Single-cell renalysis of two severe COVID-19 patients and in-depth characterization of monocyte-associated severe-stage clusters

Michał Stanowski
Faculty of Mathematics, Informatics and Mathematics, University of Warsaw
ms439001@students.mimuw.edu.pl

June 25, 2025

Abstract

A reanalysis of single-cell RNA-sequencing data from two patients with severe-stage COVID-19 was performed to better characterize a monocyte cluster previously associated with disease severity. The original study identified this cluster but did not investigate it in depth or compare it with other viral infections. In the present analysis, alternative preprocessing parameters were applied, and cluster reproducibility was confirmed. Subclustering revealed limited heterogeneity, with specific transcripts such as AP001189.1 and DDR2 indicating potential functional diversity among monocytes. Integration with external datasets showed that the cluster remained distinct when combined with influenza A data, suggesting disease specificity. However, the cluster did not appear in an independent COVID-19 dataset, raising concerns about its generalizability. Additional transcriptomic differences were identified in B cells across conditions. These findings highlight the analytical sensitivity of single-cell clustering and suggest that previously reported immune signatures may reflect cohort-specific rather than universal patterns. Broader datasets are needed for robust characterization.

Introduction

COVID-19 has affected the vast majority of the global population, either directly or indirectly. Its impact extends beyond medical consequences, such as the estimated 5,42 million deaths (2023) [1], to substantial economic disruptions [2]. The pandemic also demonstrated an unprecedented mobilization within the scientific community, leading to the rapid development of vaccines and a thorough molecular characterization of the virus itself. Studies investigating the disease progression followed shortly thereafter. Within months of the initial cases reported in China, researchers published an analysis of two patients, tracking disease progression through sampling during both the severe phase and remission [3]. They identified a monocyte cluster specifically associated with the severe stage of COVID-19. Although this cluster was briefly characterized, the primary aim of their study was to demonstrate changes in the immune cell landscape in response to tocilizumab treatment, so the monocyte cluster was not explored in depth. The authors also compared this monocyte cluster to clusters derived from sepsis patient samples but did not observe a similar cluster, leading them to conclude that it was unique to severe COVID-19. However, considering the very limited sample size, the superficial characterization of this cluster, and the absence of comparisons with other viral diseases, it was decided to reanalyze the data to provide a more comprehensive characterization of this monocyte cluster.

Methods and Materials

Computational Environment

All computations were performed locally using RStudio (version 4.4.2). The following R packages were used: Seurat (v. 5.3.0), reticulate (v. 1.42.0), ggplot2 (v. 3.5.2), dplyr (v. 1.1.4), GEOquery (v. 2.74.0), readxl (v. 1.4.5), org.Hs.eg.db (v. 3.20.0), AnnotationDbi (v. 1.0.4), purrr (v. 1.0.4), tibble (v. 3.2.1), vegan (v. 2.7.1.) and pheatmap (v. 1.0.13). Analyses were conducted on a local machine equipped with an AMD Ryzen 9 8945HS processor with Radeon 780M Graphics (16 CPUs), 32 GB RAM, and running Windows 11 Home (64-bit). Assistance with complex code implementation and challenging translations was provided using the ChatGPT language model [4].

Data Sources

All datasets were obtained from the Gene Expression Omnibus (GEO) database. The data were downloaded locally. Although the GEOquery package was initially used to retrieve the data, further analysis of the resulting data frame proved problematic. The control dataset used to obtained the same results can be found at this website: `https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.1.0/5k_pbmc_NGSC3_aggr?`

Code availability

The R code used to perform all analyses described in this study, as well as to generate the included figures, is available in the repository at the following link: `https://github.com/michalstanowski/scRNA_covid`.

