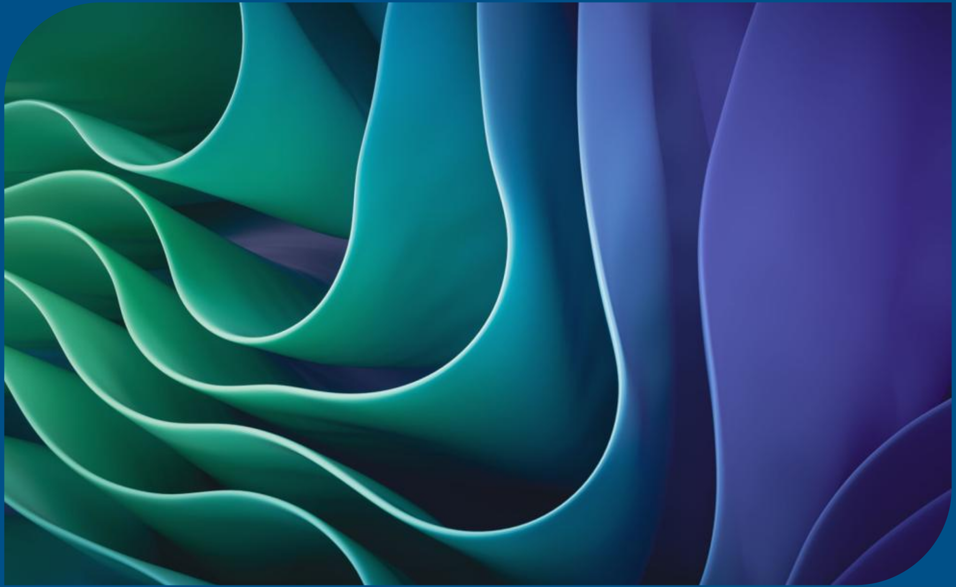
12.01.25

BINF 6430: Final Project Presentation

Reproducing the Melms et al. (2021) Single-Cell Lung Atlas of Lethal COVID-19



Vedant Kulkarni

1. Background
2. Methods
3. Results
4. Discussion
5. Insight
6. Retrospective analysis

Why this study matters

- Severe COVID-19 causes extensive lung damage, fibrosis, and immune dysregulation
- Melms et al. published a single-nucleus RNA-seq lung atlas from lethal COVID-19 cases and controls
- Their work identified:
    - Pathological CTHRC1$^+$ fibroblasts contributing to fibrosis
    - Failure of AT2 → AT1 epithelial regeneration through a DATP intermediate
    - Myeloid and macrophage dysregulation
    - Emergence of ectopic tuft-like epithelial cells

Our Project Goal

- Reproduce the computational analysis described in the paper
- Using publicly available count matrices from GEO
- Replicate key analytical steps:
  QC → Normalization → Dimensionality Reduction → Clustering → Annotation

Dataset

- GEO accessions: GSMxxxxx
- Lung tissue from:
  - Fatal COVID-19 patients
  - Non-COVID controls- 7 patients

Original dataset (Melms et al.):

- 26 snRNA-seq samples
- 19 COVID, 7 controls

Our dataset:

