Using Graph Statistics to Investigate the Properties of a Gene Regulatory Network that may Control the Cold Shock Response in *Saccharomyces Cerevisiae*

A gene regulatory network (GRN) is a set of transcription factors which regulate the level of expression of genes encoding other transcription factors. The dynamics of a GRN show how gene expression in the network changes over time. Microarray data were obtained from the wild type strain and five transcription factor deletion strains (Δcin5, Δgln3, Δhap4, Δhmo1, Δzap1) before cold shock at 13°C and 15, 30, and 60 minutes after cold shock. A modified ANOVA showed that for all networks a large number of genes had a log2 fold change significantly different than zero at any time point. These genes were submitted to the YEASTRACT database to determine which transcription factors regulated them. Data from each strain were used to generate a candidate GRN of 15 nodes and 28 edges. The edges of this network were then systematically deleted, to determine the role of each edge in the network. GRNmap was used to estimate the parameters of these networks. Gephi was used to analyze the graph properties of each network. Betweenness centrality, eigenvector centrality, eccentricity, and closeness centrality were compute. The centrality measures, when analyzed together, indicate the role of a specific node in the network. From this analysis we have found eccentricity does not vary in the edge-deleted networks, but eigenvector centrality does, suggesting that this value will be more useful for determining which transcription factors are more important in the network.

Updates after a quick review today (4/18/18) written in red

Introduction:

Saccharomyces cerevisiae (budding yeast) is an excellent model organism, especially for systems biology. The organism is very small, eukaryotic, and single celled, making it easy to grow and care for, and easy to do genetic testing between generations. This allows the mechanisms observed in yeast to be translated to other eukaryotic cells, such as human cells. The organism has a small number of genes compared to human genome, at only 6000 compared to 22,000. eukaryotic and single celled organism. Due to these features, the yeast genome has been thoroughly studied by a large group of yeast researchers, which makes deletion strains, genome datasets, and other molecular tools for yeast easily and readily accessible. (Lee 2002).

In order to gain this wealth of information on the function of individual genes, the yeast was grown in varying environmental conditions, with a variety of mutations, and in different growth conditions. Stress responses in yeast are particularly well studied, such as an investigation by ter Schure et al. in 1995 which looked into how altering the ammonia concentration of media affected the yeast's ability to metabolize nitrogen (ter Schure et al. 1995). In addition to investigations on metabolism stressors, temperature stress is another common stress studied in yeast. In investigating heat shock, specific proteins have been identified that regulate the heat shock response, and act to stabilize proteins and other macromolecules to help the yeast survive in warmer temperatures (Jakob et al. 1993). The effect of heat shock and other environmental stresses on specific cellular functions in yeast is fairly well studied, however, the yeast's response to the stress of cold shock and cold temperature stresses remains unknown.

While most information remains unknown regarding the cold shock response, it is known that when yeast are introduced to cold shock (10°C), a physiological change in the rigidity of the

phospholipid bilayer is observed (Aguilera et al. 2007). Impairment of ribosome function and protein synthesis as well as a decrease of enzymatic activities have also been observed (Schade et al. 2004). While heat shock has a unique set of proteins which govern the stress response across organisms, there is no equivalent set of proteins that controls the response to cold shock across organisms. Just as is found in other environmental responses, however, it is known that yeast respond to cold shock by changing its level of gene expression (Schade et al. 2004). In knowing this, it is possible to investigate the mechanisms in place for regulating this cold shock response.

Yeast, like any organism, use transcription factors to regulate their levels of gene and protein expression in response to different external stimuli (Chen, et al. 2007). These transcription factors can act to activate or repress expression of different genes, where activators increase gene expression, and repressors decrease gene expression (Chen, et al. 2007). The transcription factors are proteins which are also encoded by genes, so the transcription factors themselves have transcription factors that activate or repress their expression (Chen et al. 2007). This is done through combinatorial control, where the transcription factors bind to a particular segment of DNA, act as activators or repressors, and "decide" whether to alter the expression of that gene in response to the external stimuli (McKenna and O'Malley, 2002). The resulting relationships of up and down regulation of genes are called transcriptional networks.

Transcriptional networks in yeast have been rigorously studied using DNA microarrays (Bumgarner, 2013). To identify and understand what transcription factors may play a role in the cold shock response, the Dahlquist Lab utilized DNA microarray experiments. Growth experiments are first performed using *S. cerevisiae* strains that have a deletion of genes encoding

transcription factors involved in the cold shock response. If it is observed that the deletion strain yeast has impaired growth at cold temperatures compared to a wild type strain of yeast, the assumption can be made that the deleted transcription factor plays a role in regulating gene expression in response to cold shock. Microarray experiments were then performed on the wild type strain, and for deletion strains which seemed to be involved in the regulation of cold shock response. DNA microarrays work by having a large number of known genes plated on a slide, and binds to a mixture of strain DNA via hybridization, which can be detected via fluorescence (Bumgarner, 2013). Microarray data from the wild type yeast, along with five deletion strains was collected after the yeast underwent cold shock at 13°C. The microarray data was collected at four different time points, 15 minutes, 30, 45 and 60 minutes after cold shock. Once the 60 minute mark is hit, the yeast recovers as it is placed back in optimal growth conditions. Changes of gene expression as compared to the time zero time point are measured, and the data is used to generate gene regulatory networks for the regulation of cold shock.

A gene regulatory network, also known as a GRN, is the set of transcription factors that is involved in controlling the level of gene expression for genes encoding other transcription factors in the network. Out of the roughly 6000 genes in the *S. cerevisiae* genome, there are close to 250 transcription factors that regulate the entire genome. Given the nature of *S.cerevisiae* as being a highly important and well studied model organism, there are many databases and tools available for researchers to utilize in a variety of investigations. One such database, called YEASTRACT, contains information on GRNs which originates from DNA-binding evidence, gene expression evidence, and regulatory motif sources, and also uses and combines different conditions to construct the network (YEASTRACT). Using this database and the microarray

data, GRNs were constructed from clusters of genes that had similar changes in expression. From this database, the Dahlquist Lab generated six small candidate GRNs, all around 15-20 genes in size, with 27-30 connections each.

In a gene regulatory network, the genes are interconnected in a way where when one transcription factor changes expression, the expression of that transcription factor target genes are also affected by the expression change. These directional connections can be used and visualized in a structure known as a graph, a series of nodes and edges that display connected relationships between different items or groups. Mathematics can be used to describe such relationships, and also used to model the dynamics of relationships between GRNs. Biological systems have many inputs and outputs, and non-linear operations, so mathematical models of these biological systems taking advantage of ordinary differential equations are best used to model dynamics of a biological system (Vu and Vohradsky, 2007). Ordinary differential equations were also used by a study in 2007 to model transcription factors involved in regulating the cell cycle in yeast, and the equations were able to confirm current findings on how the cell cycle is regulated (Vu and Vohradsky, 2007). In the Dahlquist Lab, a mathematical model using ordinary differential equations is also used to model the dynamics of the cold shock response in small GRNs (Dahlquist et al. 2015). These network dynamics are run using GRNmap, a modeling software generated by the Dahlquist Lab on Matlab that utilizes the ODE's as mentioned. GRNmap uses the ODE's to estimate parameters that affect expression levels of an individual gene, such as the production rates, weight parameter, and threshold expression (Dahlquist et al. 2015). The threshold b for the model is the point at which the production of the gene is switched on or off, and there is a challenge in estimating these parameters as a whole, in

fitting the equation to the gene expression data generated in the wet lab. Due to the mechanism of combinatorial control present in the GRN, the weight of each edge has an indefinite number of possibilities as to the relationship to the target gene, as does the threshold of when production is switch on or off. In using least squares error, these issues are resolved by comparing model outputs to the observed data to minimize the discrepancy in values (Dahlquist et al. 2015). Once these GRNs have been modeled, it is also possible to utilize other graphing software such as Gephi to analyze the graph statistics of the modeled networks, which reveals different properties of each node in the network.

