Assignment 2: Flor Yeast RNAseq Differential Expression Analysis

BINF*611: Genomic Methods for Bioinformatics

Vlad Bularca

March 10, 2025

Introduction

RNA-sequencing (RNA-seq) is a popular genomic technique that uses next-generation sequencing technology to determine the presence and abundance of RNA transcripts from biological samples. RNA-seq is particularly valuable for comparing differences in gene expression across different time points or experimental conditions. For example, RNA-seq can be used to quantify differences in RNA abundance between a control and treatment group through the identification of differentially expressed genes (DEGs).

Flor yeast are a strain of *Saccharomyces cerevisiae* that form biofilm/velum on the surface of fermenting wine. They are responsible for protecting the wine from oxidation and for creating unique flavours such as those found in Sherry wines (Martinez et al., 1997). Mardanov et al. (2020) used RNA-seq to identify differentially expressed genes for yeast velum from 3 stages: Early biofilm, thin biofilm and mature biofilm. To validate the results produced by Mardanov et al. (2020), this analysis aims to identify differentially expressed genes between the different stages of velum development using *edgeR.* Additionally, this analysis highlights the importance of filtering steps when processing RNA-seq data, as different parameters can significantly impact the identification of differentially expressed genes.

Methods

3 FASTQ files containing the RNA-seq reads for 3 replicates from each stage of velum development were retrieved from the /scratch/lukens/Assignment_2_Seqs directory on Graham. The reads from each file were aligned to the yeast_genome.fa genome from the /scratch/lukens/Assignment_2_Genome directory. The genome was indexed beforehand using *STAR*. The reads were also aligned to the reference genome using *STAR*. RNA molecule abundance was quantified using *featureCounts* and the results for all 9 samples were combined into one table for further processing. The differential expression analysis was completed in RStudio using the *edgeR* package.

Results

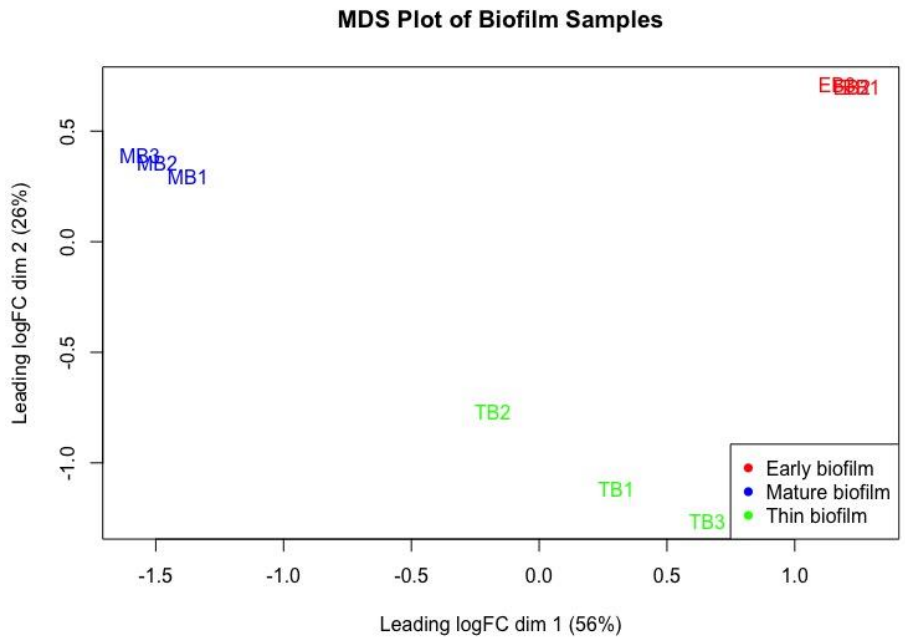

Figure 1: *MDS Plot of Biofilm Samples From 3 stages of Velum Development*
The resultant MDS plot shows clear separation between the 9 biofilm samples based on the stage of development. Axis titles indicate how much total variation is attributed to each dimension.

Table 1: *Top 10 Differentially Expressed Genes for Mature vs. Early Stages*
This table lists the top 10 genes from the Mature vs. Later pairwise comparison. The logFC value represents the magnitude of the fold change. A negative value indicates downregulation. a positive value indicates upregulation. Additional columns include p-value and FDR.

