Genomewide prediction of jasmonate-related and tissue-specific *cis* elements and their cognate transcription factors in *Nicotiana Attenuata*

Background and Significance:

Nicotiana attenuata, the wild coyote tobacco, has been characterized as a model plant that is studied for its unique ecological relationships. Found in the arid deserts of southern Utah, its abilities to efficiently manage limited resources and maintain defensive measures against predation and pathogens are of great interest. Plant jasmonates (JAs), a well-described hormone family, are essential to these 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 the transcription factors and in turn supress JA-induced response genes. One well-documented target of these JAZ proteins is MYC2, a transcription factors that is known to regulate the JA-induced response. The expression of JAZ genes is also directly regulated by the MYC2 transcription factor, revealing a negative feedback loop that affords a fine-tuned level of regulation in this response system (Chico 2008).

Plant defensins, small proteins that perform different inhibitory functions, constitute one vital arsenal of plant defense. Defensin response may be induced by pathogen attack, plant wounding, or stress, along with a host of other plant signaling molecules (Bahman 2008). One particular defensin gene in *N. attenuata,* NaDEF, showed high expression levels in the early stages of flower development. Silencing JAZi, a JAZ protein in *N.attenuata*, by virus-induced gene silencing (VIGS) revealed a significant elevation of NaDEF transcript levels that extended throughout even the later stages of flower development, confirming that JAZi acts as a repressor

of floral defensins (unpublished data[1]). Since JAZ proteins target transcription factors, these results prompt the question of which transcription factor is regulating defensin expression and being targeted by JAZi. Furthermore, such a relationship would open possibility that the targeted transcription factor regulates JAZi expression, like the feedback regulation of MYC2.

The question raised above may be approached from a computational perspective by integrating multiple bioinformatics tools. In this pursuit, a robust and flexible algorithm could not only answer this case of JAZ regulated transcription but also generate a genomewide library of predicted transcription factors and their binding sites for *N. attenuata*. This database could potentially predict the transcriptional regulators involved in any of the plant's systems. This would be a great asset to molecular biologists researching *N. attenuata* while also answering our initial questions.

Statement of Intent:

In order to better understand the transcriptional regulation of defensin proteins by JAZ in *Nicotiana attenuata,* I will create and validate a computer tool that can predict transcription factor (TF) to transcription factor binding site (TFBS) pairs and generate a database of tissue-specific, jasmonate-inducible, and genome-wide TF-TFBS pairs in *Nicotiana Attenuata*. My tool will use the basic approach used by Yu et. al on the maize genome (2014) with some creative modifications in the clustering method. This tool will also provide the results to predict answers to the following questions:

1) What are the putative transcription factors for a given gene in *N. attenuata?*

[1] These results came from the research of my advisor, Ran Li, at the Max Planck Institute for Chemical Ecology Summer 2016. See source (5) in the "Annotated Bibliography" section below

2) Which genes does a given transcription factor regulate, and what are its putative binding sites?

I will then validate these results using experimental methods (EMSA testing to demonstrate transcription factor-DNA binding) in coordination with the Max Planck Institute for Chemical Ecology.

Methodology and Procedures

Following the general approach used by Yu et al. on the maize genome in 2014, I will construct a computational tool (here, called a *pipeline*) that follows the structure described below (outlined in Figure 2).

1. Condense available gene expression data (21 sets of RNA-seq data from various tissues and 44 sets of microarray results) and use it to perform gene co-expression analysis. This will be done using the Weighted Gene Co-Expression Network Analysis (WGCNA) R package (Langfelder 2008) and Gini correlation coefficient adjacency data. This analysis

