Dissertations, Master's Theses and Master's Reports

2021

CONSTRUCTION AND ANALYSIS OF GENETIC REGULATORY NETWORKS WITH RNA-SEQ DATA FROM ARABIDOPSIS THALIANA

Tessa Kriz
Michigan Technological University, tkkriz@mtu.edu

CONSTRUCTION AND ANALYSIS OF GENETIC REGULATORY NETWORKS
WITH RNA-SEQ DATA FROM ARABIDOPSIS THALIANA

By

Tessa K. Kriz

A REPORT

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

In Statistics

MICHIGAN TECHNOLOGICAL UNIVERSITY

2021

This report has been approved in partial fulfillment of the requirements for the Degree of MASTER OF SCIENCE in Statistics.

Department of Mathematical Sciences

Report Advisor: *Dr. Kui Zhang*

Committee Member: *Dr. Hairong Wei*

Committee Member: *Dr. Qiuying Sha*

Department Chair: *Dr. Jiguang Sun*

Table of Contents

List of Figures

List of Tables

Acknowledgements

List of Abbreviations

ADMM: Alternating Directions Method of Multipliers

BIC: Bayesian Information Criterion

DiffNetFDR: Differential Network Analysis with False Discovery Rate Control

FDR: False Discovery Rate

GRN: Gene Regulatory Network

JRmGRN: Joint Reconstruction of multiple Gene Regulatory Networks

MeJa: Methyl Jasmonate

SA: Salicylic Acid

Abstract

Reconstruction of gene regulatory networks (GRNs) is a fundamental aspect of genetic engineering and provides a deeper understanding of the biological processes of an organism. Two methods were implemented to reconstruct the gene regulatory networks of *Arabidopsis thaliana* under two treatments: methyl jasmonate (MeJa) and salicylic acid (SA). The Joint Reconstruction of multiple Gene Regulatory Networks (JRmGRN) method was utilized to construct a joint network for identifying hub genes common to both conditions in addition to networks specific to each condition. The Differential Network Analysis with False Discover Rate Control method constructed a network of connections unique to only one of the two conditions. Both methods produced biologically significant results and complemented each other. The two methods were tested and found to have efficacy to identify notable hub genes involved in MeJa/SA signaling pathways and downstream response.

1 Introduction

Gene regulatory networks (GRNs) are collections of molecular regulators that interact with each other to control a variety of developmental and cellular functions in plants and animals. These regulators influence cellular functions by governing the levels of gene expression, basically telling an organism what features to turn on and to what degree. Gene expression data is collected by measuring messenger RNA (mRNA) levels or by directly measuring protein levels. GRNs can respond to external environmental changes, making them ideal subjects of study for genetic engineering. Therefore, the ability to reconstruct these GRNs becomes critical for serious study, and the ability to simultaneously reconstruct multiple GRNs across different conditions provides even deeper insight into their functions.

Gene regulatory networks for multiple conditions or tissues can be broken down into two subgroups: networks that span across all conditions/tissues, and networks that are unique to each specific condition. Genes involved in the network for multiple conditions are called hub genes and can be thought of as the interstate highways connecting many cities across a state. Within network genes are genes analogues to streets and freeways specific to each individual city in a state. The reconstruction of these gene regulatory networks can be pictured roughly as mapping out these streets and highways.

Numerous methods for reconstructing gene regulatory networks have been developed over the past several decades, each taking a slightly different approach and having unique strengths. Some methods, such as the maximized L1 regularized log-likelihood function method, developed by Yuan and Lin (Lin *et al*., 2007), only model within network relations but do not model hub genes. The methods proposed by Liu and Ihler (2011) and Tan *et al*. (2014) model hub genes in addition to within network genes but cannot handle more than one dataset simultaneously. The joint reconstruction of multiple gene regulatory networks (JRmGRN) method proposed by Deng *et al*. (2018) efficiently models both within network genes and hub genes for more than one dataset simultaneously. The JRmGRN method has been shown to be more accurate and more effective at reconstructing gene regulatory networks than the competing methods, which is why it has been chosen for this analysis.

By allowing for joint reconstruction of GRNs, the JRmGRN method provides a means for the analysis of *Arabidopsis* genetic expression data between two different conditions. This method is also the most recent available, taking advantage of previous work in the field and providing the most developed and well-rounded method for reconstructing gene networks.

The output of JRmGRN for two conditions consists of three networks: one network for the hub genes and their connections between themselves and all other genes and two networks for each respective condition. Although the differential network can be constructed through the output of JRmGRN indirectly, JRmGRN is not a method that can directly infer the differential network of two conditions. A differential network analysis compares individual networks from different conditions and identifies the connections that are condition specific. The differential network can reveal the different gene roles and their associations among different conditions, thus enhancing the understanding of underlying mechanisms of biological processes. With the availability of a large volume of gene expression data, differential network analysis has become increasingly popular, and a number of methods have been developed. In this study, DiffNetFDR, a tool for the differential network analysis with false discovery rate control (Zhang *et al*., 2019) was used. DiffNetFDR is an R package (https://github.com/Zhangxf-ccnu/DiffNetFDR) that implements two methods for testing differences in different Gaussian graphical models. It also provides the visualization tools to display the identified differential network and has several other advantages compared with other tools for the differential network analysis, including the use of data to determine the tuning parameter, the use of the false discovery rate to control the multiple testing, and the use of partial correlation coefficients.

The efficacy of JRmGRN was first exploited in identifying common and unique hub regulatory genes across two tissues or conditions through building two gene regulatory networks using one model. The data used was the RNA-seq data from *Arabidopsis thaliana Columbia* ecotype plants, which were treated with two hormones called methyl jasmonate (MeJA) and salicylic acid (SA) and the plant materials were harvested in a time series for transcriptome profiling. *Arabidopsis thaliana* was chosen because it is a model plant

species with well annotated genes and publicly available gene expression data. The main focus was on identifying the hub regulators in each treatment condition (MeJA/SA) in comparison with the control condition (Mock). In addition to JRmGRN, the use of DiffNetFDR was also explored to identify the differential network and compared it with the differential networks found from JRmGRN.

2 Methods

2.1 Description of JRmGRN Process

The joint reconstruction of multiple gene regulatory networks (JRmGRN) method used in this project was developed by Deng *et al.* (2018) as a new graphic Gaussian model that would allow for the construction of a network from common hub genes as well as tissue or condition specific gene regulatory networks using gene expression data. Although previous methods have been created to construct gene regulatory networks (GRNs) for hub genes or tissue specific genes, no method effectively modeled both simultaneously for multiple conditions or tissues. Hub genes are genes within the network that are highly connected to other genes and are often a small minority in comparison to the total number of genes. Networks built around hub genes provide insight into common networks that many tissues share. Similarly, tissue or condition specific networks provide insights into GRNs that are unique to that tissue or condition and how its development is distinguished from other tissues.

One assumption behind the JRmGRN method is that each tissue/condition network can be decomposed into a network that is tissue unique and a network that is composed of hub genes shared between several tissues/conditions. Identifying these networks aids in the ability to understand complex biological systems and how they can be altered.

The JRmGRN method is based on the Gaussian graphical model to infer the precision matrix of genetic expression data from K conditions, where $K \geq 2$. The gene expression data from each condition or tissue can be represented as an $n_k \times p$ matrix $Y^{(k)}$ where n_k corresponds to the number of samples for that condition and p corresponds to the number of genes. Each row in $Y^{(k)}$, for $k = 1, \dots, K$, is assumed to be independent and follow a multivariate normal distribution with the covariance matrix Σ_k. For each gene expression data matrix $Y^{(k)}$, the covariance matrix is Σ_k and the precision matrix, denoted by $\Theta^{(k)}$, is inferred to obtain the gene regulatory network. This inference is made by having the non-zero elements in $\Theta^{(k)}$ indicate a conditional correlation between genes i and j $(i \neq j)$ in the gene regulatory network for the kth condition or tissue. Most elements of the precision

matrix are anticipated to be 0, while a select few genes are highly correlated with many other genes. These highly correlated genes are called hub genes in the network.

Since these hub genes can be divided into two groups (the hub genes that are common across all tissues or conditions and the hub genes that are specific to a tissue or condition) Deng *et al.* (2018) proposed the precision matrix, $\Theta^{(k)}$, could be broken into two components: $\Theta^{(k)} = Z^{(k)} + V + V^T$ for $k = 1, \dots, K$, where $Z^{(k)}$ corresponds to the network that is specific to the kth tissue or condition, and V represents the hub genes and their association with other genes across several networks.

The JRmGRN method determines that the gene regulatory network can be reconstructed by solving the following penalized log-likelihood function:

$$argmin_{\{\Theta^{(k)} \in S^+, k=1,\dots K\}} \sum_{k=1}^{K} -n_k \left(log \left(det(\Theta^{(k)}) \right) - trace(S^{(k)}\Theta^{(k)}) \right) + P(\Theta) \quad (1)$$

Where $S^{(k)}$ is the sample covariance matrix estimated from $Y^{(k)}$, S^+ denotes the set of symmetric positive semidefinite matrices, and $P(\Theta)$ is the penalty term defined by equation 2:

$$P(\Theta) = \lambda_1 \sum_{k=1}^{K} \left\| Z^{(k)} - diag(Z^{(k)}) \right\|_1 + \lambda_2 \sum_{k<k'} \left\| Z^{(k)} - Z^{(k')} \right\|_1 + \lambda_3 \|V\|_1 + \lambda_4 \|V\|_{1,2} (2)$$

Four penalties have been added to the model in the form of L1 and L2 norms. The penalties of the function (equation 2) occur with the help of the four non-negative tuning parameters, $\lambda_1, \lambda_2, \lambda_3$, and λ_4, each serving a unique purpose. The first penalty encourages the off-diagonal elements of $Z^{(k)}$ to be zero, the second encourages the elementary networks across different conditions/tissues to be the same, the third encourages some elements of a hub gene column in V to be zero so that a hub gene is not connected to every gene, and the fourth penalty forces some nearly zero columns in V to be zero columns.

For a set of fixed values of the tuning parameters, Deng *et al*. (2018) developed an alternating directions method of multipliers (ADMM) algorithm using a scaled augmented Lagrangian to find $Z^{(k)}$ and V that minimize the objective function in (1).