	Geneid	logFC	logCPM	PValue	FDR
1	YNR071C	7.937183	9.838766	3.976943e-13	1.426625e-09
2	YGR087C	-5.229461	8.663135	5.738571e-13	1.426625e-09
3	YKL164C	4.410209	8.242811	9.638126e-13	1.426625e-09
4	YHR094C	-5.175216	9.030082	1.056368e-12	1.426625e-09
5	YEL070W	8.829073	7.296968	3.057247e-12	3.303049e-09
6	YNR072W	5.809192	9.086940	4.899594e-12	3.764905e-09
7	YGL055W	-4.675852	10.919975	5.386645e-12	3.764905e-09
8	YJL158C	4.320165	7.576998	5.575571e-12	3.764905e-09
9	YEL069C	7.599045	7.768151	8.047008e-12	4.829993e-09
10	YJL052W	-5.353127	13.264373	1.065456e-11	5.755591e-09

Table 2: *Top 10 Differentially Expressed Genes for Mature vs. Thin Stages*
This table lists the top 10 genes from the Mature vs. Thin pairwise comparison. The logFC value represents the magnitude of the fold change. A negative value indicates downregulation. a positive value indicates upregulation. Additional columns include p-value and FDR.

	Geneid	logFC	logCPM	PValue	FDR
1	YCR105W	-12.999718	8.891108	3.584560e-12	1.936379e-08
2	YBR117C	4.814817	7.489681	8.655064e-12	2.337733e-08
3	YEL070W	5.720703	7.296968	2.119791e-11	3.226108e-08
4	YDR403W	4.293290	6.349145	2.737029e-11	3.226108e-08
5	YPR127W	3.557965	8.187702	2.986031e-11	3.226108e-08
6	YDR085C	4.081245	7.865396	3.725996e-11	3.354639e-08
7	YEL071W	-11.511329	7.713826	7.474313e-11	5.768034e-08
8	YMR105C	3.127885	8.362911	1.264706e-10	8.539928e-08
9	YMR009W	3.399390	7.660013	1.800893e-10	9.847774e-08
10	YGR043C	3.420108	8.244651	1.837287e-10	9.847774e-08

Table 3: *Top 10 Differentially Expressed Genes for Thin vs. Early Stages*
This table lists the top 10 genes from the Mature vs. Thin pairwise comparison. The logFC value represents the magnitude of the fold change. A negative value indicates downregulation. a positive value indicates upregulation. Additional columns include p-value and FDR.

	Geneid	logFC	logCPM	PValue	FDR
1	YGR087C	-4.639845	8.663135	1.232921e-12	6.660240e-09
2	YEL069C	7.954030	7.768151	3.204041e-12	8.654116e-09
3	YHR094C	-4.274993	9.030082	4.842766e-12	8.720207e-09
4	YKR097W	5.194769	8.263584	6.850697e-12	9.251866e-09
5	YCR105W	7.106111	8.891108	8.923112e-12	9.640531e-09
6	YPL095C	-3.811992	8.458841	4.529955e-11	3.412613e-08
7	YNR072W	4.808912	9.086940	4.917994e-11	3.412613e-08
8	YOR273C	-3.378544	8.478678	5.115904e-11	3.412613e-08
9	YGR088W	-4.866786	9.156351	5.685582e-11	3.412613e-08
10	YDL243C	4.184721	8.866863	7.396880e-11	3.995794e-08

Table 4: *Top 10 Differentially Expressed Genes for Early vs. Later Stages*
This table lists the top 10 genes from the Mature vs. Later pairwise comparison. The logFC value represents the magnitude of the fold change. A negative value indicates downregulation. a positive value indicates upregulation. Additional columns include p-value and FDR.

	Geneid	logFC	logCPM	PValue	FDR
1	YGR087C	4.934653	8.663135	1.739681e-14	9.397757e-11
2	YHR094C	4.725104	9.030082	4.680069e-14	1.264087e-10
3	YOR273C	3.670600	8.478678	6.949629e-13	1.251396e-09
4	YPL095C	3.880624	8.458841	1.143257e-12	1.543968e-09
5	YGR088W	4.537296	9.156351	2.745108e-12	2.516890e-09
6	YEL069C	-7.776538	7.768151	2.795509e-12	2.516890e-09
7	YNR071C	-6.344522	9.838766	3.655393e-12	2.742092e-09
8	YJL052W	4.104765	13.264373	4.086723e-12	2.742092e-09
9	YJL212C	4.192159	7.707184	4.568462e-12	2.742092e-09
10	YGL055W	3.361516	10.919975	5.715518e-12	3.087523e-09