Methodology

Obtaining the same results

In the original study, the authors used Cell Ranger (v. 3.1.0) to process raw scRNA-seq data. In this study, all analyses were performed exclusively in RStudio, primarily using the Seurat package. Only transcripts expressed in more than 3 cells were retained, and cells expressing fewer than 200 transcripts were removed. Following the filtering criteria described in the original paper, for the control dataset, cells with more than 500 and fewer than 600 detected transcripts were selected. For the experimental dataset, cells with more than 300 and fewer than 500 transcripts were retained. To remove potential doublets using the same approach as in the original study, Scrublet was employed via the reticulate package, which enabled the use of

Python code within the R environment. In total, 132 doublets were identified in the experimental dataset and 1,096 in the control dataset. These results were considered acceptable compared to the reference values from the original study (50 and 997, respectively). Subsequently, the data were normalized using the NormalizeData function. Highly variable genes were identified using the ”vst” (variance stabilizing transformation) method, and the data were scaled using ScaleData. Principal component analysis (PCA) was performed on both datasets with 40 principal components. The datasets were then integrated using FindIntegrationAnchors and IntegrateData. The resulting integrated object was further analyzed in Seurat using RunUMAP, FindNeighbors, and FindClusters. The final outcome of this methodological pipeline was a DimPlot visualization illustrating the clustering of cells.

Subsequently, all cluster-specific marker genes were identified using the FindAllMarkers function. A ViolinPlot was generated to visualize the expression of 16 marker genes previously used in the original study to assign clusters to specific cell types. Based on this visualization, individual clusters were annotated with corresponding cell type identities. As a result, a DimPlot was created, now displaying clusters labeled with cell type annotations. Using the available metadata, it was also possible to assign a disease status (e.g., severe, remission, control) to each cell.

An additional analysis was performed using an alternative mitochondrial gene expression cutoff, based on VlnPlot visualizations. For the control dataset, a threshold of 25% was applied, while for the experimental dataset, a threshold of 15% was used. The analysis of this dataset followed the same workflow as described above. To compare the UMAP visualizations obtained from the two analysis variants, the procrustes function from the vegan package was used, allowing for alignment of the embeddings and reduction of geometric variability between them.

Identification of monocyte cluster and analysis of its subclusters

A monocyte cluster characteristic of the severe disease state was visually identified, as these cells formed a distinct and well-separated cluster from the rest. To determine the optimal number of subclusters, an elbow plot was generated for each value of k from 1 to 15. The chosen k corresponded to the inflection point of the curve. However, in this case, the curve was relatively smooth, making the inflection point ambiguous; therefore, k = 3 was selected arbitrarily. The clusters were then obtained using the K-means clustering method.

Comparison to other datasets

An attempt was made to identify sufficiently large datasets to enable comparison between the studied dataset and another viral disease. The goal was to determine whether the monocyte cluster specific to the severe stage of COVID-19 is unique to this disease or also present in other viral infections, as well as to verify whether additional PBMC data from COVID-19 patients confirm the existence of this severe-stage-specific cluster. The viral infection chosen for comparison was caused by the influenza A virus. The corresponding dataset was obtained from the GEO database under the accession number GSE24629. The COVID-19 dataset used for comparison was also retrieved from GEO, with accession number GSE188172. The comparison with these datasets was not limited to UMAP visualization; it also included identification of the most differentially expressed genes, which were subsequently visualized using pheatmaps and DotPlots.

Results

The clustering results successfully recapitulated the cell populations identified in the original study (Fig. 1A). Notably, the monocyte cluster (containing 1002 cells) characteristic of the severe stage was reliably reproduced (Fig. 1B). Similarly, when the analysis was repeated using alternative mitochondrial gene expression thresholds, comparable clusters were obtained, with the monocyte cluster (1036 cells) still predominantly composed of cells from the severe-stage group (Fig. 1C). K-means subclustering visualization of the monocyte population, along with ViolinPlots of the most highly expressed genes, is shown in Fig. 1D. Furthermore, integration with additional datasets was successfully performed, including one derived from patients infected with influenza A virus and another independent COVID-19 dataset. The UMAP visualization of the integrated data with PBMC samples from influenza A-infected patients is shown in Figure 2A. A pairwise differential gene expression analysis was conducted to compare control samples, PBMCs from influenza-infected individuals, and COVID-19 samples in both remission and severe stages. The results of this analysis are presented in Figure 2B. Additionally, Figure 2C displays a DotPlot highlighting the top differentially expressed genes between the monocyte cluster characteristic of severe-stage COVID-19 and other monocyte clusters identified following integration with influenza A-infected patients dataset.

