Differential Gene Expression Analysis

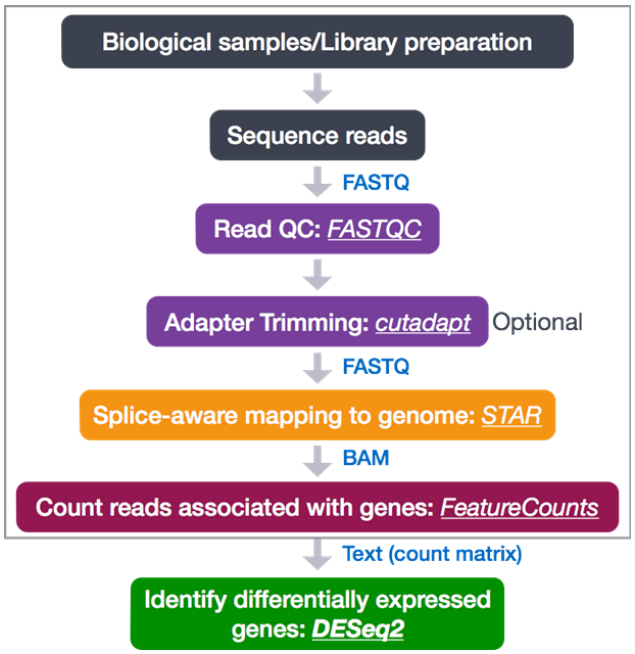
(Gonnabhakthula Akshith - 20BT30009)

INTRODUCTION:

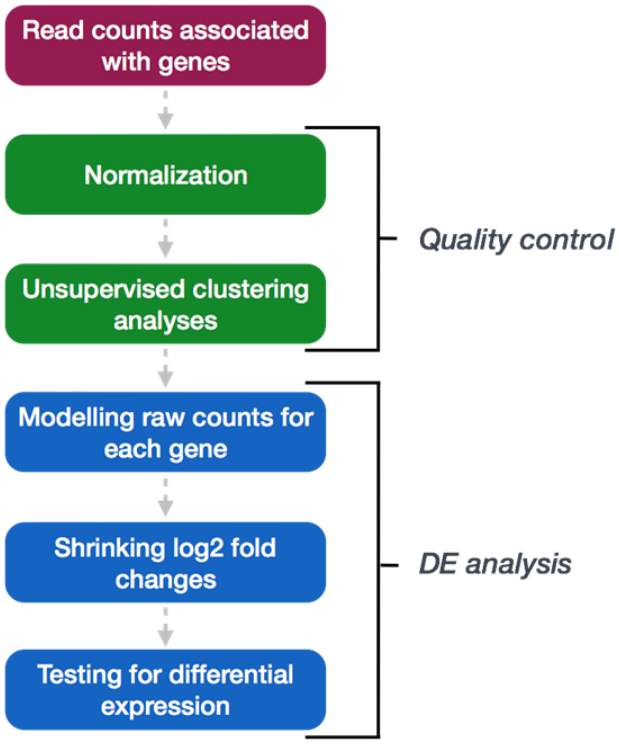
What is Differential Gene Expression Analysis?

The goal of RNA-seq is often to perform differential expression testing to determine which genes are expressed at different levels between conditions. These genes can offer biological insight into the processes affected by the condition(s) of interest.

To determine the expression levels of genes, our RNA-seq workflow followed the steps detailed in the image below. All steps were performed on the command line (Linux/Unix) through the generation of the read counts per gene. The differential expression analysis and any downstream functional analysis are generally performed in R using R packages specifically designed for the complex statistical analyses required to determine whether genes are differentially expressed.



To model counts appropriately when performing a differential expression analysis, there are a number of software packages that have been developed for differential expression analysis of RNA-seq data. Even as new methods are continuously being developed a few tools are generally recommended as best practice, e.g. **DESeq2**.



SOFTWARE AND PACKAGES REQUIRED:

1.RStudio: Provides an interactive environment to code in R.
2.BiocManager: To install the packages required in the DGE analysis.
3.GEOquery: To obtain the required datasets from GEO NCBI.
4.DESeq2: To perform differential gene expression analysis.
5.ggplot: To make plots for DGE analysis visualisation.

DATASET:

GEO Accession: GSE229611

Title: Exhaustion-associated cholesterol deficiency dampens the cytotoxic arm of antitumor immunity

Organism: Homo sapiens; Mus musculus

Experiment Type: Expression profiling by high throughput sequencing

- The concept of targeting cholesterol metabolism to treat cancer has been widely tested in clinics but the benefits are modest, calling for a complete understanding of cholesterol metabolism of intratumoral cells.
- Low cholesterol levels inhibit T-cell proliferation and cause autophagy mediated apoptosis, particularly for cytotoxic T cells. In the tumor microenvironment, oxysterols mediate reciprocal
- alterations of the LXR and SREBP2 pathways to cause cholesterol deficiency of T Cells, subsequently leading to aberrant metabolic and signaling pathways that drive T cell exhaustion/dysfunction.
- LXR depletion in CAR-T cells led to improved antitumor function against solid tumor.

