* **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**