Genome-wide DNA Microarray analysis of Yeast Treated with Thiuram

A Reassessment of Data Published by Kitagawa et al.

By Michael B. Armas, Iliana E. Crespin, Kaitlyn N. Nguyen, Emma R. Young

Biological Databases (BIOL-367-01)

Submitted on December 13th, 2019

**Introduction**

Pesticide pollution is a worldwide problem that is harmful to human health due when consumed through drinking water and residues found of foodstuff1,2. Thiuram is a fungicide that is widely used to protect vegetable crops from diseases and is thus found in drinking water due to the runoff from crop fields3. While other studies1,2 have researched the effects of heavy metal pollutants on yeast (*Sacharomyces cerevisae)* gene expression, the use of yeast DNA microarray analysis to identify a biomarker for thiuram pollution was first introduced by Kitagawa et al. (2002)3.

The applications of DNA Microarrays have been vastly growing over the last two decades4. Microarray analysis allows for the observation of gene expression and regression in a treated sample. The microarray chip is placed into an instrument that collects data of the change of gene expression. In using bioinformatics and computer software, the mass datasets can be analyzed to observe significant relationships in gene expression. While the processing of this data is expected to be reliable, analysis of large datasets can come with intricate mistakes that skew the results of an entire project. The credibility of bioinformatics research comes with open ended discussion of published data and results. Especially in studies published in the early days of microarray analysis, it is important to reanalyze results to assess the current credibility of the paper.

Herein, a paper published by Kitagawa et al. (2002)2 was reassessed to question the credibility of the results which were presented. This publication claims to have identified biomarkers of Thiuram pollution in natural and processed water sources by observing the change in gene expression in yeast cells treated with the pesticide thiuram. Since many methods of identifying pollutants can be poor5, identifying biomarkers of pollutions can be plausible indicators of pollution. In reading this publication, our group believed that some results were skewed such that the paper loses credibility. Thus, the reanalysis of the data is performed to confirm results from this paper.

A variety of bioinformatics software is used to process the data in a similar fashion. The original dataset for this publication was retrieved and used to run our own ANOVA and STEM analyses. Genes with significant change in gene expression were selected and identified with their gene ontology definition in a grouping of profiles. From the Gene Ontology (GO) terms list and gene list obtained from the significant profiles, an adjacency matrix of transcription factor relations was formed. A program called GRNsight is used to visualize adjacency matrices as connected graphs6. The adjacency matrix was also used to create an input workbook for a MATLAB program called GRNmap. GRNmap uses differential equations to model the dynamics of gene regulatory networks7. The output data from this tool produces a weighted adjacency matrix, which was also visualized using GRNsight.

By using bioinformatics to find the relationship of the change in gene expression, the effects of cell stress can be analyzed and contrasted. In Kitagawa et al.2, the change in gene expression is used to identify biomarkers of Thiuram. Using bioinformatics to assess the data obtained from this group, we reassess the biomarker candidates and significantly altered cellular functions caused by thiuram toxicity.

**Methods, Results, and Discussion**

This section of the experiment includes the results of experiment and an analysis of these results comparable to the original microarray data from the journal article.

**Figure 1- Flowchart of Methods.** This flow chart shows the method steps that were taken when working with the data from the Kitagawa et al. 2002. The first step was retrieving the data, then that data was analyzed using ANOVA testing. Using the B-H p-values the data was narrowed down to only significant numbers and those were run through STEM, the profiles from STEM were combined giving us gene ontology IDs and the gene lists for YEASTRACT. YEASTRACT was run on the gene list and the results were viewed in GRNsight. The MS Access Database was created and then used with the YEASTRACT results to create the input workbook for GRNmap. The workbook was run through GRNmap, using MATLAB and the results were visualized in GRNsight.



**Retrieving Original Data: Strains.** Data for the reassessment of Kitagawa et. al's experiment on yeast treated with thiuram was available for download through Loyola Marymount University’s Biological Databases Wikipedia page8. *Saccharomyces cerevisiae* S288C was used as the indicator strain to detect viable biomarker candidates. In the download of the journal article, as well as the dataset, the microarray dataset was manipulated through statistical analysis to reexamine and compare the new results with the original experiment’s results. The original file contained 4466 genes, its gene ID, and its corresponding data at 15 minutes, 30 minutes, and 120 minutes - each timepoint replicated 3 times.

