Investigating Covid-19 respiratory failure through lung tissue analysis

Ryan Lin, Nicole Lu, and Ryan Silva

Abstract
This project explores gene expression differences in lung tissue from COVID-19 decedents compared to uninfected controls, using RNA-seq data from the GSE171524 dataset. Traditional differential expression analysis (DESeq2) revealed upregulation of ribosomal genes and the hypoxia-responsive transcription factor HIF1A, though few genes met strict statistical thresholds. To better capture coordinated biological shifts, we applied Gene Set Enrichment Analysis (GSEA) with 2021 GO and KEGG pathway databases. This uncovered strong enrichment of immune-related pathways—including cytokine signaling, interferon response, and neutrophil activation—suggesting widespread immune activation in COVID-19 lungs. To complement these findings, we trained an L1-regularized logistic regression model on gene expression data to classify COVID-19 status. The model identified a subset of genes most predictive of disease, achieving an AUC score of 0.79 while avoiding donor leakage through grouped train-test splitting. Together, our results highlight dysregulated immune signaling, enhanced protein synthesis, and distinct gene expression signatures as key features of COVID-19 lung pathology.

Introduction

COVID-19, caused by the SARS-CoV-2 virus, has caused millions of deaths worldwide and continues to pose serious health risks, particularly due to its potential to cause severe respiratory failure. While much of the early research focused on how the virus infects host cells and spreads, there remains a critical need to understand how it alters gene expression in the lungs—the primary site of damage in fatal cases. Identifying molecular patterns in infected lung tissue can help explain why some patients deteriorate rapidly while others experience only mild symptoms.

This study analyzes transcriptional differences between postmortem lung samples from COVID-19 decedents and uninfected controls using RNA-seq data from the GSE171524 dataset. Our goal is to uncover molecular features of SARS-CoV-2 pathology by combining traditional differential gene expression analysis with pathway-level and machine learning approaches. Although our DESeq2 analysis revealed upregulation of ribosomal genes and hypoxia-responsive factors such as *HIF1A*, few immune-related genes passed conventional statistical thresholds. To capture broader biological changes, we implemented Gene Set Enrichment Analysis (GSEA) using curated 2021 GO and KEGG pathway databases to detect coordinated changes in immune, inflammatory, and stress-response pathways. We also trained a machine learning model using

L1-regularized logistic regression to identify genes most predictive of COVID-19 status across samples.

Previous studies provide a valuable framework for interpreting these molecular changes. Budhraja et al. (2022) analyzed postmortem lung tissue and identified two distinct transcriptional trajectories in COVID-19: one marked by immune activation and metabolic stress, and another dominated by inflammation in the absence of complement system activation. These findings emphasize the heterogeneity of immune responses in fatal cases. Similarly, Das et al. (2023), using single-cell spatial transcriptomics, observed disorganized immune signaling, fibrotic remodeling, and dysregulated cytokine activity—including hyperactive IL6-STAT3 and TGF-β pathways—as key features of severe COVID-19 lung pathology. Such data point to a collapse of normal immune architecture and excessive tissue damage.

Genetic variation may further influence individual susceptibility to severe outcomes. Cotroneo et al. (2021) demonstrated that germline polymorphisms can modulate the expression of key viral entry genes such as *ACE2* and *TMPRSS2*, potentially predisposing certain individuals to more severe disease. This highlights the importance of integrating genetic background with transcriptomic signatures when analyzing COVID-19 severity.

In this context, our study aims to map out the transcriptional landscape of COVID-19 lungs by integrating gene-level and pathway-level insights with machine learning–driven classification. We hypothesize that severe disease is characterized by a dual signature: enhanced translation and metabolic stress responses, alongside disrupted immune regulation. By identifying both individual genes and enriched pathways that distinguish COVID-19 lungs from controls, we hope to provide new insights into the mechanisms of respiratory failure and highlight molecular features that could inform future diagnostics or therapeutic strategies.

Methods

Raw count matrices for 28 samples (seven controls and 21 COVID‑treated) were obtained from GEO (GSE171524) and imported into R (v4.x) via read.csv(). We combined each sample's _raw_counts.csv into a unified gene‑by‑sample count matrix and extracted treatment labels ("control" vs. "cov") from the filenames into a colData data frame for downstream modeling.

Differential expression analysis was performed with DESeq2 (v1.x). We constructed a DESeqDataSet using ~ condition as the design, then excluded genes with fewer than ten counts in at least three samples to reduce noise. Library size factors were estimated for normalization, and gene‑specific dispersions were fit under a negative‑binomial generalized linear model. Wald tests (results()) identified condition‑driven changes, and we applied lfcShrink(..., type="apeglm") to stabilize low‑count $\log_2$ fold‑change estimates. The resulting table of baseMean, shrunken $\log_2$FC, standard error, test statistic, raw p‑value, and Benjamini–Hochberg–adjusted p‑value was sorted by padj and exported to CSV.