In my investigation, I performed data and statistical analyses of the six candidate GRN's generated by the Dahlquist Lab. I ran Gephi to investigate the closeness centrality, betweeness centrality, eigenvector centrality, and eccentricity of each of the six networks. Out of the six networks, I focused on looking at and analyzing what could be interpreted from each of the graph statistics for each node of the GRN derived from the dhap4 deletion strain data, a GRN referred to as db5. I then performed an experiment in which each edge was systematically deleted from db5, one-by-one to investigate the impact and importance of each edge on the structure of the network. It was found that when edges were deleted from nodes that acted as hubs, there was more likely to be a significant difference in eigenvector centrality between the intact network and the edge deletion networks. This might indicate that eigenvector centrality is the statistic most impacted by changes in network structure.

Materials and Methods:

Microarray Data and Network Creation:

Data was formatted for use in GRNmap according to protocol found in the Dahlquist Lab website on OpenWetWare. The microarray data was obtained from six deletion strains, including the dhap4 deletion strain, and that data was formatted to be run on GRNmap. R was used to normalize the data, using a code which can also be found on the Dahlquist LAb website at https://openwetware.org/wiki/Dahlquist:Microarray_Data_Analysis_Workflow. The data was then formatted again, using a within-strain ANOVA test in Excel. A Benjamini-Hochberg p-value correction was used on the data to identify genes with significantly different log fold expression changes to compensate for the multiple testing problem.

The significant genes were then clustered using STEM software to generate potential groups of genes operating together. Significant profiles for each of the six strains were chosen, and then submitted to YEASTRACT to determine regulators of the significant gene targets found using STEM. The most significant regulators were chosen, and if a deleted transcription factor (e.g. HAP4) was not present, it was added to the group of significant regulators. An adjacency matrix was produced for the set, with "1" indicating a regulatory relationship between transcription factors, and a "0" indicating no relationship. These adjacency matrices were then used to generate input workbooks for the GRNmap software.

Input Workbook Creation for GRNmap:

The matrix was copied into an Excel workbook with sheets containing production rates for each gene, degradation rates for each gene, and a threshold b value. The log2 fold expression for each deletion strain was also contained on seperate worksheets, as well as an optimization

parameters sheet which contain instructions and parameters for the GRNmap model run. The protocol for creating this worksheet can be found at https://openwetware.org/wiki/Dahlquist:Microarray_Data_Analysis_Workflow#Create_the_Input_Excel_Workbook_for_the_Model and in Github: https://github.com/kdahlquist/GRNmap/wiki/How-to-format-the-input-file-for-GRNmap-v1.4-and-above

Creation of Edge-Deletion Experiment Input Workbooks:

Once the "intact" db5 network input workbook was generated, the adjacency matrix in both the "network" and "network_weights" worksheets had a single "1" changed to a zero. For example the ASH1-ACE2 "1" was deleted, and the workbook was saved and named based on the deletion. Ex. dASH1-ACE2_15-genes_27-edges_db5-MO-LK_Sigmoid_Estimation.xlsx. This was done for each of the 28 edges in the network. In the case of nodes where a deletion of an edge resulted in a free-floating node, as was the case for ZAP1→ ACE2, the free-floating node (ZAP1) was deleted from the network as well.

GRNmap Model Structure and Running the Model:

GRNmap stands for "Gene Regulatory Network modeling and parameter estimation" and can be found both as MATLAB code, and as executable code under an open source license at https://github.com/kdahlquist/grnmap.

$$\frac{dx_i(t)}{dt} = \frac{P_i}{1+\exp\left(-\sum_j \left(w_{ij}x_j(t) - b_j\right)\right)} - d_i x_i(t)$$

Each gene in the network has a differential equation that models the change in expression over time as production – degradation, shown above. Degradation rates for each gene were

taken from mRNA half life data from Neymotin et al. (2014). A sigmoidal production function where P*i* is mRNA production rate for gene *i* and d*i* is the mRNA degradation rate for gene *i*. w is the weight term, determining the level of activation or repression of j on i, and b is a threshold of expression for each gene. Positive weights to the edges represent activation, negative weights to edges represent repression. The magnitude of the weight parameter represents the strength of the regulatory relationship. The production rate (Pi), weight (w), and threshold (b) values were estimated from DNA microarray data using a penalized least squares approach.

$$E = \alpha\|\theta\|^2 + \frac{1}{Q}\sum_{\tau=1}^{Q}[z^d(t_\tau) - z^c(t_\tau)]^2$$

E represents the Least Squares Error (LSE) error and is the difference between the microarray data (experimental) values and simulated values derived from solving the differentical equation with the estimated parameters. The LSE can be compared to the minimum theoretical LSE (minLSE) achievable given the experimental data to compare the goodness of fit of different network models.

More detailed information on the mathematics behind the model can be found in the paper by Dahlquist et al.(2015).

GRNmap version used in the edge deletion model run is the v1.8 beta available on January 22, 2018. The specific script for the model has been uploaded to a Github/DahlquistLab/data/Spring2018/MO.LK Edge Deletion Data/ subfolder

Gephi Usage:

A properly formatted document can be found on github at https://github.com/kdahlquist/DahlquistLab/blob/master/data/Spring2017/Gephi_output/MO.GephiProtocol.1.docx, but the following details how files may be uploaded to Gephi:

1. Upload the desired excel file to GRNsight

2. Export the file to a weighted GraphML format using the file-> export data tab in GRNsight. Save the file in an easily accessibly location, such as the desktop
3. Open Gephi and select "open graph file"
4. Select your file, and the graph should appear on the main page
5. On the right-hand column of the Gephi window, select the desired graph statistics you wish to run
6. To view the output of these statistics, select the "Data laboratory" at the top of the screen
7. To export the table, hit the "export table" button on the menu bar for the data laboratory
8. Select the data columns you want to export, hit "Ok", save the file and then view the data in excel

Closeness centrality, betweeneness centrality, eigenvector centrality, and eccentricity were all used and investigated. This is horribly embarassing - I thought I added the papers and information regarding how each statistic was calulated here, but forgot to cite anything, and forgot to include paired t-test methods

$$C(x) = \sum_{y} \frac{(n-1)}{d(x,y)}$$

Closeness centrality was calculated using the above equation where n = the number of shortest paths going through the node; y = the node in question, x = the node passing through node y. Closeness centrality calculates the average length of the shortest path between the node and all other nodes in the graph. The more central a node is, the higher the closeness centrality value. Closeness centrality is an unweighted measure, where the value of choice ranges between 0-1, where 0 means that no shortest paths pass through the node y, and 1 means the node is fully connected to all other nodes in the graph.

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Betweeness centrality is caluclated by Gephi using the above, where v is the node of interest, and sigma is the number of paths from nodes st, and sigma(v) is the number of paths from st that pass

through node v. Betweenness centrality calculates how often a node appears on the shortest path between other nodes in the network. The higher the betweeness centrality value, the more it is being used as a stepping-stone from one node to the next.