**Table 1- Table of ANOVA results for** ***Saccharomyces cerevisiae* strain S288C of 44467 total gene set examined at variable p-values in a test for significance.**



**Analysis of Variance (ANOVA) Test.** To determine if any genes within the 4466 gene dataset had a gene expression change that was significantly different than zero at any given timepoint, the data was manipulated via Microsoft Excel by filtering the data according to the criterion we set (Table 1). In a new worksheet labeled “S288C\_ANOVA,” the first column contained the “ID,” with the following dataset columns to the right in increasing chronological order by time and replicate number. In order to make all research assignments cohesive, the Coder/Project Manager and Quality Assurance confirmed the benchmark in which to label the dataset columns with their respective guilds. The format of the column names was as follows, “(STRAIN)\_(NAME)\_(CONCENTRATION)\_LogFC\_(TIME)-(REPLICATES). For example, in the wildtype *Saccharomyces cerevisiae strain* S288C in pesticide thiuram concentration 75uM, the logarithm at timepoint 15min. of replicate 3 is labeled as follows: “S288C\_thiuram\_75uM\_LogFC\_t15m-3.” The data from these columns was directly transferred and copied over from the original dataset. Following these dataset columns, the average of each timepoint was calculated, labeled as “S288C\_thiuram\_75uM\_AvgLogFC\_(TIME),” for three of the different timepoints: at 15 minutes, 30 minutes, and 120 minutes. To the right of the last column, the following columns included additional sets of statistical analysis, more specifically, the Benjamini & Hochberg and Bonferroni tests.

With an ANOVA p-value of 0.05, 1661 genes (37.2%) were statistically significant at a 95% confidence level. Less than 1% of the time, 810 genes (18.26%) had a gene expression change. ANOVA tests were run at a p-value less than 0.001 and 0.0001 as well. To apply a more stringent application to these unadjusted p-values, a Benjamini & Hochberg and Bonferroni tests were run. This allowed us to see a gene expression change for at least one of the timepoints by chance in about 5% of our 4466 hypothesis tests. With both criterions set at p < 0.05, the Benjamini & Hochberg corrected p-value had a significance of 16.36%, while the Bonferroni p-value had a significance of 0.11%. The Benjamini & Hochberg and Bonferroni tests specifically allowed us to select genes that had a significant change in expression and narrow down the genes that would later be put through the Java program, STEM (Short Time-series Expression Miner). The Quality Assurance double checked these numbers and made sure it was ready for STEM. In comparison to the original methods used to measure significance, the original microarray experiment did not use statistical analysis, rather it set its standard by conceptually observing fold change between greater than 2.0 or less than 0.05.

**Figure 2-This chart shows all the profiles resulting from the STEM results. The different colors show profiles with common trends of gene expressions. They are shown in order of significance.**



**STEM Analysis.** A Thuriam\_stem worksheet was added to the ANOVA workbook and the data was taken from the Thuriam\_ANOVA sheet. The columns were renamed so that the first was renamed “SPOT” which was originally named Master Index. Then, the column “ID” was renamed “standard name”. The data was filtered by B-H corrected p values the results greater than 0.05 were removed from the data. The rest of the columns, besides “Spot” and “ID”, were the average log for each time point. The average of each of the time points was renamed 15m, 30m, and 120m. The results were then saved as a text (Tab-delimited)(\*.txt) file. The STEM web site was used to download the software. The stem.zip file was then unzipped, which created a new folder called stem. The downloaded files, gene\_ontology.obo and gene\_association.sgd.gz, were placed into this file. Inside the unzipped folder, downloaded from the stem site, the Stem.jar file was clicked and opened. The Stem interface was opened. The file was opened in the interface. The options “*No Normalization/add0”* and *“Spot ID’s included in the data file”* were selected and the Gene info of the main STEM interface window was left as the default “User provided”. For the Gene Annotation file, the gene\_association.sgd.gz was selected. The Clustering Method said, “STEM Clustering Method”, and all other defaults were kept.

