The Transcriptional Response to Oxidative Stress in Yeast Treated by Cumene Hydroperoxide: A GRN Analysis

Hailey C. Ivanson, Charlotte N. Kaplan, Katherine M. Miller, Andrew H. Sandler, Natalija Stojavonic, Dean Symonds

BIOL 367

3 May 2024

Introduction

When exposed to reactive oxidative species, or ROS, at an unmanageable level, an oxidative stress response is induced in *Saccharomyces cerevisiae* (Jamieson, 1998). The oxidative stress response, or OSR, in *S. cerevisiae,* attempts to prevent cellular damage by maintaining a homeostatic redox state (Jamieson, 1998). Some mechanisms of this response include glutathione (GSH), a radical scavenger; thioredoxin, a ROS-reducing protein; and trehalose, an antioxidant disaccharide (Jamieson, 1998). The *S. cerevisiae* OSR is at least partially transcriptionally regulated, as are many other stress responses in *S. cerevisiae*, which allow the organism to tailor its response to its stressors as necessary (Jamieson, 1998). Transcription factors regulate gene expression in response to environmental changes. In eukaryotes like *S. cerevisiae,* most transcription factors work by binding to DNA at specific parts of the sequence, such as the promoter or enhancer regions (Pierce, 2020). These transcription factors can alter both whether transcription occurs at all, as well as the rate of transcription (Pierce, 2020).

A 2013 study by Sha et al. focused on the *S. cerevisiae* early OSR to cumyl hydroperoxide, or CHP. The experiment involved growing a control culture of yeast, with no CHP exposure, parallel to a CHP-exposed culture of yeast under specific conditions, and then comparing their transcriptional changes. The researchers used the Affymetrix GeneChip system to perform microarray hybridization and ran a gene-by-gene ANOVA to determine the significance of the expression changes found.

The Sha et al. study focused on only the first 20 minutes of data collected for a number of reasons, including that it was found that the CHP-exposed culture was able to eliminate the CHP and convert it to its nontoxic alcohol form within 20 minutes; further, the researchers noted

unexpected happenings at the 40 minute mark in the control culture, likely due to collecting yeast too late in their growth cycle, leading to stress related to overpopulation, which likely differs transcriptionally from oxidative stress. Because this other source of stress was likely also affecting the CHP-exposed culture, the data past 20 minutes was deemed unreliable and not used. Sha et al. were able to focus on the early transcriptional response of yeast to CHP, finding gene clusters with different functions, and identifying the formerly unknown functions of genes Yap3, Yap5, and Yap7 as part of the yeast OSR. The Sha et al. paper from 2013, while only using the data from the first 20 minutes for analysis, uploaded the data from their entire experiment's duration, 120 minutes, to the Gene Expression Omnibus (GEO) database.

The transcription factors that belong in the gene regulatory network, or GRN, that controls the gene expression of the OSR in *S. cerevisiae* is still unknown. In order to solve this, the DNA microarray data from Sha et al., 2013 will be re-analyzed using transcriptional regulator data from Harbison et al., 2004. A GRNmap can be made to model the GRN that is responsible for the OSR in *S. cerevisiae*; this model can be made by processing the Sha et al. data to create a GRNmap input workbook with a Microsoft Access database. The Sha et al. data was downloaded and processed to find which transcriptional changes were significant for further analysis by running an ANOVA and corrections. Clusters of the significant genes were formed, and the most significant cluster was analyzed further. For the significant cluster profile, Yeastract was used to predict transcription factors that most likely regulate the cluster. A Microsoft Access database was formed using the Harbison et al. transcriptional regulator data. The database was loaded with an expression table from the processed Sha et al. data.

The objective of this project is to use an Access database to create a gene regulatory network map, or GRNmap to draw new conclusions about the yeast OSR to CHP over a

120-minute period. The GRNmap would also allow quality checking of the Sha et al. study. This process also allowed for quality checking of the database that was created, as well as the methodologies used, such as the ways the data was organized, and the comparisons that were drawn.

Methods, Materials, Results, and Discussion:

ANOVA results:

Table 1: Table of Benjamini and Hochberg ANOVA results with p-values.

B&H ANOVA	Control Genes	CHP Genes
p<0.05	3699 (78.7%)	2856 (60.8%)
p<0.01	3219 (68.6%)	2390 (50.9%)
p<0.001	2558 (54.4%)	1857 (39.5%)
p<0.0001	1921 (40.9%)	1419 (30.2%)
p<0.00001	1325 (28.2%)	1058 (22.5%)

The Sha et al. data was downloaded and processed to assess the log2 fold change data. ANOVAs were run, and adjustments were made using Bonferroni and Benjamini and Hochberg corrections of p-values. After assessing the significant Benjamini and Hochberg (B&H) p-values for both the control and CHP genes, it was discovered that Sha et al.'s control group exhibited many more genes and a higher percentage of significant p-values below their respective threshold. This indicates that the treatment may not be as effective at producing significant changes in gene expression compared to the control group, which was neither expected nor true,

as CHP did invoke the OSR in the exposed yeast, but not in the control yeast. It is thought that there may be overpopulation stress affecting the control yeast detected at the 40 minute mark causing a transcriptional response. Because the 0-120 minute time frame was utilized and the yeast was collected at or after the mid-log phase, there cannot be accurate representation of gene expression as the negative control is not an effective negative control after the 20 minute mark. At $p<0.05$ 78.7% of control genes and 60.8% of CHP-treated genes expression are significantly changed. 25% significance was the approximate aim when selecting a p-value for significance, so a p-value of 0.00001 was used as a cutoff. The Sha et al paper recognized $p<0.01$ as significant, which differs from this conclusion.

Figure 1:

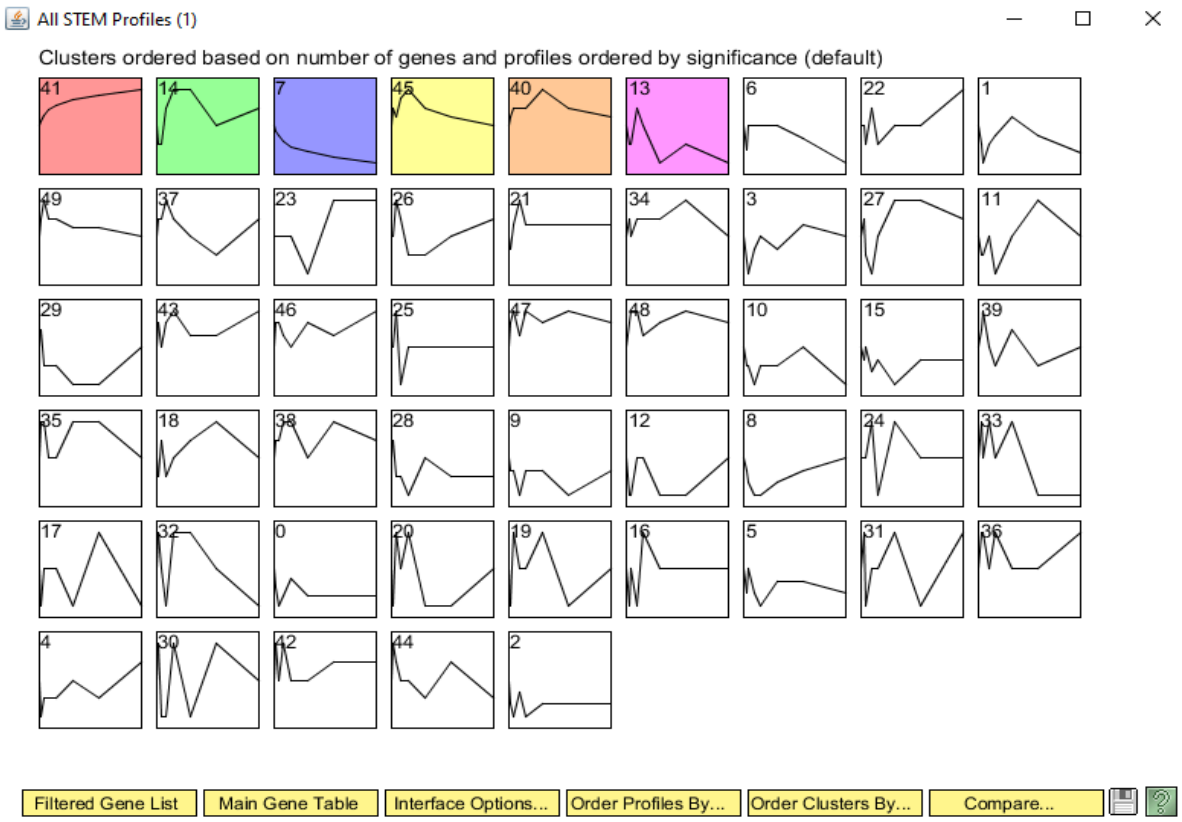


Figure 1. STEM profiles from CHP data. Colored backgrounds indicate significance.

Using STEM, 6 Significant cluster profiles were determined. The Sha et al. paper exhibited clusters different from the ones obtained using Stem. The figures in this paper exhibit transcriptional response in yeast treated with cumene hydroperoxide for the 0-20 minute time period, whereas our figure exhibited a transcriptional response from a much larger span of time, this being 0-120 minutes. Therefore, the clusters cannot be directly compared.

Figure 2:

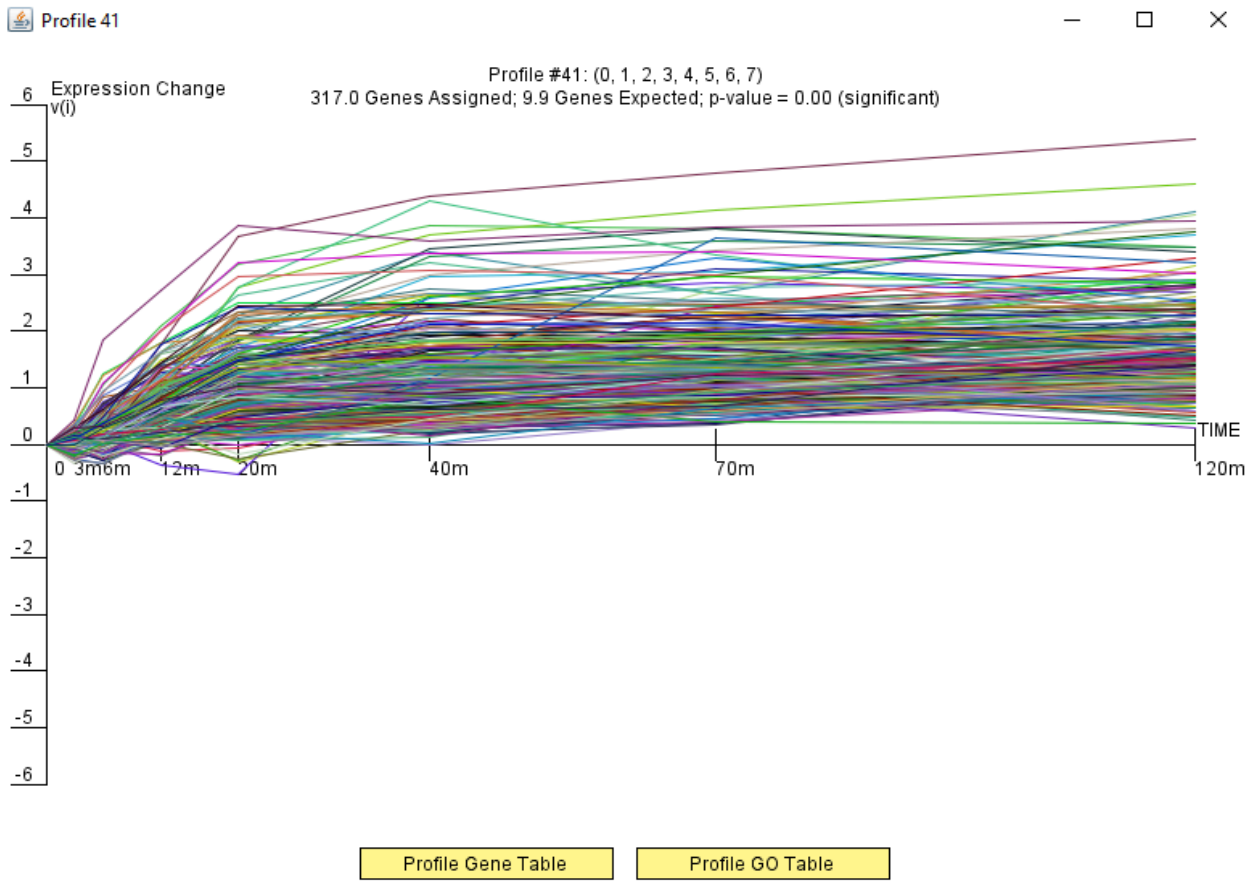


Figure 2. Profile 41 Stem cluster expression change over time. The general trend appears to be an increase in expression change at some middle time point that stays above 0 until 120 minutes.

Profile 41 was chosen to be examined due to having the highest amount of genes and a p-value equaling 0.0. Between 0-20 minutes we notice that some of the genes are repressed, but they are likely included in the cluster because their general trend is activation after 20 minutes.