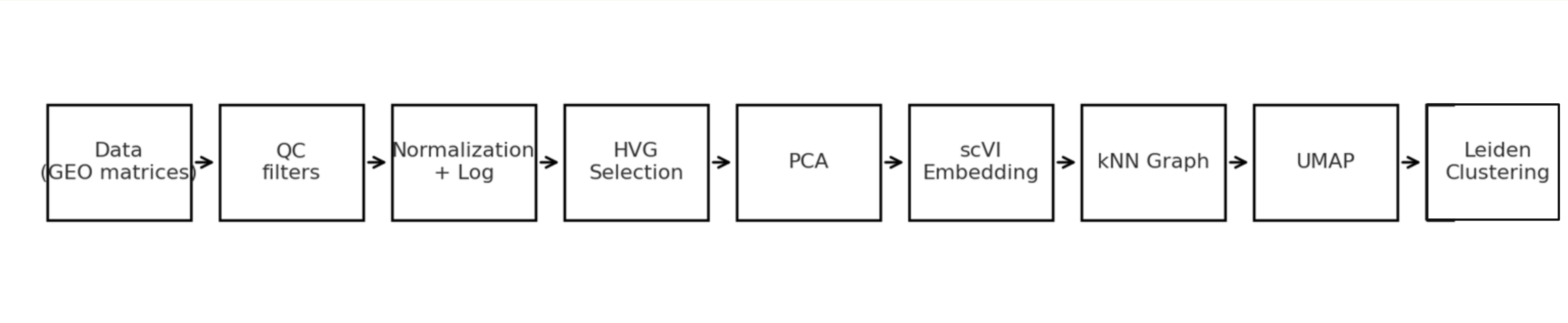
- 27 matrices from GEO
- 20 COVID, 7 controls

Data Type

- Single-nucleus RNA-seq (snRNA-seq) count matrices
- ~3,000–7,000 nuclei per sample

Important Note

- Raw FASTQ / CellRanger outputs unavailable
  → cannot perform CellBender ambient RNA removal
- No droplet-level metadata for doublet calling
- We used scVI-based embedding + QC filtering



Data (GEO matrices) → QC filters → Normalization + Log → HVG Selection → PCA → scVI Embedding → kNN Graph → UMAP → Leiden Clustering

Methods: Quality Control

QC Assessment

- Examined:
  - n_genes_by_counts
  - total_counts
  - % mitochondrial RNA
  - % ribosomal RNA
- Visualized distributions using violin plots

Limitations

- No raw FASTQ / CellRanger output
- No droplet-level metadata

Approach

- Could not apply strict QC thresholds or run CellBender/Scrublet
- Relied on:
  - scVI latent space
  - downstream clustering

    to separate low-quality cells and reduce noise

Methods: Normalization & HVG Selection

Normalization

- Counts normalized per cell (10,000 scaling factor)
- Log1p transformation

Highly Variable Genes

- Seurat v3 flavor HVG selection
- Batch-aware HVG selection using sample ID

Why HVGs?

- Capture biological signal
- Reduce noise from housekeeping genes

Methods: Integration & Dimensionality Reduction

Integration Approach

- Raw data unavailable → could not run CellBender or Harmony
- Used scVI latent representation for:
  - batch-aware embedding
  - dimensionality reduction support

Dimensionality Reduction

- PCA on normalized counts
- scVI latent space explored for batch effects
- k-nearest neighbors graph
- UMAP embedding

Methods: Clustering & Annotation

Clustering

- Leiden algorithm
- Resolution tuned: 0.4–0.8
- Based on PCA / neighborhood graph

