Vedant Kulkarni
NUID: 002087700

Final Project: Individual write-up

Introduction:

Single-cell and single-nucleus RNA sequencing (sc/snRNA-seq) have revolutionized our power to investigate human diseases at cellular resolution. In the setting of the COVID-19 pandemic, investigators utilized these technologies in efforts to understand why severe SARS-CoV-2 infection precipitates respiratory failure, an impaired lung regenerative response, and rapid fibrotic remodelling of the lungs. Among these, Melms et al., Nature 2021 provided one of the early and most impactful lung atlases describing cell-type-specific disruptions in patients who had died from COVID-19. The authors integrated post-mortem single-nucleus profiles from COVID-19 and control lung donors and identified inflammatory macrophage activation, dysfunctional epithelial repair, and expansion of pathological fibroblasts.

Our group's objective was to reproduce part of this landmark analysis namely, the integration, annotation, and differential expression analyses using the 10X Genomics–derived raw count matrices available on GEO. We implemented the workflow in Python, Scanpy, and scvi-tools in Northeastern University's Explorer HPC cluster, while paying close attention to reproducibility, QC, and environment management.

Statement of Need:

The paper was written to address an immediate and urgent scientific problem during the COVID-19 pandemic: Why do some patients progress to lethal respiratory failure, while others recover? In early 2020, clinicians observed:

- Persistent inflammation despite viral clearance
- Failure of alveolar cells to regenerate
- Rapid fibrosis in severe forms
- Ineffective immune–epithelial communication

These results, however, were obtained with bulk tissue analyses and histology that did not reach cellular resolution. What was urgently needed was a high-resolution atlas of the human lung in COVID-19 that could identify which cell types break down, which respond abnormally, and which pathways drive severe disease.

Therefore, it was the goal of the authors to create a cellular-level map for COVID-19 lungs to explain the mechanisms underlying mortality.

Original Researcher's Hypothesis:

From Melms et al.'s text, the core hypothesis was the lethal COVID-19 is dominated by integrated dysfunction across epithelial, immune, and stromal compartments, inclusive of failure in regeneration, hyperinflammation, and the activation of profibrotic fibroblasts. More specific hypotheses include:

- The AT2 epithelial cells do not regenerate AT1 cells.
- Macrophages assume a hyperinflammatory IL-1β–producing state.

- Expansion of pathological CTHRC1^{+} fibroblasts contributes to fibrosis.
- Together, these cellular programs promote acute respiratory failure.

Corroborating Evidence Used by the Authors:

The authors further reinforce their findings with multiple layers of biological evidence, including defective AT2 to DATP transitions evidenced by single-cell transcriptional signatures, strong IL-1β upregulation in macrophages, and the activation of ECM and TGF-β–related programs in fibroblasts. These findings are corroborated by a significant increase in proinflammatory cytokines, fibrotic remodeling genes, and loss of epithelial markers in COVID-19 lungs. Changes in cell-type proportions enforce these findings, with an expansion of macrophages and pathological fibroblasts at the expense of decreased epithelium. Similarly, ligand–receptor analyses reveal IL-1β–driven macrophage–epithelium signaling and TGF-β communication toward fibroblasts, further suggesting dysfunctional intercellular crosstalk. Finally, these changes are visually confirmed by histology and immunohistochemistry against a background of extensive fibrosis and immune infiltration. These combined multimodal observations support strongly the conclusions from this study.

Statement on the Validity of the Methods and Results:

The methods in the original study are scientifically sound and consistent with accepted practices for scRNA-seq analysis:

- Single-nucleus RNA-seq was used appropriately for post-mortem tissue.
- Use of batch-corrected integration (Seurat reciprocal PCA)
- Robust QC metrics
- Differential expression with appropriate statistical tests
- External validation by histology

Since the reproduction was done in Python/Scanpy/scvi-tools instead of Seurat, it was not possible to replicate all of the intermediate results exactly, but the trends we saw were consistent:

- Abundance of fibroblasts and macrophage-like cells.
- Elevated ECM-related gene signatures
- Distinct separation between COVID and control samples in latent space

Thus, the findings of Melms et al. seem scientifically valid and reproducible on a broad level, even though fine-grained annotation could not be replicated because of differences in the dataset and lower gene coverage.

Drawbacks:

Below are unique, non-obvious limitations that we identified:

1. Although they reported PMI, they did not deeply analyze how transcriptional degradation might bias immune or epithelial signatures.
2. The treatments-which likely differed among patients, possibly including steroids and/or antivirals-were not modeled in the analysis. Such differences have the potential to substantially affect transcriptional profiles.

3. Although fibroblasts were highlighted, other stromal cells (pericytes, mesenchymal progenitors) were not deeply dissected, thereby probably masking subtle biology.
4. The integration approach may favor abundant cell types, reducing resolution of rare populations, such as tuft cells, ionocytes, and DC subsets.
5. Often, COVID samples had higher UMIs; however, this was not normalized thoroughly in the downstream analyses.

These limitations do not invalidate the paper but rather give directions for refinement.

Reflection on What Went Wrong:

The analysis faced multiple compounding challenges: the GEO matrices contained only ~161 highly variable genes rather than the full transcriptome which severely limited cell-type annotation capabilities; extensive troubleshooting was required to align raw count layers with integrated cells and resolve naming inconsistencies; many canonical markers from the Nature study were absent from the dataset resulting in weak annotation power; conflicting versions of anndata, requests, and torch libraries caused installation failures; and enrichment analysis via Enrichr/GSEA required manual workarounds due to the HPC cluster blocking certain HTTP methods and the requests library being shadowed by an HPC module.

What Could Have Been Done Better:

a. Begin with complete Seurat object- We should have downloaded the official .rds object and converted it to AnnData; this would have provided the full gene set.
b. Reproduce Seurat integration directly- Using Seurat's reciprocal PCA would have allowed closer replication of results.
c. Pre-build a Docker or Conda environment- Many installation issues could have been avoided if a reproducibility container was prepared from the very beginning.
d. Version control with more care- This would have saved debugging time by pinning all versions (scanpy, torch, etc.) earlier.
e. Parallelization of SCVI- Training SCVI models on the cluster with batch submission would be faster and more reproducible.

Conclusion:

This work has reproduced the wide-ranging biological findings of Melms et al., Nature 2021, by using a Python/Scanpy/scvi-tools pipeline developed independently. Although full replication was limited by the differences in dataset and gene coverage, our results support the major conclusions about fibroblast activation, immune dysregulation, and distinct COVID-19 pathology patterns.

This experience further solidified my skills in computational pipeline design, HPC workflows, reproducibility practices, and critical evaluation of single-cell studies. This project also showed the importance of strict environment management and clear documentation if one attempts to reproduce high-impact biological research

Citation:

Melms, J. C., Biermann, J., Huang, H., et al. (2021). *A molecular single-cell lung atlas of lethal COVID-19.* **Nature**, 595, 114–119. https://doi.org/10.1038/s41586-021-03569-1