Gene Ontology Enrichment Analysis Section (GO Terms):

Table 2. Significant GO Results for Cluster 41, their place in the *Saccharomyces cerevisiae* reference list from GeneOntology.org, and their FDR corrected p-value

GO Biological Process	Place in *Saccharomyces cerevisiae* Ref. List	FDR Corrected p-value
glutamate catabolic process (GO:0006538)	3	1.46E-06
protein localization to endoplasmic reticulum exit site (GO:0070973)	4	2.57E-06
regulation of fungal-type cell wall organization (GO:0060237)	20	3.50E-06
nucleotide transport (GO:0006862)	39	6.97E-06
cellular lipid catabolic process (GO:0044242)	50	1.89E-05

late endosome to vacuole transport via multivesicular body sorting pathway (GO:0032511)	51	9.23E-05
organonitrogen compound catabolic process (GO:1901565)	415	2.95E-04
cellular response to chemical stimulus (GO:0070887)	311	4.82E-04
RNA metabolic process (GO:0016070)	919	1.76E-02
ribonucleoprotein complex biogenesis (GO:0022613)	478	1.79E-02

After collecting the stem clusters, Cluster 41 was selected for further analysis of its associated GO terms. The GO enrichment tool at GeneOntology.org was used to do so. After copying all of the genes within 41's gene list, the genes were pasted into the "Go Enrichment Analysis" box, "Saccharomyces cerevisiae" was selected from the specie's drop-down menu, and "Launch" was clicked to obtain a table of the results. The table was then exported and opened in Excel, where the values were sorted by ascending p-value. The GO terms within the table were chosen by their low values of significance and their specificity, so that each GO term describes a unique process and not a more general one that would contain many significant processes due to its position in the GO term hierarchy. The p-values indicate a probability of there being a certain

number of genes out of the cluster's total genes that belong to that specific GO term, according to the number of genes within the total genome that belong to the GO term. The FDR correction refers to the false discovery rate method, which controls for false positives. The first GO Term that describes glutamate catabolic process may be due to glutamates role in metabolizing and contributing to the biosynthesis of new nucleic acids and proteins (Yelamanchi et al., 2016). This response has been previously associated with several types of stress. There are several terms within the table that are associated with molecular transport in the cell, like protein localization to endoplasmic reticulum exit site, and late endosome to vacuole transport via multivesicular body sorting pathway. In the presence of oxidative stress, the cell is likely altering which proteins are being translated for their stress environment, and the transport processes within the cell are similarly impacted. The paper did not mention the specific GO terms aside from their GO analysis table in their supporting information section, but they did observe genes that were up-regulated due to the CHP treatment, but not the H2O2 treatment, were genes that played a role in the cell wall. The GO term findings reflect this observation with the regulation of fungal-type cell wall organization. Sha et. al. believed this was due to CHP's large size, which would cause more damage to peripheral cell structures, as CHP was required to spend more time outside of the cell when attempting to penetrate the cell wall and plasma membrane and would potentially spend more time damaging these peripheral structures in the process.

GRN Construction and Analysis

Table 3. Transcription factors and their p-values from Yeastract.

Transcription Factor	p-value

Rpn4	0
Gcn4	0
Pdr1	0
Xbp1	0
Met28	0
Spt23	0
Bas1	0
Yap1	0
Sok2	0
Msn2	0
Fhl1	0
Pdr3	0
Cbf1	1.80E-14
Rph1	7.60E-14
Stp1	1.21E-13
Msn4	1.39E-12
Tec1	6.80E-12
Rgm1	8.50E-12
Stp2	9.49E-11

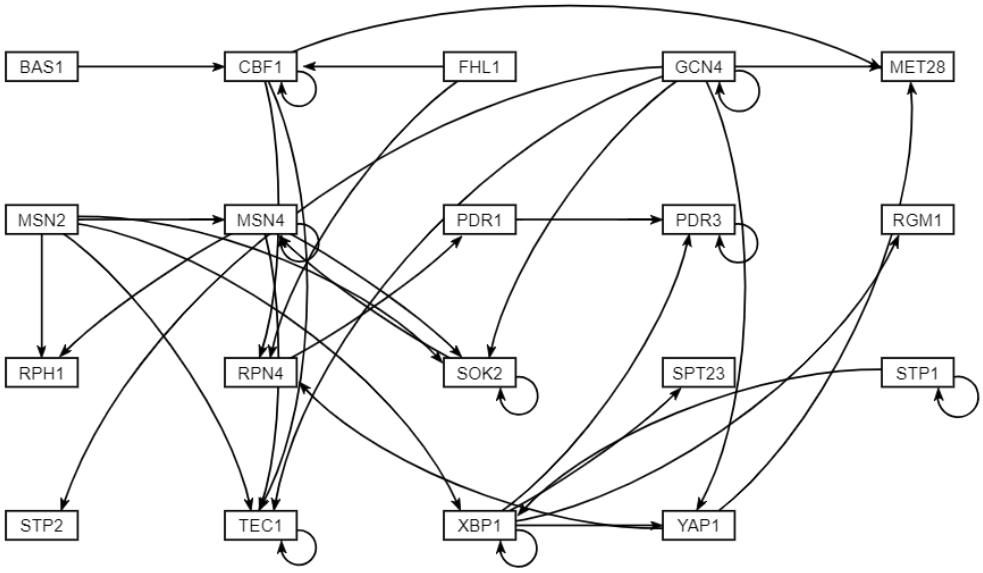
23 transcription factors were chosen from Yeastract because they had the most significant p-values. After running GRNsight, it was found that CRZ1, MET31, MGA2, and RLM1 lacked

connections to other transcription factors, so these were removed. The transcription factors mentioned in the Sha et al. paper include early up-regulated genes being HMS2, MET28, YAP5, NUT2, ROX1, and SUT2, as well as MET28, which regulates sulfur metabolism. According to the Sha et al. paper, MET28 targets the specific transcription factors MET1, MET12, MET16, MET22, MET3, MET8, CYS3 and STR3, with all of these transcription factors being activated during the 20 minute time. Furthermore, Sha et al. stated that the role of Yap1 could not be discovered, leaving the circumstances as to why unclear.

GRNmap and GRNsight Results:

Figure 3:

a.



b.

