Visualizing the Impact of Arsenite on *Saccharomyces cerevisiae* Strain: W303-1A Through Multiple Classification Methods

### By: Marcus Avila, Ivy Macaraeg, DeLisa Madere,

### John “Joey” Nimmers-Minor, Naomi Tesfaiohannes

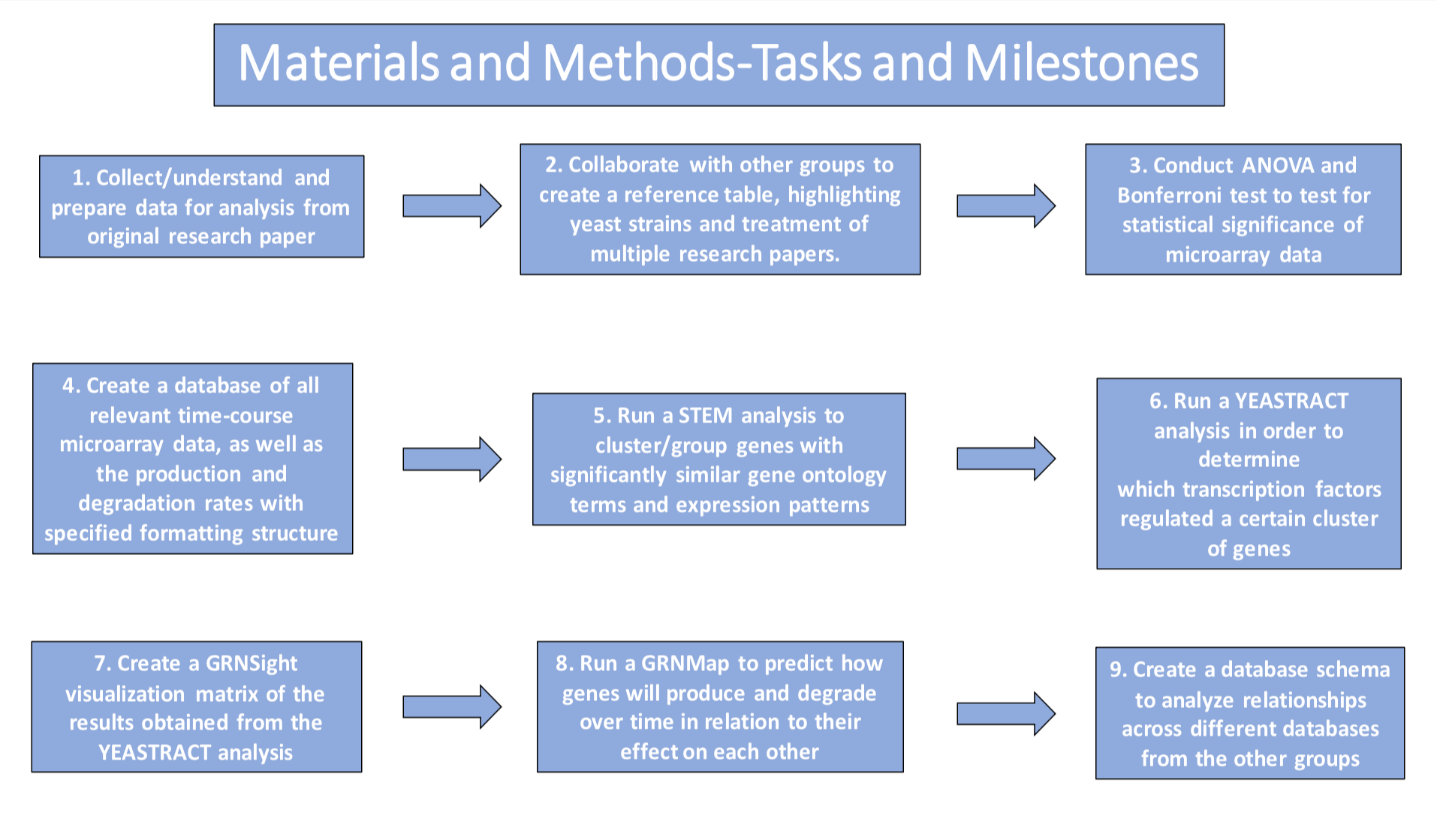
**Introduction**

This investigation observed how *Saccharomyces cerevisiae* changed gene expression tendencies in response to arsenic exposure. Previous literature has found that arsenic is present in nature and therefore cells have been able to evolve to tolerate the toxicity. Arsenic is a highly toxic metalloid that can threaten the environment and human health. Transcription of genes encoding functions is affected by arsenite exposure. The researchers found that different functions affected were protein biosynthesis, arsenic detoxification, oxidative stress defense, redox maintenance, and proteolytic activity. *Saccharomyces cerevisiae* is able to avoid arsenic toxicity by increasing efflux of As(III) with the protein Acr3p. Acr3p then sequesters glutathione-conjugated As(III) in the vacuole with the transporter Ycf1p. The transcription regulator Yap1p can control the transcription of genes encoding proteins with antioxidant properties. Yeast cells take up extracellular sulfate and metabolize it through sulfur assimilation. The researchers found that As(III)-exposed cells can channel a large part of assimilated sulfur into GSH biosynthesis. Two transcriptional regulators Yap1p and Met4p can control the response of GSH biosynthesis. Met4 is significant in cadmium tolerance, controlling sulfur assimilation, and glutathione biosynthesis. It can play a role in As(III) tolerance but is more significant with cadmium. Yap1p can regulate transcription for protein encoding genes with antioxidant properties, including *YCF1* and *ACR3*. After reviewing the researcher’s results, the Sulfiknights decided to reanalyze the yeast microarray data by providing our own contributions that include our own data analysis.

