



Mini Project Bioinformatics Lab:

***Differential Gene Expression Analysis of
Accession Id:GSE207456 Dataset***

Submitted By:

Abhinay Kumar Pandey

21BT3EP11



Table of Contents:

Page Number	Contents
3-4	Introduction
4	Software and Packages
4-5	Information on Plots used for DGE Analysis
5	About the dataset used(GSE207456)
5-6	Information on Plots Used for DGE Analysis
6-14	Plots and Inferences
15	Results
16	Conclusions
17	References

Introduction:

Differential gene expression (DGE) analysis is like a gene detective, widely used in research. It is used to find variation in gene expression across two or more than two samples. It involves taking raw count data (counts when you align to the genome and count how many reads (generally discounting multi mappers) align to each gene), normalizing it and performing statistical experiments to discover quantitative changes in expression levels between experimental groups. For example, we use statistical testing to decide whether, for a given gene, an observed difference in read counts is significant that is, whether it is greater than what would be expected just due to natural random variation.

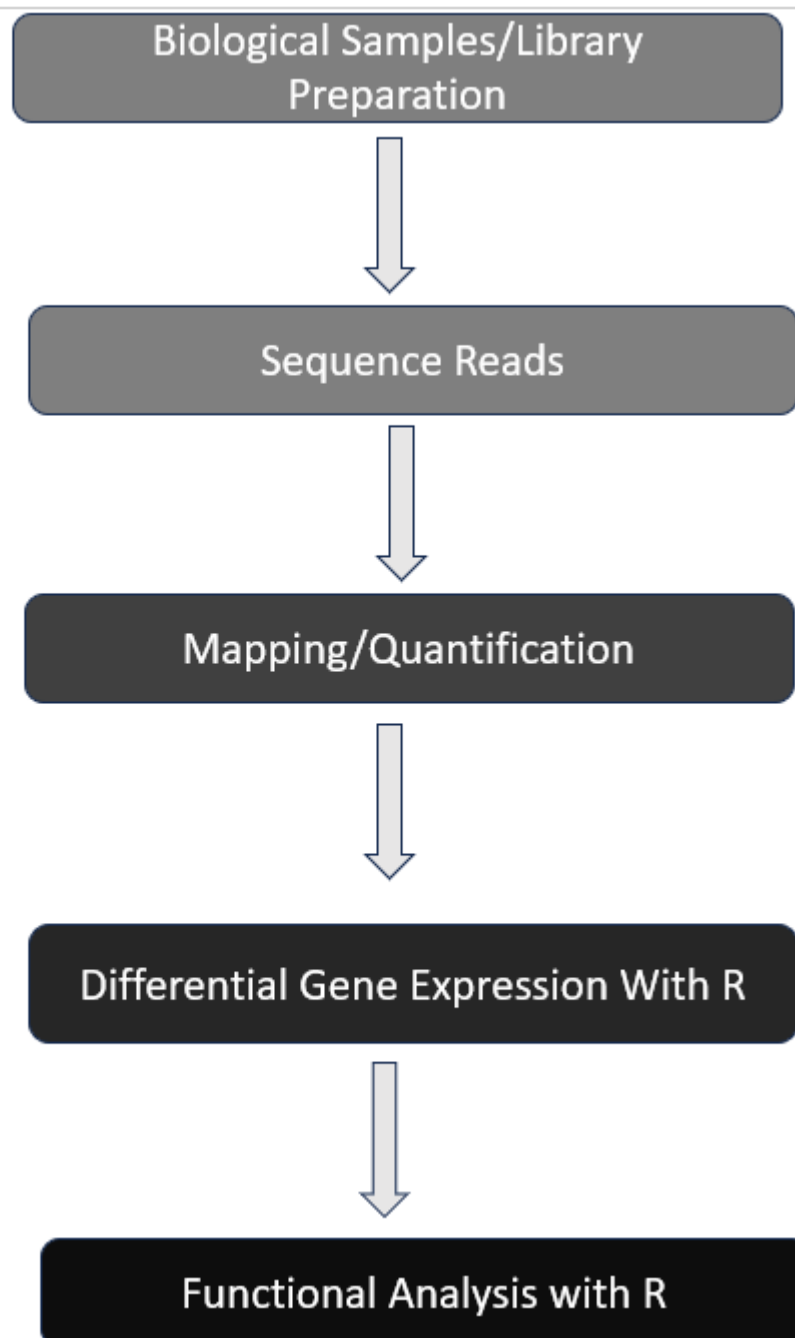



Figure 1



The primary goal of DGE Analysis is to find the genes expressed at different levels between different samples, find if the difference is significant enough and list these genes to perform their functional analysis. DESeq2 performs for each gene a hypothesis test to see whether evidence is sufficient to decide against the null hypothesis that there is no effect of the treatment on the gene and that the observed difference between treatment and control was merely caused by experimental variability (i.e., the type of variability that you can just as well expect between different samples in the same treatment group). These genes further can give an insight into the biological processes affecting/affected by the disease sample. Figure 1 has a flowchart of the steps from preparation of the samples in the laboratory leading up to DGE Analysis and finally performing functional analysis.

• **Softwares and Packages Used:**

R- Programming language for statistical computing and graphics

BiocManager: To install the packages required in the DGE analysis.

DESeq2- Package for differential analysis of count data

gplots and ggplot: To make plots for DGE analysis visualisation.

