# Bioinformatics Lab

## Shashank Dubey(20BT30027)

### Report-Differential Gene Expression (DGE) Analysis

**Aim:** To perform differential gene expression analysis on a dataset obtained from the GEO database.

**Dataset:** GSE215456

The study aims to investigate the impact of haemoglobin depletion on the outcome of differential gene expression analysis in RNA-seq data from 58 human tuberculosis (TB) patient whole blood samples. The samples were divided into two groups: 29 were subjected to haemoglobin depletion using a hybridisation-based method, and 29 were not depleted and served as controls. The samples were taken from the same patients at the time of TB diagnosis and six months after TB treatment. The study compared the effects of haemoglobin depletion by hybridisation-based removal and bioinformatic removal of reads on differential gene expression analysis results.

## Introduction:

- Differential gene analysis is a process used in molecular biology to identify differentially expressed genes between two or more biological conditions or groups. It involves comparing the gene expression levels of samples from different conditions, such as healthy vs diseased tissue or treated vs untreated cells.

- The analysis typically involves using high-throughput technologies, such as microarrays or RNA sequencing (RNA-seq), to simultaneously measure the expression of thousands of genes. The data is then analysed using statistical methods to identify genes that are significantly differentially expressed between the conditions being compared.

- The goal of differential gene analysis is to identify genes that are involved in the biological processes that distinguish the conditions being studied. These genes can serve as potential biomarkers for disease diagnosis, drug targets for therapy, or provide insights into the mechanisms underlying a particular biological process.

## Methodology:

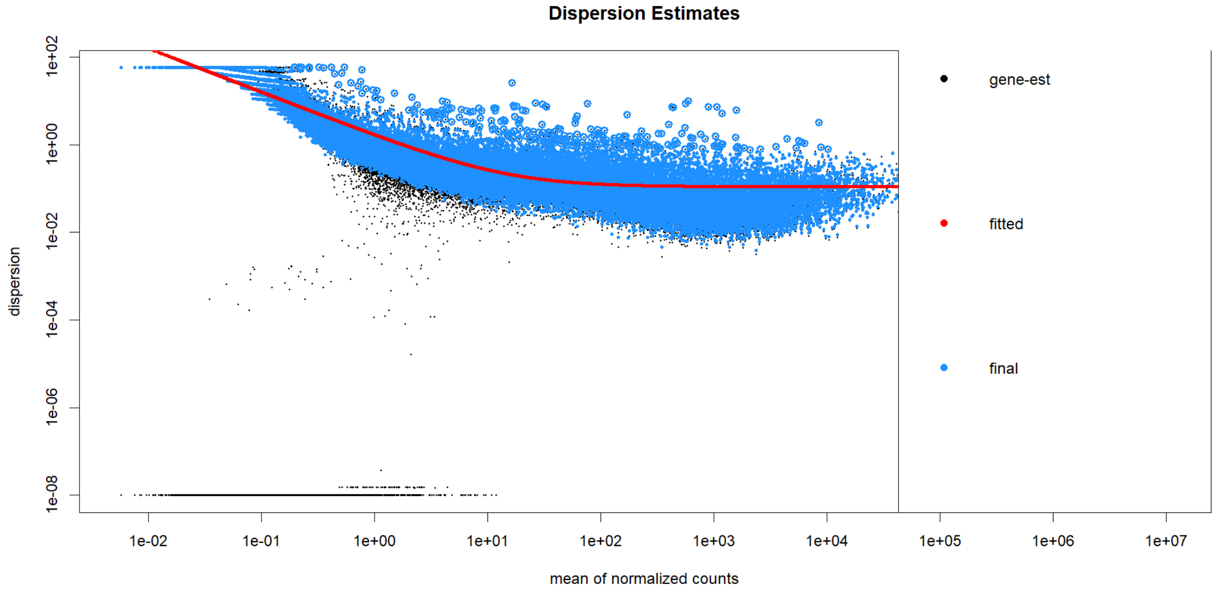
- Search for a dataset from the NCBI GEO datasets section.
- Preprocessing of the data using DE Seq2. DESeq2 performs an internal normalisation where the geometric mean is calculated for each gene across all samples. The counts for a gene in each sample are then divided by this mean. The median of these ratios in a sample is the size factor for that sample.
  the formula for the Geometric mean for a set of n elements is:

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \cdots x_n}$$

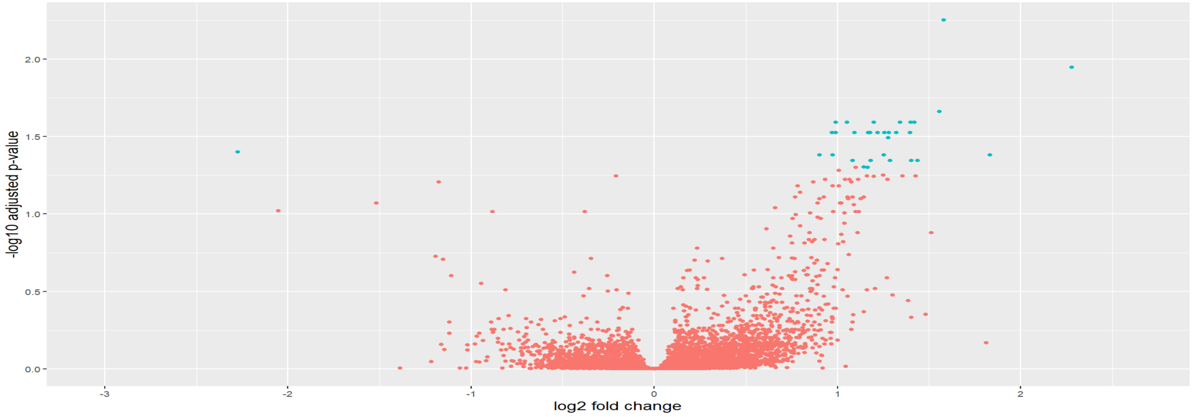
- Performing gene analysis finding out genes showing **FDR cutoff of 10%**.
- Visualisation of results using different plots like MA plot, Heatmap, and Volcano Plot.

## Results:

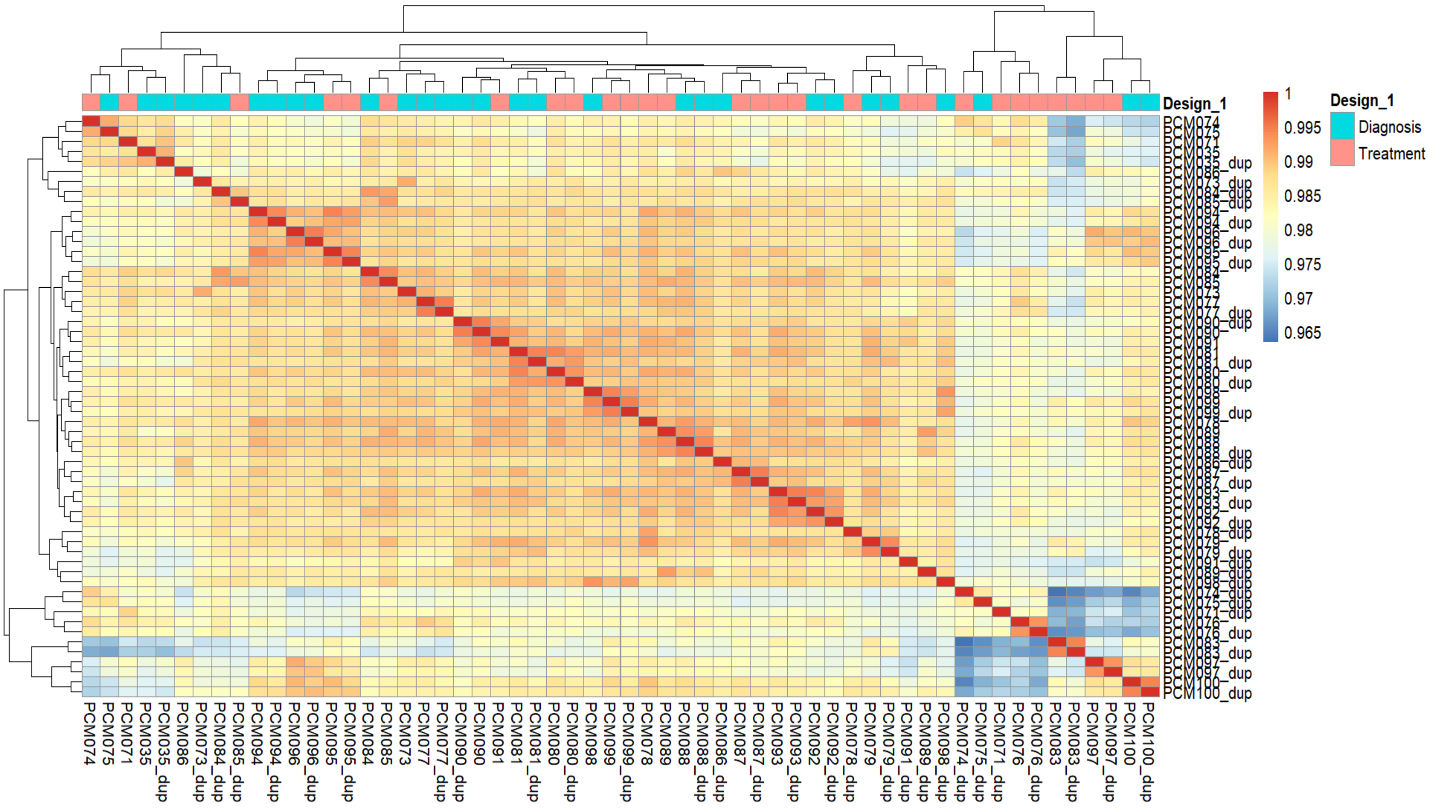
### Dispersion Plot:



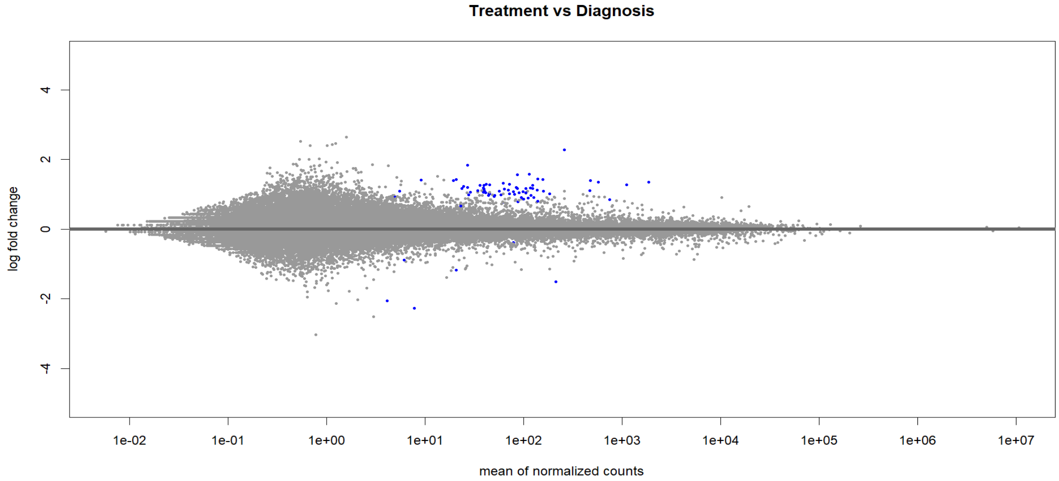
## Volcano Plot:



## Heatmap:



## MA Plot:



## Inference:

To summarise, differential expression (DE) analysis is a commonly utilised method to detect genes expressed differently under varying experimental conditions. Its primary purpose is to identify genes whose expression levels vary significantly between different groups, making it a valuable tool for researchers in various fields of study.

Like using the dataset, we studied differential gene expression analysis performed using RNA-seq data from 58 human tuberculosis (TB) patients or contact whole blood samples-29 globin kit-depleted and 29 matched non-depleted-a subset of which were taken at TB diagnosis and at six months post-TB treatment from the same patient.

The analysis of differential gene expression (DE) is an effective technique for identifying genes that are expressed differently between groups, which can aid in identifying genes associated with disease pathology.