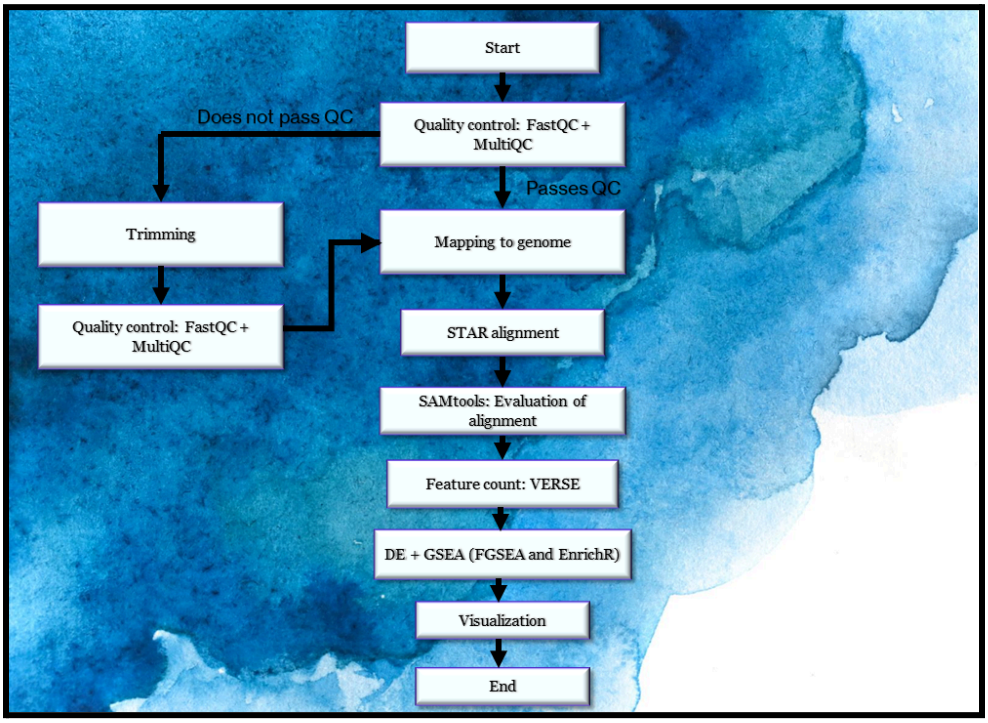
BF 528 FINAL PROJECT
RNA-SEQ ANALYSES



PROJECT PROPOSAL

1.1 Introduction

In this project, I will perform a comprehensive analysis of RNAseq data derived from 6 samples, including 3 control (CTL) and 3 knock-out (KO) samples from a human source. The primary objective is to conduct a basic differential expression analysis comparing the CTL and KO conditions. Additionally, I will perform quality control (QC) at both the read and alignment levels, followed by exploratory data analysis and functional enrichment analysis to elucidate biological insights.

1.2 Methods

1.2.1 Quality Control (QC):

- Utilize FastQC (version 0.11.9) [1] and MultiQC (version 1.20) [2] for initial read quality assessment.

- Perform trimming and filtering using Trimmomatic (version 0.39) [3] to remove low-quality bases and adapters.
- Assess read quality post-trimming with FastQC (version 0.11.9) [1] and MultiQC [2].
- Use STAR (version 2.7.9a) [4] for read alignment against the human reference genome.
- Evaluate alignment statistics including mapping rates, duplication rates, and coverage using SAMtools (version 1.9) [5].

1.2.2 Differential Expression Analysis:

- Generate count matrices using Verse v2.0.5.
- Conduct differential expression analysis using EdgeR v3.26.0.
- Determine appropriate fold change and false discovery rate (FDR) thresholds.
- Subset significant differentially expressed (DE) genes based on chosen statistical thresholds.

1.2.3 Exploratory Data Analysis:

- Perform principal component analysis (PCA)
- Interpret the plot to understand sample relationships and potential batch effects.

1.2.4 Functional Enrichment Analysis:

- Employ enrichR (version 3.0) for functional annotation and enrichment analysis.

1.3 Deliverables

1.3.1 Quality Control Assessment:

- Provide QC results.
- Address any concerns regarding sequencing read quality and alignment statistics.
- Decide on sample exclusion based on QC results.

