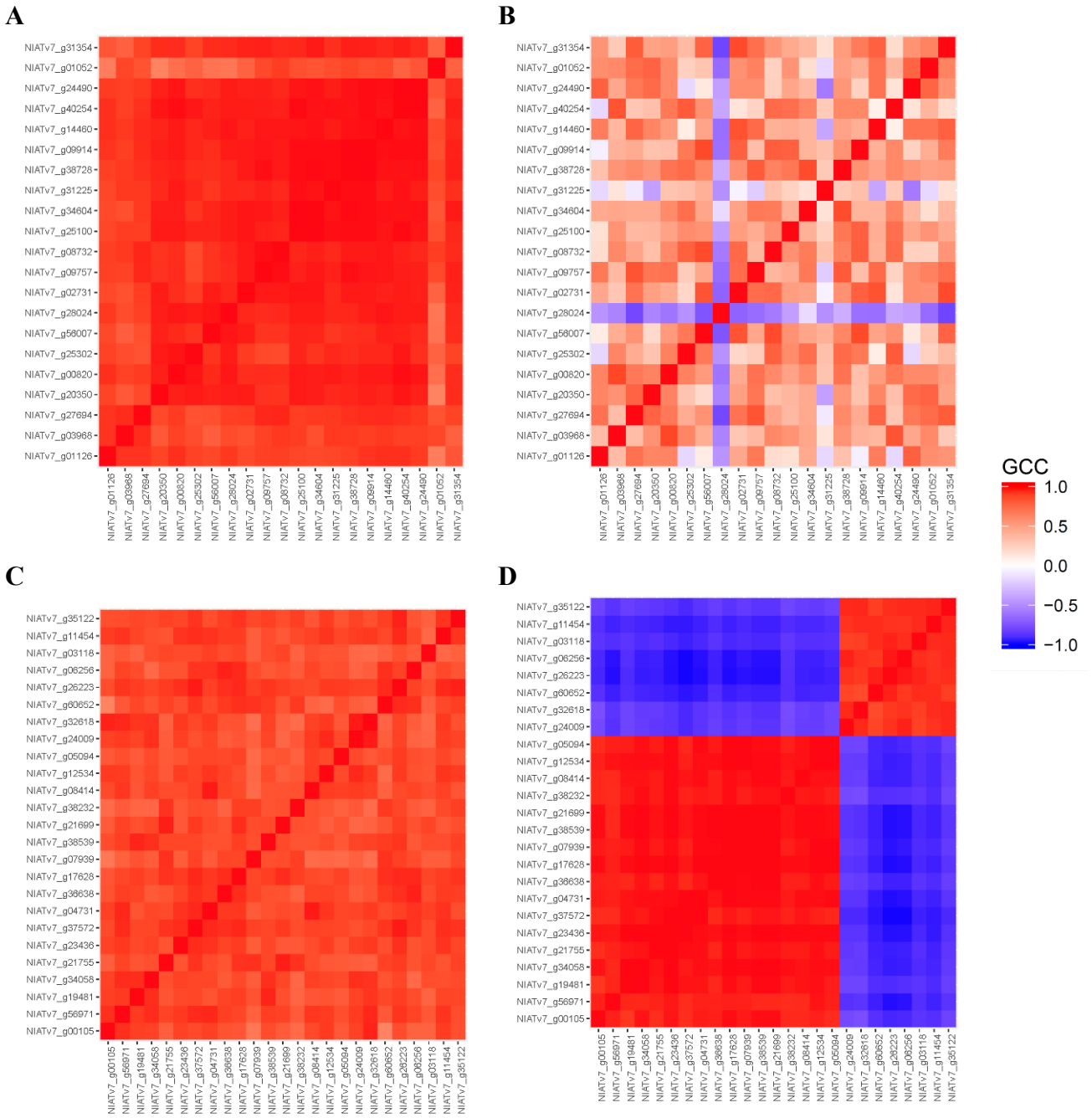

Figure 2: Correlation plots of the highest connectivity genes in *N. attenuata* based on microarray data (A, C) and RNA-seq data (B,D). We identified the top 25 genes using the *softConnectivity* function of the WGCNA R package for each dataset (RNA-seq, microarray) separately. The selected genes were then hierarchical clustered by Gini correlation coefficient (GCC) to reflect connectivity patterns within the top genes. GCC is coded for by color, with red representing positive correlation and blue negative correlation between gene transcript levels. **A and B)** Top 21 genes based on GCC calculated from microarray data. 4 of the initial 25 genes identified did not appear in the RNA-seq dataset and so were omitted from both plots. **A** is based on transcript level correlations from micro-array results; **B** is based on RNA-seq results. **C and D)** Top 25 genes based on GCC calculated from RNA-seq data. **C** shows transcript level correlations based on microarray data; **D** shows the same from the RNA-seq datasets. The clear disparity in expression patterns between the datasets reflect the different experimental conditions of the varied datasets.

The clear heterogeneity in gene correlation between the microarray and RNA-seq datasets (Figure 2A compared to 2B, and 2C compared to 2D) reflects the diverse nature of the experimental conditions and tissue types from which the datasets were drawn. This is highlighted by the fact that the 25 most connected genes identified by each data type had no genes in common between them. This suggests we have a broad and diverse sampling of genes for a robust downstream co-expression analysis. In accordance with these differences, we treated RNA-seq and microarray-based results separately throughout each step of the prediction pipeline.

Gene co-expression analysis

To identify groups of genes with similar expression profiles and potentially similar pathways of regulation, we performed gene co-expression analysis also using the WGCNA[9] package (see Methods section). Clustering resulted in 31 gene sets (also called modules) for the RNA-seq data, and 33 for the microarray-based set, with an average of 312.5 and 294.1 genes per gene set, respectively. Each gene set was assigned an arbitrary color by the package (Figure 3) for easy reference.

We then created smaller gene subsets within gene sets by selecting set genes most correlated to predicted transcriptional regulators, for a total of total of 623 RNA-seq-based and 681 microarray-based subsets. The distribution of these subsets was not necessarily proportional to the number of genes in each set, such that the initial size of a gene set was not reflected in final database results (Figure 3).

