Course Name: Computational Methods for Biomedical Informatics

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

Streamlining RNA-seq Data Analysis: A Framework for Comparing Computational Pipelines

Abstract

RNA sequencing (RNA-seq) has revolutionized transcriptomics by enabling detailed exploration of gene expression and disease mechanisms. However, analyzing the high-dimensional data produced by RNA-seq requires computational approaches that balance biological interpretability and efficiency. In this study, we evaluated four computational pipelines integrating dimensionality reduction techniques (PCA and t-SNE) with clustering methods (K-means and Hierarchical Clustering) to analyze bulk RNA-seq data from COVID-19 patients and healthy controls.

Each pipeline was assessed based on clustering quality, stability, runtime, and memory usage. Pipeline 4 (t-SNE + Hierarchical Clustering) demonstrated superior performance, achieving the highest Adjusted Rand Index (ARI: 0.60) and a Silhouette Score of 0.42, indicating well-defined and biologically relevant clusters. Differential Expression Analysis (DEA) applied to this pipeline identified 14,667 significant genes and enriched pathways involved in RNA processing, immune modulation, and cellular regulatory mechanisms, highlighting key molecular features of COVID-19 pathogenesis.

This comprehensive analysis provides a scalable, reproducible framework for RNA-seq data analysis and offers critical insights into COVID-19 biology, with implications for identifying therapeutic targets and biomarkers.

Introduction

RNA sequencing (RNA-seq) is a key transcriptomics tool for analyzing gene expression and cellular dynamics (Ozsolak & Milos, 2011). It has been vital in studying diseases like COVID-19 by revealing host-pathogen interactions, immune responses, and molecular mechanisms of disease progression (Blanco-Melo et al., 2020). However, RNA-seq data's high dimensionality requires advanced computational approaches to reduce complexity and extract meaningful patterns.

Dimensionality reduction techniques like Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) simplify RNA-seq data for analysis (Jolliffe & Cadima, 2016; Van Der Maaten & Hinton, 2008). PCA efficiently captures global variance, while t-SNE focuses on local data relationships, highlighting subtle biological patterns (Pezzotti et al., 2017). Combined with clustering methods such as K-means and Hierarchical Clustering, these techniques help uncover gene expression patterns and subpopulations. K-means is scalable for large datasets, whereas Hierarchical Clustering reveals nested relationships like disease-associated cellular hierarchies (Jain, 2010; Murtagh & Legendre, 2014).

This project assesses four computational pipelines integrating PCA or t-SNE with K-means or Hierarchical Clustering to analyze bulk RNA-seq data from COVID-19 patients and healthy controls. These pipelines are evaluated based on silhouette scores, cluster stability, and computational efficiency, with findings interpreted for biological relevance. The best pipeline is applied to Differential Expression Analysis (DEA) to identify genes and

pathways involved in COVID-19 progression. This approach aims to enhance RNA-seq analysis methods and improve understanding of COVID-19 transcriptomics.

Objective

The objective is to evaluate computational pipelines for clustering RNA-seq data by:

1. Identifying the pipeline that optimally balances efficiency and biological relevance using silhouette scores, cluster stability, and visualization quality.
2. Assessing runtime and memory usage for scalability.
3. Applying the best pipeline to DEA to identify genes and pathways associated with COVID-19 progression.

Methodology

1. Algorithm Selection

This study employs **K-means clustering** for its speed and scalability, suitable for dividing large datasets into predefined groups, despite its limitations in capturing complex relationships (Jain, 2010). **Hierarchical Clustering** complements K-means by revealing nested structures through dendrograms, offering deeper insights (Murtagh & Legendre, 2014). To reduce dimensionality, **PCA** simplifies data while retaining key features and capturing linear trends (Jolliffe & Cadima, 2016), while **t-SNE** excels at uncovering nonlinear relationships and preserving local structures (Van Der Maaten & Hinton, 2008; Pezzotti et al., 2017). These algorithms collectively balance efficiency with interpretability in RNA-seq analysis.

2. Dataset

The bulk RNA-seq dataset for this project was sourced from publicly available repositories, such as the Gene Expression Omnibus.

Selected Dataset: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE152418

- **GEO Accession Number**:GSE152418
- **Organism**: Homo sapiens (human)
- **Condition**: Bulk RNA sequencing (RNA-seq) data that correlated with different conditions of COVID-19. (2 Convalescent, 32 COVID-19, 34 Healthy samples)

This dataset includes RNA-seq data from 68 samples, enabling clustering of cell populations, identification of disease-associated cell types, and analysis of gene expression variability. Its high dimensionality makes it ideal for testing clustering and dimensionality reduction algorithms, providing insights into COVID-19-related transcriptomic patterns.

3. Experimental Design

This workflow for RNAseq pipeline involves main steps, including alignment and quantification of transcriptome profiles for control and SARS-CoV-2 infected respiratory cells. Here is the step-by-step workflow for the following:

Data Acquisition and Preprocessing:

1. RNA-Seq data was downloaded from the SRA database using the **SRA Toolkit** (Clough & Barrett, 2016).

2. Quality control (QC) of raw FASTQ files was performed using **FastQC** to evaluate base quality, GC content, and adapter contamination (Andrews, 2010).