Regulation of genes in a stimulus-generated response cause for the increased or decreased transcription of certain genes in order to either increase the stimulus’ effect or decrease it. The papers in our annotated bibliography enforced our knowledge on the subject of gene regulation in transcription as we moved towards analyzing our journal club paper. In the example of our journal club article, the stimulus that drives the response is arsenite toxicity being introduced to wild-type *S. cerevisiae*. Through microarray re-analysis, we hope to better understand how the genes directly affect each other in its path or induction and repression as well as looking to see what the pattern of gene expression will be over time given the production and degradation rates of the gene products. The paper describes what transcription factors played a major role in the arsenite response; however, we hope to find information on the gene regulatory network that the original authors could not find given their older dated equipment. In order to find unknown/unaccounted for genes in the regulation matrix, we will utilize GRNsight, which allows for a visualization of the gene regulatory matrix created by YEASTRACT. Further, in order to test how the genes will act on each other in the future, GRNmap provides a well-detailed explanation of gene regulation over a time-course, and we will be utilizing it for that reason. The original journal club paper hoped to answer the question: How does a cell use gene regulation in order to respond and combat toxicity stemming from the metalloid, arsenic within a compound known as arsenite? In order to accomplish the goal we are attempting to achieve, we will utilize Microsoft (MS) Access, a database application, in order to centralize all data on *S. cerevisiae* gene expression, product degradation and production, as well as the results collected from the Thorsen et al. (2007) paper. This database is also necessary to create the GRNmap analysis that was mentioned earlier because factors including gene expression data, rate of production and degradation, and other predicting aspects are needed in order to create an accurate prediction of how the gene regulatory matrix will continue to effect itself over the course of time, using the time data of the original journal article to guide how long we plan for this prediction to go. The original journal article took time points at 15 minutes, 30 minutes, 60 minutes and 1080 minutes, therefore, these time points were the ones accounted for in the GRNmap.

**Methods, Results, and Discussion**

The next sections describe the workflow for completing this project as well as the results from the project that Sulfiknights came up with from their Data Analysts, Quality Assurances, and Designer.



**Figure 1**: Tasks used to keep track of Sulfiknights’ progress as well as to analyze and organize data. Milestones were completed over a course of a few weeks.

The materials and methods flowchart above demonstrate the steps we took to create a re-analysis of the data. In step 1 we collected and prepared for analysis of the paper by downloading the data into excel. In step 2 we worked with the other groups to have consistent column headings for the datasheets. We also highlighted the yeast strains and what treatments were being use for each of the groups. In step 3 we conducted an ANOVA and Bonferroni test. These were used to test for statistical significance of microarray data. In step 4 we created a database of all the relevant data as well as included two other sheets, the production and degradation rates. In step 5 we ran a STEM analysis to observe clusters/groups of genes with significantly similar gene ontology terms and expression patters. In step 6 we ran a YEASTRACT analysis. This determined what transcription factors regulated a cluster of genes. In step 7 we created a GRNsight visualization matrix of the results that occurs from the YEASTRACT analysis. In step 8 we ran a GRNMap to predict how genes produce and degrade over time in relation to each other. In step 9 we created a database schema to analyze relationships across different databases from other groups.

