Differential Expression Analysis on Lethal Covid -19 Single Cell RNA seq data.

INTRODUCTION:

Infection from SARS-CoV-2 pandemic led to global health crisis which resulted in severe respiratory problems in millions of people. The main aim of this analysis is to understand the lethal COVID 19 at the tissue level using single cell RNA sequencing methods. We use approximately 116000 cells from the lungs of 19 individuals that died due to COVID-19 and 7 from uninfected individuals(Control).

In this, I perform differential expression analysis to see which gene is expressed significantly in which cluster. Additionally, Go enrichment analysis is performed to understand the Biological pathways associated with these clusters. Python was used for the differential expression analysis and visualization. For this analysis, I have taken reference from the research paper “A molecular single-cell lung atlas of lethal COVID-19” and used publicly available downloaded from NCBI using the GEO accession number GSE171524.

METHODS:

Data preprocessing and Integration

The data was loaded using Scanpy. Initial filtering was applied and genes containing less than 10 cells were removed. Again, only the top 2000 most variable genes were retained for downstream analysis using Seurat v3 for noise reduction. Cells containing more than 20% mitochondria ,more than 2% ribosome content and less than 200 genes were removed from the dataset.

Doublet detection was done using scVI. Cells with confidence difference greater than 1 were removed from the dataset to ensure there are no misclassifications or distortions in downstream analysis. After preprocessing was done, all the samples were concatenated into one Ann data object. The data was saved as combined.h5ad format for future use.

Data Normalization and log transformation:

Counts were normalized to 1e4 total counts per cell and log transformed. An scVI model was further set up and trained to generate a latent representation of cells, stores in the X_scVI slot of AnnData.

Clustering and Annotation:

Nearest neighbors were calculated and dimensionality reduction using UMAP was done for visualization of clusters at a resolution 1. Differential expression analysis was done using scVI, grouping based on Leiden algorithm. Cell type annotation is done manually taking reference from similar studies and using PanglaoDB. Following the research study, I also performed differentially expressed genes between cell types AT1 and AT2 and conditions(COVID-19 and Control). Genes with absolute logFC less than 0.5 and adjusted p value greater than 0.05 were removed and others retained.

GO enrichment Analysis:
In python, GO enrichment analysis was done using gseapy function to identify the biological pathway of the significantly differentially expressed genes.

RESULTS:

The analysis identified 10 predominant cell types across the sample clusters. Significant differences in cell type composition were observed between COVID-19 and control samples. Key findings include:

- Reduction in Epithelial Cells: A marked reduction in the epithelial cell compartment, notably a loss of AT1, AT2, and airway epithelial cells in COVID-19 samples. This reduction may explain the severe lung dysfunction observed in COVID-19 patients and aligns with findings from previous studies on respiratory infections.
- Increase in Specific Cell Populations: A significant increase in fibroblasts, macrophages, and plasma cells was observed in COVID-19 samples. The increased presence of fibroblasts suggests tissue remodeling or fibrosis, a common feature in severe respiratory diseases.
- Minimal Changes in Other Cell Types: B cells, CD4+ T cells, CD8+ T cells, neuronal cells, monocytes, and pericytes did not show significant differences in abundance between COVID-19 and control samples. This stability may indicate that adaptive immune responses remain relatively unchanged in the lungs during acute COVID-19.
- Myeloid Cell Expansion: Increased expression of myeloid cells (macrophages, dendritic cells, and monocytes) was detected in COVID-19 samples. This result suggests that myeloid cells are a major source of inflammation in COVID-19, playing a critical role in the cytokine storm often observed in severe cases. Their heightened activity could contribute to hyperinflammation and subsequent lung damage.
- T Cell Expression: Mean T cell expression did not differ significantly between COVID-19 and control samples. This observation raises questions about the role of T cells in the immediate response to SARS-CoV-2 in the lungs, suggesting that localized innate immune responses may be more significant drivers of inflammation.

The result from this analysis enables to look further into the genes being highly expressed in the COVID-19 sample, which may further help us understand the long-term complications of COVID-19 survivors and provides important resource for development of therapeutic treatment specific to patient and their condition. By integrating single-cell data with clinical outcomes, researchers could establish predictive biomarkers and identify novel therapeutic targets.

CHALLENGES AND LIMITATIONS

Data Size and Computational Load: The large size of the dataset significantly increased processing time and memory usage, making data handling and analysis challenging. Loading and storing the large AnnData object required substantial computational resources and extended processing times.

Manual Annotation: Manual annotation of cell clusters was time-consuming and labor-intensive. While reference databases like PanglaoDB and previous studies were used for guidance, the process still required a significant manual effort. Employing automated annotation methods or machine learning-based classification tools in future analyses could improve efficiency.

