**Discriminating transciptomics differences of critical COVID-19 and IPF by single cell profiling of lung tissues**

**JongHyun Kim, MS1,**MinDong Sung, MD1, Yu Rang Park, PhD1,

1 Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Korea

**Correspondence**  
Yu Rang Park, Ph.D  
Department of Biomedical Systems Informatics,  
Yonsei University College of Medicine,  
Yonsei-ro 51, Seodaemun-gu, Seoul, South Korea  
  
E-mail: yurangpark@yuhs.ac

**Abstract**

Understanding the mechanism that leads to immune dysfunction in severe COVID-19 and IPF is crucial for the development of effective treatment of pulmonary fibrosis. Here, using open published single-cell RNA sequencing data, we characterized the lung tissues from healthy controls, COVID-19, and IPF patients and cells in lung tissues. We discovered significant changes in cellular composition, transcriptional cell states using integrated analyses, revealing new insight into the biology of fibrosis lung. In both severe COVID-19 and IPF individuals, we find heterogeneous interferon stimulated gene signature and HSF1-activation pathway that appears in patients with acute respiratory failure and pulmonary fibrosis. Importantly, we found that the epithelial-mesenchymal transition (EMT) pathway related gene sets have expressed substantial amounts in severe COVID-19, which is known as a pulmonary fibrosis factor in IPF. Collectively, we provide new insight into therapeutic targets for preventing pulmonary fibrosis in severe COVID-19 infection. Suggesting that blocking pulmonary fibrosis in severe COVID-19 can also be treated with the drug used in IPF since the same related pathway are both related.

**Keywords**: pulmonary fibrosis, severe COVID-19, IPF

**Word Count**:

# Introduction

Pulmonary fibrosis (PF) means scarring of the lung and can be seen in many types of ILD. The most common idiopathic ILD is Idiopathic pulmonary fibrosis.

Idiopathic pulmonary fibrosis (IPF) is a chronic lung disease that causes permanent scarring in the distal lung. Respiratory failure and death1 are the most common outcomes of IPF, with a median survival of fewer than 5 years2. Although decades of research have been conducted to evaluate the cause of the IPF, the etiology of IPF remains still unknown, thus the development of successful treatments has lagged other fields in medicine3. Currently, only two drugs are FDA approved for the treatment of IPF which include ninetedanib4 and pirfenidone5, which are anti-fibrotic agents that consistently slow down the progression of the disease. Although the progression of the disease can be slowed by drugs, still the ultimate curative treatment is lung transplantation6.

The first reports of a new severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in Wuhan, China, surfaced in December 20197. Coronavirus-induced disease 2019 (COVID-19), a virus that causes atypical pneumonia that can lead to severe lung damage and acute respiratory distress syndrome (ARDS), expanded quickly over the global regions. Symptoms vary in severity and include ARDS, thrombosis, and organ failure like pulmonary fibrosis. Observational studies indicate that fibrosis was clinically confirmed in 71% of patients with severe COVID-19 patients8, which could lead to death. The mechanism of COVID-19 inducing pulmonary fibrosis has yet been discovered.

Our goal is to compare pulmonary fibrosis in IPF and in severe-COVID-19. Openly published scRNA-seq and snRNA-seq data was used to analysis on lung samples from individuals from COVID-19, IPF, and healthy control individuals, and identified the similarity or the difference between pulmonary fibrosis in COVID-19 and IPF patients. Especially, we focused on Macrophages, Monocytes, T cells, NK cells, AT1, AT2, and Fibroblasts have been identified as pathogenic fibrotic response regulators we are going to specify each cell type in more details9–13. Also, insight will be provided for Epithelial-Mesenchymal Transition (EMT)14 signaling pathway, which seems to be related to pulmonary fibrosis151617.

# Results

## Characteristics of the study population

Total of 61 (scRNA = 34, snRNA = 27) lung tissue samples were analyzed in our study. Here, we performed publicly available scRNA-seq on lung tissues from 10 samples with healthy controls (HC1 - HC10), 20 samples with IPF (I1 – I20), and 4 samples with severe COVID-19 (C1\_1 – C1\_4) (**Fig. 1a**, **Supplementary** **Table 1**) .