Theory:

Normalization:

DESeq2 is a tool that helps researchers analyze RNA-seq data by detecting differentially expressed genes between two or more samples. Normalization of raw counts is a crucial step in this analysis pipeline. DESeq2 calculates normalization factors for each sample using the median of ratios method, which accounts for differences in sequencing depth and RNA composition between samples. This method creates a pseudo-reference sample for each gene in the dataset, calculating the ratio of the sample expression level to the pseudo-reference expression level for each gene in a given sample. The median of ratios for each sample is then used as the normalization factor. This ensures that the total counts for each sample are adjusted to be equal, allowing for meaningful comparisons of gene expression levels between samples. DESeq2's normalization method is robust to large numbers of differentially expressed genes and enables accurate analysis of RNA-seq data.

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

Shrinkage Estimation of fold change and dispersion by EmpiricalBayes:

Yes, that's correct. DESeq2 uses shrinkage estimation to improve the accuracy and stability of dispersion and fold change estimates, especially when working with datasets that have a small number of replicates. In DESeq2, the dispersion value for each gene is estimated using a model fit procedure that accounts for both the mean and variance of gene expression across samples. Shrinkage estimation is then applied to the dispersion estimates to produce more accurate and stable estimates.

Shrinkage estimation is a statistical technique that borrows information from other genes in the dataset to produce better estimates for a given gene. This is especially useful when working with small datasets, where the variance estimates for individual genes can be unreliable due to limited sample sizes. By using information from other genes in the dataset, shrinkage estimation allows DESeq2 to produce more reliable and robust dispersion and fold change estimates, even when working with small numbers of replicates. Overall, the use of shrinkage estimation is an important aspect of the DESeq2 analysis pipeline, as it helps to improve the accuracy and stability of gene expression estimates, especially in datasets with limited sample sizes.

Model Fitting:

After dispersion estimation, DESeq2 fits a negative binomial generalized linear model for each gene. The count data y_ij for gene i in sample j is modeled by:

$$y_{ij} \sim NB(mu_{ij}, size_i)$$

where μ_ij s the expected value of y_ij and size_i is the size factor for gene i. The size factor is included to account for differences in sequencing depth across samples.

The log2 fold change between two conditions (e.g.treatment vs control) is estimated as the difference in the expected log2 counts, i.e.

$$log_2(FC_i) = log_2(mu_i, treatment) - log_2(mu_i, control)$$

In summary, DESeq2 performs a negative binomial generalized linear model to estimate log2 fold changes between conditions and perform hypothesis testing, while adjusting for sequencing depth and correcting for multiple testing.

Hypothesis Testing:

Hypothesis testing is done to determine which genes are significantly differentially expressed between groups. DESeq2 uses a Wald test to calculate a p-value for each gene, which is adjusted for multiple testing using the Benjamini-Hochberg procedure to control the false discovery rate.

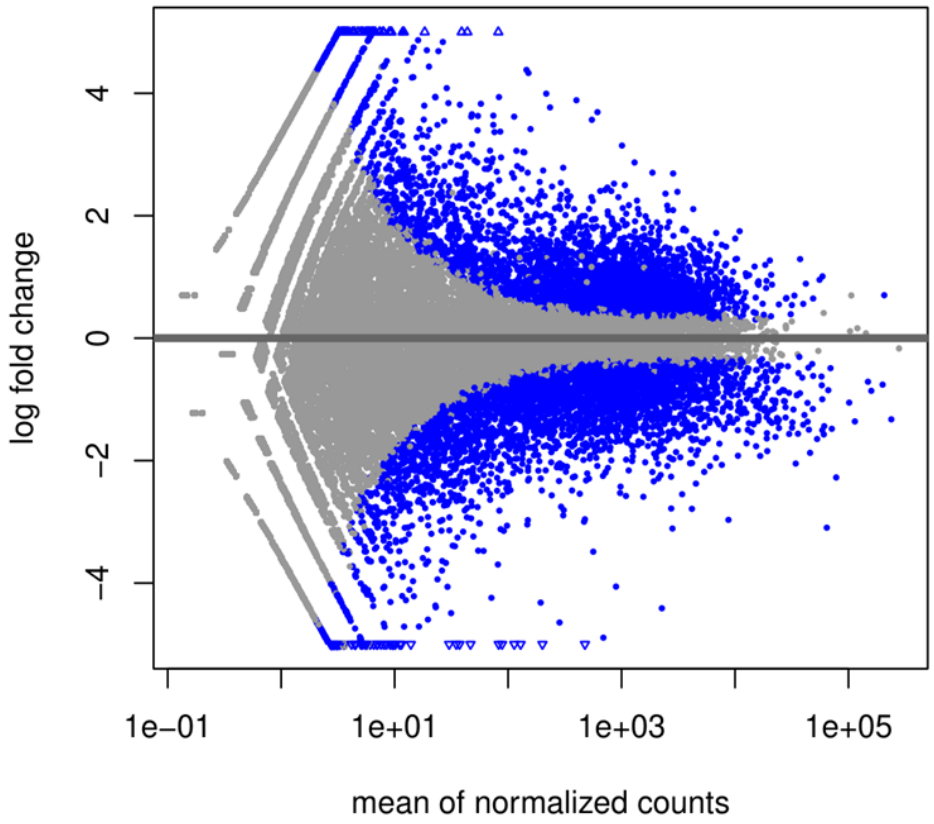
CODE:

The code for DE Analysis of GSE229611 as 20BT30009.R in code folder.