• **Information on Plots used for DGE Analysis**

→ **MA-plot:** The MA-plot shows the log₂ fold changes from the treatment over the mean of normalized counts, i.e. the average of counts normalized by size factor. The DESeq2 package incorporates a prior on log₂ fold changes, resulting in moderated estimates from genes with low counts and highly variable counts, as can be seen by the narrowing of spread of points on the left side of the plot.

→ **Heatmap:** The normalized values of *all* the significant genes are extracted and a heatmap of their expressions across the samples is created. The heatmap has a color gradient throughout according to the specified scale. This helps in a clear visualization of the change in expression levels, both across samples as well as genes.

→ **Volcano Plot:** Heatmaps are great to look at the expression levels of a fairly large number of genes, but for more of a global view we can use the volcano plot. Here, the log transformed adjusted p-values are plotted on the y-axis and log₂ fold change values on the x-axis.



→ Histogram of p-values:

This histogram is to be created *before* performing multiple hypothesis test correction, false discovery rate control, or any other means of interpreting many p-values. This graph lets you get an immediate sense of how your test behaved across all your hypotheses, and immediately diagnose some potential problems. histogram by looking at how tall the peak on the left is: the taller the peak, the more p-values are close to 0 and therefore significant. Similarly, the “depth” of the histogram on the right side shows how many of your p-values are null.