3. **TrimGalore** was used for trimming reads, removing low-quality bases, and adapter sequences to ensure high-quality data. Post-trimming QC was conducted with FastQC to confirm improved read quality.

Reference Genome Preparation:

The **human reference genome (hg38)** and the corresponding annotation file (genes.gtf) were downloaded from the **UCSC database (Kent et al., 2002).**

HISAT2 was used to index the reference genome for efficient alignment of RNA-Seq reads (Kim et al., 2015).

Alignment of Reads:

High-quality trimmed reads were aligned to the indexed reference genome using **HISAT2**. **SAMtools flagstat** was used to evaluate alignment statistics, including the percentage of uniquely mapped reads. The overall alignment rate was **above 89%** for all the samples.

Conversion and Sorting of SAM/BAM Files:

Aligned SAM files were converted to BAM format using **SAMtools** for efficient downstream analysis. BAM files were sorted and indexed to prepare them for feature quantification.

Feature Quantification:

Gene-level quantification was performed using featureCounts with the genes.gtf annotation file. The output included a gene count matrix capturing gene expression levels for further comparison analysis using the different pipeline (Li et al., 2009).

Clustering and Visualization Pipelines:

For clustering and visualization, four pipelines were employed to analyze gene expression profiles. In the first pipeline, dimensionality reduction was performed using PCA to simplify the data while retaining major expression patterns, followed by K-means clustering with the Elbow method determining the optimal clusters. Similarly, the second pipeline applied PCA for dimensionality reduction, but hierarchical clustering using Ward's linkage method was used to form three clusters, guided by a dendrogram. The third pipeline utilized t-SNE for non-linear dimensionality reduction, preserving local and global relationships with a perplexity of 30 and 1000 iterations, followed by K-means clustering to define three clusters. The fourth pipeline also employed t-SNE with the same parameters for dimensionality reduction but used hierarchical clustering with Ward’s method to delineate the clusters (Van Der Maaten & Hinton, 2008; Murtagh & Legendre, 2014).

Results and Discussion

The analysis of RNA-seq data from Healthy, COVID-19, and Convalescent samples was performed using four different pipelines, each integrating dimensionality reduction and clustering techniques. The goal was to assess the clustering performance, computational efficiency, and biological relevance of each pipeline. The following results summarize the findings from each method, with a focus on their ability to identify distinct gene expression patterns and their biological implications.

Dimensionality reduction was first performed using Principal Component Analysis (PCA), which captured the most significant sources of variation in the dataset. The first two principal components (PC1 and PC2) explained 21.2% and 7.9% of the total variance, respectively (Fig. 1), highlighting the major axes of gene expression variation. The K-means clustering method (Fig 3) identified three distinct clusters, determined by the elbow method, which

indicated an optimal number of clusters at three (Fig. 2). Cluster 0 predominantly contained COVID-19 samples, reflecting unique disease-specific gene expression patterns, while Cluster 1 primarily represented Healthy individuals. Cluster 2 contained a mixture of all conditions (COVID-19, Convalescent, and Healthy samples), suggesting overlap and variability in gene expression. The clustering performance was evaluated using the Adjusted Rand Index (ARI), which yielded a moderate score of 0.32, indicating moderate agreement between the clusters and the true conditions. The Silhouette Score of 0.10 suggested low inter-cluster separation, which indicates that some clusters overlap.