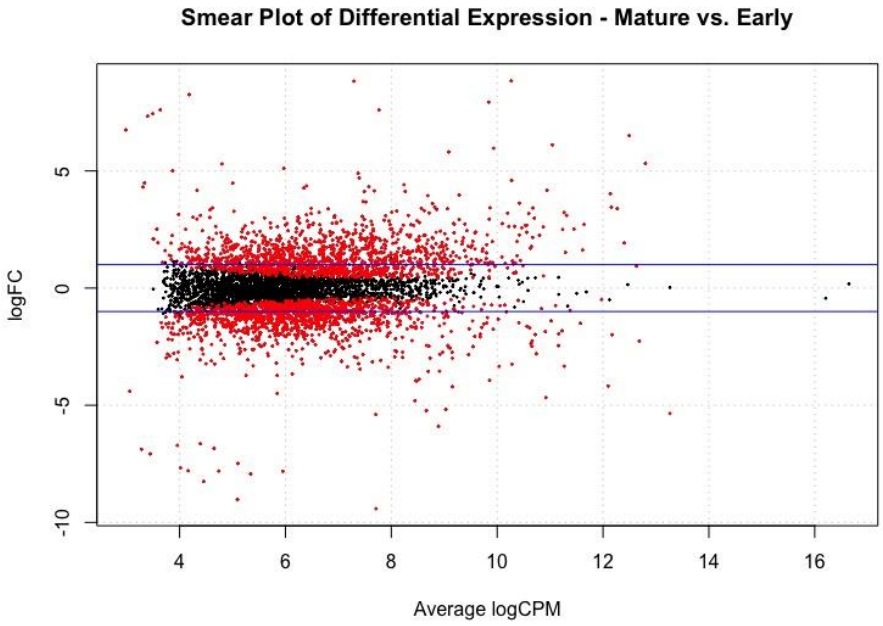

Figure 2: *Smear Plot for Mature vs. Early Comparison*
The smear plot illustrates the distribution of log fold changes (logFC) against the average log counts per million (logCPM), which represents expression level. The red points are genes that are differentially expressed, while the black pints are non-differentially expressed genes.

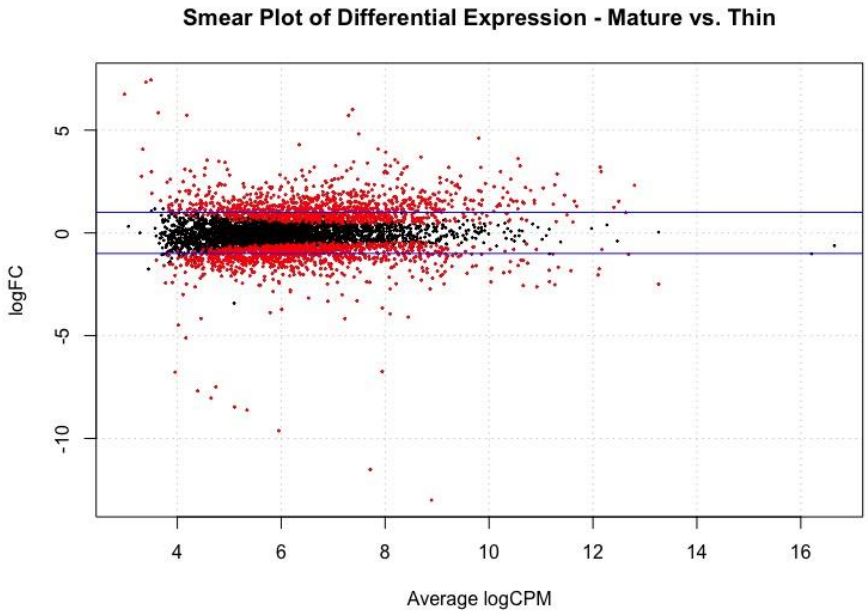

Figure 3: *Smear Plot for Mature vs. Thin Comparison*
The smear plot illustrates the distribution of log fold changes (logFC) against the average log counts per million (logCPM), which represents expression level. The red points are genes that are differentially expressed, while the black pints are non-differentially expressed genes.