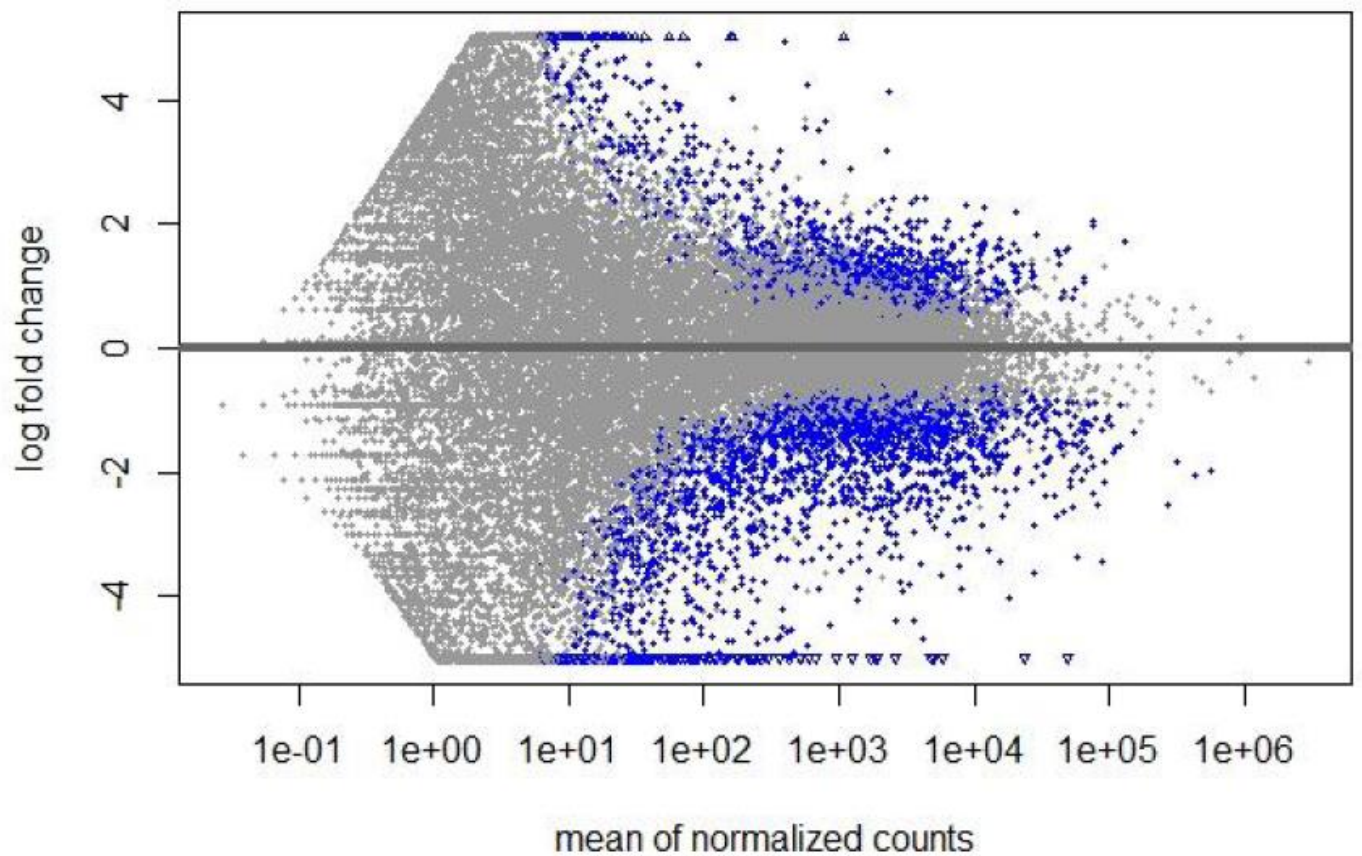
● About the dataset GSE207456

The dataset GSE207456 was taken from the NCBI GEO- a public functional genomics data repository. This dataset’s sequencing method is **Expression profiling by high throughput sequencing**. The dataset is titled ‘Human pluripotent stem cell-derived macrophages host *Mycobacterium abscessus* infection’. The organism utilized for the study is *Homo sapiens*. A reliable model for *M. abscessus* infection using human pluripotent stem cell-derived macrophages (hPSC-macrophages) has been established. Electron microscopy demonstrated that *M. abscessus* was present in the vacuoles of hPSC-macrophages. RNA-sequencing analysis revealed a time dependent host cell response to *M. abscessus*, with differing gene and protein expression patterns observed at 3-hours, 24-hours and 48-hours post-infection. The study describes the first hPSC-based model for *M. abscessus* infection, which represents a novel platform for studying *M. abscessus*-host interaction and an accessible tool for drug discovery. This dataset has been used for primarily two reasons: 1. I have an immense interest in Antibiotic resistance and tuberculosis is one of the diseases whose prevention and cure are deeply affected by the same. 2. There are few available treatments and the search for effective antibiotics against *M. abscessus* has been hindered by the lack of a tractable in vitro intracellular model of infection.

- **Plots and inferences**

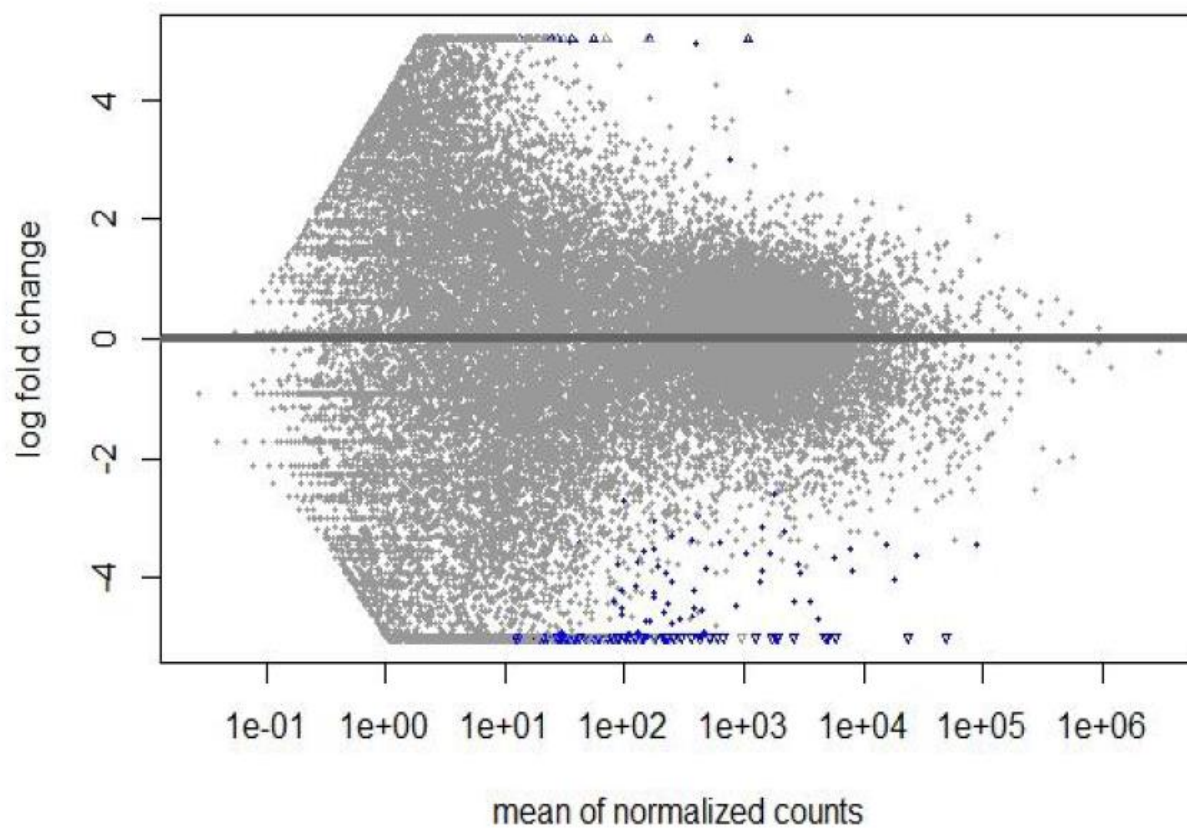
- **MA-plot:**

1. MA Plot when FDR cutoff = 10%



It is more useful to visualize the MA-plot for the shrunken log₂ fold changes, which remove the noise associated with log₂ fold changes from low count genes without requiring arbitrary filtering thresholds.