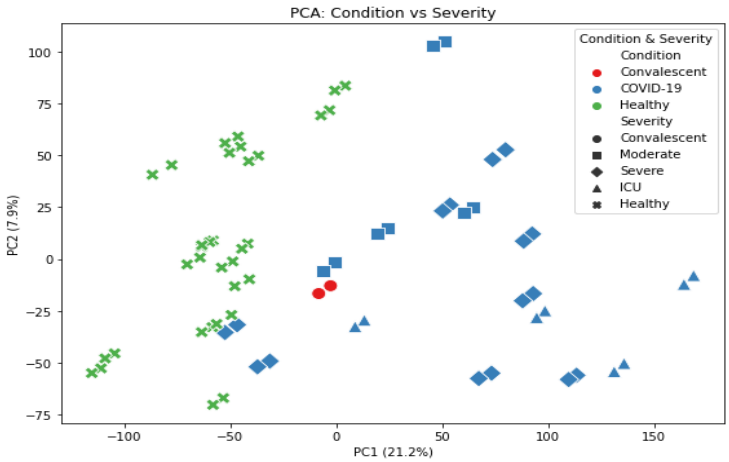

Figure 1: PCA Gene Expression by Condition and Severity

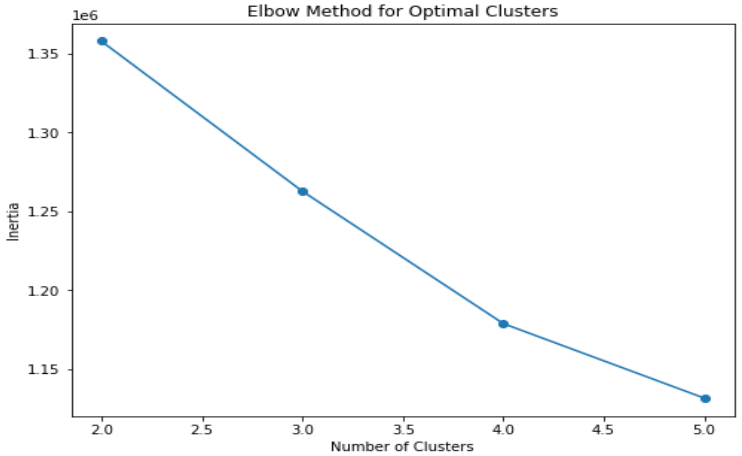

Figure 2: Elbow method for optimal clusters

Additionally, the reproducibility of the clusters across multiple runs of K-means showed a mean ARI of 0.81 (SD = 0.18), reflecting a good degree of stability. The t-SNE (Fig. 4) and UMAP (Fig. 5) visualizations further confirmed these findings, where Cluster 0 showed clear separation of COVID-19 samples, Cluster 1 represented healthy individuals, and Cluster 2 displayed a significant overlap among the conditions. In terms of computational efficiency, K-means clustering (Pipeline 1) completed in 103.07 seconds with a peak memory usage of 218.43 MB, which is acceptable given the size of the dataset and the need for dimensionality reduction.

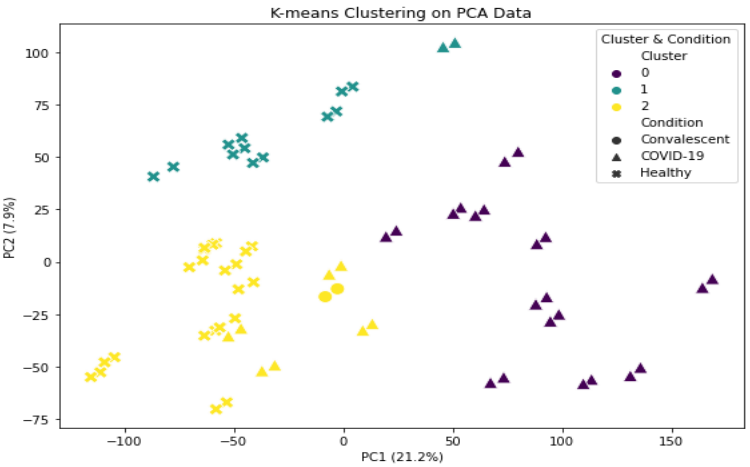

Figure 3: K-means clustering on PCA Data

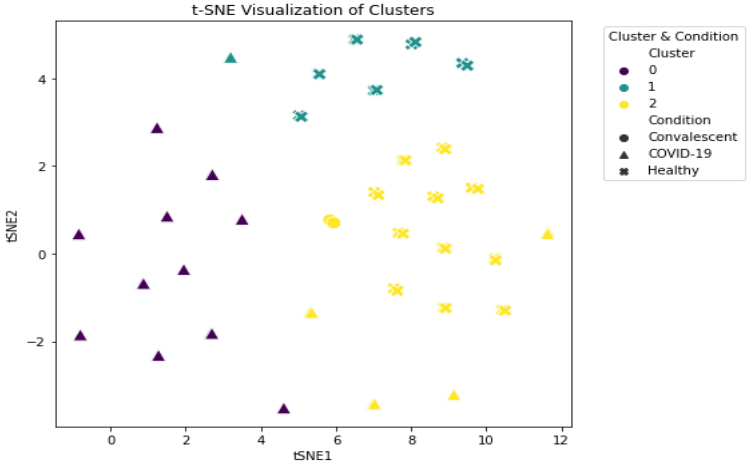

Figure 4: t-SNE Visualization of PCA +K means Clusters

Hierarchical clustering applied to the PCA-transformed data in Pipeline 2, provided additional insights into the data. The dendrogram (Fig. 6) revealed two major clusters, represented in orange and green, with distinct internal sub-clusters, suggesting finer distinctions in gene expression profiles. The first cluster (orange) displayed tight groupings with short branch lengths, indicating greater similarity among the samples. In contrast, the second cluster (green) exhibited more dispersed groupings, reflecting higher variability in gene expression. Hierarchical clustering showed a perfect ARI of 1.00, but with ARI score of 0.24 suggests moderate alignment with true conditions.