*Data Analysts: Data Organization*

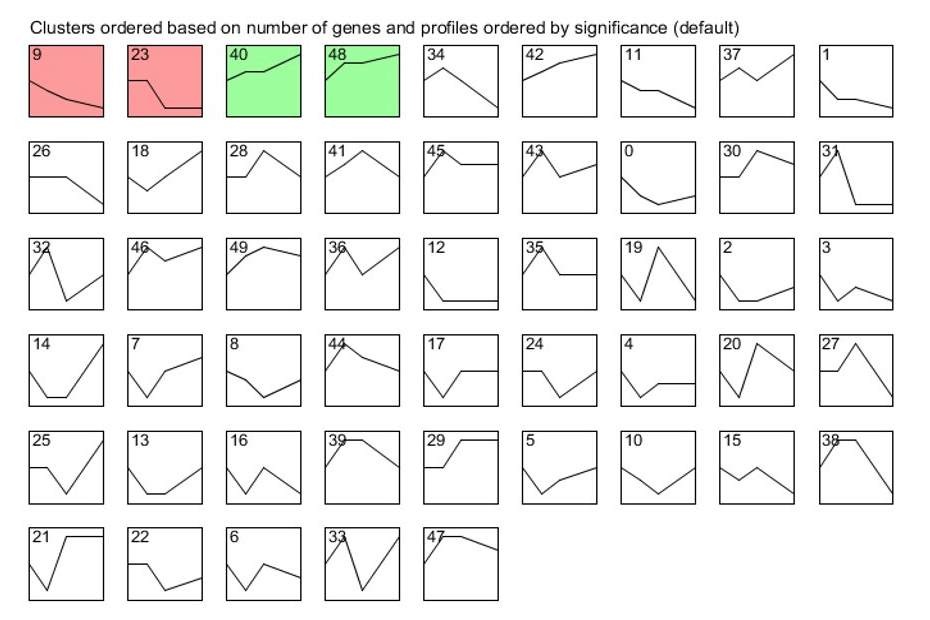
The first step in analyzing the data was to perform statistical analyses. The Excel data file that contained all the data sets was downloaded, and each of the time point headers were changed to fit a more comprehensive layout. From here, we only performed analysis on the “GSE6068\_setA\_family” data sheet. This sheet contained data concerning non-stressed wild type versus stressed wild type at the 15, 30, 60, and 1080-minute time points. “Stressed” referred to the exposure of the *S. cerevisiae* cells to 1mM As(III), and there were three replicates for each of the 15, 30 and 1080 minute time points and six replicates for the 60 minute time points. ANOVA analysis performed on this data included the calculation of p-values and the correction of these p-values by the Benjamini & Hochberg and Bonferroni standards. Based on our analyses, only a certain number of genes were statistically significant (Table 1).

|  |  |
| --- | --- |
| ANOVA | SwtVnwt\_1mM (numbers out of 4785) |
| p-value< 0.05 | 1067 (22.3%) |
| p-value < 0.01 | 534 (11.2%) |
| p-value < 0.001 | 239 (4.99) |
| p-value < 0.0001 | 112 (2.34) |
| B&H p-value < 0.05 | 375 (7.84%) |
| Bonferroni p-value < 0.05 | 25 (.0522%) |

**Table 1**: ANOVA results showing significant changes in gene expression. ANOVA results find varying levels of statistical significance for stressed WT and non-stressed WT.

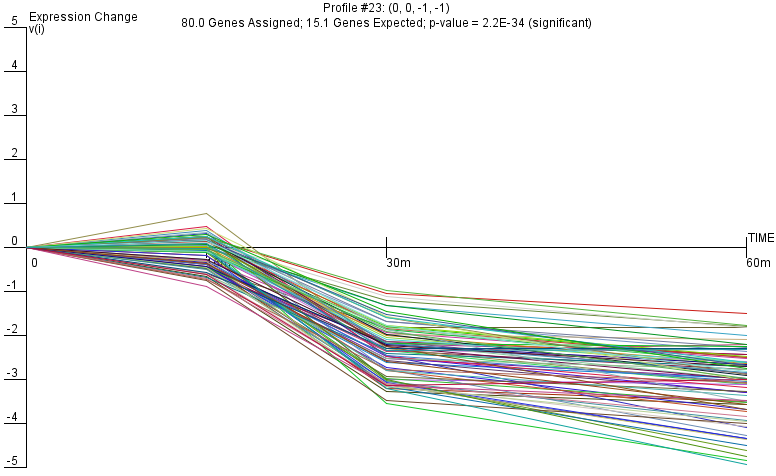
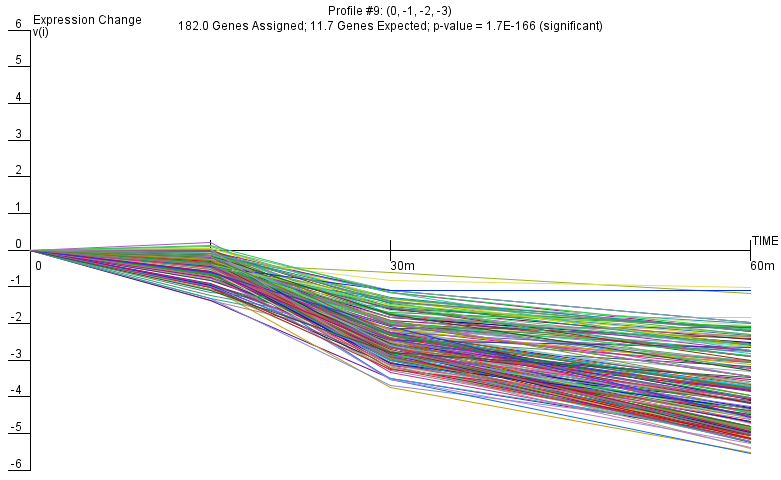
Only about 20% of the initial, unadjusted p-values were significant, and this was a much smaller ratio than the p-values of other groups. Based on the most stringent parameters, the Bonferroni calculation, less than 1% of the genes contributed to a significant change in gene expression. However, the authors in the paper did not use standard ANOVA methods to test for significance because they used a general applied model. This involved the use of the equation: log[(Prob(Xg=1))/(1-Prob(Xg=1)] = βYg+F(Lg) where “Prob” is probability, “Xg” is the variable that indicates the presence of a promotor motif, “Yg” is the variable that indicates if the gene is regulated, “Lg” is the length of the promotor, and “β” and “f” are based off of statistical language R. P-values were based on β=0. This method of analysis was used to get rid of biases within p-values based on the promotor regions of genes involved in up/down-regulation; however, the paper does not mention how many of the 5,000 analyzed where considered significant (Thorsen et al., 2007).