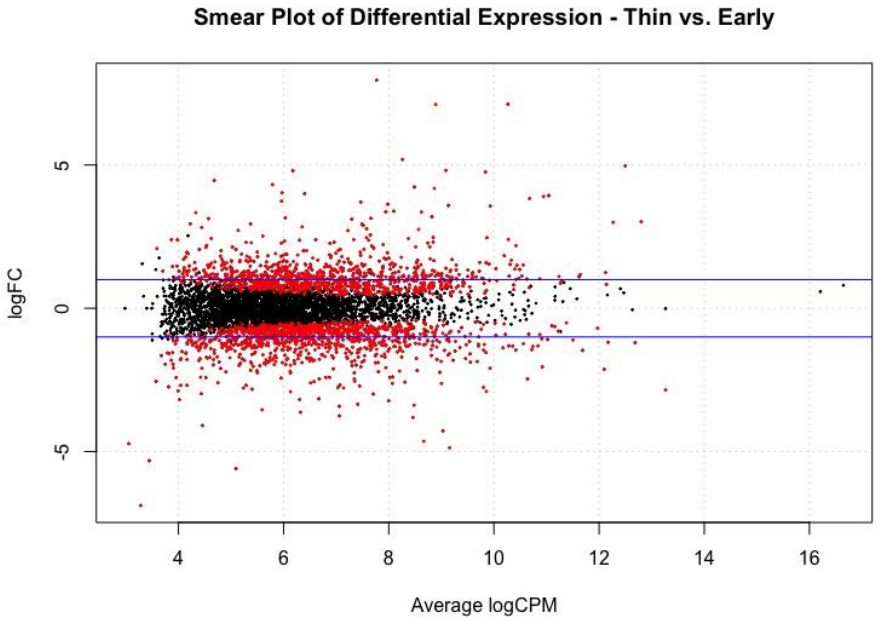

Figure 4: *Smear Plot for Thin vs. Early Comparison*
The smear plot illustrates the distribution of log fold changes (logFC) against the average log counts per million (logCPM), which represents expression level. The red points are genes that are differentially expressed, while the black pints are non-differentially expressed genes.

Discussion

This analysis successfully identified differentially expressed genes between samples from 3 stages of yeast velum development. Figure 1, a MDS plot based on the raw gene counts for each sample, clearly shows that samples from the same developmental stages form distinct clusters. The distance between points is determined based on the similarity of gene expression profiles between samples, and the generation of a MDS plot is a crucial quality control step used to validate sample quality.

The results for the differential expression analyses are as follows and were derived using False Discovery Rate (FDR) < 0.05 as the threshold: Mature vs. Thin – 1110 downregulated, 1113 upregulated & 2213 total DEGs, Mature vs. Early – 1462 downregulated, 1405 upregulated & 2857 total DEGs, Thin vs. Early – 1075 downregulated, 998 upregulated & 2073 total DEGs, Early vs. Later Stages – 1244 downregulated, 1359 upregulated & 2063 total DEGss.

The top 10 DEGs for each comparison can be found in Tables 1-4. The comparison of thin vs. early stages produced the lowest number of DEGs. There were more DEGs for the mature vs. thin stage when compared to the total DEGs for the thin stage vs. the early stage, and several genes had log fold changes greater than 7 (e.g., YEL071W). This is highlighted by the smear plots in Figures 2-4. This result may suggest that flor yeast undergo more pronounced biological changes to adapt to new conditions when transitioning from the thin stage to the mature stage. It is believed that the formation of a biofilm is a method to promote growth and maintain oxygen availability during the wine aging process (Zara et al., 2020).

Overall, these results alone do not provide much insight without further investigation. To get meaningful insights out of these lists, next steps would be to identify gene names and function based on the gene IDs for the DEGs from each comparison. Genes with large log fold changes (+/- 4) should be prioritized, as suggested by Mardanov et al. (2020). Additionally, it would be valuable to identify genes specific to each stage of development, such as the genes expressed solely in the mature stage, and understand their developmental role. Mardanov et al. (2020) identified fives genes exclusive to the mature biofilm stage: *YJL218W* (acetyltransferase), *AIF1* (apoptosis-inducing factor1), *ENB1* (siderophore iron transporter), *REE1* (regulator of enolase expression) and *CSS3*(hypothetical protein).

In conclusion, *edgeR* is a powerful tool for identifying DEGs between yeast velum samples from different developmental stages. This analysis provides further insight into the molecular mechanisms and functional roles of key genes involved in velum development.

References

Mardanov, A. V., Eldarov, M. A., Beletsky, A. V., Tanashchuk, T. N., Kishkovskaya, S. A., & Ravin, N. V. (2020). Transcriptome Profile of Yeast Strain Used for Biological Wine Aging Revealed Dynamic Changes of Gene Expression in Course of Flor Development. *Frontiers in microbiology*, *11*, 538. https://doi.org/10.3389/fmicb.2020.00538

Martínez, P., Pérez Rodríguez, L., & Benítez, T. (1997). Factors which affect velum formation by flor yeasts isolated from sherry wine. *Systematic and Applied Microbiology*, *20*(1), 154–157. https://doi.org/10.1016/S0723-2020(97)80060-4

Zara, S., Gross, M. K., Zara, G., Budroni, M., & Bakalinsky, A. T. (2010). Ethanol-independent biofilm formation by a flor wine yeast strain of Saccharomyces cerevisiae. *Applied and environmental microbiology*, *76*(12), 4089–4091. https://doi.org/10.1128/AEM.00111-10