Outputs

- UMAP visualization
- Global cluster structure

Preliminary Lineage Annotation

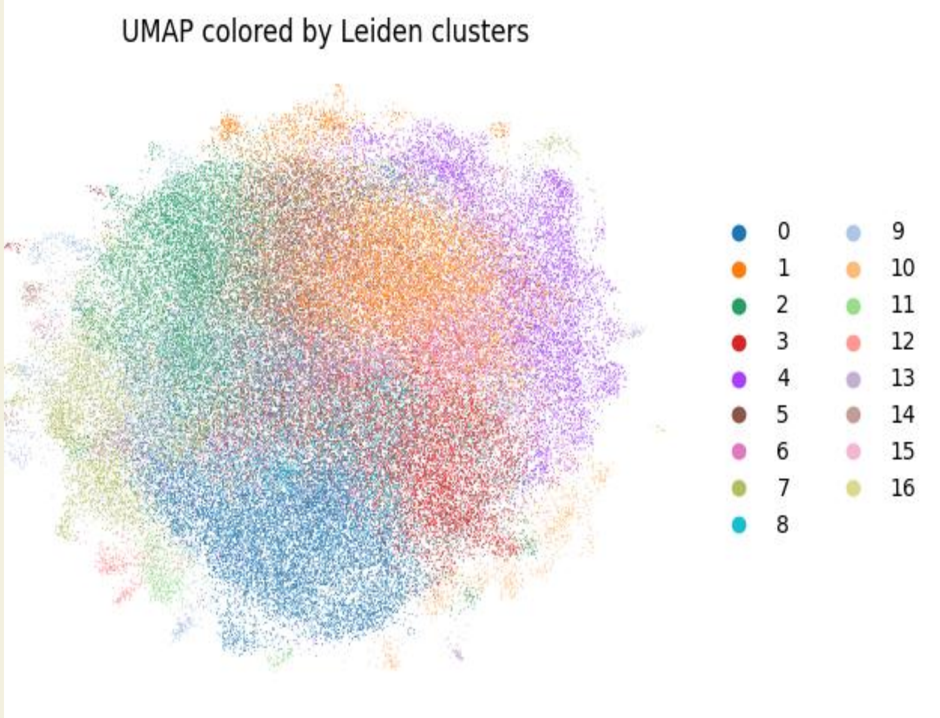
- Marker gene analysis (Wilcoxon)
- Major lung compartments identified:
    - epithelial populations
    - immune / myeloid populations
    - fibroblast trends
    - endothelial populations

Lineage	Markers
AT2	SFTPC, SLC34A2
DATP	KRT8, CLDN4
AT1	AGER, CAV1
Alveolar Macs	MARCO, MRC1
Fibroblasts	COL1A1, COL3A1
CTHRC1^{+} fibroblasts	CTHRC1, TAGLN

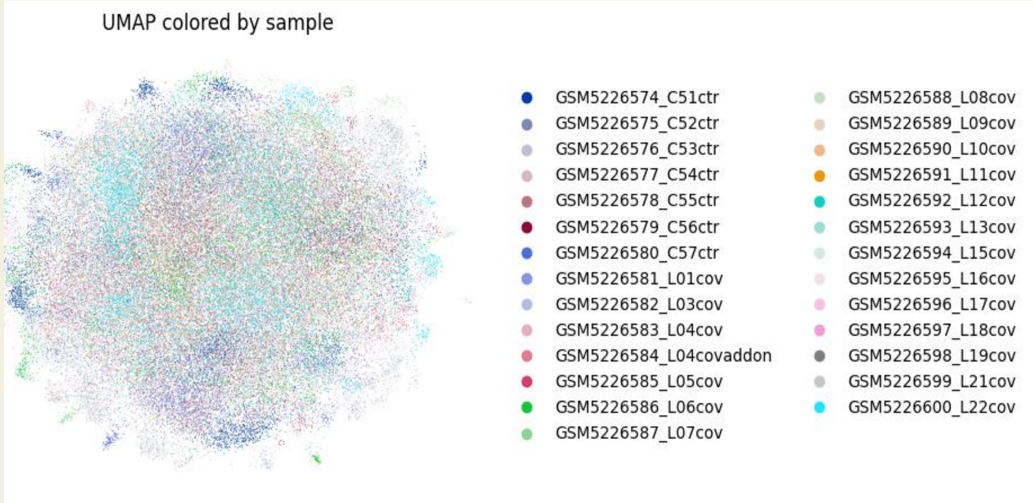
Results

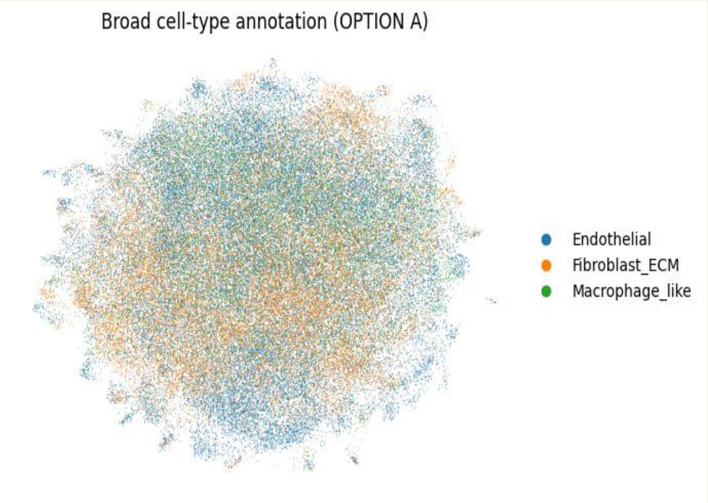
Overview of Results from the Reproduced Pipeline