However, the selection of tuning parameters is complex as each combination is selected specifically for each set of tissues/conditions and must be determined based on the gene expression data used. Tuning parameters play a crucial role in ensuring the true networks are identified, as well as a reasonable number of hub genes. Deng *et al*. (2018) introduced two theorems to aid in the selection of tuning parameters, noting that the matrix V would be all zeros if the condition $\lambda_1 \le \frac{\lambda_4}{2K\sqrt{p}} + \frac{\lambda_3}{2K}$ was met, thereby reducing the number of hub genes to zero. A Bayesian information criterion (BIC) like quantity can be used to evaluate the selection of tuning parameters, though it is not the only selection criterion. Until further research can be done on the best approach to select these parameters, Deng *et. al* (2018) suggested using a uniformed grid search of log space for each of the four parameters independently. The JRmGRN method is implemented in R and can be used to search the best gene regulatory network across a set of tuning parameters.

2.2 Description of the DiffNetFDR Process

The differential network analysis was also conducted using DiffNetFDR (Zhang *et al*., 2019), an R package that implements the Gaussian graphical model to identify the gene network that is specific to a condition/tissue. Since different methods can construct vastly different gene networks, it is of great interest to compare the condition/tissue specific gene regulatory networks found from JRmGRN and DiffNetFDR.

The differential network analysis with false discovery rate control method (DiffNetFDR), constructed by Zhang *et al.* (2019), is the compilation of several influential Gaussian graphical methods into one user friendly package. The methods included in the construction of DiffNetFDR are based on regularization techniques that require a selection of tuning parameters. The goal of these tuning parameters is to control the false discovery rate (FDR), the rate at which differential edges are incorrectly identified, and keep it within a target level. However, selection of these tuning parameters is often complicated or

unclear. Two previous methods, developed by Xia *et al*. (2015) and Liu (2017) proposed methods to test the differences between two Gaussian graphical models with test-based techniques to select tuning parameters in a data-driven manner and control FDR at a predetermined level. Since the methods developed by Xia *et al.* and Liu are difficult to implement, Zhang *et al.* (2019) applied concepts from both to implement the "DiffNetFDR" R package that even those without strong backgrounds in statistics can utilize.

The DiffNetFDR method consists of two hypothesis test-based differential network analysis methods: the difference of precision matrices and the difference of partial correlations. In order to make the best comparison with the JRmGRN method, the difference of precision matrices method, defined as the difference between two condition specific precision matrices, was used for analysis. The assumption behind both types of DiffNetFDR analysis is that the data is sampled from two different normal distributions each with precision matrix $\Theta^{(k)}$ for $k = 1, 2$. This method states that for the $(i, j)^{th}$ entries of the precision matrices, $\theta_{i,j}^{(1)}$ and $\theta_{i,j}^{(2)}$, a differential edge exists if and only if $\theta_{i,j}^{(1)} - \theta_{i,j}^{(2)} \neq 0$. Therefore, the corresponding hypothesis statements are $H_{0,ij}: \theta_{i,j}^{(1)} = \theta_{i,j}^{(2)}$ verses $H_{1,ij}: \theta_{i,j}^{(1)} \neq \theta_{i,j}^{(2)}$ for $1 \leq i < j \leq p$. Like any other hypothesis test, a test statistic $T_{i,j}$ is constructed for each individual hypothesis $H_{0,ij}$. To construct these statistics, lasso estimators are used to construct initial estimations of regression coefficients. Here, a selection of the tuning parameter is required to control the sparsity level of the estimators. This selection is done adaptively to make the total number of false rejections and the estimated total of false rejections, $\frac{(2-2\Phi(t))(p^2-p)}{2}$, as close as possible, where p is the number of variables and $\Phi(t)$ is the standard normal cumulative distribution function. When $|T_{i,j}| > t$ for some threshold $t > 0$, the null hypothesis $H_{0,ij}$ will be rejected. The FDR (α) can be controlled by choosing a reasonable threshold t. For a target FDR, $0 \leq \alpha \leq 1$, t can be chosen through a multiple testing procedure:

$$\hat{t} = \inf\left\{0 \le t \le 2\sqrt{log\,p} : \frac{\frac{2\big(1 - \Phi(t)\big)(p^2 - p)}{2}}{R(t)\lor 1} \le \alpha\right\} \quad (3)$$

Where $R(t) = \sum_{1 \le i < j \le p} I\{|T_{i,j}| \ge t\}$ is the total number of rejections. If $\hat{t}$ does not exist, it is set at $\hat{t} = 2\sqrt{\log p}$. The differential network can then be constructed based on the set $\{(i,j): |T_{i,j}| \ge \hat{t}\}$.

2.3 Arabidopsis Data

The genetic expression data chosen for analysis was that of the *Arabidopsis* plant, a quick growing weed found in Eurasia and Africa ideal for genetic experimentation due to its short life cycle, small genome, and well documented genetic characterization. Raw data, as well as normalized data, was downloaded from the public repository, the Sequence Read Archive (SRA), NCBI (https://www.ncbi.nlm.nih.gov/sra), with Run Sector IDs from SRR1038651 – SRR1039336. The data consisted of the gene expressions of 32,833 genes over a series of time sequence trials from a total of 670 samples. The samples were from three treatments: Mock, Methyl Jasmonate (MeJa), and Salicylic Acid (SA). Each of these treatments represented a different condition and each had measurements taken on hourly intervals beginning at 1 hour through 8 hours, and then after every two hours until 12 hours. In addition to these time increments, data was also provided for measurements at 0.25 hours, 0.5 hours, 1.5 hours, and 16 hours. Each time increment had 16 sample measurements taken per gene, with the exception of the Mock treatment at 16 hours which only had 14 sample measurements (Table 2.1).

2.4 Groups of Samples

Since only data from two conditions/tissues can be used by JRmGRN and DiffNetFDR and the interest was in how the gene regulatory networks differ in different time periods and different conditions/tissues, the data was divided into two groups: early stage and late stage. The early stage was composed of data collected between 0.25 hours and 7 hours, while the late stage was composed of data collected between 8 hours and 16 hours. For each stage,

three pairs of conditions (Mock and MeJa, Mock and SA, and MeJa and SA) were used to construct the gene regulatory networks. Therefore, the JRmGRN and DiffNetFDR were used for six combinations of conditions and stages: Mock verses MeJa at early stage, Mock verses MeJa at late stage, Mock verses SA at early stage, Mock verses SA at late stage, MeJa verses SA at early stage, and MeJa verses SA at late stage.

Table 2.1. Number of samples per condition and time

Time (hrs)	Stage	Mock	MeJa	SA
0.25	Early	16	16	16
0.5	Early	16	16	16
1	Early	16	16	16
1.5	Early	16	16	16
2	Early	16	16	16
3	Early	16	16	16
4	Early	16	16	16
5	Early	16	16	16
6	Early	16	16	16
7	Early	16	16	16
8	Late	16	16	16
10	Late	16	16	16
12	Late	16	16	16
16	Late	14	16	16

2.5 Differentially Expressed Genes

Before any gene regulatory network reconstruction analysis could be conducted, a number of data pre-processing steps were performed. The first of these steps was to identify which of the 32,833 genes were differentially expressed. Genes that had mostly zero entries with a few nonzero entries would be identified as being highly differentially expressed despite having very low levels of gene expression (as indicated by the high proportion of zero entries). To guard against this, the number of zero entries for each gene, for each condition at each time point, was counted. If any gene was found to have more than 10% zero entries, the gene was removed from the analysis regarding that specific time and condition. Once these genes had been removed, the "edgeR" package in R was used to identify the level of differential expression in each gene at each time point for every combination of condition (i.e., Mock verses MeJa, Mock verses SA, and MeJa verses SA). The smaller the p-value

calculated by edgeR, the more likely that gene was differentially expressed. Conversely, large p-values indicated that genes were less likely differentially expressed, suggesting those genes would not provide as many insights into gene regulatory networks.

The next question was that of selecting a subset of genes to run through the JRmGRN and DiffNetFDR processes. For each combination of condition, each gene had a number of different p-values to be considered, corresponding to the multiple time points in either early or late stage. In order to combine these p-values into a single statistic that could be used to measure how likely a gene was to be differentially expressed, Fisher's method (also called Fisher's combined probability test) was applied. Fisher's method combines the p-values from n independent hypothesis tests to form a new statistic:

$$-2 \sum_{i=1}^{n} \ln(p_i) \tag{4}$$

When the n tests are independent and all n null hypotheses are true, the above test statistic has a chi-square distribution with n degrees of freedom. Because the p-values to be combined all originated from the same gene, the independence assumption of Fisher's method (equation 4) would not be satisfied. Therefore, the Fisher's test statistic from (4) for this study did not have a chi-square distribution when all null hypotheses were true. However, this was not a problem since the Fisher's test statistic was used mainly to rank genes so a set of genes could be selected for the subsequent network analysis.

In addition, since any zero valued p-value would instantly cause the Fisher test statistic to become undefined, any p-value of zero was replaced with a value of $1 * 10^{-50}$.

The genes for each of the six combinations were then ranked according to the associated Fisher test statistic and the 500 genes with the largest Fisher statistic were selected for analysis. Finally, the rank-based normal score transformation (also called the rank-based inverse normal transformation) (Altman, 1991) was applied to these 500 genes using the blom function from the R "rcompanion" package to ensure the data followed a standard normal distribution. Two matrices for each combination of condition and stage were formed: one matrix consisting of transformed genetic expression data for the selected 500

genes one condition, and one matrix of the transformed data from another condition at the same stage. This step resulted in 12 input matrices: two matrices for each of the six sets of top 500 genes for each treatment. These matrices were then used in JRmGRN and DiffNetFDR to construct the gene regulatory networks.

Before running these matrices through the JRmGRN or DiffNetFDR code, the intersections and unions of the six sets of genes were examined. Pairwise intersections and unions were conducted, as well as intersections and unions for all six sets of genes together. Large intersections would indicate a set of genes was highly expressed in several conditions and stages whereas small intersections would indicate different genes are expressed throughout different conditions and stages.