Betweeness centrality is an unweighted measure, with values being expressed as fractions, or simplified to integers. The higher the betweenness measure, the greater the number of shortest paths that go through the node of interest. One downside to using betweeness centrality as a measure of importance or connectedness of a node is that the measure requires an input and output for each node in order for the measure to be calculated. So a node that might be the start of transcription regulation, and a highly important node in the network would still get a betweenness measure of 0, because there is nothing regulating that node, so it cannot be on the shortest path for any other nodes.

Eccentricity is calculated using an algorithm identifying the max{dist(i,j)} where *i* is the node of interest, and *j* is any other node in the network. This algorithm used by Gephi is detailed in a paper by Ulrik Brandes, *A Faster Algorithm for Betweenness Centrality* in the Journal of Mathematical Sociology. Eccentricity shows how accessible a node is from other nodes, or the distance from the starting node to the farthest node from it in a network. To have a high eccentricity measure means that the node is indirectly connected to other nodes in the network. Nodes with higher eccentricity have a higher impact/influence on other nodes in a network than nodes with a low eccentricity.

Eccentricity is an unweighted, directional statistic which only takes into account a node's out degree. Eccentricity is expressed as a positive integer, with an eccentricity of 0 indicating

that a node has no out degrees. A high integer means the node is highly connected, or has a far reach across the network.

$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t$$

Eigenvector centrality is calculated by using the adjacency matrix, where the relative centrality of a vertex *v* is defined as the above. *Av* = 1 if vertex *v* is linked to vertex *t* and = 0 otherwise. *M(v)* is a set of neighbors of vertex *v* and is a constant. Eigenvector centrality measures the influence of a node in a network. The measure assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question thank equal connections to low-scoring nodes. Gephi offers the option to run any number of iterations on Gephi, with the number of iterations automatically set to 100.

Eigenvector centrality is an unweighted statistic, with values between 0 and 1. Based on playing with the iteration count, it appears as though this centrality measure acts similarly to a limit, with the higher number of iterations causing the statistic to reach higher values. A high eigenvector centrality would indicate that the node in question has a high level of influence over the graph. This statistic is a node based measure rather than one that is indicative of edge importance, so it might prove interesting to look at this node in relation to network structure.

Results:

Comparing graph statistics of db5 to those of other data-derived candidate networks shows similar motifs and trends in node importance.

Six candidate networks were compared to determine how graph statistics may play a role in analyzing the importance of different nodes in a graph. The six candidate networks are shown in Figure 1, and will be referred to as db# for the rest of this paper.

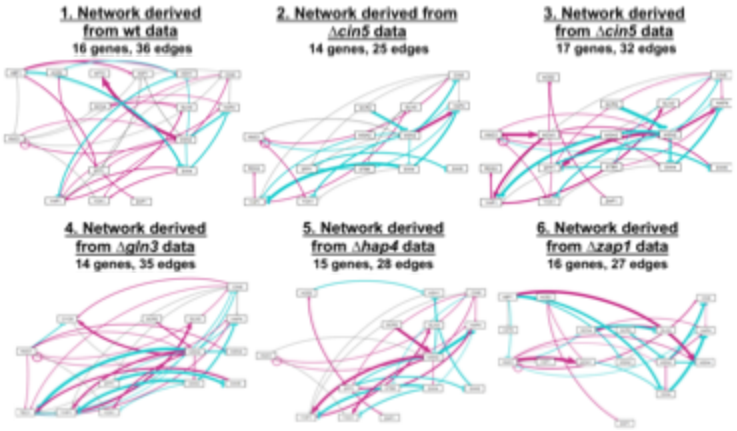

Figure 1. Weighted network visualizations of six candidate gene regulatory networks. Cyan edges indicate a repressive edge, while magenta indicates an activating edge. Going to find a better version of this figure for the final version of the thesis

The graph statistics of each network were found, and computed to determine the role of each statistic in determining the importance of an edge in the network regulating cold shock. It was thought that if there was a similar trend in statistic values for a gene across several networks in which is appeared, that might determine the overall role in the response to cold shock. As seen in Figure 2, eccentricity centrality was the first statistic calculated. The eccentricity centrality of a network shows how easily accessible a node is from other nodes The eccentricity is calculated using an algorithm for identifying the max{dist(i,j)} where i is the node listed in the table and j is any other node in the network. Eccentricity centrality is a directional statistic, which only takes a node's out degree into account. To have a high eccentricity centrality means that the gene is connected indirectly to many other genes in the network. This indicates these genes with high eccentricities have a greater impact on other nodes than a node with low eccentricity. As seen in

the figure, there were many nodes that had much higher eccentricity measures than others, such as CIN5, GCR2 and HMO1, indicating they are highly influential in their networks.

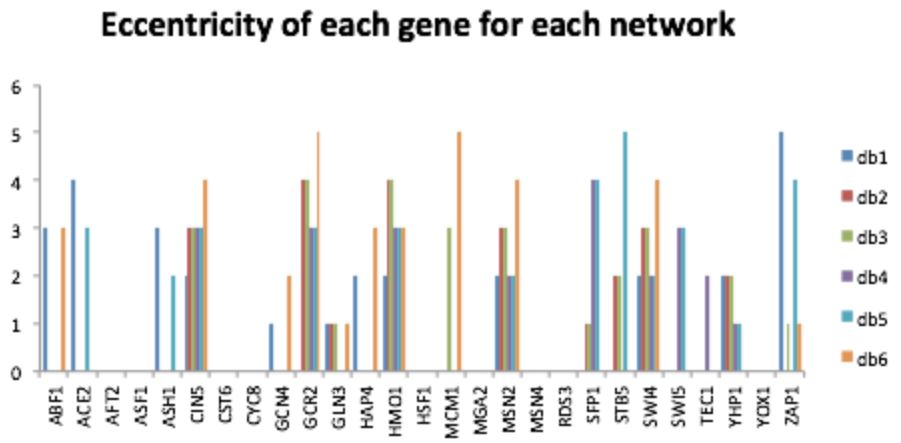

Figure 2. The eccentricity value of each gene is shown for the six data-derived gene regulatory networks.

Closeness centrality was the next measure calculated. Closeness centrality calculates the average length of the shortest path between the node and all other nodes in the graph. The more central a node is, the higher the closeness centrality value. Closeness centrality is an unweighted measure, where the value of ranges between 0-1, where 0 means that no shortest paths pass through the node y, and 1 means the node is fully connected to all other nodes in the graph. The directional aspect of the closeness centrality measure means those genes and nodes with no out-degree connections have a closeness centrality of 0. As seen in Figure 3, nodes such as GLN3, SFP1, and YHP1 seem to have the highest closeness centralities at 1, which might indicate they are frequently used by other nodes in the graph to act as a short-cut to other nodes of interest.