Discussion

Obtaining the same results

Unfortunately, the authors of the original publication did not provide the code that would allow mapping functions to directly compare the UMAP visualizations obtained in this analysis with those presented in the study. However, the clusters appear to be visually similar to those reported in the original work. Notably, the monocyte cluster characteristic of the severe stage of the disease is clearly visible. Interestingly, this cluster appears to be slightly less compact and consists of a slightly smaller proportion of severe-stage cells (89.96% vs. 90.12%) when the analysis is conducted using an alternative, and arguably more reasonable, threshold for mitochondrial gene expression based on ViolinPlot inspection. This highlights the first limitation of this cluster: even a seemingly minor change in one analytical parameter can noticeably impact the cluster's composition. As shown in Figure 1C, the severe-stage monocyte cluster does not exhibit the same degree of compactness as in Figure 1B. While the cluster still exists, severe-stage cells appear more intermixed with those from the remission phase. Furthermore, the overall UMAP clustering pattern is visibly different from the original. After extracting the embeddings and comparing the two clusterings, a Pearson correlation of 0.47 for the X-axis and -0.44 for the Y-axis was observed, and an average Euclidean distance between the two UMAP embeddings of 11.36. This indicates moderate agreement along the X-axis and similar agreement on the Y-axis, with evidence of rotation, as seen in Fig. 1B-1C. The Procrustes RMSE analysis further confirms this: it reveals that more than half of the observed differences between the UMAPs result from geometric transformations—rotation, scaling, and translation — rather than underlying biological differences. Therefore, while the severe-stage monocyte cluster may appear visually less cohesive, all quantitative analyses suggest that it is highly similar to the one observed in the original study. The overall clustering structure is preserved across UMAP embeddings, with differences primarily reflecting spatial rearrangements rather than changes in biological identity, as illustrated in Figures 1B and 1C.