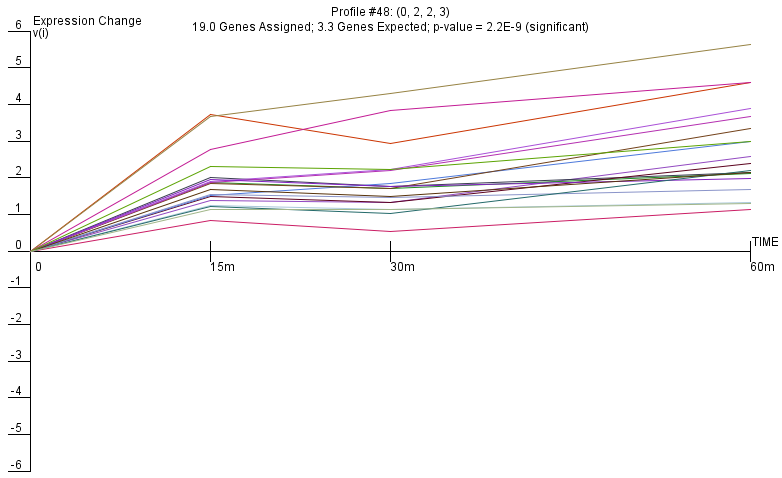
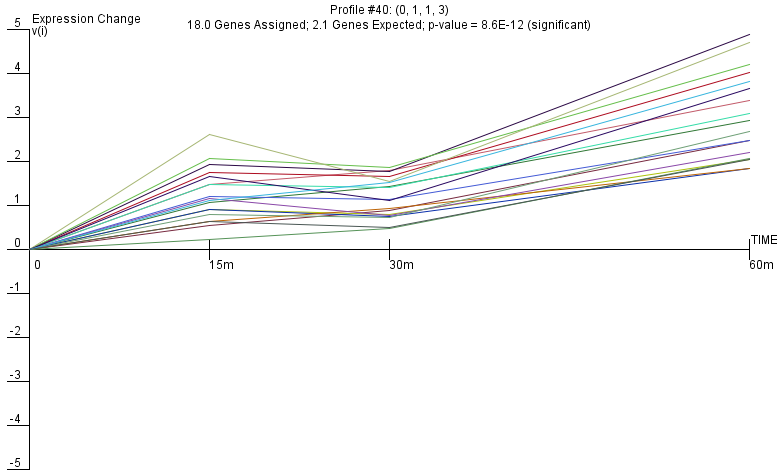
After ANOVA analysis, a STEM worksheet was created and inputted into the STEM program. For graphic visualization ease, the 1080-minute time point was deleted, but gene expression for each of the profiles returned to normal by this time. All resulting profiles are shown in Figure 1.



**Figure 2**: STEM produced 4 significant profiles. Profiles 9 and 23 both show similar gene repression, while Profiles 40 and 48 both show similar gene expression increase so each pair was combined for further analysis.

Out of the 50 clusters, only four of them were considered significant. These were Profiles 9, 23, 40, and 48. Profiles 9 and 23 contained many more genes, about less than 200 in total, than Profiles 40 and 48 which was about 50 genes in total. These pairs of clusters, Profiles 9 and 23 and Profiles 40 and 48, were involved in down-regulation and up-regulation respectively of gene expression (Figure 2). Instead of running their analysis through STEM, researchers in the paper instead used previous literature as well as the *Saccharomyces* Genome Database to identify 132 regulatory motifs that pertained to about 90 transcription factors. Within these 90, they chose to analyze the most overrepresented transcription factors, the ones that were conserved across other *Saccharomyces* species. These included thirteen transcription factors involved in up-regulation and six involved in down-regulation.





**Figure 3:** Gene repression over time for 182 genes of Profile 9 (upper left, p = 1.7x10-166) and gene repression over time after exposure to arsenic for 80 genes of Profile 23 (upper right, p = 2.2x10-34). These were the most significant profiles. Lower left and right graphs show a gene expression increase over time for 18 genes of Profile 40 (p = 8.6x10-12) and for 19 genes of Profile 48 (p = 2.2x10-9) respectively.

After STEM analyses, Profiles 9 and 23 were combined and Profiles 40 and 48 were combined and each gene list was put into Gene Ontology to observe the biological relevance of the clusters. Profiles 9 and 23, the clusters involved in down-regulation in gene expression, played a major part in translational termination or modification. There were many GO terms under these two clusters, and their p-values were very significant. On the other hand, Profiles 40 and 48, involved in up-regulation, contributed to oxidation-reduction metabolism and abiotic response. These p-values were still significant, but not as significant as the terms found in Profiles 9 and 23 (Figure 3). The p-values from GO and from STEM differ in the sense that the STEM cluster values correlated to how the clusters themselves statistically differed from each other, whereas the GO p-values show which term is most significant in that cluster. The paper written by Thorsen et al. did not correlate any transcription factors with GO terms; however, they compared the function of their factors with previous literature and showed that these factors were involved with sulfur metabolism, GSH assimilation, transcription/translation, and cellular protection. These are processes are very similar to the GO terms we produced.

