BINF 6110 Assignment #2

Jacob Hambly

**Introduction**

***What is RNA sequencing?***

RNA sequencing is a powerful methodology most commonly used for the transcriptome-wide analysis of differential gene expression between different cells (Stark et al., 2019). RNA sequencing has contributed to the understanding of phenotypic variation amongst organisms under different conditions/treatments (Ritchie et al, 2015). This has given scientists an extensive glimpse into how cells operate under both normal and diseased states (Ritchie et al, 2015). RNA sequencing has been used extensively in many different research domains such as cancer research and the study of cell development (Ritchie et al, 2015). The standard RNA sequencing workflow is as follows: extracting and enriching mRNA reads from cells, synthesis of cDNA from the mRNA, ligating sequence adaptors to the cDNA fragments, amplification of the sequences through PCR to construct a library, sequencing of the fragments and then finally bioinformatic analysis of the sequence reads (Stark et al., 2019).

***Challenges of RNA sequencing***

Many bioinformatic software tools have been developed to overcome the particular challenges of aligning cDNA to a reference genome. For instance, due to the nature of mRNA where exons in a gene region can be alternatively spliced with the removal of introns, regular DNA aligner tools would struggle to contiguously map the cDNA byproducts of mRNA (Dobin et al., 2012). However, specialized software tools for cDNA alignment such as STAR have been developed that account for alternate splicing events (Dobin et al., 2012). This has led to the accurate identification of genes involved in certain cellular processes as well as the accurate quantification of gene expression levels (Costa-Silva et al., 2017). Another challenge for RNA sequencing is the intricate nature of the methodology for the construction of libraries of cDNA fragments (Shi et al., 2021). The cDNA fragments must be amplified through PCR prior to sequencing for the most used next-generation sequencing platforms such as Illumina and Pacific Biosciences (Shi et al., 2021). However, PCR amplification has been shown to introduce bias in amplifying certain sequences over others. This can lead to erroneous measures of gene expression during downstream analysis (Shi et al., 2021). Thus, researchers must consider the possibility of unintentional biases in their results (Shi et al., 2021).

***Goal of this analysis***

The goal of this analysis is to capture the patterns of differential gene expression of the flor strain of *Saccharomyces cerevisiae* during velum development. The flor strain of *S.cerevisiae* is an important contributor during the aging process of different types of wine such as sherry (Mardanov et al., 2020). The development of the velum by *S.cerevisiae* gives the wine protection from oxidation during the aging process and enables biochemical alterations important in the production of wine (Mardanov et al., 2020). This analysis will study the gene expression levels of *S.cerevisiae* under three different velum development stages: early biofilm development, thin biofilm development, and mature biofilm development.

**Methods and Results**

***Workflow of the analysis***

The data for this analysis is from the 2020 study, *Transcriptome Profile of Yeast Strain Used for Biological Wine Aging Revealed Dynamic Changes of Gene Expression in Course of Flor Development,* by Mardanov et al. Nine fastq files that represent the cDNA fragment reads for three biological replicates for each velum development phase as well as the indexed yeast genome are the inputs for this analysis. To begin the analysis, the nine fastq were each mapped to the reference yeast genome using the ‘alignReads’ function from the STAR alignment software. This created nine ‘.sam’ files which contained information about where the cDNA fragments are mapped to the yeast reference genome. To decrease the time required to transfer these files to my local computer, I converted each of these files into a binary ‘.bam’ file.

 I then created a ‘feature counts’ object in R using the featureCounts function from Rsubread. The featureCounts function assigns the mapped sequence reads from the ‘.bam’ files to specific genes in the yeast genome based on where the reads are aligned using the annotated yeast genome gtf file. Using the ‘feature counts’ object, a differential gene expression (DGE) object is created using the DGEList function in edgeR. This object type allows for further downstream analysis of the counts data. Next, data preprocessing is conducted to filter out low expression across the samples followed by normalization of the counts data to account for different library sizes. Design matrices were then constructed to represent the experimental groups and the observed differential expression data. Dispersion was then calculated to estimate the variability in gene expression for each gene across the samples. Finally, a linear model was then fit for each gene to identify differential gene expression between the groups using a quasi-likelihood f test.