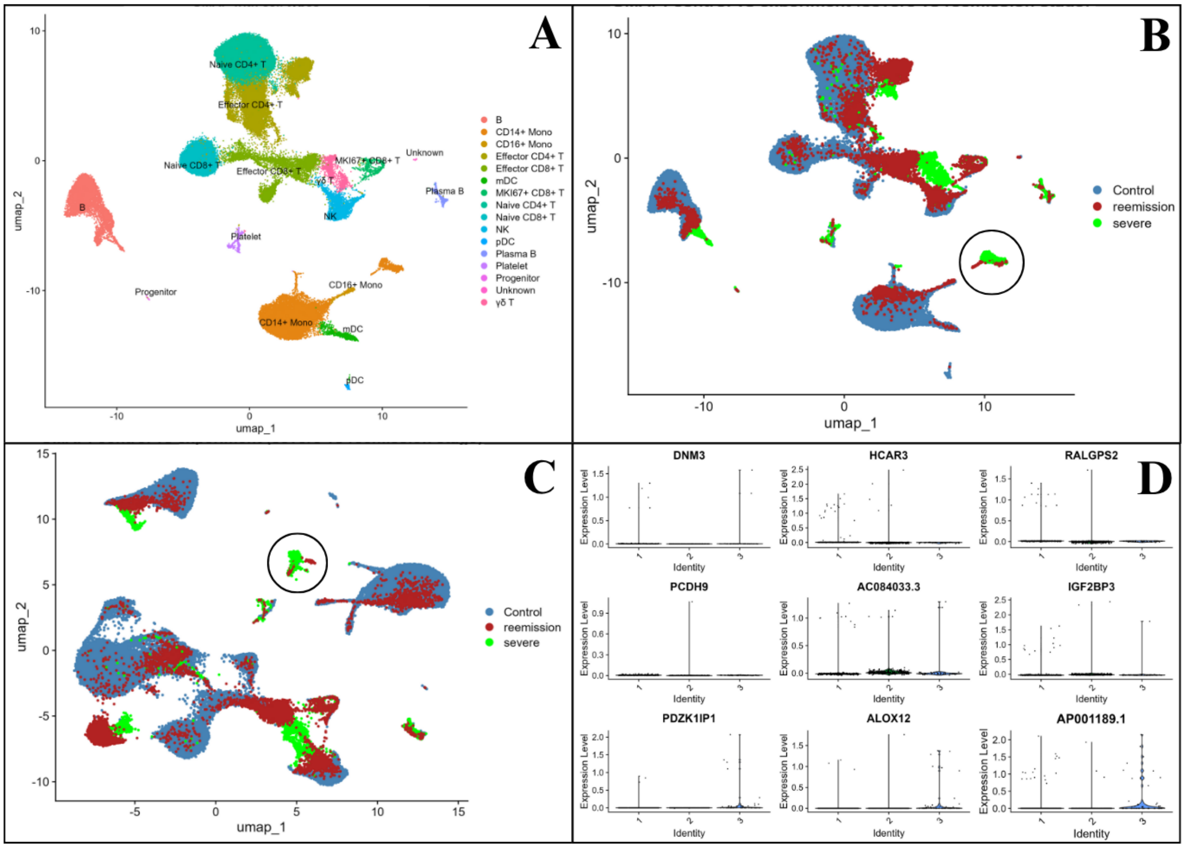

Figure 1: Clustering analysis recapitulates original cell populations and identifies a monocyte cluster associated with severe disease. (A) Overview of cell populations identified by clustering, matching those in the original study. (B) Monocyte cluster characteristic of the severe stage (1002 cells) reliably reproduced and marked with a circle. (C) Clustering results using alternative mitochondrial gene expression thresholds, showing a comparable monocyte cluster (1036 cells) predominantly from the severe-stage group (marked with a circle). (D) K-means subclustering of the severe-stage specific monocyte population with ViolinPlots of the top expressed genes.

Severe-stage monocyte subclusters analysis

Fig. 1D presents the nine most differentially expressed genes between the clusters, indicating that this cluster appears relatively homogeneous. Notably, K-means subclustering (not shown) did not reveal meaningful structure, with cluster 3 distinguished mainly by a single transcript — AP001189.1. This long non-coding RNA (lncRNA) is known as a pyroptosis-related lncRNA (PRlncRNA) and is part of a five-lncRNA prognostic signature in lung squamous cell carcinoma [5]. This finding is intriguing, given that pyroptosis has repeatedly been implicated in SARS-CoV-2 infection [6]. Pyroptosis is a form of programmed cell death mediated by gasdermin proteins, leading to the lytic destruction of infected cells to expose viral particles to immune detection. Could macrophages, particularly those associated with the severe stage of disease, play a role here? Several studies suggest that macrophages themselves undergo pyroptosis, releasing proinflammatory cytokines [7]. Thus, this finding may be biologically meaningful — macrophages expressing pyroptosis-related lncRNAs might self-destruct as a defensive mechanism against viral infection. Previous research has indeed reported overexpression of dozens of lncRNAs in COVID-19 [8]. Although expression of AP001189.1 in cluster 3 is modest, its elevation compared to other clusters warrants further investigation, ideally in a larger cohort.