1.3.2 PCA Plot:

- Provide PCA plot.
- Interpret the plot to understand potential batch effects.

1.3.3 Differential Expression Analysis Results:

- Provide a CSV containing DE analysis results.
- Generate a histogram showing the distribution of log2FoldChanges.
- Create a volcano plot distinguishing significant DE genes.

1.3.4 Functional Enrichment Analysis Results:

- Perform enrichR (GSEA) analysis.
- Provide analysis results summarized in a table and figure.
- Discuss implications of results on biological functions related to the factor of interest.

METHODS

2.1 Quality Control (QC):

- FastQC (version 0.11.9) [1] and MultiQC (version 1.20) [2] were utilized for initial read quality assessment. This step aimed to identify potential issues such as overrepresented sequences, adapter contamination, and base call quality scores.

2.2 Trimming:

- Trimming and filtering were performed using Trimmomatic (version 0.39) [3] to remove low-quality bases and adapters. Parameters were set to remove bases with a Phred score < 20 from the leading and trailing ends of reads, as well as to perform adapter trimming and remove reads shorter than 36 base pairs.

2.3 Post-trimming Quality Control (QC):

- Read quality post-trimming was reassessed with FastQC (version 0.11.9) [1] and MultiQC [2]. This step ensured that trimming improved the overall quality of the reads.

2.4 Read Alignment and Assesment:

- Trimmed reads were aligned to the human reference genome (GRCh38) using STAR v2.7.9a [4]. The alignment parameters were set to account for the paired-end nature of the reads, and alignment files were generated in BAM format.
- SAMtools v1.9 [5] was utilized to extract alignment statistics, including mapping rates, duplication rates, and coverage. These statistics provided insights into the quality and efficiency of the alignment process.

2.5 Count Matrix Generation:

- Gene-level count matrices were generated from the aligned reads using Verse v2.0.5 [6]. This step involved quantifying the number of reads mapping to each gene in the genome.

2.6 Differential Expression Analysis:

- EdgeR v3.26.0 [7] was employed to perform differential expression analysis between control and knockout conditions. Statistical thresholds for fold change and false discovery rate (FDR) were determined, and genes showing significant differential expression were identified.

RESULTS AND DISCUSSION

3.1 Briefly remark on the quality of the sequencing reads and the alignment statistics, make sure to specifically mention the following:

Are there any concerning aspects of the quality control of your sequencing reads?

Are there any concerning aspects of the quality control related to alignment?

Based on all of your quality control, will you exclude any samples from further analysis?

3.1.1 Quality Control of Sequencing Reads

- Based on the MultiQC report, all sequences had passed basic quality checks (e.g., per-base sequence quality, per-sequence quality scores, etc.), whereas a couple of "per_base_n_content" fails indicated that there might be an excessive number of ambiguous bases (N) in the sequences, which could affect downstream analysis. The presence of warnings for "overrepresented_sequences" for some samples suggested that there might be sequences that are overrepresented in the dataset, which could be contaminants or artifacts.
- This warranted for trimming and filtering of the data to remove low-quality bases and adapters.
- Compared to the pre-trimming data, the percentage of sequences flagged as poor quality is now zero for all samples. Additionally, the average and median sequence lengths are consistent across samples post-trimming.

3.1.2 Quality Control of Aligned Reads

- All reads are mapped, with a high percentage properly paired, indicating reliable alignment to the reference genome. Additionally, the absence of QC-failed reads and mate pairs mapped to different chromosomes indicates good quality data.

