BINF 6110 Assignment 2:

Introduction:

The performance of differential gene analysis is an essential approach in RNA sequencing studies, which allows research to help identify genes that may exhibit change in gene expression during experimental conditions. The data was obtained from a study focusing on Saccharomyces cerevisiae change in gene expressions between two different treatments conditions across different time intervals. It was observed that after treatment 816 genes were upregulated and 674 genes were downregulated by 2-fold after mature biofilm stage compared to early biofilm stage (1). The understanding of these altered gene expressions can provide an insight to various biological and molecular mechanism the yeasts use to adapt to wine making conditions (1).

During the conduction of this project, the primary objective was to analyze whether the featureCount output file generated in Linux demonstrated gene expressions being altered significantly due to different treatment conditions, intervals, and interactions. The primal goal of this assignment is to determine if gene expressions altered when exposed to different treatments and whether gene expression altered during the treatment. This was conducted in effort to determine if certain genes were downregulated or upregulated to adjust to experimental conditions. The differential gene analysis is a great way of understanding the molecular level mechanisms to improve the quality of perhaps medicine being developed or in this case improve the quality of winemaking (2). However, differential gene analysis can provide time consuming and non-intuitive results due to an influx of various tools and methods developed (2).

Even though, edgeR and RNA-seq are robust tools for detecting and analyzing differential gene expressions, they also have some limitations. The accuracy of the analysis is heavily depended on the quality of sequencing data and the statistical cutoffs, such as false discovery rate (FDR). Having a keen understanding of the data is very essential to determine the statistical cut off range, which in return affects the ability to obtain data of altered gene expression, therefore careful consideration of statistical thresholds and biological significance of the data is essential.

Methods:

File path: /home/ebad/scratch/BINF6110/Yeast/

- contains the aligned sequence and feature counts output

File path: /home/ebad/scratch/Assignment_2/

-contains indexed genome and RNA-seq sequence data

- **Load required modules**

```
module load StdEnv/2023

module load gcc/12.3.0

module load star/2.7.10

module load subread/2.0.3

module load r/4.2.2
```

- **Navigate to assignment directory**

```
cd /home/lukens/scratch/Assignment_2/
```

- **Navigate to genome directory**

```
cd /home/lukens/scratch/Assignment_2/Genome/
```

- **Navigate to user's scratch directory for working with data**

```
cd /scratch/ebad/Assignment_2/
```

- **Create a text file listing all SRA sample IDs**

```
nano sra_list.txt
```

- **Add the following sample IDs:**

```
SRR10551657

SRR10551658

SRR10551659

SRR10551660

SRR10551661

SRR10551662

SRR10551663

SRR10551664

SRR10551665
```

- **Download yeast reference genome**

```
wget -O /scratch/ebad/Assignment_2/Genome/yeast_genome.fa \
ftp://ftp.ensembl.org/pub/release-110/fasta/saccharomyces_cerevisiae/dna/Saccharomyces_cerevisiae.R64-1-1.dna.toplevel.fa.gz
```

- **Decompress genome file**

```
gunzip /scratch/ebad/Assignment_2/Genome/yeast_genome.fa.gz
```

- **Download genome annotation file (GTF)**

```
wget -O /scratch/ebad/Assignment_2/Genome/genome.gtf \

ftp://ftp.ensembl.org/pub/release-110/gtf/saccharomyces_cerevisiae/Saccharomyces_cerevisiae.R64-1-1.110.gtf.gz
```

- **Decompress GTF file**

```
gunzip /scratch/ebad/Assignment_2/Genome/genome.gtf.gz
```

- **Create directory for alignment results**

```
mkdir -p /scratch/ebad/BINF6110/Yeast/Aligned/
```

- **Loop through FASTQ files and align them using STAR**

```
for file in /scratch/ebad/Assignment_2/*_1.fastq.gz; do

    base=$(basename "$file" _1.fastq.gz)  # Extract sample ID

    STAR --runThreadN 8 \

        --genomeDir /scratch/ebad/Assignment_2/Genome/ \

        --readFilesIn /scratch/ebad/Assignment_2/${base}_1.fastq.gz \

        --readFilesCommand zcat \
        --outFileNamePrefix /scratch/ebad/BINF6110/Yeast/Aligned/${base} \
        --outSAMtype BAM SortedByCoordinate
done
```