What is particularly worth highlighting is that, upon adjusting the mitochondrial gene expression cutoff thresholds, the monocyte cluster — although still highly similar to the previously identified one — exhibited distinct characteristics when subjected to K-means clustering (visualization not shown). At k=3, two subclusters demonstrated significantly elevated expression of two transcripts: AC025569.1 and DDR2. Visually, K-means clustering yielded more clearly separated cell groups than in the analysis with the original parameters. The first transcript, AC025569.1, is a long non-coding RNA (lncRNA) that has been identified as one of lncRNAs with prognostic potential in lung adenocarcinoma [9]. However, no information was found suggesting its involvement in pyroptosis or other immune-related processes. The second transcript, DDR2, is far more extensively studied. DDR2 is expressed in macrophages and modulates their polarization — specifically promoting the transition from the pro-inflammatory M1 phenotype to the anti-inflammatory M2 phenotype — thereby exerting a regulatory effect on inflammation. This is particularly relevant in the context of severe-stage COVID-19, where inflammatory modulation is critical [10]. These findings raise an important question: does the apparent homogeneity of the monocyte subcluster reported in the original study reflect a biologically meaningful structure, or is it unstable and sensitive to analysis parameters that might seem trivial? These results suggest the latter may be true, as altering a single threshold can change the functional characterization of the cluster. It is thus likely that the true nature of this cluster is functional heterogeneity, representing diverse macrophage subpopulations with distinct immunological roles. Further investigation with a broader range of parameters and datasets is warranted to assess the stability of this cluster and better understand the functional spectrum of the monocytes it contains.

Comparison with other datasets

In the original publication, the authors compared the analyzed dataset to data derived from patients with sepsis. Based on this integration, they concluded that the observed monocyte cluster is specific to COVID-19. However, this is a bold conclusion, considering the data was obtained from only two patients and compared with a single external dataset. In the present study, a broader comparative approach was taken by integrating the COVID-19 dataset with data from patients infected with influenza A virus, as well as with an independent COVID-19 dataset from patients in the severe stage of the disease.

Fig. 2A illustrates that the monocyte cluster remained clearly distinct from other clusters, even after integration with PBMC data from influenza-infected patients. Notably, this cluster exclusively consisted of cells originating from the severe-stage COVID-19 group, further supporting the original hypothesis of its COVID-19 specificity. Interestingly, a separate cluster comprising B lymphocytes displayed a distinct behavior. As shown in Fig. 2B, pairwise differential gene expression analysis revealed stark differences between B cell profiles in the context of influenza and COVID-19. Surprisingly, the expression patterns of B cells in COVID-19 closely resembled those of healthy controls, which contrasts with the distinct profile observed in B cells from influenza-infected individuals. Some genes exhibited elevated expression specifically in COVID-19 B cells compared to both control and influenza groups. For example, NBPF20, a gene typically not expressed at meaningful levels in B lymphocytes according to the Human Protein Atlas, showed unexpected upregulation [11]. NBPF20 is primarily associated with neurogenesis and brain development, making its expression in this context puzzling and potentially incidental. Another gene, HES1, was also selectively upregulated in COVID-19-associated B cells. HES1 encodes a transcriptional repressor that inhibits the differentiation of mature B cells. On the other hand, certain genes were uniquely upregulated in

B cells from influenza A-infected patients, such as NIBAN3 and RBFOX2. NIBAN3 encodes a protein involved in cellular stress responses, which aligns with the epithelial cell damage induced by influenza A and its associated oxidative and ER stress. RBFOX2, known for its role in regulating alternative splicing in B cells, may reflect the strong transcriptional rewiring triggered by the influenza virus [12]. Despite the virological similarities between influenza A and SARS-CoV-2, the transcriptional landscapes of B lymphocytes in response to these pathogens diverge substantially. The underlying reasons for these differences likely involve distinct viral tropism, immune activation patterns, and signaling cascades, although some mechanisms remain speculative.

