ScRNA-seq Expression of *IFI27* and *APOC2* Identifies Four Alveolar Macrophage Superclusters in Healthy BALF

Xin Li[1], Fred W. Kolling[2], Daniel Aridgides[3], Diane Mellinger[3], Alix Ashare[1,3*], and Claudia V. Jakubzick[1*]

[1]Department of Microbiology and Immunology, Dartmouth Geisel School of Medicine, Hanover, NH [2]Department of Biomedical Data Science, Dartmouth Geisel School of Medicine, Hanover, NH [3]Department of Medicine, Dartmouth Hitchcock Medical Center, Lebanon, NH [*]Co-senior authors

Correspondences:
Claudia Jakubzick, PhD
Geisel School of Medicine at Dartmouth
Department of Microbiology and Immunology
626W Borwell
One Medical Center Drive
Lebanon, NH 03756
Phone: 303-898-3791
claudia.jakubzick@dartmouth.edu

Alix Ashare, MD, PhD
Geisel School of Medicine at Dartmouth
Dartmouth Hitchcock Medical Center
Department of Medicine
Department of Microbiology and Immunology
One Medical Center Drive
Lebanon, NH 03756
alix.ashare@hitchcock.org

Conflict-of-interest Statement:
The authors have declared that no conflict of interest exists.

Abstract

Alveolar macrophages (AMs) reside on the luminal surface of the airways and alveoli, ensuring proper gas exchange by ingesting cellular debris and pathogens, and regulating inflammatory responses. Therefore, understanding the heterogeneity and diverse roles played by AMs and recruited monocytes is critical for treating airway diseases. We performed single-cell RNA sequencing on 113,213 bronchoalveolar lavage cells from four healthy and three uninflamed cystic fibrosis subjects and identified four AM superclusters (families) based on the expression of *IFI27* and *APOC2* genes. These 4 AM families have at least eight distinct functional members (subclusters) named after their differentially expressed gene(s): IGF1, CCL18, CXCL5, Cholesterol, Chemokine, Metallothionein, Interferon and small-cluster AMs. Interestingly, the Chemokine cluster further divides with each subcluster selectively expressing a unique combination of chemokines. One of the most striking observations, besides the heterogeneity, is the conservation of AM family members in relatively equal ratio across all AM superclusters and individuals. Finally, transcriptional data and TotalSeq technology were used to investigate cell surface markers that distinguish resident AMs from recruited monocytes. Overall, functional specializations of the AM subclusters suggest that there are AM niches with defined programming states, highlighting a clear division of labor.

Introduction

Alveolar macrophages (AMs) are the primary phagocytes within the airspace of the lungs. In response to inflammatory stimuli, AMs secrete cytokines and chemokines to initiate the immune response. Subsequently, AMs and recruited monocytes contribute to the resolution of the inflammatory response to prevent excess damage (1-4). In addition, they play a major role in maintaining lipid homeostasis in the lung, which is essential for adequate gas exchange and alveolar epithelial integrity.

AMs may hold keys to understanding, preventing, and curing several diseases. They derive from prenatal monocytes, self-renew, and have shown remarkable plasticity by altering their transcriptome in response to the environmental changes (3, 5-8). Their transcriptional profile is altered in various inflammatory lung diseases, including asthma (2, 9), chronic obstructive pulmonary disease (10), cystic fibrosis (CF) (11), and cancer (12-17). Therefore, enhancing our current understanding of the normal cellular composition and functional diversity of AM subtypes will help characterize disease-related transcriptional changes more precisely and assess targeting specifically a given AM subtype to restore immune balance.

Single-cell RNA sequencing (scRNA-seq) technology is critical for accomplishing these goals. AMs have traditionally been thought to be a uniform population of cells that can be activated by different disease states (5). However, recent scRNA-seq studies have revealed a rich diversity in AMs in bronchoalveolar lavage fluid (BALF) from healthy subjects, with multiple subpopulations that have yet to be characterized other diseases (18).

Our study was designed to investigate the immune cell populations of BALF and potential differences in individuals with mild CF compared to healthy controls (HC). Interestingly, unbiased analysis of our scRNA-seq dataset identified four AM superclusters based on the expression of *IFI27* and *APOC2* genes, which has not been previously described. Each supercluster contains four shared subclusters named after their differentially expressed genes (DEGs): IGF1.AMs, CCL18.AMs, CXCL5.AMs, and Cholesterol.AMs. Beyond the four AM superclusters, we found four additional clusters, which can categorize into the AM superclusters. The most striking AM subcluster was the Chemokine cluster, which further divides into four subclusters with each selectively expressing a distinct combination of chemokines. Hence, there appears to be a clear division of labor in leukocyte recruitment. Moreover, the Chemokine subclusters suggest that there are distinct transcription factors (based on our regulon analysis) that regulate their differentiation and programming. This pattern is analogous to what is found in T cell biology, where distinct transcription factors regulate the cell differentiation pathway to produce a defined combination of cytokines known as Th1, Th2, Th17, etc. Overall, our dataset opens many new areas for further investigation, including addressing how resident macrophage and monocyte subtypes are altered during disease and what factors regulate their programming.

Results

BAL cell populations are conserved across healthy and mild CF subjects

To assess the cellular composition in the non-diseased airspace versus that with CF, four HCs and three subjects with mild CF underwent BAL (donors, Table 1). All three CF subjects had two copies of the F508del *CFTR* mutation. Study subjects were closely age matched. CF subjects were selected for this study based on well preserved lung function and general good health. To investigate the impact of aberrant CFTR function on the immune cell composition of the lung, we enrolled CF subjects without significant lung inflammation and who were not on CFTR modulator treatment. Two subjects had been on CFTR modulator treatment previously but stopped more than 12 months prior to enrollment due to side effects, while the third subject declined treatment with a CFTR modulator due to overall good health. The CF subjects in this study were on a stable medication regimen and had no exacerbations within the past 24 months.

BALF samples were spun, washed, and immediately loaded onto the 10X platform for scRNA-seq. Sequenced cells were processed, normalized, and integrated using the Seurat package (details in Methods). Uniform manifold approximation and projection (UMAP) of all seven samples illustrates a similar cluster distribution across all subjects (Figure 1A). Using a curated gene list, we identified twelve major cell types (Complete list of genes: Table S1 and Figure 1B-1D) (18-29). As anticipated, the most abundant cell type is the alveolar macrophage, followed by monocytes, cycling myeloid cells, dendritic cell (DC) 2, FOLR2 and SPP1 interstitial macrophages (30), lymphocytes, epithelial cells, migratory DCs, DC1, plasmacytoid DCs, and cycling lymphoid cells (Figure 1C-1D). Next,

MetaNeighbor analysis was used to confirm the classification and replicability of the identified clusters within our dataset and with two other recently published datasets, Mould et al., and Schupp et al. (18, 19, 31). Within our dataset, samples derived from HC and CF subjects demonstrated a strong correlation across all cell types (Figure 1E). Similarly, our HC sample macrophages, monocytes, and epithelial cells strongly correlated with healthy cell types from Mould et al., as did our CF sample cell types with CF sputum cell types (Figure 1F-1G). All in all, curated genes and robust alignment of our cells with previous studies support the identification of twelve cell types in healthy and CF BALF.