- **Verify that alignment BAM files are created**

```
ls -lh /scratch/ebad/BINF6110/Yeast/Aligned/*.bam
```

- **Perform feature counting on one sample**

```
featureCounts -T 8 -t exon -g gene_id \ -a /scratch/ebad/Assignment_2/Genome/genome.gtf \
-o /scratch/ebad/BINF6110/Yeast/Counts/featureCounts_output.txt \
/scratch/ebad/BINF6110/Yeast/Aligned/*Aligned.sortedByCoord.out.bam
```

- **Verify feature count output**

```
head -n 20 /scratch/ebad/BINF6110/Yeast/Counts/featureCounts_output.txt
```

- **Summarize featureCounts output**

```
cat /scratch/ebad/BINF6110/Yeast/Counts/featureCounts_output.txt.summary
```

R Codes:

- **Read in the featureCounts output (raw count matrix)**

```
counts <- read.table("featureCounts_output_cleaned.txt", header=TRUE, row.names=1, sep="\t")
```

- **Remove metadata columns (Chr, Start, End, Strand, Length)**

```
counts <- counts[, -(1:5)]
```

- **Check the data structure**

```
head(counts)

dim(counts)

str(counts)

colnames(counts)
```

- **Define experimental metadata (Condition: TreatmentA/TreatmentB, Time: Before/After)**

```
sample_metadata <- data.frame(
  row.names = colnames(counts),
  Condition = c("TreatmentA", "TreatmentA", "TreatmentB", "TreatmentB",
                "TreatmentA", "TreatmentA", "TreatmentB", "TreatmentB", "TreatmentA"),
  Time = c("Before", "After", "Before", "After",
           "Before", "After", "Before", "After", "Before")
)
```

- **Convert to factor variables**

```
sample_metadata$Condition <- factor(sample_metadata$Condition)

sample_metadata$Time <- factor(sample_metadata$Time)
```

- **Check metadata**

```
print(sample_metadata)
```

- **Create a DGEList object for edgeR analysis**

```
dge <- DGEList(counts=counts, group=sample_metadata$Condition)
```

- **Filter lowly expressed genes**

```
keep <- filterByExpr(dge, group=sample_metadata$Condition)

dge <- dge[keep, , keep.lib.sizes=FALSE]
```

- **Normalize data using TMM normalization**

```
dge <- calcNormFactors(dge)
```

- **Check samples and normalization factors**

```
dge$samples
```

- **Create design matrix for modeling effects of Condition and Time**

```
design <- model.matrix(~ Condition * Time, data=sample_metadata)
```

- **Check column names in design matrix**

```
colnames(design)
```

- **Estimate dispersion for differential expression analysis**

```
dge <- estimateDisp(dge, design)
```

- **Fit a quasi-likelihood negative binomial generalized log-linear model**

```
fit <- glmQLFit(dge, design)
```

- **Test for differentially expressed genes between TreatmentA and TreatmentB**

```
res_treatment <- glmQLFTest(fit, coef="ConditionTreatmentB")
```

- **View top differentially expressed genes**

```
topTags(res_treatment)
```

- **Test for differential expression across TimeBefore and TimeAfter**

```
res_time <- glmQLFTest(fit, coef="TimeBefore")
```

- **View top differentially expressed genes**

```
topTags(res_time)
```

- **Test for interaction effects (Treatment * Time)**

```
res_interaction <- glmQLFTest(fit, coef="ConditionTreatmentB:TimeBefore")
```

- **View top differentially expressed genes for interaction**

```
topTags(res_interaction)
```

- **Generate an MA plot for TreatmentA vs. TreatmentB**

```
ggplot(res_treatment$table, aes(x=logCPM, y=logFC, color=F < 0.05)) +

  geom_point(alpha=0.7, size=1.3) +

  scale_color_manual(values=c("black", "red")) +

  geom_hline(yintercept=0, linetype="dashed", color="gray50") +

  theme_minimal() +

  labs(title="MA Plot: TreatmentA vs TreatmentB",
     x="Average log CPM", y="Log Fold Change") +
 theme(
  legend.position="top",
  plot.title = element_text(hjust=0.5, size=16, face="bold"),
  axis.title = element_text(size=14),
  axis.text = element_text(size=12)
 )
```

