Final Project DSCI 512

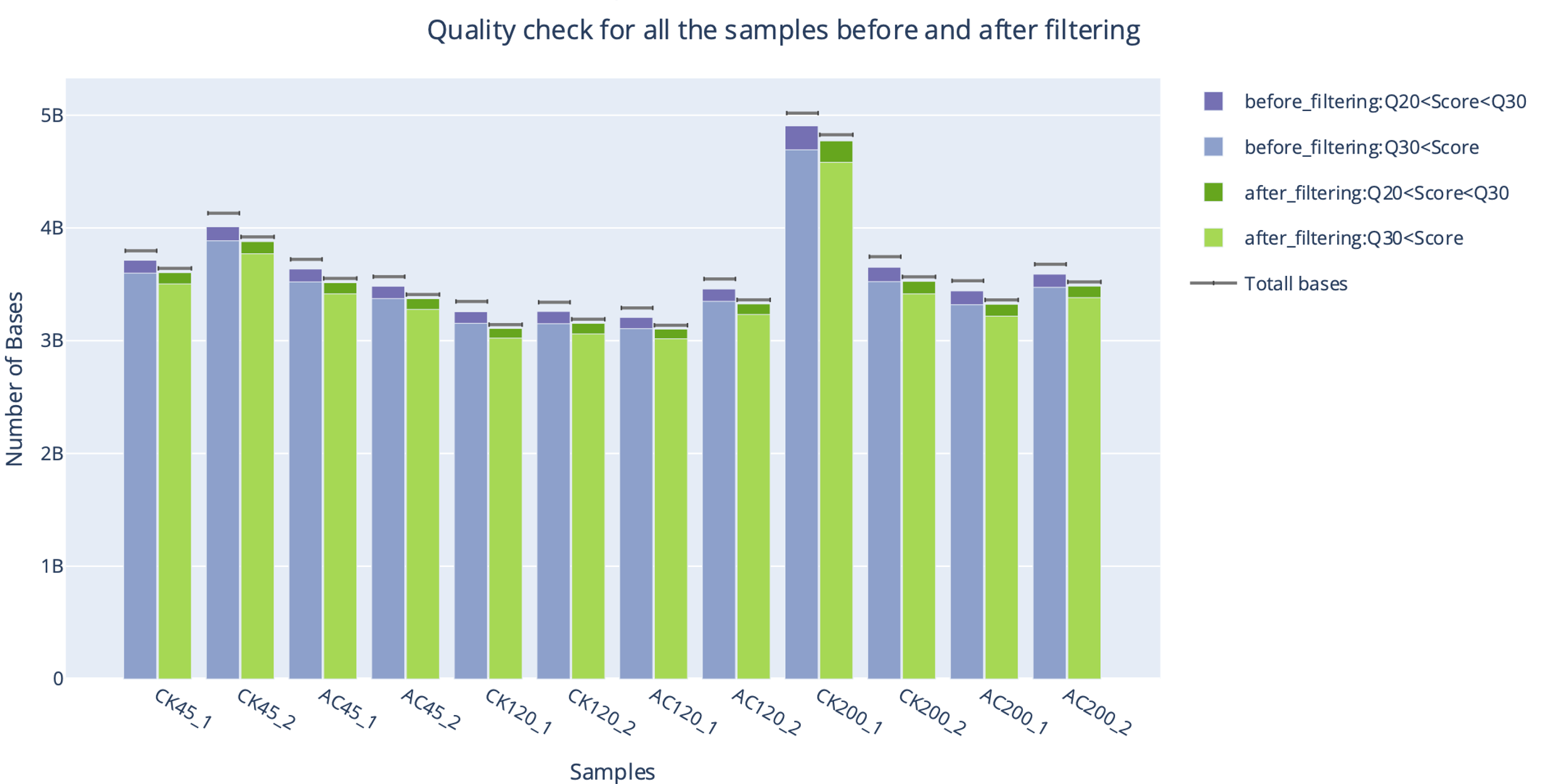
Parsa Ghadermazi

**Q1- Introduction**

I picked [1] as the dataset for my final project. This paper studied the effect of acetic acid exposure on transcriptome and other type of phenotypic data. The authors of this paper exposed S. cerevisiae to 150 mM concentration of acetic acid at pH 3. This means that acetic acid is in molecular form, protonated, in the environment and could enter the cell easily. They picked this level to make sure that the cells cannot restore pH to neutral which is observed for bellow 60 mM. They ran **three identical experiments for both controls and acetate treated reactors** and took samples at 45 min and 120 min and 200 min. Two samples from the replicates per time point. The main research question in this study was to explore the changes across the entire transcriptome after acetic acid exposure. The authors have made this study available on SRA by the accession id of: SRP075510.