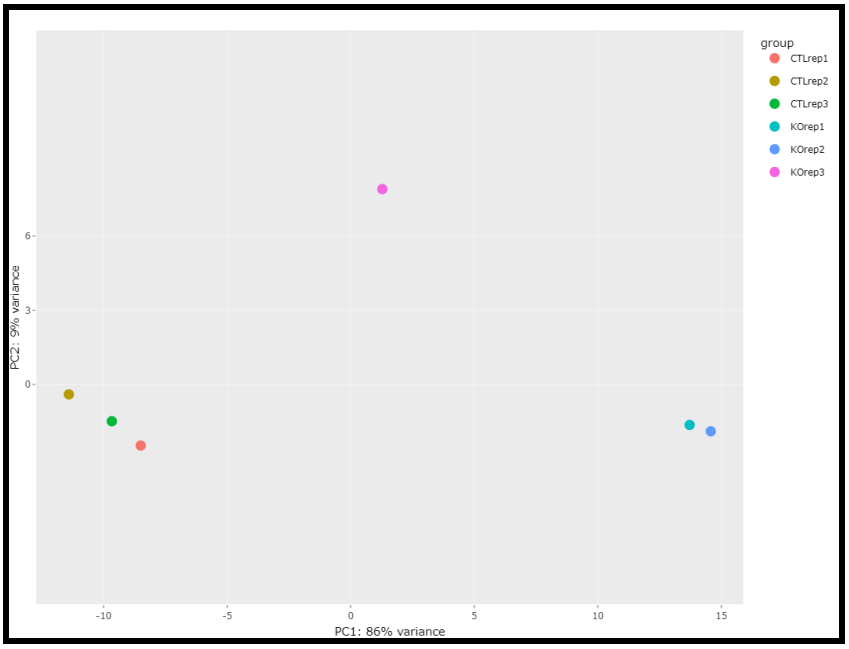
3.1.3 Exclusion of samples based on Quality Control

- Following stringent quality control measures, samples that failed to meet predefined quality thresholds were excluded from downstream analysis to ensure data integrity and reliability.
- Specifically, samples exhibiting persistent issues such as excessive ambiguous bases (N) or the presence of overrepresented sequences were flagged for exclusion.
- This step was crucial to maintain the overall quality of the dataset and to minimize the potential introduction of artifacts or biases that could compromise the accuracy of subsequent analyses.

3.2 After generating your counts matrix, perform a PCA or produce a sample-to-sample distance plot as described in the DESeq2 vignette.

Briefly remark on the plot and what it indicates to you in terms of the experiment.

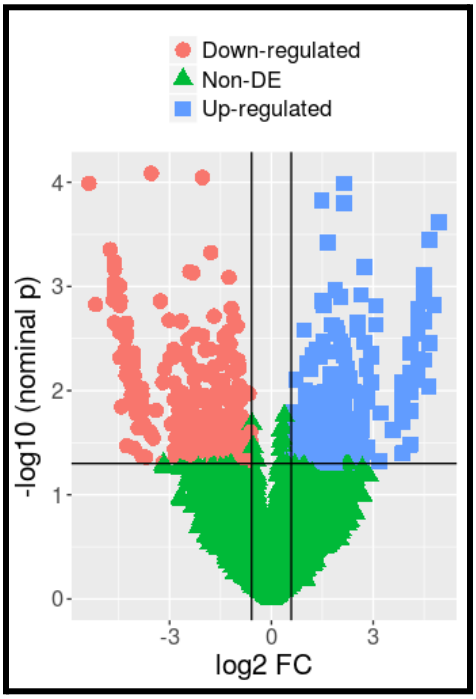
- The control (CTL) and knockout (KO) groups are separated along the PC1 axis, which suggests differences in gene expression patterns between the control and knockout samples. The CTL samples have negative PC1 values, while the KO samples have positive PC1 values.
- The KOrep3 sample has a high PC2 value and a low PC1 value, which separates it from the other knockout samples. This could indicate a unique gene expression pattern for this sample, possibly due to biological or technical factors.
- In the given PCA plot, there is a clear separation between the CTL and KO groups along the PC1 axis, which is the main focus of the analysis. The separation between the groups is not random and is related to the biological condition. Therefore, it does not appear to be a batch effect at first glance.