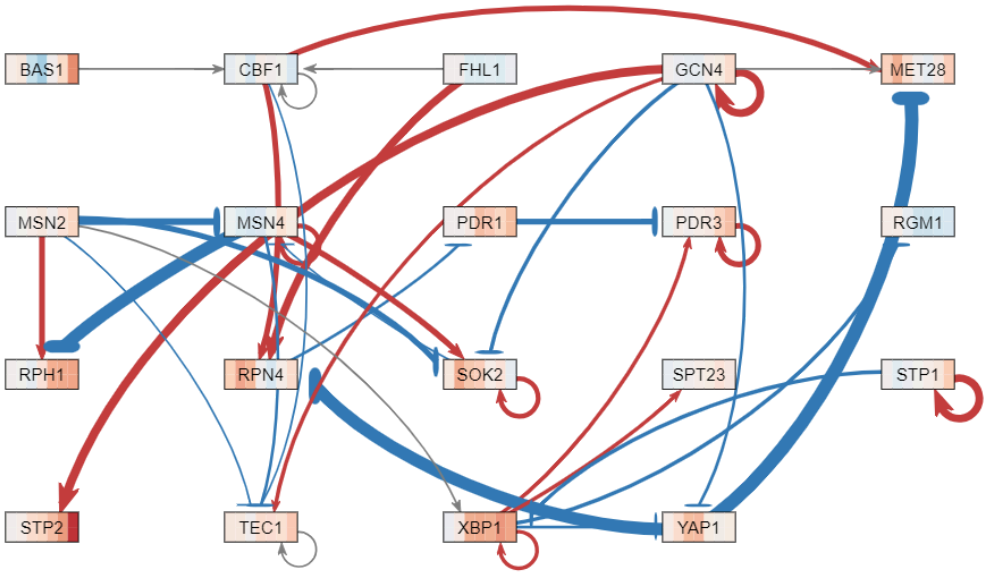


Figure 3. GRNmaps created using GRNsight. A. is the unweighted GRNmap base network. B. is the weighted GRNmap with node and edge coloring. Each node has vertical stripes indicating the log foldchange for timepoints 3, 6, 12, 20, 40, 70, and 120 minutes; red represents a foldchange above zero and blue represents a folchange below zero, and the color intensity corresponds to the foldchange absolute value. Red arrows indicate activation and blue blunt ends indicate repression. The edge thickness corresponds to the magnitude of the influence. Gray arrows indicate a small weight value of <5% of the value of the highest magnitude weight.

The LSE: minLSE ratio of the GRNmap was calculated as 0.114/0.044 = 2.59, which indicates how well the model performed. The model seems to indicate both similar and dissimilar findings with the Sha et. al. paper. MSN2/MSN4 and YAP1, which are important transcription factors in the response to different types of cell stress, are notable transcription factors in the network. In their study, Sha et al. found that 54% of the genes controlled by MSN2/MSN4 displayed a statistically significant change in the response to CHP treatment. The role that MSN2/MSN4 play in CHP response is indicated in the network, with both MSN2 and

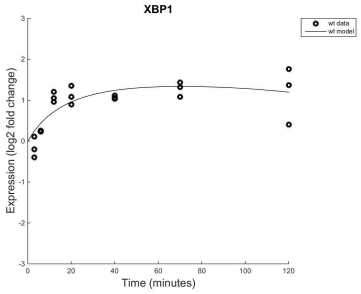
MSN4 showing several connections. MSN4 has a thick blue edge linking it to RPH1, indicating a high level of repressive influence. YAP1 also had two thick blue edges linking it to RPN4 and MET28, which shows it exerts a higher level of repression onto several transcription factors of the network. YAP1 is also experiencing a repressive influence from XBP1 and GCN4, though those edges are much thinner and indicate a lower magnitude of influence. Dissimilarities with the paper were concerning the transcription factors TEC1 and XBP1, which were not discussed in the Sha et. al. paper. In the network, TEC1 has the most incoming edges, and is therefore influenced by many transcription factors. Within the network, it is influenced by both activation and repression, which is shown by both red and blue incoming edges. TEC1 is associated with the cell stress response (Laloux et al., 1990). XBP1 had the most outgoing edges, meaning it influences many transcription factors. Within the network, it is shown to influence both activation and repression, as indicated by the red and blue edges stemming from it. XBP1 is involved in the glucose deprivation response, so its significance is likely due to the experimental conditions of glucose availability. As glucose is used by the cell and the supply of glucose is depleted, the cell must activate the functional processes for when glucose is low, so XBP1 acts as a transcriptional factor for these processes (Mai & Breeden, 1997).

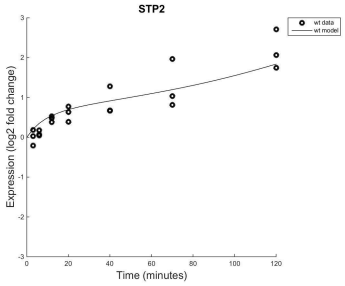
Table 4. Transcription factors and associated production rates, threshold_b parameters, and weights.

id	production_rate	threshold_b	BAS1	CBF1	FHL1	GCN4	MET28	MSN2	MSN4	PDR1	PDR3	RGM1	RPH1	RPN4	SOK2	SPT23	STP1	STP2	TEC1	XBP1	YAP1
BAS1	0.23928011	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CBF1	0.141918455	-0.004054	0.0023606	0.0261864	0.0077906	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
FHL1	0.165993127	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GCN4	0.179448151	2.2959737	0	0	0	1.4737829	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MET28	1.180343353	-0.436781	0	0.9072414	0	0.0550158	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-2.679106
MSN2	0.499693452	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MSN4	0.588067924	0.1937545	0	0	0	0	0	-1.302357	0.7000179	0	0	0	0	0	-0.156178	0	0	0	0	0	0
PDR1	0.397167218	-0.058333	0	0	0	0	0	0	0	0	0	0	0	-0.248708	0	0	0	0	0	0	0
PDR3	0.463526326	1.1879687	0	0	0	0	0	0	0	-0.997209	0.91425	0	0	0	0	0	0	0	0	0.5215401	0
RGM1	0.673868281	-0.082231	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.453609	0
RPH1	0.99517551	0.3203179	0	0	0	0	0	0.9023783	-2.565043	0	0	0	0	0	0	0	0	0	0	0	0
RPN4	1.311144413	0.0997009	0	0.917696	1.3826574	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-2.62295
SOK2	1.44568562	0.5262992	0	0	0	-0.459003	0	-1.086029	0.9763196	0	0	0	0	0	0.8455469	0	0	0	0	0	0
SPT23	0.220168809	0.4339373	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5399173	0
STP1	0.272479149	2.1967252	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.4498513	0	0	0	0
STP2	0.704644667	3.0721186	0	0	0	1.7523658	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TEC1	0.708734639	0.1242618	0	-0.16308	0	0.628896	0	-0.205878	-0.296196	0	0	0	0	0	0	0	0	0	0.102119	0	0
XBP1	0.361774639	0.3605913	0	0	0	0	0	0.0604992	0	0	0	0	0	0	0	0	-0.601095	0	0	0.6361475	0
YAP1	0.473803172	-0.189354	0	0	0	-0.314917	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.331754	0

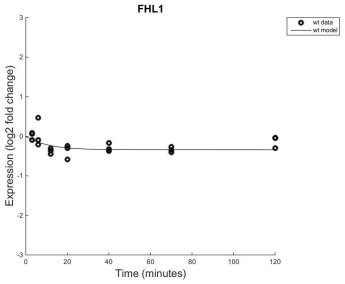
Figure 4:

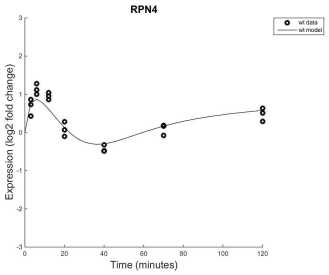
In Cluster 41:

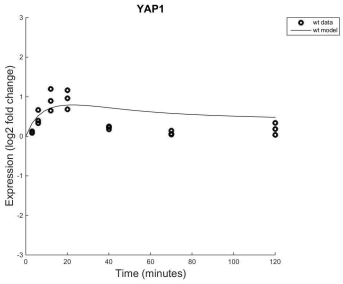


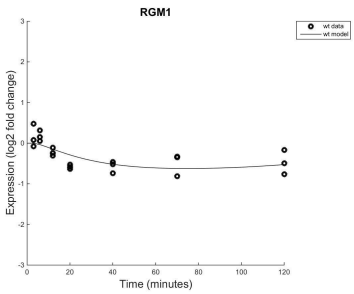


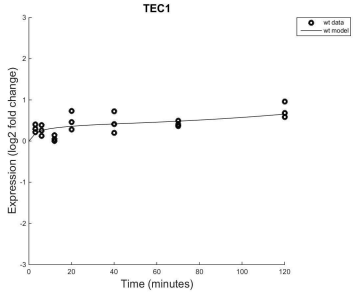
Not in Cluster 41:

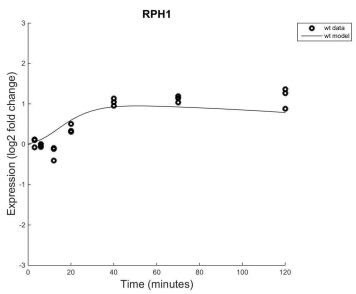


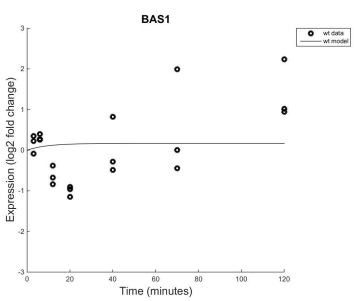


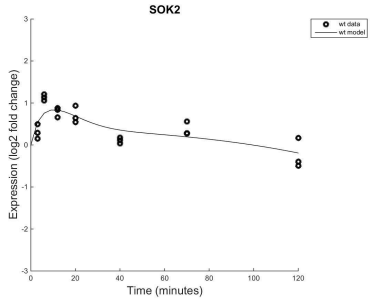


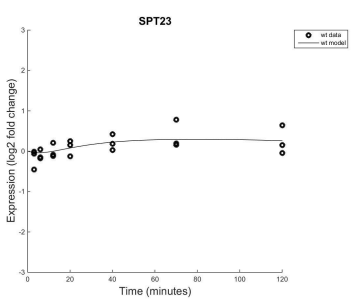












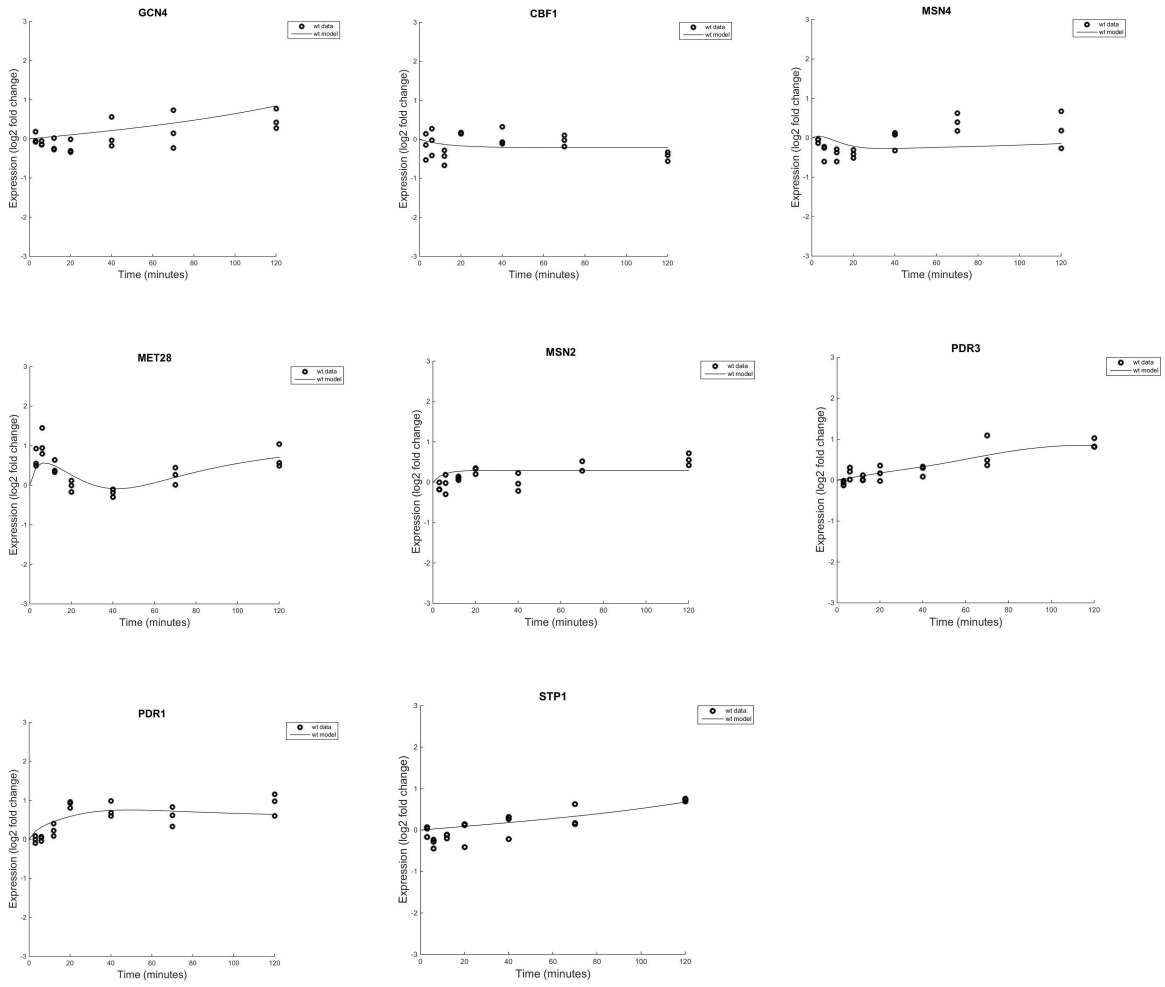


Figure 4. Individual expression plots of each transcription factor. Of the 19 genes, only XBP1 and STP2 were within Cluster 41 while acting as a transcription factor for that cluster.