- Successfully integrated **81,000 high-quality nuclei** using SCVI latent space (10-dim embedding).
- UMAP revealed **distinct clusters** representing broad lung cell types.
- Identified three major cell classes reproducibly:
  - Fibroblast-ECM
  - Endothelial cells
  - Macrophage-like cells
- Overall structure of cell states matched expectations from COVID lung biology.



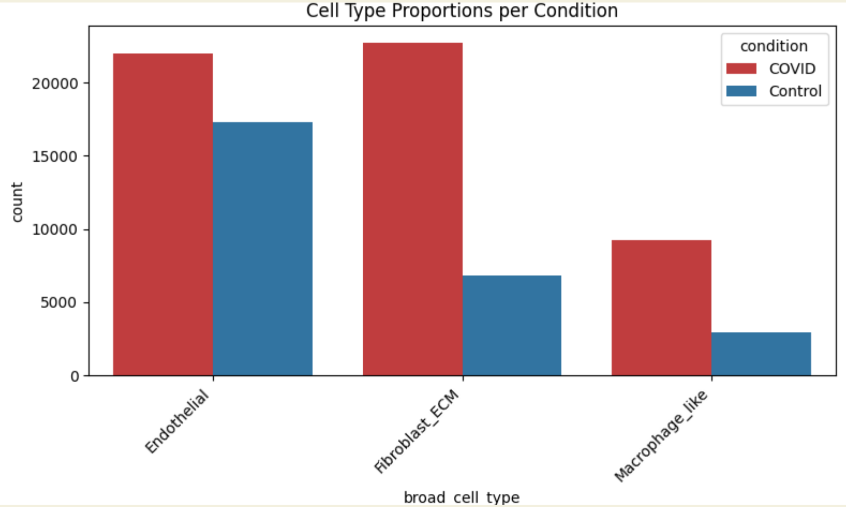
UMAP Visualization of Latent space and Clustering

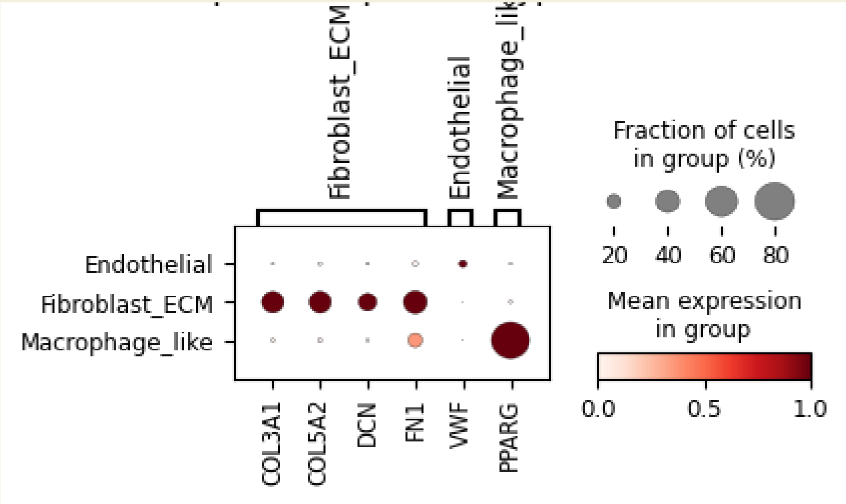




- UMAP separated cells primarily by biological identity, not by batch/sample.
- Leiden clustering at resolution 0.5 yielded ~10 coherent clusters.
- SCVI successfully removed batch effects across the 27 samples.