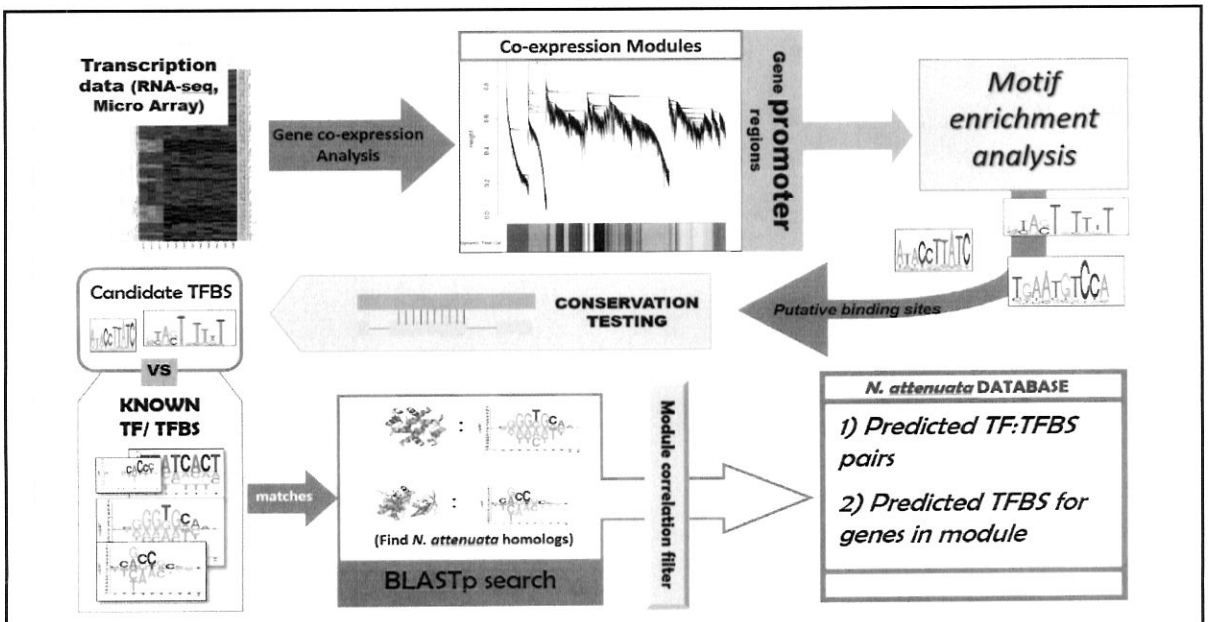

Figure 1: Diagrammatic representation of predictive pipeline

will generate a series of gene co-expression clusters, called "modules," that provide groups of genes that are co-expressed based on the transcription data.

2. Extract the promoter regions of the genes in each of these modules for motif analysis to identify likely transcriptional regulatory motifs within each module.
3. Perform motif analysis on the promoter sequences from each individual gene module using the HOMER motif analysis tool (Heinz 2010), and select the top motif candidates as putative transcription factor binding sites (pTFBS).
4. Perform conservation testing of the identified motif candidates against orthologous genes in a related plant species, *Solanum lycopersicum*. Validate these results by running the conservation analysis on randomly selected promoters for comparison.
5. Search conserved pTFBS in a library of known transcription factor binding sites in other well-documented plant species and identify their transcription factors.
6. Select the top motif matches from the library search to generate a list of transcription factors and binding motif identifiers for the other known plant species
7. Map these known transcription factors and their binding sites back onto *N. attenuata* using a BLASTp search to identify the most likely protein homologs associated with the pTFBS.
8. Assemble the results according to most relevant parameters (i.e. e-value, p-value, conservation score, percent frequency in module, etc.) into a database
9. Compare predicted modules and TF-pTFBS pairs with known results of transcription factor activity (i.e. gene-expression microarray data from transcription factor MYB8 silenced *N. attenuata* lines).

If the results aren't particularly satisfying, I will modify which input data I select (i.e. focus the dataset to JA-induced response genes) to improve the algorithm's prediction accuracy.

10. Work with the Max Planck Institute for Chemical Ecology to perform electrophoretic mobility shift assay (EMSA) testing on up to 3 identified TFs and their cognate pTFBS to experimentally validate predicted pairs, including a JAZ-target TF/defensin regulatory pair.
11. Finalize results and compile into a database for submission to the Max Planck Institute's online *N. attenuata* repository. Using this database, I hope to answer the initial question of regulation of defensin genes by JAZi-repressed TFs.