3.3 After performing DE analysis, choose an appropriate FDR threshold to subset your DE results.

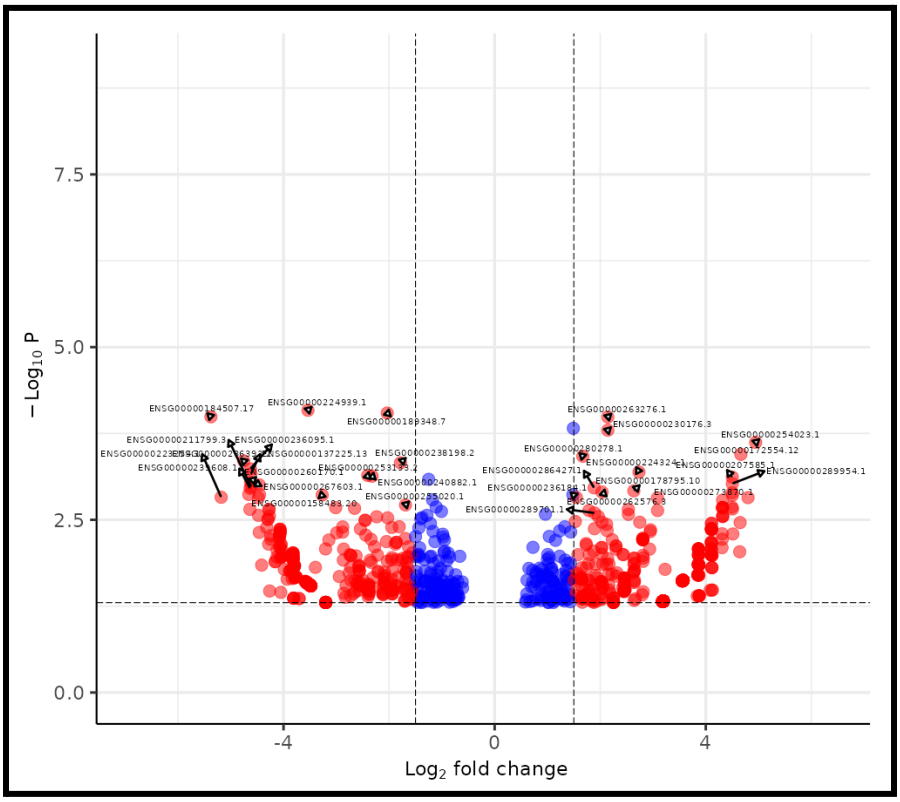
How many genes are significant at your chosen statistical threshold?