- **Generate a Volcano Plot for Before vs. After Treatment**

```
ggplot(res_time$table, aes(x=logFC, y=-log10(PValue), color=F < 0.05)) +
 geom_point(alpha=0.8, size=1.5) +
 scale_color_manual(values=c("black", "red")) +
 theme_minimal() +
 geom_hline(yintercept=-log10(0.05), linetype="dashed", color="blue", size=0.8) +
 geom_vline(xintercept=c(-1,1), linetype="dashed", color="gray50", size=0.8) +
 labs(title="Volcano Plot: Before vs After Treatment",
     x="Log2 Fold Change", y="-Log10 P-Value") +
 theme(
  legend.position="top",
  plot.title = element_text(hjust=0.5, size=16, face="bold"),
  axis.title = element_text(size=14),
  axis.text = element_text(size=12)
)
```

Results:

Figures:

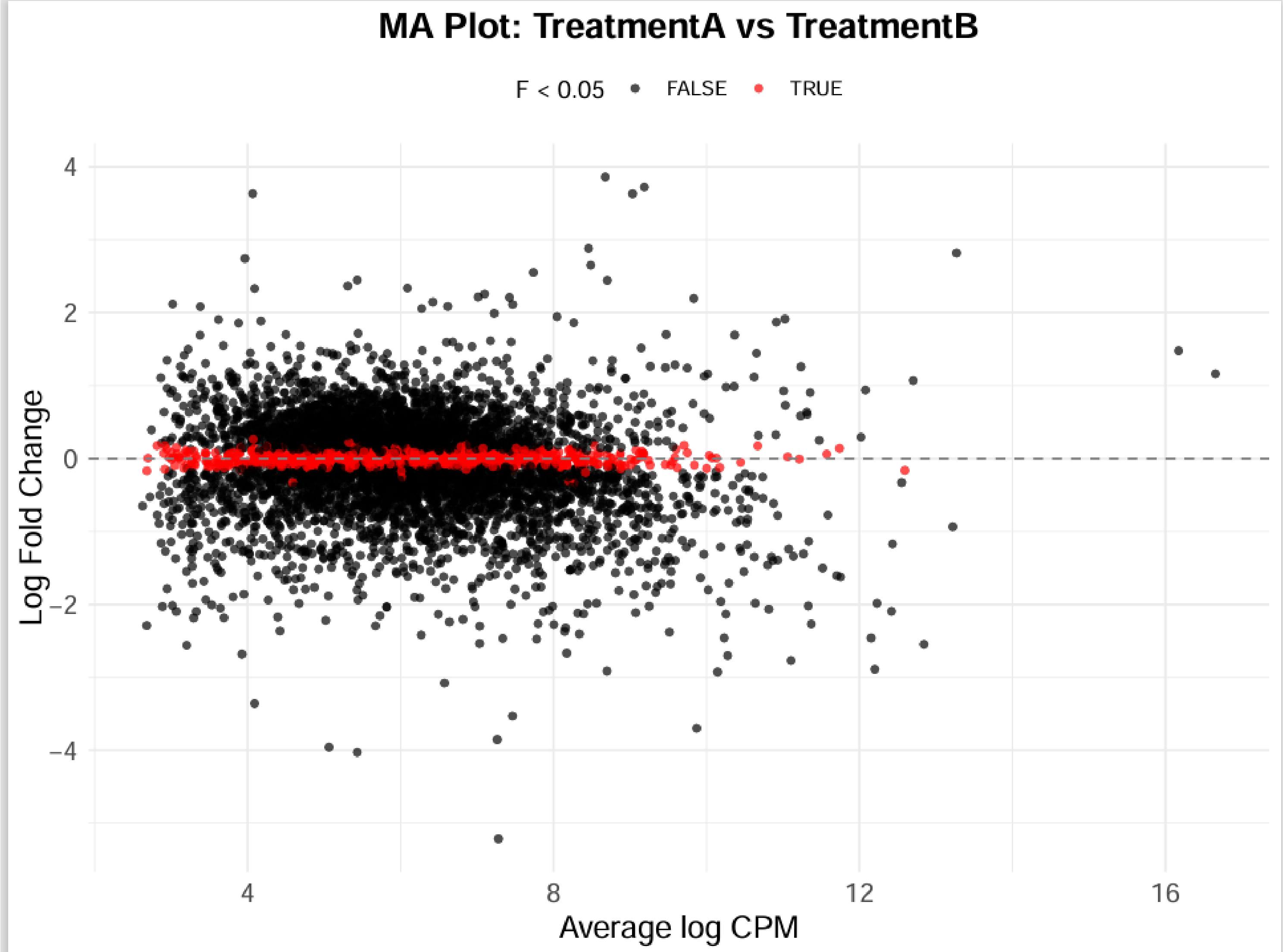


Figure 1: MA plot (TreatmentA vs TreatmentB)

This plot shows the log fold changes of gene expression levels across samples. Most genes cluster around a log fold change of 0, suggesting relatively small differences in expression. Genes marked in red exceed the FDR threshold of 0.05, indicating significant differential expression.