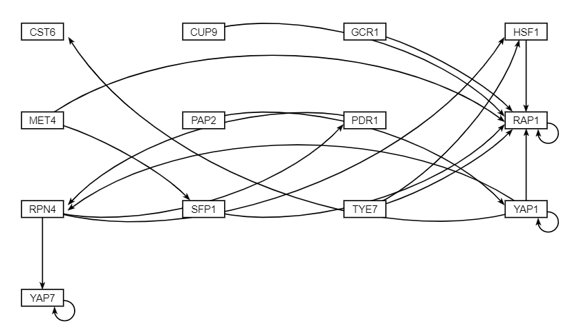
**Table 2:** Biological processes for each profile according to the GeneOntology database. Three of the most significant terms were selected except for Profile 48 because it only had one term associated with it.

|  |  |  |  |
| --- | --- | --- | --- |
| Profile | GO Biological Process | GO ID | P-value |
| 9 | Translational termination | GO:0006415 | 1.66x10-106 |
| Cellular protein complex disassembly | GO:0043624 | 1.04x10-103 |
| Cytoplasmic translation | GO:0002181 | 8.51x10-105 |
| 23 | Organonitrogen compound biosynthetic process | GO:1901566 | 2.12x10-19 |
| Cytoplasmic translation | GO:0002181 | 2.54x10-17 |
| Biosynthetic process | GO:0009058 | 3.43x10-16 |
| 40 | Oxidation-reduction process | GO:0055114 | 2.41x10-7 |
| Cellular aldehyde metabolic process | GO:0006081 | 6.18x10-3 |
| Cysteine biosynthetic process | GO:0019344 | 6.73x10-3 |
| 48 | Response to abiotic stimulus | GO:0009628 | 5.90x10-6 |

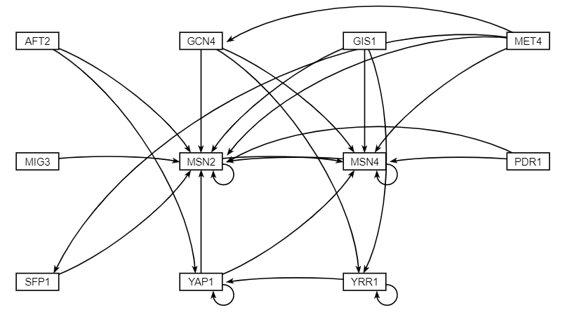
Transcription factors were chosen based on their p-value after being run through YEASTRACT to ensure only the most significant transcription factors were used for the GRNsight network visualization. The article focused on transcription factors MET4 and YAP1, but also mentioned that MSN2, MSN4, RAP1, and RPN4. The transcription factors MET4 and YAP1 have been found to control GSH biosynthesis pathways for sulfate assimilation (Thorsen et al., 2007). Also, MSN2, MSN4, and RAP1 have been shown to regulate environmental stress response (Gasch, 2003), while RPN4 has been found to control expression of proteasome genes in response to arsenite exposure (Haugen, 2004).

**Table 3**: Results of ANOVA and YEASTRACT analyses of genes in combined Profiles 9 & 23 and combined Profiles 40 & 48, highlighting the transcription factors present during the introduction of As(III) to Yeast strains.

|  |  |  |  |
| --- | --- | --- | --- |
| Combined Profiles 9 and 23 | | Combined Profiles 40 and 48 | |
| Transcription Factor | YEASTRACT p-value | Transcription Factor | YEASTRACT p-value |
| CST6 | <0.0000000001 | AFT2 | 0.000003352678561 |
| CUP9 | <0.0000000001 | GCN4 | <0.0000000001 |
| GCR1 | <0.0000000001 | GIS1 | 0.00000058671095 |
| HSF1 | <0.0000000001 | MET4 | 0.000000000043537 |
| MET4 | <0.0000000001 | MIG3 | 0.000000505193164 |
| PAP2 | <0.0000000001 | MSN2 | 0.000010362345153 |
| PDR1 | <0.0000000001 | MSN4 | 0.000010853606216 |
| RAP1 | <0.0000000001 | PDR1 | 0.000008995591101 |
| RPN4 | <0.0000000001 | SFP1 | 0.000012505499121 |
| SFP1 | <0.0000000001 | YAP1 | <0.0000000001 |
| TYE7 | <0.0000000001 | YRR1 | 0.000006593243698 |
| YAP1 | <0.0000000001 |  |  |
| YAP7 | <0.0000000001 |  |  |

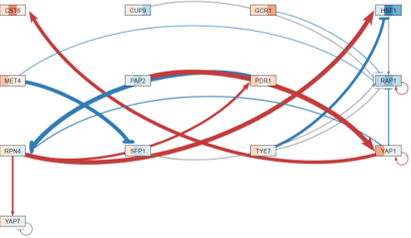


**Figure 4:** Resulting unweighted network from inputting combined Profiles 9 and 23 to GRNsight. Suggests highest gene regulation for RAP1.