Using an integrated quality control pipeline, we generated a lung atlas that profiled 271073 nuclei (scRNA = 161990, snRNA = 109083), including 68105 (scRNA = 38570, snRNA = 32511) from control lungs, 112059(scRNA = 32511, snRNA = 79548) from COVID-19 infected lungs, and 90909(scRNA = 90909, snRNA = 0) from IPF lung tissues. (Supplementary **Table** )

The cell type assignment is reported at two degrees of granularity: major cell types and sub-cell types. Five major cell types were identified: epithelial cells, endothelial cells, immune cells, mesenchymal cells, and neuronal cells. (Supplementary Fig. 1a, b) Integrated analysis showed 32 distinct clusters and 20 different cell types at the most granular level. (**Supplementary Fig. 1c**) Initial integrated tests showed some different proportions of cell types between scRNA-seq and snRNA-seq, and since we don’t have any snRNA-seq data of IPF we performed our analysis with only scRNA-seq data first. We used uniform manifold approximation and projection to show scRNA-seq data with dimensionality reduction. (UMAP) (Fig. 1b, c) Both globally and within the immune and non-immune compartments, there were substantial variations in cell fractions between severe COVID-19 and IPF, as well as controls lung. (Fig. 1d) There was a decrease in epithelial cells due to the loss of both AT2 and AT1 cells in severe COVID-19 and IPF, as well as an increase in macrophages, macrophages, and fibroblasts. (Supplementary Fig. 2a, b) In the case of severe COVID-19, although the number of samples is smaller than that of others, there were significantly more distributions in the immune cell types.

## Macrophages

Macrophages and monocytes were re-clustered into 19 clusters to further dissect their heterogeneity and we used UMAP to show the difference between each diagnosis group. (**Supplementary Fig. 3a, b)**) The macrophages and monocytes were divided into six groups based on their expression patterns of FABP4, SPP1, FCN1 as follows: FABP4+, FABP4+FCN1+, SPP1+, FCN1+, SPP1+FCN1+, and uncertain group, according to recent classification criteria18,19. (**Fig. 2a, Supplementary Fig. 3c**) The composition and expression of FABP4, FCN1, and SPP1 in macrophages and monocytes were varied greatly by each diagnosis. (Fig. 2b, Supplementary Fig. 3c, d) Healthy controls were more likely to express FABP4, as it accounted for 45% of the total ratio, while severe COVID-19 patients with only 5.2% and IPF patients with 24.5% of the total ratio. The expression percentage of severe COVID-19 and IPF were significantly lower than that of the control group, but there was also a significant difference between COVID-19 and IPF. SPP1+ expressed high levels in severe COVID-19 as a ratio of 66.6% of the total, 25.1% in IPF, and only 9.3% in the control group. Our data show that severe COVID-19 and IPF patient lung proportions have a common distribution pattern compared to healthy controls. FCN1+ distribution was similar for all three diagnosis groups, 22.4% in control; 19.7% in COVID-19; 20.3% in IPF. However, the average expression amount of FCN1 in our datasets was significantly differed by each diagnosis lungs. As a result of calculating the average expression values for gene sets (FABP4 group = FABP4, INHBA, FCN1 group = FCN1, IL1B, SPP1 group = SPP1, MERTK), we found that, even if the distribution of each group might be similar, the expression level of each corresponding group was different. (Fig. 2d, Supplementary Table () ) Although it was confirmed that the expression of FABP4 group was less in the case of the fibrosis-related lung in the case of healthy lung, as mean expressed 1.54 in healthy control, 0.52 in severe COVID-19, and 1.15 in IPF, and it refers that expression level differed significantly by each diagnosis. FCN1 group means expressed 0.91 exp in COVID-19 and 0.77 in IPF, while 0.41 exp in control which suggests that FCN1 is the main factor that causes pulmonary fibrosis. SPP1 group, which is known as the key genes of pulmonary fibrosis as well as Group2, we found that they are more pronounced in the severe COVID-19 patient lung microenvironment than in IPF.