A graph with colored dots

Description automatically generated

Figure 1. PCA plot for the normalized counts for each of the different sample groups of velum development

Table 1. Number of genes identified as being significantly differentially expressed using the quasi-likelihood f-test between each of the velum development groups.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Early vs Mature** | **Early vs Thin** | **Thin vs Mature** | **Early vs Later (Thin + Mature)** |
| **Downregulated** | 1733 | 1423 | 1297 | 980 |
| **Not Significant** | 2387 | 3214 | 2936 | 3976 |
| **Upregulated** | 1652 | 1135 | 1539 | 816 |
| **Total Significant** | 3385 | 2558 | 2836 | 1796 |

Table 2. Number of significant genes with a fold change greater than 2 between each of the velum development groups being compared.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Early vs Mature** | **Early vs Thin** | **Thin vs Mature** | **Early vs Later (Thin + Mature)** |
| **Downregulated** | 1042 | 702 | 700 | 847 |
| **Not Significant** | 3839 | 4567 | 4428 | 4433 |
| **Upregulated** | 891 | 503 | 644 | 492 |
| **Total Significant** | 1933 | 1205 | 1344 | 1339 |

A graph of different colored dots

Description automatically generated with medium confidence

A graph of different colored dots

Description automatically generated with medium confidence

Figure 2. Volcano plots for each of the comparisons between the velum development groups.

A screenshot of a graph

Description automatically generated

Figure 3. Heatmap of gene count for the top 100 differentially expressed genes between the early and not early samples.

***Results***

A PCA plot was created to show the separation of the samples based on the normalized gene expression counts. PC1 explains 83.9% of the total variance between the samples while PC2 explains 11.78% of the total variance. Thus, taken together, PC1 and PC2 explain 95.68% of the total variance in the normalized gene expression data set.

The results from the quasi-likelihood f-test between each of the velum development groups are shown in table 1.  A quasi-likelihood f-test was chosen over the pairwise test due to more robust error controls compared to the pairwise test (Chen et al., 2016). Table two represents a subset of values from table 1 that filters for a fold change greater than 2.  The fold change is calculated as the log2 of the ratio of the mean normalized count of a gene in one of the biological groups and the mean normalized count of the same gene in a different biological group. A log2 fold change of 1 and -1 was chosen as this represents at least a two-fold increase or decrease in gene expression respectively (log2(2) = 1).

Figure 2 shows four volcano plots that represent the differential expression between the velum development groups.  The x-axis represents the log2 fold change in the gene expression levels between the two velum development groups and the y-axis shows the -log10 of the adjusted p-value associated with each gene. The p-values were adjusted for multiple testing using the Benjamini-Hochberg (BH) procedure. Genes with a log2 fold change less than -1 and a -log10(adjusted p-value) greater than 1.3 were identified as being significantly downregulated and coloured in green. Genes with a log2 fold change greater than 1 and a -log10(adjusted p-value) greater than 1.3 were identified as being significantly upregulated.

Figure 3 shows a heat map representing the counts for the top 100 differentiated genes identified using the quasi likelihood f-test based on adjusted p-values for the early compared with the not early samples. The specific group for the ‘not early’ samples are shown as well. The z-score of the normalized counts data for the top 100 genes was used to facilitate easier comparison of the normalized count data by shrinking the scale of the counts. A value closer to two and the colour red represents a higher count and a value closer to -2 and the colour blue represents a lower count.

**Discussion**

Based on the results, this analysis was successful in capturing the patterns of differential gene expression between each of the possible velum development groups. As you can see from figure 1, the early samples cluster together while the later samples (thin and mature) cluster with themselves. Looking closer, the PC2 seems to separate the early cluster of samples from the thin and mature cluster of samples more effectively than the PC1. Additionally, when looking at the PC1, the thin samples are not clustered well when compared with the other two groups. Since the early samples clustered well together and showed separation from the other two samples, it was logical to also examine the differential gene expression data of the early group compared to the other two groups taken together.