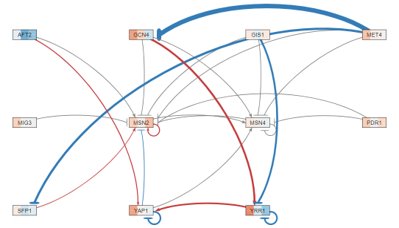


**Figure 5:** Resulting network from inputting combined Profile 40 and 48 to GRNsight. Suggests highest gene regulation for MSN2 and MSN4.

The GRNsight network visualizations suggest that RAP1 was the most important transcription factor for combined Profiles 9 and 23, which were both groupings of transcription factors that promote gene down regulation. Likewise, MSN2 and MSN4 were suggested as the most important transcription factors for combined Profiles 40 and 48, which were both groupings of transcription factors that promoted gene up regulation. The LSE:miniLSE ratio for combined Profiles 9 and 23 was found to be 1.080112. The LSE:miniLSE ratio for combined Profiles 40 and 48 was found to be 1.085448367. These LSE:miniLSE ratios suggest both models were very good estimations. The model for Profiles 9 and 23 were slightly better than that of Profiles 40 and 48 according to the LSE/miniLSE ratio.



**Figure 6:** Weighted network results from inputting combined Profiles 9 and 23 to GRNsight, including gene repression (blue) and increased expression (red).



**Figure 7:** Weighted network results from inputting combined Profiles 40 and 48 to GRNsight, including gene repression (blue) and increased expression (red).

|  |  |  |
| --- | --- | --- |
| Profiles 9 and 23 | | |
| id | production\_rate | threshold\_b |
| CST6 | 0.205301 | 3.692632445 |
| CUP9 | 0.074452 | 0 |
| GCR1 | 0.401079 | 0 |
| HSF1 | 0.092695 | 2.751155798 |
| MET4 | 0.3055 | 0 |
| PAP2 | 0.131969 | 0 |
| PDR1 | 0.184824 | 0.717116899 |
| RAP1 | 0.499368 | 0.149793822 |
| RPN4 | 3.18864 | -4.681528114 |
| SFP1 | 2.915812 | -3.243882769 |
| TYE7 | 0.58294 | 0 |
| YAP1 | 1.185729 | 6.847922918 |
| YAP7 | 0.3219 | 1.997074754 |

**Table 5:** Output production rate and threshold\_b from the combined Profiles 40 and 48 GRNmap workbook.

**Table 4:** Output production rate and threshold\_b from the combined Profiles 9 and 23 GRNmap workbook.

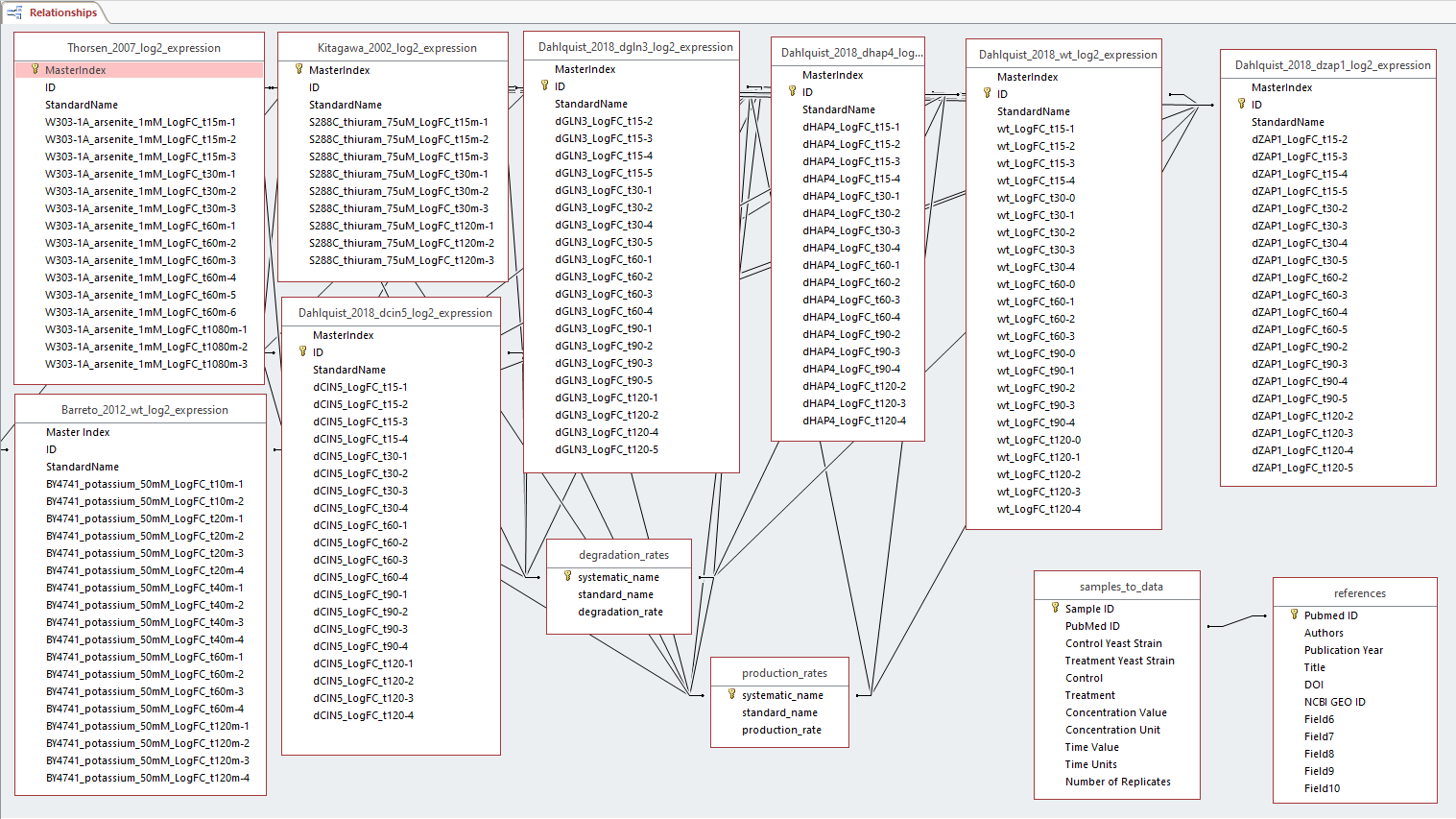
|  |  |  |
| --- | --- | --- |
| Profiles 40 and 48 | | |
| id | production\_rate | threshold\_b |
| AFT2 | 0.147589 | 0 |
| GCN4 | 4.114486 | -7.227724657 |
| GIS1 | 0.419276 | 0 |
| MET4 | 0.33723 | 0 |
| MIG3 | 0.316831 | 0 |
| MSN2 | 0.871062 | 0.10765237 |
| MSN4 | 0.509943 | 0.200965003 |
| PDR1 | 0.290892 | 0 |
| SFP1 | 2.862198 | -3.094744394 |
| YAP1 | 0.897875 | 1.999265272 |
| YRR1 | 1.416894 | 2.33362044 |