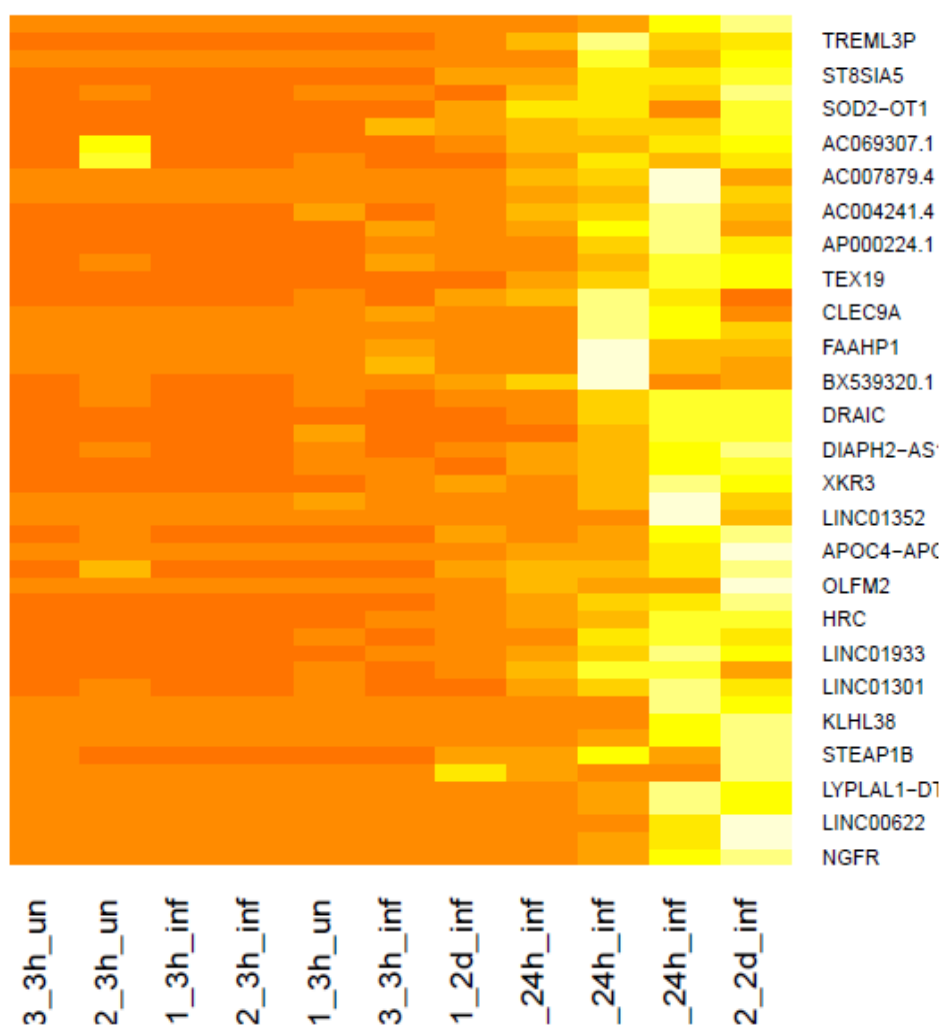
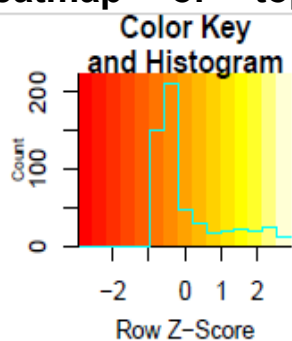
2. MA Plot when LFC threshold is 10%



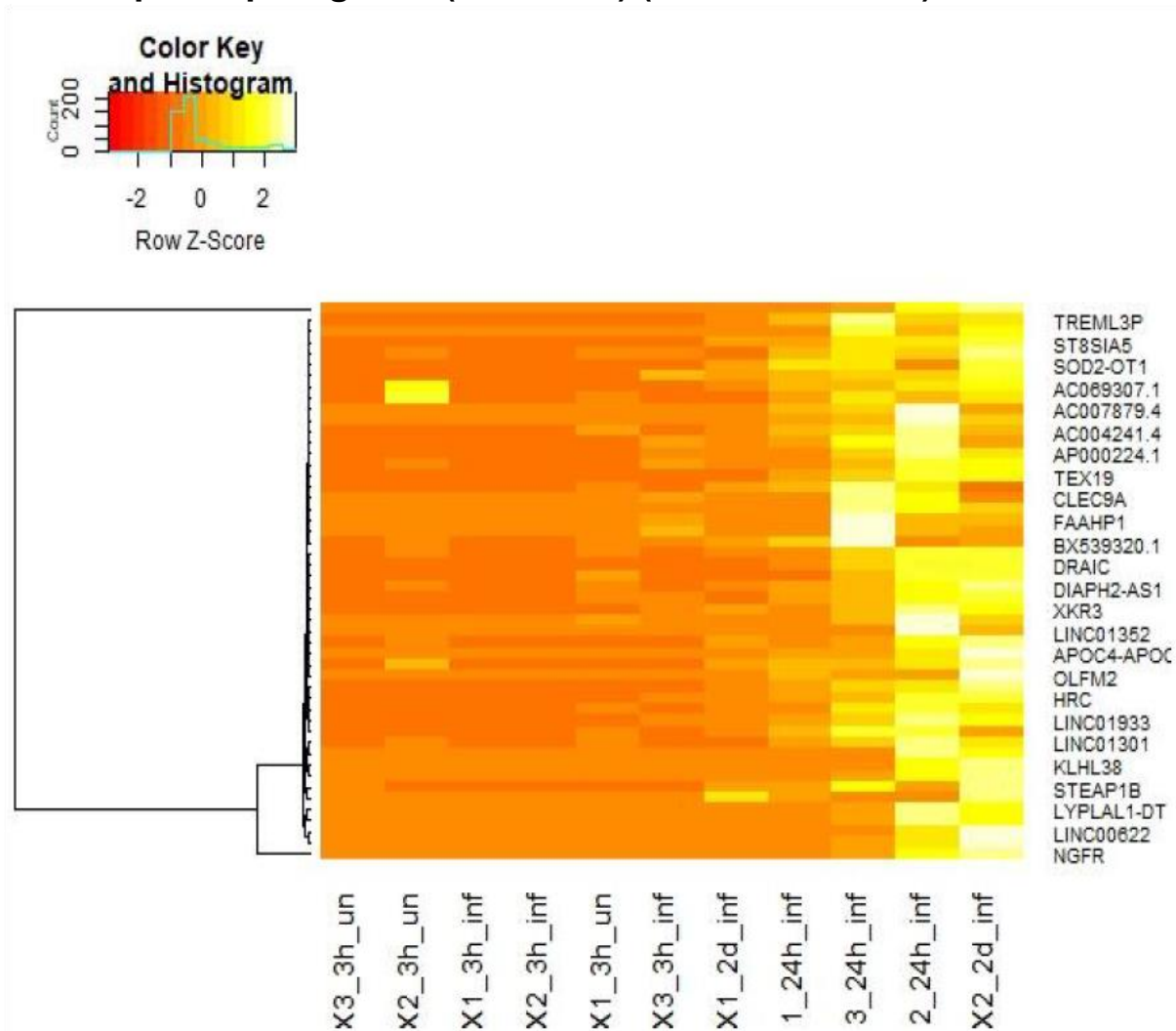
As mentioned previously, this MA Plot is easier to interpret and has lesser points in a more streamlined manner due to the LFC threshold being added on to the FDR cutoff.