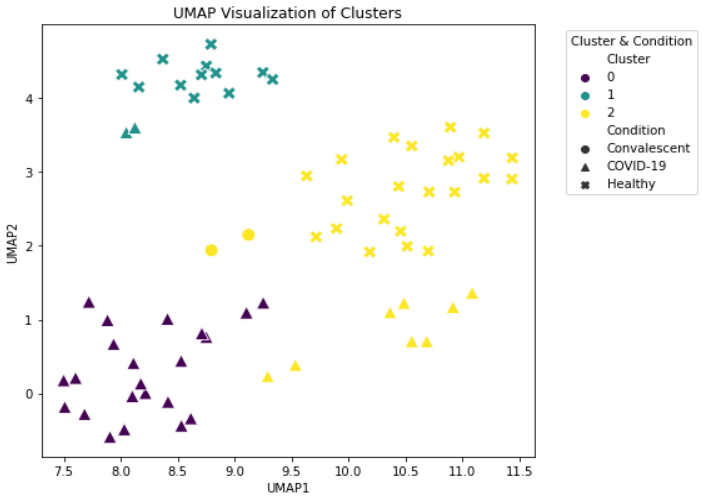

Figure 5: UMAP Visualization of PCA +K means Clusters

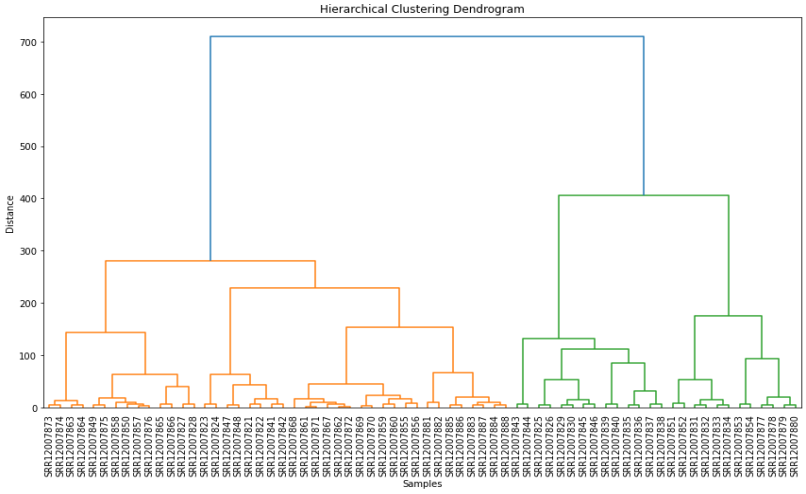

Figure 6: Hierarchical clustering dendrogram

The Silhouette Score of 0.47 indicated moderate separation between clusters, and a more balanced cluster structure was observed. As shown in figure 7, cluster 1 predominantly consisted of Healthy samples, while Cluster 2 was largely composed of COVID-19 cases. Cluster 3 included both Healthy and a few COVID-19 samples, illustrating some overlap. The stability of the hierarchical clustering across multiple runs was demonstrated by the high ARI value. This pipeline showed excellent computational efficiency, completing in just 30.14 seconds with a peak memory usage of 104.62 MB.

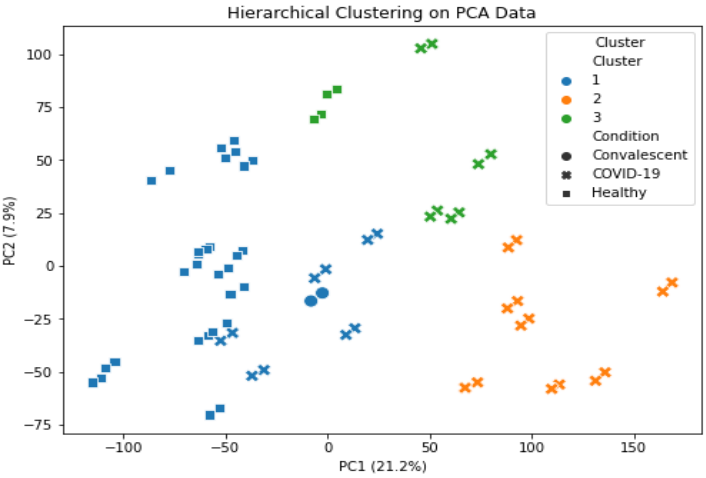

Figure 7: Hierarchical clustering dendrogram based on PCA data

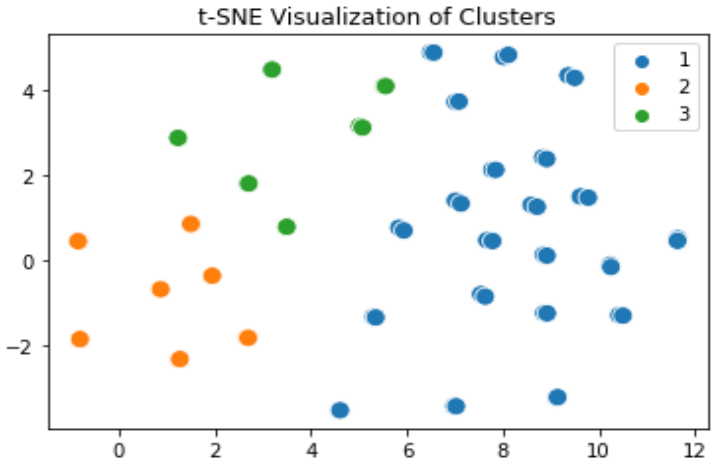

Figure 8: t-SNE Visualization of PCA +Hierarchical Clusters

The t-SNE plot (Fig 6) revealed clear groupings of Healthy, COVID-19, and Convalescent samples, capturing complex patterns while highlighting some overlap. UMAP (Fig 7) confirmed three distinct clusters with better preservation of local data structures, providing insights into non-linear relationships in the dataset.