*Designer: Schema and Metadata Design*

The Designer worked with the QA to create a database gathering all the wildtype data and collaborated with other Coders to create the metadata tables including the sample-to-data sheet as well as the citations from each of the groups papers. There were multiple databases created to better understand and compile all the group’s datasets together to form a larger database. Before designing the database, metadata tables were created to compile the information given in each of the groups’ papers. The samples\_to\_data sheet incorporated the experimental design, including the PubMed ID, the yeast strain used, the treatment yeast strain, the control, the treatment used, the concentration of the treatment and concentration units, the time values and time units, and the number of replicates for that time value among all of the groups. The references sheet included the citations of the paper, containing the PubMed ID, authors, title of the paper, DOI, and NCBI GEO ID. With the gene expression tables provided from the other groups as well as from Dahlquist’s research lab, each expression sheet was labeled with a specific header including the author, publication year, log2 fold change, and expression. For the data in each of the groups’ expression sheets, the titles of the timepoint datasets were standardized to the yeast strain, treatment used, concentration of the treatment, log fold change, the time value and the replicate number to match the formatting in the samples\_to\_data sheet. The degradation rates and production rates of the genes were also provided and unchanged. These newly formatted sheets were the same across each group to remain consistent with the data provided. With these sheets, the database was then created through linking all the standard IDs together (ID, systematic name, PubMed ID) to see how the data can be related to each other.

In the schema provided in Figure 8, the metadata tables sample\_to\_data and references are linked only to each other, but in the final database, these metadata tables will be linked to the timepoints of each of the datasets to show which dataset they came from originally.

*QA: Project Overview*

The QA worked alongside the designer and data analysts by reviewing the microarray dataset. We first began by downloading the microarray data associated with Thorsen et al. (2007) as the data analysts did. We referred to the sample-data relationship table to go over the formatting of all the samples, including the concentration of the arsenite, timepoints, and the replicate number. The QA also went over the details of the microarray dataset that was being analyzed.

**Figure 8**: The final database schema shows the relationships between different datasets among each group, connected through their standard IDs across the degradation rates, production rates, and fold change expressions data sheets.

The QA process for the database included the assimilation of the data from the Dahlquist, (2018) work as well as collected degradation and creation rates. Data from the Thorsen, et al., (2007) was then added to this created database, including the data of gene expression as a result of the introduction of the toxic arsenite compound. In order to format and regularize the data tables and the headers of each data column, the QA members were tasked with communicating with the QAs and Coders/Designers of other groups and finding similar formatting patterns of each column. From there, the column headers as well as the titles of the database tables themselves were renamed in order to provide the most informative and concise representation of the data that was attempting to be presented. Using the sample-data table, the column headers were formatted to include the stimulus being tested, the concentration of the stimulus, the yeast strain, and the amount of time the stimulus was affecting the yeast. Once the database was complete in terms of formatting and data inclusion, the QAs worked with the Coder/Designer in order to create a query that connected each relevant data table with each other using the Gene IDs as the similar data column persistent throughout all of the tables. There were no encountered issues in creating the database and all the data that was collected and assimilated was complete. When formatting the genes to their ORF names, all genes IDs were successfully imported into the database and no formatting issues arose at the time.