We performed differentially expression gene (DEG) analysis among control, COVID-19, and IPF. To find the pulmonary fibrosis-related novel genes, both up and down-regulated genes in COVID-19 and IPF lungs were selected as candidate genes. We found 186 up-regulated genes in macrophages and 147 down-regulated genes; 133 up-regulated genes in monocytes and 86 down-regulated genes. (**Table 4**) (**Fig. 2e**)

(뒤에 어떠한 유전자들이 나왔는지 정리해서 작성하기, 아래 example)

<< Ex) Both diagnosis groups had gene expression patterns suggestive of ~ . expressed the immunoregulatory genes (. ) and the profibrotic genes (. ) >>

Next, we conducted gene ontology (GO) analysis, based on the result of DEG analysis. Severe COVID-19 and IPF both up-regulated genes were related to interleukin 4, 7, and 10 pathways, which have been known to be cause pulmonary fibrosis20,21. Down-regulated genes were related to the vitamin A-related pathway, which was known that vitamin A deficiency exacerbates pulmonary fibrosis22. (Fig. 2f)

## T cell

To get insight into humoral immunity infection about severe COVID-19 and IPF in the lung, we re-clustered T cells and NK cells as the previous method. Total 18 clusters of T and NK cells were identified and we distinguished CD4+ T cells, CD8+ T cells, NK cells, NKT cells, Proliferating T cells, and T regulatory cells by manual annotation method. In terms of proportion, there were no obvious similarities between severe COVID-19 and IPF lungs. (Boxplot부분이 좀 더 표현이 좋을듯? , 아니면 barplot 둘다 보여줘도 되고 ) Proliferating T cell seems to be severe COVID-19 specific cell types as 8.1% in COVID-19 and only 2.4% in control, and 0.5% in IPF. Also, CD8+ T cells proportion was similar between the healthy controls and severe COVID-19 as proportion 40.6%, 37%, respectively, while IPF only 21.6%. However, we find that the CD4+ T cells proportion was both commonly distributed to COVID-19 and IPF as the result of 23.1%, 22.2%, respectively. (Supplementary Table3) CD4 + T cells are the only cell types that occupy a more distribution compared to the healthy control group in common in both groups.

For further exploration, we generated differentially expression gene (DEG) analysis between the control, severe COVID-19, and IPF. By performing DEG analysis we identified the novel profibrotic gene sets, which were not found in the macrophage/monocytes. In total T cells proportion, 78 gene sets were up-regulated, and 63 gene sets were down-regulated. (98, 124, 73, 200, 102, 4 up-regulated; 130, 94, 73, 419, 190, 48 down-regulated in CD8+ T cell, CD4+ T cell, NK cell, NKT cell, Treg, and Proliferating T cell, respectively.) (Fig. 3c, Supplementary Table 4)

The GO enrichment analysis of DEGs between severe COVID-19 and IPF patients shows cytokine-related pathways and HSF1 activation pathway, which is related to heat shock proteins (HSPs) cycle activation23. (Fig. 3d) To develop an HSPs signature, we used the expression of each HSPs marker gene (HSP90 = HSP90AA1, HSP90AA1, and HSP90AB1; HSp27 = HSPB1) and found that HSPs markers from severe COVID-19 and IPF lungs scored significantly higher for expression of this signature compared to the control group. (Fig. 3e) The mean expression level in ~~~

but we also found that COVID-19 patients have a higher expression than IPF patients as it’s known to be an acute disease. Overall, their data suggest that HSPs express higher in fibrotic lungs and imply the positive correlation between pulmonary fibrosis and the EMT signaling pathway.

## EMT part

We focus on lung epithelial cells and fibroblasts to discuss common pathways and processes that may be implicated in how such fibrotic changes occur within the lung of pulmonary fibrosis patients, whether the changes will persist long term or and capable of resolving, and whether post-COVID-19 pulmonary fibrosis has the potential to become progressive, as in other fibrotic lung diseases. We hypothesized that the presence of a profibrotic element could lead to an EMT pathway, in which AT2 cells are transmitted to mesenchymal cells rather than to AT1 cells.

