* **Downloading the file**
* Create a folder on your desktop and name it exp\_data\_cleaning
* Download the excel file from the following link [expression\_data](https://github.com/dondi/GRNsight-archive/blob/main/documents/Source_data/expression_Data/GSE33097_s257_final-4.xlsx)
* Add this excel file to exp\_data\_cleaning folder
* **Editing the excel file**
* Change the headings (row 1) of columns D-U from wt-matA\_(time)min\_nbr029-a format to (time)m format (e.g., wt-matA\_0min\_nbr029-a to 0m)
* Change the heading of A1 cell to GeneSymbol
* Delete column B & C
* Delete row 2
* Place the timepoint in order of least to most with replications next to each other as such 0m, 0m, 3m, 3m, 7.5m, 7.5m …
* Insert a column after the replicates of the first timepoint
* Using the AVERAGE formula in excel compute the average of the first timepoint in the column created in the previous step
* Repeat the last two steps for each timepoint
* Rename this file to Average\_exp\_data.txt
* Go to file > Save as > Tab-delimited Text (.txt)

(This file should have 12,983 records)

* **Run ID-sorting-code.py code**
* make sure to give the correct file paths for code to work

(A file named sys\_SGD\_ID.txt should be created with 5,624 records with a mix of SGD and systematic names)

* **Download Yeast nomenclature from Yeast mine website (to convert sgd ID’s)**
* Visit [https://yeastmine.yeastgenome.org/yeastmine/bagDetails.do?scope=all&bagName=ALL\_Verified\_Uncharacterized\_Dubious\_ORFs#](https://yeastmine.yeastgenome.org/yeastmine/bagDetails.do?scope=all&bagName=ALL_Verified_Uncharacterized_Dubious_ORFs)
* Click on Export
* Select All columns from options on the left
* De-select Organism short name, Standard name, and name
* Click on Download file

(This gives a tsv file format)

* Save a copy in tab-delimited text format and name it YeastNomenclature.txt
* **Converting SGD ID to systematic ID’s**
* Open sys\_SGD\_ID.txt file in excel
* Select columns A-J
* Click on sort & filter
* Choose sort A to Z
* Select row #2 to row #722 including all columns A through J and cut it
* Open a new blank excel workbook and click on A1 cell and paste the SGD ID’s
* Name this file SGD\_incorrect\_format.txt
* Go to file > Save as > Tab-delimited Text (.txt)
* Run SGD-remover.py code
* A file named SGD\_CORRECT\_format.txt is created which has the SGD ID correct formats so they can be replaced with their respective systematic ID’s
* Use SGD\_CORRECT\_format.txt file in sgd-to-systematic-ID.py code as the template and then run the code
* A new file named replaced\_SYS\_IDs.txt should be created in exp\_data\_cleaning folder

(replaced\_SYS\_IDs.txt file contains the systematic ID instead of SGD ID’s) [there are fewer records than original file because the database isn’t complete]

* **Replacing the SGD ID’s not in the database**
* Open SGD\_CORRECT\_format.txt file in excel
* Open YeastNomenclature.txt in excel and copy column A
* Paste column A into SGD\_CORRECT\_format.txt column L
* Select column A and column L
* Select conditional formatting > Highlight Cells Rules > Duplicate Values > keep default setting and click ok (this will highlight the ID’s that have been replaced by the code)
* The ID’s that are not highlighted in column A are looked up in [YeastMine](https://www.yeastgenome.org/) using the search option on the top right of the website (there are 34)
* Open replaced\_SYS\_IDs.txt file in excel
* For each unhighlighted cell in column A copy the row from column A to J and add it to the next available row in replaced\_SYS\_IDs.txt file and replace the ID in column A with the systematic ID found of YeastMine website
* Do the previous step for the remaining 33 unhighlighted SGD ID’s
* **Creating a file with systematic names**
* Copy all the 721 rows from replaced\_SYS\_IDs.txt from column A to J and paste it into sys\_SGD\_ID.txt file
* Rename this file to systematic\_IDs.txt

(This gives 5590 records; however, some of them are aliases)

* **Remove the aliases**
* Run the removing\_duplicates.py code to remove aliases
* This code created a file named unique\_sys\_IDs.txt which contains all unique systematic IDs

(it doesn’t matter which ID is removed because YEASTRACT can identify the any ID’s including aliases)