Preliminary Outline of the Finished Thesis

I plant to submit my findings for publication in a bioinformatics or plant biology journal, such as BMC Bioinformatics. Thus, my final paper will follow standard scientific format

- Abstract: A summary of the findings and a report on the database created
- Introduction: A summary of the initial question and the JA regulation pathway in *N. attenuata* and of our approach
- Results and Discussion: A summary of the pipeline and a gauge on its effectiveness, a sample clipping of the database produced and answers to the initial question about JAZ and defensin. This section will include diagrams summarizing results from the database, such as one illustrating certain transcription factors active in certain tissue types. This section will also include a summary of the experimental validation results.

- Conclusion: A brief description of what I learned in response to the initial question, the general usefulness of the database of TF-pTFBS for plant-wide systems, and where to access the database.
- Materials and Methods: A more detailed account of the pipeline mechanics, including the parameters used for various functions, a description of the tools used and any novel approaches included. This will also describe the experimental validation method in detail.

Preliminary Research

Please see the annotated bibliography below. While my preliminary research has come from reviewing several of the papers listed here, my understanding of *N. attenuata*, its transcription factors, and this project have been improved by discussions and presentations by researchers studying this plant during my internship at the Max Planck Institute.

Qualifications of the Investigator

I am pursuing a major in Bioinformatics and have been successful in my major coursework endeavors, maintaining a 4.0 GPA. My coursework has included an emphasis in molecular biology, computer science and mathematics. As a result, I have programming skills in Java, Python and R. I also spent the summer of 2016 working closely with lab and bioinformatics experts on *N. attenuata* at the Max Planck Institute for Chemical Ecology in Jena, Germany. Under their tutelage, I was exposed to real data results, lab procedures, and relevant literature regarding this topic and species. They provided the ideas, guidance and data for the bulk of my work on this project.

This experience, coupled with my academic coursework provide me with the resources and the skillset to complete this project, with support from advisors.

Qualifications of the Faculty Advisor

Dr. Piccolo, my BYU faculty advisor, has performed a wide variety of bioinformatics research, with a focus on transcriptomics, cancer, and machine learning, and currently leads a mentored research lab on-campus. His experience in effectively mentoring students and communicating challenging technical topics in peer-reviewed publications will be invaluable to me as I attempt to elucidate complex computational and biological concepts in academic prose.

Reader Because my project was initiated outside of BYU, I will also be working closely with researchers from the Max Planck Institute (MPI) for Chemical Ecology. Specifically, I will be coordinating with Ran Li to perform the EMSA validation testing and Shuqing Xu to help with the plant-specific datasets. Both have extensive experience with *N. attenuata*; Shuqing is involved in analyzing recently published genomic data for this plant, while Ran is investigating its transcription factors and their regulation of defense responses.

Schedule

Predicted Graduation Date: April 2018

Task	**Completion Date**
Review pipeline coded thus far and finish generating results *Complete genome-wide analysis on RNA-seq data* *Complete genome-wide analysis on microarray data* *Merge and cross-validate the genome-wide analysis sources*	Oct. 1-Dec. 10, 2016

Run pipeline on specific tissue expression datasets (leaf, petal, root) and compile results	
Coordinate with MPI to perform EMSA tests to validate selected results	December, 2016
Produce informative diagrams and begin 1st draft of manuscript	December, 2016
Final draft of manuscript complete	January 31, 2017
Thesis defense	February 2017
Submit final copies of thesis and submit to journal	March 2017

Budget:

Since the bulk of the research work has already been performed and is computationally focused, there are no immediate budget requirements for this project. However, the EMSA testing portion, will be performed at the Max Plank Institute for Chemical Ecology in Jena, where access to the plant and related laboratory resources are readily available. Because this work began as an internship under the MPI's supervision, this testing will fall under their budget.

Other notes:

This project was inspired by the question presented in the "Background and Significance" section above. To date, I have already coded the pipeline and run it to completion on 71/623 gene modules generated from RNA-seq data for *N. attenuata*. However, there yet remains significant work in fine-tuning the approach and running the pipeline on the gene-expression microarray datasets.