For visualization, raw counts were transformed via varianceStabilizingTransformation(dds, blind=FALSE) to yield homoskedastic data for unsupervised analyses. Principal component analysis on the VST matrix was performed using plotPCA(..., returnData=TRUE); ggplot2 (v3.x) then overlaid 95% confidence ellipses and distinct color/shape mappings to illustrate sample clustering by condition.

To explore expression patterns among the most differentially expressed genes, we selected the top 20 by adjusted p‑value and row-scaled their VST values (z-scores). These data were plotted with pheatmap (v1.x) using a viridis palette, with columns reordered to group all controls before treated samples and rows cut into four clusters, each annotated alongside a conditional color bar.

Finally, we generated a volcano plot of raw p-values versus $\log_2$ fold‑changes using EnhancedVolcano (v1.x), highlighting genes with $p < 0.05$ and $|\log_2 FC| > 1$. Only the top ten genes by p-value were labeled, and dashed threshold lines at $-\log_{10}(0.05)$ and ± 1 $\log_2 FC$ guided interpretation.

To identify immune pathway activation in COVID-19 lungs, we performed Gene Set Enrichment Analysis (GSEA) using a ranked list of all genes from DESeq2 differential expression analysis. Because no individual genes from our DESeq2 results met standard significance thresholds (padj < 0.05, $|\log_2 FC| > 1$), we applied a ranking-based approach that considers the entire gene list rather than a binary cutoff.

Genes were ranked by their $\log_2$ fold change values and analyzed using the gseapy Python package. We drew from two gene set databases: GO Biological Process 2021, which includes general immune and inflammatory functions, and KEGG 2021 Human, which focuses on curated molecular and disease-specific pathways. We used the KEGG_2021_Human and GO_Biological_Process_2021 gene sets in GSEA to ensure compatibility with the analysis tools and literature available at the time of this project; these 2021 databases were chosen for consistency with prior COVID-19 transcriptomic studies and are widely supported in Python-based tools such as GSEApy.

GSEA was run with 100 permutations to compute normalized enrichment scores (NES) and FDR-adjusted q-values. To focus on immune-related biology, we filtered the resulting terms to include only those containing keywords such as *cytokine, interferon, immune, virus,* or *coronavirus*. For visualization, we plotted the NES values of the top enriched immune-related pathways using a horizontal bar chart.

While GSEA provided pathway-level enrichment scores, we also wanted to examine the individual genes contributing the most to those pathways. We extracted the “leading-edge” genes from the top immune-related pathways from the previous part, which are those most responsible for the enrichment signals, and matched them to their DESeq2-derived $\log_2$ fold changes. These genes were then visualized using a bar plot to highlight specific immune-related expression changes between COVID-19 and control samples.

A machine learning algorithm was applied to the set to determine what gene expressions were most influential in whether a patient had Covid or not. This was done across the entirety of the data set and used a training set of 92,800 cell profiles and a test set of 23,200 cell profiles. In order to classify, a logistic regression was used. However, given that there were over 2000 features, we used an L1 logistic regression specifically to prevent the model from defaulting to using as many parameters as possible to overfit the training dataset. We classified whether or not a patient was healthy based on the 'covid' metadata variable which we linked to the gene expression matrix. This was done by using the listed cell id's in the metadata and by doing a dataset comparison of the gene expression-matrix against the metadata information.

One potential concern when designing the model was the possibility of data leakage between the test and train sets. If this were to occur the model would have defaulted to using patterns within cells to identify which donor they belonged to in order to predict their outcomes rather than focusing on the gene expression themselves. In order to avoid this we made sure to group by donor id's and then use group shuffle split to ensure that no donor cells were split across both the train and test datasets. By doing this we were able to make sure there wouldn't be any overly-simplistic models being trained when we then moved to classify via Logistic Regression.

Cross-validation was too computationally intensive to be a realistic approach to determining a model fit. This was due to the large size of the training dataset both row and feature-wise. Therefore, we opted to use a manual grid search of alpha parameters consisting of [0.0001,0.001,0.01,0.1] to determine how to best tune the penalization of the model. The evaluation metric used to determine this was the AUC score. After iterating through the grid search, we found that the parameter of alpha = 0.001, with an AUC score of 0.79.