The STEM results were saved. This included the figure (Figure 2) that showed all the STEM Profiles, graphs of each of the significant profiles. The Gene Ontology lists were saved for each of the profiles, as well as the gene lists for each of the profiles. The Quality Assurance made sure STEM profiles were accurate.

**Graph 1:**This graph shows the results of the gene clustered in Profile 9 RED. These results were achieved through running the Gene data from Kitawaga et al. 2002 through the STEM program.



**Graph 2:** This graph shows the results of the gene clustered in Profile 34 Red. These results were achieved through running the Gene data from Kitawaga et al. 2002 through the STEM program.

 

**Graph 3:** This graph shows the results of the gene clustered in Profile 26 Red. These results were achieved through running the Gene data from Kitawaga et al. 2002 through the STEM program.



**Graph 4:** This graph shows the results of the gene clustered in Profile 11 Red. These results were achieved through running the Gene data from Kitawaga et al. 2002 through the STEM program.



**Graph 5:** This graph shows the results of the gene clustered in Profile 40 Green. These results were achieved through running the Gene data from Kitawaga et al. 2002 through the STEM program.



**Graph 6:** This graph shows the results of the gene clustered in Profile 18 Green. These results were achieved through running the Gene data from Kitawaga et al. 2002 through the STEM program.



**Graph 7:** This graph shows the results of the gene clustered in Profile 42 Green. These results were achieved through running the Gene data from Kitawaga et al. 2002 through the STEM program.



**Graph 8:** This graph shows the results of the gene clustered in Profile 29 Blue. These results were achieved through running the Gene data from Kitawaga et al. 2002 through the STEM program.



The profiles were put into two different large profiles, based on their overall trend. The red profiles, #9,26, 34, and 11, were all combined into a group. This was called the Red profile. All the different profiles showed a downward trend, meaning a repression of gene expression. The green profiles, #40, 42, and 18, were combined with blue profile, #29, to make one big one called, Green Profile. The profiles that were combined into the Green profile all showed an upward trend, representing an increase in expression of the genes in this profile. The Red profile had 289 genes, only 51.5 genes were expected to be present in this profile. The Green profile had 214 genes, only 48.5 were expected.

**Figure 3A - GO results from Red profile (repressed)** from Geneontology.org. Small portion of full table of results.

**Figure 3B - GO results from Green profile (induced)** from Geneontology.org. Full table of results.

**Gene list and GOlist for each significant profile.** Using the same STEM.jar application from the previous step, each profile that was clicked on contained a “Profile Gene Table” and “Profile GO Table,” both of which were downloaded for each profile and zipped into one file. In a new Excel workbook, each gene list was uploaded onto either the Green or Red profile worksheet. It is important to note that STEM did not include the Gene term names of the following dataset; therefore, the following clustered gene lists were pasted onto <http://geneontology.org/> in the Go Enrichment Analysis box on the right-hand side of the homepage. From the species drop-down menu, Saccharomyces Cerevisiae was selected, and the application was launched. This prompted to a results page, where the button “Table” was saved and imported into Excel. This would later be used for YEASTRACT.

**YEASTRACT.** The gene lists were generated from the profiles from running STEM. The gene lists were combined and prepared to be examined. This was done for both the Red and Green profiles. The list of the Red profile was copied and the YEASTTRACT Database was opened. The option of “Rank by TF” in the left panel of the site was selected. The copied gene list was pasted in the box Labeled “ORFs/Genes”. The box for “Check for all TFs” was selected, and the defaults for the regulations filter were kept. The filter for “Filter Documented Regulations by environmental condition” was not applied, but the Rank genes by TF using “the % of genes in the list and in YEASTRACT regulated by each TF” was applied. The information was then run through the database. The results were then saved. The same steps were repeated with the Green profile. In the Red profile, the results showed 30 genes that were considered significant; whereas, the results of the Green profile showed 31 genes that were considered significant. The 17 genes with the most significance was chosen from each profile. They were put back through YEASTRACT. The “Generate Regulation Matrix” was selected from the left panel. The list of 17 genes was reselected and pasted into both the "Transcription factors" field and the "Target ORF/Genes" windows. The "Regulations Filter" options of "Documented" and "**Only** DNA binding evidence" were selected. The program was then run. Of the results that popped up, the “Regulation matrix (semicolon separated Values (CSV) file)” was saved to the desktop.

