Course Name: Genomic Data Analysis And Precision Medicine

Group Members: Asra Tasneem Shaik; Muni Manasa Vema; Mahima Mahabaleshwar Siddheshwar

Comparative Transcriptomic Profiling of COVID-19 and RSV: Insights from RNA Seq and Weighted Gene Co-expression Network Analysis

Abstract

The COVID-19 pandemic emphasizes the need to identify molecular mechanisms underlying severe respiratory viral infections. This study combines RNA sequencing (RNA-Seq) and Weighted Gene Co-expression Network Analysis (WGCNA) to compare transcriptomic profiles of COVID-19 and Respiratory Syncytial Virus (RSV) infections. RNA-Seq datasets from COVID-19 patients (n = 32), healthy controls (n = 34), convalescent (n= 2) and RSV-infected individuals were analyzed. WGCNA identified key gene modules, with MEyellow and MEgrey strongly correlated with COVID-19 severity. Comparative analysis showed some shared antiviral pathways, including genes such as OASL, TXN and RBCK1, while revealing COVID-19-specific mechanisms linked to immune evasion and viral replication. These findings suggest overlapping and distinct molecular pathways between COVID-19 and RSV, providing insights into potential biomarkers and therapeutic targets for severe respiratory infections.

Keywords: RNA-Seq, WGCNA, COVID-19, RSV, Transcriptomics, Antiviral Pathways, Immune Evasion, Biomarkers, Therapeutic Targets

Introduction

The COVID-19 pandemic, caused by SARS-CoV-2, has emphasized the important knowledge gaps related to respiratory viral diseases and their molecular causes. There are several different clinical symptoms of COVID-19, ranging from mild to severe systemic inflammation and respiratory distress (Huang et al., 2020). Similarly, Respiratory Syncytial Virus (RSV) is a significant respiratory pathogen, primarily affecting infants and the elderly. While both target the respiratory system, their pathogenesis and host responses differ, making comparative molecular studies essential for uncovering shared and distinct mechanisms of disease progression (Vabret et al., 2020). RNA sequencing (RNA-Seq) has emerged as a powerful tool for profiling host gene expression changes during viral infections. By combining RNA-Seq with Weighted Gene Co-expression Network Analysis (WGCNA), clusters of co-expressed and potentially co-regulated genes (modules) can be identified, providing insights into host responses and pathways driving infection severity. Such integrative approaches are vital for identifying biomarkers and therapeutic targets for managing severe respiratory viral infections like COVID-19 and RSV (Mogensen & Paludan, 2021)**.**

Objective

This project aims to find molecular insights into respiratory viral infections, focusing on COVID-19 and Respiratory Syncytial Virus (RSV). Using Weighted Gene Co-expression Network Analysis (WGCNA), it seeks to identify key gene modules associated with COVID-19 severity. Comparative transcriptomic profiling will reveal shared and distinct molecular pathways between COVID-19 and RSV, shedding light on their pathogenesis and host responses (Shen et al., 2020). By integrating these findings, the study aims to identify potential biomarkers and therapeutic targets for managing severe cases of both infections. This comprehensive approach will contribute to a deeper understanding of respiratory viral infections and inform therapeutic strategies.

Methodology

The study employed a structured sequencing pipeline to preprocess RNA-seq data, normalize gene expression values, construct co-expression networks, below are the detailed steps:

Data Acquisition and Preprocessing

RNA-Seq data were obtained from the GEO database, focusing on dataset GSE152418, which includes 32 COVID-19 patients, 34 healthy controls, and 2 convalescent individuals. An additional dataset was sourced for RSV comparison (Huang et al., 2020). Raw sequencing data were downloaded using the SRA Toolkit. Quality control was performed using FastQC to assess base quality, GC content, and adapter contamination. TrimGalore was used for adapter trimming and removal of low-quality bases, with post-trimming quality confirming improved read quality. The dataset is available at GSE152418. The human reference genome (hg38) and annotation file (genes.gtf) were downloaded from UCSC. HISAT2 was employed for genome indexing and alignment, achieving alignment rates exceeding 89% across samples, with quality verified using SAMtools. For feature quantification, SAM files were converted to BAM format, sorted, and indexed using SAMtools. Gene expression levels were quantified using featureCounts, generating a gene count matrix for downstream analyses. This pipeline ensured high-quality, reliable data for comparing transcriptomic profiles of COVID-19 and RSV infections, providing a foundation for exploring shared and distinct molecular pathways.