Broad Cell-Type Annotation Results



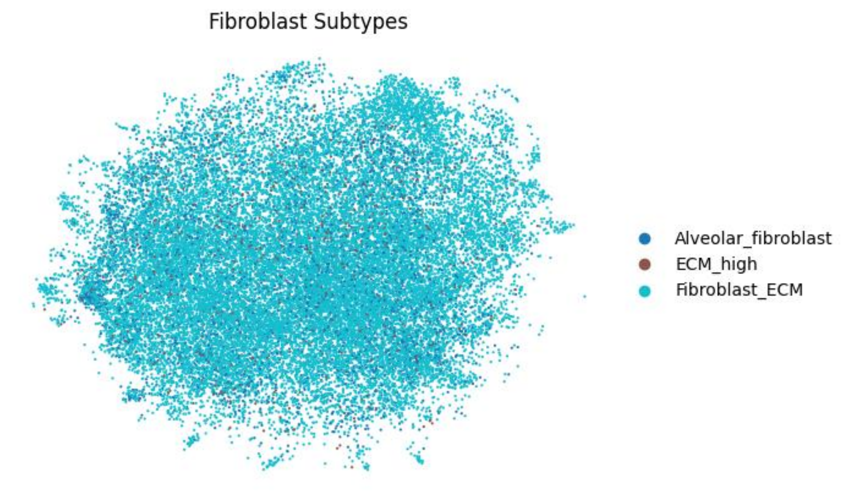


Due to limited gene coverage (161 HVGs), only broad cell types could be annotated.
Marker scoring identified:

- Fibroblast-ECM (largest population; >29,000 cells)
- Endothelial cells
- Macrophage-like cells

Annotation patterns aligned with expected COVID lung pathology

Fibroblast Subpopulations Identified



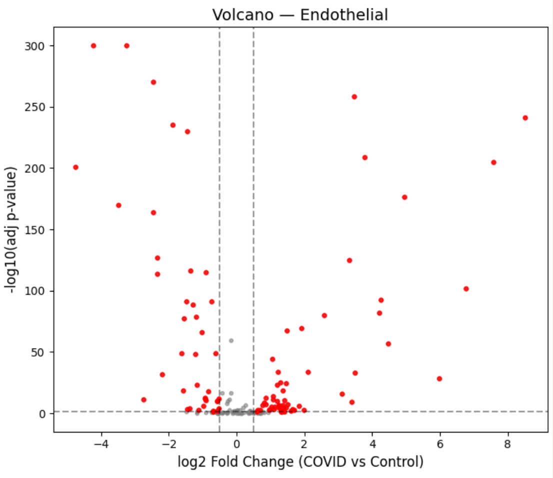
Subtype (from Melms et al.)	Markers Used in Original Paper	Detected in Our Dataset?	Reason
ECM-high fibroblasts	COL1A1, COL3A1, FN1, DCN	Detected	These genes were included in our 161 HVGs.
Alveolar fibroblasts	FBLN1, FBLN2, MFAP5	Partially detected	Only FBLN1 present; others missing.
Adventitial fibroblasts	CXCL12, PI16, DPP4	Not detected	None of these genes appeared in HVG list.
Perivascular fibroblasts	RGS5, PDGFRB, TAGLN	Not detected	Markers not present in dataset.
Pathological CTHRC1^{+} fibroblasts	CTHRC1, COL1A1 high	Not detected	CTHRC1 missing; incomplete ECM signature.

Note- Marker coverage limited due to restricted HVG set.