Works Cited:

Bahramnejad, Bahman, L. R. Erickson, A. Chuthamat, and P. H. Goodwin. "Differential Expression of Eight Defensin Genes of N. Benthamiana following Biotic Stress, Wounding, Ethylene, and Benzothiadiazole Treatments." Plant Cell Rep 28 (2009): 710-17.

Chico, Jose M., Andrea Chini, Sandra Fonseca, and Roberto Solano. "JAZ Repressors Set the Rhythm in Jasmonate Signaling." Current Opinion in Plant Biology 11.5 (2008): 486-94. Web.

Chi-Nga Chow, Han-Qin Zheng, Nai-Yun Wu, Chia-Hung Chien, Hsien-Da Huang, Tzong-Yi Lee, Yi-Fan Chiang-Hsieh, Ping-Fu Hou, Tien-Yi Yang, and Wen-Chi Chang "PlantPAN 2.0: an update of plant promoter analysis navigator for reconstructing transcriptional regulatory networks in plants", Nucleic Acids Res. 2015 : gkv1035v1-gkv1035.

Heinz S, Benner C, Spann N, Bertolino E et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell 2010 May 28;38(4):576-589. PMID: 20513432

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

Lay, F., and M. Anderson. "Defensins - Components of the Innate Immune System in Plants." Current Protein & Peptide Science CPPS 6.1 (2005): 85-101.

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, Zheng H, Goity A, van Bakel H, Lozano JC, Galli M, Lewsey MG, Huang E, Mukherjee T, Chen X, Reece-Hoyes JS, Govindarajan S,

Shaulsky G, Walhout AJ, Bouget FY, Ratsch G, Larrondo LF, Ecker JR, Hughes TR. Cell. 2014 Sep 11;158(6):1431-43. doi: 10.1016/j.cell.2014.08.009.PMID: 25215497

Yu, Chun-Ping, Sean Chen Chun-Chang, Yao-Ming Chang, Wen-Yu Liu, Hsin-Hung Lin, Jinn-Jy Lin, Hsiang Chen June, Yu-Ju Lu, Yi-Hsuan Wu, Mei-Yeh Lu Jade, Chen-Hua Lu, Arthur Shih Chun-Chieh, Maurice Ku Sun-Ben, Shin-Han Shiu, Shu-Hsing Wu, and Wen-Hsiung Li. "Transcriptome Dynamics of Developing Maize Leaves and Genomewide Prediction of Cis Elements and Their Cognate Transcription Factors." Proceedings of the National Academy of Sciences Proc Natl Acad Sci USA 112.19 (2015): n. pag. Web.

Annotated Bibliography

1) Chico, Jose M., Andrea Chini, Sandra Fonseca, and Roberto Solano. "JAZ Repressors Set the Rhythm in Jasmonate Signaling." Current Opinion in Plant Biology 11.5 (2008): 486-94. Web.

   Outlines the role of JAZ repression in jasmonate (JA) signalling in *Arabidopsis.* JAZ proteins are targeted for degradation in response to JA signaling, which is triggered by as a stress response to attack or resource deficiency. This allows transcription factors, like MYC2 to transcribe other JA response genes, including JAZs in a negative feedback loop.

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

This paper describes an R package for performing gene co-expression analysis using a scale-free topology approach. This package also includes tools to predict genes interconnectedness as well as tools to create visualizations of the clustering While the tool itself is presented in this article, I will use the tutorials and other resources available on the website "WGCNA: an R package for weighted correlation network analysis" which accompanies this paper.

3) Heinz S, Benner C, Spann N, Bertolino E et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell 2010 May 28;38(4):576-589. PMID: 20513432

This paper is the formal citation for the HOMER motif analysis and searching toolkit more specifically described at http://homer.salk.edu/homer/index.html. The Hypergeometric Optimization for Motif Enrichment or HOMER toolkit allows for *de novo* identification of over-expressed motifs in strings of DNA and offers tools to analyze several other bioinformatics datasets, including ChIP-Seq and RNA-Seq data.