Four AM superclusters and shared subclusters co-exist in the airspace of the lungs

To define AM heterogeneity, we isolated and re-clustered the AM population. Unbiased clustering defined 8 AM clusters. Some clusters were previously described (32), such as the Metallothionein, Chemokine, IFN-reacting, and Cycling AM clusters. (Figure 2A-2D). In addition, we identified 4 novel AM superclusters based on the differential expression of two genes: *IFI27* and *APOC2* (Figure 2B-2E): *IFI27*$^{+}$*APOC*$^{+}$ AMs.S1, *IFI27*$^{+}$*APOC*$^{-}$ AMs.S2, *IFI27*$^{-}$*APOC*$^{+}$ AMs.S3, and *IFI27*$^{-}$*APOC*$^{-}$ AMs.S4. Besides *IFI27* and *APOC2*, the four AM superclusters share the same top DEGs *IFI6*, *CTSC*, *RPS4X*, *RBP4*, *FBP1*, and *GSN*, which suggest anti-microbial and metabolic function (i.e., common macrophage phenotypes) (Figure 2D). Each AM supercluster contains 4 shared subclusters. The subclusters are named after their top DEG(s): IGF1, CCL18, CXCL5 and Cholesterol (based on the expression of cholesterol-biosynthesis-related genes) (Figure 3A-3C). AM Supercluster 4 (AMs.S4) contained two additional subclusters, GF and SNHG. Cells in

the GF subcluster expressed the epithelial growth factor *AREG* and TGF-β superfamily ligands including, *GDF15,* whereas cells in SNHG expressed genes encoding small nucleolar RNA host genes (SNHGs) and genes encoding nuclear transcriptional regulators and transcription factors. The remaining cluster in each supercluster is referred to as a small-subcluster within the supercluster, noting a small 's' after AMs, for example, AMs.s#. This small-subcluster is the supercluster minus all four shared subclusters (Figure 3D).

In addition to the AM superclusters, four other AM clusters can be defined based on their expression of Metallothioneins (MT.AMs), Chemokines (CK.AMs), IFN-related genes (IFN.AMs), and Cycling genes (CK.AMs) (Figure 2B). The chemokine cluster can be further divided into 4 subpopulations, which we denote as CK.AMs.c1-4. The top 6 DEGs and curated genes illustrate the unique chemokine expression pattern within each subcluster (Figure 4D). CK.AMs highly express *CCL18*, *CCL23* and chemokine receptor *CXCR4* in CK.AMs.c1; *CCL3*, *CCL4*, *CXCL8* and *CCL20* in CK.AMs.c2; *CXCL9*, *CXCL10* and *CXCL11* (IFN-gamma inducible chemokines) in CK.AMs.c3, and *CXCL3* in CK.AM.c4 (Figure 4B-4D). The expression of unique combinations of chemokines in each CK.AMs suggests that different subtypes of AMs are programmed to chemoattract distinct leukocytes into the airspace, depending on the inflammatory stimuli they encounter. This is a novel finding since the levels of division of labor of resident AMs secreting chemokines have not been previously appreciated.

Next, given the number of AM subtypes identified, we performed Gene Ontology (GO) enrichment analysis to infer functionality based on their set of DEGs (Figure 4E).

Predictably, the Chemokine AM subclusters contained GO terms regarding cell interaction and response to stimuli, together with other enriched GO terms consistent with the top DEGs in each AM subtype. For instance, the top GO terms for Metallothionein AMs pertain to cellular responses to metal ions (Figure 4E and Figure S1). Likewise, the top 4 enriched terms for Cholesterol AMs contained "*cholesterol biosynthetic process*" and "*regulation of cholesterol biosynthetic process*" (Figure S1).

Lastly, to understand the transcription factors and co-factors driving each AM subtype, we performed single-cell level regulon analysis using the SCENIC (Single-Cell rEgulatory Network Inference and Clustering) package in R and its Python implementation pySCENIC (33). Co-expressed genes were used to identify and score regulons enriched in each cluster. For example, the proinflammatory regulator of the cytokine-induced apoptotic pathway, *BACH2*, is the top regulon identified for CK.AMs.c2, while the *IRF7* regulon is enriched in IFN-reacting AMs (Figure 4F and Figure S1). This analysis offers yet another vantage point for understanding the distinct AM subtypes and corroborates their defined DEGs and GO terms.

Like AM Chemokine clusters, monocytes also display unique chemokine-expressing clusters

We performed unbiased clustering for monocytes, which outlined 8 distinct populations (Figure 5A-5C). The largest cluster was denoted as conventional monocytes (Conv.Mono) that expresses classic myeloid cell markers shared with the other seven clusters (Figure 5D). An additional six populations can be defined as functional monocytes based on the expression of specific genes, including FCN1.Mono,

CCL2.Mono, CXCL5.Mono, CXCL9.Mono, CCL13.Mono, and VEGFA.Mono (Figure 5D). To note, we also identified an intermediate monocyte cluster (Int.Mono) with distinct DEGs that encompasses cells transitioning from Conv.Mono to other functional monocyte phenotypes. Thus, monocytes appear to display a division of labor in orchestrating and recruiting leukocytes into the airspace, similar to what is observed for chemokine-expressing AMs.

IL10 is selective secreted by two IM clusters

Other than AMs and monocytes, we observed several unexpected cell type clusters in the BALF, including IMs and DCs. The IMs in our dataset form two distinct clusters with distinct gene expression signatures, which we refer to as FOLR2.IMs and SPP1.IMs (Figure 1D and Figure 5D-5E). Genes such as *LGMN*, *MARCKS* (Figure 5E), *TMEM37* and *MERTK* (data not shown) are specifically expressed by IMs, serving as potential markers to distinguish this population from other macrophage/monocyte cell types. It is worth mentioning that these two IM clusters are the main producer of IL10, suggesting their immunoregulatory role in the airspace.

Next, we performed gene ontology and regulon analysis on the monocyte and IM clusters (Figure S2). The most notable GO term was for FCN1.Mono, whose genes were enriched for bacterial defense mechanisms suggesting the importance of this monocyte subtype during pulmonary bacterial infections (Figure S2A). The GO analysis also indicates the exogenous antigen presentation function of FOLR2.IMs and the inflammatory response gene ontology terms were enriched in SPP1.IMs (Figure S2A).

Pseudotime analysis suggests monocytes differentiate into AMs and IMs

ScRNA-seq captures all cell types at a given time point, producing a series of expression profiles across cells as they transition from one state to another. To define the transitional processes of macrophage development, we performed pseudotime analysis using all the monocytes, AMs, and IMs in our dataset (30, 34-36). The trajectory analysis revealed starting points of origin in the monocyte cluster (purple region and white open dots, Figure 6A), revealing two directions of differentiation: one toward IMs and the other toward cycling AMs and AMs.S4 (Figure 6A-6B). AMs.S4 appears to have its own starting branch point leading to all other AM clusters (Figure 6A-6B), which suggests that AMs.S4 may contribute to all other AM subtypes in the airspace.