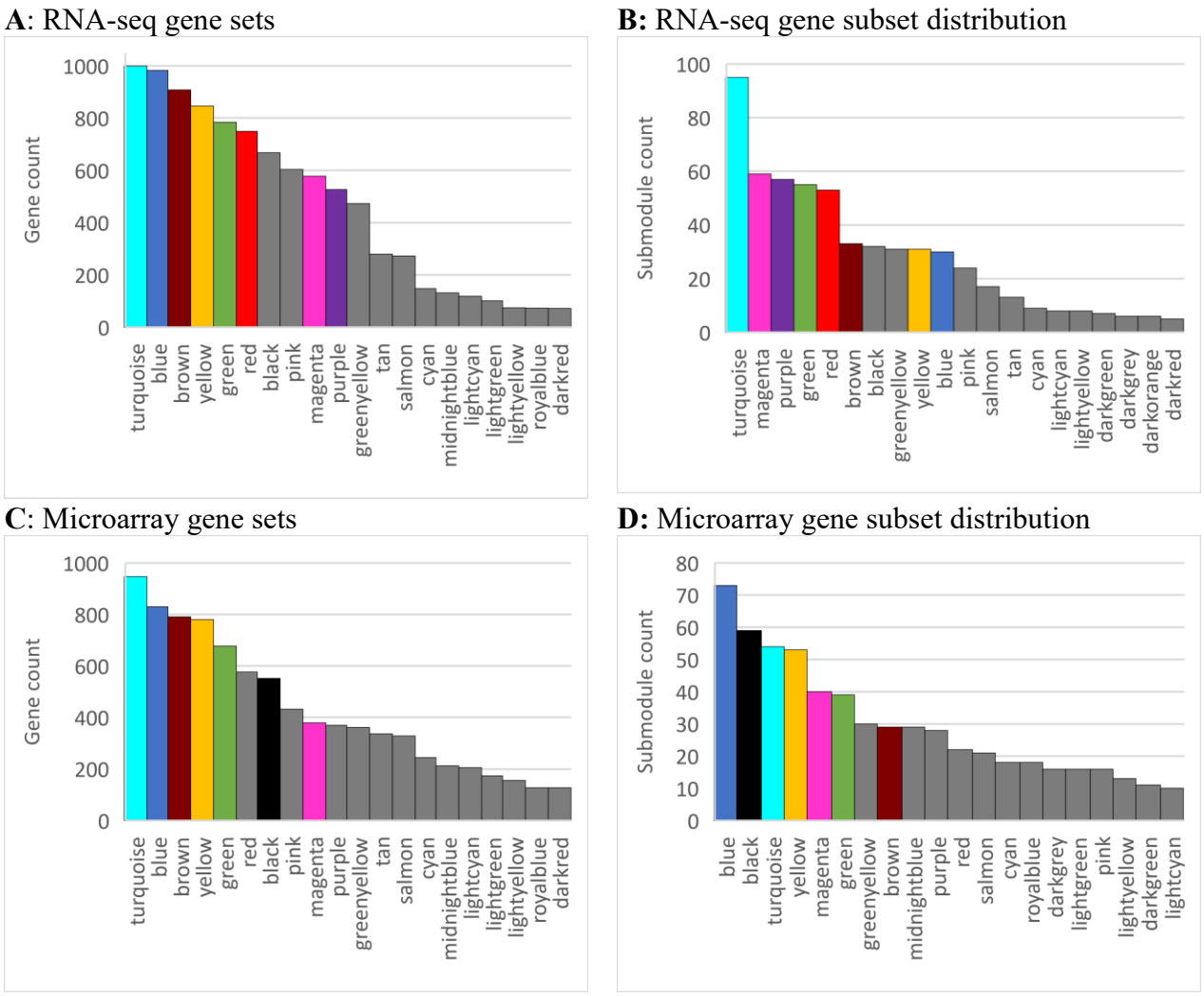

Figure 3: WGCNA gene set sizes (A, C) and distribution of subsets (B,D) for RNA-seq (A, B) and microarray (C, D) data. Subsets were selected based on highest Gini correlation coefficient surrounding predicted transcriptionally relevant genes[12] within a gene set. Colored bars indicate groups in the top 5 largest gene set or subset categories. **A)** The number of genes in each WGCNA set (from RNA-seq results). The *turquoise* gene set is the largest with 999 genes, while the smallest had only 34 genes. The average number of genes per set was 312.5 **B)** Number of gene subsets in each WGCNA set (from RNA-seq analysis). Note that while the *turquoise* module contains the greatest number of subsets (95), the second-largest *blue* set does not contain the second-most number of subsets. **C)** The gene count in WGCNA gene sets (from microarray results). *Turquoise* was the largest with 946 genes, while the smallest had only 54. The average gene set size was 294.1. **D)** Subset counts across gene sets (from microarray results).

To determine the biological relevance of our clustering technique, we examined the distribution of Gene Ontology (GO) terms among the gene sets and subsets. We observed that at least 5 of the RNA-seq based gene sets and 6 of the microarray-derived gene sets had non-random distribution of GO term assignments, as pictured in Figure 4 below ($p < 0.001$). Of particular interest was the overlap of GO enrichment across the sets identified by microarray and RNA-seq data; we found gene sets non-randomly associated with chloroplast structure, ribosomal structure and function, heme binding, membrane components, and ATP binding from both sources.

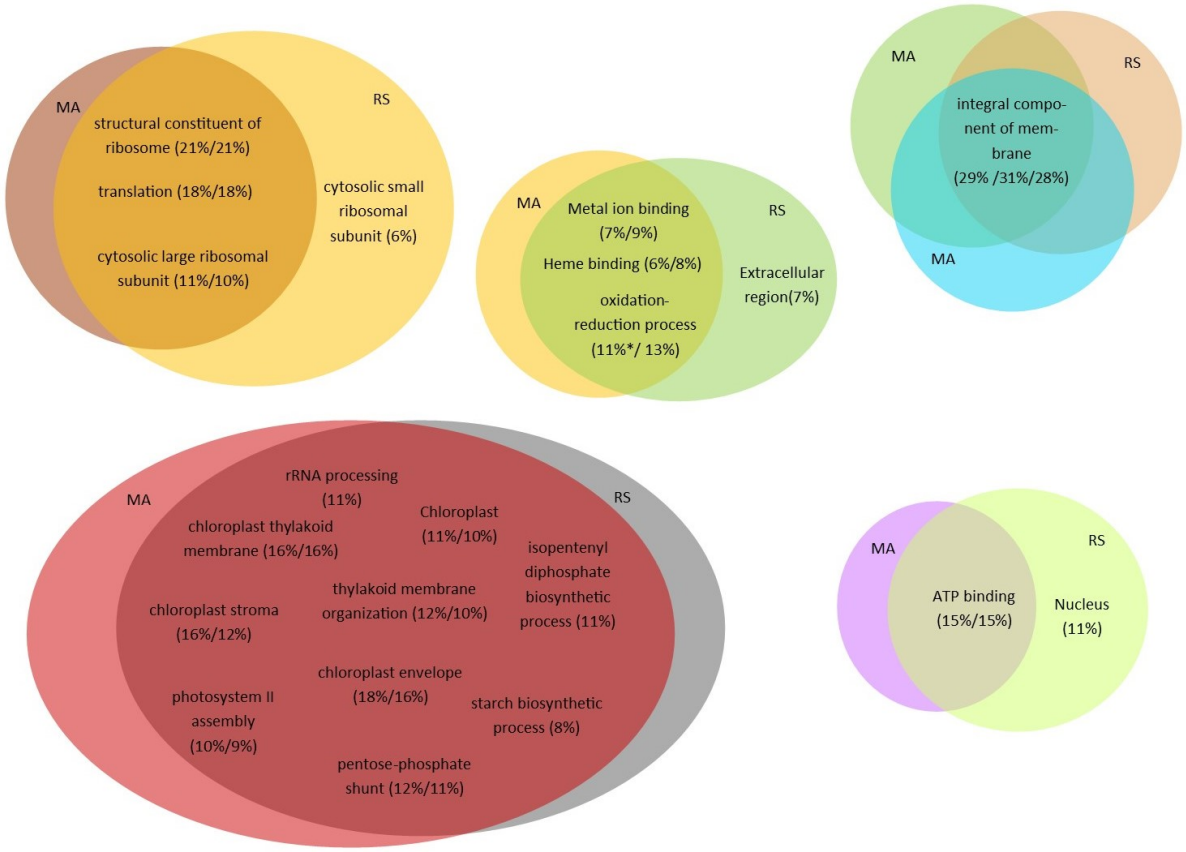

Figure 4: Gene sets enriched with GO terms. In the 11 GO-enriched gene sets identified by our co-expression analysis, we observed a high degree of overlap in annotation groups, suggesting that independent co-expression analysis on RNA-seq and microarray-based data created functionally similar gene groupings. MA refers to results from microarray analysis, RS from RNA-seq analysis. All enrichments have $p < 0.001$, unless otherwise noted. Colors indicate WGCNA assigned set color. (* $p = 0.001$)

Likewise, when we performed the same GO enrichment analysis procedure on the 27 gene subsets yielding top TF-TFBS candidate pairs, we observed 10 subsets which were statistically enriched for structurally or functionally-related GO terms (see Table 1). These groupings suggest that our step of sub-setting gene sets by GCC provides greater functional granularity then considering WGCNA gene sets alone.