Biofuel production is a promising way mitigate climate change. S. cerevisiae provides a robust platform for bioethanol production. Acetic acid as a microbial product can lower the efficiency of ethanol production and in turn the feasibility of the entire process. Acetate poisoning is a well-known problem in anaerobic digestion as well. The effect of acetate on metabolic pathways can extend beyond S. cerevisiae and be applied to other organisms. Some researchers attribute the lower efficiencies of fermentation products at high acetic acid level to thermodynamic limitations. However, this study shows that the effect is more related to changes in transcriptome and other phenotypes rather than thermodynamic limits. To me, the effect of acetic acid on metabolic pathways are the most central which is also addressed by the authors.

**Q2- QC plot**



**Figure 1** Comparison between number of bases before and after filtering for all samples. These information is extracted from individual outputs of Fastp for each sample. On the x axis individual samples are shown. Bar plots are grouped by before (Blue) and after filtering (Green). The y axis shows the number of bases in each group. First item in each stack shows the number of bases in that group that had a Phred score higher than 30, and the second item shows items with Phred score between 20 and 30. Also shown in the plot with solid lines are the total number of bases in each group. Overall, the number of bases kept after trimming is reasonable and ready for processing.

For making this plot, all the .json files from fastp outputs that were generated in the pipeline were used in python through a jupyter notebook. Important information was extracted from each .json file and were inserted into a large dictionary that held these information for all the samples. Finally, Plotly library was used to generate Figure 1 from the large dictionary mentioned. Also, a comma separated file for the important information was created by using pandas library. The jupyter notebook walking step-by-step through all the process is provided in the link bellow:

[Jupyter Notebook](https://github.com/ParsaGhadermazi/RNA_Seq/blob/e3ddf2c010b84249516409b9a174763258dd0468/Final%20Project/QC.ipynb)