3.3.1 Number of significant genes at chosen statistical threshold

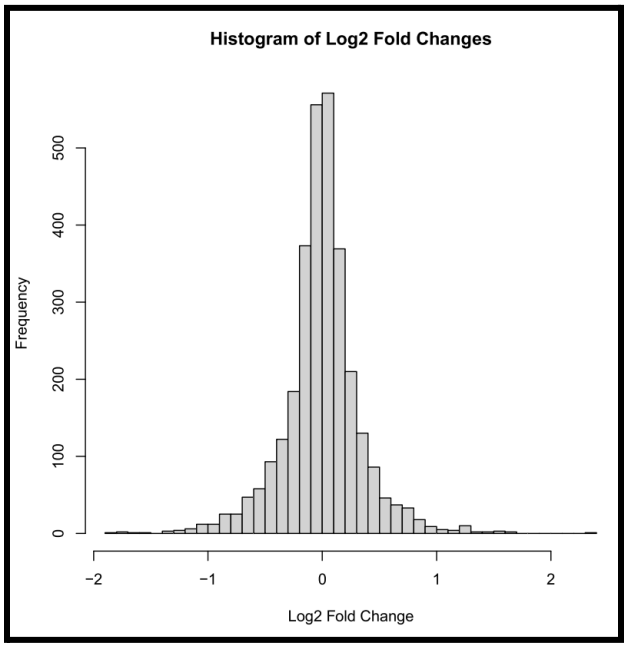


- Down-regulated DEGs: 506

- Non-DEGs: 38265
- Up-regulated DEGs: 739



Volcano plot with labels



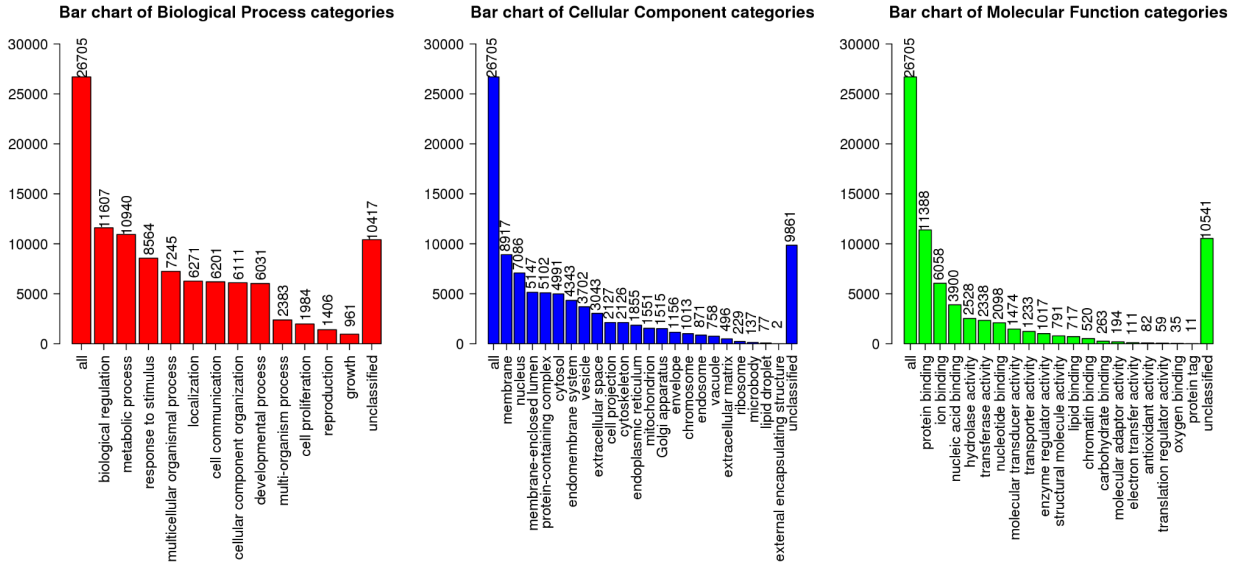
- The shape of the histogram gives us an idea about the distribution of log2 fold changes. In this case, the histogram appears to be roughly symmetrical around 0, suggesting that the log2 fold changes are approximately normally distributed.

- The histogram does not show any bars extending significantly beyond the main body of the distribution, suggesting that there are possibly no extreme outliers in the log2 fold changes.

3.4 After performing FGSEA (GSEA) using a ranked list of all genes in the experiment and performing gene set enrichment using your list of statistically significant DE genes, please answer the following questions:
How similar are the results from these two analyses? Are there any notable differences?
Do you expect there to be any differences? If so, why?
What do the results imply about potential biological functions of the factor of interest?

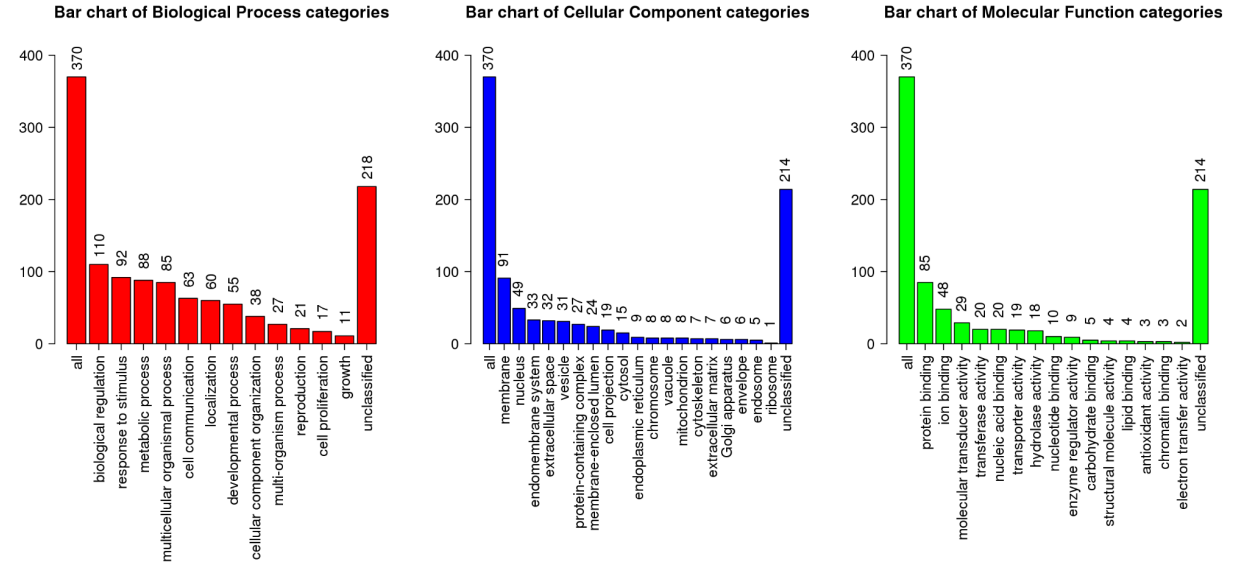
3.4.1 Similarity and differences between results