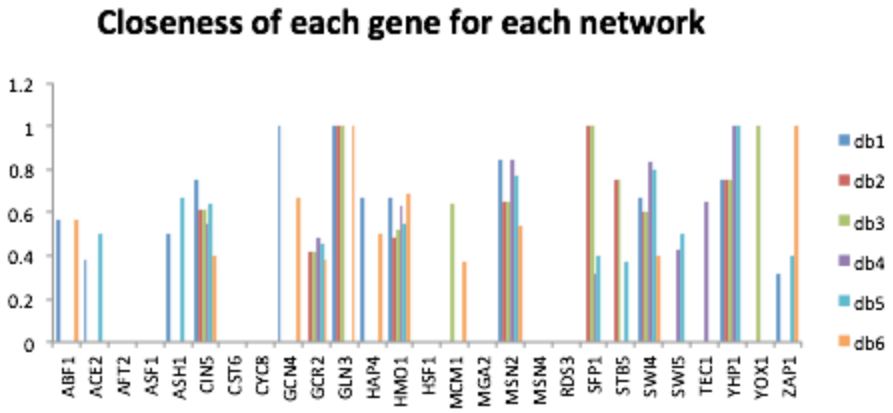

Figure 3. The closeness centrality of each gene is shown for the six data-derived gene regulatory networks.

Betweeness centrality was the last measure calculated for all six candidate networks. Betweenness centrality calculates how often a node appears on the shortest path between other nodes in the network. The higher the betweeness centrality value, the more it is being used as a stepping-stone from one node to the next. Betweeness centrality is an unweighted measure, with values being expressed as fractions, or simplified to integers. The higher the betweenness measure, the greater the number of shortest paths that go through the node of interest. One downside to using betweeness centrality as a measure of importance or connectedness of a node is that the measure requires an input and output for each node in order for the measure to be calculated. So a node that might be the start of transcription regulation, and a highly important node in the network would still get a betweenness measure of 0, because there is nothing regulating that node, so it cannot be on the shortest path for any other nodes. As is seen in Figure 4, MSN2 almost exclusively has the highest betweeness centrality of all nodes in all 6 networks, indicating that this transcription factor might be a highly important hub in the regulation of cold shock in *S. cerevisiae.*

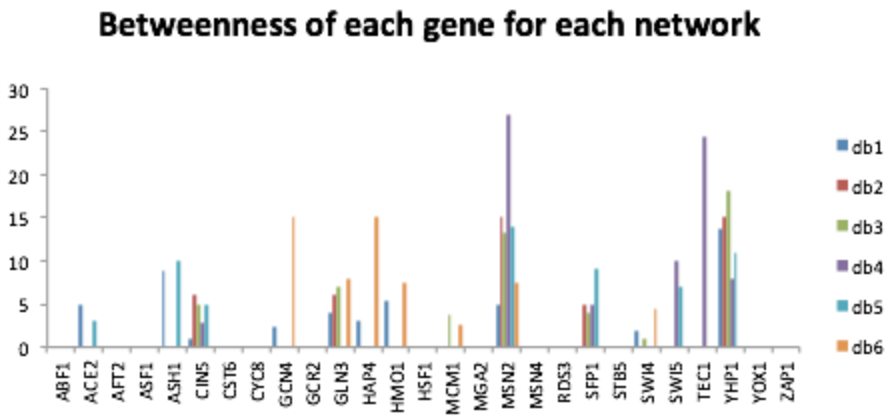

Figure 4. The closeness centrality of each gene is shown for the six data-derived gene regulatory networks.

An in depth analysis of each graph statistic in the network can reveal the properties of each node in the network

In looking just at the graph statistics of the db5 network, it was determined that the statistics, when combined can tell a lot about the relation of a specific node to the rest of the nodes in the network. Nodes with a low eigenvector centrality seem to be those that are deemed by other statistics to be the "most central to the graph," or the starting of regulatory pathways. This might indicate that the level of "node importance" may be partially based on a relationship between in degrees and out degrees shown in the network. Nodes with a high eigenvector centrality seem to be those that have fewer out degrees, and thus are more important because more nodes are regulating them. The eigenvector centrality measure does not seem to be particularly informative to investigating our networks. What is interesting from the graph as seen in Figure 5, is that it might indicate where regulatory pathways in networks end, as the highest measures are those with no out degrees. What these statistics show overall as well is that there seems to be something odd going on with the node YHP1. There is a statistic of 1 seen for almost all

measures, which would indicate that there might be an error generated for most of the centrality measures based on the calculations being done for each statistic. This relationship will be investigated further, and is something to keep in mind before judging the importance of YHP1 in the network.

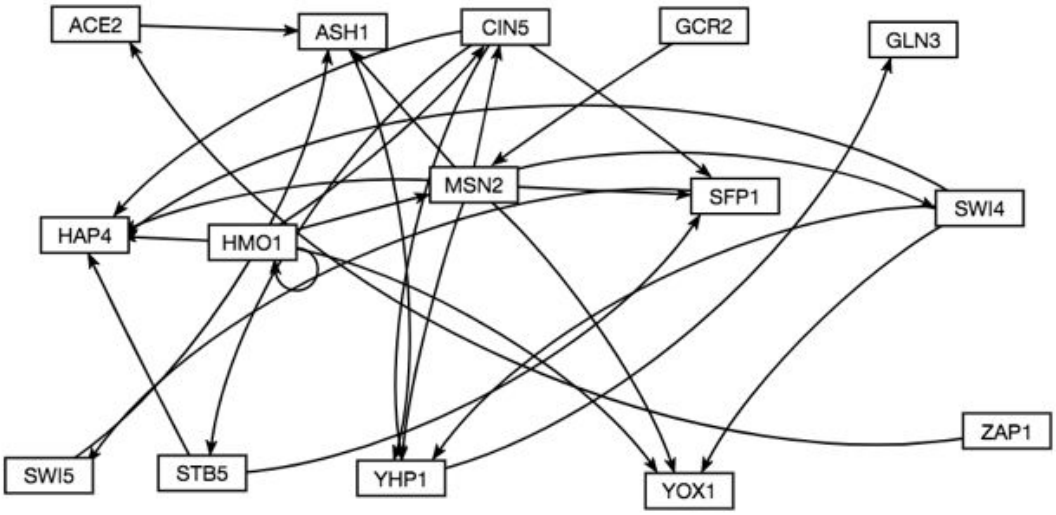

Figure 5. An unweighted visualization of the db5 network generated by GRNsight (Dahlquist, et al. 2015)

As seen in table 1, when combined the graph statistics can describe what the place of a specific node is in a graph. For example, looking at ACE2, a betweeness of 3 means that ACE2 is contained in 3 shortest paths on the network, so it is being used as a hub for a small number of nodes to reach other nodes in the graph. A closeness of 0.5 is relatively high, which means many nodes have paths going to, or going through ACE2, which makes sense in that it acts as a hub for several nodes. An eccentricity of 3 means that compared to other nodes in the graph, ACE2 has a similar "reach" or influence. This moderate influence indicates that the hub nature of the node is of moderate importance to the network. An eigenvector centrality of 0.008418 is incredibly low, and indicates that because the node is deemed central

by other measures, and because the in:out degree ratio is 1:1, the eigenvector statistic is labeling the node as unimportant. This analysis can be done for each node in the network, revealing the various levels of importance of each node.

Table 1. Compilation of graph statistics as computed by Gephi for the network db5

Gene	Closeness Centrality	Betweeness Centrality	Eigencentrality	Eccentricity
ACE2	0.5	3	0.008418	3
ASH1	0.666667	10	0.575118	2
CIN5	0.636364	5	0.249597	3
GCR2	0.458333	0	0	3
GLN3	0	0	0.8377	0
HAP4	0	0	0.861994	0
HMO1	0.55	0	0.11352	3
MSN2	0.769231	14	0.121938	2
SFP1	0.4	9	0.605438	4
STB5	0.375	0	0.248138	5
SWI4	0.8	0	0.136077	2
SWI5	0.5	7	0.52969	3
YHP1	1	11	1	1
YOX1	0	0	0.392633	0
ZAP1	0.4	0	0	4

The following is a thorough analysis of each node, detailing the role of the node in the network's overall structure:

Not sure how it was wanted for me to include this information in the results section, so I just listed everything node by node.