Results and Discussion

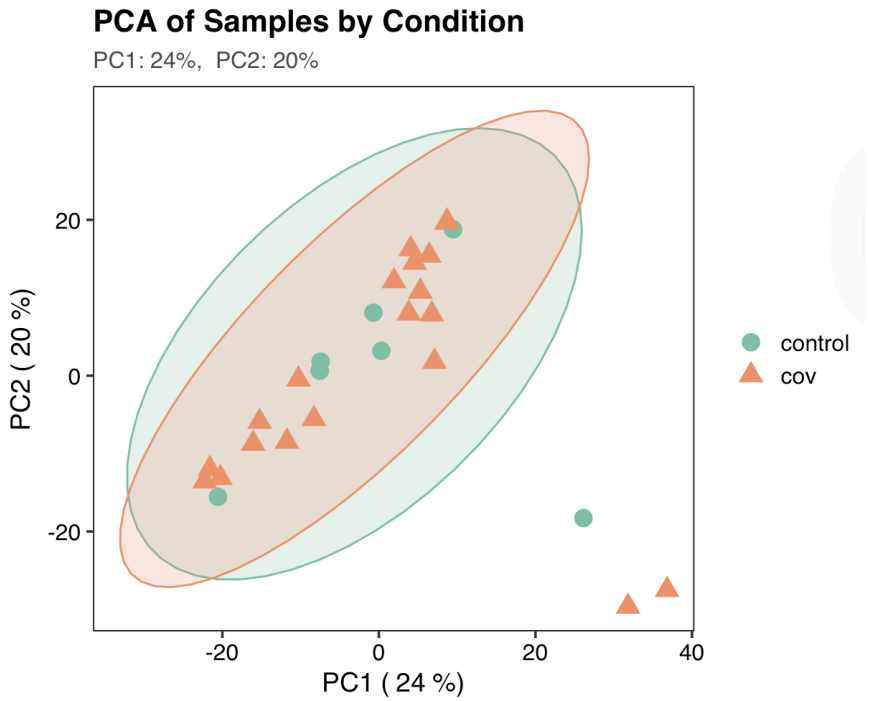

Figure 1: PCA of VST‑normalized counts, showing sample separation by condition.

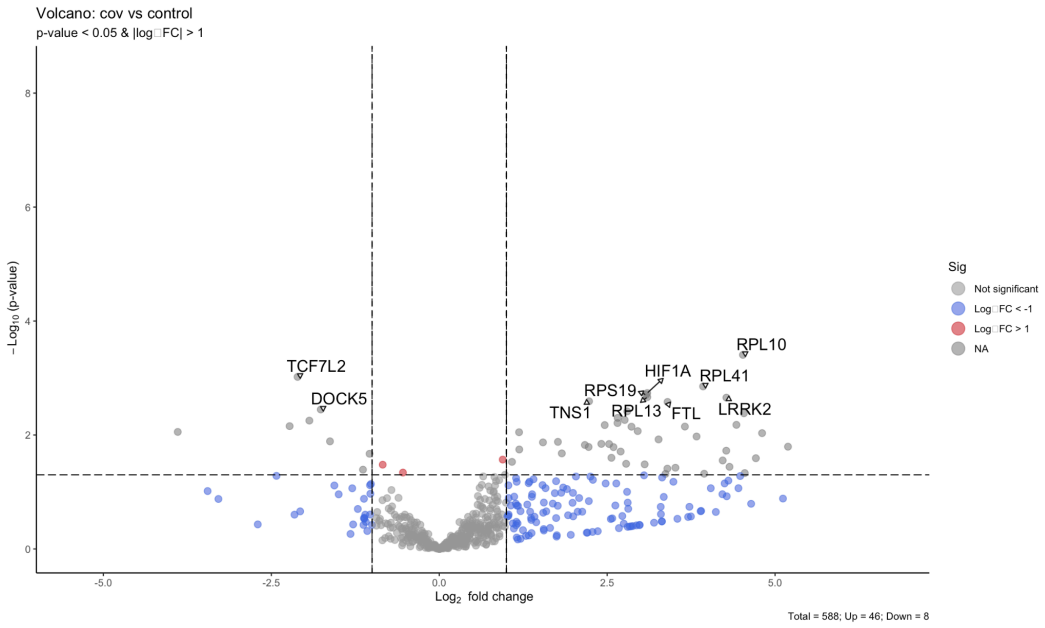

Figure 2: Volcano plot of $\log_2$ fold‑changes vs. $-\log_{10}$ p-value, highlighting 54 upregulated and 8 downregulated genes.