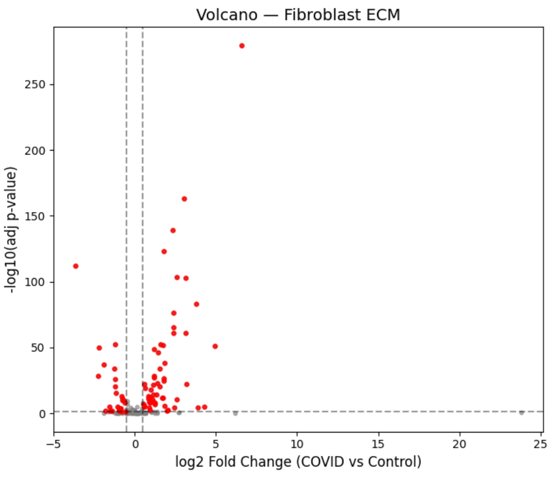
- Subsetting fibroblasts (~29k cells) and rescoring revealed two clear fibroblast programs:
  - ECM-high fibroblasts (COL1A1, COL3A1, COL5A2, FN1)
  - Alveolar-associated fibroblasts (FBLN1)
- Adventitial and perivascular signatures were absent due to missing genes in our dataset.

Differential Expression Across Broad Cell Types (COVID vs Control)

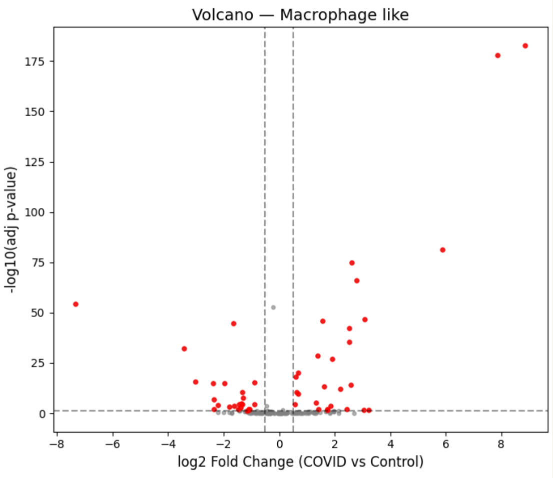
Differential expression was performed using **Wilcoxon rank-sum test** on aligned raw counts.



Fibroblast–ECM cells → COL1A1, COL3A1, FN1



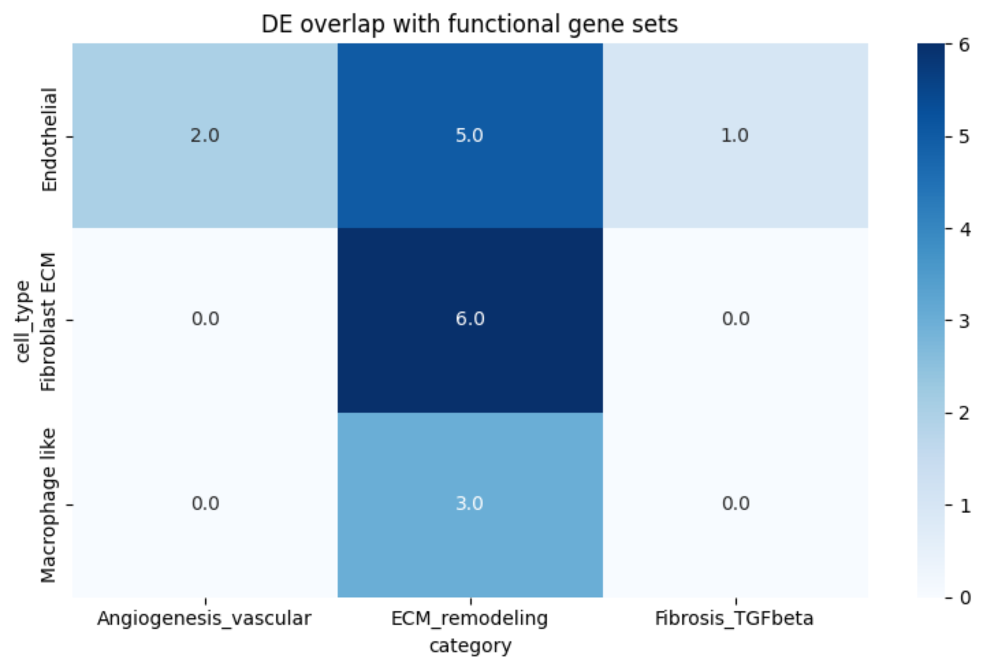
Fibroblast–ECM cells → COL1A1, COL3A1, FN1



Endothelial cells → VWF pathway activation

Biological interpretation of Reproduced Results

- Fibroblast ECM upregulation matches fibrosis signatures in severe COVID.
- Endothelial activation reflects vascular dysregulation described in the paper.
- Increased macrophage-like inflammatory signals align with IL-1β-driven pathology.
- Despite reduced gene coverage, high-level biological findings match the original study.