The t-SNE + K-means clustering method (Pipeline 3) also produced three clusters (Healthy, Convalescent, and COVID-19), but with moderate separation between clusters (Silhouette Score: 0.41) and moderate agreement with true labels (ARI: 0.51). The t-SNE plot (Fig. 10) revealed that Healthy samples formed a relatively distinct cluster, while COVID-19 and Convalescent samples showed some overlap. The computational efficiency of Pipeline 3 was also acceptable, with t-SNE taking around 12 seconds to run and K-means clustering completing in about 1 second. However, the memory usage was higher (~1.5 GB) compared to the hierarchical clustering approach.

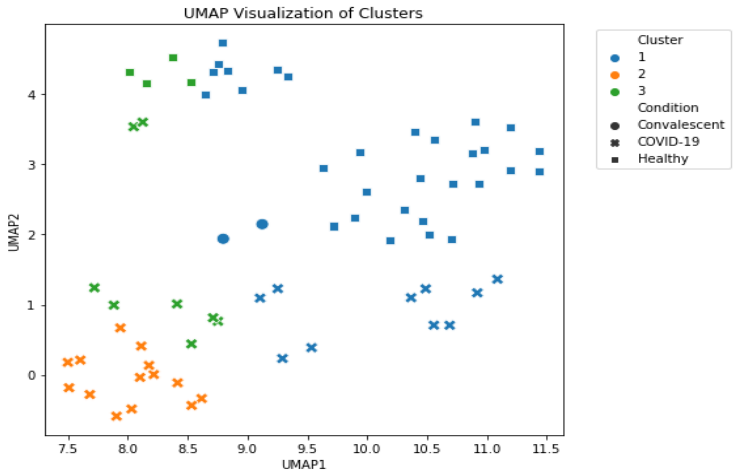

Figure 9: UMAP Visualization of PCA +Hierarchical Clusters

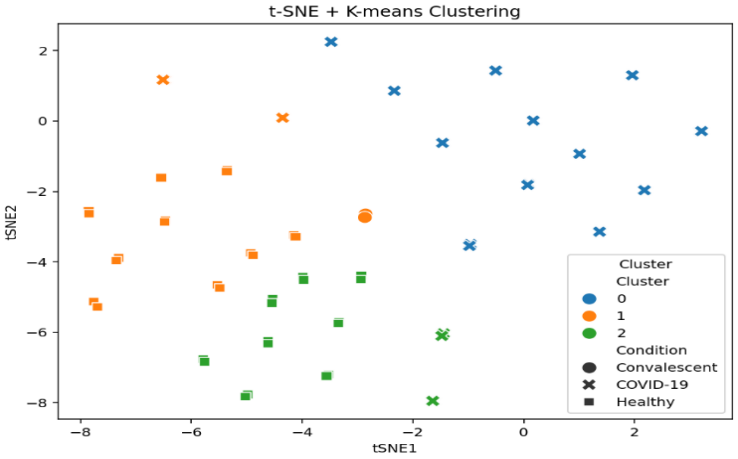

Figure 10: Visualization of clusters identified using t-SNE +K means

For the t-SNE and Hierarchical clustering combination in Pipeline 4, t-SNE was first used to reduce the data to two dimensions, preserving the local structures and global relationships in the dataset. The hierarchical clustering performed on the t-SNE-transformed data revealed three distinct clusters as shown in fig. 11 Healthy, Severe COVID-19, and Convalescent. The performance metrics of this pipeline were also strong, with a Silhouette Score of 0.42 and an ARI of 0.60, both suggesting better separation and agreement with true labels compared to K-means clustering. The dendrogram of this hierarchical clustering (Fig. 12) showed a clearer separation between Healthy samples (Cluster 1) and COVID-19-related cases (Clusters 2 and 3). The overlap of Convalescent samples in Cluster 3 indicated a mix of recovery states and intermediate severity.