Core Subset Gene	GO ID	GO Description
WRKY9 (NIATv7_g03410)	GO:0005576	extracellular region
	GO:0046872	metal ion binding
WRKY (NIATv7_g12711)	GO:0006355	regulation of transcription, DNA-templated
MYB-DIVARICATA (NIATv7_g17075)	GO:0009535	chloroplast thylakoid membrane
	GO:0006098	pentose-phosphate shunt
	GO:0006364	rRNA processing
	GO:0010207	photosystem II assembly
	GO:0019252	starch biosynthetic process
	GO:0009941	chloroplast envelope
	GO:0019288	isopentenyl diphosphate biosynthetic process, methylerythritol 4-phosphate pathway
	GO:0000023	maltose metabolic process
	GO:0010027	thylakoid membrane organization
	GO:0015995	chlorophyll biosynthetic process
	GO:0009570	chloroplast stroma
	GO:0009773	photosynthetic electron transport in photosystem I
	GO:0043085	positive regulation of catalytic activity
	GO:0009902	chloroplast relocation
	GO:0010218	response to far red light
	GO:0016117	carotenoid biosynthetic process
WRKY (NIATv7_g21131)	GO:0005509	calcium ion binding
WRKY65 (NIATv7_g27755)	GO:0046872*	metal ion binding
	GO:0005576	extracellular region
WRKY61 (NIATv7_g29978)	GO:0005576	extracellular region
	GO:0046872	metal ion binding
	GO:0020037	heme binding
	GO:0004601	peroxidase activity
	GO:0006979	response to oxidative stress
	GO:0042744	hydrogen peroxide catabolic process
	GO:0098869	cellular oxidant detoxification
GATA12 (NIATv7_g34810)	GO:0003735	structural constituent of ribosome
TRAF (NIATv7_g40277)	GO:0016021*	integral component of membrane
	GO:0005524	ATP binding
	GO:0006468	protein phosphorylation
	GO:0004672	protein kinase activity
	GO:0004674	protein serine/threonine kinase activity
	GO:0006612	protein targeting to membrane
	GO:0010363	regulation of plant-type hypersensitive response
bHLH (NIATv7_g41243)	GO:0005576	extracellular region
MYC2a† (NIATv7_g16429)	GO:0004672	protein kinase activity
	GO:0005524	ATP binding
	GO:0006468	protein phosphorylation
	GO:0009738	abscisic acid-activated signaling pathway
	GO:0035556	intracellular signal transduction

Table 1: Ten of 27 gene subsets yielding TF-TFBS pairs are enriched for GO terms. Enriched subsets are listed by their central TF. Colors of rows indicate original source gene set a subset came from. Only GO terms enriched in at least 10% of the submodule genes, with $p < 0.001$, are listed, unless otherwise indicated. (*p = 0.001). †MYC2a enrichment from the MYC2a microarray subset is also included as a point of interest but did not yield a final TF-TFBS pair.

For instance, the gene subset centered on the MYB/DIVARICARTA transcription factor (NIATv7_g17075) came from the black RNA-seq gene set enriched for GO terms relating to chloroplast structure and function. Our subset was additionally enriched with genes ($p < 0.001$) relating to photosystem I, response to far-red light, and chloroplast relocation. Likewise, the MYC2a microarray subset came from the magenta microarray gene set enriched for genes with ATP binding functionality but was additionally enriched for protein kinase activity. Interestingly, it has been shown that MYC2 phosphorylation is "required for MYC2 regulation of gene transcription," with the proposal that phosphorylation marks MYC2 as "spent" for proteolysis, thereby allowing other MYC2 molecules to interact with gene promoter regions and stimulate further transcription[13]. WRKY3, JAZa, and JAZf gene subsets had no meaningful GO term enrichment ($p >$ 0.001 or appearing in less than 5% of all genes).

Motif enrichment analysis

Over the course of our analysis, we identified 7528 unique motifs that passed our selection threshold ($p \leq 1 \times 10^{-5}$, appearing in at least 20% of all subset promoters) with a statistically significant conservation score ($p \leq 0.001$). 4038 unique motifs came from the microarray data, and 3672 from the RNA-seq data. Of these, only 980 motifs matched with a TF in our database lookup step. This large number of motifs with no match constitute a valuable space for further investigation into potentially novel cis-regulatory elements in *N. attenuata* in the future.

Motif analysis on the WRKY3 microarray subset identified a highly conserved W-box-like motif (GTTGAC) in both the 1kb and 2kb regions upstream of transcription start sites (TSS) in close to half of all subset genes (Figure 5). Interestingly, this binding site was identified in the promoter region of the WRKY3 (NIATv7_g07696) gene at 833 and 193 base pairs upstream of the TSS as well. WRKY3 is a known transcription factor associated with plant stress response in the jasmonic acid pathway that appears to be upregulated in response to plant wounding.[2] Members of the WRKY family are known to bind to W-box motifs, and in many cases even contain W-boxes in their own promoters. In fact, the homolog of WRKY3 in *Oryza sativa* WRKY70 has been shown to bind to a W-box its own promoter region as a potential regulatory factor.[14] Our results suggest a similar mechanism may be taking place in *N. attenuata*. For a list of WRKY3-subset genes containing the W-box motif see *Supplementary Tables 2* and *3*.

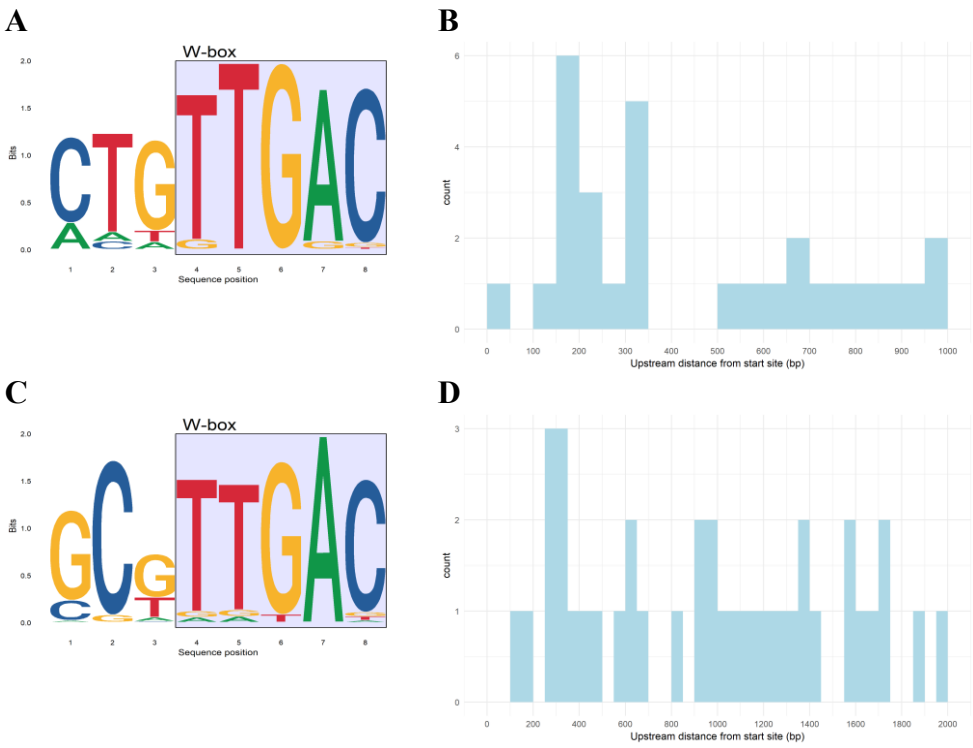

Figure 5: Top putative binding motifs and their distributions for the WRKY3 microarray gene set. A and C) Both motifs contain a conserved W-box motif associated with WRKY transcription factor binding. **A** was identified in the 1kb promoter analysis, with a conservation score of 0.529, and appears in 44% of promoter sequences. **C** was identified in the 2kb promoter analysis, with a conservation score of 0.583 and appearing in 58% of promoter sequences. **B)** Histogram of distribution of binding site A across the 1000kb promoter by distance from transcription start site (TSS). 57% of the binding sites appear in the 350 bps closest to the TSS, with the largest number between 150-200bp from the TSS. **D)** Histogram of distribution of binding site B across the 2000kb promoter by distance from the gene's transcription start site. 58% of binding sites appear in the 1100 bp closest to the TSS, with the highest number between 250-350 bps.