→ Heatmap:

1. Heatmap of top 50 genes (unclustered) (FDR cutoff 10%)



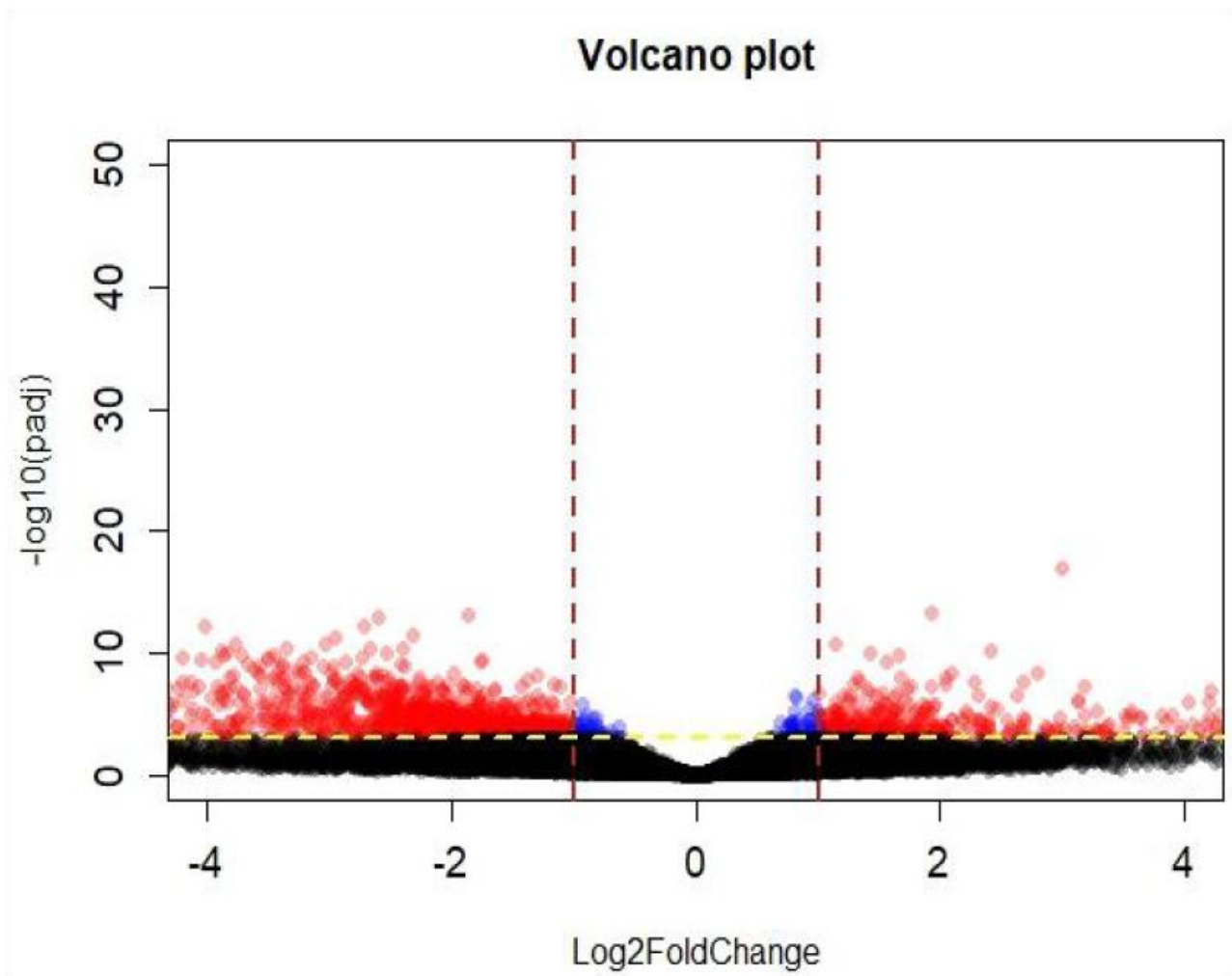
2. Heatmap of top 50 genes (clustered) (FDR cutoff 10%)