Gene Set	Description	Size	Expect	Ratio	P Value	↑ FDR
hsa01100	Metabolic pathways	1305	1292.6	1.0080	0.00014519	0.047332
hsa04151	PI3K-Akt signaling pathway	354	350.63	1.0096	0.031300	1
hsa05200	Pathways in cancer	526	521.00	1.0077	0.035128	1
hsa05165	Human papillomavirus infection	339	335.78	1.0096	0.036375	1
hsa05206	MicroRNAs in cancer	299	296.16	1.0096	0.054220	1
hsa04010	MAPK signaling pathway	295	292.20	1.0096	0.056421	1
hsa04060	Cytokine-cytokine receptor interaction	294	291.21	1.0096	0.056985	1
hsa04144	Endocytosis	244	241.68	1.0096	0.093528	1
hsa04014	Ras signaling pathway	232	229.79	1.0096	0.10528	1
hsa04714	Thermogenesis	229	226.82	1.0096	0.10844	1



The above are the GSEA results with all the genes. From the biological processes perspective, we can infer that the category “Metabolic pathways” stands out followed by Pl3K-Akt signaling pathways.

Gene Set	Description	Size	Expect	Ratio	P Value	↑ FDR
hsa04740	Olfactory transduction	448	4.3786	2.7406	0.0012093	0.39424
hsa04514	Cell adhesion molecules (CAMs)	144	1.4074	3.5526	0.013068	1
hsa00480	Glutathione metabolism	56	0.54733	5.4812	0.017151	1
hsa00590	Arachidonic acid metabolism	63	0.61575	4.8721	0.023411	1
hsa04918	Thyroid hormone synthesis	74	0.72326	4.1479	0.035433	1
hsa04742	Taste transduction	83	0.81122	3.6981	0.047222	1
hsa04080	Neuroactive ligand-receptor interaction	277	2.7073	2.2162	0.052932	1
hsa05320	Autoimmune thyroid disease	53	0.51801	3.8609	0.094434	1
hsa04927	Cortisol synthesis and secretion	64	0.62552	3.1973	0.12919	1
hsa05160	Hepatitis C	131	1.2804	2.3431	0.13609	1



While including only the statistically significant DE genes as the GSEA input, we observe two new categories that appear on the top, olfactory transduction and CAMs. Additionally, we gain more insights about the metabolic pathways, i.e. glutathione and arachidonic acid metabolism pathways.

The GSEA results suggest that the data has to do primarily with metabolism and olfactory pathways.

3.4.2 Expectations of differences

- The advantage of the first approach is that it takes into account the expression levels of all genes in the experiment, rather than just the DE genes.
    - This can help to identify subtle changes in gene expression that may not be statistically significant on their own, but that may still be biologically relevant.
    - Additionally, the first approach can help to identify pathways or processes that are affected by changes in gene expression that are not necessarily statistically significant, but that may still be important for understanding the underlying biology.
- The advantage of the second approach is that it focuses on the genes that are most likely to be biologically relevant, based on their statistically significant changes in expression.
    - This can help to reduce the noise and complexity of the data, and can make it easier to identify pathways or processes that are significantly affected by the experimental condition.

REFERENCES

[1] Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[2] Philip Ewels, Måns Magnusson, Sverker Lundin, Max Käller, MultiQC: summarize analysis results for multiple tools and samples in a single report, Bioinformatics, Volume 32, Issue 19, October 2016, Pages 3047–3048, https://doi.org/10.1093/bioinformatics/btw354

[3] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114-20. doi: 10.1093/bioinformatics/btu170. Epub 2014 Apr 1. PMID: 24695404; PMCID: PMC4103590.

[4] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15-21. doi: 10.1093/bioinformatics/bts635. Epub 2012 Oct 25. PMID: 23104886; PMCID: PMC3530905.