**Table 2: Red Profile Regulatory Transcription Factors.** These are the initial 17 genes ranked most significant from YEATRACT that were chosen for further examination (some were removed later on for being floaters or having incomplete data ). The first column is the gene name and the second column in the corresponding p-value calculated in YEASTRACT.



**Table 3: Green profile regulatory transcription factors** These are the initial 17 genes ranked most significant from YEATRACT that were chosen for further examination (some were removed later on for being floaters or having incomplete data ). The first column is the gene name and the second column in the corresponding p-value calculated in YEASTRACT.



**GRNsight.** The Red matrix and Green matrix, saved from YEASTRACT, were then put into Excel worksheets. The networks were then formatted into a format compatible for the GRNsight modeling software and the GRNsight visualization software. This was done by copying the matrix into a new sheet. The copied matrix was then transposed. In order to have the labels for the genes match, the letter p at the end and capitalizing all the letters were completed. In the cell A1, it was titled “rows genes affected/cols genes controlling”. Finally, the rows and columns were alphabetized. The new sheet was then labeled “network” and was saved. This was done for both the Green and Red profiles.

The GRNsight home page was opened, the file option, then the created regulation matrix was selected. The matrix was then run through the GRNsight software and the graph was created from the network. The network was then realized for the Red profile for that network, GCR1and SFP1 were found to be floaters and removed. For the Green profile BAS1 and MSN were found to be a floater and the genes were removed.

**Figure 4A – Unweighted Red profile on GRNsight.**



**Figure 4B – Unweighted Green profile on GRNsight.**



**Create MS Access Database.** That database is created using Microsoft Access. A metadata workbook is made to describe the samples from the microarray dataset. The metadata sheets were titled “references” and “samples\_to\_data.” The “references” table contains information on the three papers presented by the three groups. The PubMed ID of each article is used to link the sample information found in the “samples\_to\_data” sheet to each of the papers found in the references sheet. The “samples\_to\_data” sheet presents information on the data of each sample. These data were used to make universal column headers between all groups. The headers are in the format:

yeastStrain\_treatmentOrMutation\_concentrationWithUnits\_LogFC\_t+timeWithUnits-replicate

Expression tables are included from the three in-class groups as well as the Dahlquist Lab data. These tables, as well as the tables for production and degradation rates, are linked by the systematic gene name. The column headers were made using the parameters found in the metadata. This was the only relationship between the metadata tables and the microarray sample data. There is no relationship shown in the relationship report found in MS Access. This is due to the inability to link the data and metadata sheets. However, a ReadMe document is included with the database that explains the metadata and data relationships.

**Creating the input workbook for GRNmap.** A new workbook was created for all the genes in the network, each worksheet serving different purposes for interpretation. For example, “production\_rates” worksheet contained the initial guess for the production rate parameters, while “degradation\_rates” included an initial guess on the degradation rate. The data for these workbooks was found using queries in the MS Access Database. Due to a lack of data some genes were removed, PDR3 and MGA2 from the Red Profile and PDR3 from the green profile. This workbook was later called the “Input Workbook” as it was to be plugged into GRNmap for further analysis. While YEASTRACT allowed us to determine significance and draw a connection between transcription factors, GRNmap’s software uses the network and input workbook created to determine the rate of the gene expression, and therefore the weighted connection. This program and algorithm allows us to determine which genes are repressed and induced, how much, and at what rate. Following running GRNmap, the outbook workbook was available for download, and put into GRNsight again for a final visualization of the results.

**Final visualization in GRNsight.** The GRNmap .xlsx file was then uploaded into GRNmap to view the results. The GRNsight home page was opened, the file option, then the created regulation matrix was selected. The matrix was then run through the GRNsight software and the graph was created from the network. This created a weighted network for both of the Red and Green profiles.