ACE2

- Betweeness Centrality: 3

  A Betweeness of 3 means that ACE2 is contained in 3 shortest paths on the network, so it is being used as a hub for a small number of nodes to reach other nodes in the graph.
- Closeness Centrality: 0.5

  A Closeness of 0.5 is relatively high, which means many nodes have paths going to, or going through ACE2, which makes sense in that it acts as a hub for several nodes.
- Eccentricity: 3

  An eccentricity of 3 means that compared to other nodes in the graph, ACE2 has a similar "reach" or influence. This moderate influence indicates that the hub nature of the node is of moderate importance to the network.
- Eigenvector Centrality: 0.008418

  An eigenvector centrality of 0.008418 is incredibly low, and indicates that because the node is deemed central by other measures, and because the in:out degree ratio is 1:1, the eigenvector statistic is labeling the node as unimportant.

ASH1

- Betweeness Centrality: 10

  With a Betweeness of 10, ASH1 is shown to be acting like a large hub for the network, with many shortest paths having to pass through ASH1
- Closeness Centrality: 0.666667

  A closeness centrality of 0.666667 is relatively high, which makes sense in conjunction with the high Betweeness measure, as many paths in the network have to pass through or go to ASH1
- Eccentricity: 2

  An eccentricity of 2 means that ASH1 has slightly less influence on the graph than the majority of the nodes, which would indicate that in acting as a hub, it is more of a way station than a command center in sending out activating or suppressing influences across the network.
- Eigenvector Centrality: 0.575118

  With a relatively high eigenvector centrality of 0.575118 and a Betweeness value also so high, this means that ASH1 has more in degrees than out degrees, and that more nodes are trying to regulate it than it is regulating other nodes.

CIN5

- Betweeness Centrality: 5

  With a Betweeness centrality of 5, CIN5 is operating as a moderately sized hub in the network, with several shortest paths passing through the node.
- Closeness Centrality: 0.636364

  A closeness of 0.636364 makes sense in this node, as it means it is highly central to the network, which further confirms the hub nature of the node.
- Eccentricity: 3

  An eccentricity of 3 indicates that CIN5 has a moderate level of influence over the network, as an eccentricity of 3 is about average for the network.
- Eigenvector Centrality: 0.249597

  An eigenvector centrality of 0.249597 means that the node has more out degrees than in degrees, and is having more influence on other nodes in the network than nodes are having on it.

GCR2

- Betweeness Centrality: 0

  With a Betweeness centrality of 0, and looking at the network, GCR2 is at the start of a regulatory chain, and not a hub in the network.
- Closeness Centrality: 0.458333

  With a closeness centrality of 0.45833, this means that GCR2 is moderately connected to the rest of the network, and through it's connection to MSN2, it has many shortest paths connecting it to other nodes.
- Eccentricity: 3

  With an eccentricity of 3, it can be determined that while at the start of a regulatory pathway, GCR2 has an average level of influence over the rest of the network, which when compared to other nodes at the start of regulatory pathways, might help determine the importance of GCR2.
- Eigenvector Centrality: 0

  GCR2 has an eigenvector centrality of 2, which means nothing is influencing the node (no in degrees), which makes sense seeing as how the node only has one out degree.

GLN3:

- Betweeness Centrality: 0

  With a Betweeness centrality of 0 and looking at the graph, it can be determined that GLN3 is at the end of a regulatory pathway.
- Closeness Centrality: 0

A closeness centrality of 0 makes sense, as there are no out degrees for GLN3, and therefore no edges are emanating from it that form a shortest path.

- Eccentricity: 0

  An eccentricity of 0 makes sense for GLN3, as there are no nodes for it to influence, since there is no out degree for this node.

- Eigenvector Centrality: 0.8377

  As the in:out degree ratio for GLN3 is 1:0, it makes sense for this node to have a very high eigenvector centrality at 0.8377, as nodes are regulating it, and it is regulating no nodes.

HAP4:

- Betweeness Centrality: 0

  With a Betweeness centrality of 0 and looking at the network, HAP4 is at the end of a regulatory pathway, which makes sense why no shortest paths are passing through the node.

- Closeness Centrality: 0

  With a closeness centrality also at 0, this makes sense as there are no edges emanating out from the node.

- Eccentricity: 0

  With an eccentricity of 0, it makes sense that HAP4 has no influence over other nodes in the network.

- Eigenvector Centrality: 0.861994

  HAP4 has a high eigenvector centrality at 0.861994 because the ratio of in degree:out degree is 5:0, which shows many nodes are influencing HAP4.

HMO1:

- Betweeness Centrality: 0

  With a Betweeness centrality of 0, it can be determined that the Gephi measure for Betweeness does not take self-regulation into account as an in degree. Therefore, HMO1 does not act as a hub, and there are no shortest paths that go through HMO1, making it the start of a regulatory pathway.

- Closeness Centrality: 0.55

  With a closeness centrality of 0.55, HMO1 is moderately connected to the rest of the network, with many shortest paths emanating from HM01.

- Eccentricity: 3

With an eccentricity value at 3, HMO1 has the same level of influence as the majority of the genes in the network, indicating that while it is the start of a regulatory pathway, it is not necessarily of the most importance.

- Eigenvector Centrality: 0.11352

  With a low eigenvector centrality, HMO1 is influencing more nodes than are influencing it, which makes sense with the other centrality measures calculated for this node/network.

MSN2:

- Betweeness Centrality: 14

  With a Betweeness centrality of 14, MSN2 is shown to be the biggest hub in the network, with many nodes containing MSN2 on a shortest path. This means it is a highly important node in that it acts as a step stone that is incredibly central to the structure of the network.

- Closeness Centrality: 0.769231

  With a very high closeness centrality, it is clear that not only do shortest paths go through MSN2, but it also has a number of shortest paths emanating from it, further showing the importance of this node as a hub for many edges in the network.

- Eccentricity: 2

  With an eccentricity of 2, MSN2 might not have the farthest reach across the graph, but in being incredibly central to the graph, it is also possible that this number is lower than average because MSN2 does not have to reach as far as other nodes to get to the furthest node from it.

- Eigenvector Centrality: 0.121938

  With many in and out degrees, it makes sense that MSN2 has a very low eigenvector centrality, as the number of nodes regulating it, as it is regulating is fairly similar.

SFP1:

- Betweeness Centrality: 9

  With a Betweeness of 9, SFP1 is acting as a moderate sized hub for the network, with several shortest paths going through the node.

- Closeness Centrality: 0.4

  With a closeness centrality of 0.4, SFP1 has a moderate closeness centrality, which means it is moderately connected to the rest of the network, which makes sense given the hub-like nature of the node being described by the high Betweeness centrality value.

- Eccentricity: 4

  With an eccentricity of 4, SFP1 is able to reach and influence more nodes in the graph than is average of this network. This means it is a network of high influence, and based on the weight of

the edges emanating from this node, it might be determined what the influence of this node is on the network.

- Eigenvector Centrality: 0.605438

  With a relatively high eigenvector centrality of 0.605438, it can be determined that this hub has more in degrees than out degrees, and thus the few edges that are coming out of the node have a far reach over the graph, and are important to the structure of the network.