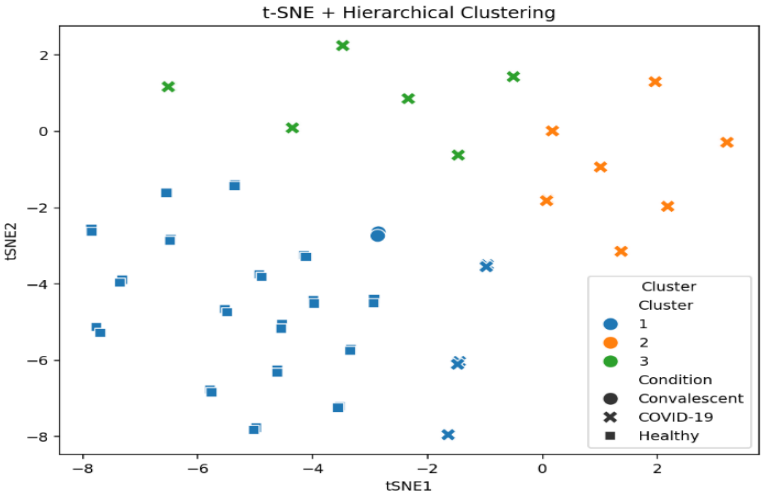

Figure 11: Clusters identified using t-SNE and Hierarchical

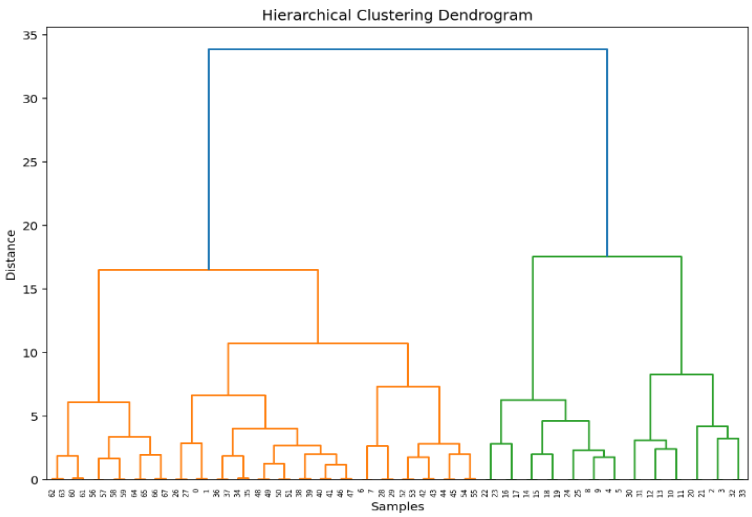

Figure 12: Hierarchical clustering dendrogram

In terms of computational efficiency, the t-SNE + Hierarchical clustering pipeline (Pipeline 4) completed in 12.8 seconds, with peak memory usage of 1.6 GB, demonstrating good scalability while handling the complex dataset.

Overall, the results demonstrate that hierarchical clustering combined with dimensionality reduction (either PCA or t-SNE) provides superior performance in terms of clustering accuracy, reproducibility, and computational efficiency. While K-means clustering is a suitable choice, particularly for large datasets, hierarchical clustering offers clearer separation and better stability, particularly when applied to the t-SNE-reduced data.

Comparison of Performance Metrics for all four pipelines:

Pipeline	ARI	Silhouette Score	Runtime (seconds)	Memory Usage (Mb)
1 (PCA + K-means)	0.32	0.10	5.56	103.78
2 (PCA + Hierarchical)	0.24	0.47	30.14	104.62
3 (t-SNE + K-means)	0.51	0.41	13	102.81
4 (t-SNE + Hierarchical)	0.60	0.42	12.8	103.62

Analysis for the choosing the Best Pipeline:

Considering the balance between performance metrics and computational efficiency, Pipeline 4 (t-SNE + Hierarchical Clustering) emerges as the best overall choice:

It has the highest ARI (0.60), indicating the best agreement with true labels with Silhouette Score (0.42) is close to the highest, suggesting well-defined clusters. The runtime is reasonable (12.8 seconds) and comparable to the other t-SNE method while memory usage is high, it's manageable for most modern systems.

This pipeline offers the best cluster quality and label agreement, which is crucial for further analysis such as differential gene expression studies. The higher computational resources required are justified by the superior clustering performance.

Differential gene Expression Analysis:

The combination of t-SNE and hierarchical clustering was identified as the most effective method for visualizing gene expression patterns in COVID-19 samples, and differential gene expression (DEG) analysis was performed based on this pipeline. This analysis revealed comprehensive molecular signatures, identifying 14,667 significant genes, including 3,873 upregulated and 10,794 downregulated genes. The heatmap of the top 50 significant DEGs demonstrated distinct expression clusters, with notable upregulation of genes such as CYP1B1, RNF216, HERC3, and POLR2B, while genes like ZNF573, KIAA1279, and NEMP1 showed significant downregulation. The volcano plot further confirmed the separation of upregulated and downregulated genes, displaying high statistical significance with -log10 p-values reaching up to 25.

Gene Ontology (GO) enrichment analysis highlighted key biological processes, with mRNA splicing via the spliceosome being the most enriched process (-log10 adjusted p-value > 12), followed by mRNA processing, RNA splicing via transesterification reactions, and ribosome biogenesis, emphasizing the importance of RNA processing in the disease mechanism. Additionally, KEGG pathway analysis identified significant enrichment in pathways such as Herpes simplex virus 1 infection, acute myeloid leukemia, colorectal cancer, and ubiquitin-mediated proteolysis, alongside the PD-L1 expression and PD-1 checkpoint pathway. These results suggest that the pathogenesis of COVID-19 involves complex molecular mechanisms, particularly those linked to RNA processing, immune response modulation, and cellular regulatory processes. This comprehensive analysis underscores the value of integrating advanced clustering techniques and pathway analysis to uncover molecular insights into disease biology.