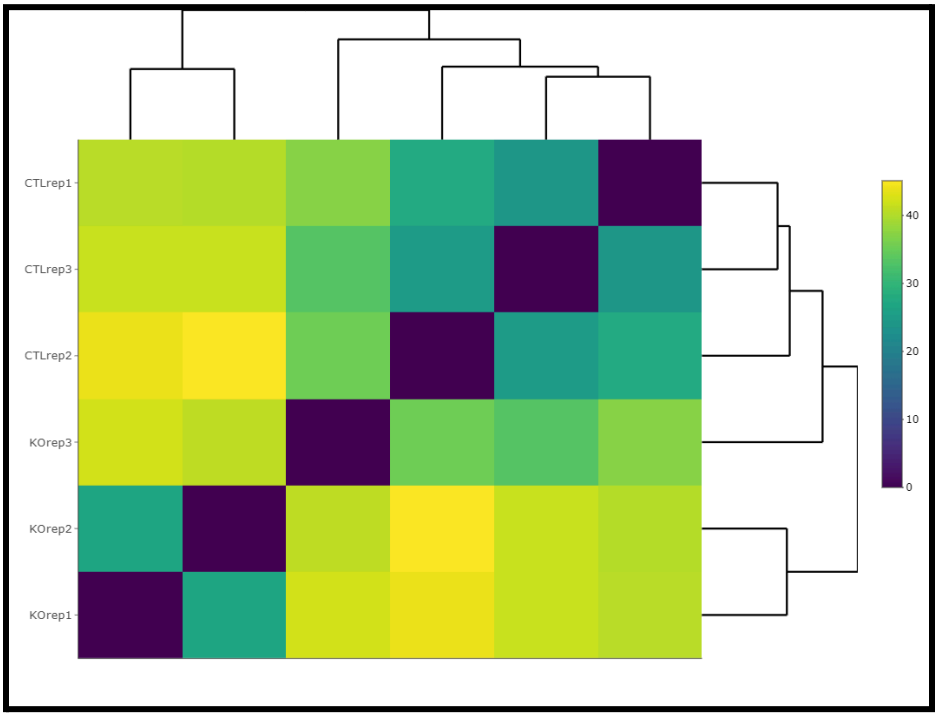
[5] Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, Heng Li, Twelve years of SAMtools and BCFtools, GigaScience, Volume 10, Issue 2, February 2021, giab008, https://doi.org/10.1093/gigascience/giab008

[6] Zhu Q, Fisher SA, Shallcross J, Kim J. VERSE: a versatile and efficient RNA-Seq read counting tool. bioRxiv; 2016. DOI: 10.1101/053306.

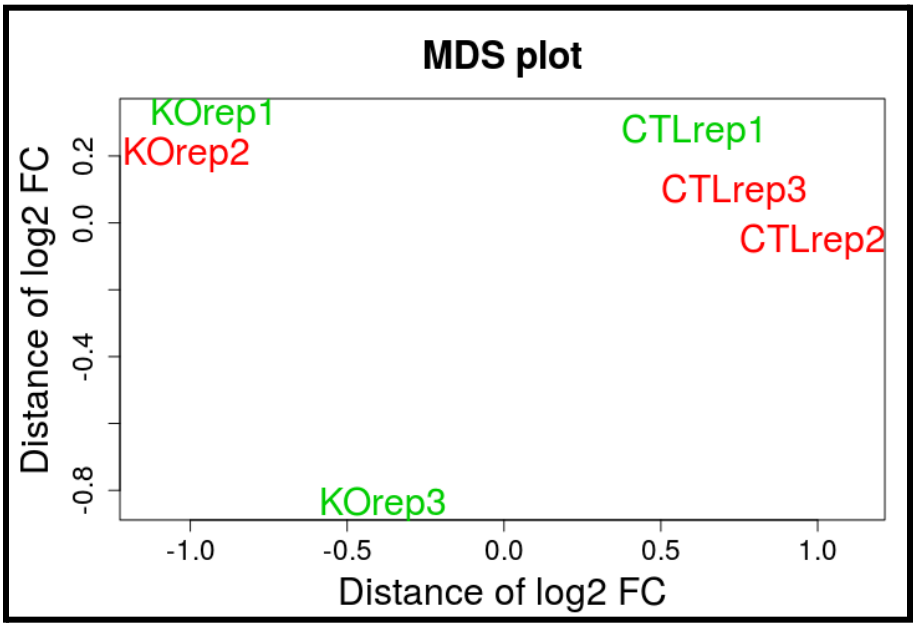
[7] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139-40. doi: 10.1093/bioinformatics/btp616. Epub 2009 Nov 11. PMID: 19910308; PMCID: PMC2796818.