To pinpoint the genes that change as a function of pseudotime trajectory, DEG analysis was applied. As monocytes differentiate into clusters of AMs and IMs, different genes are up-regulated sequentially, likely as a cause or consequence of changes in their cell state over time (Figure 6C). Genes for cell proliferation and myeloid cell function were among the top DEGs (Figure 6C). For example, *ISG15* is highly expressed during the later progression of pseudotime and is mainly upregulated in IFN-reacting AMs. Whereas *another among the top 10 DEGs, MT2A,* change along pseudotime and associate with Metallothionein AMs. Interestingly, AM superclusters AMs.S1 and AMs.S2 seem to evolve much later than their counterparts AMs.S3 and AMs.S4, even though they share 4 subclusters and most gene expression, except *IFI27* and *APOC2* (Figure 6C).

Although scRNA-seq has revolutionized our knowledge of macrophage heterogeneity, there are instances where investigators need to isolate bulk populations to perform other functional or morphological analyses. Therefore, flow cytometric antibodies that can be used to sort and enrich bulk AMs and recruited monocytes are in need. Using flow

cytometry, we demonstrate that, in addition to high side-scatter for AMs and low side-scatter for monocytes, investigators can use antibodies against CD43 (SPN) and CD169 to define AMs and CD93, CD36, and CD14 to identify monocytes (Figure 6D) (37-41). ScRNA-seq and TotalSeq (for single-cell level protein expression assessment) also support the use of CD169 and CD43 for AMs compared to CD93, CD36, and CD14 for monocytes (Figure 6E and Figure S3). Thus, several forms of technology validate that higher expression of CD43 and CD169 in BAL cells define AMs over recruited monocytes.

Bivariate expression of *IFI27* and *APOC2* also exist in other AM subtypes

Although our unbiased AM analysis in Figure 2B resulted in 8 clusters, four AM superclusters, and four other clusters (i.e., Chemokine, Metallothionein, Interferon, and Cycling AMs), we questioned whether the four other non-supercluster AMs express all four combinations of *IFI27* and *APOC2*. Indeed, we observed a quadrant distribution of *IFI27* and *APOC2* for Chemokine, Metallothionein, and Interferon AMs (Figure 7A). This suggests that non-supercluster AMs clusters are a part of the four AM superclusters. And the reason this was not observed in the analysis in Figure 2B is because of the overpowering expression of other functional genes in the non-supercluster AMs. Based on this observation, we hypothesize that discrete compartments in the lung dictate the expression of *IFI27* and *APOC2* and within these compartments, a family of AMs, all the AM subtypes, exist to ensure proper physiological functions (Figure 7B).

Lastly, we examined whether AM superclusters occurs in aggregated HC or CF, or for each individual sample. Unbiased analysis of each sample, or combined samples, demonstrated the formation of four unique AM superclusters (Figure S4). Thus, the robust

recurrence of AM superclusters in each sample further validates our forementioned discovery and analysis.

Discussion

AMs have multiple functional roles, including clearing surfactant components, cellular debris, inhaled particulates, and pathogens. Therefore, it is not surprising that in the airspace, there would be multiple AM and monocyte subtypes with different specializations. scRNA-seq technology has revolutionized our understanding of AMs, and we are only beginning to appreciate the depth and complexity of airway macrophages and monocytes in healthy and diseased lungs. Here, we used scRNA-seq to show that AMs form 4 superclusters based on the expression of two genes, *IFI27* and *APOC2*, whether integrated or independent. Each supercluster contains subclusters defined by characteristic DEGs, including those for cytokines and chemokines. In addition to AM superclusters, we observed four other AM clusters. Of particular interest the Chemokine AM cluster, which divides further into four subclusters. Each of these chemokine clusters expressed a combination of chemokines that selectively chemoattract different leukocytes into the environment. Strikingly, there appears to be a defined niche for each AM subtype given the tightly regulated frequency of each AM subtype across all samples.

Why do the superclusters and subclusters exist? In addition to the hypothesis that each AM supercluster exists in a discrete compartment of the lung, AM superclusters might regulate their environment in a tissue-specific niche and functionally specific manner. For example, it would be scientifically and clinically interesting to know whether certain CK.AM clusters regulate the leukocyte environment in the airspace during and after infection with respiratory viruses like SARS-CoV-2.

Chemokines are chemoattractant cytokines, and this type of cytokine profiling resembles what has been observed for decades in T cell biology. Each T cell type, such as Th1 and Th2, selectively produces a characteristic set of cytokines. It is known that the differentiation of T cells is driven by distinct transcription factors like T-bet and Gata-3. Likewise, each AM profiled in this study outlined a distinct set of transcription factors, observed by our regulon analysis. Therefore, an important future goal in studying the AM system will be to investigate which transcription factors regulate the individual Chemokine AM subtypes and beyond. Overall, the findings presented here open the door to identify new therapeutic targets for inflammatory lung disease, while significantly enhancing our understanding of the complexity that exists with airspace macrophages.

Other immune cell types that are not normally thought to reside in airway space were also captured within our BALF analysis. For example, different DC types and IMs, which makes this study more inclusive and comprehensive. IMs have been mainly investigated in mouse lung and multiple subtypes of IMs has been reported by different groups (20, 42-45). Among those subtypes, the well described Folr2 positive IMs also exist in our human BAL cell population, suggesting the potential homology across species. One other IM population is also found and denoted as SPP1.IMs for their specific expression of SPP1, which is mainly expressed in mouse AMs but not mouse IMs (29). This IM subtype seem to be consistent with previously reported SPP1/MERTK-expressing macrophages in idiopathic pulmonary fibrosis (46). Other than those, we also report several common markers for both human IM populations, for example, LGMN and MARCKS, consistent with previous study (47).

Our study was originally designed to investigate the cellular and transcriptional differences in cells extracted from BALF between a cohort of CF subjects with normal lung function and minimal lung inflammation compared to healthy controls, in order to investigate the impact of aberrant CFTR function. We designed this study to investigate HC versus mild-CF due to many lines of evidence point to impaired innate immune cell function in the CF lung (32, 48-54). However, although we observed some minor differences between CF and HC BALF cells (Figure S5), it is unclear that these differences are true across all CF patients. What is striking is that even though all three CF subjects had two copies of the F508del *CFTR* mutation, when the CF lung is in a healthy-like state, the macrophage, monocyte, IM and DC heterogeneity is highly conserved and its niche appropriately filled. All in all, the CF subjects in our study had preserved lung function and limited inflammation by design. Future studies could include subjects with more significant lung disease to determine the impact of inflammatory milieu on CF immune cell gene expression. Additionally, comparisons of CF subjects with and without lung disease could help identify factors in airspace immune cells that predispose one to disease progression. Overall, several fundamental questions require further investigation, including these clusters' locations, functions, transcription factors that regulate their distinct programming, and whether similar macrophage superclusters can be found in other organs or species.