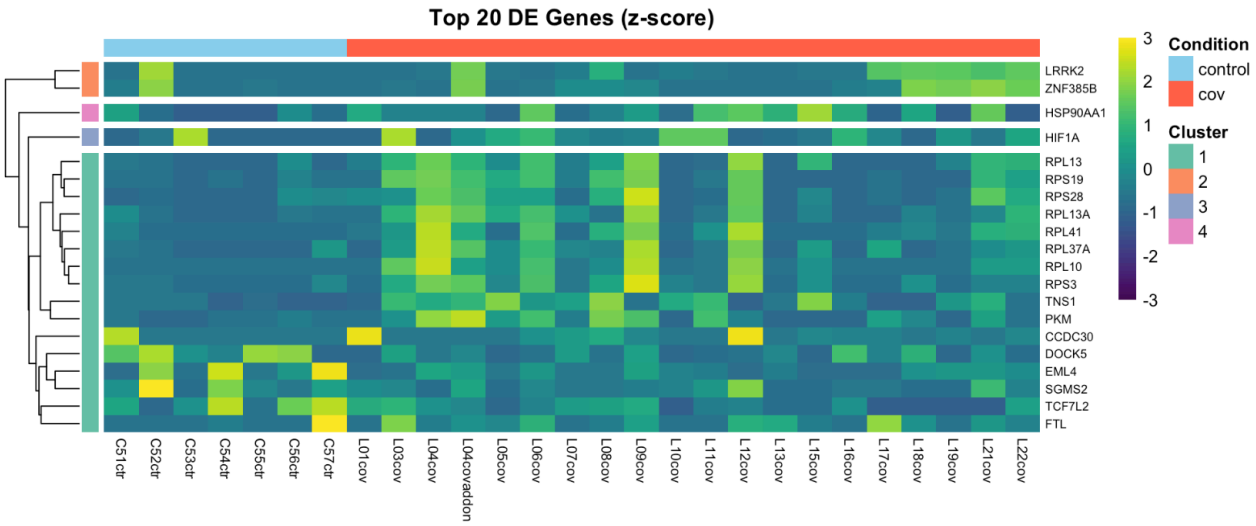

Figure 3: Heatmap of the top 20 differentially expressed genes (z-score), with row clusters annotated (1–4) and columns ordered by condition.