4) Ma, C., and X. Wang. "Application of the Gini Correlation Coefficient to Infer Regulatory Relationships in Transcriptome Analysis." Plant Physiology 160.1 (2012): 192-203. Web.

Ma et. al present a comparison the Gini correlation coefficient as an alternative to traditional correlation coefficients (i.e. Pearson and Spearman correlation coefficients) in for considering transcriptional relationships in gene expression

data. Gini's consideration of rank and value allow it to capture non-linear relationships between transcription factors and their target binding sites and appears to be more robust for this purpose than the other correlation coefficients. Tools to calculate the Gini correlation coefficient are given in the R package *rsgcc.*

5) Transcriptome dynamics of developing maize leaves and genomewide prediction of cis elements and their cognate transcription factors. Chun-Ping Yu - Sean Chun-ChangChen - Yao-Ming Chang - Wen-Yu Liu - Hsin-Hung Lin - Jinn-Jy Lin - Hsiang JuneChen - Yu-Ju Lu - Yi-Hsuan Wu - Mei-Yeh JadeLu - Chen-Hua Lu - Arthur Chun-ChiehShih - Maurice Sun-BenKu - Shin-Han Shiu - Shu-Hsing Wu - Wen-Hsiung Li - Proceedings of the National Academy of Sciences Proc Natl Acad Sci USA – 2015

   Here, the maize genome is analyzed using a computer tool to predict transcription factors and their putative binding sites. Their approach employs gene set enrichment analysis for gene clustering, conservation testing and comparison to multiple known transcription factor databases. Their analysis predicted 1340 novel transcription factor binding sites and over 300 new TF-TFBS pairs with high reliability.

6) Lay, F., and M. Anderson. "Defensins - Components of the Innate Immune System in Plants." Current Protein & Peptide Science CPPS 6.1 (2005): 85-101.

   Plant defensins are small proteins typically between 45-54 amino acids long and involved in plant defense response. While their amino acid sequences vary, they reflect a structural conservation similar to insect defensin and scorpion toxins. Defensin production may be induced by pathogen attack or stress conditions. The

variety in inducible conditions reflect an intricate gene expression and signal transduction network behind defensin production. Different plant defensin may perform different defense activities, such as enzymatic inhibition, protein translation inhibition, or the blocking of ion channels

7) Yu, Chun-Ping, Sean Chen Chun-Chang, Yao-Ming Chang, Wen-Yu Liu, Hsin-Hung Lin, Jinn-Jy Lin, Hsiang Chen June, Yu-Ju Lu, Yi-Hsuan Wu, Mei-Yeh Lu Jade, Chen-Hua Lu, Arthur Shih Chun-Chieh, Maurice Ku Sun-Ben, Shin-Han Shiu, Shu-Hsing Wu, and Wen-Hsiung Li. "Transcriptome Dynamics of Developing Maize Leaves and Genomewide Prediction of Cis Elements and Their Cognate Transcription Factors." Proceedings of the National Academy of Sciences Proc Natl Acad Sci USA 112.19 (2015).

Using two computational approaches, Yu et. al predicted 253 new TF-TFBS pairs and 1340 novel transcription factor binding sites in the maize genome, and validated selected results using EMSA testing. They describe two computational approaches: the first clusters data using gene set enrichment analysis, performs motif analysis on gene set promoter regions and compares these motifs to TF-TFBS libraries to identify likely TF binding candidates. TF homologs in maize were found from these search hits, and these homologs were used to identify their co-expressed genes and then predict their TFBS. The second approach begins with known TF-TFBS pairs in other species, and then searching for homologous TFs in maize with similar DNA-binding domains. These TFs were then used to identify a co-expressed gene set from which they predicted new TF-TFBS pairs.

8) Ran Li, unpublished results. My advisor at the MPI shared with me several of his findings as the foundation for my original research project there. This included investigation into the role of defensins in *N. attenuata* and their regulation by members of the JAZ family. Specifically, he demonstrated through VIGS testing that JAZi acts as a repressor of the NaDEF gene.