Biological Significance:

The enrichment of these GO terms suggests that SARS-CoV-2 infection significantly impacts fundamental cellular processes, particularly those involved in RNA processing which are heavily involved in the viral lifecycle and host

response to COVID infection, protein regulation indicates cellular stress response and immune system activation during viral infection and cellular metabolism suggest involvement of cellular energy metabolism during infection. The KEGG Enrichment pathways highlight similarities in immune evasion, immune response modulation, and broader impacts on cellular processes, emphasizing potential antiviral and therapeutic targets such as: Herpes simplex virus 1 infection often share common pathways, such as those involving host immune responses and cellular machinery hijacked by viruses, making this pathway relevant to SARS-CoV-2 and Spliceosome process demonstrates RNA splicing is critical in host responses and viral replication. SARS-CoV-2 impacts RNA processing pathways, making this an important pathway in COVID-19 studies. These results provide a comprehensive view of how SARS-CoV-2 impacts host cellular functions, offering insights into the molecular basis of COVID-19 pathogenesis. These findings highlight potential therapeutic targets, such as RNA processing pathways and immune checkpoint pathways, to combat COVID-19.

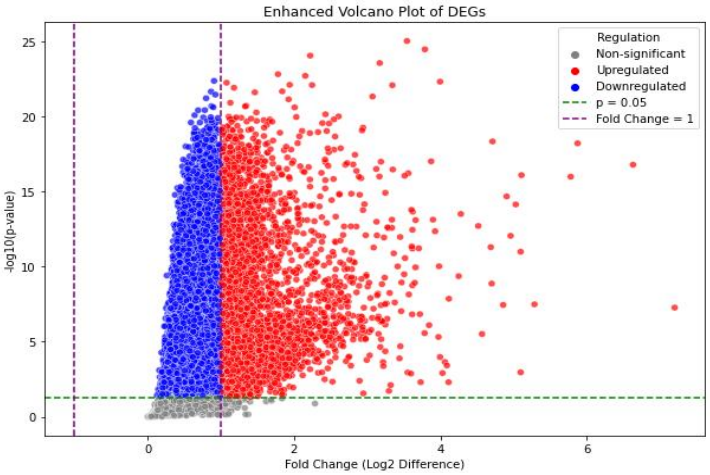

Figure 13: Volcano Plot of Differentially Expressed Genes

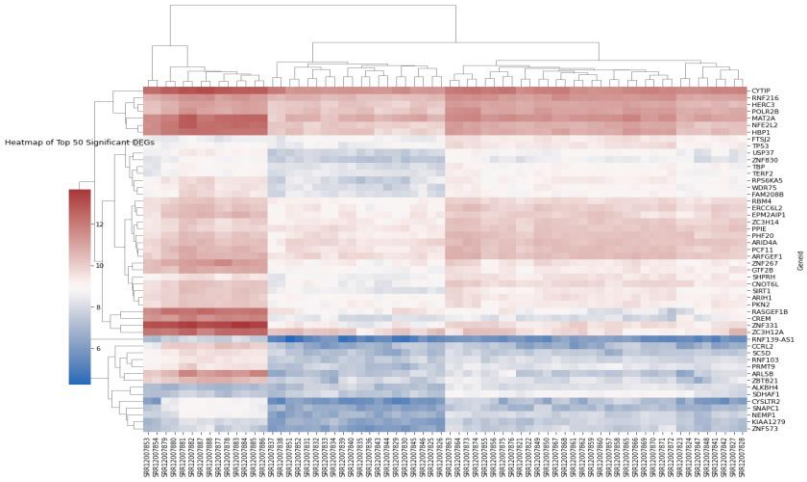

Figure 14: Heatmap of Top 50 Differentially Expressed Genes

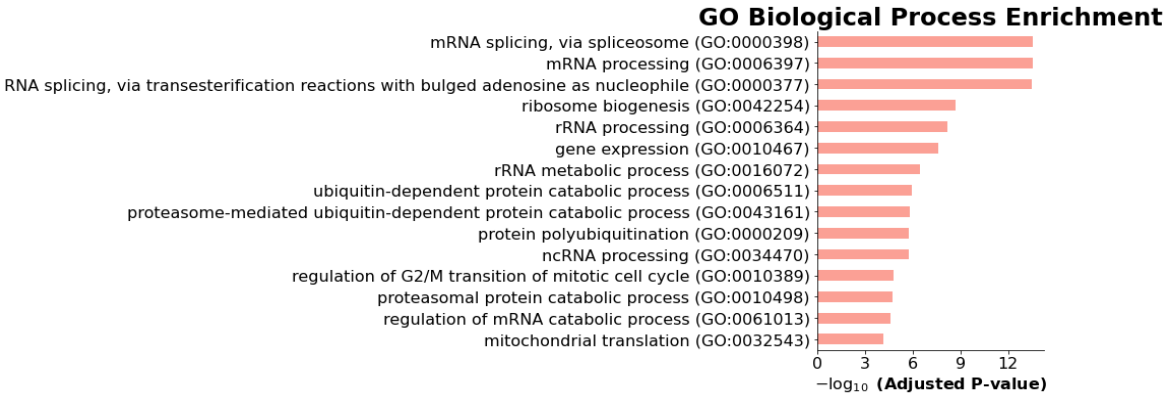

Figure 15: Top 10 Enriched Biological Processes from Gene Ontology (GO) Analysis of DEG