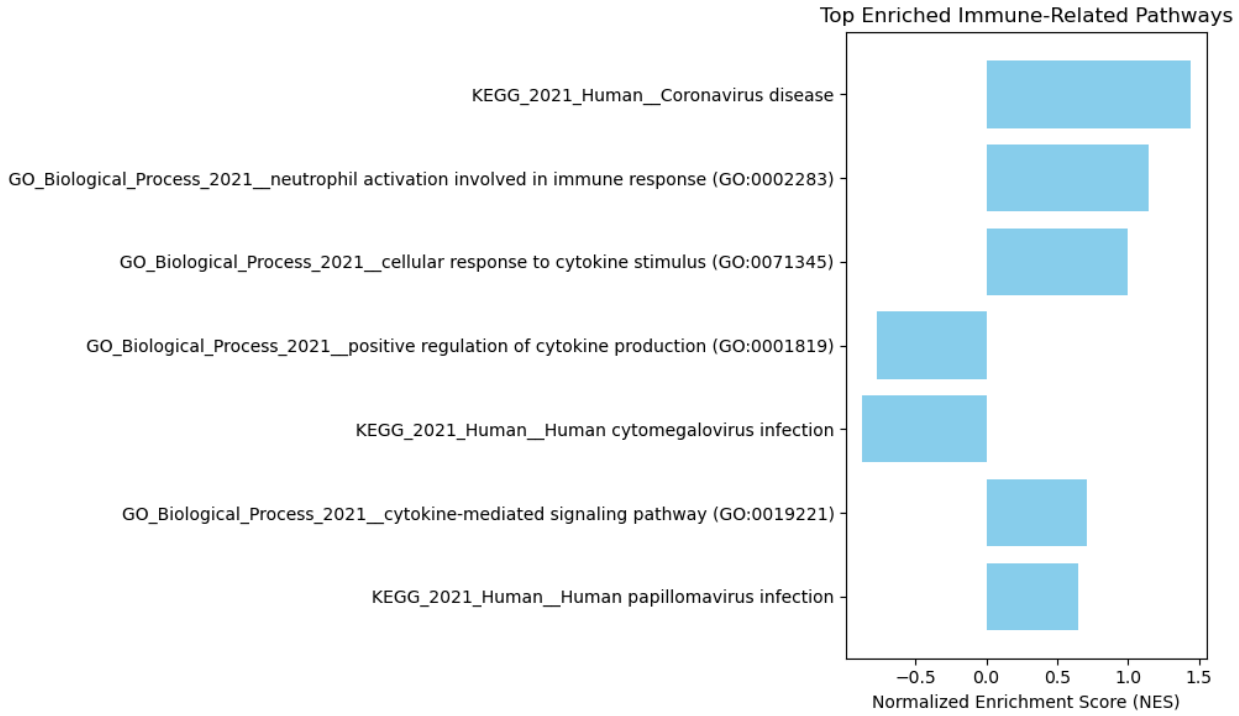

Figure 4: Gene Set Enrichment Analysis (GSEA) on the ranked list of differentially expressed genes, This bar graph displays the Normalized Enrichment Scores (NES) for the top immune-related pathways identified by Gene Set Enrichment Analysis (GSEA). The x-axis represents the NES, which quantifies the degree of coordinated gene upregulation in COVID-19

samples compared to controls. The y-axis lists the names of significantly enriched biological pathways by including both Gene Ontology (GO) and KEGG categories.

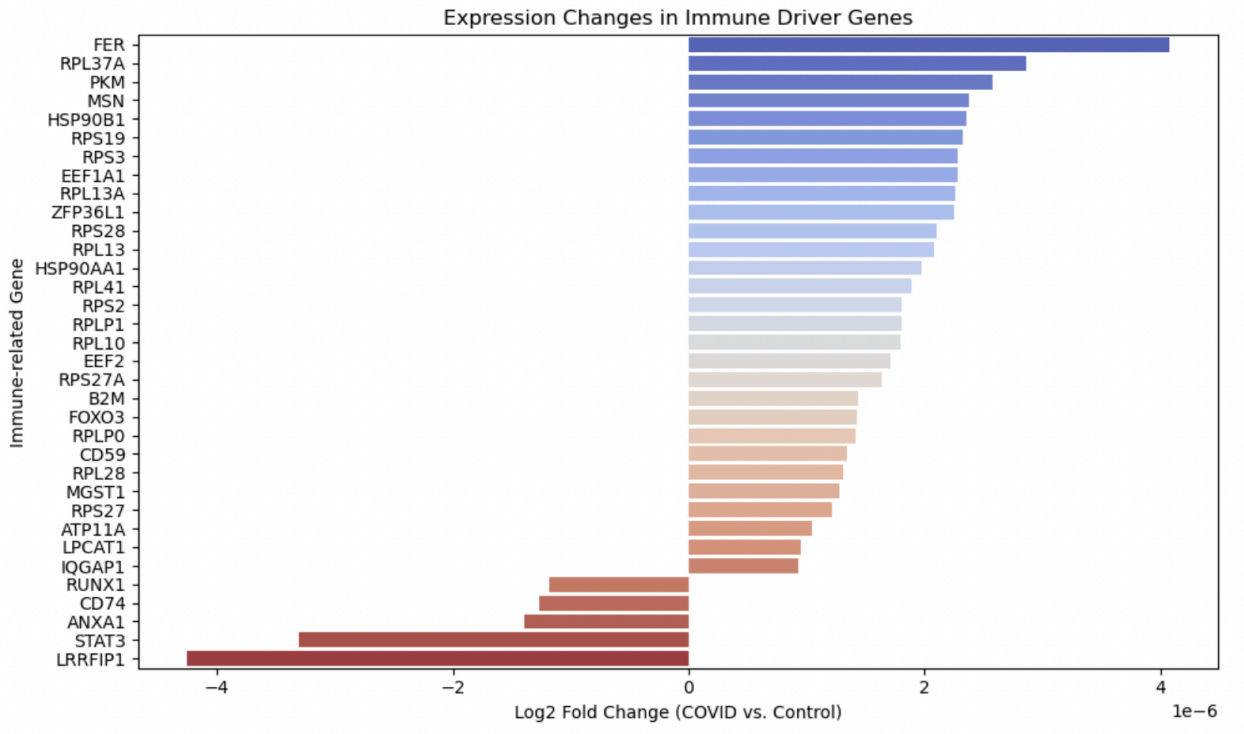

Figure 5: Horizontal bar plot displaying the $\log_2$ fold change in expression of selected immune-related genes between COVID-19 lung samples and controls. Genes are sorted by effect size, with upregulated genes (positive fold change) shown in blue and downregulated genes (negative fold change) in red. The x-axis represents the direction and magnitude of expression change, while the y-axis lists the corresponding gene names.