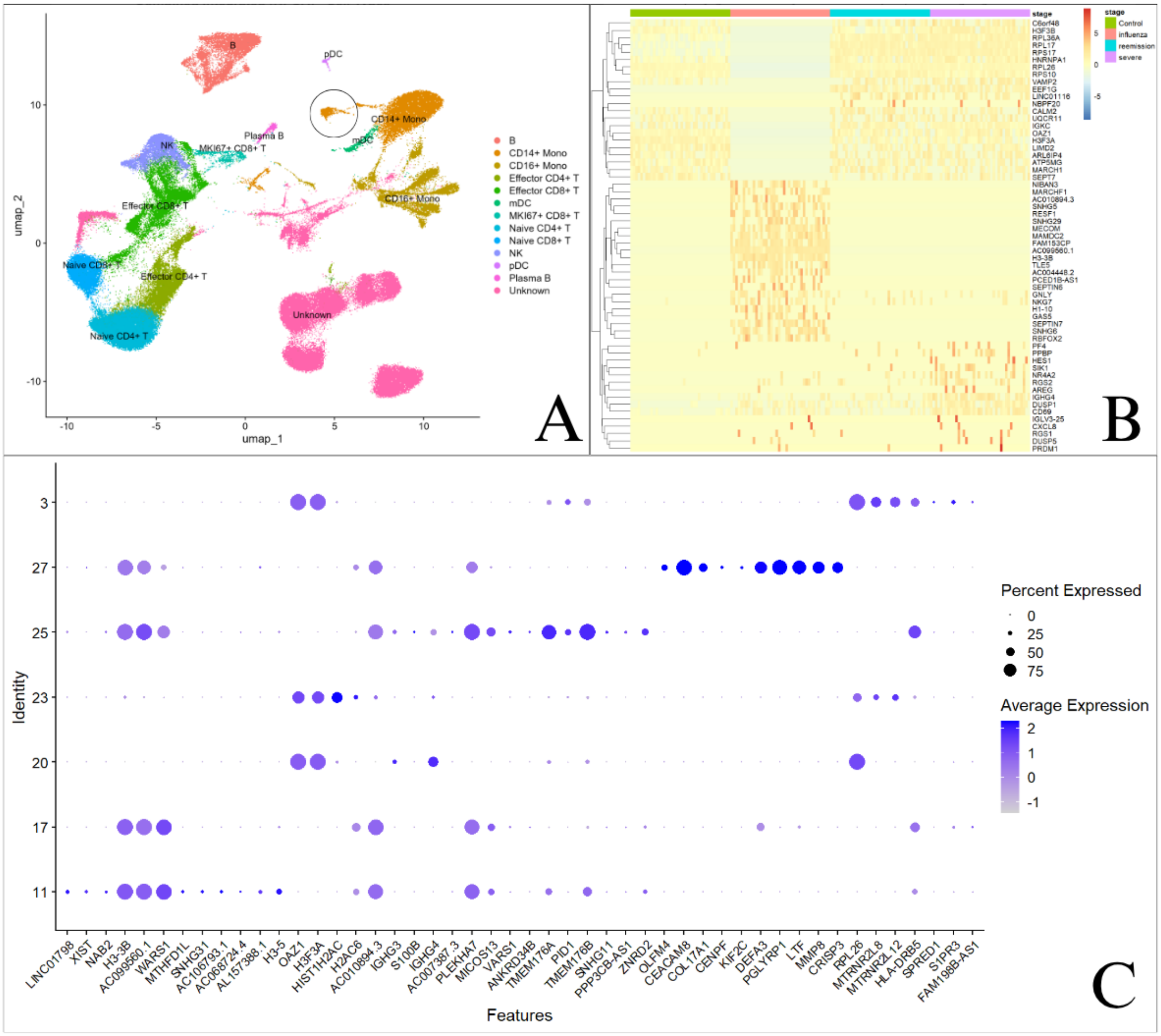

Figure 2: Integration of datasets and differential gene expression analysis across conditions. (A) UMAP visualization of integrated PBMC samples from influenza A-infected patients and COVID-19 cohorts. (B) Pairwise differential gene expression analysis comparing control samples, influenza-infected PBMCs, and COVID-19 samples in remission and severe stages visualized using heatmap. (C) DotPlot showing the top differentially expressed genes in the monocyte cluster characteristic of severe-stage COVID-19 (20) compared to other monocyte clusters identified after integration with the influenza A dataset (11, 17, 25, 27). Cluster 3 consisted of control cells and cluster 11 consisted of cells from different sources.