2.6 JRmGRN

At this stage, the only step left to complete for the JRmGRN method was the selection of tuning parameters. Based on the methods applied by Deng *et al.* (2018), the four tuning parameters were constructed from a space similar to a uniformed grid of log space, with λ_1 taking values 0.01, 0.05, 0.10, 0.25, 0.50 and 1, λ_2 becoming 0.5, 1, and 2 factors of λ_1, λ_3 becoming 0.5, 1, 2, 3, and 4 factors of λ_1, and λ_4 becoming 0.1, 0.5, 1, 2, and 4 factors of λ_1. Thus, for each value of λ_1, there were 75 possible combinations of tuning parameters and a total of 450 possible combinations of parameters.

If the values of the tuning parameters are too large, the penalties will overpower the function itself; conversely, if the tuning parameters are too small, they will not be significant enough to produce any notable effect on the optimization function. Running the JRmGRN process for all 450 combinations of parameters allowed for an adequate set of informative results to select from.

For each combination of condition and stage, two matrices were then run through JRmGRN for each of the 450 combinations of tuning parameters and a BIC value for each combination was calculated. The set of tuning parameters selected for the final analysis of the early stage Mock verses MeJa and Mock verses SA conditions were $\lambda_1 = 80$, $\lambda_2 = 40$, $\lambda_3 = 40$, and $\lambda_4 = 320$, and $\lambda_1 = 16$, $\lambda_2 = 16$, $\lambda_3 = 48$, and $\lambda_4 = 64$ for the early

stage MeJa verses SA condition. For the late stage Mock verses MeJa and Mock verses SA conditions, the set of tuning parameters selected for the final analysis were $\lambda_1 = 16$, $\lambda_2 = 16$, $\lambda_3 = 32$, and $\lambda_4 = 64$. For the late stage MeJa verses SA condition, the tuning parameters used were $\lambda_1 = 8$, $\lambda_2 = 16$, $\lambda_3 = 24$, and $\lambda_4 = 32$.

2.7 DiffNetFDR

To run the DiffNetFDR analysis, the R function DiffNet.FDR in the "DiffNetFDR" package with the difference of precision matrices method was used. The desired false discovery rate was set at 0.01. Since the DiffNet.FDR function only used one matrix of gene expression data as input, the two rank-based normal score transformed matrices for each combination of condition and stage were combined into one large matrix with columns corresponding to genes and rows corresponding to the treatment and time of each condition. A vector was created to identify the treatment of each condition and was also used as input for the DiffNet.FDR function.

2.8 Gene Ontology Enrichment Analysis

Gene ontology enrichment analysis is a commonly used method to find which gene ontology terms are over-represented (or under-represented) for a given set of genes. Such analysis can greatly facilitate the understanding of underlying biological processes. To perform the gene ontology enrichment analysis, the gene ID's for the top 500 differentially expressed genes from each comparison were obtained and then entered into the Generic Gene Ontology (GO) Term Finder (https://go.princeton.edu/cgi-bin/GOTermFinder/). Tree images were generated using the biological process ontology aspect with the TAIR – A. thaliana (Common wallcress) annotation. In addition to the three images, the GO Term Finder also generated a table of frequent gene ontology terms and the respective cluster frequency. According to the finder, the *p*-value cutoff was set at 0.01 and Bonferroni correction was used for multiple testing correction.

3 Results

The process to remove genes from analysis due to high levels of zero entries resulted in between 50.1% and 51.0% of genes being removed from analysis for each of the six comparisons (Table 3.1). Specifically, 16,461 of the 32,833 genes were removed from the analysis of early stage Mock verses MeJa, 16,634 genes were removed from early stage Mock verses SA, 16,742 genes were removed from early stage MeJa verses SA, 16,469 genes were removed from late stage Mock verses MeJa, 16,645 genes were removed from late stage Mock verses SA, and 16,742 genes were removed from late stage MeJa verses SA.

Table 3.1. Number of genes removed from analysis for each condition and stage

Stage	Mock v MeJa	Mock v SA	MeJa v SA
Early	16,461	16,634	16,742
Late	16,469	16,645	16,742

After examining the pairwise intersections and unions of the six sets of top 500 differentially expressed genes identified by the Fisher test statistics (Figure 3.1 and Table 3.2), it was noted that the overlap between treatment pairs (Table 3.2) was minimal. Only three genes were shared between all six combinations of conditions and stages, while the union of all six combinations contained 1,668 genes. The intersections among the early stage conditions and the late stage conditions (Figure 3.1) were also not strong.
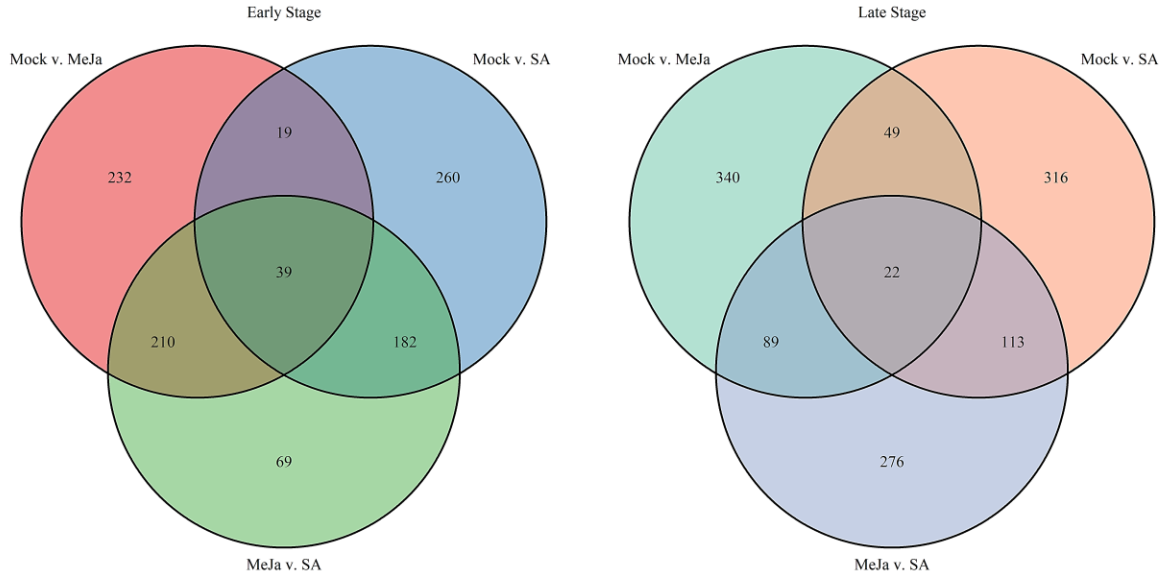
Figure 3.1. Left: Venn diagram of top 500 differentially expressed genes for treatment combinations for early stage. Right: Venn diagram of top 500 differentially expressed genes for treatment combinations for late stage.

Table 3.2. Pairwise intersections and unions between combinations of conditions and stages. Intersections are marked in red, and unions are marked in blue.

		Early Stage			Late Stage		
		Mock v. MeJa	Mock v. SA	MeJa v. SA	Mock v. MeJa	Mock v. SA	MeJa v. SA
Early Stage	Mock v. MeJa	500	58	249	253	30	93
	Mock v. SA	942	500	221	70	154	175
	MeJa v. SA	751	779	500	172	57	191
Late Stage	Mock v. MeJa	747	930	828	500	71	111
	Mock v. SA	970	846	943	929	500	135
	MeJa v. SA	907	825	809	889	865	500

Upon running the early stage Mock verses MeJa and early stage Mock verses SA data through all 450 possible selections of tuning parameters with the JRmGRN code in R, it was noticed that many combinations of tuning parameters provided the same, or very

14

similar, results regarding BIC and the number of hub genes identified. Additionally, the majority of tuning parameter combinations identified all 500 genes as hub genes or no genes as hub genes. For a set of λ_1, λ_3, and λ_4, changes in the value of λ_2 often had no impact on the BIC or the number of hub genes identified. Since these trends were replicated with two different sets of input matrices and condition combinations, these results were used to reduce the number of possible tuning parameter selections necessary to apply to each condition combination. Instead of using all 450 possible selections, only those selections in which $\lambda_2 = 16$ or $\lambda_2 = 40$ were chosen, reducing the number of tuning parameter combinations to 100. Of those 100 sets, a subset of the most informative 20 was selected and the final set of parameters was chosen from this subset. This procedure significantly reduced the run time needed for each condition combination.

With six condition combinations, carefully examining the results of each would be very extensive and beyond the scope of this report. Therefore, the results from MeJa verses SA at the early stage and late stage were the focus for the rest of the analysis as they were believed to be the most interesting and informative results.