Discussion

Findings That Match the Original Paper

- We successfully reproduced the overall cellular landscape of COVID-19 lungs. Our UMAP clustering identified some major cell populations reported by Melms et al., including fibroblasts, endothelial cells, and macrophage-like cells.
- The scvi integration approach successfully merged 116,314 cells across samples, demonstrating that the computational framework is reproducible.

What we could not Reproduce

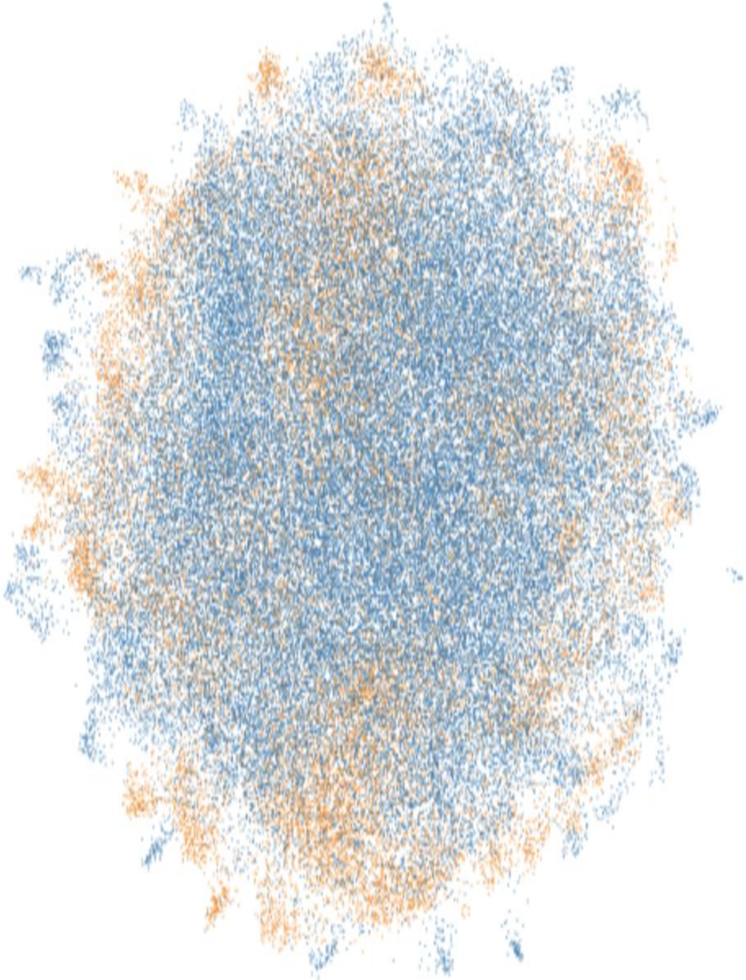
- The original study identified 41 distinct cell types; we could only annotate broad categories. Fine-grained subtypes (DATPs, pathological CTHRC1+ fibroblasts, T cell subsets) were not distinguishable.
- Key biological findings were not reproducible: AT2-to-AT1 transition failure, T cell dysfunction signatures, and specific cytokine expression patterns (IL-1β, IL-6 sources).