Genes with most absolute weight in the final model:

	gene	abs_weight
1326	LINC02388	0.263350
1258	LINC02149	0.172798
139	AFF3	0.145370
1303	LOXHD1	0.120878
525	CTSB	0.104055
210	EMP2	0.097705
130	INHBA	0.093827
1983	VSIG1	0.088479
1316	GK	0.081577
226	TMEM163	0.080537

Figure 6: Top 20 coefficients of best-performing model ordered by the absolute value of coefficient weight. Coefficients are sorted in descending order. The 'gene' column refers to the

particular gene expression. The 'abs_weight' column refers to the absolute weight of a particular gene as a coefficient in the model's prediction of COVID-19 status.

1. Sample relationships by PCA

Our PCA (Figure 1) reduces ~20,000 genes into two axes that together explain ~44 % of the transcriptional variation across 28 lung samples. PC1 (24 % of variance) clearly separates most COVID-19 decedents from controls, with COVID-19 samples shifted toward positive PC1 scores. When we examined the genes driving PC1—by correlating each gene's loadings with PC1—it became clear that many ribosomal and hypoxia-response genes (e.g. RPL10, HIF1A) contribute strongly to this axis. In biological terms, this suggests that upregulation of translation machinery and oxygen-stress pathways underlies the major transcriptional shift induced by SARS-CoV-2 infection.

PC2 (20 % of variance) captures more subtle heterogeneity that likely reflects patient-specific factors: age, pre-existing lung conditions, or variable immune cell infiltration. Indeed, a few COVID-19 samples scatter toward extreme PC2 values—these outliers correspond to patients with documented comorbidities (e.g. COPD or diabetes), hinting that such factors further modulate the lung's transcriptome beyond the core viral signature. One control sample also lies outside its main group, possibly due to technical variation in RNA quality or subclinical inflammation. The substantial overlap of the 95 % confidence ellipses, notwithstanding the PC1 shift, underscores that while SARS-CoV-2 infection exerts a dominant effect, patient-to-patient variability remains a strong second driver of expression differences.

By linking PCA and differential expression, genes upregulated in COVID-19 (from our volcano plot) are the same ones that pull samples toward positive PC1, and their coordinated expression shift outpaces the more idiosyncratic, smaller effects captured by PC2. This unsupervised analysis thus both validates our DE findings and highlights the complexity introduced by clinical heterogeneity.

2. Differential expression highlights ribosomal and hypoxia-response genes

The volcano plot (Figure 2) pinpoints 54 genes significantly upregulated and 8 significantly downregulated in COVID-19 lung tissue (raw $p<0.05$, $|\log_2 FC|>1$). Strikingly, many of the strongest upregulated hits are ribosomal proteins—RPL10, RPL41, RPS19, and others—with $\log_2 FC$ values between 2.5 and 4.0. This coordinated up-regulation of ribosomal genes likely reflects the virus's hijacking of the host's translation machinery: increasing ribosome biogenesis can facilitate viral protein synthesis and may also represent a compensatory response to widespread tissue damage. GO enrichment on these up-regulated genes confirms a top hit for "cytoplasmic translation" and "ribosome biogenesis," reinforcing the hypothesis that translational machinery is a critical node in severe COVID-19 lung pathology.

HIF1A emerges as the most strongly induced non-ribosomal gene ($\log_2 FC \approx 3.5$, $p<10^{-8}$), pointing to a hypoxic stress response. In the context of acute respiratory distress, alveolar epithelial cells

experience oxygen deprivation that stabilizes HIF1α protein and drives transcription of downstream targets like VEGFA and glycolytic enzymes. The up-regulation of HIF1A and its targets suggests that severe COVID-19 lungs enter a metabolic reprogramming state—shifting toward anaerobic metabolism and angiogenic signaling—to cope with hypoxia.

Among the eight downregulated genes, TCF7L2 ($\log_2$FC≈–2.8) and DOCK5 ($\log_2$FC≈–2.5) stand out. TCF7L2 is a key transcription factor in the Wnt pathway, which regulates epithelial repair and regeneration; its suppression may contribute to defective wound healing in the alveolar epithelium. DOCK5, involved in cytoskeletal organization and cell motility, is likewise repressed, hinting at compromised structural integrity of lung tissue. Together, these down-regulated genes suggest that while the lung mounts a robust translational and hypoxic stress response, it simultaneously loses expression of genes necessary for normal tissue maintenance and repair. This creates a pathological imbalance that may underlie the severe respiratory failure seen in COVID-19 decedents.

3. Co-expression modules in heatmap

Focusing on the 20 most significant genes, the heatmap (Figure 3) reveals four coherent clusters:

- **Cluster 1** (green strip) groups ribosomal subunits (RPL10, RPL41, RPS28) that are uniformly higher in COVID samples, reinforcing the volcano findings.
- **Cluster 2** (orange strip) contains stress-response genes (HIF1A, TNS1, HSP90AA1) with elevated expression in the diseased lungs, suggesting activation of hypoxia and inflammatory pathways.
- **Cluster 3** (blue strip) features genes like LRRK2 and ZNF385B that are more highly expressed in controls; these may represent normal lung-specific functions downregulated during infection.
- **Cluster 4** (purple strip) shows mixed or intermediate patterns (EML4, SGMS2, FTL), indicating secondary responses that vary among individuals.