STB5:

- Betweeness Centrality: 0

  With a Betweeness of 0, STB5 is not being used as a hub for the graph, which makes sense, as in the network STB5 is at the start of a regulatory pathway

- Closeness Centrality: 0.375

  With a closeness centrality of 0.375, STB5 has a moderate closeness centrality, which makes sense as STB5 has two out degrees, which are both shortest paths.

- Eccentricity: 5

  With an eccentricity of 5, STB5 has the highest eccentricity in the network. This means it has the highest level of influence over the network, and the farthest "reach" across the network. This makes sense, as the edge between STB5 and SFP1 connects STB5 to the rest of the network.

- Eigenvector Centrality: 0.248138

  An eigenvector centrality of 0.248138 means that the node is not very important in the graph, which makes sense as it only has out degrees, with no in degrees. This means it is inherently influencing more nodes than are influencing it.

SWI4:

- Betweeness Centrality: 0

  With a Betweeness centrality of 0, and both in and out degrees, this means that while pathways exist going through SWI4, those pathways are not the shortest pathways that exist between nodes, and so there exists more direct pathways between genes such as MSN2 and YOX1 (a pathway that goes through SWI4)

- Closeness Centrality: 0.8

  SWI4 has the highest closeness centrality in the network, which would indicate that it has the largest amount of shortest paths emanating from the node. Following the edges coming from SWI4, this makes sense, as the edges lead to CIN5, and other nodes with a high Betweeness.

- Eccentricity: 2

With an eccentricity of 2, SWI4 does not have the furthest reach across the network. This might indicate that SWI4 is centrally located in the connection of edges in the network.

- Eigenvector Centrality: 0.136077

  With an eigenvector centrality of 0.136077, SWI4 has a low eigenvector centrality, meaning that SWI4 is of low importance in the graph, and the ratio of in to out degrees is close to 1:1.

SWI5:

- Betweeness Centrality: 7

  With a Betweeness of 7, SWI5 is used by the network as a fairly central hub between nodes, with many shortest paths going through SWI5.

- Closeness Centrality: 0.5

  With a closeness of 0.5, SWI5 has a moderate number of paths emanating from the node, through a connection from SWI5 to ASH1.

- Eccentricity: 3

  With an eccentricity of 3, SWI5 has a moderate level of influence over the rest of the network, with 3 being about average for eccentricity measures across the network.

- Eigenvector Centrality: 0.52969

  With an eigenvector centrality of 0.52969, SWI5 is of moderate importance to the graph, which might indicate, given that its in:out degree ratio is 1:1, that SWI5 is closer to the end of the regulatory pathway it is on than to the beginning of the pathway.

YHP1:

- Betweeness Centrality: 11

  With a Betweeness centrality of 11, YHP1 is being used frequently as a hub between nodes, which makes sense given that it has in degrees coming from MSN2, CIN5, and ASH1, which also are large hubs in the network.

- Closeness Centrality: 1

  With a value of 1, YHP1 has the highest closeness centrality measure in the network. This might be due to the large number of hubs that it is connected to, which would lead to the largest number of edges emanating from YHP1 than any other node in the graph.

- Eccentricity: 1

  With an eccentricity of 1, YHP1 does not have much reach across the network, which would make sense, as the other graph statistics seem to indicate that YHP1 is the most central hub in the graph, which would mean it doesn't have to reach far to access the furthest node from it.

- Eigenvector Centrality: 1

With an eigenvector centrality of 1, YHP1 is evidently the most important node in the graph, which would make sense given the information provided by the other graph statistics.

YOX1:

- Betweeness Centrality:
  With a Betweeness of 0, YOX1 is not being used as a hub for any nodes in the graph. This makes sense, as in looking at the graph it is at the end of a regulatory pathway.
- Closeness Centrality: 0
  YOX1 has a closeness of 0, which makes sense as there are no out degrees emanating from the graph, and it is at the end of a regulatory pathway.
- Eccentricity: 0
  With an eccentricity of 0, YOX1 has no reach across the network, which makes sense given that the node is at the end of a regulatory pathway, and thus has nothing that it can reach to.
- Eigenvector Centrality: 0.392633
  With an eigenvector centrality of 0.392633, YOX1 has a much lower eigenvector centrality than other nodes that are also at the end of regulatory pathways. This might be because the nodes regulating YOX1 are not the most connected in the graph.

ZAP1:

- Betweeness Centrality: 0
  With a Betweeness centrality of 0, ZAP1 is not being used as a hub for any nodes in the graph. Looking at the network, this is due to ZAP1 only having one out degree, regulating ACE2.
- Closeness Centrality: 0.4
  With a closeness centrality of 0.4, the path emanating from ZAP1 is deemed to be moderately important to the network, which makes sense given that the edge eventually connects to YHP1, which statistically, seems to be the most important node in the network
- Eccentricity: 4
  With an eccentricity of 4, ZAP1 has above average reach across the graph, which makes sense given that in regulating ACE2, it is indirectly influencing some of the important hubs in the network.
- Eigenvector Centrality: 0
  With an eigenvector centrality of 0, ZAP1 is deemed to be very unimportant in the graph, which makes sense given that the in:out degree ratio for this node is 0:1.

Comparing production rates and betweenness centrality measures of randomly generated networks to those of db5 reveals significant differences in how networks are structured.

The db5 network was then compared to 30 random networks generated from the db5 data. Based on a paired t-test that was corrected using a Benjamini-Hochburg test, several random networks were found to have significant variation in production rates from the production rates of db 5 (Table 4) and the betweeness centrality measures of db5 (Table 3). There was a correlation between random networks with production rates that varied significantly from the db5 production rates and those that had a signifcant difference in betweenness centrality, though further analysis of variation between other graph statistics might determine whether or not this is a spurious correlation.

Table 2. This would show the results of the betweeness centrality paired t-test comparison of db5 to random networks. Not sure the best method for concise visualization of this data, however below is the link to the workbook https://github.com/kdahlquist/DahlquistLab/blob/master/data/Fall2017/MOdb5_rand1-31_BetweenessComparison.xlsx

Table 3. A similar issue on concise visualization as found for table 3, however the production rate comparison workbook can be found below https://github.com/kdahlquist/DahlquistLab/blob/master/data/Fall2017/MOdb5_rand1-31_ProductionRates.xlsx

Edge deletion experiments reveal the impact of specific edges on the performance of the network

Each edge deletion network was run through the beta version of GRNmap available on January 23, 2018, and then the estimated parameters of each edge deletion network were compared to those of db5. It was found that depending on which edge was deleted, the performance of how

the network was modeled changed. As shown in Figure 5, the deletion of an edge had a large impact on the LSE:minLSE ratio of a network. It was found that edges such as SWI5-ASH1 were highly important to the network being modeled well, while edges such as ZAP-ACE2 actually hindered the modeling of the network. Using a paired t-test and Benjamini-Hochburg correction, the edge deletion variants of db5 were compared to the intact db5 network for production rates, as well as a variety of graph statistics (Figure 6). Through doing this test, it was found that in looking at eccentricity, there was the least change in variation between the intact db5 network and networks with a deleted edge. This might indicate that it is not the most useful statistic in determining the importance of nodes in a network. Eigenvector centrality on the other hand, showed the most significant variation between db5 and the edge deletion networks, indicating that this measure may be most useful in determining the importance of nodes in the network.