Weighted Gene Co-expression Network Analysis (WGCNA) and Visualization

The WGCNA framework was applied to a normalized expression matrix to identify gene modules and their associations with clinical traits (Mogensen & Paludan, 2021). A soft-threshold power was selected to ensure a scale-free network, with a power value of 14 chosen based on the scale-free topology model fit and mean connectivity analysis across power values (1–50). Using the blockwiseModules function, a signed co-expression network was constructed with a minimum module size of 30 genes and a merge threshold of 0.25 to combine related modules. Hierarchical clustering grouped genes into distinct modules, each represented by a unique color, visualized using a cluster dendrogram.

Module eigengenes (MEs) were calculated to summarize gene expression patterns within modules. These MEs were correlated with clinical traits, including Condition (COVID-19, Convalescent, Healthy), Severity (Moderate, Severe, ICU, Healthy), and Gender (Male, Female). Pearson correlation coefficients and p-values were computed, and heatmaps were generated to illustrate the strength and significance of module-trait associations, highlighting key interactions.

Hub genes within significant modules were identified based on module membership (MM) scores, with MM > 0.9 defining hub genes. Key modules, such as MEyellow (associated with COVID-19 severity), and MEgrey and MEred (linked to other traits), were prioritized. For the selected RSV dataset, Differentially Expressed Genes (DEGs) were identified based on log2 fold change and p-value thresholds. Shared and unique hub genes across COVID-19 and RSV datasets were identified using Venn diagrams.

To visualize findings, cluster dendrograms, heatmaps, volcano plots, and Venn diagrams were generated, showcasing module-trait correlations, differentially expressed genes, and shared pathways.

Results and Discussion

1. Soft-Threshold Power Selection

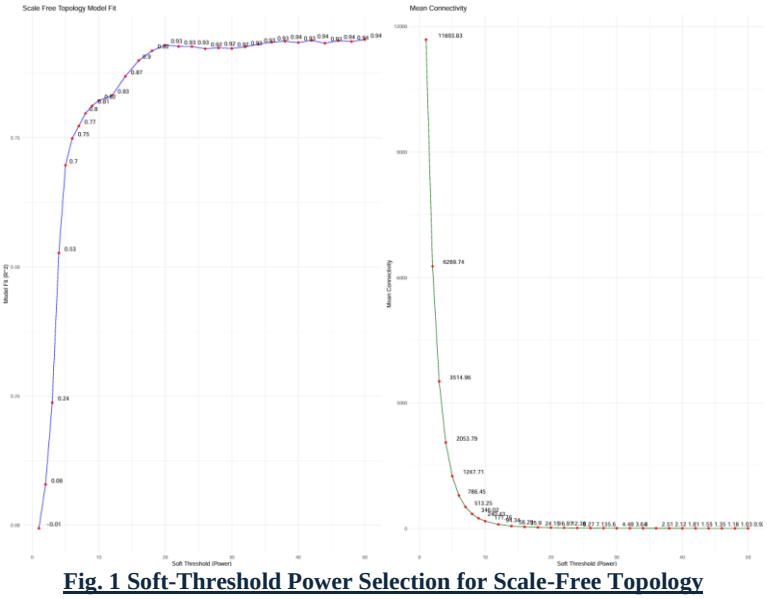

Fig. 1 Soft-Threshold Power Selection for Scale-Free Topology

Fig.1 shows the soft-threshold power selection process in WGCNA to ensure a scale-free gene co-expression network. The left panel illustrates the scale-free topology model fit (R^2), which increases with power and plateaus at ~ 0.94, indicating the optimal threshold at power 14. The right panel shows mean connectivity, which decreases sharply as power increases, ensuring the network focuses on strong gene interactions. Power 14 achieves a balance between high model fit and reduced connectivity, ensuring a biologically meaningful, scale-free network. This selection is critical for robust module detection and reliable hub gene identification in downstream analyses.