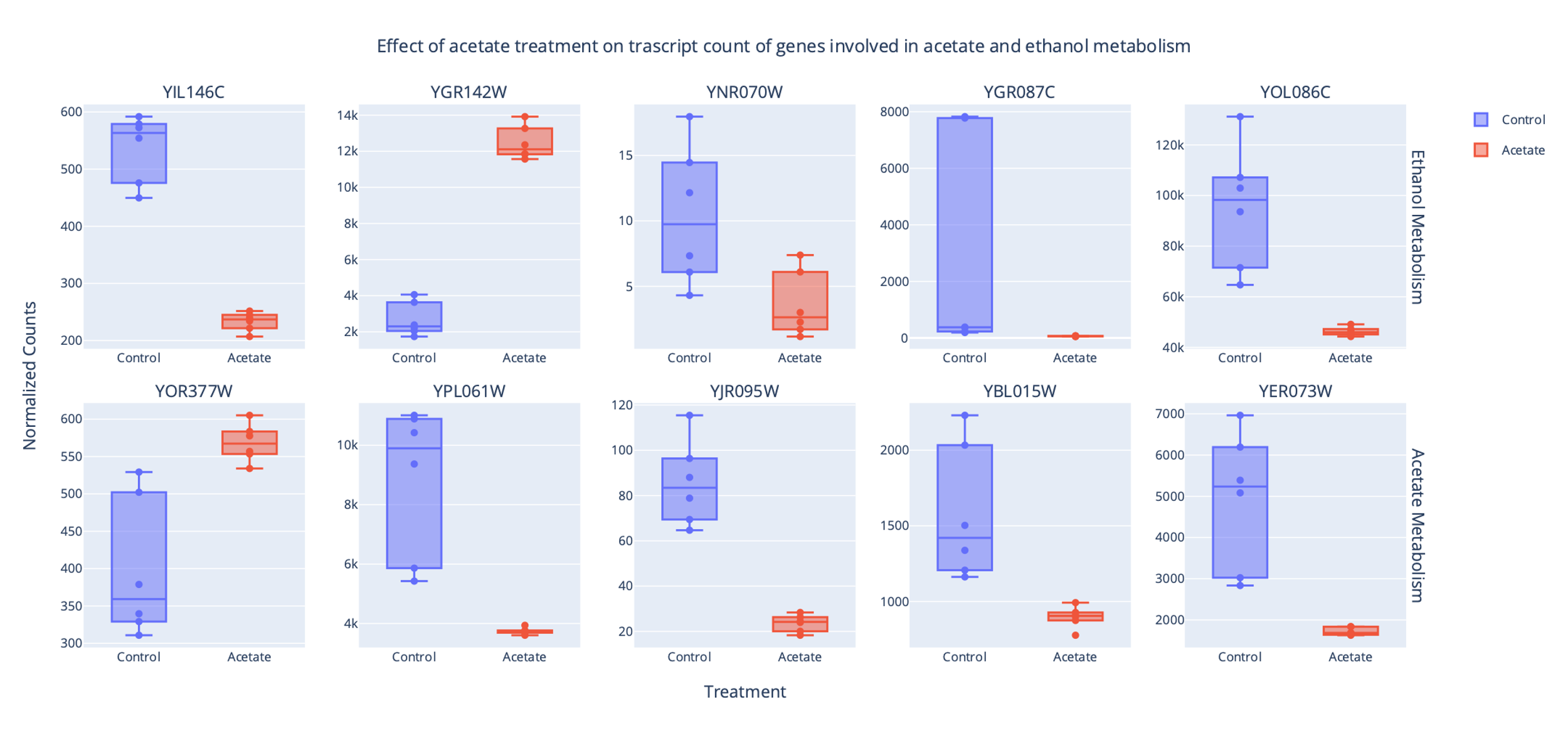
**Q3- DESeq Plot**



**Figure 2** Heatmap showing the Euclidian distance between different samples. Brighter colos represent higher **dissimilarity** between the samples. Rows and column show the samples. Dendograms show the clustering of different samples.

This plot shows the dissimilarity between the samples and shows how the samples cluster together based on their similarities. It is notable that samples in the same treatment group are all clustered together. Interestingly, the replicates are also clustered together. This means that treatment with acetate has changed the transcriptome. This plot was made using DESeq2 library in R by following the template code given in the class.

**Q4- Wildcard plot**



**Figure 3** Boxplot showing normalized transcript counts for genes involved in acetate and ethanol metabolism. Each plot shows transcript count comparison for individual genes between acetate treatment and control. Rows show whether the gene is involved in ethanol or acetate metabolism. Titles for each subplot shows the gene identifiers in Ensembl genome browser annotation file.

This plot shows how acetate stress changes the transcription of genes related to acetate and ethanol metabolism. To make this plot I extracted the normalized counts from DESeq pipeline. Then fetched the .cds file with gene information from ensemble genes which describes what is the function of each gene in the genome. I collected 5 genes involved in ethanol and acetate metabolism from this file and plotted the box plot using these genes and my normalized counts. Specifically, I picked genes which have less known functionality. With the hope that this study can add to our understanding of these genes. I created this plot using Pandas and Plotly libraries. The procedure for generating the plot is provided in the Jupyter notebook for this question:

Jupyter notebook

**Q5- Discussion**

This study aimed to explore the effect of acetate stress on transcriptome. The figure in Q3 clearly shows what was expected. Samples in each treatment group are clustered together. Setting 0.5 as shrinkage threshold in DESeq, detected 2000 DEGs. This number is significantly higher than what is reported in the manuscript. One possible reason can be that I used HISAT with DESeq while they used Tophat and FPKM normalization. Both factors can cause large changes between analysis results. The fact that so many genes are detected as DEG in my analysis can rise the question that some of the underlying assumptions of DESeq might not be valid anymore. However, all of the genes I selected for plotting are detected with very high confidence, absolute value of the logfold changes are around 1 or even higher. Also s-value in all cases are less than 0.005.

We can see interesting patterns in fig Q4. Information for the gene identifiers can be extracted from the .cds files for SC3 genome or using info\_dict python dictionary in the jupyter notebook for Q4. YNR070W, YIL146C, YGR087C, YOL086C are all considered to be involved in ethanol production and they are all down regulated in acetate treatment. This means that lower ethanol production can be attributed to down regulation of these enzymes. YGR142W is known to be expressed under ethanol stress. It is possible that acetate has the same effect. Supplementary table 1 at the end of this document shows the data retrieved for each identifier from the cds file.

On other class of analysis that I think in interesting is to compare different timepoint. I did the analysis in DESeq for comparing different timepoints but did not do any visualization. Most interesting group to consider here is a comparison between time 200 min and 45 min in the group that were exposed to acetate. Additionally, taking the interaction of time and acetate treatment could be and interesting study to reveal positive or negative feedback loops in gene regulation.

A future experiment could be genetically modifying the yeast. For example, adding chain elongation genes so that acetate could be channeled to other value-added products. This by itself might mitigate the effect of stress and increase the economic benefit from fermentation process.

**Q6 Notebook**

I have a notebook for this project which I have provided in on canvas along with this document and in the Final\_Project directory in the GitHub repository for this project.

[Notebook](https://github.com/ParsaGhadermazi/RNA_Seq/edit/main/notebook.md)

**References**

[1] Y. Dong, J. Hu, L. Fan, and Q. Chen, “RNA-Seq-based transcriptomic and metabolomic analysis reveal stress responses and programmed cell death induced by acetic acid in Saccharomyces cerevisiae,” *Sci Rep*, vol. 7, Feb. 2017, doi: 10.1038/srep42659.