* **The final cleaned file contains 5569 genes**
* This was confirmed by running a test to determine all unique expression data on the original data regardless of their ID’s (Meaning there are 5569 genes with unique expression data)
* **The ID’s were then used in YEASTRACT to obtain their standard names**
* **Running STEM**
* Open stem.jar
* In section 1, navigate the file containing the expression data
* Check Normalize data and Spot IDs included in the data file
* In section 2, leave the default for Gene Annotation source, Cross reference source, Gene location source as user provided
* Click the browse button to the right of “Gene Annotation file” and browse to STEM folder and select “gene\_association.sgd.gz” and click open
* In section 3, do not change the default and leave the clustering method as “STEM Clustering Method”
* In section 4, click on execute to run STEM

(You are expected to get 8 significant clusters for this dataset)

* **Viewing and saving STEM results**
* Click on Interface options on the button on the window
* At the bottom of the window popped up changed the “X-axis scale should be” from uniform to “Based on real time” and the close the window
* Take a screenshot of the window containing all clusters and save it
* Then take a screenshot of all the significant profiles (the colored ones) and save them
* For each significant profile, open the profile and at the bottom of the window popped up click on “Profile Gene Table” and after that click on “Save Table” and rename the file as “wt\_profile#\_genelist.txt” where # is the number of profiles e.g. wt\_profile2\_genelist.txt
* For each significant profile, open the profile and at the bottom of the window popped up click on “Profile GO Table” and after that click on “Save Table” and rename the file as “wt\_profile#\_GOlist.txt” where # is the number of profiles e.g. wt\_profile2\_GOlist.txt
* **Using YEASTRACT to infer which TF regulate a cluster of genes**
* Choose the most significant profile from stem (in this case I performed same steps for the first 4 profiles)
* Copy column C6 (included) to the end of all gene ID’s (depends on each profile)
* Open [YEASTRACT database](http://www.yeastract.com/)
* Click on “Rank by TF” on the left panel
* Paste the ID’s from your clipboard to the box labelled “Target ORF/genes”
* Check the box for “check for all TFs”
* Leave the default settings
* Click on search button
* A table appears that contains the Transcription factor and its corresponding p-value
* Copy all the Transcription factors that have a significant p-value (the green highlighted cells)
* Open a new sheet in the appropriate genelist profile excel file and paste the TF in A1 cell
* Copy column A
* Open another sheet
* Paste special with values in cell A1 in the new sheet
* Delete the trailing p at the end of each TF
* Open [Generate regulation matrix](http://www.yeastract.com/formregmatrix.php)
* Copy the first 15 TF
* Paste it into the box labelled “Transcription factors” and “Target ORF/genes”
* Choose “DNA binding AND expression evidence” from the regulation filter box
* Click “Generate” button
* A new window pops up and select “Regulation matrix (semicolon separated values (CSV) file)”
* Rename it into “Regulation\_matrix\_profile#” where # is the number of profile used
* **Visualizing your Gene Regulatory Networks with GRNsight**
* First the file needs to be correctly formatted so open the file in excel
* Select column A
* Go to “Data” tab and select “Text to columns”
* In the window appeared select “Delimited” then “Next”
* In the next window select “Semicolon” then “Next”
* In the next window leave data format to “General” and click “Finish”
* Save the file in .xlsx format
* Insert a new sheet in the excel workbook and name it “Network”
* Copy everything from the original matrix and paste special “Paste special” into A1 cell of the new sheet
* A new window appears and select “Transpose” at the bottom then click “OK”
* In row 1 delete the trailing “p” at the end of each gene name
* In cell A1 copy and paste “cols regulators/rows targets”
* Sort Column A from cell A2 from A-Z
* Sort row 1 from cell B1 from sort and “filter option” and select “options” and choose “Sort from left to right”
* Select B1 cell and using formula “=UPPER” in cell B20 make the gene name upper case
* Then drag the black cross marker to the right to apply the formula to all gene names in the first row
* Copy row 20 and paste special values into row 1 starting at column B
* Delete row 20
* Save the file
* Visualize the matrix in GRNsight, if there are some genes with no connection add or subtract until you get a gene regulatory network with all nodes connected
* Repeat these steps for all 4 profiles
* When all the adjacency matrices are obtained and correctly formatted open [GRNsight](https://dondi.github.io/GRNsight/) website
* Click on “File” from the top bar
* Then click on “Open file”
* Navigate through your adjacency matrix file that you saved in .xlsx
* Then click on “upload”
* You should now be able to visualize your GRN (you will get some warnings, but you can ignore them)