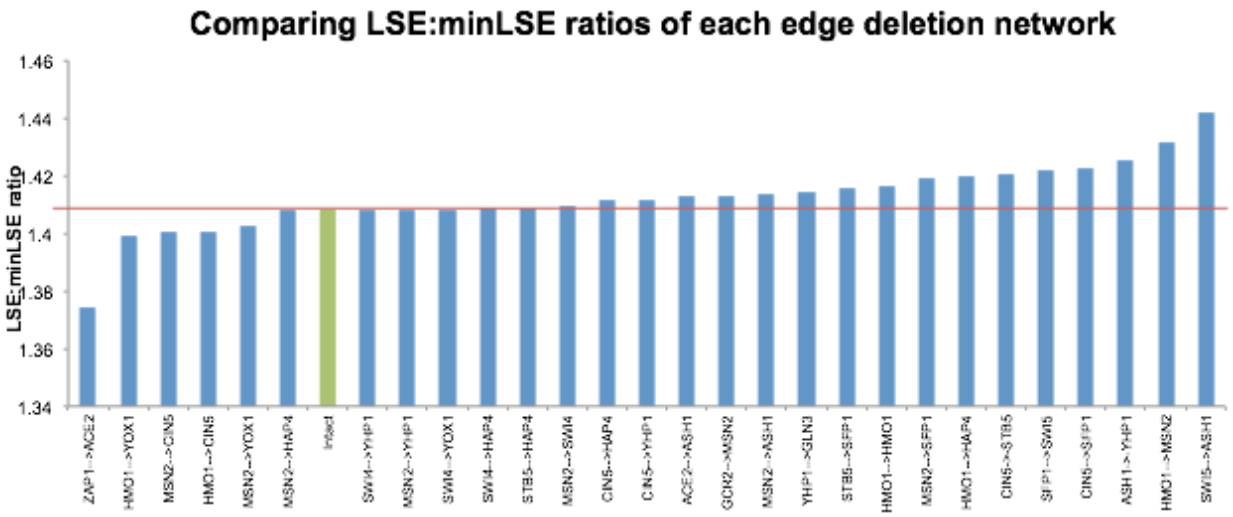

Figure 5. This figure compares the LSE:minLSE ratios of the intact db5 network as compared to the edge deletion networks. The red line indicates the LSE:minLSE ratio of db5, to illustrate how well the edge deletion networks were modeled as compared to db5.

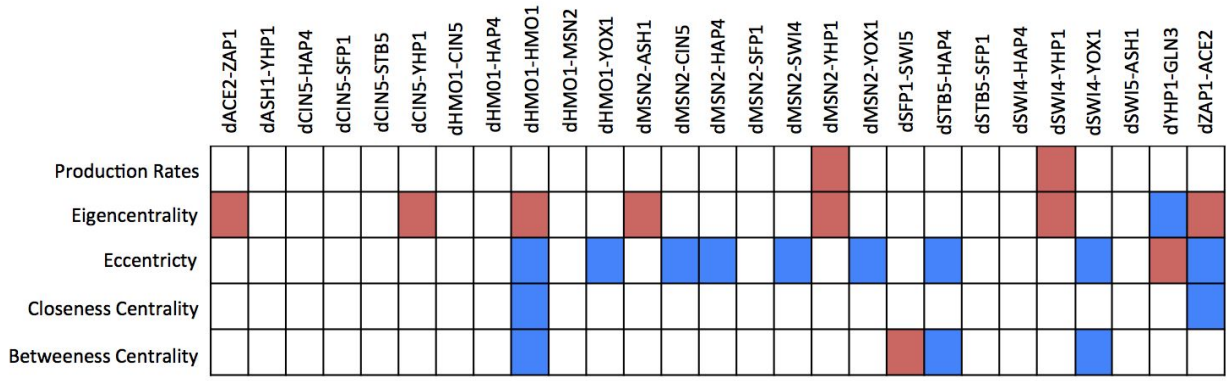

Figure 6. Paired t-tests were performed comparing the graph statistics and production rates of db5 to each of the random networks. Red indicates a significant change in values, while a blue box indicates there was no variation between db5 and the deletion strain.

After performing the paired t-tests for the full range of graph statistics, an analysis of edge weights was also performed, as shown in Figure 7. In comparing the weights across networks, it was shown that weights changed most when the edge deletion was connected to a major hub in the network, such as CIN5, MSN2, or HMO1. The edges that seemed to cause the most change in edge weight modeling are the HMO1 → CIN5, MSN2 → CIN5, HMO1 → YOX1, and ZAP1 → ACE2 edge deletions. It is interesting to note, going back to Figure 5, that of these four edge deletions, ZAP1 → ACE2 is the only deletion that resulted in improved model performance from the intact network according to the LSE:minLSE ratio. The other three edge deletion networks all performed worse than the intact network according to the LSE:minLSE ratio. This might indicate that the edge weight plays a role in the overall importance of the network, or that edge weights may be changed in the model in order to compensate for structural changes in the network.

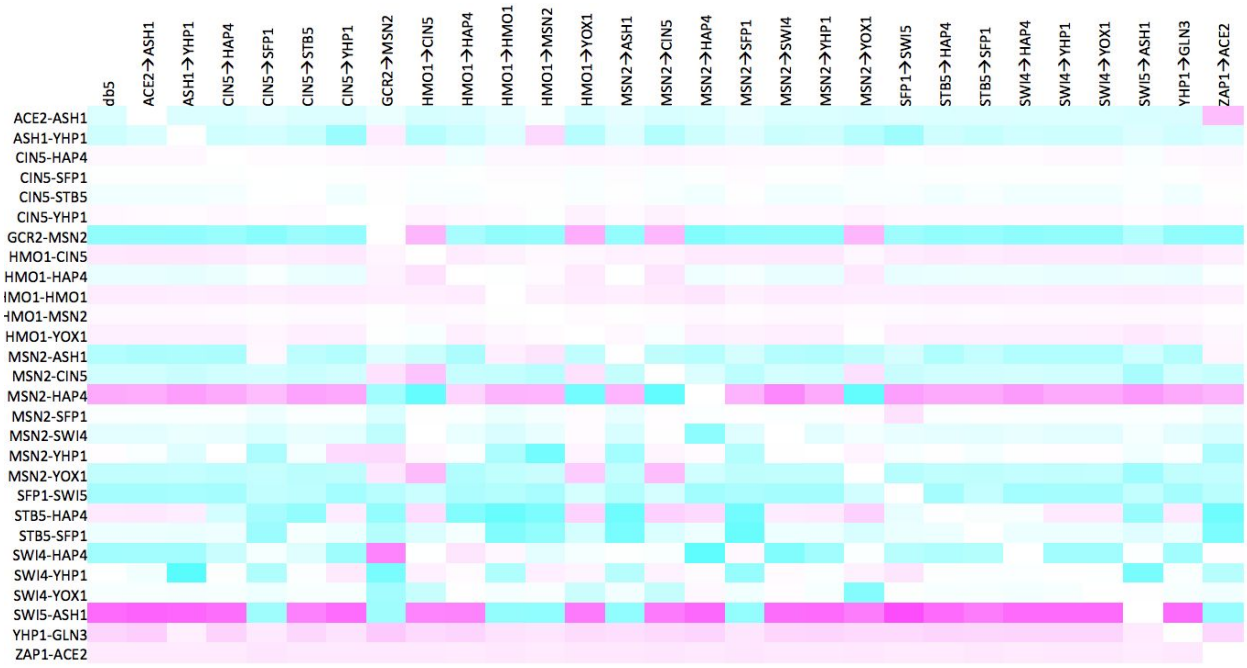

Figure 7. This figure shows the normalized weights of each edge in each deletion network as modeled by GRNmap.