2. Co-expression Network Construction

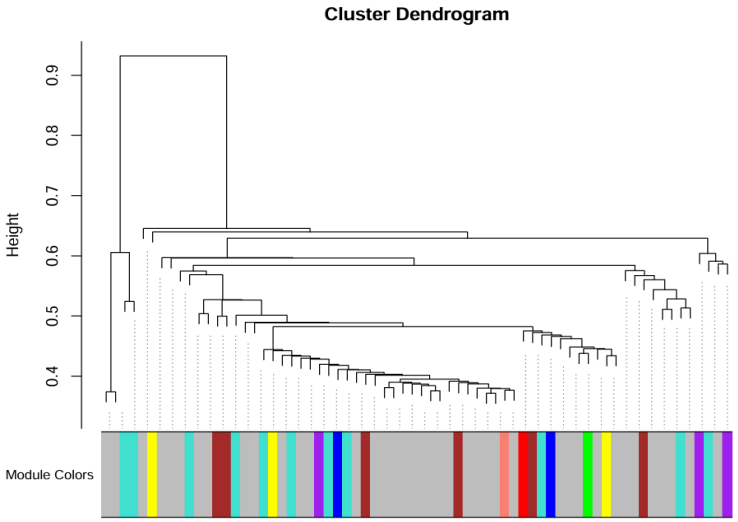

Fig. 2 Cluster Dendrogram of Gene Co-Expression Modules

Using a soft-threshold power of 14, a signed co-expression network was constructed within the WGCNA framework. **The cluster dendrogram (Fig.2) visualizes the** hierarchical clustering of genes into distinct co-expression modules, with the Y-axis representing the dissimilarity between clusters. Closely related genes merge at lower heights, and each module is represented by a unique color in the colored bar, such as grey, yellow, turquoise, and red. These modules highlight distinct co-expression clusters and serve as the basis for further analysis. Correlating these modules with clinical traits uncovers biologically significant relationships and identifies candidate hub genes linked to disease progression.

3. Module-Trait Correlation Analysis

The three heatmaps (Fig. 3, 4 & 5) depict the correlation between gene modules (rows) identified through WGCNA and various clinical traits (columns) such as Condition, Gender, and Severity. The color gradient ranges from blue (negative correlation) to red (positive correlation), with numerical values and p-values showing the strength and significance of the relationships. In the Condition-based heatmap, the MEyellow module shows a strong positive correlation with COVID-19 samples (r=0.89, p<2e -24) and a significant negative correlation with Healthy controls (r=−0.85). This suggests MEyellow is highly associated with COVID-19 progression. Conversely, the MEred module exhibits a positive correlation with Healthy samples, indicating its role in homeostasis. The Gender-based heatmap highlights MEtan with a strong positive correlation with Male (r=0.44) and negative correlation with Female (r=−0.44), while MEgreenyellow correlates positively with Female (r=0.37). This demonstrates gender-specific differences in gene expression patterns. In the Severity-based heatmap, MEyellow correlates positively with Severe (r=0.46) and ICU cases (r=0.39), and negatively with Healthy (r=−0.85). Other modules, such as MEtan and MEmidnightblue, also correlate with moderate and severe conditions (Han et al., 2020).