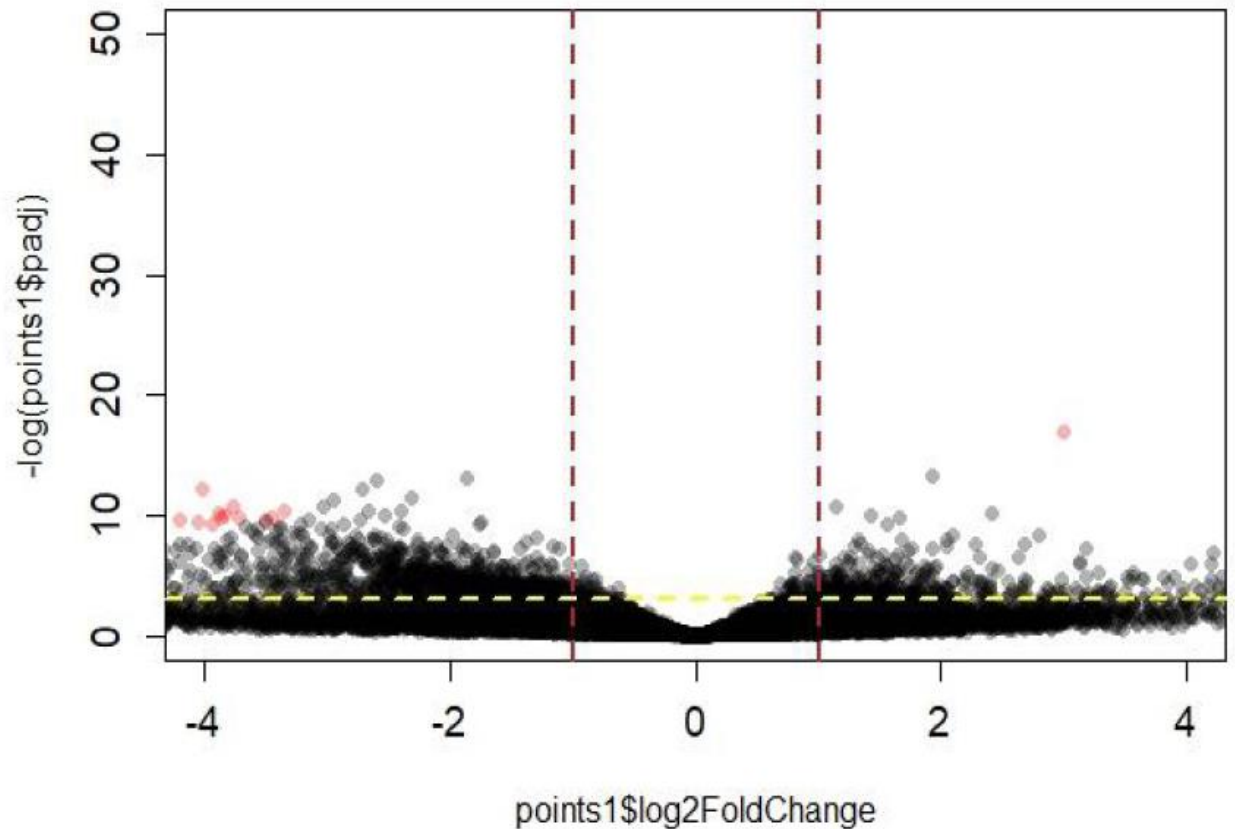
Since the set of genes for LFC + FDR cutoff = 10% is much lesser than those for FDR Cutoff = 10% and are mostly coinciding with the top genes of FDR cutoff = 10%, a heatmap of these genes has not been created as it could lead to redundancy in observations. The heat maps also show a nearly similar count data for most genes through 3 hours samples and some 24 hours samples of both infected and uninfected strains. The highest change occurs at the intersection of 24 hours infected samples of the second and third replicate and the 2nd day replicate as well. Hence to study the variation patterns as well as perform functional analysis, the data where the colour change in the heatmap is stark should be utilized. This is how the heatmap helps analysis by clear visualization.