All monocyte clusters were aggregated to assess their differences, with particular emphasis on cluster 20, which was a defining feature in the prior analysis. Differential gene expression analysis for each monocyte cluster is presented as a DimPlot in Fig. 2C. The data indicate that monocyte clusters exhibit distinct gene expression profiles. Specifically, clusters 11, 17, 25 and 27 correspond to influenza virus infection, cluster 23 comprises mixed cell populations, and cluster 3 represents the control group. Notably, cluster 20 demonstrates a gene expression pattern closely resembling that of the control group. There is a paucity of genes shared in expression patterns between cluster 20 and clusters associated with influenza infection. The gene IGHG4 stands out as being most highly expressed in cluster 20 relative to other clusters. IGHG4 encodes the heavy chain of the IgG4 immunoglobulin subclass, one of four IgG isotypes. The reduced or absent expression of this gene in influenza-associated clusters may reflect differences in disease pathophysiology, as influenza is characterized by an acute onset and brief clinical course, whereas COVID-19 infection can persist for several weeks with prolonged antigen exposure. Furthermore, the genes expressed within cluster 20 are largely similar to those observed in the control group; however, this analysis focused only on the top three most highly expressed genes per cluster. It is plausible that differential gene expression occurs only in a subset of cells within the cluster, which might not be fully captured by DotPlot visualizations. This interpretation aligns with previous observations from K-means subclustering, which revealed gene expression differences limited to specific subclusters within the severe-stage cluster. Additionally, potential functional changes may occur independent of transcriptomic alterations, for example through post-translational modifications. Overall, these findings suggest a significant transcriptomic divergence between macrophages characteristic of severe-stage COVID-19 and those associated with influenza infection, supporting the hypothesis that cluster 20 is unique to COVID-19 pathogenesis.

The second comparative analysis conducted in this study involved integration with an independent COVID-19 dataset. Interestingly, the monocyte cluster identified in the original dataset appeared to be specific only to that particular cohort. Unlike in the comparison with the influenza dataset, in this case, the absence of the cluster in the new dataset does not support its general relevance to COVID-19. Rather, it suggests that the cluster may reflect unique features of the two original patients rather than a universal characteristic of the disease. Furthermore, although not shown in this report, UMAP visualization revealed several other distinct clusters associated with the severe stage of COVID-19 in the independent dataset — clusters that were entirely absent in the original dataset. This observation indicates that the presence of the originally described severe-stage cluster may have been coincidental, specific to those two individuals. It also raises the possibility that the immune response to SARS-CoV-2 is far more individualized than previously assumed. Alternatively, it may reflect redundancy in immune signaling pathways: the cytokine storm, for instance, is a non-specific inflammatory motif often driven by common mechanisms such as macrophage pyroptosis. Given the considerable variability observed between patient samples, analyzing single clusters from small datasets may be of limited value. Drawing general conclusions from such outliers, as was done in the original study, is methodologically problematic. Broader datasets — comprising hundreds or even thousands of patients — are necessary to reliably characterize the transcriptional landscape of immune responses, including monocyte subpopulations specific to the severe stage of COVID-19. To further investigate whether the original ”severe-stage” cluster might have been driven by cells from a single patient, the distribution of cells within the cluster was assessed. Interestingly, cells were evenly derived from both individuals, and K-means clustering did not reflect any batch or patient-specific artifacts. This suggests that technical bias is unlikely to explain the observation. It is also worth considering potential biological sources of variation, such as the year and geographic origin of the samples. The original dataset was collected in China in 2020, whereas the independent COVID-19 dataset was obtained from South Korea in 2022. Given the rapid evolution of SARS-CoV-2 and the known geographic and temporal heterogeneity of viral strains and immune responses, biological differences in the nature of infection likely contribute to the observed discrepancies.