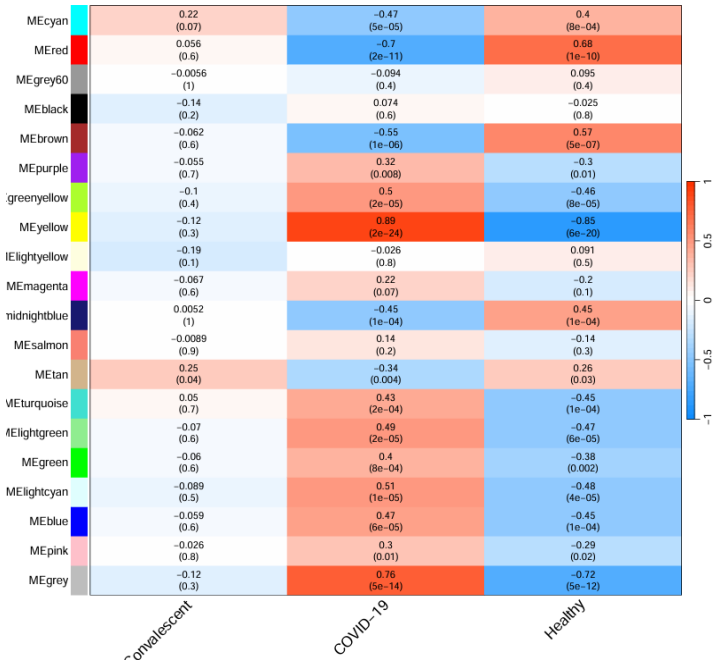

Fig.3 Module-Trait Correlation Heatmap for COVID-19 Severity

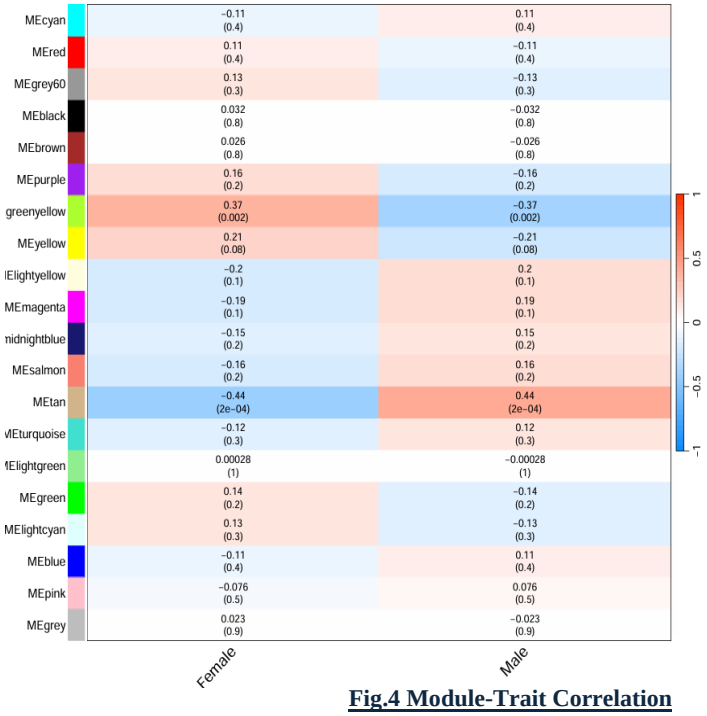

Fig.4 Module-Trait Correlation Heatmap for Gender

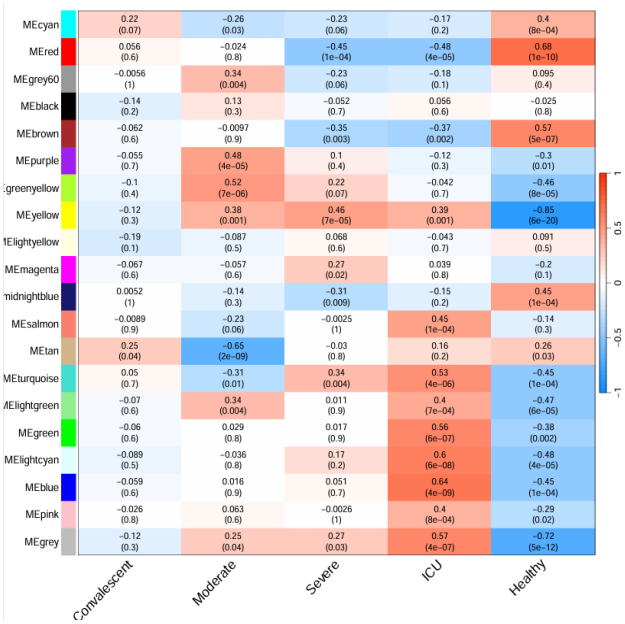

Fig.5 Module-Trait Correlation Heatmap for Condition (COVID-19 vs. Healthy)

4. Differential Expression Analysis

The volcano plot (Fig.6) visualizes differential gene expression, plotting log2 fold change (X-axis) against -log10 p-value (Y-axis). Genes are categorized into upregulated (red), downregulated (blue), and non-significant (grey) groups. Upregulated genes, with positive fold changes, show higher expression in the condition of interest, while downregulated genes, with negative fold changes, indicate reduced expression. Non-significant genes cluster near the center, reflecting minimal changes or high p-values. Genes at the plot's edges represent significantly altered expression, with upregulated genes on the right and downregulated on the left. This plot highlights candidate genes for further analysis as biomarkers or therapeutic targets.