Materials and Methods

Human Subjects

Four healthy control (HC) subjects and three subjects with CF underwent research bronchoscopy as previously described (55, 56). Briefly, after local anesthesia to the posterior pharynx and intravenous conscious sedation, a flexible fiberoptic bronchoscope was inserted transorally and passed through the vocal cords into the trachea. The bronchoscope was sequentially wedged into tertiary bronchi in the right upper lobe (RUL), right middle lobe (RML), and right lower lobe (RLL), saline was instilled, and BALF was collected. An aliquot of 10 mL of BALF from the RML was used for the studies outlined here.

Sample Processing

Sample processing and library construction were performed immediately after lavage. 1X phosphate-buffered saline (PBS) containing 0.5 mM EDTA and 0.1% bovine serum albumin (BSA) was added to the BALF. Samples were filtered through 100-micron cell strainers and centrifugated at 300 x *g* for five minutes at 4 °C, and supernatants were discarded. TotalSeq antibodies were added at concentrations recommended by the manufacturer and stained for 30 min on ice (57), details in online supplement. After staining, chilled Hank's balanced salt solution (HBSS) supplemented with 0.5% BSA was used to wash the sample. Samples were again filtered through 100-micron cell strainers and spun as above. Supernatants were aspirated, and pellets were resuspended with HBSS plus 0.5% BSA at an approximate concentration of 7.5×10^5 cells/mL. Cell quality and viability were assessed with a Cellometer K2 (Nexcelom Bioscience, Lawrence, MA).

All samples had viability >80%. Single cells were then processed using the Chromium Next GEM Single Cell 3' Platform (10X Genomics, Pleasanton, CA). Approximately 30,000 cells were loaded on each channel with an average recovery rate of 24,000 cells. Libraries were sequenced on NextSeq 500/550 (Illumina, San Diego, CA) with an average sequencing depth of 50,000 reads/cell.

Data Preparation

Raw sequencing reads were demultiplexed, mapped to the GRCh38 human reference genome, and gene-expression matrices were generated using CellRanger v6.1 (10X Genomics, Pleasanton, CA). The following analyses were conducted in R 4.1 (58) and Python 3.6. Seurat package v4.0 was used for downstream data analyses (59), and figures were produced using the package ggplot2. Following a standard workflow, the gene-expression matrix was filtered to discard cells with less than 200 genes, as well as genes that were expressed in less than 3 cells. Samples then underwent quality control to remove cells with either too many or too few expressed genes (average around 2000 and 7000) and cells with too many mtRNA (average around 10%), resulting in a total of 113,213 cells. Then, "SCTransform" was applied with the "glmGamPoi" method to normalize gene expression data (60, 61), and a CLR transformation was applied to normalize protein data within each cell. After individual preparation, all the samples were introduced into a combined Seurat object via “FindIntegrationAnchors” and “IntegrateData” functions (60). Then scaled values of variable genes were then subject to principal component analysis (PCA) for linear dimension reduction. A shared nearest neighbor network was created based on Euclidean distances between cells in multidimensional PC space (the first 50 PC were used) and a fixed number of neighbors

per cell (30 neighbors). This was used to generate a two-dimensional Uniform Manifold Approximation and Projection (UMAP) for visualization as a harmonized atlas to dissect the cell populations in BALF of healthy and CF donors.

TotalSeq Antibodies

TotalSeq antibodies were added at concentrations recommended by the manufacturer and stained for 30 min on ice (57). CD88 (B1046), CD1c (B0160), CD1a (B0402), CD64 (B0162), CD14 (B0081), CD93 (B0446), CD197 (B0148), XCR1 (B0208), CD163 (B0358), CD36 (B0407), CD206 (B0205) HLA-DR (B0159), CD169 (B0206) TotalSeq B (BioLegend, San Diego, CA). TotalSeq antibodies were added to the last four samples (HC3, HC4, CF2, CF3).

Differentially Expressed Genes

Differentially expressed genes (DEGs) were calculated with "FindAllMarkers" function of Seurat in R 4.1 to study the different expression profiles in different cell types and up-regulated genes in HC or CF cohort (58, 59). The "data" matrices of "SCT" assay were used, and the minimal log fold change was set to 0.25. Only genes that were detected in more than 25% of cells in either of the two populations were used to compute the DEGs with the Wilcoxon rank-sum test. Markers were identified as genes exhibiting significant upregulation when compared against all other clusters and defined by having a Bonferroni adjusted p-value < 0.05. The DEGs are ranked by the adjusted p--value to select the top DEGs for downstream analysis.

Cell Type Identification

To identify cell types, "FindClusters" function with the Leiden algorithm with the various resolution from 0.5 to 2.0 in the Seurat package were used for clustering (59). "FindAllMarkers" function was then applied. The top DEGs of individual clusters were examined for well-studied marker genes across literatures and the clusters were then annotated for the most likely identity. For re-clustered AMs and Mono subtypes, "FindClusters" was performed with different resolution until the right resolution is reached so that each cluster has unique gene expression pattern.

Cross-sample Cell Type Replicability

To compare cross-sample and cross-dataset transcriptomes, we performed MetaNeighbor (31) analysis between individual samples from our dataset and publicly available datasets addressing similar questions, Mould et al. and Schupp et al. (18, 19). datasets were downloaded from UCSC Cell Browser (62) (https://ams-supercluster.cells.ucsc.edu) and gene expression omnibus (GEO; GSE151928; GSE145360). SummarizedExperiment objects were constructed using combined "data" matrices of SCT assays and the "variableGenes" function of MetaNeighbor package was used to identify the highly variable genes as input of similarity assessment (31). The cell–cell similarities were represented by the mean area under the receiver operator characteristic curve (AUROC) scores, keeping the original cell type from every dataset. The correlation matrices were then passed to ComplexHeatmap package and clustered with "spearman" method (63). Heatmap splits were applied to highlight the focus of the study.

Gene Ontology Analysis

The gene set functional analysis was conducted with TopGO (64) (Fisher's exact test) to get a general overview of the more represented functional categories with the DEGs in each cluster. Background genes of gene enrichment assays were defined as all the detected genes in datasets and the biological process was selected as the target of ontology analysis. The DEGs of different cell types were used as the enrichment input. Top 5 gene ontology (GO) terms ranked by P value (65) for every cell type was selected and compared with other cell types. Overlaps of those GO terms were removed from the final visualization.

Regulon Analysis

All the cell types were downsampled to maximal 1000 cells. TF activity was analyzed using SCENIC package of R 4.1 and pySCENIC of Python 3.8 for the cell types with downsampled count matrices as the input (33, 66, 67). The regulons and TF activity (area under the curve) for each cell were calculated with motif collection version v9 as well as hg38__refseq r80__10kb_up_and_down_tss.mc9nr.featherdatabases from the cisTarget (https://resources.aertslab.org/cistarget/). The transcription factor regulons prediction was performed using the pySCENIC default parameters (33, 66, 67). The resulting AUC scores per each cell and adjacency matrices were used for downstream quantification of regulon specificity score. Top 5 transcription factor ranked by specificity score for every cell type was selected and compared with other cell types. Overlaps of those transcription factors were removed from the final visualization.

