# Bioinformatics Report: Differential Gene Expression Analysis in Long COVID

## 1. Introduction

This report documents the bioinformatics workflow for analyzing a gene expression dataset (GSE270045) to identify differentially expressed genes (DEGs) in patients with Long COVID compared to a control group. The analysis was performed using the R programming language and several key bioinformatics packages, including DESeq2, EnhancedVolcano, and ComplexHeatmap. The objective of this analysis was to identify genes with significant changes in expression, which can provide insight into the biological pathways affected by the disease.

## 2. Methodology

### 2.1 Data Acquisition and Preprocessing

The analysis began by downloading two key data files from the Gene Expression Omnibus (GEO) database: GSE270045\_counts.tsv (raw gene counts) and GSE270045\_metadata.csv (sample information).

A critical initial step was to ensure that the sample identifiers were consistent across both files. The counts file used CCI IDs, while the metadata file used GSM IDs. To resolve this, the metadata file was manually corrected to match the CCI IDs, which was a vital step for ensuring data integrity and alignment.

The R script then loaded the data, filtered out low-expression genes (counts less than 10 in at least 3 samples), and converted the count data into an integer matrix suitable for DESeq2 analysis. The metadata was also prepared by converting the Conditions variable into a factor and setting "normal" as the reference level for comparison.

### 2.2 Differential Gene Expression Analysis with DESeq2

The core of the analysis was performed using the DESeq2 package. A DESeqDataSet object was created to store the count data and metadata. The DESeq() function was then applied to perform the statistical analysis, which normalizes the data and fits a negative binomial model to test for differential gene expression.

After the analysis was complete, the results() function was used to extract a table of results, including:

* baseMean: The average expression level of the gene.
* log2FoldChange: The magnitude of gene expression change between the two groups.
* padj: The adjusted p-value, which accounts for multiple testing and is used to determine statistical significance.

### 2.3 Identification of Significant DEGs

Significant DEGs were identified by applying a dual filtering criterion:

* **Adjusted p-value (padj) < 0.05**: This ensures that the findings are statistically significant.
* **Absolute log2FoldChange > 1**: This ensures that the genes have a biologically meaningful expression change (at least a two-fold change).

The list of significant DEGs was sorted by adjusted p-value and then exported to a CSV file (significant\_degs.csv) for further review and downstream analysis.

### 2.4 Visualization

To visualize the results of the analysis, two plots were generated: a volcano plot and a heatmap.

#### 2.4.1 Volcano Plot

A volcano plot was created using the EnhancedVolcano package. The plot visualizes the log2 fold change versus the adjusted p-value for each gene. Genes that met the significance criteria were highlighted in a distinct color, providing a clear and comprehensive overview of the overall gene expression changes.

#### 2.4.2 Heatmap

A heatmap was generated to visually represent the expression patterns of the top 10 most up-regulated and top 10 most down-regulated genes. This was performed using the ComplexHeatmap package, which is robust for this type of visualization. The heatmap was scaled by row (Z-score), allowing for a direct comparison of relative expression levels for each gene across all samples. This visualization helps to confirm the distinct expression patterns between the "covid" and "normal" groups.

### 2.5 Combining Top Genes

For additional analysis and to provide a focused list of the most impactful genes, separate CSV files were created for the top 15 up-regulated genes (top15\_up\_regulated\_genes.csv) and top 15 down-regulated genes (top15\_down\_regulated\_genes.csv). These two lists were then combined into a single, comprehensive data frame and exported to a new file named top30\_combined\_genes.csv. This combined file is ideal for subsequent pathway analysis and can serve as a primary input for tools like Cytoscape.

### 2.6 Protein-Protein Interaction Network Analysis

Protein-protein interaction (PPI) network analysis can be used to understand how the identified DEGs interact with each other within a biological system. This type of analysis is typically performed using network visualization tools such as Cytoscape.

The top30\_combined\_genes.csv file, which contains a focused list of the top 30 most impacted genes, can be imported directly into Cytoscape. Using plugins like STRING or GeneMania, an interaction network can be constructed by mapping the genes to a reference database of known protein interactions. The log2FoldChange and padj values from the top30\_combined\_genes.csv file can be used to visually customize the network, for example, by coloring nodes based on fold change or by sizing them based on significance. This allows for the identification of key hub genes and interaction modules.

### 2.7 Functional Enrichment Analysis

Functional enrichment analysis is a crucial next step to understand the biological context of the identified DEGs. This analysis determines if the DEGs are overrepresented in specific biological pathways, molecular functions, or cellular components. The top30\_combined\_genes.csv file was used as the input for the **Enrichr web-based tool**.

By providing the list of genes to Enrichr, the tool queries a wide range of databases, including Gene Ontology (GO), KEGG, and others, to identify enriched pathways and terms. The results are displayed as interactive charts and tables, providing a high-level view of the biological processes most affected by Long COVID. This analysis helps to pinpoint the key molecular mechanisms and pathways that may be driving the observed gene expression changes.

## 3. Conclusion

This report details a complete bioinformatics workflow for differential gene expression analysis using publicly available data. By carefully preprocessing the data, applying a rigorous statistical analysis with DESeq2, and generating clear visualizations and gene lists, we successfully identified and documented genes with significant expression changes. The final outputs—including the volcano plot, heatmap, comprehensive gene lists, and a foundation for subsequent network and enrichment analysis—provide a solid basis for further biological interpretation and research into the molecular mechanisms of Long COVID.