Computational Complexity: The computational demands of scRNA-seq analysis, especially when integrating and processing large datasets, presented technical challenges. Optimizing computational workflows and leveraging high-performance computing resources would streamline analysis and reduce processing time.

Appendix:

Fig 1: Cluster formation based on grouping by Leiden Algorithm and Samples

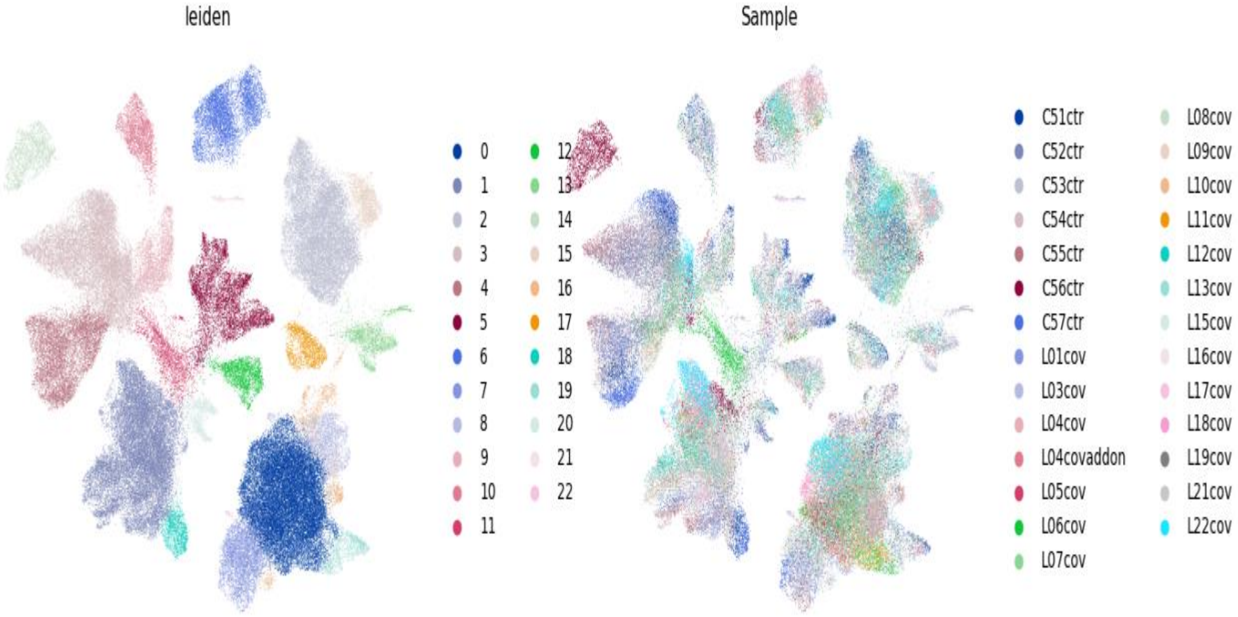