3.1 Differentially Expressed Genes

There were 191 genes that were differentially expressed between MeJa and SA at both the early and late stages (Table 3.2). No significant GO terms were found for these 191 genes; however, the GO finder did identify some significant terms for the top 500 differentially expressed genes for these two combinations. For the early stage, these significant terms fell into the overarching categories of cellular processes and response to cadmium ion, while the overarching terms for the late stage were cellular processes, organonitrogen compound metabolic processes, and cellular metabolic processes. The tree graphs constructed by the GO finder for the early and late stages of the MeJa verses SA conditions are below.
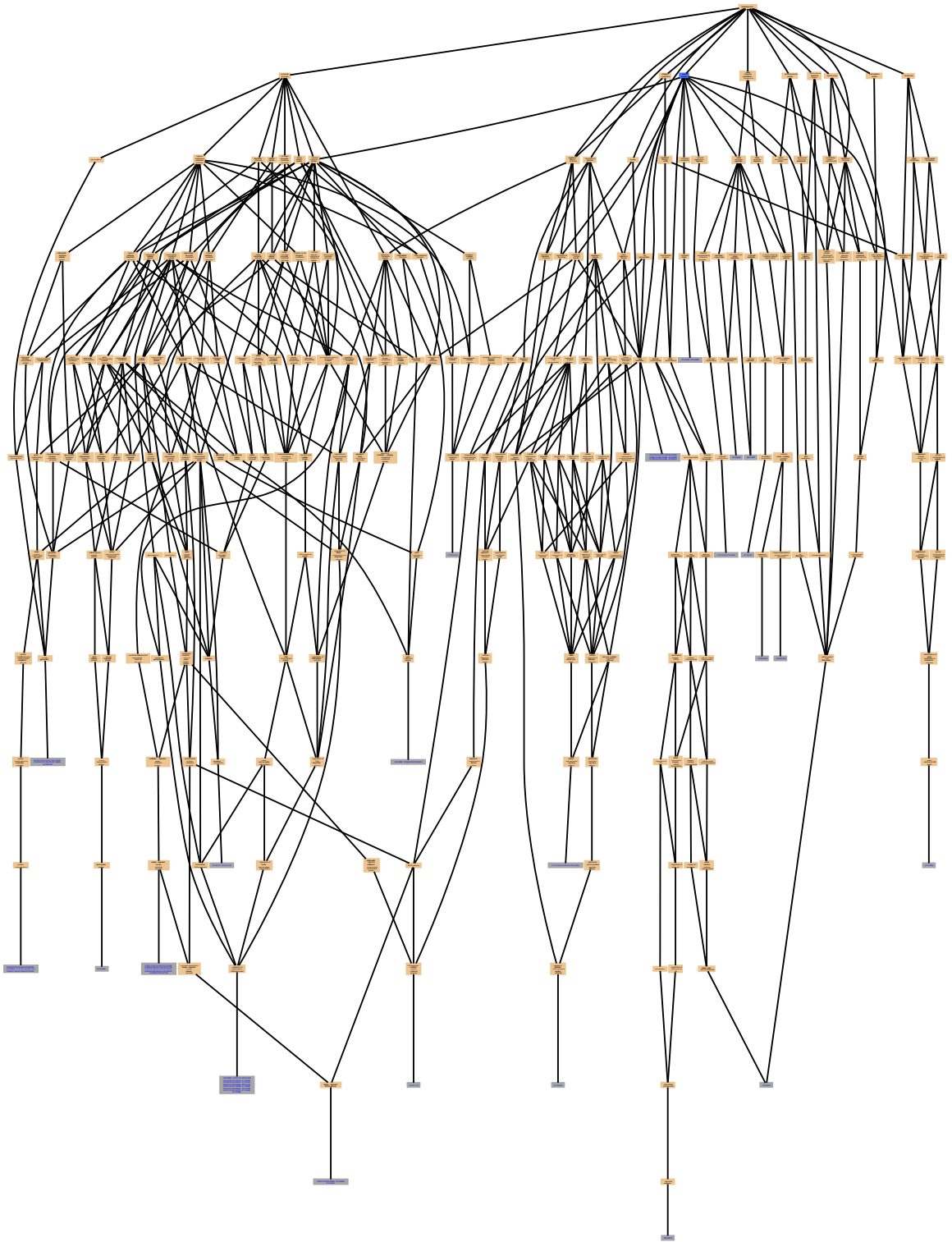
Figure 3.2. Early stage MeJa verses SA gene ontology tree image. Text in blue represents specific genes; text in black corresponds to biological processes.
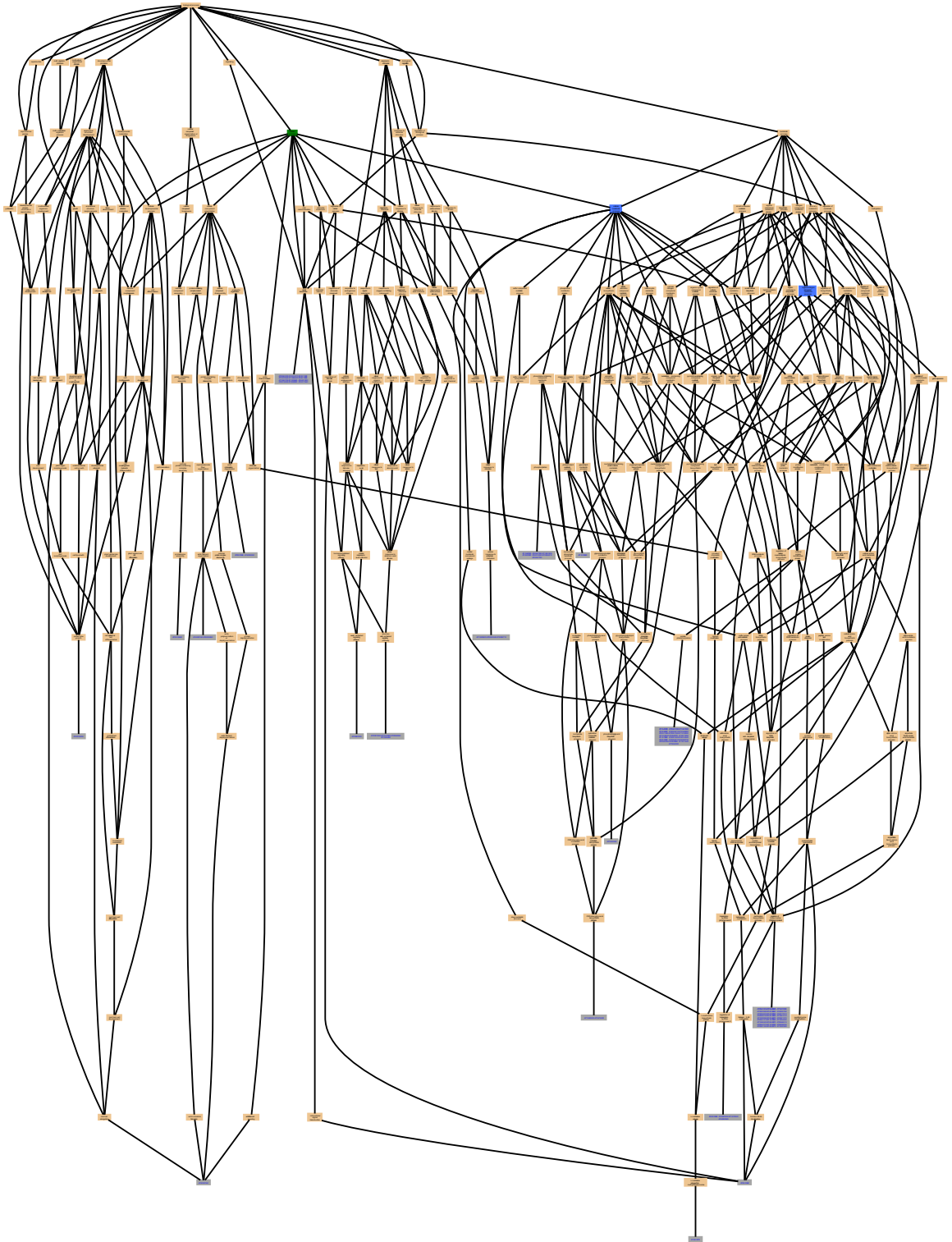
Figure 3.3. Late stage MeJa verses SA gene ontology tree image. Text in blue represents specific genes; text in black corresponds to biological processes.

3.2 JRmGRN Results

Based on the JRmGRN results, there were 46 hub genes in the early stage MeJa verses SA network (Table 3.3). These hub genes had very high connectivity, as demonstrated by the corresponding 2,470 edges. Each hub gene had between 30 and 149 connections, thus connecting to between 6.0% and 29.8% of the total genes. The median number of connections for these 46 hub genes was 55. For the late stage MeJa verses SA condition, only 13 hub genes were identified. Each of these hub genes was connected to 39 to 169 other genes, approximately 7.8% to 33.8% of the total genes. The median number of connections for these hub genes was 56.

The condition specific networks for MeJa and SA had a similar number of hub genes for both the early and late stages (Table 3.3), though these genes were not necessarily the same for both conditions. For example, the 122 genes identified as significant in the early stage MeJa network were not necessarily the same as the 119 genes identified as significant in the early stage SA network. The condition specific networks had far fewer edges than the common hub networks, as anticipated.

Table 3.3. Network summary for MeJa verses SA conditions.

Stage	Network	Genes	Edges
Early	MeJa	122	231
Early	SA	119	214
Early	Hub	46	2470
Late	MeJa	55	87
Late	SA	55	85
Late	Hub	13	786

The identified networks were graphed using tools in Cytoscape version 3.8.2. The networks for the MeJa condition, SA condition, and common hubs were all graphed separately, in addition to being graphed together (Figure 3.4 through Figure 3.11).
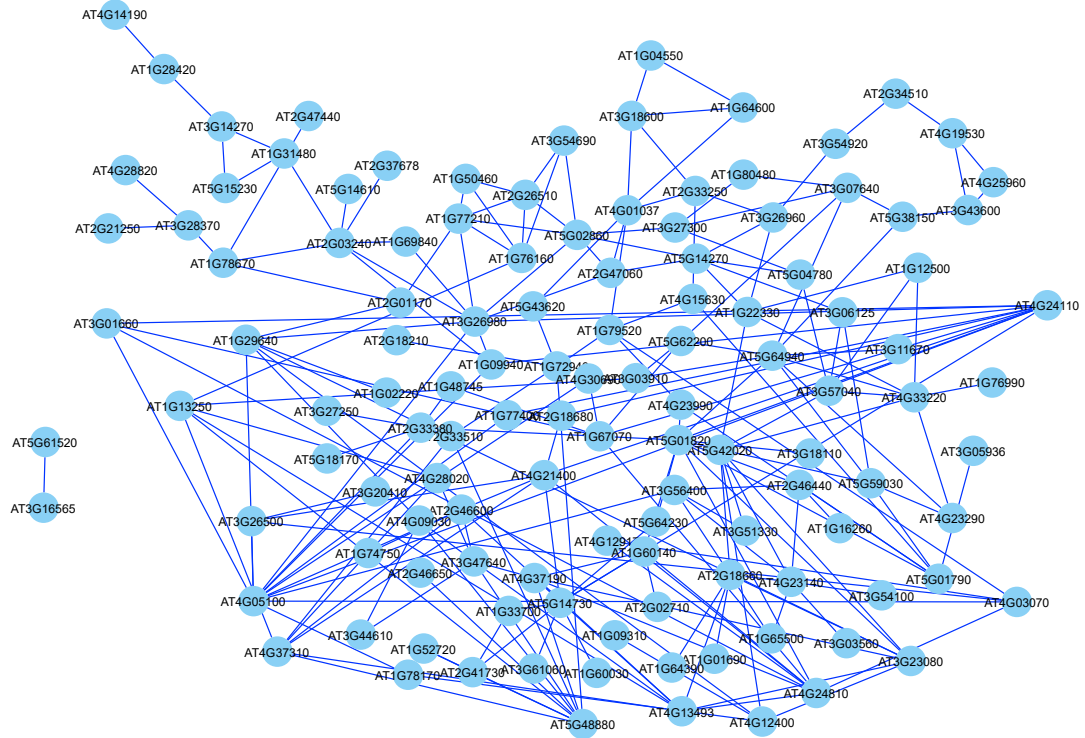
Figure 3.4. Network from MeJa at the early stage.