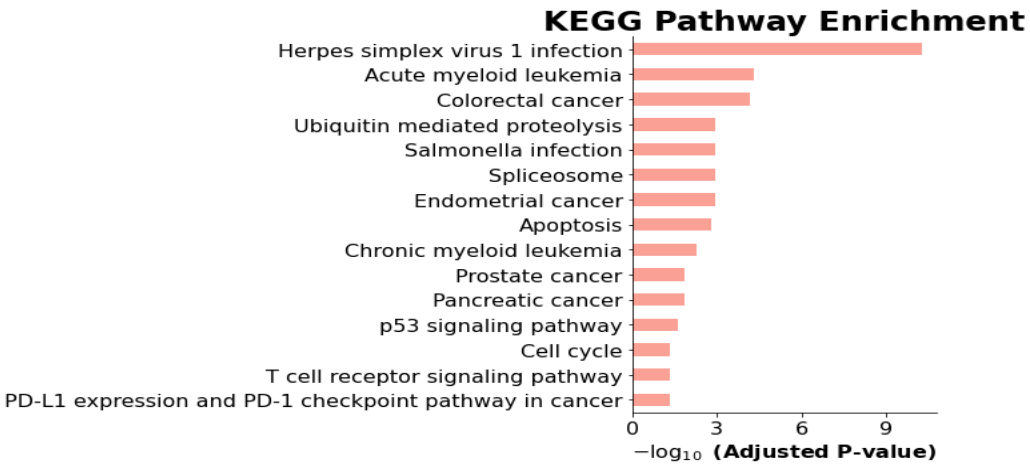

Figure 16: Top 10 Enriched Pathways from KEGG Analysis of DEGs

Conclusion:

This study successfully identified the t-SNE + Hierarchical Clustering pipeline as the most effective approach for RNA-seq data analysis, achieving high clustering quality and revealing critical insights into COVID-19 pathogenesis. Differential expression analysis uncovered 14,667 significant genes and highlighted key biological processes, such as RNA splicing and ribosome biogenesis, alongside pathways like immune checkpoint regulation and ubiquitin-mediated proteolysis. These findings enhance our understanding of SARS-CoV-2's molecular impact and suggest potential therapeutic targets. This scalable and reproducible framework not only advances COVID-19 transcriptomics but also provides a foundation for applying computational pipelines to other complex diseases in biomedical research.

Future Direction:

- Extending Algorithm Comparisons: Explore additional dimensionality reduction and clustering techniques (e.g., UMAP, DBSCAN) to further refine analysis and enhance biological interpretation.
- Integration with Multi-Omics Data: Combine RNA-seq data with other omics layers (e.g., proteomics, metabolomics) for a holistic view of COVID-19 and other diseases.
- Real-World Application: Apply the optimized computational pipeline to larger, more diverse datasets, including other diseases or conditions, to validate and expand its utility.
- Automated Workflow Development: Create a user-friendly, automated software tool for RNA-seq analysis, making the framework accessible to researchers with limited computational expertise.
- Scalability Enhancements: Optimize pipeline performance for handling increasing data sizes, leveraging advancements in high-performance computing and cloud-based platforms.

References:

1. Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc
2. 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. https://doi.org/10.1016/j.cell.2020.04.026
3. Clough, E., & Barrett, T. (2016). The Gene Expression Omnibus database. *Statistical Genomics, 93–110.* https://doi.org/10.1007/978-1-4939-3578-9_5
4. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011
5. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202
6. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Research*, *12*(6), 996–1006. https://doi.org/10.1101/gr.229102
7. Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360. https://doi.org/10.1038/nmeth.3317
8. Kobak, D., & Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. *Nature Communications*, *10*(1), 1–14. https://doi.org/10.1038/s41467-019-13056-x
9. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352
10. Van Der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605. Retrieved from https://jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf
11. Pezzotti, N., Lelieveldt, B. P. F., van der Maaten, L., Hollt, T., Eisemann, E., & Vilanova, A. (2017). Approximated and user-steerable tSNE for progressive visual analytics. IEEE Transactions on Visualization and Computer Graphics, 23(7), 1739–1752. https://doi.org/10.1109/TVCG.2016.2570755
12. Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? Journal of Classification, 31(3), 274–295. https://doi.org/10.1007/s00357-014-9161-z