Pseudotime Analysis

The package Monocle3 was performed to analyze single-cell trajectories (30, 34-36). Pseudotime values were determined for all myeloid cells except DCs, which have a different ontogeny. All the cell types were downsampled to maximal 1000 cells. Then Monocle3 was used to learn the sequence of gene expression changes each cell must go through as part of a dynamic biological process and then arrange each cell in the trajectory (30, 34-36). Two nodes of monocytes population were set as the start allowing the visualization of the differentiation process of monocytes to macrophages. To investigate the dynamic gene expression between monocytes and macrophages, DEGs over the pseudotime were calculated by the "graph_test" function. By ranking the P value, top 100 pseudotime DEGs were identified. By further downsampling the cell types to maximal 100 cells, the expression levels of the top 100 pseudotime DEGs were then visualized and the top 10 were labeled.

Flow Cytometry

BAL cells were stained with the following monoclonal Abs: anti human CD45 BUV805 clone HI30, anti-human CD14 V500, clone MϕP9, anti-human HLA-DR APC-Cy7, clone L243, anti-human CD206 PerCp-Cy5.5, clone 15-2, anti-human CD36 APC, clone AC106, anti-human CD169 FITC and anti-human CD93 PE-Cy7. The viability dye DAPI (#D9542, Sigma, Milwaukee, WI) was added immediately before each sample acquisition on a BD Symphony A3 analyzer (BD Biosciences, San Diego, CA). Data were analyzed using FlowJo (Tree Star, Ashland, OR). Antigen-specific antibodies and isotype controls

were obtained from BioLegend (San Diego, CA), eBioscience (Ebioscience, AUT) and BD Biosciences (San Diego, CA).

Data Availability

The sequencing raw data and processed data used in this paper are available in Gene Expression Omnibus, accession code GSE193782 (not released until publication). The processed data is also available for online visualization at https://ams-supercluster.cells.ucsc.edu (visualization for Figure 2-4 can be viewed with the "AMs Subset UMAP" layout).

Study Approval

All subjects provided written informed consent, and this study was approved by the Institutional Review Board of Dartmouth-Hitchcock Medical Center (protocol #22781).

Author Contributions

XL, AA, and CJ prepared the manuscript; XL, FK and CJ performed the experiments; DA and AA enrolled all subjects, collected samples, and curated clinical data; DM processed clinical samples; XL performed the scRNA-seq and other bioinformatics analyses; all authors provided intellectual input, critical feedback, discussed results, and designed experiments.

Acknowledgement

This work was supported by National Institutes of Health grants R01 HL115334, R01 HL135001, and R35 HL155458 (CJ). 10x Genomics and Illumina sequencing workflows were performed in collaboration with the Genomics Shared Resource and Single Cell Genomics Cores at Dartmouth, which receives support from NCI CCSG grant P30 CA023108, NIGMS P20 GM130454, and S10 OD025235.

Thank you to Dr. Tor Wager and William T. King for critical feedback.

References

1. W. Fujii *et al.*, Alveolar macrophage transcriptomic profiling in COPD shows major lipid metabolism changes. *ERJ Open Res* **7** (2021).
2. M. Hetzel, M. Ackermann, N. Lachmann, Beyond "Big Eaters": The Versatile Role of Alveolar Macrophages in Health and Disease. *Int J Mol Sci* **22** (2021).
3. W. J. Janssen *et al.*, Fas Determines Differential Fates of Resident and Recruited Macrophages During Resolution of Acute Lung Injury. *Am J Respir Crit Care Med* 201011-1891OC [pii]

10.1164/rccm.201011-1891OC (2011).

4. K. J. Mould *et al.*, Cell Origin Dictates Programming of Resident versus Recruited Macrophages during Acute Lung Injury. *Am J Respir Cell Mol Biol* **57**, 294-306 (2017).
5. K. J. Mould, N. D. Jackson, P. M. Henson, M. Seibold, W. J. Janssen, Single cell RNA sequencing identifies unique inflammatory airspace macrophage subsets. *JCI Insight* **4** (2019).
6. S. Yona *et al.*, Fate mapping reveals origins and dynamics of monocytes and tissue macrophages under homeostasis. *Immunity* **38**, 79-91 (2013).
7. C. Jakubzick *et al.*, Minimal differentiation of classical monocytes as they survey steady-state tissues and transport antigen to lymph nodes. *Immunity* **39**, 599-610 (2013).
8. A. V. Misharin *et al.*, Monocyte-derived alveolar macrophages drive lung fibrosis and persist in the lung over the life span. *J Exp Med* **214**, 2387-2404 (2017).
9. M. Fricker, P. G. Gibson, Macrophage dysfunction in the pathogenesis and treatment of asthma. *Eur Respir J* **50** (2017).

10. S. L. O'Beirne *et al.*, Alveolar Macrophage Immunometabolism and Lung Function Impairment in Smoking and Chronic Obstructive Pulmonary Disease. *Am J Respir Crit Care Med* **201**, 735-739 (2020).

11. H. I. McFarland, A. S. Rosenberg, Skin allograft rejection. *Current protocols in immunology / edited by John E. Coligan ... [et al.]* **Chapter 4**, Unit 4 4 (2009).

12. L. V. Deriy *et al.*, Disease-causing mutations in the cystic fibrosis transmembrane conductance regulator determine the functional responses of alveolar macrophages. *J Biol Chem* **284**, 35926-35938 (2009).

13. T. L. Bonfield, Macrophage Dysfunction in Cystic Fibrosis: A Therapeutic Target to Enhance Self-Immunity. *Am J Respir Crit Care Med* **192**, 1406-1407 (2015).

14. J. Hey *et al.*, Epigenetic reprogramming of airway macrophages promotes polarization and inflammation in muco-obstructive lung disease. *Nat Commun* **12**, 6520 (2021).

15. E. M. Bruscia, T. L. Bonfield, Cystic Fibrosis Lung Immunity: The Role of the Macrophage. *J Innate Immun* **8**, 550-563 (2016).

16. X. Li, K. Rawat, C. V. Jakubzick, Targeting resident macrophages in cancer. *Nat Immunol* **22**, 1078-1079 (2021).

17. E. D. María Casanova-Acebes, Andrew Leader, Jessica LeBerichel, Jovan Nikolic, Blanca M. Morales, Markus Brown, Christie Chang, Leanna Troncoso, Steven T. Chen, Ana Sastre-Perona, Matthew D. Park, Alexandra Tabachnikova, Maxime Dhainaut, Pauline Hamon, Barbara Maier, Catherine M. Sawai, Esperanza Agulló-Pascual, Markus Schober, Brian Brown, Boris Reizis, Thomas Marron, Ephraim Kenigsberg, Christine Moussion, Philippe Benaroch, Julio A. Aguirre-Ghiso, and Miriam Merad, Tissue-resident macrophages provide a pro-tumorigenic niche to early NSCLC cells. *Nature* (2021).