In the gene subset centered on JAZb, we identified conserved G-box-like motifs as the most highly conserved and highest frequency motifs in subset gene promoters (Figure 6). MYC2, a known regulator of JAZ family proteins, has been shown to bind to G-box motifs such as those we identified. Interestingly, analysis on promoters regions of JAZ genes in *A. thaliana* are enriched with the same motif pattern identified as a top motif candidate in this module (ACACGTGT)[7]. While our analysis didn't identify this exact motif in the promoter region of the JAZb gene, the motif CACGT appears with some frequency in the JAZb promoter and may interact with MYC2.

We also identified a motif containing the G-box (TCCACGTG) enriched for in promoters of the JAZd gene subset (from microarray data), appearing in 35% of promoter regions. Our analysis located this motif in the promoter region of the JAZd gene (Figure 6D).

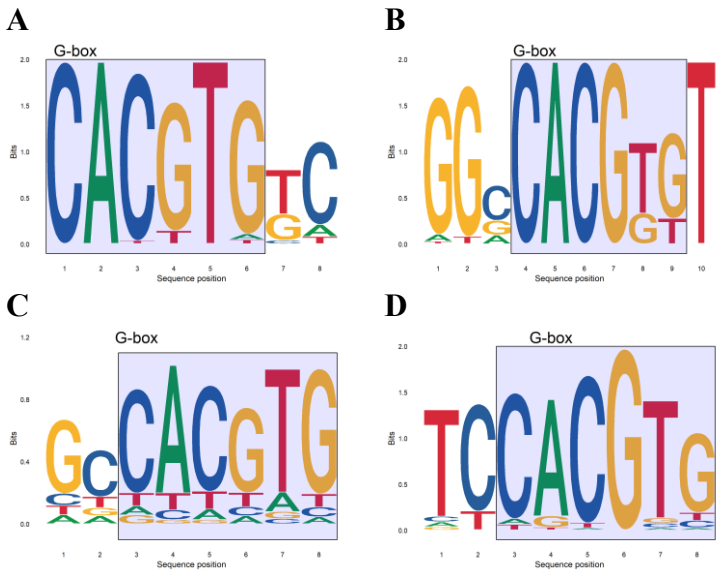

Figure 6: G-box motifs enriched in JAZb (A,B,C) and JAZd (D) gene subset promoters. All 4 motifs contain a conserved G-box motif (CACGTG). MYC2, which binds to G-box motifs, is a known regulator of JAZ repressor transcription, suggesting a potential regulatory relationship. **A** and **B** were identified in the 1000 kb promoter analysis of the microarray subset, with conservation scores of 0.5, and 0.6, respectively. **C** was identified in the 1000 kb promoter analysis from the RNA-seq subset with a conservation score of 0.61 and appearing in 46% of 2kb promoter sequences. **D** came from the 1000kb promoter analysis of the JAZd microarray subset, with a conservation score of 0.41 and appearing in 35% of promoters.

Database lookup and BLASTp search

From motifs used to search the PlantPan2.0 and CIS-BP databases, we identified 573 unique transcription factor candidates in *N. attenuata*. The top three most predicted transcription factors were all predicted to be members of the WRKY family, highlighting its central role in plant defense regulation.

Final results

Following the filtering and back-validation procedure described in the Methods section, we produced a list of 27 unique transcription factor – binding site pairs, including members of 10 distinct transcription factor families. 46% of these pairs come from the WRKY family.

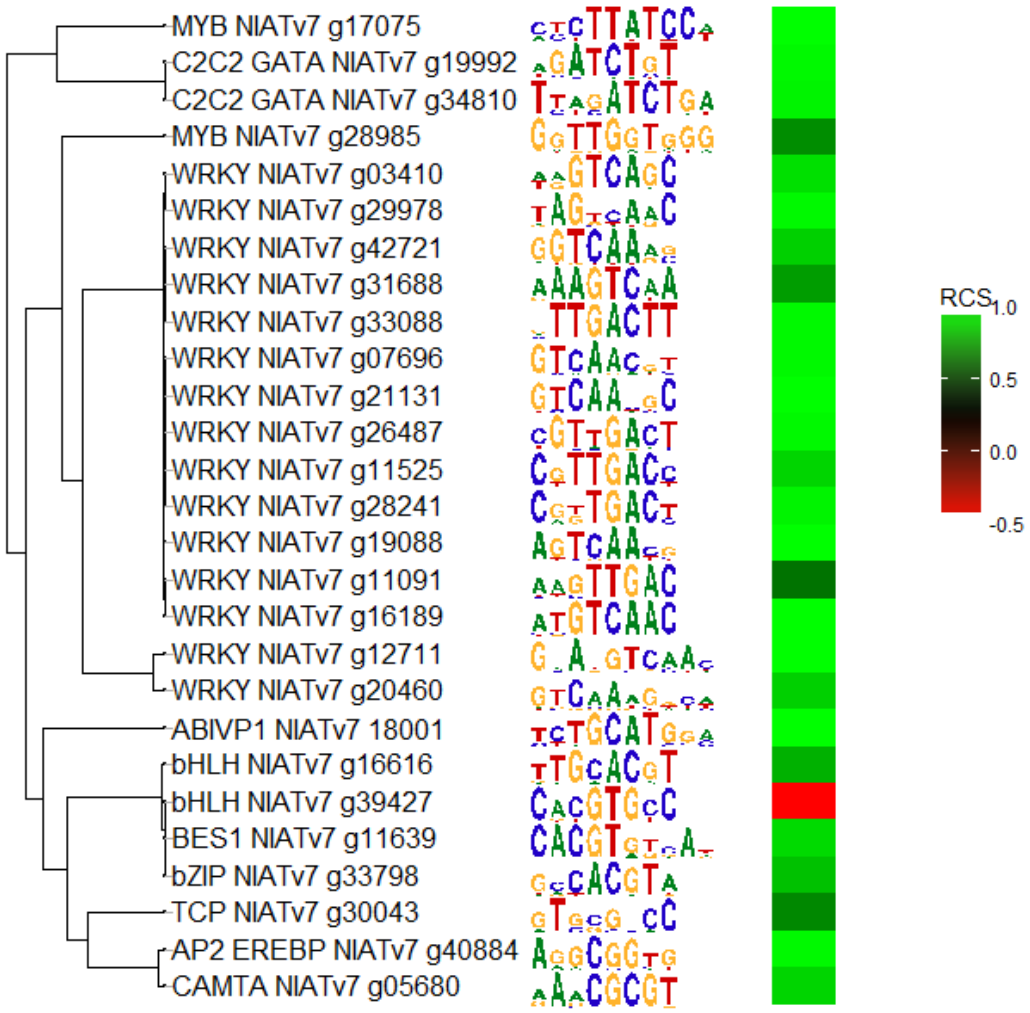

Figure 7: Predicted TF-TFBS pairs have high correlation with regulating gene sets. Predicted TFs are listed on the left with their family and gene names, grouped by motif similarity. Notice that the WRKY TFs group together due to similarities in binding motifs. Motif charts are aligned beside predicted TFs, and the heatmap on the far right shows the Relative Correlation Score (RCS) of the TF gene with the genes in the subset they were associated with as a predicted regulatory element. Note that the bHLH TF (NIATv7_g39427) has a negative RCS (-0.56), suggesting this element may be part of a repressive regulator loop.

As Figure 7 highlights, most of these transcription factors have high Relative Correlation Scores (RCSs) with the gene sets with which they are associated (see also Supplementary Table 1; RCS is a measure of correlation between the GCC scores of transcription factors and submodule genes, given on a scale of -1 to 1). Remarkably, 15 of the transcription factors identified were those around which their gene set centered, suggesting that co-expression analysis and subset selection had some power to identify units of transcriptionally regulated genes. 2/3 of the predicted TF binding sites came from 1kb upstream promoter regions, indicating that transcriptional binding patterns may be more easily inferable in this region.

Of particular interest among these motifs is the pattern identified for WRKY3 binding. The original WRKY3 gene set (NIATv7_g07696) from microarray data in isolation identified CTGTTGAC as a candidate binding motif. However, after back-validating and selecting for best matching motifs with this transcription factor, our final predicted binding site is GTCAACGT, its reverse complement. This final binding site more closely resembled motif database hits yielding this TF-TFBS pair, where this motif pattern (GTCAA) was listed as the forward sequence. See Supplementary Tables 2 and 3 for a list of all WRKY3 subset genes containing these binding sites.