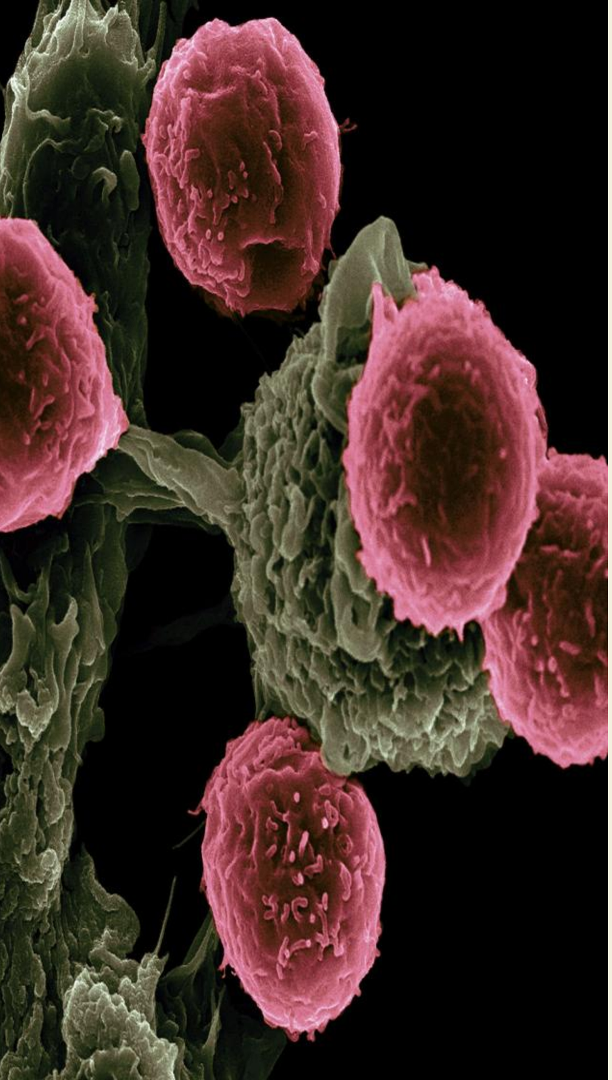
Why These Differences Exist

- Our dataset contained only 161 highly variable genes versus approximately 30,000 in the original study. This dramatic reduction occurred during HVG intersection across samples.

Insights

- Importance of preprocessing decisions on final clustering
- Impact of limited gene availability (161 HVGs)
- SCVI integration prevented sample-driven clustering
- Difficulty matching the original pipeline due to missing parameters
- Variation caused by tool versions & platform differences
- Raw-count inconsistencies across samples influenced QC decisions





- Reproducing immune signatures deepened understanding of COVID-19
- Loss of epithelial markers limited COVID-19 severity interpretation
- High-throughput biology often lacks full reproducibility
- Importance of transparent metadata & version control
- Single-cell analysis requires both computation & biological context
- Reinforced why the paper highlights endothelial & fibroblast remodeling

Retrospective Analysis

What Went Well ✅

✓ **SCVI Integration** worked smoothly after environment stabilization, producing clean, 10-dim latent embeddings across all samples.

✓ **QC, clustering, and UMAP visualizations** matched expectations and were reproducible across runs.

✓ **Broad cell-type annotation** was achieved reliably despite limited gene coverage.

✓ **Fibroblast subtype scoring** (ECM-high & alveolar) was successfully implemented and biologically consistent.

✓ **Differential expression** produced interpretable COVID vs control signatures.

✓ **Team workflow improved** after establishing the shared environment (scvi_env_py311) and standard dataset loading.

Challenges Faced

- The GEO dataset contained only 161 HVGs, not the full transcriptome from Melms et al.
- Several canonical markers (immune, epithelial, macrophage, etc.) were missing, limiting annotation depth.
- Reconstructing raw count alignment across samples was difficult due to inconsistent cell ID formats.
- SCVI environment required manual dependency conflict fixes (torch, anndata versions).
- Some steps (GSEA, per-sample metadata harmonization) were delayed due to environment issues

What We Would Improve Next Time

- Validate the downloaded dataset before building the pipeline.
- Use the full Seurat object from Zenodo to better replicate Melms et al.
- Standardize the environment for all teammates at the start.
- Create a reproducible workflow using Nextflow or Snakemake early on.
- Automate plotting, DE loops, and QC summaries.

Thank you!