18. K. J. Mould *et al.*, Airspace Macrophages and Monocytes Exist in Transcriptionally Distinct Subsets in Healthy Adults. *Am J Respir Crit Care Med* **203**, 946-956 (2021).

19. J. C. Schupp *et al.*, Single-Cell Transcriptional Archetypes of Airway Inflammation in Cystic Fibrosis. *Am J Respir Crit Care Med* **202**, 1419-1429 (2020).

20. S. L. Gibbings *et al.*, Three Unique Interstitial Macrophages in the Murine Lung at Steady State. *Am J Respir Cell Mol Biol* **57**, 66-76 (2017).

21. A. C. Villani *et al.*, Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356** (2017).

22. M. Collin, V. Bigley, Human dendritic cell subsets: an update. *Immunology* **154**, 3-20 (2018).

23. M. Collin, N. McGovern, M. Haniffa, Human dendritic cell subsets. *Immunology* **140**, 22-30 (2013).

24. Y. Cai *et al.*, Single-cell transcriptomics of blood reveals a natural killer cell subset depletion in tuberculosis. *EBioMedicine* **53**, 102686 (2020).

25. C. C. Brown *et al.*, Transcriptional Basis of Mouse and Human Dendritic Cell Heterogeneity. *Cell* **179**, 846-863 e824 (2019).

26. S. L. Gibbings *et al.*, Transcriptome analysis highlights the conserved difference between embryonic and postnatal-derived alveolar macrophages. *Blood* **126**, 1357-1366 (2015).

27. E. L. Gautier *et al.*, Gene-expression profiles and transcriptional regulatory pathways that underlie the identity and diversity of mouse tissue macrophages. *Nat Immunol* **13**, 1118-1128 (2012).

28. X. Jin *et al.*, Characterization of dendritic cell subtypes in human cord blood by single-cell sequencing. *Biophysics Reports* **5**, 199-208 (2019).

29. S. M. Leach *et al.*, Human and Mouse Transcriptome Profiling Identifies Cross-Species Homology in Pulmonary and Lymph Node Mononuclear Phagocytes. *Cell Rep* **33**, 108337 (2020).

30. C. Trapnell *et al.*, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381-386 (2014).

31. M. Crow, A. Paul, S. Ballouz, Z. J. Huang, J. Gillis, Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat Commun* **9**, 884 (2018).

32. S. J. Gardai *et al.*, Cell-surface calreticulin initiates clearance of viable or apoptotic cells through trans-activation of LRP on the phagocyte. *Cell* **123**, 321-334 (2005).

33. S. Aibar *et al.*, SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14**, 1083-1086 (2017).

34. J. Cao *et al.*, The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496-502 (2019).

35. X. Qiu *et al.*, Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* **14**, 979-982 (2017).

36. X. Qiu *et al.*, Single-cell mRNA quantification and differential analysis with Census. *Nat Methods* **14**, 309-315 (2017).

37. A. Bharat *et al.*, Flow Cytometry Reveals Similarities Between Lung Macrophages in Humans and Mice. *Am J Respir Cell Mol Biol* **54**, 147-149 (2016).

38. Y. A. Yu *et al.*, Flow Cytometric Analysis of Myeloid Cells in Human Blood, Bronchoalveolar Lavage, and Lung Tissues. *Am J Respir Cell Mol Biol* 10.1165/rcmb.2015-0146OC (2015).

39. Y. A. Yu, R. M. Tighe, Isolation and Characterization of Human Lung Myeloid Cells. *Methods Mol Biol* **1809**, 111-119 (2018).

40. A. N. Desch *et al.*, Flow Cytometric Analysis of Mononuclear Phagocytes in Non-diseased Human Lung and Lung-draining Lymph Nodes. *Am J Respir Crit Care Med* 10.1164/rccm.201507-1376OC (2015).

41. S. L. Gibbings, C. V. Jakubzick, A Consistent Method to Identify and Isolate Mononuclear Phagocytes from Human Lung and Lymph Nodes. *Methods Mol Biol* **1799**, 381-395 (2018).

42. D. Bedoret *et al.*, Lung interstitial macrophages alter dendritic cell functions to prevent airway allergy in mice. *J Clin Invest* **119**, 3723-3738 (2009).

43. S. Chakarov *et al.*, Two distinct interstitial macrophage populations coexist across tissues in specific subtissular niches. *Science* **363** (2019).

44. J. Schyns *et al.*, Non-classical tissue monocytes and two functionally distinct populations of interstitial macrophages populate the mouse lung. *Nat Commun* **10**, 3964 (2019).

45. B. B. Ural *et al.*, Identification of a nerve-associated, lung-resident interstitial macrophage subset with distinct localization and immunoregulatory properties. *Sci Immunol* **5** (2020).

46. C. Morse *et al.*, Proliferating SPP1/MERTK-expressing macrophages in idiopathic pulmonary fibrosis. *Eur Respir J* **54** (2019).

47. P. A. Reyfman *et al.*, Single-Cell Transcriptomic Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis. *Am J Respir Crit Care Med* **199**, 1517-1536 (2019).

48. H. F. Hazlett *et al.*, Altered iron metabolism in cystic fibrosis macrophages: the impact of CFTR modulators and implications for Pseudomonas aeruginosa survival. *Sci Rep* **10**, 10935 (2020).

49. J. L. Bessich, A. B. Nymon, L. A. Moulton, D. Dorman, A. Ashare, Low levels of insulin-like growth factor-1 contribute to alveolar macrophage dysfunction in cystic fibrosis. *J Immunol* **191**, 378-385 (2013).

50. Y. Chen *et al.*, Genome-wide DNA methylation profiling shows a distinct epigenetic signature associated with lung macrophages in cystic fibrosis. *Clin Epigenetics* **10**, 152 (2018).

51. A. Di *et al.*, CFTR regulates phagosome acidification in macrophages and alters bactericidal activity. *Nat Cell Biol* **8**, 933-944 (2006).

52. L. Paemka *et al.*, Monocyte derived macrophages from CF pigs exhibit increased inflammatory responses at birth. *J Cyst Fibros* **16**, 471-474 (2017).

53. B. S. Murphy *et al.*, Characterization of macrophage activation states in patients with cystic fibrosis. *J Cyst Fibros* **9**, 314-322 (2010).

54. P. Ruytinx, P. Proost, J. Van Damme, S. Struyf, Chemokine-Induced Macrophage Polarization in Inflammatory Conditions. *Front Immunol* **9**, 1930 (2018).

55. D. A. Hogan *et al.*, Analysis of Lung Microbiota in Bronchoalveolar Lavage, Protected Brush and Sputum Samples from Subjects with Mild-To-Moderate Cystic Fibrosis Lung Disease. *PLoS One* **11**, e0149998 (2016).

56. D. S. Aridgides *et al.*, Functional and metabolic impairment in cigarette smoke-exposed macrophages is tied to oxidative stress. *Sci Rep* **9**, 9624 (2019).

57. M. Stoeckius *et al.*, Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* **14**, 865-868 (2017).

58. R. C. Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL

. http://www.R-project.org (2014).