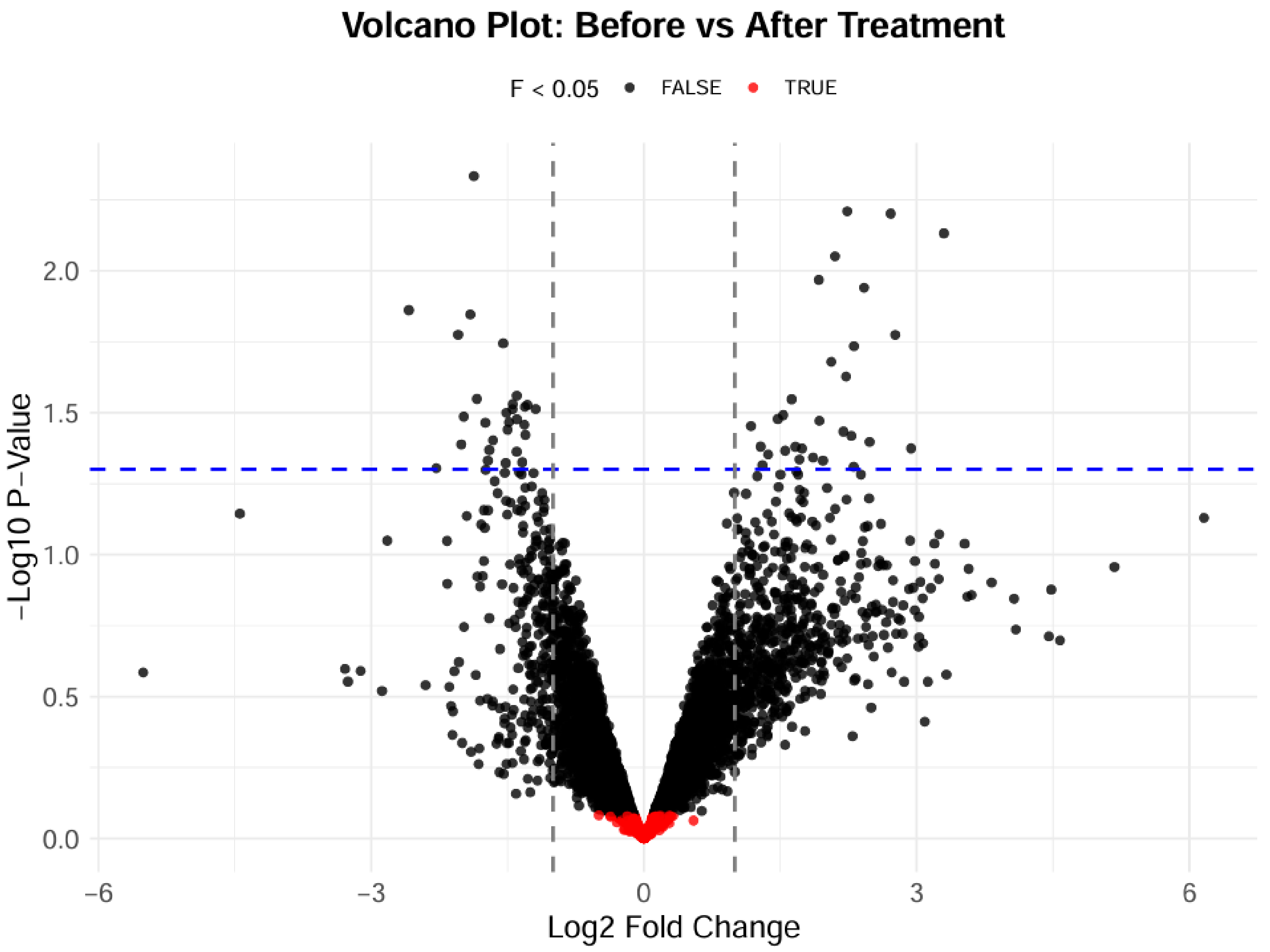
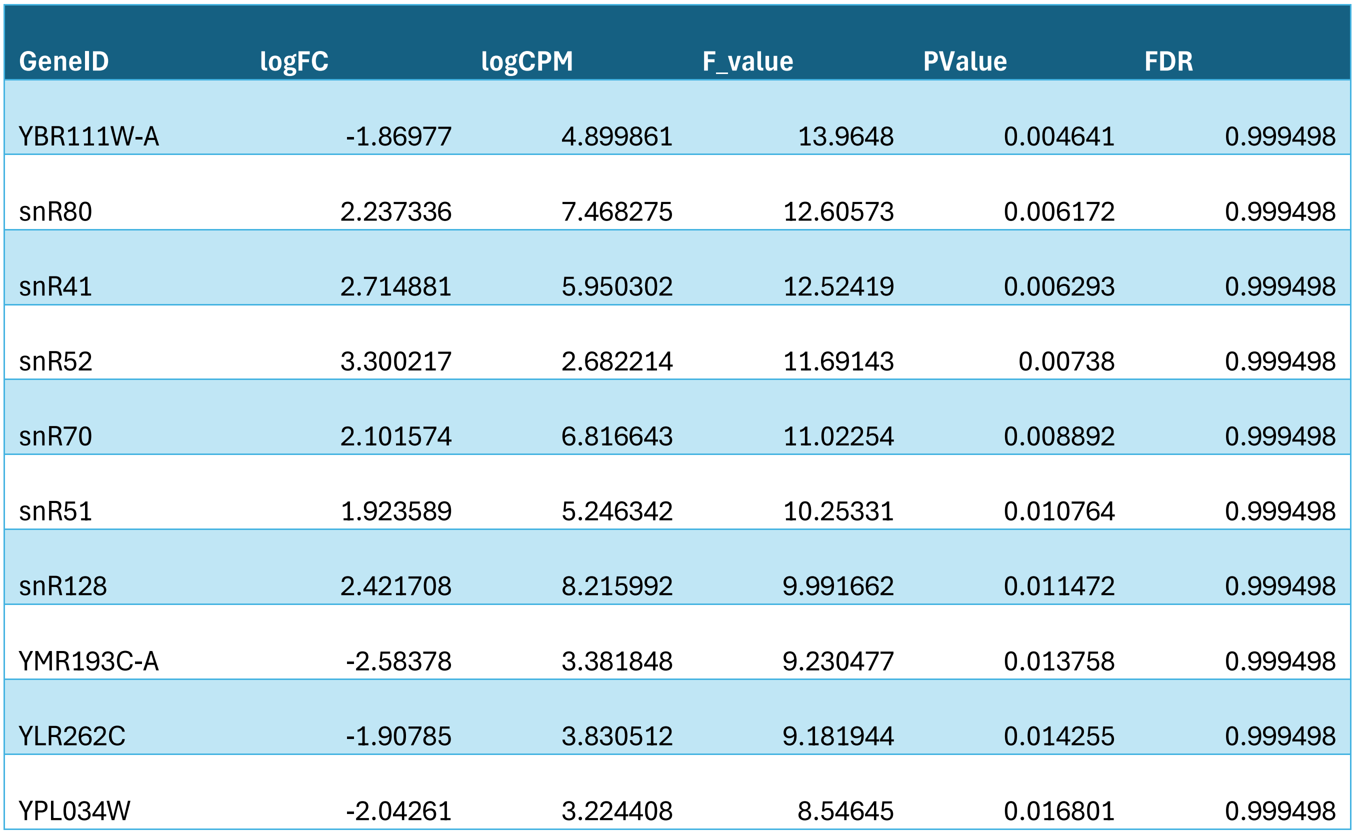


Figure 2: Volcano Plot (Before and After Treatment)

This volcano plot highlights genes with significant differential expression after treatment. The red dots represent genes that both a large log-fold change and a statistically significant p-value.