Figure 3.5. Network from SA at the early stage.

Figure 3.6. Network for common hub genes of MeJa and SA at the early stage.
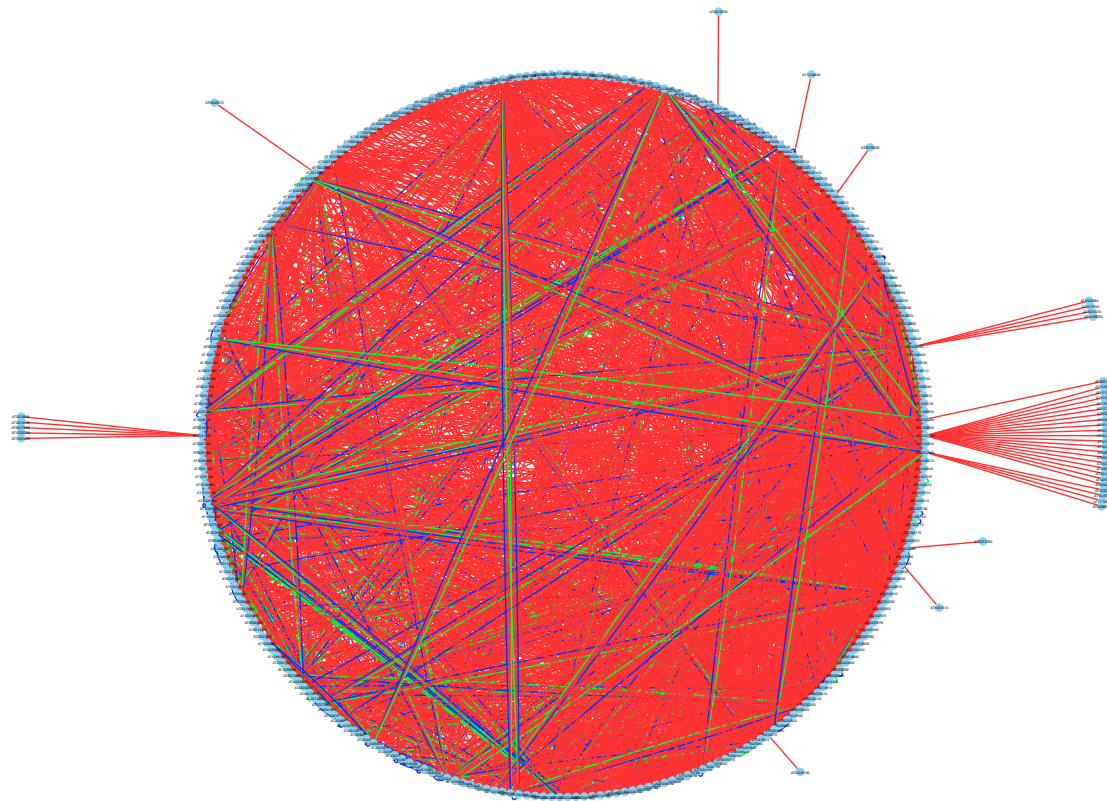
Figure 3.7. Network from the early stage. The edges in blue, green, and red are from MeJa, SA, and the common hub genes, respectively.

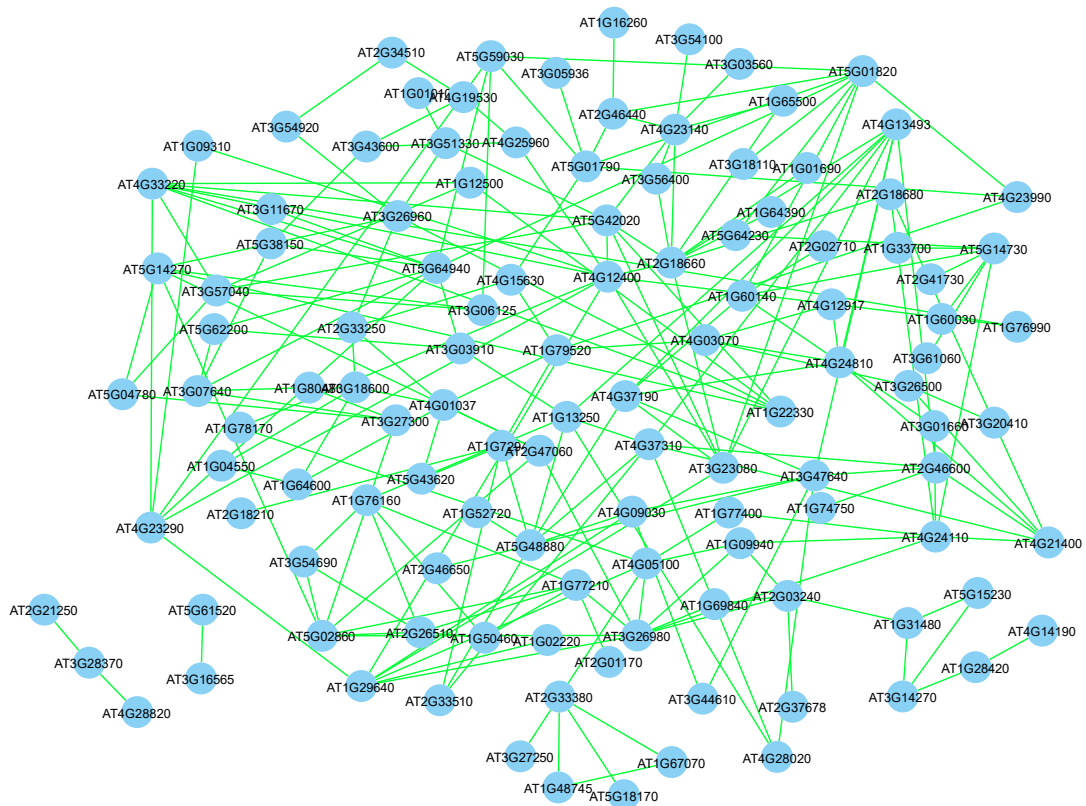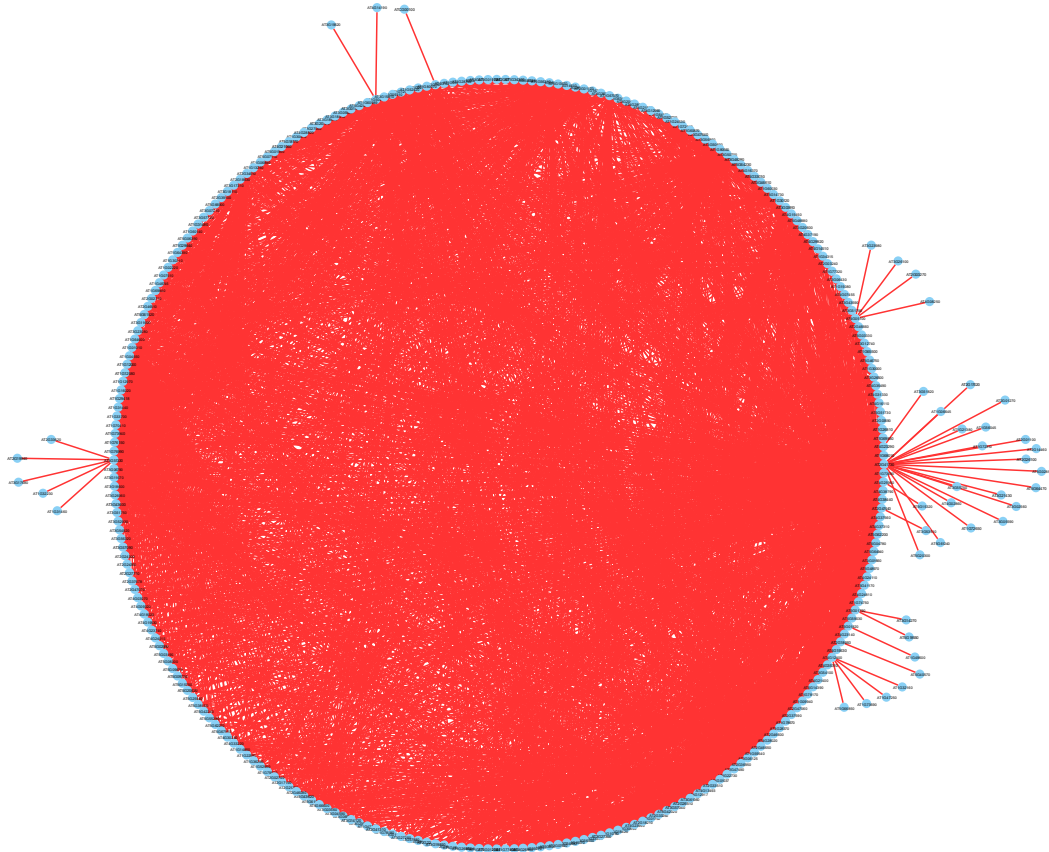**Figure 3.8**. Network from MeJa at the late stage.