AT1 and AT2 cells were re-clustered into 18 clusters to get an insight into the EMT pathway. (Fig. 4a) The proportion of AT1 cells in each diagnosis group was 11.3% in control, 9.8% in COVID-19, and 9.4% in IPF, however, it was confirmed that each distribution was distributed differently by each diagnosis group. ( Supplementary Table3 ). In COVID-19 lungs, however, AT1 and AT2 cell clustering were less defined, with a large proportion of cells not overlapping with their control counterparts. (Fig. 4b) To compare the full transition to AT1 cells from AT1 cells, we check the mean expression levels of CAV1, a marker of late AT1 maturation24, and was expressed significantly lower levels in AT1 cells from COVID-19 lungs as expression level 1.1, and 2.2 exp in IPF, compare to control, which expressed 3.1. (Fig. 4c , Supplementary Table5) ETV5, incomplete AT1 transition markers25, displayed decreased expression in COVID-19 and IPF. (Fig. 4d , Supplementary Table5)

We next generated EMT scoring (see method) by using the EMT hallmark gene sets between the control, COVID-19, and IPF patients to validate our hypothesis. (Fig. 4e , Supplementary Table5) Also, Gene set enrichment analysis (GSEA) of variance-ranked genes was done by using the demonstrate enrichment for the MSigDB hallmark EMT gene set in all conditions. (Fig. 4f) Both results are in line with the discovered in the previous analysis, and it adds to the evidence that these changes are linked to an EMT response. Overall, these data suggest, the presence of an EMT signaling pathway in the COVID-19.

In IPF lungs, there were considerably more fibroblasts than in COVID-19 and control lungs. For further analysis, we annotated fibroblasts into fibroblasts and myofibroblasts manually. The main difference in the fibroblasts cluster was the proportion of myofibroblasts in each diagnosis group. It was revealed that the proportion of myofibroblasts in IPF was significantly lower than the control group, and COVID-19 had a lower percentage than the control groups, as proportion 43.5% in IPF, 61.4% in COVID-19, and 66.7% in control, respectively. To check the characteristic of fibrosis, we identify the deposition of extracellular matrix proteins (ECM) in each lung by using the MsigDB Reactome ECM gene sets. A total of 76 gene sets was used for ECM scoring (see method), and the result of ECM score was 0.42 exp, 0.51 exp, and 0.58 exp at control, COVID-19, and IPF, respectively. (Fig. 4\*, Supplementary Table5) CTHRC1, COL1A1, COL3A1, markers of pathological fibroblasts which have been reported to cause pulmonary fibrosis were defined.26 The mean expression of pathological fibroblasts markers was 1.95 in COVID-19, 1.79 exp in IPF, and 1.23 exp in controls. (Fig. 4\*, Supplementary Table5) Finally, we also discovered 348 gene sets as the potential targets for pulmonary fibrosis by using DGE analysis (see method), such as COL6A3, COL5A1, COL5A2, COL18A1, COL4A1. (Supplementary Table4)

# Discussion

Here we integrated scRNA-seq analysis of lung tissues and increased our mechanistic and systemic understanding of pulmonary fibrosis by depicting unparalleled insights regarding the altered immune cell landscape in COVID-19 and IPF patient lungs. Even though this research contributed to the discovery of various facets of pulmonary fibrosis development, a comprehensive picture of the disease’s etiology has yet to emerge.

Firstly, we identified specific cell types such as Macrophages, Monocytes, T cells, NK cells, AT1, AT2, and Fibroblasts to find the interactions with pulmonary fibrosis and the difference between the adaptive immune responses and innate immune responses, and to check the presence of EMT signaling pathway, that are predicted to replenish pulmonary fibrosis in COVID-19. Secondly, our analysis about macrophages suggests that in the lung of patients with severe COVID-19 and IPF have both highly inflammatory microenvironments, even if the IPF is classified as chronic disease and COVID-19 as an acute respiratory disease. Also, the result of our analysis matches with the previous knowledge of macrophage explosive inflammatory response that could cause pulmonary fibrosis. Recent studies have shown that HSP9027–29, and HSP2730–32 are abnormally expressed in fibrotic disease, which can participate in the activation of Epithelial-Mesenchymal Transition (EMT) signaling pathways3334. The result of HSP90 and HSP27 expressed higher in fibrotic lungs, consistent with our study, which implies the relation with EMT pathway and the pulmonary fibrosis also in COVID-19.