Inferring regulatory relationships in JAZ and WRKY associated gene subsets

In an effort to determine the ability of our pipeline to infer regulatory relationships between proposed transcription factors and their associated gene subsets, we examined the gene subsets created in association with JAZb, JAZg, and WRKY3. Our observations demonstrate the limitations of our pipeline in predicting such relationships.

While not assigned specific TFBS by our final filtering steps, both MYC2a and MYC2b were identified as top transcription factor candidates for the JAZb gene subset (Supplementary Table 4; RCS = 0.79 and 0.65 respectively). In *Nicotiana tabacum,* these basic helix-loop-helix (bHLH) transcription factors form nuclear complexes with the NtJAZ1 repressor[15]. Given the strong evidence of transcriptional feedback regulation between JAZ genes and MYC2 transcription factors[7], these findings suggest a similar mechanism may be taking place for the JAZb repressor and its co-expressed genes in *N. attenuata*.

Both JAZd and JAZg were also associated with gene subsets in our final analysis. While MYC2 was not identified as a regulatory transcription factor candidate in either case, both JAZb and JAZd predicted bHLH62 (NIATv7_11555, the same family of TFs as MYC2) as a possible regulating TF. Interestingly, transcription factors predicted in association with the JAZg module largely came from the WRKY family and included WRKY3. This finding highlights that many families of transcription factors are involved in plant defense response and that these TFs may be co-induced and even interacting, as has been observed in *N. attenuata* and other plant species.[1]

The gene subsets (from microarray and RNA-seq data) associated with WRKY3 contained 13 genes in common and both predicted WRKY3 as a transcription factor. Both submodules also contained the mitogen-activated kinase 3 gene. MPKs operate as part of signaling pathways responding to external stress in plants, in which MPK kinases phosphorylate along a signal cascade, activating other MPKs which act on substrate proteins include transcription factors. In *Arabdiopsis*, for instance, perception of bacterial flagellin triggers an MPK pathway which activates WRKY family TFs, positively regulating defense gene expression. In *Arabidopsis,* MPK3 was shown to phosphorylate AtWRKY46 as part of AtWRKY46 degradation regulation.[5] Our association of MPK3 with the WRKY3 gene opens the possibility that a related mechanism may be taking place in *N. attenuata* and warrants further investigation. In both the RNA-seq and microarray gene subsets, the G-box binding site associated with WRKY3 was identified

in the MPK3 promoter region, suggesting the possibility that WKRY3 may also transcriptionally regulate MPK3.

Discussion and future work

While the examples described previously provide us some confidence of our pipeline's ability to predict putative TF-TFBS pairs, we ultimately need to perform experimentation to validate our findings. The variation we observe in our gene subset predictions highlight the need for strong knowledge of plant regulatory and response systems to accompany data-based predictions. In particular, inferring the regulatory relationships of genes and TFs is more challenging and requires more targeted experimentation. As the previous examples highlight, our analysis provides the potential to inform hypothesis formation regarding regulatory relationships between TFs and genes but cannot be extended beyond this without additional experimentation. For this reason, over the coming weeks, I hope to focus on testing these predictions by sampling gene expression in a WRKY3-silenced line of *N. attenuata* to validate submodule genes regulated by WRKY3 (such as MPK3), as well as quantitative PCR analysis of JAZb transcripts in MYC2-silenced plants to validate our prediction of MYC2 regulation of JAZb.

Considering the overall complexity of transcriptional regulation in plant species, identification of only 27 TF-TFBS pairs seems to be lower than expected. Indeed, Yu et al.'s [8] similar analysis on the maize genome yielded well over a100 new TF-TFBS pairs. We explain this difference in part by the many variable cutoffs and thresholds associated with these analysis, in addition to the more granular subsetting approach we adopted. We note, however, that the 27 pairs reported here represent only high confidence associations and not all likely associations suggested by our analysis. The large amount of data generated associating bindings sites, transcription factors, and gene lists is a fertile ground for further investigation and prediction of regulatory pathways in *N. attenuata* research. However, before such conclusions can be made, these top candidates should be experimentally validated.

Once validation has taken place, this data will be condensed for posting on the online *Nicotiana attenuata Data Hub* (http://nadh.ice.mpg.de/NaDH/).

Methods

Gene co-expression analysis

We first performed gene co-expression analysis on two large mRNA transcript datasets available for *N. attenuata* by the Mac Planck Institute for Chemical Ecology (MPI-ICE). When we started analysis in June 2016, this included 21 samples of RNA-seq data as well as 41 microarray experiments. The RNA-seq experiment samples came from plants in the 30th generation of an inbred line of a 1996 collection of native plants from Washington, Utah,[15] and spanned 11 different plant tissue types (ovary, nectary, anther, stigma, flower bud, corolla, root, leaf, stem, pedicel and flower). Samples were under

different biotic and abiotic stress treatments with different sampling schemes; a more complete description of experimental conditions and raw reads for each are available in the NCBI database accession number PRJNA317743, as well as the online *Nicotiana attenuata* Data Hub.[11] RNA was isolated using TRIZOL® (Thermo Fisher Scientific), and RNA-seq libraries were sequenced on an Illumina 2000 HiSeq platform with pair-end sequencing. Sequence reads were trimmed using Adapter Removal (v1.1) and then aligned to *N. attenuata* genome using TopHat2 (v2.1.0)[16]. Transcripts per million (TPM) was calculated using RSEM v1.2.20; a complete description of bioinformatic procedures is provided by Xu *et al*.[15] in the online Supplementary Information Appendix (section 2.5). Microarray samples were prepared on the Agilent platform GPL13527 and included wild type samples from leaves, roots and flowers. These samples were treated with either oral secretion from *M. sexta*, wounding, or no treatment, and were measured at various timepoints, including 3 and 6-hour intervals following treatment. Microarray probes were annotated based on gene predictions; a more complete description is provided by Brockmöller *et al.*[11] (see Additional file 1 in source 11).

We performed $\log_2$ transformation on both the RNA-seq and microarray data before further analysis. We then performed hierarchical clustering to identify and remove gene clusters that were extreme outliers based on visual examination. Top connectivity genes for each cluster were identified based on expression count results using the WGCNA[9] package *softConnectivity* function; we selected the 10,000 genes with the highest connectivity values. In the WGCNA package, connectivity acts as a measure of "how correlated a gene is with all other network genes," given as "the sum of connection strengths with the other network genes."[9] Following a methodology similar to that outlined in the online WGCNA tutorials,[17,18] we created an adjacency matrix for the 10,000 selected genes based on a Gini correlation coefficient (GCC) similarity matrix. Ma and Wang[19] demonstrated that the GCC outperforms commonly used correlation statistics (including Spearman and Pearson) in predicting regulatory relationships in transcriptome analysis in plants, and so we used it for our analysis. We used the GCC matrix to calculate a topological overlap distance matrix, which we hierarchically clustered. Afterwards, we performed tree cutting using the *cutreeDynamic* function included in the WGCNA package, selecting a *deepSplit* parameter of 3 to produce roughly 30 gene sets for each dataset and arbitrary minimum gene set sizes of 30 (RNA-seq) and 50 (microarray) genes. This resulted in 31 gene sets for the RNA-seq data, and 33 microarray-based gene sets, with an average of 312.5 and 294.1 genes per gene set, respectively. This process also assigned each gene a scaled intramodular connectivity score from 0 to 1 as a measurement of "gene set membership."

Subset creation

Gene sets were further divided by forming 'subsets' centered around transcriptionally relevant genes (within a gene set). Transcriptionally relevant genes were designated as either transcription factors or transcriptional regulators based on gene domain identification using the iTAK tool[12]. This produced a list 2509 transcriptionally relevant 'core' genes (2112 transcription factors and 397 transcriptional regulators)

around which subsets were centered. A subset was created if the 'core' gene: 1) had received a gene set assignment and 2) had an intramodular connectivity score > 0.5. Subsets were formed by selecting the top 10% genes most positively correlated (by GCC) with the 'core' subset gene within the given gene set. All subsets were required to have an arbitrary minimum of 30 genes for the RNA-seq based results or 50 for the microarray-based results. This resulted in a total of 623 RNA-seq-based and 681 microarray-based subsets.