Figure 3.9. Network from SA at the late stage.
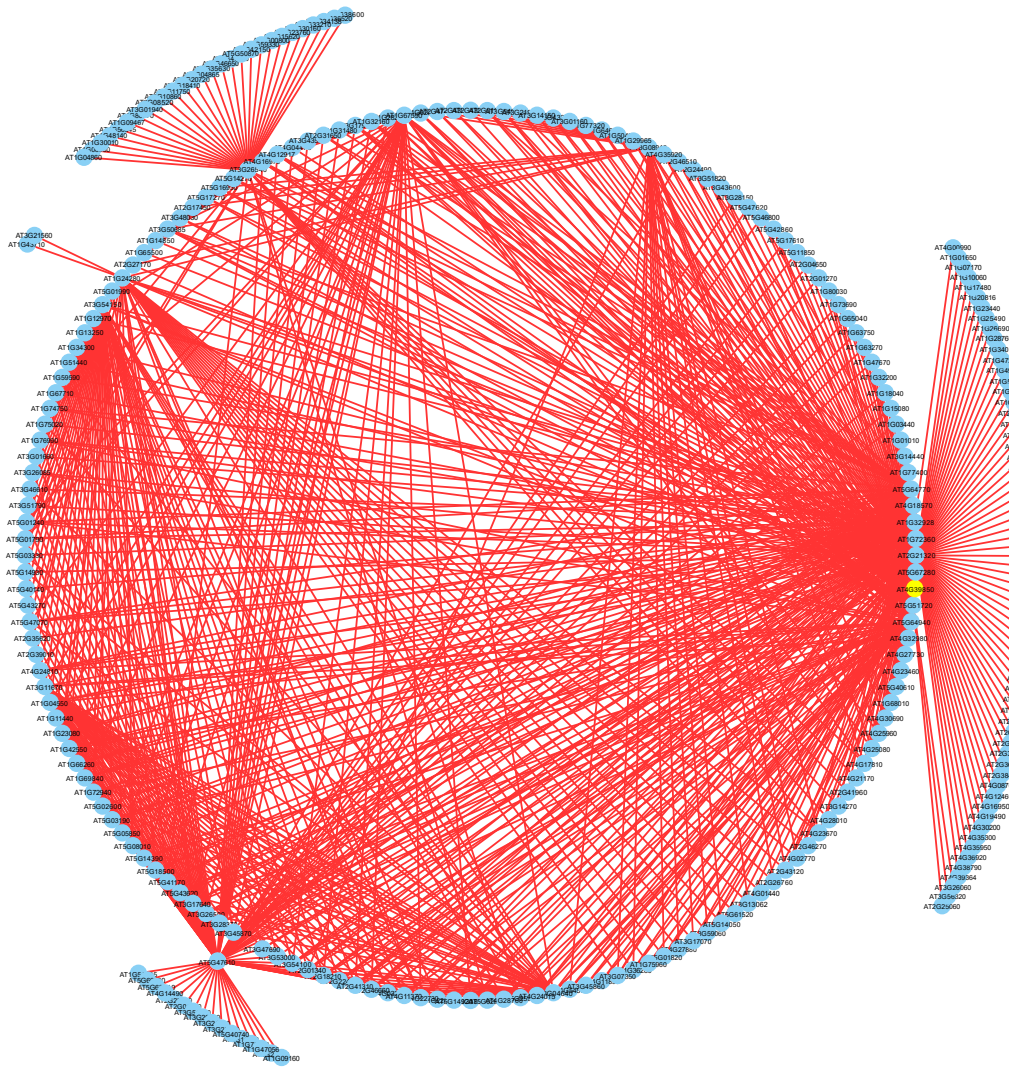
Figure 3.10. Network for common hub genes of MeJa and SA at the late stage.
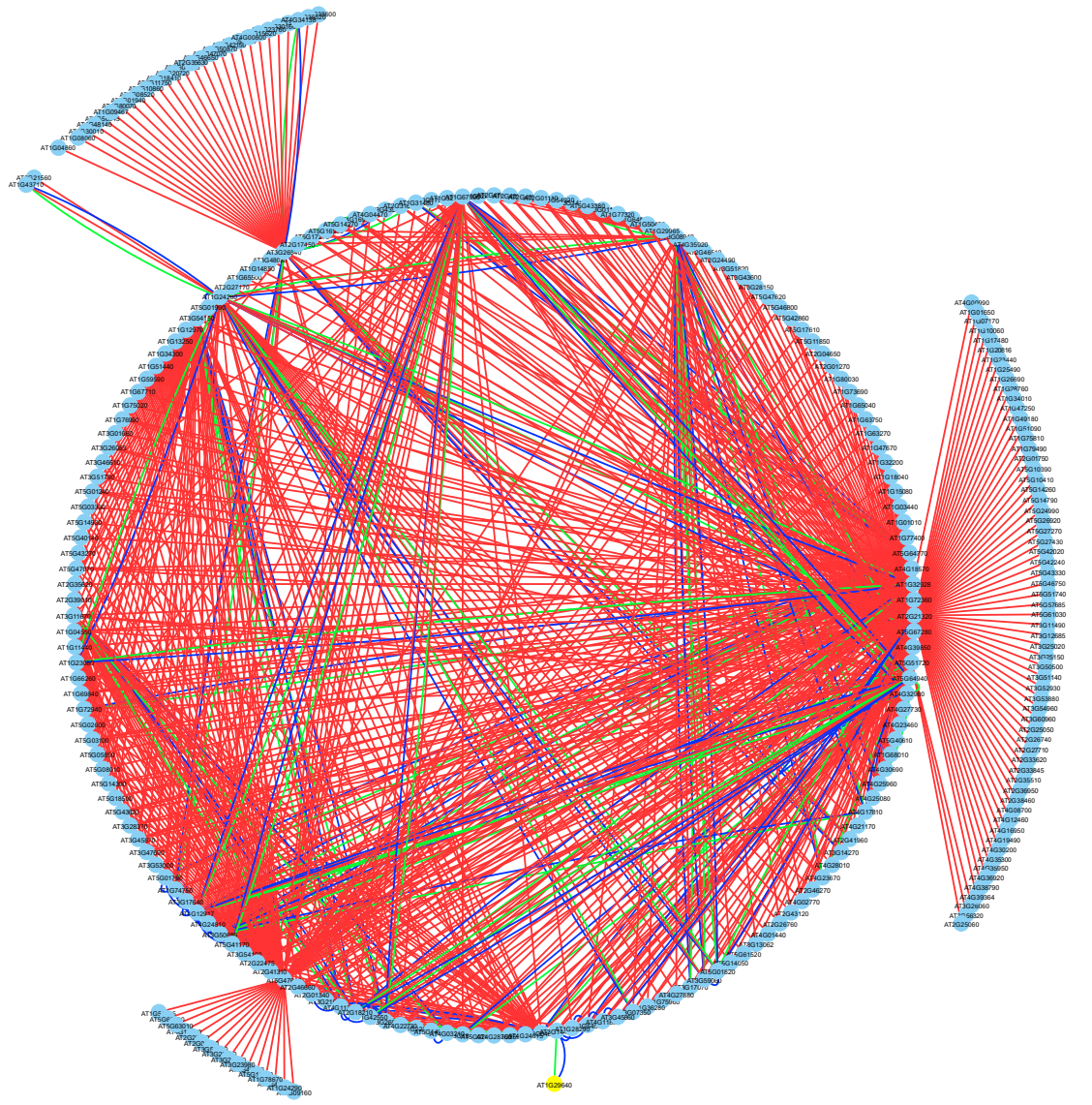
Figure 3.11. Network from the late stage. The edges in blue, green, and red are from MeJa, SA, and the common hub genes, respectively.

3.3 DiffNetFDR Results

The differential networks identified by the DiffNetFDR method were significantly different from those identified from the JRmGRN method. There were only two networks: one network for the early stage MeJa verses SA comparison and one network for the late stage MeJa verses SA comparison. These networks are genes and edges that appear in only one of the condition specific networks, essentially the networks that are not common to

both. These networks have been plotted alongside the results from the JRmGRN method, allowing for an easy comparison of the identified networks (Figure 3.12 and Figure 3.13). The following figures were also graphed using Cytoscape version 3.8.2.
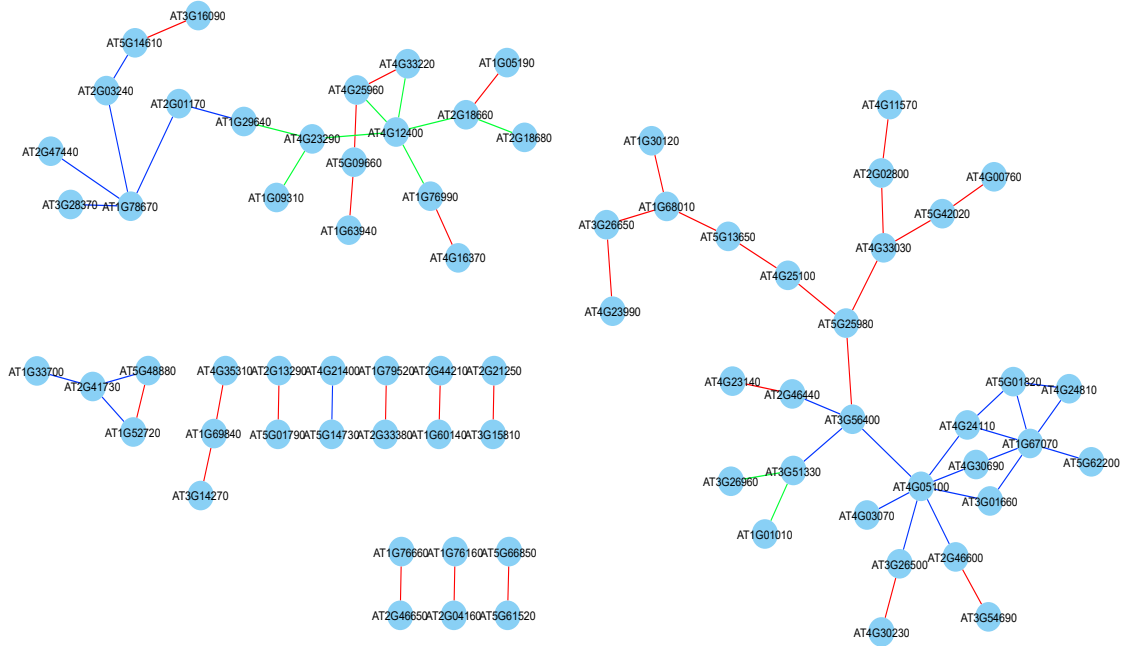
Figure 3.12. Differential network at the early stage. The edges in blue, green, and red are the unique edges for MeJa identified using JRmGRN, the unique edges for SA identified using JRmGRN, and the unique edges for MeJa or SA identified using DiffNetFDR, respectively.
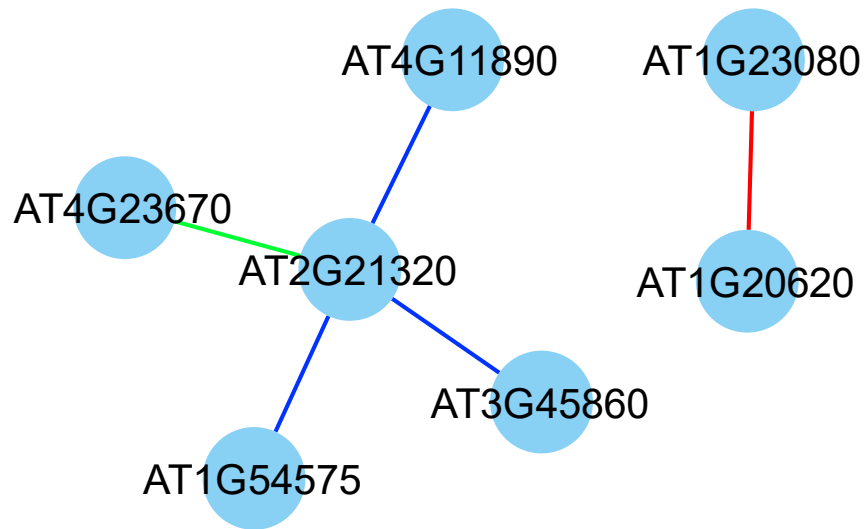
Figure 3.13. Differential network at the late stage. The edges in blue, green, and red are the unique edges for MeJa identified using JRmGRN, the unique edges for SA identified using JRmGRN, and the unique edges for MeJa or SA identified using DiffNetFDR, respectively.