Alveolar type I and type II (AT1, AT2) cells were discovered within the alveolar epithelial compartment. AT2 cells can proliferate and differentiate into AT1 cells, and they serve as progenitor cells in the healing of alveolar epithelial injury3536. We reveal that epithelium cells in severe COVID-19 and IPF are unable to fulfill the role of AT2 via translation into AT1 by following injury3536, but could result in apoptosis and favoring excessive deposition of extracellular matrix (ECM), which leads to pulmonary fibrosis37–40. Previously, several studies have reported that response to injury in the epithelium and the activation of TGF-β causes fibroblast proliferation41–43, transdifferentiating to a contractile myofibroblast phenotype44–47, and induces the production and deposition of ECM proteins4148. We report a decrease of myofibroblasts in both fibrosis lungs, but since the COVID-19 is not a chronic disease like IPF, the decrease of its proportion seems to be slightly less.

Even if our study provides insight into responses to pulmonary fibrosis lungs, the size of the COVID-19 tissue samples has limited the study. However, through validation with snRNA-seq data from Benjamin Izar25 study, we could verify our research. By using integrated data set (scRNA-seq & snRNA-seq) we performed the trajectory analysis to check whether the results of the EMT pathway were different for each diagnosis group and found that transition to fibroblasts was being more done in severe COVID-19 and IPF patient lungs compared to the healthy control lungs, (Supplementary Fig. \*) which aided in the validation of our hypothesis that pulmonary fibrosis is related with EMT pathway in COVID-19 as well as IPF.

In conclusion, we fully characterized the disturbed immunological landscapes between COVID-19 and IPF from the infected lungs. This study reveals molecular mechanisms implicated in COVID-19 immunopathogenesis and identifies EMT functional roles in pulmonary fibrosis. A related pathway has been discovered in both fibrosis lungs and it reveals potential therapeutic such as drug of IPF can also be applied to patients with severe COVID-19 who is suffering from pulmonary fibrosis.

# Materials and methods

## Patients and Data Collection

The scRNA-seq analysis for lung tissues of healthy controls and IPF were downloaded from GEO49 under the accession number GSE13589350. The downloaded scRNA-seq data includes 10 healthy donors, which were generated using 3’V2 chemistry kit on Chromium Single-cell controller(10Xgenomics). Severe COVID-19 lung scRNA-seq data were obtained from the GEO database under the series number GSE14987851, which includes data from 4 parts of lung tissues from one patient. In addition to data for 6 healthy donors and 20 severe COVID-19 patient lung tissue, snRNA-seq data were acquired from the GEO database under the accession code GSE17152425. Using 3’V3.1 Chromium Single-cell controller(10XGenomics), a total of 15,000 ~ 20,000 nuclei were loaded per channel on a Chromium controller.

## Data processing

## Quality control

The resulting expression matrices were processed respectively in R (v.4.1.0) using Seurat(v.4.0.4)52. Nuclei with 200-7,500 genes, 400-40,000 unique molecular identifiers (UMIs), and fewer than 10% mitochondrial reads were kept using filters. After filtering, samples with fewer than 1,000 nuclei were removed from further analysis. The filtered matrices were normalized using “NormalizeData” function with a default parameter. The “vst” method in “FindVariableFeatures” was used to find the top 2,000 variable genes. The “ScaleData” function was used to scale and center gene expression matrices. Next, principal component analysis (PCA) was then subjected to a default parameter before integration.