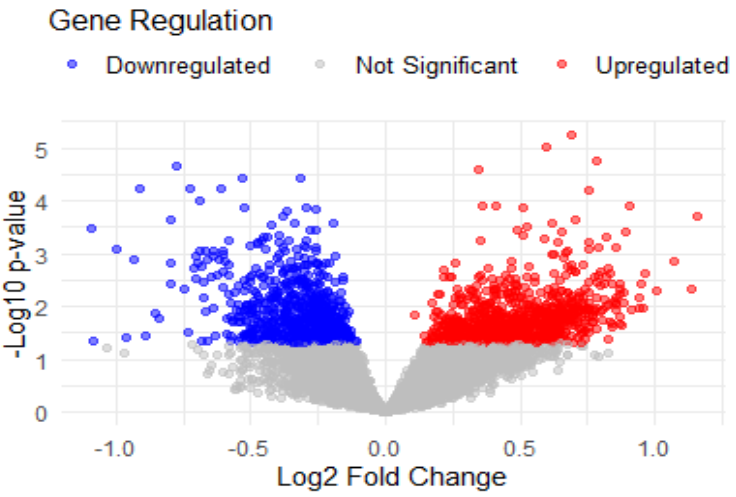

Fig.6 Volcano Plot of Differentially Expressed Genes (DEGs)

5. Overlap Between COVID-19 and RSV Hub Genes

The Venn diagram (Fig.7) illustrates the overlap of hub genes between COVID-19 and RSV datasets. The red circle represents hub genes identified in COVID-19, totaling 97 unique genes, while the green circle represents RSV hub genes, with 96 unique genes. The overlapping region contains 3 shared hub genes, indicating common immune responses activated in both infections. These shared genes are of particular interest as they may play critical roles in antiviral responses, such as immune activation, stress regulation, and viral clearance (Blanco-Melo et al., 2020; Wu & McGoogan, 2020). The presence of distinct gene sets highlights unique molecular mechanisms specific to COVID-19 and RSV infections (Brodin, 2020).

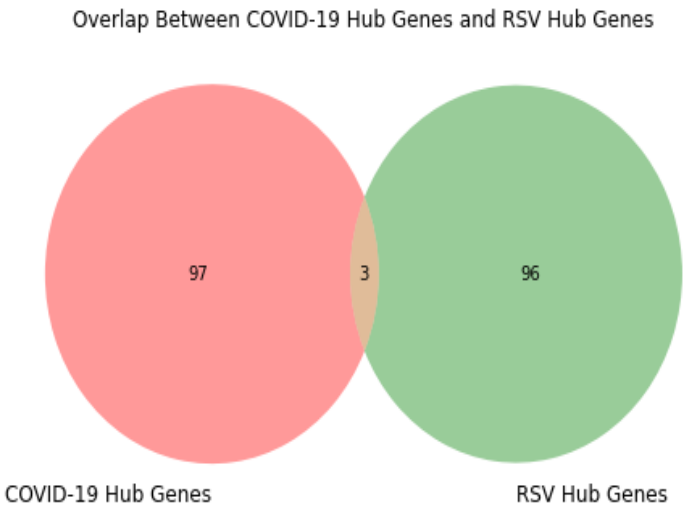

Fig.7 Venn Diagram Showing Shared and Unique Hub Genes

Hub Gene Identified

- **OASL (2'-5'-Oligoadenylate Synthetase Like):** Involved in antiviral response by degrading viral RNA. Upregulated in RSV and COVID-19 to enhance immune defense and viral clearance.
- **TXN (Thioredoxin):** Regulates oxidative stress and inflammation during RSV infection. Modulates redox balance and immune responses in COVID-19, potentially influencing viral replication and immune activation.
- **RBCK1 (RANBP2-Type and C3HC4-Type Zinc Finger Containing 1):** Regulates protein degradation through the ubiquitin-proteasome pathway, impacting viral protein turnover. Plays a role in immune regulation and antiviral defense in both RSV and COVID-19 infections.

Conclusion

This study successfully employed RNA sequencing (RNA-Seq) and Weighted Gene Co-expression Network Analysis (WGCNA) to compare transcriptomic profiles between COVID-19 and RSV infections. Co-expression networks **revealed distinct gene modules, such as MEyellow and MEgrey,** significantly associated with COVID-19 severity. Shared and unique pathways were identified, with COVID-19 exhibiting immune hyperactivation and platelet dysfunction, while RSV showed distinct immune responses. **The analysis highlighted only three shared hub genes (OASL, TXN, and RBCK1)** between COVID-19 and RSV datasets (Blanco-Melo et al., 2020). This limited overlap can be attributed to:

Biological Differences: COVID-19 and RSV differ in their pathogenesis, host immune responses, and disease progression, leading to unique gene expression patterns.