4 Discussion

Background on MeJa and SA Conditions

Two of the major pathways involved in induced defense of plants against pathogens include the salicylic acid (SA) – and jasmonic acid (JA) – mediated pathways that are thought to act primarily against biotrophs and necrotrophs, respectively (Klemme *et al.*, 2019).

Jasmonates (JAs) are a class of plant hormones that play essential roles in response to tissue wounding. JAs are known to be synthesized through the octadecanoids pathway (the starting point is the 18 C fatty acid α-linolenic acid 18:3 (α-LA)) (Schaller, 2001). They act on gene expression to slow down plant growth and development and to redirect metabolism towards producing defense chemicals and repairing damages. At present, it is known that JAZs and MYCs are the targets of JA signaling and it is assumed that cell cycle progression genes are located further downstream (Pauwels *et al.*, 2009). In recent years, JAs have been found to interact with many other plant hormones and thus have essential functions to affect plant growth and development. For example, JAs restrain leaf growth by inhibiting G1/S transition through its receptor COI1, thereby reducing both cell number and size (Noir *et al.*, 2013). Originally identified as a major component of fragrant oils, methyl jasmonate (MeJa) and its precursor jasmonic acid (JA) were first demonstrated to promote senescence in detached oat (Avena sativa) leaves (Ueda & Kato, 1980) and were subsequently shown to be a type of plant growth regulators that play pervasive roles in several other aspects of plant development, which include, but are not limited to, seed germination and pollen development, responses to mechanical and insect wounding, pathogen infection, and drought stress (Schaller, 2001).

Among various stress-related molecular components, the defense hormone salicylic acid (SA) has been implicated in plant stress response toward an array of abiotic and biotic stresses that include, but are not limited to, drought, salinity, cold, high light, and microbial pathogens (Shah, 2003; Scott *et al.*, 2004; Yang *et al.*, 2004; Mateo *et al.*, 2006; Lee *et al.*, 2010; Wan *et al.*, 20212; Miura *et al.*, 2013). Despite the pivotal role of SA in plant stress responses, when it becomes constitutively active, the spatiotemporal SA-mediated signaling generally leads to growth inhabitation and cell death. The study on *Arabidopsis*

lesion-mimic mutants, which manifested lesions associated with increased cellular SA contents and constitutive activation of SA-mediated signaling, implicated that some molecular components might counteract this pathway (Dietrich *et al.*, 1994).

As stated above, jasmonic acid (JA) and salicylic acid (SA) play important roles in plant defense systems. JA and SA signaling pathways interact antagonistically in dicotyledonous plants. Significant progress has been made in identifying the key components and understanding the role of salicylic acid (SA) and jasmonates (JA) in plant responses to biotic and abiotic stresses. Recent studies have shown that other hormones such as abscisic acid (ABA), auxin, gibberellic acid (GA), cytokinin (CK), and brassinosteroids (BR) also interact with JA and SA and play a role in plant defense signaling pathways, but their roles in plant defense is thus far not well elucidated (Yang *et al.*, 2019). The SA- and JA-mediated signaling pathways were thought to be the backbone of plant immune responses against biotic stresses or wounds, whereas ABA, auxin, BR, GA, and CK were considered to be involved in the plant immune response through modulating the SA-JA signaling pathways (Li *et al.*, 2019). In addition, JA and SA have been shown to activate a common defense system in rice (Tamaoki *et at.*, 2013). However, the study based on microarray analysis showed that more than half of the genes were upregulated by the SA analog BTH and were also upregulated by JA, which needs to be further and more carefully characterized in monocots before a firm conclusion is drawn.

Notable Hub Genes in Networks

The network analysis using the implemented JRmGRN and DiffNetFDR methods led to the identification of the following hub genes that have high connectivity in the constructed networks:

AT5G48880 encodes peroxisomal 3-keto-acyl-CoA thiolase 2, which is an enzyme that catalyzes the last step of JA biosynthesis: the conversion of OPC-8:0 to JA (He *et al.*, 2002). This gene is a common hub in the joint network using the early stage data after treatment (Figure 3.6), indicating it had some common edges in the joint network built with data from the MeJa and the SA treatments. In addition, this gene had some unique edges in the MeJa network (Figure 3.4) and the SA (Figure 3.5) network.

28

AT4G19530 encodes a TIR-NB-LRR resistance protein. Transient expression of this gene in tobacco induces cell death (Swiderski *et al*., 2009). This is a mechanism through which plants get rid of the infected cells and tissues before the biotic invaders are transmitted to heathy tissues or organs. This gene was included at the early stage but not in the late stage networks. This gene was a common element in both the MeJa and SA networks and was also a common gene in the unified network (Figure 3.4 through Figure 3.7), indicating that it can be activated by both SA and JA for defending.

AT4G23290 encodes a NRG1 protein involved in the EDS1-SAG101-NRG1 module which mediates cell death signaling by TIR-domain immune receptors (Lapin *et al.*, 2019). EDS1 (enhanced disease susceptibility1) encodes a lipase-like protein, which forms EDS1 heterodimers with PAD4 (phytoalexin deficient 4) to boost basal defense pathways. EDS1 interacts with PAD4-related SAG101 (senescense-associated gene101), but the roles of EDS1-SAG101 heterodimers function together with NRG1 coiled-coil domain helper NLRs as a coevolved TNL cell death signaling module. EDS1-SAG101-NRG1 cell death signaling activity is verified in the solanaceous species, and *Nicotiana benthamiana*. This gene showed up in the joint network of MeJa and SA at the early stage with eight edges and its connectivity was reduced to only one by the late stage. It did not show up in the MeJa and SA specific networks regardless which stage (Figure 3.6 and Figure 3.10).

AT3G56400 encodes a WRKY70, which was a common gene in the joint network of early stage, indicating that it was activated by both MeJa and SA (Figure 3.6). It also showed in the MeJa and SA specific networks, indicating that it had some unique regulatory targets in either the MeJa (Figure 3.4) or SA (Figure 3.5) network. *Arabidopsis* WRKY75 coordinates with WRKY46 and WRKY53 in basal resistance against pathogen Pseudomonas syringae (Hu *et al.*, 2012).

AT2G18660 encodes a plant natriuretic peptide A.- a functional analog to vertebrate atrial natriuretic peptides that appears to antagonize the SA-mediated plant stress responses, while loss of the receptor potentiates SA-mediated signaling (Lee *et al.*, 2020). This gene was a common gene in the joint early stage network (Figure 3.6) and also showed in the

MeJa (Figure 3.4) and SA (Figure 3.5) specific networks, indicating that it can be induced by both MeJa and SA.

AT4G03070 encodes 2-oxoglutarate (2OG) and FE(II)-dependent oxygenase involved in glucosinolate biosynthesis pathway (Kliebenstein *et al.*, 2001b). Plant tissue that is subjected to damage results in glucosinolate hydrolysis and formation of unstable inter-mediates that can be rearranged to an array of chemicals including isothiocyanates, thiocyanates, nitriles, oxazolidine-2-thiones, or epithioalkanes. Some of these products are biologically active and have been implicated in plant defense against pathogens and herbivores (Kliebenstein *et al.*, 2001a; Lambrix *et al.*, 2001). The gene was a common gene in the joint network (Figure 3.6) and also in the MeJa specific network (Figure 3.4) and SA specific network (Figure 3.5).

Relatively speaking, few hub genes were identified from the late stage. This is because there are fewer marker genes known for the late response. However, some genes with the defense functions were still identified. For example, AT2G21320 encodes a B-box 18 (BBX18), which is an important thermomorphogenesis regulator in *Arabidopsis* (Ding *et al.*, 2018). BBX18 and BBX23 mutations lead to reduced thermoresponsive hypocotyl elongation. In contrast, BBX18 overexpression promotes hypocotyl growth at elevated temperatures. This gene did not show up in the early networks; instead, it showed up in the late networks. It had approximately 168 connections in the common network (Figure 3.10), four in the MeJa network (Figure 3.8), and two in the SA network (Figure 3.9), indicating it is mainly a common hub regulator that controls growth downstream of MeJa and SA signaling.

There was also one gene, AT1G72940, which showed up in both the early and late networks. It had approximately 50 connections in the early joint network (Figure 3.6) and seven connections in the late joint network (Figure 3.10). It encodes nucleotide-binding, leucine-rich repeat (NLR) genes to prevent autoimmunity. Its expression can be induced by Pseudomonas syringae inoculation and exogenous SA treatment (Nasim *et al.*, 2020). Mutations in AT1G72940 resulted in increased susceptibility to Pseudomonas syringae,

whereas their overexpression resulted in severely stunted growth (Nasim *et al.*, 2020). Therefore, it is a gene downstream from the JA and SA signaling.