## Integrate of each sample

The three datasets were integrated using the canonical correlation analysis (CCA) method53 in Seurat. The “FindIntegrationAnchors” function was applied to find a set of anchors between a list of two datasets, which can later be used to integrate the objects using the “IntegrateData” function. Next, the “IntegrateData” function was used to integrate two datasets using the pre-computed anchor sets we applied before. The integrated dataset was scaled with “ScaleData” function, then PCA and UMAP dimension reduction were conducted using the top 30 main components. To build nearest-neighbor graphs we used “FindNeighbors” function and the top 30 dimensions of the PCA reduction were used. Last, the clustering was applied with a resolution of 0.8.

## Cell type annotation

The cell types were first subdivided into four groups manually. PTPRC+ (immune cells), EPCAM+ (epithelial cells), PECAM1+/PTPRC- (endothelial cells), PTPRC-/EPCAM-/PECAM1- (mesenchymal cells), and SEMA3A/PKHD1L1 (neuronal cells) were manually annotated each cell type by using “FeaturePlot”. We performed scaling, PCA, UMAP dimension reduction, clustering on each subdivided data to annotate each specific cell type. Next, the “FindAllMarkers” function was used to annotate the specific cell types by performing differential gene expression analysis with a minimum fraction of 25% and a log-transformed fold change criterion of 0.25. The resulting clusters were annotated manually by using cell type-specific markers. (Table) Signature markers to identify subdivided groups were obtained from past research50, and each specific cell type marker from other research here25,26,50,54.

## Subset Specific cell types

Each cell types were subset by a value in the Seurat object metadata. The “integrated” assay was reanalyzed using the Seurat workflow, which includes “ScaleData”, “RunPCA”, “FindNeighbors”, “FindClusters”, “RunUMAP” function. The top 30 dimensions were calculated, and clustering with a resolution of 0.6 was used.

## Differential gene expression analysis

Seruat’s “FindAllMarkers” MAST55 function was used to perform gene expression among control groups. The log fold change (FC) cutoff was set to 0.25, and the p-value less than 0.01 was considered differentially expressed. To find pulmonary fibrosis-related novel genes in severe COVID-19 lung, DGE analysis among control, COVID-19, and IPF lung was done. We first selected up and down-regulated genes in COVID-19 compared to control and IPF compared to control, respectively, using the MAST method. Next, only the common up/down-regulated genes from COVID-19 and IPF results were considered for further pathway enrichment analysis. To identify every gene in every cell type, we took each annotated cell type independently, for all cell types which we annotated. The DEG results are summarized in Supplementary Table (Table4).

## Pathway enrichment analysis

Pathway enrichment was performed using REACTOME(v.3.7)56 pathway browser for Homo sapiens. REACTOME pathways are highly hierarchical, with numerous small, highly specialized routes at lower levels. Differentially expressed gene sets that were both up/down-regulated in severe COVID-19 and IPF were selected as input. To select the gene sets to use in REACTOME pathway analysis, we first filtered the result of DGE. Only the genes which logFC > 0.5, and the p-value less than 0.001 were selected. We also found that data with scRNA-seq and snRNA-seq mixed data significantly reduce the result of value when we validate our result, so we only used the scRNA-seq datasets to conduct the analysis.

## Trajectory analysis

The pseudotime cell trajectory analysis was used by the R package Slingshot57. To identify the cell trajectory, the slingshot function was used with the UMAP dimensionality reduction and cluster labels as in Seurat objects. Slingshot trajectories are computed with a start point and endpoint for each trajectory. Several trajectories were displayed on the same graph and were performed to control, severe COVID-19, and IPF to each group separately.

## Gene set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA)58 was performed with the R package fgsea59. The name “HALLMARK\_EPITHELIAL\_MESENCHYMAL\_TRANSITION” in MsigDB60 was used as reference gene sets and analysis. The variance values after the “vst” were calculated using Seurat “FindVariableFeatures” function, and the corrected p-value from the differential expression analysis was used to rank the input genes.

## Gene set scoring

Gene set scoring of EMT hallmark gene set and Reactome pathway “ECM proteoglycans” was performed using the Seurat “FetchData” function. Only genes matching the corresponding gene sets were selected from our data, and. The average expression of each gene was compared for each diagnosis group. The other genes we wanted to identify in the study were also compared through this method.