59. Y. Hao *et al.*, Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 e3529 (2021).

60. T. Stuart *et al.*, Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821 (2019).

61. C. Ahlmann-Eltze, W. Huber, glmGamPoi: fitting Gamma-Poisson generalized linear models on single cell count data. *Bioinformatics* **36**, 5701-5702 (2021).

62. W. J. Kent *et al.*, The human genome browser at UCSC. *Genome Res* **12**, 996-1006 (2002).

63. Z. Gu, R. Eils, M. Schlesner, Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847-2849 (2016).

64. F. Marini, H. Binder, pcaExplorer: an R/Bioconductor package for interacting with RNA-seq principal components. *BMC Bioinformatics* **20**, 331 (2019).

65. P. Rantakari *et al.*, Fetal liver endothelium regulates the seeding of tissue-resident macrophages. *Nature* **538**, 392-396 (2016).

66. V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, P. Geurts, Inferring regulatory networks from expression data using tree-based methods. *PLoS One* **5** (2010).

67. B. Van de Sande *et al.*, A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat Protoc* **15**, 2247-2276 (2020).

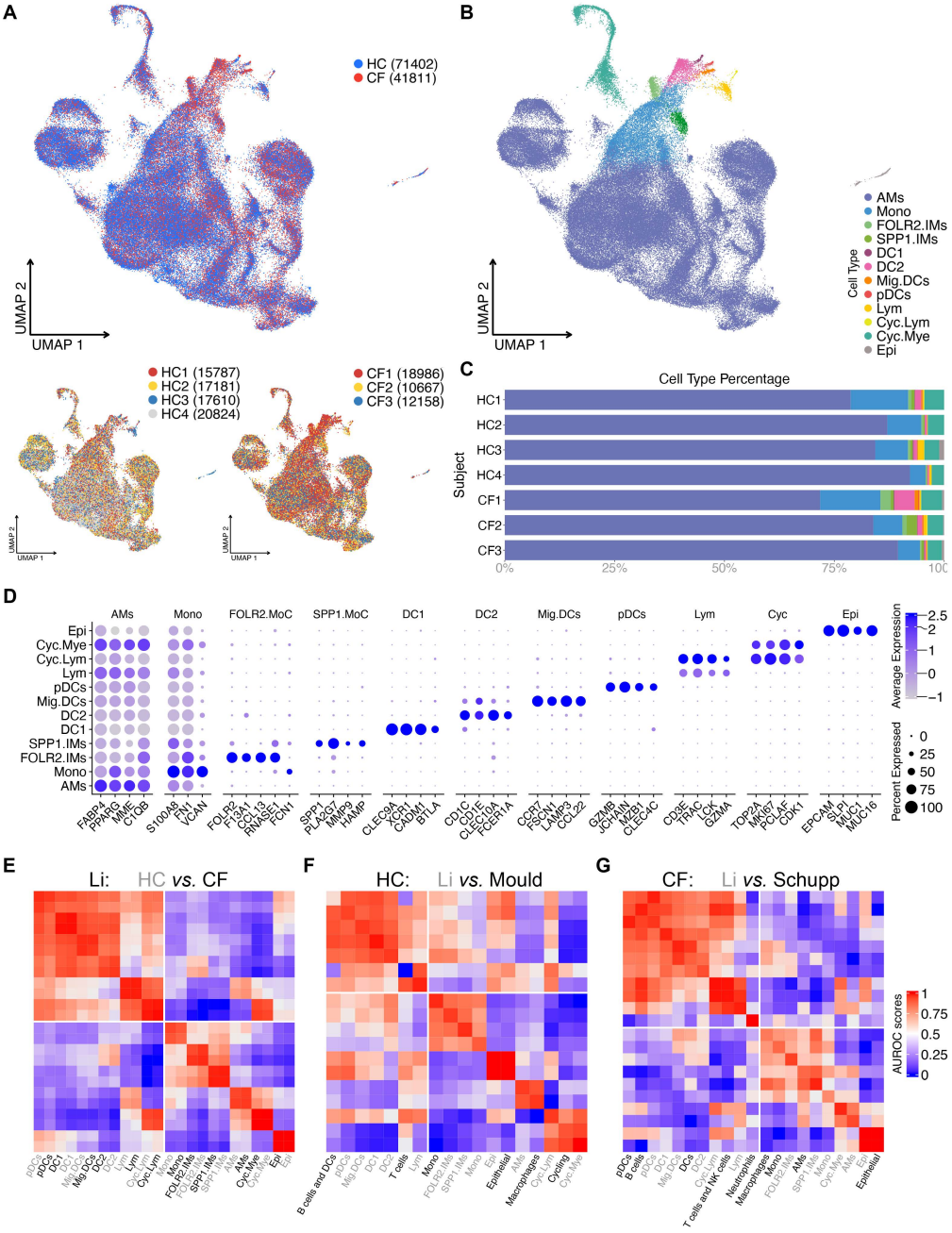

Figure.1 Twelve BAL cell types identified in healthy control (HC) and CF samples.

(A) ScRNA-seq data shows integrated UMAP with HC and CF, along with individual sample distribution. (B) UMAP demonstrating the twelve major cell types: alveolar macrophages (AMs), myeloid and lymphoid cycling cells (Cyc.Mye and Cyc.Lym), FOLR2 and SPP1 interstitial macrophages, monocytes (Mono), Dendritic cell 1 (DC1), Dendritic cell 2 (DC2), Migratory DC (Mig.DC), plasmacytoid DCs (pDCs), lymphocytes (Lym), and epithelial cells (Epi). (C) Percentage bar graph shows the distribution of cell types in each sample. (D) Dot plot shows the expression of curated genes in each individual cell type. (E) Correlation heat map of cell types against all samples in in-house dataset (Li). (F) Correlation heat map of HC cell types in in-house and Mould dataset. The lower quadrant highlights the corresponding cell types. (G) Correlation heat map of HC and CF cell types in in-house and Schupp dataset. The lower quadrant highlights the corresponding cell types.