**Figure 5A - Red profile on GRNsight.** Final visualization in GRNsight showing weighted network.



**Figure 5B - Green profile on GRNsight.** Final visualization in GRNsight showing weighted network. This profile has one floater MIG3 which was not removed from the data, due to the fact it was not known to be a floater until after this weighted network was visualized.



**Conclusion**

Overall, the comparison of the results differed between the original article and the reassessment that was conducted. Using biological databases, data analysis was facilitated by the linkage of multiple expression data sheets. The data used across many papers were connected and used to assess the expression change data for the profiles found in this study. As mentioned before, the categories of the biological processes made it very difficult to categorize the gene functions. Since, MIPS was used in the original article, it seemed very generic and outdated for today’s studies. Therefore, it was difficult to correlate the GO terms with it. In addition, the transcription factor YAP1 was observed with the secondary data analysis and the original article. This was confirmed in the gene regulatory network and adjacency matrix. However, in the original article, it was stated that they found genes with the response element consensus sequence. Besides that, 2 of the 4 biomarkers (YKL071W and YCR102C) were not found the data analysis done in the ANOVA step. This lead us to realize that the data set we downloaded was missing roughly 2,000 genes. However YLR303W (Met17) and YLL057C (JLP1) were found in the ANOVA data and both had a calculated B-H p-value < 0.05, this meant that as the article had suspected they had a significance in the data. Although these genes did not show up in any of the other analysis steps they are worth further study. Although the article seemed to have some valid findings, the credibility of the article findings is in question due to insufficient evidence on the results they acquired. The methods of finding significance was not mentioned or used. Also, there was a lack of outlined methods, which led to a rushed analysis. For any future directions regarding yeast treated with thiuram, there would be reruns of this experiment using new technology that wasn’t available in 2002. In addition, reviewing the biomarker candidates of thiuram pollution using the new findings. Finally, analyzing the gene ontology using standardized category names instead of the ones that they acquired from MIPs.

**References**

**1** Braconi, D., Bernardini, G., & Santucci, A. (2016). Saccharomyces cerevisiae as a model in ecotoxicological studies: A post-genomics perspective. Journal of Proteomics, 137, 19-34. DOI: 10.1016/j.jprot.2015.09.001

2 Fujita, K., Iwahashi, H., Kodama, O., & Komatsu, Y. (1996). Induction of heat-shock proteins in the presence of thiuram in Saccharomyces cerevisiae—For evaluating the potential toxic risks in pollutants. *Water Research*, *30*(9), 2102–2106. doi: 10.1016/0043-1354(96)00019-x

3 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: 10.1021/es015705v

4 Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J., & Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. *Information Sciences*, *282*, 111–135. doi: 10.1016/j.ins.2014.05.042

5 Iwahashi, H., Ishidou, E., Kitagawa, E., & Momose, Y. (2007). Combined Cadmium and Thiuram Show Synergistic Toxicity and Induce Mitochondrial Petite Mutants. *Environmental Science & Technology*, *41*(22), 7941–7946. doi: 10.1021/es071313y

6 Kam D. Dahlquist, John David N. Dionisio, Ben G. Fitzpatrick, Nicole A. Anguiano, Anindita Varshneya, Britain J. Southwick, Mihir Samdarshi. (2016) GRNsight: a web application and service for visualizing models of small- to medium-scale gene regulatory networks. PeerJ Computer Science 2:e85 https://doi.org/10.7717/peerj-cs.85

7 Dahlquist, K.D., Fitzpatrick, B.G., Camacho, E.T., Entzminger, S.D., and Wanner, N.C. (2015) Parameter Estimation for Gene Regulatory Networks from Microarray Data: Cold Shock Response in Saccharomyces cerevisiae. *Bulletin of Mathematical Biology*, 77(8), 1457-1492, published online September 29, 2015. DOI: 10.1007/s11538-015-0092-6.

8 Dahlquist, K. (2019, October 2). Main Page [Wikipedia Page Post]. Retrieved from https://xmlpipedb.cs.lmu.edu/biodb/fall2019/index.php/Main\_Page.