There was not an obvious trend for the expression plots of each gene, but it is notable that for most of the plots, any changes occur within 20 or 40 minutes of the treatment. Our findings were in common with the Sha et al. paper regarding transcription factor MET28, which they stated was an “early up-regulated gene”. As observed in the network, Met28 is activated at the 3 minute time point by CBF1 and GCN4 which had not been discussed by Sha et al. This is

mirrored in the expression plot. In RPN4's plot, it can be seen that the gene was activated from minutes 0-20, slightly repressed from minutes 20-40, and then resumed a general trend of activation for the remaining time points. The expression plot agrees with Sha et al., who had observed an early peak in expression for the RPN4 gene. MSN2/MSN4 show very little variation in their plots, which corroborates with the Sha et. al. paper, who had stated that while the genes regulated by these factors were significantly impacted, MSN2 and MSN4 themselves did not show any significant expression change. As noted above, Cluster 41 contained genes with a general trend of activation in response to CHP stress. Genes XBP1 and STP2 were the only transcription factors that were a part of Cluster 41, and their expression plots mirror this, with both genes showing a trend of activation in their individual plot.

Database Development:

The creation of the database required each of the following softwares: Microsoft Access, Microsoft Excel, and BOX. Each of the tables in the database were first imported into excel. The first table that was necessary was a gene table. Our main table, "Gene Table", allowed us to organize and connect all the data into a central location. The gene table contained information about each gene in the yeast genome, including the systematic ID of the gene, the standard name of the gene, the gene's primary database ID, the Gene name, the gene's feature type, as well as a brief description of the gene. The next table required was an expression table of the yeast data when treated in a control group and with CHP, which was created by the data analysts and was imported into access as well as two tables. One table was for the control group, and the other was for the CHP group. There was also a degradation rates table that included information of the degradation rate of each of the genes, this data was obtained from Neymotin et al. (2014). This table was used to calculate the respective production rates of each of the genes, which was

calculated by our professor. Another table in the database was the network table which was made from Harbison et al. (2004) data which included data on each of the genes and their respective binding capability to 203 transcriptional regulators.

Table 5: Network table in Microsoft Access, each data point in this table was a p value which was converted to either 0 or 1 based on if the p value was less than .001. A p value of 1 means that the data is significant and a data point of 0 means it is insignificant.

Systematic Name	A1 (MATA1)	ABF1	ABT1	ACA1	ACE2	ADR1	AFT2	ARG80	ARG81	ARO80
YAL001C	0	0	0	0	0	0	0	0	0	0
YAL002W	0	0	0	0	0	0	0	0	0	0
YAL003W	0	0	0	0	0	0	0	0	0	0
YAL004W	0	0	0	0	0	0	0	0	0	0
YAL005C	0	0	0	0	0	0	0	0	0	0
YAL007C	0	0	0	0	0	0	0	0	0	0
YAL008W	0	0	0	0	0	0	0	0	0	0
YAL009W	0	0	0	0	0	0	0	0	0	0
YAL010C	0	0	0	0	0	0	0	0	0	0
YAL011W	0	0	0	0	0	0	0	0	0	0
YAL012W	0	0	0	0	0	0	0	0	0	0
YAL013W	0	0	0	0	0	0	0	0	0	0
YAL014C	0	0	0	0	0	0	0	0	0	0
YAL015C	0	0	0	0	0	0	0	0	0	0
YAL016W	0	0	0	0	0	0	0	0	0	0
YAL017W	0	0	0	0	0	0	0	0	0	0
YAL018C	0	0	0	0	0	0	0	0	0	0
YAL019W	0	0	0	0	0	0	0	0	0	0
YAL020C	0	0	0	0	0	0	0	0	0	0
YAL021C	0	0	0	0	0	0	0	0	0	0
YAL022C	0	0	0	0	0	1	0	0	0	0
YAL023C	0	1	0	0	0	0	0	0	0	0
YAL024C	0	0	0	0	0	0	0	0	0	0
YAL025C	0	0	0	0	0	0	0	0	0	0
YAL026C	0	0	0	0	0	0	0	0	0	0
YAL027W	0	0	0	0	0	0	0	0	0	0
YAL028W	0	0	0	0	0	0	0	0	0	0
YAL029C	0	0	0	0	0	0	0	0	0	0
YAL030W	0	0	0	0	0	0	0	0	0	0
YAL031C	0	0	0	0	0	0	0	0	0	0
YAL032C	0	0	0	0	0	0	0	0	0	0
YAL033W	0	0	0	0	0	0	0	0	0	0
YAL034C	0	0	0	0	0	0	0	0	0	0
YAL034W-A	0	0	0	0	0	0	0	0	0	0
YAL035C-A	0	0	0	0	0	0	0	0	0	0
YAL035W	0	0	0	0	0	0	0	0	0	0
YAL036C	0	0	0	0	0	0	0	0	0	0
YAL037W	0	0	0	0	0	0	0	0	0	0
YAL038W	0	0	0	0	0	0	0	0	0	0

To help with clarity for anyone validating our data or using our database, we created a metadata table which describes and explains each of our Access worksheets. The design of the table includes these fields: Primary Key, Description, Data Source, Date Accessed, Date Uploaded, and Publication (Pubmed ID) or DOI. The Primary Key for the metadata table uniquely identifies each entry in the metadatable and corresponds to each of the entries. The description field explains what each table is. The data source section specifies where the data for

the corresponding table was accessed or downloaded. The date accessed shows when the corresponding data table was accessed and downloaded. The date updated section shows the date the most recent date of update for the related data if specified by the source and applicable. The publication (pubmed id) or DOI section has the related link to access the related file and paper.

Figure 5:

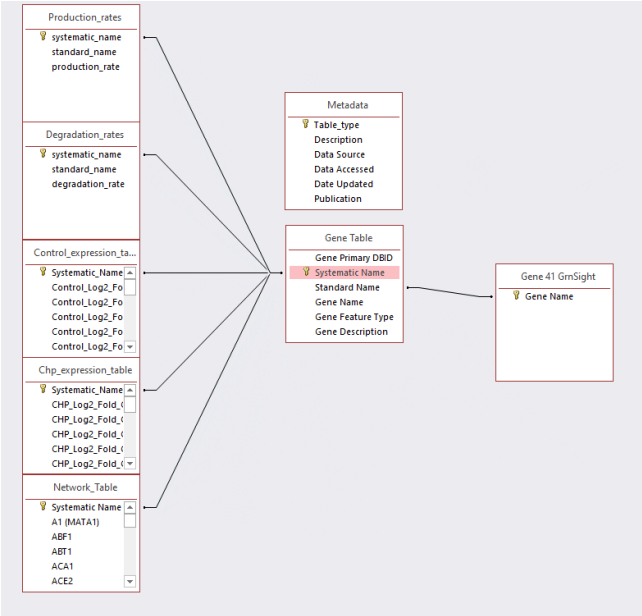


Figure 5. Relationships window of the database in Microsoft Access. Gene 41 GrnSight table was a table of Gene names that was not part of the database but was used for querying the database when creating the GRNmap input. The columns that connected each of the tables through the relationships was the gene's systematic name; this was also chosen to be each table's primary key.