In looking closer at the two statistics that changed the most and the least with the edge deletions, it was observed that overall, eigenvector centrality significantly increased when there was a significant difference in expression observed between the intact and deletion networks (Figure 8.) In looking at the edges in question which were deleted, it was seen that all but ZAP1 → ACE2 performed worse in the LSE:minLSE ratio than the intact network (Figure 9). This might indicate that significant differences in eigenvector centrality between two similar networks would generally indicate differences in overall network performance.

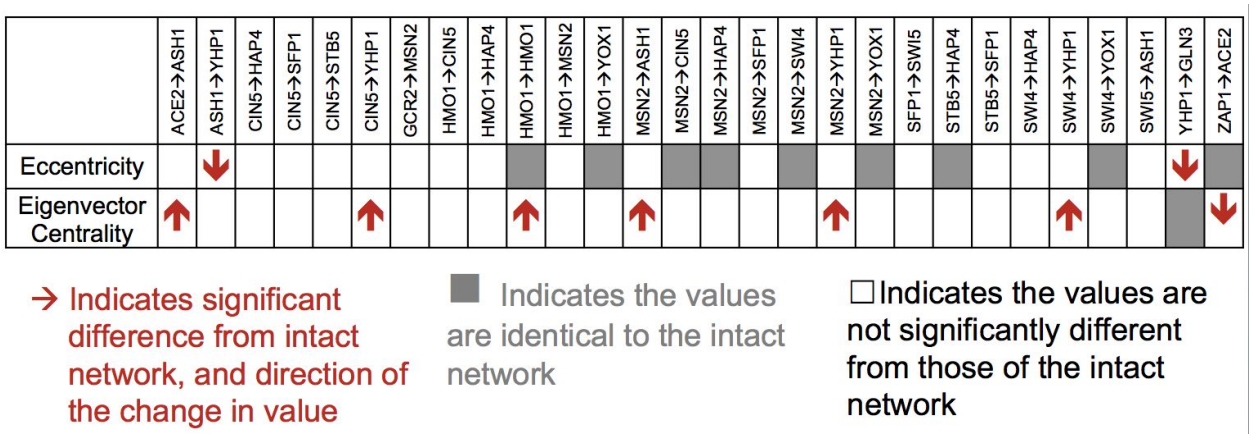

Figure 8. A close look at Figure 6, with arrows now indicating the direction of a significant shift in eccentricity and eigenvector centrality values across all nodes.

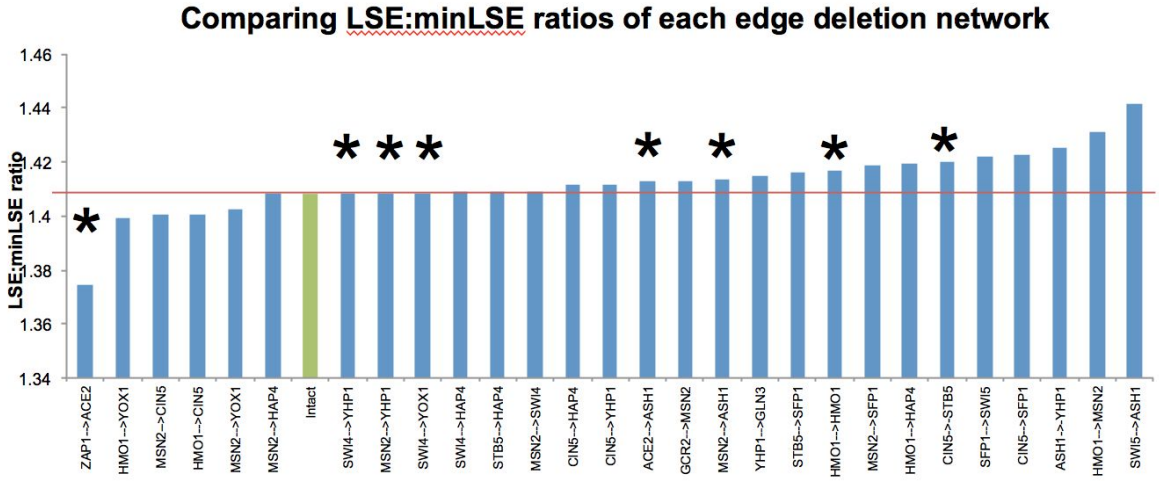

Figure 9. This figure shows the same information as Figure 5, with the asterisks added next to the edges which showed a significant difference in eigenvector centrality across the network as compared to the intact network.

Discussion:

DNA microarray data from all six strains subjected to cold shock was analyzed using an ANOVA test, the YEASTRACT database, and an ordinary differential equations model called GRNmap that modeled the dynamics of each gene in candidate gene regulatory networks. The output weight parameters were visualized using GRNsight.

The Gephi results showed that many of the centrality measures are consistent with the

in-degree, out-degree statistics, where the genes with the highest degree and overall degree measures are also found to have the highest betweenness centrality measures, and those nodes with the lowest degree measures also have the lowest betweenness centrality. The statistics from Gephi provided useful information through which to view the graphs. While MSN2 has the highest betweenness centrality and the highest degree measure, it is tied for the highest eccentricity with SWI4, which shows that high accessibility might not be directly related to high centrality in the networks.

The average in- and out-degrees across all networks reveal trends across the board, such as YOX1 having very little activation. YOX1 was also found to not be regulating any other gene across the board. This is similar to the graph statistics which show YOX1 as being least central on average to all networks. This might suggest that it should not be included in the networks moving forward, and might not play a significant role in regulation in response to cold shock. In addition to the above, future directions include comparing the Gephi statistics to the statistics from random networks. Then, comparisons of the database-derived network statistics to the random networks could be performed to determine if genes such as MSN2 were deemed to be similarly central and important in those networks. It would also be interesting to run Gephi analysis on networks of larger size in order to see how the centrality of nodes and connections change with the deletion of important nodes and edges.

In analyzing the graph statistics by hand, it was determined that with all graph statistics at hand, it may be possible to generally recreate a network structure using graph statistics alone.

My apologies for the below being very rough and not fleshed out

Looking at the edge deletion experiment, in total, 29 networks were examined. These were comprised of the intact network and the 28 networks generated from each individual edge deletion. Five of the networks overall had better LSE:minLSE ratios than was seen in the intact network, indicating that deletion of these genes resulted in a better performing network. . These five resulted from the HMO1→ CIN5, HMO1→ YOX1, MSN2→ CIN5, MSN2→ YOX1, and ZAP1→ ACE2 edge deletion networks. In the case of sixteen of the edge deletions, the LSE:minLSE ratio was worse than the intact network. These deletions included ASH1→ YHP1, HMO1 → MSN2, and SWI5→ ASH1 as the worst performing networks overall according to the LSE:minLSE ratio. The edge deletions that resulted in a higher LSE:minLSE ratio suggest that those particular edges are important to to the network.

The systematic deletions of each of the 28 edges in the intact network revealed that ZAP1ACE2 is most likely not important to the network and can be removed. The edges that cause changes in optimized expression and increase the LSE:minLSE ratio when they are deleted are likely important to the network. The edges that cause variability in other edges and decrease the LSE:minLSE ratio when they are deleted are likely not important to the network.