Severity-Specific Variability: The COVID-19 dataset included severity-based stratification (e.g., ICU, severe), introducing variability not uniformly present in RSV data.

Data-Specific Factors: Variations in sample size, experimental design, and sequencing platforms could impact hub gene detection and result in fewer shared genes.

Despite the small overlap, the identified hub genes play critical roles in antiviral responses, immune regulation, and stress pathways, offering insights into shared mechanisms. This study underscores the complexity of viral infections and highlights the need for disease-specific and comparative transcriptomic analyses to identify robust biomarkers and therapeutic targets.

Future Direction

Future studies should focus on increasing the sample size with balanced severity levels for both COVID-19 and RSV to improve statistical power. Functional validation of hub genes (**OASL, TXN, and RBCK1**) using in-vivo models will confirm their biological relevance. Pathway enrichment analysis and single-cell RNA-seq can provide deeper insights into cell-specific responses. Expanding comparisons to other respiratory viruses, such as influenza, will help identify conserved pathways. Integrating multi-omics data (proteomics, metabolomics) and **applying machine learning models** will enhance hub gene detection and therapeutic target identification, advancing precision medicine for respiratory viral infections (Vabret et al., 2020).

References

1. Blanco-Melo, D., Nilsson-Payant, B. E., Liu, W.-C., Uhl, S., Hoagland, D., Møller, R., ... & tenOever, B. R. (2020). Imbalanced host response to SARS-CoV-2 drives development of COVID-19. *Cell*, *181*(5), 1036-1045.e9. https://doi.org/10.1016/j.cell.2020.04.026
2. Brodin, P. (2020). Why is COVID-19 so mild in children? *Acta Paediatrica*, *109*(6), 1082–1084. https://doi.org/10.1111/apa.15271
3. Han, H., Ma, Q., Li, C., Liu, R., Zhao, L., Wang, W., & Zhang, P. (2020). Profiling serum cytokines in COVID-19 patients reveals IL-6 and IL-10 are disease severity predictors. *Emerging Microbes & Infections*, *9*(1), 1123–1130. https://doi.org/10.1080/22221751.2020.1770129
4. Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., ... & Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, *395*(10223), 497-506. https://doi.org/10.1016/S0140-6736(20)30183-5
5. Mogensen, T. H., & Paludan, S. R. (2021). Molecular pathways in virus-induced cytokine production. *Microbiology and Molecular Biology Reviews*, *65*(1), 131-150. https://doi.org/10.1128/MMBR.65.1.131-150.2001
6. Shen, B., Yi, X., Sun, Y., Bi, X., Du, J., Zhang, C., ... & Chen, H. (2020). Proteomic and metabolomic characterization of COVID-19 patient sera. *Cell*, *182*(1), 59-72.e15. https://doi.org/10.1016/j.cell.2020.05.032
7. Vabret, N., Britton, G. J., Gruber, C., Hegde, S., Kim, J., Kuksin, M., ... & Sinai Immunology Review Project. (2020). Immunology of COVID-19: Current state of the science. *Immunity*, *52*(6), 910–941. https://doi.org/10.1016/j.immuni.2020.05.002
8. Wu, Z., & McGoogan, J. M. (2020). Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China. *JAMA*, *323*(13), 1239–1242. https://doi.org/10.1001/jama.2020.2648

Team Contribution

- **Mahima Mahabaleshwar Siddheshwar:** Performed the upstream analysis, including data acquisition, preprocessing (FastQC, TrimGalore), genome alignment (HISAT2), and feature quantification using featureCounts. Prepared the normalized gene expression matrix and ensured quality control for downstream analyses.
- **Muni Manasa Vema:** Conducted the Weighted Gene Co-expression Network Analysis (WGCNA), including soft-threshold power selection, module detection, module-trait correlation analysis, and hub gene identification. Contributed to generating visualizations such as cluster dendrograms and heatmaps.
- **Asra Tasneem Shaik:** Performed the Differentially Expressed Gene (DEG) analysis for the RSV dataset and visualized results using volcano plots. Compared hub genes between COVID-19 and RSV datasets using Venn diagrams and contributed to pathway analysis.

"All team members actively contributed to report writing, creating GitHub repositories, and finalizing the project presentation."