Not all genes in the resulting networks are exhaustively listed above. Some median-level hub genes that were not super-hubs are not listed. But these genes provided above are sufficient to demonstrate that the methods implemented have efficacy to identify hub genes functioning in two conditions.

Comparison of the JRmGRN and DiffNetFDR Methods

There are three outputs for the JRmGRN method: a network of common hub genes, a network of genes for the first condition, and a network of genes for the second condition. The output of the DiffNetFDR method is one network of connections that occur in only one of the two specific conditions. Because of these differences, a direct comparison between the two methods is not possible. However, comparing the differences of the JRmGRN condition specific networks with the results from DiffNetFDR is possible. Due to these methods having different approaches and the lack of known results, the results from these two methods are not expected to be exactly the same, or necessarily very similar. Ideally there would be some similarities. When comparing the results from the JRmGRN and DiffNetFDR methods for this set of *Arabidopsis thaliana* data, no similar networks were identified, only several genes that appeared in both networks (Figure 3.12 and Figure 3.13). Currently, there is no known way to determine which results are more accurate, or if both are correct. From a biological standpoint, aspects of the networks constructed through both methods appear to be reasonable. Though these results are not ideal, they are a reminder that the process of identifying gene regulatory networks is complicated and has plenty of room for development.

Comments on the JRmGRN Method

Perhaps the most insightful aspect of this project was gaining a better understanding of the background necessary to implement the JRmGRN method and the complications related to running real data with this approach. The first challenge was creating a procedure to correctly identify differentially expressed genes to be used as input. Though not directly a part of the JRmGRN method, this step was the first of many setbacks addressed before

being able to get functional results. The original data had to be subdivided into several subsets and each subset had to be considered individually to find which genes were more or less likely to be differentially expressed. These subsets were determined according to the conditions that were to be compared using the JRmGRN method. Because the entire data set could not be run through the available JRmGRN code, these subsets were necessary. This step of identifying differentially expressed genes was not difficult but added an extra step to the analysis.

The most troublesome aspect was that of tuning parameter selection. Although Deng *et al.* (2018) provided some guidelines for tuning parameter selection, these guidelines were vague and left much open to experimentation. Several sets of tuning parameters had to be experimented with before finally settling on a group that seemed sufficient. For example, no limit was set on how large or small a parameter could be, aside from being a positive real number greater than zero. After some testing, it was found that large parameters were mostly uninformative, and smaller parameters were more informative. For this particular set of data, parameter values above 400 produced uninformative results, while the actual parameter values used were significantly smaller. However, this upper bound may not apply to all datasets or all parameters. In several instances, for a set of $\lambda_1, \lambda_3,$ and λ_4, varying the value of λ_2 had no effect on the results. Since the code run time for JRmGRN with 500 genes was approximately 30 minutes, experimentation with various tuning parameters was time costly. Once a set of 450 tuning parameters was generated based on previous experimentation, that set was further reduced to a smaller group of only 20 tuning parameters according to the number of hub genes identified by JRmGRN for the respective parameters. This process was time consuming and rather frustrating, particularly since the information learned might not apply to other datasets. Until an efficient and successful method of tuning parameter selection is established that can be applied to any data set, the benefits of the JRmGRN method might not be worth the work.

Other more technical difficulties included the BIC-like criteria returning a value of negative infinity for many iterations of the JRmGRN code and the Z matrices identifying no significant condition specific networks for certain sets of tuning parameters. Any

commonality between these two issues and the tuning parameters used was not clear and is an area that should be explored in the future.

Implementing the JRmGRN method and selecting a useful set of results with a statistical or biological background only would prove challenging as a solid understanding of both is necessary. So, despite Deng *et al.* (2018) demonstrating the strength and accuracy of the JRmGRN method, the useability of this method is not ideal, potentially decreasing its value and strength.

Comments on the DiffNetFDR Method

The DiffNetFDR method, on the other hand, lived up to its name as a user-friendly approach to creating differential networks in an uncomplicated manner. Code run times for this method were only a couple of minutes for the data used, allowing for easy experimentation with different false discovery rates or testing types. However, one notable challenge was that the program would not run with the rank-based normal score transformed data for certain combinations of stage and condition and had to be run with the original data. The cause of this issue was not discovered. The DiffNetFDR method is not as comprehensive as the JRmGRN method, but still functions very well at what it was designed for.

5 Conclusion

When applied to *Arabidopsis thaliana* genetic expression data, both the JRmGRN and DiffNetFDR methods reconstructed reasonable gene regulatory networks. Despite the JRmGRN method being more complex and challenging, the results obtained for the early and late stage MeJa verses SA comparisons provided insights into networks common to both conditions and networks unique to each specific condition. The DiffNetFDR method also successfully reconstructed networks for the early and late stages of MeJa verses SA, highlighting connections that were unique to just one condition. Though very simple to apply, the DiffNetFDR does not provide as many insights into the gene regulatory networks of the *Arabidopsis thaliana* and does not provide the same well-rounded results as the JRmGRN method does. These methods could potentially work to confirm one another or to support the results from one another, but based on the results obtained through this project, the JRmGRN and DiffNetFDR methods might be more suitable to complement one another. Each method highlights certain featured structures and components within the GRN, building off each other to give a well-developed picture of the networks at play.

6 References

Atlman, DG. (1991) Practical statistics for medical research. *Chapman and Hall*.

Deng, W. *et al*. (2018) JRmGRN: joint reconstruction of multiple gene regulatory networks with common hub genes using data from multiple tissues or conditions. *Bioinformatics,* 34, 3470-3478.

Ding, L. *et al.* (2018). Two B-box domain proteins, BBX18 and BBX23, interact with ELF3 and regulate thermomorphogenesis in Arabidopsis. *Cell Rep,* 25(7):1718-1728 e1714.

He, Y. *et al.* (2002) Evidence supporting a role of jasmonic acid in Arabidopsis leaf senescence. *Plant Physiol,* 128(3): 876-884.

Hu, Y. *et al.* (2012) Arabidopsis WRKY46 coordinates with WRKY70 and WRKY53 in basal resistance against pathogen Pseudomonas syringae. *Plant Sci,* 185-186: 288-297.

Klemme, S. *et al.* (2019) Selection of salicylic acid tolerant epilines in brassica napus. *Agronomy,* 9(92): doi:10.3390.

Kliebenstein, DJ. *et al.* (2001a) Genetic control of natural variation in Arabidopsis glucosinolate accumulation. *Plant Physiol,* 126(2): 811-825.

Kliebenstein, DJ. *et al.* (2001b) Gene duplication in the diversification of secondary metabolism: tandem 2-oxoglutarate-dependent dioxygenases control glucosinolate biosynthesis in Arabidopsis. *Plant Cell,* 13(3): 681-693.

Lambrix, V. *et al.* (2001) The arabidopsis epithiospecifier protein promotes the hydrolysis of glucosinolates to nitriles and influences Trichoplusia in herbivory. *Plant Cell,* 13(12): 2793-2807.

Lapin, D. *et al.* (2019) A coevolved EDS1-SAG101-NRG1 module mediates cell death signaling by TIR-domain immune receptors. *Plant Cell,* 31(10): 2430-2455.

Lee, KP. *et al*. (2020) PLANT NATRIURETIC PEPTIDE A and its putative receptor PNP-R2 antagonize salicylic acid-mediated signaling and cell death. *Plant Cell,* 32(7): 2237-2250.

Li, N. *et al*. (2019) Signaling crosstalk between salicylic acid and tthylene/jasmonate in plant defense: Do we understand what they are whispering? *Int J Mol Sci,* 20(3).

Lin, Y.-H. *et al*. (2007) Enhancement of ferromagnetic properties in Bi Fe O 3 polycrystalline ceramic by La doping. *Appl. Physics Lett.*, 90, 172507.

Lui, W. (2017) Structural similarity and difference testing on multiple sparse Gaussian graphical models. *Ann. Stat.*, 45, 2680-2707.

Lui, Q. and Ihler, A. (2011) Learning scale free networks by reweighted l1 regularization. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. Pp. 40-48.

Nasim, Z. *et al.* (2020) Role of AT1G72910, AT1G72940, and ADR1-LIKE 2 in plant immunity under nonsense-mediated mRNA decay-compromised conditions at low temperatures. *Int J Mol Sci,* 21(21).

Noir, S. et al. (2013) Jasmonate controls leaf growth by repressing cell proliferation and the onset of endoreduplication while maintaining a potential stand-by mode. *Plant Physiol,* 161(4): 1930-1951.

Pauwels, L. et al. (2009) Jasmonate-inducible gene: What does it mean? *Trends Plant Sci,* 14(2): 87-91.

Schaller, F. (2001) Enzymes of the biosynthesis of octadecanoid-derived signalling molecules. *J Exp Bot,* 52(354): 11-23.

Swiderski, MR. *et al.* (2009) The TIR domain of TIR-NB-LRR resistance proteins is a signaling domain involved in cell death induction. *Mol Plant Microbe Interact,* 22(2): 157-165.

Tamaoki, D. *et al.* (2013) Jasmonic acid and salicylic acid activate a common defense system in rice. *Plant Signal Behav,* 8(6): e24260.

Tan, K.M. *et al.* (2014) Learning graphical models with hubs. *J. Mach. Learn. Res.*, 15, 3297-3331.

Ueda, J., Kato, J. (1980) Isolation and identification of a senescence-promoting substance from wormwood (Artemisia absinthium L.). *Plant Physiol,* 66(2): 246-249.

Wei, M. *et al.* (2020) PuHox52-mediated hierarchical multilayered gene regulatory network promotes adventitious root formation in Populus ussuriensis *New Phytologist,* 228(4): 1369-1385.

Xia, Y. *et al.* (2015) Testing differential networks with applications to detecting gene-by-gene interactions. *Biometrika*, 102, 247-266.

Yang, J. *et al.* (2019) The crosstalks between jasmonic acid and other plant hormone signaling highlight the involvement of jasmonic acid as a core component in plant response to biotic and abiotic stresses. *Front Plant Sci,* 10: 1349.

Zhang, X. *et al.* (2019) DiffNetFDR: differential network analysis with false discovery rate control. *Bioinformatics,* 35, 3184-3186.