→ **Volcano Plot:**

1. Volcano Plot when FDR cutoff = 10%



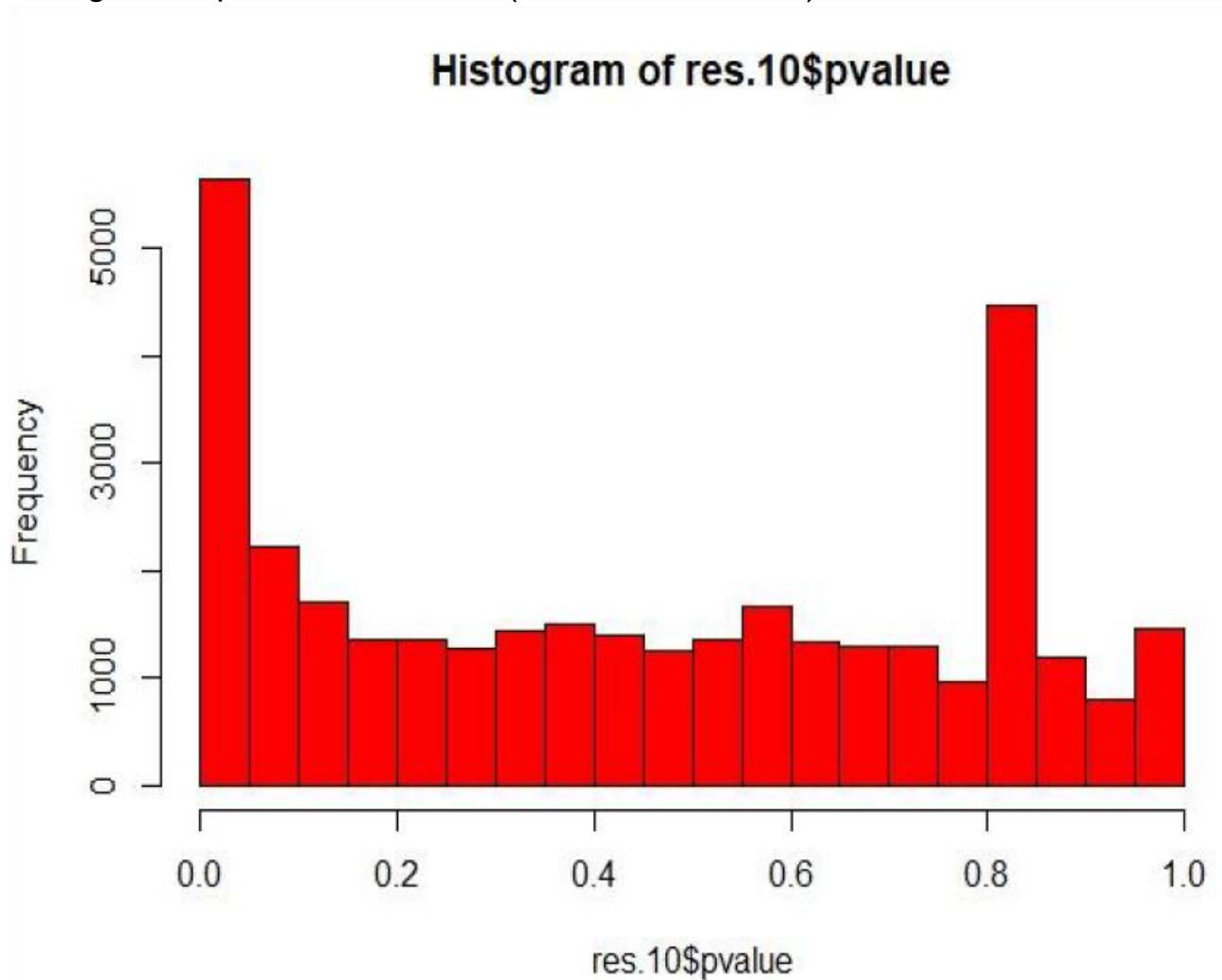
2. Volcano Plot when FDR cutoff and LFC threshold are 10%



The number of points in the volcano plot have clearly reduced in the FDR + LFC = 10% cutoff due to reduction in the number of significantly expressed genes. The other genes have stayed very similar but the significantly upregulated and downregulated genes have reduced drastically. Hence, a volcano plot helps in visualization of the difference in numbers of significantly expressed genes versus all other genes which is not possible in a heatmap.

→ Histogram of p-values:

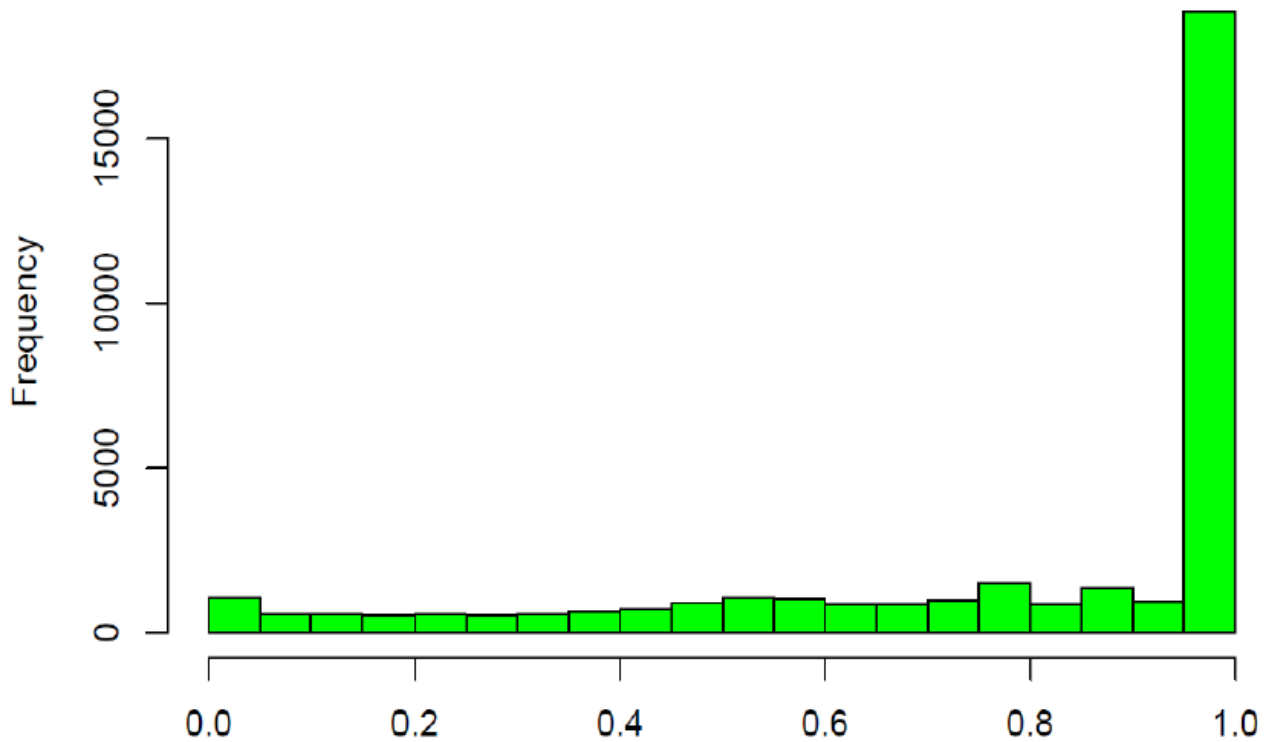
1. Histogram of p-values of res.10 (FDR Cutoff = 10%)