As you can see in table 1, there are many significantly differentially expressed genes between the velum development stages, with the largest number of significantly differentiated genes being in the early vs mature group. This would make sense as we would expect there to be distinct patterns in gene expression between the earliest and latest stages of velum development (Mardanov et al., 2020). However, looking at the results from table 2 and figure 2 could give us a better insight into the subset of genes that may be most biologically relevant. This is because, from those genes being identified as significantly differentially expressed in table 1, we are filtering for those genes that show at least a 2 fold increase or decrease in gene expression between the two groups. This would indicate a large change in expression patterns for a particular gene. When comparing the early group to both the thin and mature groups taken together, there is also a large amount of significantly differentially expressed genes, with more genes being downregulated in the early group compared to the later group. Therefore, we could conclude that as velum development progresses, we see more genes being upregulated in the later development stages compared to the early stage. These genes identified in table 2 and figure 2 that are being significantly upregulated could be contributing to the development of the velum itself as well as other cellular processes needed for yeast to survive during velum development (Mardanov et al., 2020).

When viewing figure 3, we can see that the patterns in counts closely reflect the information contained in the PCA plot (figure 1). The three early samples show clear differences in expression levels for the top 100 genes compared to the rest of the samples. Additionally, the three mature samples are very similar to one of the thin samples whereas two of the thin samples show slight distinctions in count patterns. Thus, figure 3 shows differential gene expression between the different velum development groups that match what is seen from the PCA plot. Taken together, this emphasizes a clear change in gene expression patterns between the early velum development group and that of the later velum development groups.

An extension of this analysis would be to include the gene oncology for the genes significantly upregulated and downregulated (fold change greater than 1 or less than -1) between each of the groups. This would allow for the investigation of the function of the differentially expressed genes and thus allow for a potential explanation of the mechanisms behind how the *S.cerevisiae* cells produce the velum which could have important implications for the winemaking industry.

**Citations**

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: The teenage years. *Nature Reviews Genetics*, *20*(11), 631–656. https://doi.org/10.1038/s41576-019-0150-2

Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, Gordon K. Smyth, limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Research, Volume 43, Issue 7, 20 April 2015, Page e47, <https://doi.org/10.1093/nar/gkv007>

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*. https://doi.org/10.3389/fmicb.2020.00538

Shi, H., Zhou, Y., Jia, E., Pan, M., Bai, Y., & Ge, Q. (2021). Bias in RNA-seq library preparation: Current challenges and solutions. *BioMed Research International*, *2021*, 1–11. https://doi.org/10.1155/2021/6647597

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2012). Star: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Costa-Silva, J., Domingues, D., & Lopes, F. M. (2017). RNA-seq differential expression analysis: An extended review and a software tool. *PLOS ONE*, *12*(12). https://doi.org/10.1371/journal.pone.0190152

Chen, Y., Lun, A. T., & Smyth, G. K. (2016). From reads to GENES TO PATHWAYS: Differential expression analysis of RNA-seq experiments using Rsubread and the Edger quasi-likelihood pipeline. *F1000Research*, *5*, 1438. https://doi.org/10.12688/f1000research.8987.1

Appendix

**Code used:**

**Graham**

#requesting a compute node for 1 hour with 32gb of ram

salloc --time 1:00:00 --mem 32G

#align each RNAseq file to the yeast reference genome

 for i in \*.fastq; do j=$(basename "$i"); STAR --runMode alignReads --genomeDir . --readFilesIn $i --outFileNamePrefix $j; done

#convert each fastqAligned.out.sam file into a fastqAligned.out.bam

 for i in \*.sam; do j=$(basename $i .sam); samtools view -bS $i -o $j.bam; done

#transfer .bam files to my personal pc for further analysis in R

scp [jhambly@graham.computecanada.ca](mailto:jhambly@graham.computecanada.ca):scratch/BINF6110/a2/\*.bam .

**R**

R code can be found in the accompanying rscript file titled: binf6110\_assignment2\_rfile.R

**Lists**

The lists for the differentially expressed genes can be found in gene\_diffexp\_any.csv and gene\_diffexp\_earlyvsother.csv