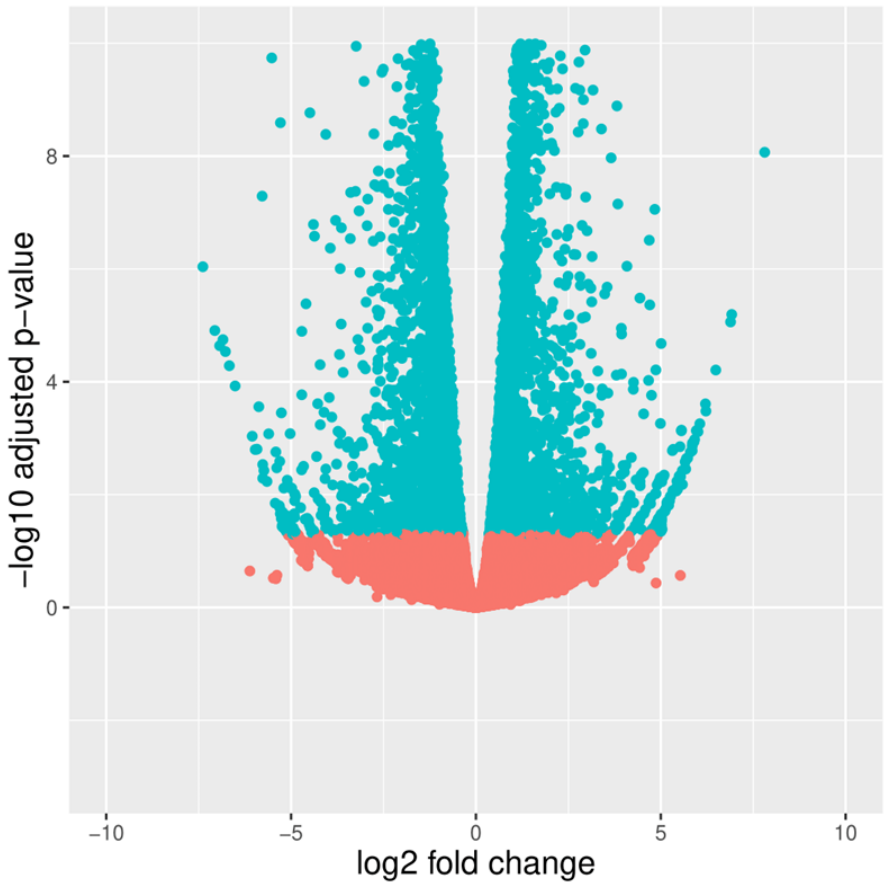
RESULTS:

Results are uploaded as output folder

PLOTS:



MA PLOT



VOLCANO PLOT

Inferences:

An MA plot is created by plotting the log2 fold-change values (M) of gene expression levels between two samples against their average expression levels (A).-The fold-change values represent the difference in gene expression levels between the two samples, while the average expression levels represent the overall gene expression level across both samples.-MA plot is useful for identifying differentially expressed genes.It can help in identifying false positives or technical artifacts in the data.

TheVolcanoplot gets its name from the shape of the plot, which looks like a volcano with the most significant differentially expressed genes or proteins represented as points located at the top of the volcano. The volcano plot allows for a quick visualization of the data and helps in identifying genes or proteins that are differentially expressed with statistical significance. The plot is usually divided into two regions: the up-regulated genes or proteins located on the right side of the plot, and the down-regulated genes or proteins located on the left side of the plot. The

most statistically significant genes or proteins are located at the top of the plot, while less significant genes or proteins are located at the bottom of the plot. The plot is further enhanced by adding colour coding to represent the magnitude of the fold-change, and highlighting specific genes of interest. Here, False Discovery Rate cut-off of 10% for the α and LFC threshold was used. The plot clearly shows the significantly expressed genes with Log Fold Change and FDR cut-offs as mentioned in the code as well.

Conclusion:

There are significant gene expression differences in T cell function between control and 27HC Treatment done on the tissue, with distinct profiles observed in each group.

-Several genes are differentially expressed, with upregulation being notable in Control.

-These findings offer insights into the molecular mechanisms of Cholesterol deficiency in intratumoral cells and potential therapeutic targets.

References:

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE118430

https://bioconductor.org/packages/release/bioc/html/DESeq2.html

https://hbctraining.github.io/DGE_workshop/lessons/01_DGE_setup_and_overview.html

http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#differential-expression-analysis