During the process of creating the database, we also encountered a few challenges. One challenge was related to how we were importing data and manipulating it. One team member programmed custom queries for data imports, but Dean mentioned that he had a preference for using the Access graphical user interface for these queries, due to his concern about the code's complexity and the extra steps required to scale the data at a later date with the statically coded method. To address the issue, we went with the GUI approach so that we would be focusing on the ease of use and future scalability of the database.

Another challenge that we had to deal with was the formatting of data and issues with importing data. One issue we encountered was an issue with our degradation rates and production rates. They were appearing as 0's and 1's in Access initially. We pretty immediately recognized this was an error with rounding, and with the help of Dr. Dahlquist, we realized the specific culprit was an import error. We solved this issue by directly importing from a text file instead of an Excel file. We also made sure the data types matched in this new upload of data.

We also had issues with the network table. The issue stemmed from some null values that were stopping us from creating a primary key. Since every primary key in Access needs to be unique and there were null values, the database would not allow us to proceed. To solve this error -which was an issue in Excel where there were empty rows - we had to go back and delete empty rows that did not hold any actual data but were at the bottom of our Excel file. When the Excel file was exported to a text file, they were remaining, and then getting imported into Access. This taught us the importance of checking the data for errors and any housekeeping that might need to be done before importing into Access.

There were also issues with preprocessing the data before we even imported anything to Access. The Harbison paper for example, was missing shorthand names, and descriptions, and it

had #REF entries on excel. We looked at different potential solutions like using YeastMine, but eventually spoke with Dr. Dahlquist and came up with the solution of just removing the #REF entries. We also encountered 'NaN' values, which led us to some questions about how the data we got from Harbison even had NaN values in it. After speaking about this together and with Dr. Dahlquist, we realized we could ignore this issue and just treat them as not significant in our analysis of the data since this would not give a false positive on influence.

We also ran into difficulties in integrating data across different sources. To make sure everything was compatible between the Harbison paper and the other datasets in the project we had to carefully consider what would be our primary key. We used the Gene ID column because it was the field that was most consistent across all of our tables since this is the Gene ID all of the different things we looked at were associated with. This allowed us to make sure everything was integrated and reduced the need to make lots of separate database connections.

There were also issues in running queries in the database. We created a GRNmap input of the data from profile 41, which was selected by the data analysts since it has the most genes of the significant profiles they found. The profile contained 23 genes in which the coders had to run queries to determine certain attributes of the genes to make a GRNmap network. There was a sample GRNmap workbook that we used as a template. The first table in the workbook was the production rates of each of the 23 genes. The second was the degradation rates of the genes. The third was the expression values for the genes treated with CHP. The fourth was the network table values for the 23 genes, there was also a network weights table that was identical to the network table. Two more tables did not require queries since we only had to fill out the fields respective to the project including the time points. And one more table which just needed a 0 next to every gene. Andrew first completed the first 3 queries, however it was noticed that his results did not

include all 23 genes. It was also noticed that his methods included writing out all of the syntax for the queries in SQL mode which was far more extensive than it needed to be. The correct way, and easier way of doing the queries was in design mode. This was done by creating another table with all of the genes from profile 41 in the database. This table only included their standard names. When it was first attempted to run queries selecting their outputs from the production rates table, it was realized that the production rates table did not include their standard names. It was then attempted to connect the profile 41 table to the gene table and then to the production rates table, which it was then realized that the gene table was missing the standard names of the genes. A new version of the gene table was created with the standard names of the genes. Once this had been done, another relationship was able to be created that connects the standard names of the genes from profile 41 to the standard names of all of the genes in the genome. Which was then able to be connected to the production rates table. Once this was done, running each of the queries was very simple and worked very smoothly.

Figure 6:

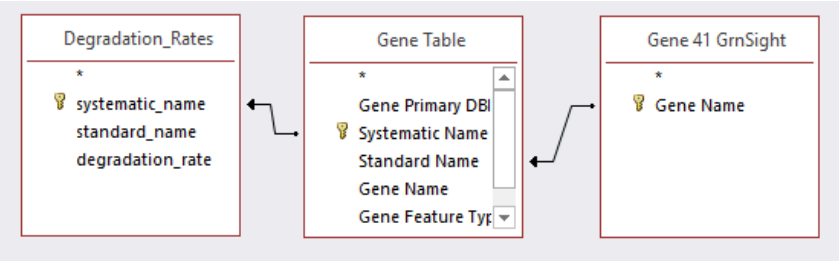


Figure 6. Query Design in Microsoft Access which connects the gene standard name from profile 41 to the standard name in the gene table and then to the systematic name in the degradation rates table

Table 6: Query results table in Microsoft Access which shows the degradation rates for all genes in profile 41.

systematic_name	standard_name	degradation_rate
YKR099W	BAS1	0.1066
YJR060W	CBF1	0.0835
YNL027W	CRZ1	0.1238
YPR104C	FHL1	0.105
YEL009C	GCN4	0.0513
YIR017C	MET28	0.0542
YPL038W	MET31	0.055
YMR037C	MSN2	0.2039
YKL062W	MSN4	0.1386
YGL013C	PDR1	0.1083
YBL005W	PDR3	0.1359
YMR182C	RGM1	0.2666
YPL089C	RLM1	0.2236
YER169W	RPH1	0.105
YDL020C	RPN4	0.1136
YMR016C	SOK2	0.4332
YKL020C	SPT23	0.1284
YDR463W	STP1	0.0845
YHR006W	STP2	0.0912
YBR083W	TEC1	0.2773
YIL101C	XBP1	0.0912
YML007W	YAP1	0.0835

QA Process:

Throughout the project, quality assurance was an important factor because it ensured that the data being analyzed by the data analysts was satisfactory in order for the coder/designers to create the database. The process began with the coders/designers inputting data from the Saccharomyces Genome Database (SDG) to create a “sample-data relationship” that contained all of the samples. The data was organized in an Excel workbook and appropriate header

columns were given to summarize the information found from the dataset including the experimental description.

The first issue encountered was whilst filtering the ‘Network Table’; we weren’t able to convert the p-values to 0 or 1 to see the significance. We tried to set it up so that if the p-value was less than 0.00, it would be 1 and if the p-value was more than 0.001, it would be 0. The =IF(pval<0.01,1,0) did not work on Dean’s, Natalija or Dr. Dahlquist's computers but it worked on Andrew’s personal computer and he got 6 results which is what Dr. Dahlquist originally found, so his file was used to complete the rest of the project. Some of the tables had errors such as #REF and blanks, however upon discussion, the columns that contained them were deleted because they weren’t important for the creation of the database. The workbooks were saved as text files and we began importing the tables into Access.