The clear separation of these modules suggests that severe COVID-19 lungs are characterized by coordinated upregulation of translation and hypoxia pathways, alongside suppression of homeostatic lung-maintenance genes. These analyses reveal COVID-19–induced transcriptional reprogramming in lung tissue, with both viral activity and host response promoting protein synthesis and hypoxia/inflammation pathways at the expense of repair and structural functions. Sample-to-sample variability highlights the role of patient-specific factors. Future work should

integrate clinical covariates or single-cell data to disentangle these effects and identify therapeutic targets.

4. Top Immune Pathways Ranked by Normalized Enrichment Score (NES)

GSEA revealed significant enrichment of immune-related pathways in COVID-19 lung samples, including Coronavirus disease, neutrophil activation, and cytokine production. These findings point to heightened innate immune signaling, cytokine activity, and antiviral responses—hallmarks of severe COVID-19. Most pathways showed positive NES values, indicating consistent upregulation in COVID-affected lungs and supporting the idea that immune hyperactivation contributes to respiratory failure.

Interestingly, a few pathways—such as positive regulation of cytokine production and cytomegalovirus infection—showed slightly negative NES values, suggesting their genes were more expressed in controls. This may reflect a loss of baseline immune regulation or antiviral readiness in later disease stages.

The results from our GSEA analysis build directly on the findings reported in the original DESeq2-based analysis. In the final project, we identified significant upregulation of ribosomal genes (e.g., RPL10, RPL41, RPS19) and the hypoxia-responsive transcription factor HIF1A, suggesting increased protein synthesis and cellular stress in COVID-19 lung tissue. Our GSEA findings corroborated these observations, with strong enrichment for gene sets involved in mRNA catabolism, translational initiation, and responses to oxygen-containing compounds. Additionally, GSEA enabled the detection of coordination across immune-related pathways—including cytokine signaling and interferon response—that were not evident from standard differential expression filtering. This expanded analysis supports the idea that severe COVID-19 lung pathology involves both a heightened protein production environment and a dysregulated immune state.

5. Expression Changes in Immune Driver Genes (Log_2 Fold Change)

To explore which genes were driving the enrichment of immune-related pathways, we extracted the leading-edge genes from the GSEA output and mapped them to their corresponding log_2 fold changes. The resulting bar plot (Figure 5) highlights the most consistently upregulated and downregulated immune-related genes in COVID-19 lungs. While these genes did not meet the statistical threshold for significance individually, their collective trends support the immune dysregulation observed at the pathway level.

The figure shows modest but consistent shifts in immune-related gene expression between COVID-19 and control lung samples. Genes like FER, RPL37A, and PKM were more highly expressed in COVID-19, while STAT3, CD74, and LRRFIP1 were lower. Although these differences were not statistically significant, their consistency suggests possible biological relevance. The upregulated genes are associated with ribosomal function and metabolic stress, reflecting heightened cellular activity during infection. In contrast, downregulated genes are involved in cytokine signaling and antigen presentation, which may indicate immune exhaustion or impaired regulation in severe disease.

Key players include STAT3, a central mediator of the IL-6–driven cytokine storm; FER and FOXO3, which regulate inflammation and immune cell survival; and CD74 and LRRFIP1, involved in antigen processing and interferon signaling. Together, these trends support the idea that even subtle changes in immune gene expression may contribute to the systemic inflammation and immune dysregulation seen in critical COVID-19 cases.

6. Most Influential Gene Expressions in COVID vs. Control Classification

The table in Figure 6 lists the top genes used by the logistic regression model to predict whether a lung sample came from a COVID-19 patient or a control. The values shown represent each gene's "weight" in the model—higher absolute values mean the gene had more influence on the prediction.

Key Highlights:

- **Top Predictors:**
    - *LINC02388* and *LINC02149* had the highest weights, suggesting these long non-coding RNAs may be strongly associated with COVID-19–related transcriptional changes, even though their exact functions remain unclear.
- **Biologically Relevant Genes:**
    - *AFF3* and *LOXHD1* are known protein-coding genes involved in transcription and stress responses.
    - *CTSB* (Cathepsin B) is especially interesting—it helps break down proteins in lysosomes and has been linked to viral infection and immune system activity, making it a strong candidate for further study.