Gene ontology enrichment analysis

To investigate the biological distribution of subsets and gene sets formed, we performed Gene Ontology (GO) enrichment analysis using a simple randomized gene permutation method. For any given module or subset, a group of genes was selected (from the group of 10,000 genes analyzed) at random equal in number to the size of the gene set. Using GO assignments made on the blast2GO platform[20] (default settings), the frequency of GO assignments appearing in a randomly created gene set were counted. This random sampling was performed 1000 times for each gene set and subset. The GO term assignment counts associated with actual gene sets/subsets were then assigned a p-value based on the frequency of a GO term count equaling or exceeding the GO term's frequency in the random permutation tests.

Motif analysis

For each subset, we performed motif analysis on candidate promoter regions both 1kb and 2kb upstream of gene TSSs, searching for motifs of both 8 and 10 bp lengths. Genomic sequences were provided by the MPI-ICE and sequenced using both Illumina HiSeq2000 and PacBio technologies as described by Xu et al[15]. Promoter regions for all genes in a given subset were extracted and then passed to *findMotifs.pl*, a motif searching tool in the HOMER suite version 4.8.3[10]. Promoter regions for all *N. attenuata* genes were used as background space in this analysis. Only motifs with enrichment $p\text{-}value \leq 1 \times 10^{-5}$ and appearing in at least 20% of all subset promoter sequences were considered in downstream analysis as putative cis-regulatory elements. Motif candidates were located within subset gene promoter regions using the HOMER *scanMotifGenomeWide.pl* tool with default parameters for later analysis.

Motif conservation testing

To assign each motif candidate a conservation score, we compared subset genes containing motif candidates to orthologous genes in *S. lycopersicum*. The MPI-ICE provided a one-to-one gene ortholog mapping for over 18,000 *N. attenuata* genes to genes in *S. lycopersicum* based on a BLAST reciprocal best-hits algorithm (see *Supplementary Information Appendix,* Source 15). Genes that both had ortholog assignments and contained a candidate motif hit were aligned to their *S. lycopersicum* ortholog using the YASS genomic similarity tool version 1.14[21] with default parameters and *-d 4* to provide the output in bed format. We then located intersections of the aligned region with known motif locations using the *intersectBed* tool from the BEDtools suite[22] version 2.25.0, with a required minimum overlap fraction of at least 90% of the motif

candidate with the aligned region to count as a 'conserved' hit (*intersectBed* -f 0.9; see Figure 8). The number of conserved hits (*c)* for a given motif (*n*) across all genes in a subset (*s)* were totaled and then divided by the total number of subset genes that both: 1) had ortholog gene assignments and 2) contained the motif in question (g_n). This gave a conservation score for each motif in each subset (S_{mn}) between 0 and 1, as shown below:

$$S_{mn} = \frac{\sum_{g_n \epsilon s} c_{g_n}}{\sum_{g_n \epsilon s} g_n}$$

Biologically, this score indicates how frequently a motif is conserved as a proportion of the total number of gene subset genes with orthologs in *S. lycopersicum*. For instance, a score of 0.5 would indicate that among 50 subset genes having orthologs, a given motif was considered "conserved" in 25 instances among all those genes.

To approximate the likelihood of encountering such scores for a given subset motif by random chance, we repeated this alignment and calculation process using randomly assigned gene orthologs from *S. lycopersicum* in the alignment step. This was repeated 1000 times for each subset candidate motif to generate a sample distribution. We then assigned a p-value to each score (S_{mn}) based on the frequency of a score $\geq S_{mn}$ appearing in the sample distribution.

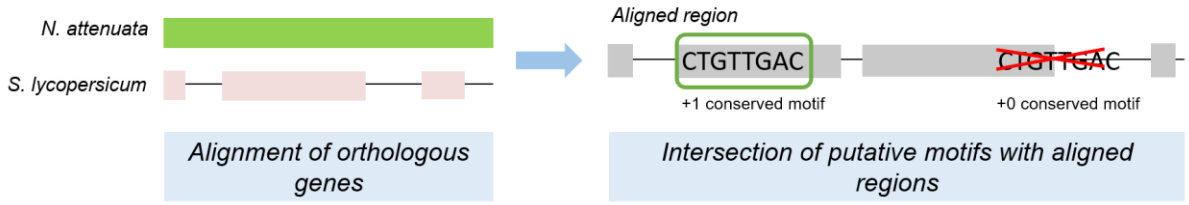

Figure 8: Identifying conserved putative motifs to calculate conservation score. We assigned each binding motif candidate a conservation score between 0 and 1 and an associated significance score (p-value) for later filtering steps. *N. attenuata* genes were first assigned a one-to-one gene ortholog in *S. lycopersicum*. Orthologs were then aligned using YASS; genes with no ortholog assignment were not included in analysis. We examined YASS aligned regions for intersection with motif candidate locations. Regions intersecting with at least 90% of the motif candidate were considered 'conserved' motifs. The total number of 'conserved' appearances of a motif in a given gene subset was divided by the number of genes that both contained the candidate motif and had an ortholog to calculate the conservation score. P-values were assigned to conservation scores by the frequency of conservation scores ≥ the assigned conservation score per 1000 permutations with random gene ortholog assignments.

Motif database search

Using the identified candidate motifs as queries, we searched for transcription factors binding to similar binding sites in related plant species. For this search, we used two large online TF/TFBS databases, PlantPan2[23] and CIS-BP[24] to identify TF-TFBS pairs. We downloaded both databases in August of 2016 and searched them specifically for cis-factors found in *Arabidopsis thaliana, Cucumis sativus*, *Populus trichocarpa,* and *Oryza sativa*. Data downloaded from PlantPan2 included 13079 unique TFs and 1149 unique TFBS; CIS-BP included 135134 unique TFs and 6094 unique TFBS. We used the *compareMotifs.pl* tool from the HOMER[10] suite v2.8.3 to compare subset motif candidates with known TFBSs in both libraries, requiring a minimum match threshold of 0.7 and accepting similar search queries with up to 0.9 similarity (reduction threshold setting; -matchThresh 0.70 -reduceThresh 0.9 -cpu 2). All TFBS-TF matches were given a match score from 0-1, with 1 being a perfect match.

BLASTp search

The MPI-ICE provided results for an all-by-all BLASTp search from several plants, including *A. thaliana, C. sativus*, *P. trichocarpa,* and *O. sativa,* against *N. attenuata.* From these search results, we selected TFs in *N. attenuata* with e-values $< 1 \times 10^{-10}$ when matched with corresponding TFs in the other plant species. These TFs became our putative TF candidates for further analysis.

Back-validation and filtering of TF-TFBS Pairs

To increase our confidence in predicted TF-TFBS pairs and eliminate multiple binding site assignments per TF, we compared the proposed TFBSs for a given TF to the motifs identified for that same TF's own gene subset. Comparisons were made across datatypes (i.e. motifs identified from RNA-seq data were compared to TF subsets derived from both microarray and RNA-seq subsets, where they existed). Comparing data across all subsets, we kept only the predicted binding sites with the highest match score to subset motifs. Ties between motifs were broken first by motif conservation score and then by TF-motif association scores from database lookup (by TF). Only TFs with a BLASTp percent identity of 60% or higher were kept. This resulted in 27 unique TF-TFBS pairs.

Additional metrics we used to discriminate between top TF candidates included BLASTp e-value, BLASTp percent identity similarity, and a Relative Correlation Score (RCS). We assigned RCS by first calculating the average Gini correlation coefficient for each of the genes in the dataset of 10,000 genes relative to all genes in a gene subset of interest (for a total of 10,000 averaged scores). We then calculated the Pearson correlation coefficient of these averaged scores and the Gini correlation scores of the candidate transcription factor to produce a final score between -1 and 1.