With Access, the first problem we ran into was when inputting the data, specifically the “Degration_rates” table and the “Production_rates” table. The rates for both tables kept showing up as “0” even though in the Excel workbook, they were showing as “0.21”. We thought that this was because of the decimal place setting, and tried to change it. However, we later realized that it was because instead of importing the data, we copy-pasted it. After importing the text file into Access, we no longer had that issue. The Coders/Designer created the following tables: Gene, Expression, Degration_rates, Production_rates, network and metadata. We ran into another issue when inputting the expressions data from the Data Analysts,we realized that the data contained some duplicate columns and the Data Analysts fixed it before it was imported again. Once everything was imported quality assurance was performed to make sure that all of the information was there and that it was correct. More specifically, we made sure that all of the

tables had the appropriate rows and that the Expressions Table had both the ID and Systematic Names included, and that the schema contained all of the appropriate relationships.

Conclusion:

The process of analyzing the transcriptional response to oxidative stress in yeast was outlined in this project, involving various stages of experimental design, database creation and quality assurance to construct a final GRNmap model to determine which transcription factors belong in the gene regulatory network that controls the OSR to CHP in yeast. This project began by exploring the transcriptional response over time to CHP-induced oxidative stress in *S. cerevisiae* by Sha et al. Using their DNA microarray data and Harbison et al.'s transcriptional regulator data, this project set out to create a GRNmap to show the regulatory mechanisms involved in the yeast OSR to CHP.

The Microsoft Access database that we created allowed us to identify transcription factors and their regulatory interactions to construct the GRNmap in the form of a GRNmap input workbook. While we were unable to generate enough edges using the Harbison et al. data, we used 2024-03-19 version of the Saccharomyces Genome Database regulation data and GRNsight to generate a GRNmap. We were then able to identify and verify transcription factors within our cluster that are involved in the OSR to CHP. Out of the 19 predicted connected transcription factors, XBP1 and STP2 were the only two that were verifiably present in the GRN.

In the future it would be useful to compare CHP to control data directly, performing t-tests for each timepoint. It could also be useful to look at narrower time sets, like before 20 minutes, and after 20 minutes, and compare the two to determine exactly if there are two distinct stress responses before and after the time between 20 and 40 minutes.

Upon the completion of the project, we gained a better understanding of the transcriptional response of oxidative stress in yeast and also learned important skills such as communication, organization, and time management when working as a team.

Acknowledgments

Dr. Kam Dahlquist taught the BIOL 367 course and co-created the curriculum with Dr. John Dionisio.

Dr. Kam Dahlquist assisted our team with understanding the material as well as creating the GRNmaps for us based on our GRNmap input workbook.

Dr. John Dionisio walked us through how to use Access.

References

Balakrishnan, R., Park, J., Karra, K., Hitz, B. C., Binkley, G., Hong, E. L., Sullivan, J., Micklem, G., & Michael Cherry, J. (2012). YeastMine—An integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit. *Database*, *2012*. https://doi.org/10.1093/database/bar062

Ernst, J., Patek, D., & Bar-Joseph, Z. (n.d.). *STEM: Short Time-series Expression Miner* (1.2) [Computer software].

GRNsight v6.0.7. (2022, April 19). GRNsight. http://dondi.github.io/GRNsight/

Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J.-B., Reynolds, D. B., Yoo, J., Jennings, E. G., Zeitlinger, J., Pokholok, D. K., Kellis, M., Rolfe, P. A., Takusagawa, K. T., Lander, E. S., Gifford, D. K., Fraenkel, E., & Young, R. A. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature*, *431*(7004), 99–104. https://doi.org/10.1038/nature02800

Jamieson, D. J. (1998). Oxidative stress responses of the yeast Saccharomyces cerevisiae. *Yeast*, *14*(16), 1511–1527. https://doi.org/10.1002/(SICI)1097-0061(199812)14:16<1511::AID-YEA356>3.0.CO;2-S

Laloux, I., Dubois, E., Dewerchin, M., & Jacobs, E. (1990). TEC1, a Gene Involved in the Activation of Ty1 and Ty1-Mediated Gene Expression in Saccharomyces cerevisiae: Cloning and Molecular Analysis. *Molecular and Cellular Biology*, *10*(7), 3541–3550. https://doi.org/10.1128/mcb.10.7.3541-3550.1990

LMU BioDB 2024. (2024). Final Project. Retrieved May 2, 2024, from https://xmlpipedb.cs.lmu.edu/biodb/spring2024/index.php/Final_Project

Mai, B., & Breeden, L. (1997). Xbp1, a Stress-Induced Transcriptional Repressor of the *Saccharomyces cerevisiae* Swi4/Mbp1 Family. *Molecular and Cellular Biology*, *17*(11), 6491–6501. https://doi.org/10.1128/MCB.17.11.6491

Microsoft Access (Access 2016). (2016). [C++]. Microsoft.

Pierce, B. A. (2020). *Genetics: A conceptual approach* (Seventh edition). Macmillan Learning.

Saccharomyces Genome Database | SGD. (n.d.). Retrieved May 2, 2024, from https://www.yeastgenome.org/

Sha, W., Martins, A. M., Laubenbacher, R., Mendes, P., & Shulaev, V. (2013). The Genome-Wide Early Temporal Response of Saccharomyces cerevisiae to Oxidative Stress Induced by Cumene Hydroperoxide. *PLoS ONE*, *8*(9), e74939. https://doi.org/10.1371/journal.pone.0074939

Timoshina, N. (2023, February 8). *What is SQL Database: Structure, Types, Examples*. https://www.alphaservesp.com/blog/what-is-sql-database-structure-types-examples

Teixeira, M. C., Viana, R., Palma, M., Oliveira, J., Galocha, M., Mota, M. N., Couceiro, D., Pereira, M. G., Antunes, M., Costa, I. V., Pais, P., Parada, C., Chaouiya, C., Sá-Correia, I., & Monteiro, P. T. (2023). YEASTRACT+: A portal for the exploitation of global transcription regulation and metabolic model data in yeast biotechnology and pathogenesis. *Nucleic Acids Research*, *51*(D1), D785–D791. https://doi.org/10.1093/nar/gkac1041

Yelamanchi, S. D., Jayaram, S., Thomas, J. K., Gundimeda, S., Khan, A. A., Singhal, A., Keshava Prasad, T. S., Pandey, A., Somani, B. L., & Gowda, H. (2016). A pathway map of glutamate metabolism. *Journal of Cell Communication and Signaling*, *10*(1), 69–75. https://doi.org/10.1007/s12079-015-0315-5