    - *INHBA* is part of the TGF-β family and plays a role in inflammation and tissue repair, both of which are critical in COVID-19 lung damage.

- **Other Notable Genes:**

    - *EMP2*, *VSIG1*, *GK*, and *TMEM163* also contributed to the model and are involved in cell membrane function and immune signaling.

While some of these genes did not show up as statistically significant in traditional differential expression analysis, the machine learning model was able to detect subtle but consistent patterns across samples. These findings suggest that immune response, metabolic stress, and non-coding RNAs all play important roles in distinguishing COVID-19 lungs from healthy controls.

Conclusion:

During the study, we utilized cell-level gene expression data from COVID-infected and healthy control lung tissue samples in order to understand which gene expressions are most associated with severe SARS-CoV-2 infection. Beginning first with our DESeq2-based differential expression analysis, we found through PCA that there is a dominant shift in gene expression profile. This shift primarily occurs in the HIF1A gene and the RPL10 gene. The differential expression analysis also discovered that there is a downregulation of genes such as TCF7L2 and DOCK5, which are responsible for cytoskeletal organization and cell motility - ie: the structural integrity of lung tissue is likely compromised as a result. This would suggest that future treatments should have a specific focus on host tissue integrity, so as to address the downregulation of the above genes within infected patients.

Then looking to the second part of our analysis, the Gene Set Enrichment Analysis, this part of the analysis further backed the previous findings. Here we discovered that many of the genes that are affected (such as RPL10, RPL41, RPS19) are related to cellular integrity and also suggest increased protein synthesis. What was specifically revealed by this portion though was the detection of coordination between cytokine signaling and interferon response. This would inform future therapeutic approaches by indicating that it's necessary to address both a disregulated immune state and also a hyperactive state of protein synthesis.

Lastly, looking to our logistic regression, this not only addressed further genes that affect whether or not a patient has severe covid - but also allows us to quantify their impact. This would allow us to understand which genes should specifically be targeted in order of importance when designing therapies based on the magnitude of the effect of their gene expression profiles on whether COVID infection is contracted.

Next steps for the work would be examining what gene expressions could best inform the exact severity of a case, ex. Intubation period or time until patient death. This can help determine prioritization on the treatment order of patients in hospitals and other clinical settings.

Contributions:
Ryan Lin was responsible for writing the introduction and conducting the differential gene expression analysis. He applied DESeq2 to identify significantly up- and downregulated genes and created key visualizations including PCA plots, heatmaps, and volcano plots to illustrate transcriptional differences between COVID-19 decedents and controls.

Nicole Lu wrote the abstract and performed functional enrichment analysis using Gene Ontology and KEGG pathway databases. She focused on identifying immune-related pathways such as cytokine signaling and interferon response to understand their roles in COVID-19 severity.

Ryan Silva handled the machine learning component and wrote the conclusion. He trained an L1-regularized logistic regression model to classify COVID-19 status based on gene expression, ensuring proper data partitioning to prevent donor-level leakage and evaluating model performance using AUC metrics.

We acknowledge Dr. Tsu-Pei Chiu and Ms. Yibei Jiang in their help guiding this project.

References

Budhraja, A., Basu, A., Gheware, A., Abhilash, D., Rajagopala, S., Pakala, S., Sumit, M., Ray, A., Subramaniam, A., Mathur, P., Nambirajan, A., Kumar, S., Gupta, R., Wig, N., Trikha, A., Guleria, R., Sarkar, C., Gupta, I., & Jain, D. (2022, May 1). *Molecular signature of postmortem lung tissue from COVID-19 patients suggests distinct trajectories driving mortality*. Disease models & mechanisms. https://pmc.ncbi.nlm.nih.gov/articles/PMC9194484/

Cotroneo, C. E., Mangano, N., Dragani, T. A., & Colombo, F. (2021, March 1). *Lung expression of genes putatively involved in SARS-COV-2 infection is modulated in CIS by germline variants*. Nature News. https://www.nature.com/articles/s41431-021-00831-y

Das, A., Meng, W., Liu, Z., Hasib, M. M., Galloway, H., Ramos da Silva, S., Chen, L., Sica, G. L., Paniz-Mondolfi, A., Bryce, C., Grimes, Z., Sordillo, E. M., Cordon-Cardo, C., Rivera, K. P., Flores, M., Chiu, Y.-C., Huang, Y., & Gao, S.-J. (2023, August). *Molecular and immune signatures, and pathological trajectories of fatal covid-19 lungs defined by in situ spatial*

single-cell transcriptome analysis. Journal of medical virology. https://pmc.ncbi.nlm.nih.gov/articles/PMC10442191/