Code availability

Scripts used to generate figures and perform pipeline analysis are posted on the author's public GitHub repository (*https://github.com/aomdahl/N_attenuata_TF_TFBS_pipeline*).

Sources

1. Woldemariam, M. G. *et al.* NaMYC2 transcription factor regulates a subset of plant defense responses in Nicotiana attenuata. *BMC Plant Biol.* **13,** (2013).

2. Skibbe, M., Qu, N., Galis, I. & Baldwin, I. T. Induced Plant Defenses in the Natural Environment: Nicotiana attenuata WRKY3 and WRKY6 Coordinate Responses to Herbivory. *PLANT CELL ONLINE* **20,** 1984–2000 (2008).

3. Gfeller, A., Liechti, R. & Farmer, E. E. Arabidopsis jasmonate signaling pathway. *Science Signaling* **3,** (2010).

4. Dombrecht, B. *et al.* MYC2 Differentially Modulates Diverse Jasmonate-Dependent Functions in Arabidopsis. *PLANT CELL ONLINE* **19,** 2225–2245 (2007).

5. Sheikh, A. H. *et al.* Regulation of WRKY46 Transcription Factor Function by Mitogen-Activated Protein Kinases in Arabidopsis thaliana. *Front. Plant Sci.* **7,** (2016).

6. Chico, J. M., Chini, A., Fonseca, S. & Solano, R. JAZ repressors set the rhythm in jasmonate signaling. *Current Opinion in Plant Biology* **11,** 486–494 (2008).

7. Chini, A. *et al.* The JAZ family of repressors is the missing link in jasmonate signalling. *Nature* **448,** 666–671 (2007).

8. Yu, C.-P. *et al.* Transcriptome dynamics of developing maize leaves and genomewide prediction of *cis* elements and their cognate transcription factors. *Proc. Natl. Acad. Sci.* **112,** E2477–E2486 (2015).

9. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559 (2008).

10. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38,** 576–589 (2010).

11. Brockmöller, T. *et al.* Nicotiana attenuata Data Hub (NaDH): an integrative platform for exploring genomic, transcriptomic and metabolomic data in wild tobacco. *BMC Genomics* **18,** 79 (2017).

12. Zheng, Y. *et al.* iTAK: A Program for Genome-wide Prediction and Classification of Plant Transcription Factors, Transcriptional Regulators, and Protein Kinases. *Molecular Plant* **9,** 1667–1670 (2016).

13. Zhai Q, Yan L, Tan D, Chen R, Sun J, Gao L, et al. Phosphorylation-Coupled Proteolysis of the Transcription Factor MYC2 Is Important for Jasmonate-Signaled Plant Immunity. *PLoS Genet.* **9,** (2013).

14. Li, R. *et al.* Prioritizing plant defence over growth through WRKY regulation facilitates infestation by non-target herbivores. *Elife* **4,** (2015).

15. Xu, S. *et al.* Wild tobacco genomes reveal the evolution of nicotine biosynthesis.

Proc. Natl. Acad. Sci. **114,** 6133–6138 (2017).

16. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).

17. Langfelder, P. & Horvath, S. Tutorial for the WGCNA package for R : I . Network analysis of liver expression data in female mice 1 . Data input and cleaning. (2014).

18. Langfelder, P. & Horvath, S. Tutorial for the WGCNA package for R : I . Network analysis of liver expression data in female mice 2 . b Step-by-step network construction and module detection. (2014).

19. Ma, C. & Wang, X. Application of the Gini Correlation Coefficient to Infer Regulatory Relationships in Transcriptome Analysis. *Plant Physiol.* **160,** 192–203 (2012).

20. Götz, S. *et al.* High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **36,** 3420–3435 (2008).

21. Noé, L. & Kucherov, G. YASS: Enhancing the sensitivity of DNA similarity search. *Nucleic Acids Res.* **33,** 540–543 (2005).

22. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

23. Chow, C. N. *et al.* PlantPAN 2.0: An update of Plant Promoter Analysis Navigator for reconstructing transcriptional regulatory networks in plants. *Nucleic Acids Res.* **44,** D1154–D1164 (2016).

24. Weirauch, M. T. *et al.* Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell* **158,** 1431–1443 (2014).

Supplementary Resources

Supplementary Table 1: Complete list of 27 predicted TF-TFBS

Predicted TF	TF family	TF functional annotation	Predicted TFBS	TF-subset correlation	TFBS frequency in subset promoters (%)	TFBS Conservation Score
NIATv7_g05680	CAMTA	calmodulin-binding transcription activator 3-like	AAACGCGT	0.87	30	0.53
NIATv7_g31688	WRKY	probable WRKY transcription factor 65 isoform X1	AAAGTCAA	0.68	44.16	0.41
NIATv7_g11091	WRKY	WRKY transcription factor 22-like	AAGTTGAC	0.56	55.84	0.54
NIATv7_g19992	C2C2-GATA	GATA transcription factor 5-like	AGATCTGT	0.97	28.36	0.69
NIATv7_g40884	AP2-EREBP	dehydration-responsive element-binding 2A-like	AGGCGGTG	0.96	26.67	0.29
NIATv7_g19088	WRKY	probable WRKY transcription factor 57	AGTCAACG	0.98	36.84	0.71
NIATv7_g16189	WRKY	probable WRKY transcription factor 24	ATGTCAAC	0.99	50.65	0.59
NIATv7_g39427	bHLH	transcription factor bHLH104	CACGTGCC	-0.56	35.56	0.43
NIATv7_g11639	BES1	NA	CACGTGTCAW	0.88	36	0.62
NIATv7_g28241	WRKY	probable WRKY transcription factor 7	CGKTGACT	0.96	40.35	0.5
NIATv7_g11525	WRKY	probable WRKY transcription factor 11	CGTTGACC	0.86	60	0.68
NIATv7_g26487	WRKY	probable WRKY transcription factor 65 isoform X1	CGTTGACT	0.98	41.82	0.45
NIATv7_g17075	MYB	transcription factor DIVARICATA	CTCTTATCCW	0.99	21.21	0.5
NIATv7_g12711	WRKY	probable WRKY transcription factor 7	GAATGTCAAC	0.99	23.08	0.33

NIATv7_g33798	bZIP	G-box-binding factor 1-like isoform X1	GCCACGTA	0.80	44.44	0.61
NIATv7_g42721	WRKY	probable WRKY transcription factor 7	GGTCAAAS	0.85	52	0.42
NIATv7_g28985	MYB	myb-related 308-like	GGTTGGTGGG	0.65	22	0.4
NIATv7_g20460	WRKY	NA	GTCAAAGKC W	0.84	30	0.67
NIATv7_g21131	WRKY	probable WRKY transcription factor 7	GTCAACGC	0.99	28.07	0.38
NIATv7_g07696	WRKY	probable WRKY transcription factor 26	GTCAACGT	0.98	56.67	0.57
NIATv7_g30043	TCP	transcription factor TCP23-like	GTGCGYCC	0.63	42.68	0.25
NIATv7_g33088	WRKY	probable WRKY transcription factor 61	GTTGACTT	0.97	38.96	0.48
NIATv7_g29978	WRKY	probable WRKY transcription factor 61	TAGTCAAC	0.97	44.16	0.22
NIATv7_g18001	ABI3VP1	B3 domain-containing transcription factor ABI3-like	TCTGCATGGA	0.99	20.45	0.43
NIATv7_g34810	C2C2-GATA	GATA transcription factor 12-like	TTAGATCTGA	0.96	25.3	0.47
NIATv7_g16616	bHLH	transcription factor bHLH18-like	TTGCACGT	0.75	35.06	0.33
NIATv7_g03410	WRKY	probable WRKY transcription factor 9	WRGTCAGC	0.89	46.75	0.41

Supplementary Table 2: WRKY3 gene subset genes with G-box binding motif from RNA-seq results (GTCAACGT)