Fig 2: Clusters formation at resolution 1

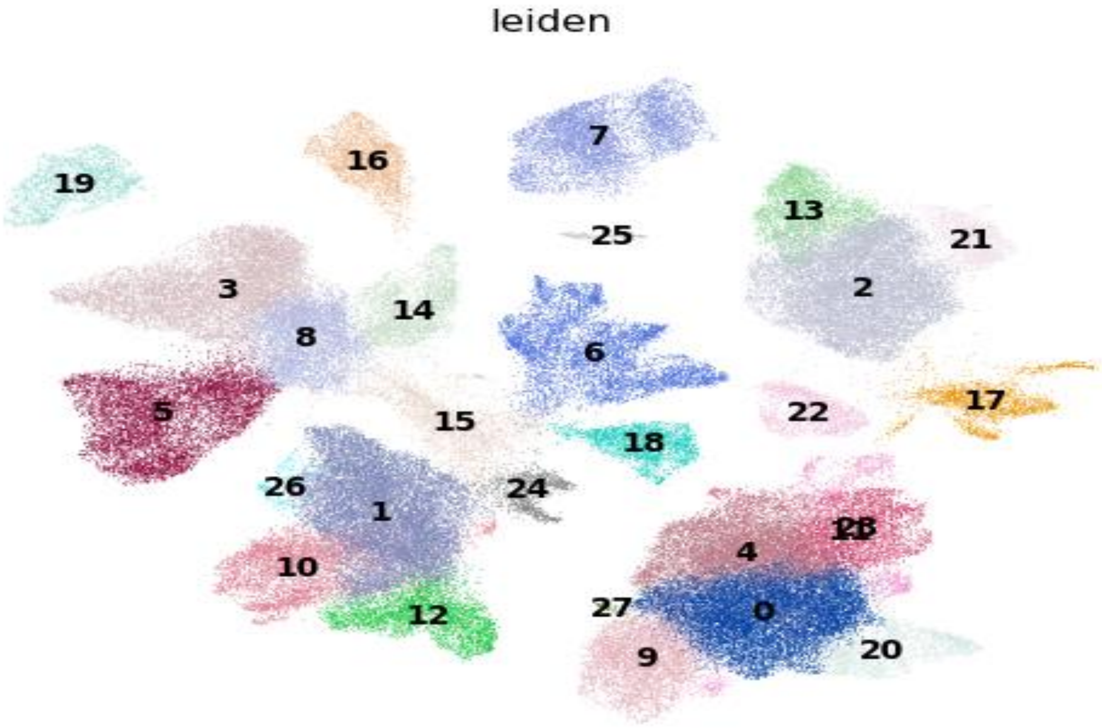
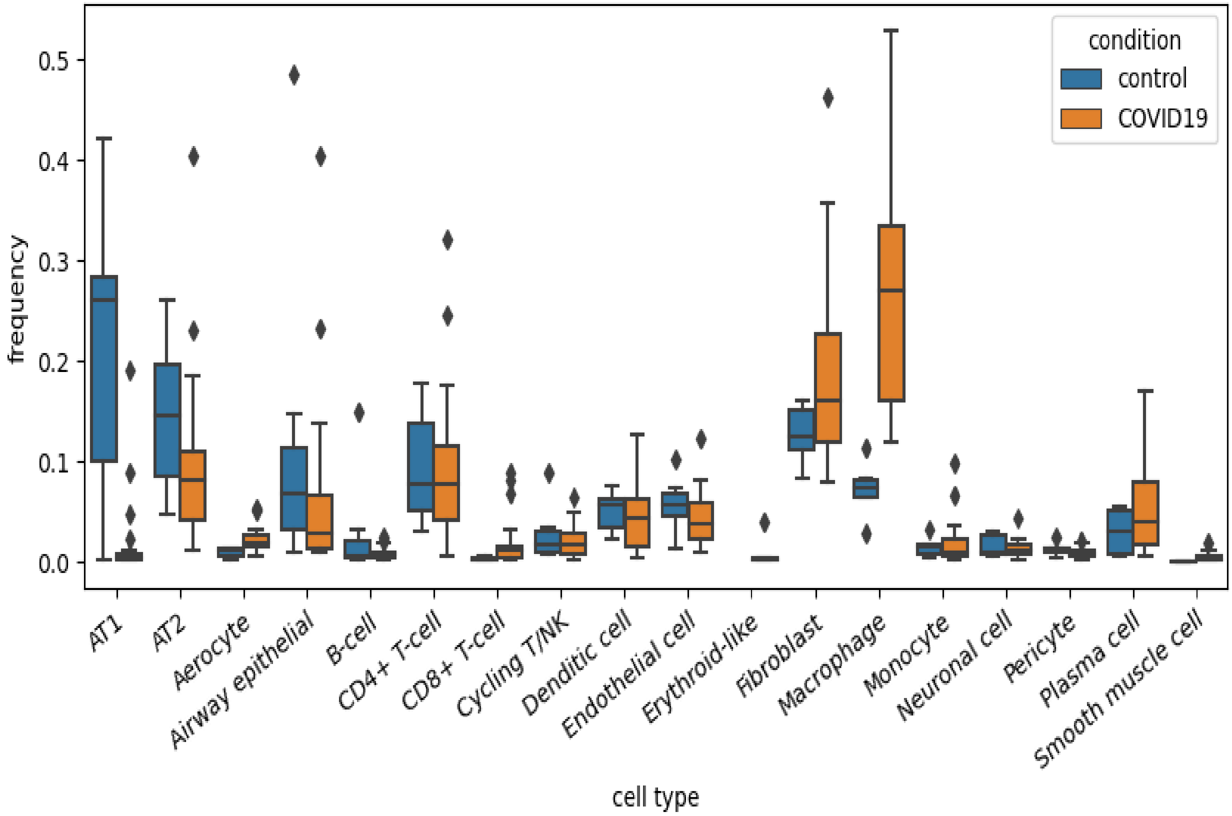


Fig 3: Boxplot representing cells expression in Control vs COVID-19 samples


condition
control
COVID19
frequency
AT1
AT2
Aerocyte
Airway epithelial
B-cell
CD4+ T-cell
CD8+ T-cell
Cycling T/NK
Denditic cell
Endothelial cell
Erythroid-like
Fibroblast
Macrophage
Monocyte
Neuronal cell
Pericyte
Plasma cell
Smooth muscle cell
cell type