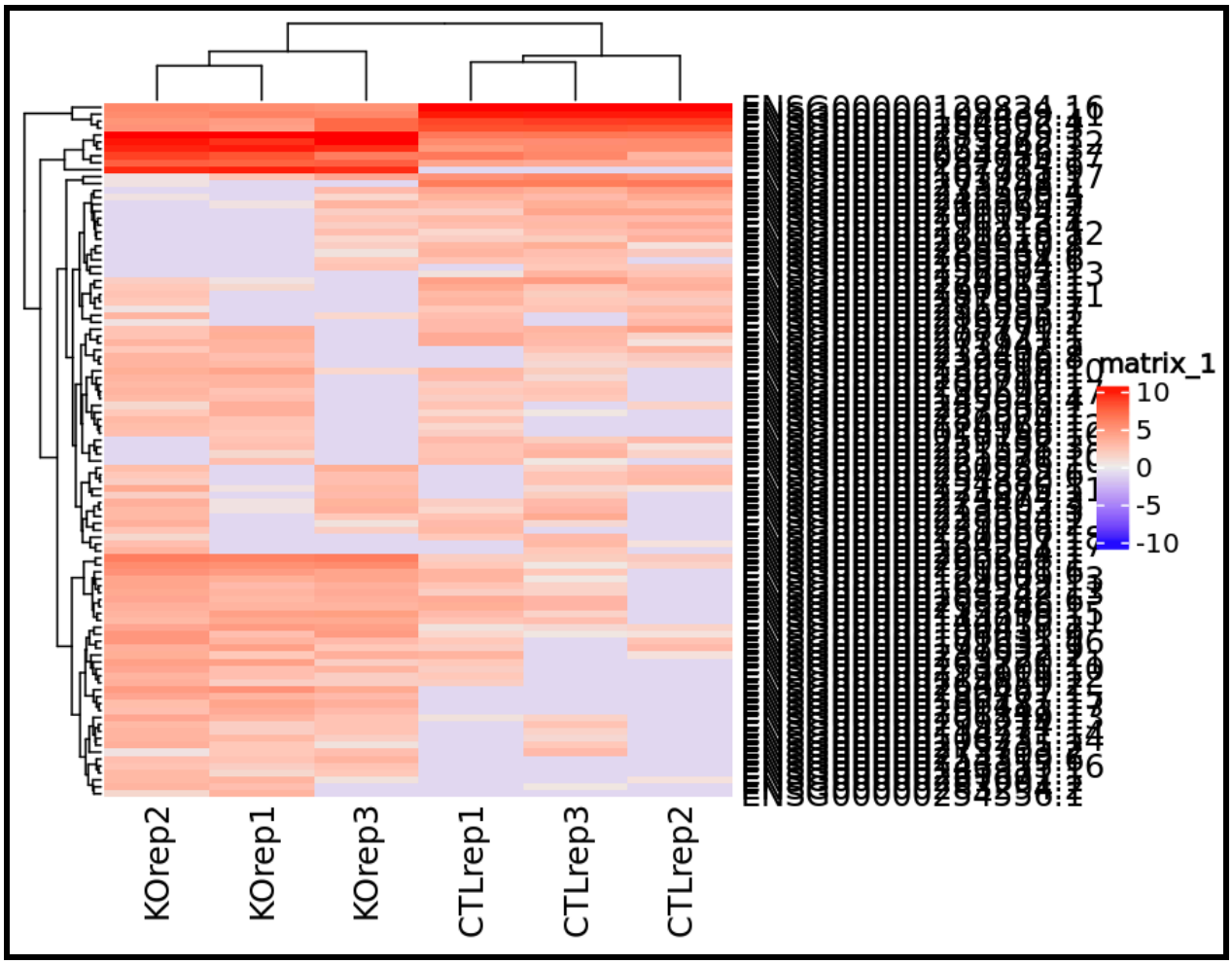
[8] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. doi: 10.1186/s13059-014-0550-8. PMID: 25516281; PMCID: PMC4302049.

SUPPLEMENTARY PLOTS



Sample heatmap





Heatmap

(since the labels are clustered, the results have been provided in table format)