Gene ID	Functional Annotation	Binding site upstream from TSS (bp)	Strand
NIATv7_g01329	-	169	-
NIATv7_g01329	-	98	-
NIATv7_g02607	lysM domain receptor-like kinase 4	811	+
NIATv7_g02607	lysM domain receptor-like kinase 4	523	+
NIATv7_g02607	lysM domain receptor-like kinase 4	267	-
NIATv7_g02779	probable receptor kinase At5g47070 isoform X2	105	+
NIATv7_g06351	quinone-oxidoreductase homolog, chloroplastic	178	-
NIATv7_g07696	***Na*WRKY3**	**116**	**+**
NIATv7_g10571	lysM domain receptor-like kinase 4	779	+
NIATv7_g10571	lysM domain receptor-like kinase 4	263	+
NIATv7_g15248	probable receptor kinase At5g39020	601	+
NIATv7_g15285	mitogen-activated kinase 3	305	-

NIATv7_g18543	phospholipase D alpha 1-like	263	+
NIATv7_g19262	probable WRKY transcription factor 40	629	-
NIATv7_g19440	aspsartic protesase in guard cell 2-like	174	-
NIATv7_g19804	BRASSINOSTEROID INSENSITIVE 1-associated receptor kinase 1-like	868	+
NIATv7_g23441	premnaspirodiene oxygenase-like	134	-
NIATv7_g38552	subtilisin-like protease	985	+
NIATv7_g38552	subtilisin-like protease	678	+
NIATv7_g38951	isoflavone 2 -hydroxylase-like	599	+
NIATv7_g38951	isoflavone 2 -hydroxylase-like	57	+
NIATv7_g39472	probable phosphatase 2C 4	351	+
NIATv7_g39472	probable phosphatase 2C 4	178	+
NIATv7_g40325	anthocyanidin 3-O-glucosyltransferase 2-like	402	+

Supplementary Table 3: WRKY3-subset genes with G-box binding motif from microarray results ((C/G)TGTTGAC)

Gene ID	Functional Annotation	Upstream from TSS (bp)	Strand
NIATv7_g02578	receptor 12	205	+
NIATv7_g04487	PLANT CADMIUM RESISTANCE 2-like	192	+
NIATv7_g04487	PLANT CADMIUM RESISTANCE 2-like	1225	-
NIATv7_g06351	quinone-oxidoreductase homolog, chloroplastic	275	-
NIATv7_g06351	quinone-oxidoreductase homolog, chloroplastic	363	+
NIATv7_g06351	quinone-oxidoreductase homolog, chloroplastic	865	-
NIATv7_g06491	G-type lectin S-receptor-like serine threonine- kinase At1g11300	17	+
NIATv7_g06491	G-type lectin S-receptor-like serine threonine- kinase At1g11300	1280	+
NIATv7_g07696	**probable WRKY transcription factor 26**	**195**	**-**
NIATv7_g07696	**probable WRKY transcription factor 26**	**845**	**+**
NIATv7_g10571	lysM domain receptor-like kinase 4	203	+
NIATv7_g10571	lysM domain receptor-like kinase 4	271	-
NIATv7_g10571	lysM domain receptor-like kinase 4	603	-
NIATv7_g10671	U-box domain-containing 28-like	769	+
NIATv7_g10851	U-box domain-containing 28-like	1402	+
NIATv7_g12923	G-type lectin S-receptor-like serine threonine- kinase SD2-5	271	-
NIATv7_g13625	MACPF domain-containing CAD1	331	-
NIATv7_g13625	MACPF domain-containing CAD1	500	+
NIATv7_g13806	YLS9-like	838	+
NIATv7_g13806	YLS9-like	1853	+
NIATv7_g15247	probable receptor kinase At1g67000	171	+
NIATv7_g15247	probable receptor kinase At1g67000	1325	-
NIATv7_g15285	mitogen-activated kinase 3	305	+
NIATv7_g15285	mitogen-activated kinase 3	1589	+
NIATv7_g15931	F-box At1g78280	601	-

NIATv7_g15931	F-box At1g78280	925	+
NIATv7_g19262	probable WRKY transcription factor 40	1068	-
NIATv7_g19262	probable WRKY transcription factor 40	1702	+
NIATv7_g19440	ASPARTIC PROTEASE IN GUARD CELL 2-like	174	+
NIATv7_g20186	MLO 6	314	-
NIATv7_g20186	MLO 6	926	+
NIATv7_g20770	BPS1, chloroplastic-like	341	+
NIATv7_g20770	BPS1, chloroplastic-like	1194	+
NIATv7_g20961	sigma factor binding 1, chloroplastic-like	1395	+
NIATv7_g21618	probable phosphatase 2C 10	960	+
NIATv7_g23176	hydroquinone glucosyltransferase-like	1632	+
NIATv7_g23441	premnaspirodiene oxygenase-like	134	+
NIATv7_g23441	premnaspirodiene oxygenase-like	170	-
NIATv7_g23441	premnaspirodiene oxygenase-like	1700	+
NIATv7_g23468	Calcineurin-like metallo-phosphoesterase superfamily isoform 1	686	+
NIATv7_g23660	exocyst complex component EXO70B1	535	-
NIATv7_g23660	exocyst complex component EXO70B1	577	-
NIATv7_g24088	synaptotagmin-4 isoform X1	321	+
NIATv7_g27413	AP2 ERF and B3 domain-containing transcription factor RAV1-like	267	+
NIATv7_g32512	aspartic ase 1	135	-
NIATv7_g33395	phospholipid-transporting ATPase 1-like	640	+
NIATv7_g33943	methylesterase 11, chloroplastic	174	-
NIATv7_g34361	transmembrane	154	+
NIATv7_g34361	transmembrane	239	-
NIATv7_g34361	transmembrane	439	+
NIATv7_g34361	transmembrane	587	-
NIATv7_g34361	transmembrane	692	+
NIATv7_g35369	NAC transcription factor 29-like	322	-
NIATv7_g35369	NAC transcription factor 29-like	1134	+
NIATv7_g35369	NAC transcription factor 29-like	1956	-
NIATv7_g36048	FK506-binding 4-like	1554	+
NIATv7_g36323	multiple C2 and transmembrane domain-containing 2-like	315	-
NIATv7_g36323	multiple C2 and transmembrane domain-containing 2-like	719	+
NIATv7_g36323	multiple C2 and transmembrane domain-containing 2-like	941	+
NIATv7_g38552	subtilisin-like protease	678	-
NIATv7_g38552	subtilisin-like protease	985	-
NIATv7_g38552	subtilisin-like protease	1006	+
NIATv7_g39620	transmembrane ascorbate ferrireductase 1	988	-
NIATv7_g39620	transmembrane ascorbate ferrireductase 2	1395	+
NIATv7_g39620	transmembrane ascorbate ferrireductase 3	1731	-

Supplementary Table 4: Top transcription factor candidates for JAZb microarray gene subset (including MYC2a and MYC2b)

Gene	BLASTp e-value	BLASTp Percenty identity	Functional annotation	Predicted Binding Site	RCS
NIATv7_g23317	**0**	**62.02**	**transcription factor MYC2-like (MYC2b)**	**CACGTGTC**	**0.646649304**
NIATv7_g42868	7.00E-60	72.73	transcription factor ICE1-like	CACGTGTC	0.578483203
NIATv7_g16429	**5.00E-56**	**57.95**	**transcription factor MYC2-like (MYC2a)**	**CACGTGTC**	**0.787168699**
NIATv7_g15722	9.00E-52	58.1	basic leucine zipper 43-like	GGCCACGTGT	0.74265316
NIATv7_g31247	5.00E-36	70.48	ABSCISIC ACID-INSENSITIVE 5 5	GGCCACGTGT	0.57342476
NIATv7_g32085	3.00E-26	78.26	transcription factor SPATULA-like	TGCCACGTGT	0.586531843
NIATv7_g13774	1.00E-22	61.33	transcription factor SPATULA isoform X2	TGCCACGTGT	0.541747846
NIATv7_g11555	4.00E-22	72.73	transcription factor bHLH62	TGCCACGTGT	0.854109812
NIATv7_g02164	1.00E-21	72.73	transcription factor bHLH62-like	TGCCACGTGT	0.636721259