GeneID	logFC	logCPM	F_value	PValue	FDR
YAL044W-A	1.079678	3.399483	10.3977	0.009144	0.998524
YPL245W	-0.97026	5.969645	7.895877	0.018542	0.998524
YIL067C	-0.72319	4.994678	7.88657	0.018626	0.998524
YOR177C	1.164595	3.754874	7.401409	0.021734	0.998524
YOL110W	-1.26683	3.285408	7.385431	0.021824	0.998524
YDR531W	-0.653	5.433425	7.14516	0.023456	0.998524
YPR137W	-0.64207	5.049046	6.834177	0.025962	0.998524
YBR071W	-0.78912	5.410905	6.495381	0.029033	0.998524
YOR237W	1.006685	3.27694	6.444525	0.029487	0.998524
YHR172W	-0.61924	4.087502	6.285337	0.031288	0.998524

Table 1: Genes Differentially Expressed Between Treatment A and Treatment B



GeneID	logFC	logCPM	F_value	PValue	FDR
YBR111W-A	-1.86977	4.899861	13.9648	0.004641	0.999498
snR80	2.237336	7.468275	12.60573	0.006172	0.999498
snR41	2.714881	5.950302	12.52419	0.006293	0.999498
snR52	3.300217	2.682214	11.69143	0.00738	0.999498
snR70	2.101574	6.816643	11.02254	0.008892	0.999498
snR51	1.923589	5.246342	10.25331	0.010764	0.999498
snR128	2.421708	8.215992	9.991662	0.011472	0.999498
YMR193C-A	-2.58378	3.381848	9.230477	0.013758	0.999498
YLR262C	-1.90785	3.830512	9.181944	0.014255	0.999498
YPL034W	-2.04261	3.224408	8.54645	0.016801	0.999498

Table 2: Genes Differentially Expressed After Treatment

Discussion:

The analysis of differential gene expression conducted in this project was completed in hopes of assessing the impact of different treatments applied and the duration of those treatments had on gene expression. The results obtained during this study indicates that there were some genes exhibiting fold changes in expression, however the adjusted p-values (FDR) remain high across all the comparisons, which suggests that no genes examined depicts significant changes in gene expression.

Figure 1 illustrates the MA plot comparing Treatment A and Treatment B, showcased that most genes clustered around a log fold change (logFC) of zero, which is indicative of the minimal differences in gene expression between the two treatments being compared. The MA plot depicts very little upregulated and down regulated genes, indicating majority of genes had a small fold change. Figure 2, the volcano plot demonstrated a similar trend, where the plot for time effects shows a symmetrical distribution of genes with very few crossing the threshold. The blue line represents significant difference, and grey line represents 2-fold change. As demonstrated in the plot fewer genes had 2-fold change and went over the significance line. The results visualized through these two plots indicate that perhaps experimental conditions did not induce strong differential gene expression or could very well perhaps be that the statistical power was not sufficient to detect changes.

The methods applied during this project, were successfully implemented to complete the RNA-seq pipeline from applying STAR alignment tool to utilizing edgeR to perform differential gene expression analysis. Through the use of featureCounts in Linux, the feature counts output file was generated and successfully depicted the number of aligned sequences. The visual data was successfully generated in R studio. The result obtained through of the performance of various computational software used, demonstrated no significant differential gene expression, which necessarily does not leave room to perform further analysis, however using alternative computational methods such as weighted gene co-expression network analysis, may provide new insights into the data being analyzed.

References:

1. 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
2. McDermaid, A., Monier, B., Zhao, J., Liu, B., & Ma, Q. (2019). Interpretation of differential gene expression results of RNA-seq data: review and integration. *Briefings in bioinformatics*, *20*(6), 2044–2054. https://doi.org/10.1093/bib/bby067