This is a **bimodal** p-values histogram distribution. It primarily states that the peak at the beginning (p value nearing 0) is accompanied by one at the end of the histogram (p value nearing 1). The peak close to 0 is where the alternative hypotheses are along with some potential false positives. The flat distribution along the bottom is all the null p-values, which are uniformly distributed between 0 and 1.

2. Histogram of p-values of resLFC1 (FDR and LFC Cutoff = 10%)

Histogram of resLFC1\$pvalue



This histogram clearly shows that LFC and FDR (multiple hypotheses corrections) cutoffs have been applied to the dataset as the highest frequency of p values is near the latest mark of p=1.



• Results

→ summary(res.10)

Out of 35208 with nonzero total read count
adjusted p-value < 0.1
LFC > 0 (up) : 1191, 3.4%
LFC < 0 (down) : 2320, 6.6%
outliers [1] : 261, 0.74%
low counts [2] : 15116, 43%

Total differentially expressed genes: 3511

→ summary(resLFC1)

Out of 35208 with nonzero total read count
adjusted p-value < 0.1
LFC > 1.00 (up) : 11, 0.031%
LFC < -1.00 (down) : 159, 0.45%
outliers [1] : 261, 0.74%
low counts [2] : 13152, 37%

Total differentially expressed genes: 170

The lists of these differentially expressed genes, sorted in the descending order of their LFC values is attached along with this report.



Conclusions from Differential Gene Expression Analysis:

In our investigation, we delved into the differential gene expression within our dataset, uncovering intriguing insights. Initially, we identified a substantial pool of 3511 genes that exhibited differential expression among a total of 35208 genes with non-zero read counts, under the condition of a 10% False Discovery Rate (FDR) cutoff. However, to refine our analysis and focus on more pronounced changes, we imposed an additional condition: a Log Fold Change (LFC) threshold of 1. This stringent criterion drastically reduced the number of differentially expressed genes to 170, highlighting the effectiveness of incorporating LFC thresholds in gene expression studies.

Our findings underscore the challenge of sifting through vast lists of significant genes to extract meaningful biological insights. We recognize the importance of balancing sensitivity and specificity in such analyses. While FDR thresholds alone may not suffice in narrowing down significant genes substantially, our incorporation of LFC thresholds provided a valuable means of enhancing stringency.

Moreover, we utilized various graphical representations, such as MA Plots, Volcano Plots, Heatmaps, and Histograms of p-values, to glean crucial information from our data. These visualizations offer diverse perspectives on the significantly expressed genes, facilitating not only the identification of potential targets but also aiding in subsequent functional analyses.

Published Conclusions from the Dataset:

The culmination of our dataset analysis yielded profound insights into the molecular mechanisms underlying heightened antimicrobial activity. Leveraging transcriptomic and metabolomic analyses, we conducted a comparative study between a wild-type strain and a zinc efflux mutant to elucidate the cellular targets of zinc intoxication.

Our investigation illuminated the multifaceted disruptions induced by zinc intoxication within cellular processes. These disruptions, when coupled with frontline antibiotics, exhibited a synergistic effect capable of overcoming antibiotic resistance. Furthermore, our findings suggest that this approach may hold promise in mitigating the emergence of resistance in microbial populations, potentially offering a preemptive strategy against future resistance developments.

References:

1. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/>.
2. McDermaid A, Monier B, Zhao J, Liu B, Ma Q. Interpretation of differential gene expression results of RNA-seq data: review and integration. Brief Bioinform. 2019 Nov 27;20(6):2044-2054. doi: [10.1093/bib/bby067](https://doi.org/10.1093/bib/bby067). PMID: 30099484; PMCID: PMC6954399.
3. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository Nucleic Acids Res. 2002 Jan 1;30(1):207-10
4. <https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/rna-sequencing/performing-a-rna-seq-experiment/data-analysis/differential-gene-expression-analysis/>
5. https://hbctraining.github.io/Training-modules/planning_successful_rnaseq/lessons/sample_level_QC.html
6. https://hbctraining.github.io/Intro-to-R-with-DGE/lessons/B1_DGE_visualizing_results.html
7. <http://varianceexplained.org/statistics/interpreting-pvalue-histogram/>
8. <https://bioc.ism.ac.jp/packages/2.14/bioc/vignettes/DESeq2/inst/doc/beginner.pdf>
9. <http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#ma-plot>