**Conclusions**

In the scientific article we learned that the yeast cells were exposed to two concentrations of Sodium arsenite (0.2 mM and 1 mM). The researchers found that some genes were related to arsenic detoxification and structural components of the sulfur assimilation and GSH biosynthesis pathway. This supports what they introduced in the beginning with the yeast cells channeling a large part of assimilated sulfur to when being exposed to arsenite. Overall, we found that some transcription factors were present in all four profiles and that the same biological processes were found in all three of the groups. This indicated a similarity in how the yeast strain responds to stress with different stressors. The paper focused on two transcription factors, MET4 and YAP1, but after re-analyzing the information we found many more transcription factors. During the making of our presentations we had a difficult time navigating the strain name of the yeast we used. It was not clearly stated in the paper. Based on the Bonferroni criteria only 0.522% of regulatory genes at the 15, 30, 60, and 1080 minutes were significant. Our STEM results showed that the most significant transcription factors were involved in the down-regulation in gene expression, and these factors were mainly involved with translational termination and cellular component disassembly. This was seen in profiles 9 and 23. Other factors controlled the cell’s response to oxidative-reduction metabolism and to abiotic stressors. This was seen in profiles 40 and 48. For future directions we would want to analyze the datasheets that were not included in this assignment. We would want to have the database accessible on GRNsight for future research. Lastly, we would want to find IDs that can be linked within the Access schema.

Throughout the course of the data analysis and database creation, we learned that cells, namely yeast cells have properties and mechanisms in place that allow them to activate or deactivate certain genes in an attempt to combat toxic compounds that present themselves in the environment. The activation of these certain genes causes for the creation and release of compounds from the cell in order to keep the toxic compound from completely killing the cell itself. Besides having all the relevant data in one place, the new database allows for the comparison of gene expression with degradation and creation rates, allowing for the creation of a GRNMap which allows for the prediction of gene expression over a period of time. Given more time with this experiment, we would like to Analyze the datasheets that were not included in this assignment (the multiple database tables from the Thorsen et al. (2007) report would be analyzed under mutated yeast strain conditions over time), have this database accessible on GRNsight for other future research to be done by scientists and our peers, and find IDs that can be linked within the Microsoft Access schema that we created (If we wanted to compare gene function in one yeast strain as opposed to another, we could use this database as a tool to find any differences, whether it be differences with the other research groups or scientists currently in the field). The results that we obtained from this project didn’t greatly differ from the paper that we presented during the journal club. Most, if not, all of the results in terms of genes controlling the toxicity response lined up with the original research paper’s results and conclusion. We learned a lot about how extensively gene expression affects the cellular processes carried out as a result of arsenite exposure, however, in terms of information that was left out of the research paper itself, we could not think of any brand new information created through our independent research besides the fact that we found out the connections between genes and how each gene effects other genes in order to effect the cell as a whole. Instead of it simply being a gene or a group of genes working to solve the toxicity problem, there is a complex mesh framework of genes that regulate each other whether it be through repression or induction.

**Acknowledgments**

We would like to acknowledge Dr. Dahlquist for providing the instructions needed to re-analyze the data from this scientific article as well as Loyola Marymount University for giving us the resources needed to conduct this project. We would also like to acknowledge our classmates, specifically the coders from both groups for working with us in creating the sample to data sheet, data analysists, quality assurance, and project managers.

**References**

Barreto, L., Canadell, D., Valverde‐Saubí, D., Casamayor, A., & Ariño, J. (2012). The short‐term response of yeast to potassium starvation. *Environmental microbiology*, 14(11), 3026-3042. DOI: https://doi.org/10.1111/j.1462-2920.2012.02887.x

Gasch, Audrey P. (2003). The environmental stress response: a common yeast response to diverse environmental stresses. *Yeast stress responses*. 11-70. DOI: https://doi.org/10.1007/3-540-45611-2\_2

Haugen, A. C., Kelley, R., Collins, J. B., Tucker, C. J., Deng, C., Afshari, C. A., ... & Van Houten, B. (2004). Integrating phenotypic and expression profiles to map arsenic-response networks. *Genome biology*, 5(12), R95. DOI: https://doi.org/10.1186/gb-2004-5-12-r95

Kitagawa, E., Takahashi, J., Momose, Y., & Iwahashi, H. (2002). Effects of the pesticide thiuram: genome-wide screening of indicator genes by yeast DNA microarray. *Environmental science & technology*, 36(18), 3908-3915. DOI: https://doi.org/10.1021/es015705v

Thorsen, M., Lagniel, G., Kristiansson, E., Junot, C., Nerman, O., Labarre, J., & Tamás, M. J. (2007). Quantitative transcriptome, proteome, and sulfur metabolite profiling of the Saccharomyces cerevisiae response to arsenite. *Physiological genomics*, 30(1), 35-43. DOI: https://doi.org/10.1152/physiolgenomics.00236.2006