Summary

This study aimed to replicate and expand upon findings from a previously published single-cell RNA-seq study on peripheral blood mononuclear cells in COVID-19. Alternative preprocessing choices, such as mitochondrial gene expression thresholds, influenced the compactness and composition of the monocyte cluster. Procrustes analysis and UMAP alignment indicated that major differences between UMAPs resulted from geometric transformations rather than biological divergence. Subclustering of the monocyte population using K-means revealed limited heterogeneity. However, minor subclusters were associated with distinct transcripts, such as AP001189.1, a pyroptosis-related lncRNA, and DDR2, involved in macrophage polarization. These findings suggest functional diversity within the seemingly homogeneous monocyte cluster, potentially reflecting different immune roles or responses. Comparative analyses with external datasets, including influenza A virus-infected patients and an independent COVID-19 cohort, yielded divergent results. The monocyte cluster characteristic of severe-stage COVID-19 was preserved when integrated with the influenza dataset, but not with the independent COVID-19 dataset. This suggests that the originally reported cluster may reflect patient-specific or cohort-specific immune responses rather than a universal signature of severe COVID-19. Differential gene expression analyses across B cells further demonstrated divergent transcriptional responses to SARS-CoV-2 and influenza A, with genes such as HES1 and NBPF20 upregulated selectively in COVID-19. These discrepancies point to virus-specific immune activation and potentially different regulatory pathways. In conclusion, while the severe-stage monocyte cluster observed in the original study was broadly reproducible, its functional composition appears sensitive to analytical parameters. Its absence in independent COVID-19 datasets calls into question its generalizability. These findings highlight the need for large-scale, diverse datasets to draw robust conclusions about immune cell states in COVID-19 and other viral infections.

References

[1] Msemburi, W., Karlinsky, A., Knutson, V. & et al. The who estimates of excess mortality associated with the covid-19 pandemic. *Nature* **613**, 130–137 (2023).

[2] Rothwell, J. T., Cojocaru, A., Srinivasan, R. & Kim, Y. S. Global evidence on the economic effects of disease suppression during covid-19. *Humanities and Social Sciences Communications* **11** (2024).

[3] Wu, P. *et al.* Real-time tentative assessment of the epidemiological characteristics of novel coronavirus infections in wuhan, china, as at 22 january 2020. *Nature Communications* **11**, 618 (2020).

[4] OpenAI. Chatgpt (june 2025 version). https://chat.openai.com (2025). Accessed: 2025-06-25.

[5] Zhou, W. & Zhang, W. A novel pyroptosis-related lncrna prognostic signature associated with the immune microenvironment in lung squamous cell carcinoma. *BMC Cancer* **22** (2022).

[6] Ferreira, A. C. *et al.* Sars-cov-2 engages inflammasome and triggers pyroptosis in human monocytes. *Cell Death Discovery* **7** (2021).

[7] Ni, L., Chen, D., Zhao, Y., Ye, R. & Fang, P. Unveiling the flames: macrophage pyroptosis and its crucial role in liver diseases. *Frontiers in Immunology* **15**, 1338125 (2024).

[8] Heydari, R., Tavassolifar, M. J., Fayazzadeh, S., Sadatpour, O. & Meyfour, A. Long non-coding rnas in biomarking covid-19: a machine learning-based approach. *Virology Journal* **21**, 134 (2024).

[9] Li, Q. *et al.* A cuproptosis-related lncrnas risk model to predict prognosis and guide immunotherapy for lung adenocarcinoma. *Annals of Translational Medicine* **11**, 198 (2023).

[10] Liu, Q., Wang, X., Chen, Y. & et al. Ablation of myeloid discoidin domain receptor 2 exacerbates arthritis and high fat diet induced inflammation. *Biochemical and Biophysical Research Communications* **649**, 47–54 (2023).

[11] The Human Protein Atlas. Nbpf20 gene expression summary – the human protein atlas. https://www.proteinatlas.org/ENSG00000162825-NBPF20 (2025). Accessed: 2025-06-25.

[12] Quentmeier, H., Pommerenke, C., Bernhart, S. H. & et al. Rbfox2 and alternative splicing in b-cell lymphoma. *Blood Cancer Journal* **8**, 77 (2018).