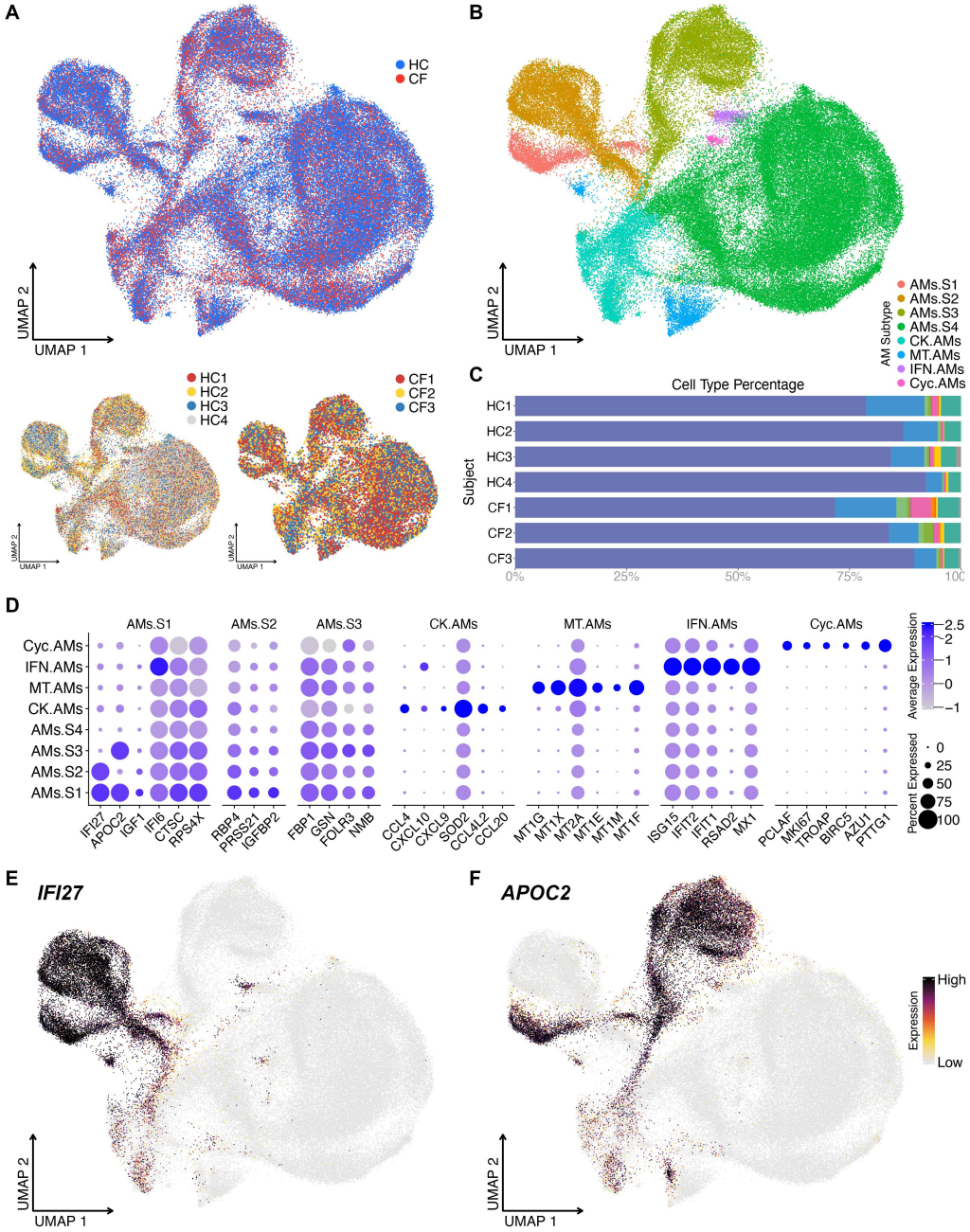

Figure.2 Unbiased analysis of scRNA-seq data illustrates four AM superclusters based on IFI27 and APOC2 expression.

(A) UMAP of re-clustered AMs shows sample source distribution. (B) UMAP shows distinct AM clusters: AM Supercluster 1 (AMs.S1), AMs.S2, AMs.S3 AMs.S4, Metallothionein (MT.AMs), Chemokine (CK.AMs), IFN-reacting (IFN.AMs) and Cycling AM (Cyc.AMs) clusters. (C) Percentage of individual AM clusters for each subject. (D) Dot plot shows the expression of the top 6 DEGs in individual AM clusters, duplicates removed. For example, AMs.S4's top 6 DEGs are already present in the first three AM superclusters; therefore, AMs.S4 is seen on the Y axis but not the X axis. (E) Feature plot shows the expression of IFI27 and APOC2, which define the four AM superclusters.

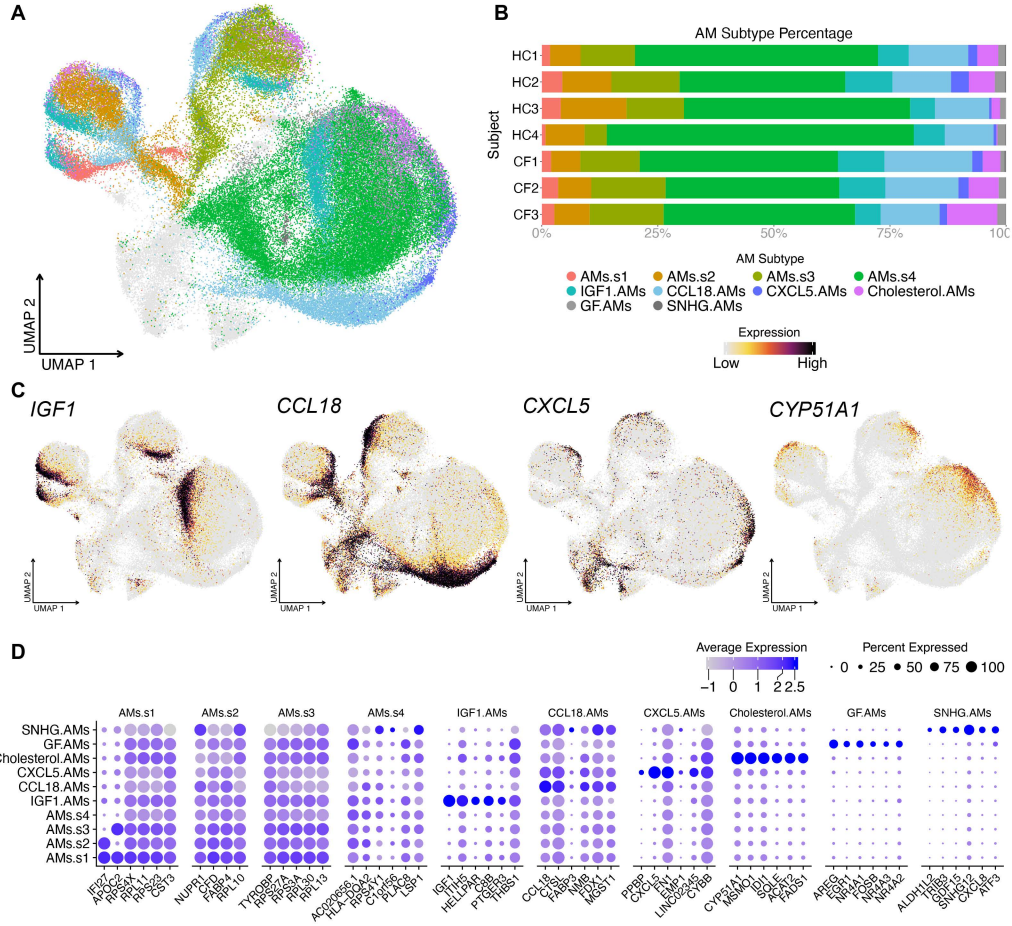

Figure.3 AM superclusters contain 4 shared subclusters and a small-subcluster.

(A) UMAP illustrates the shared AM subclusters. (B) Percentage of individual AM subclusters for each subject. (C) Featured plots show four representative genes for each shared AM subcluster. (D) Dot plot illustrates the expression of the top 6 DEGs in each individual AM subcluster, duplicates removed.

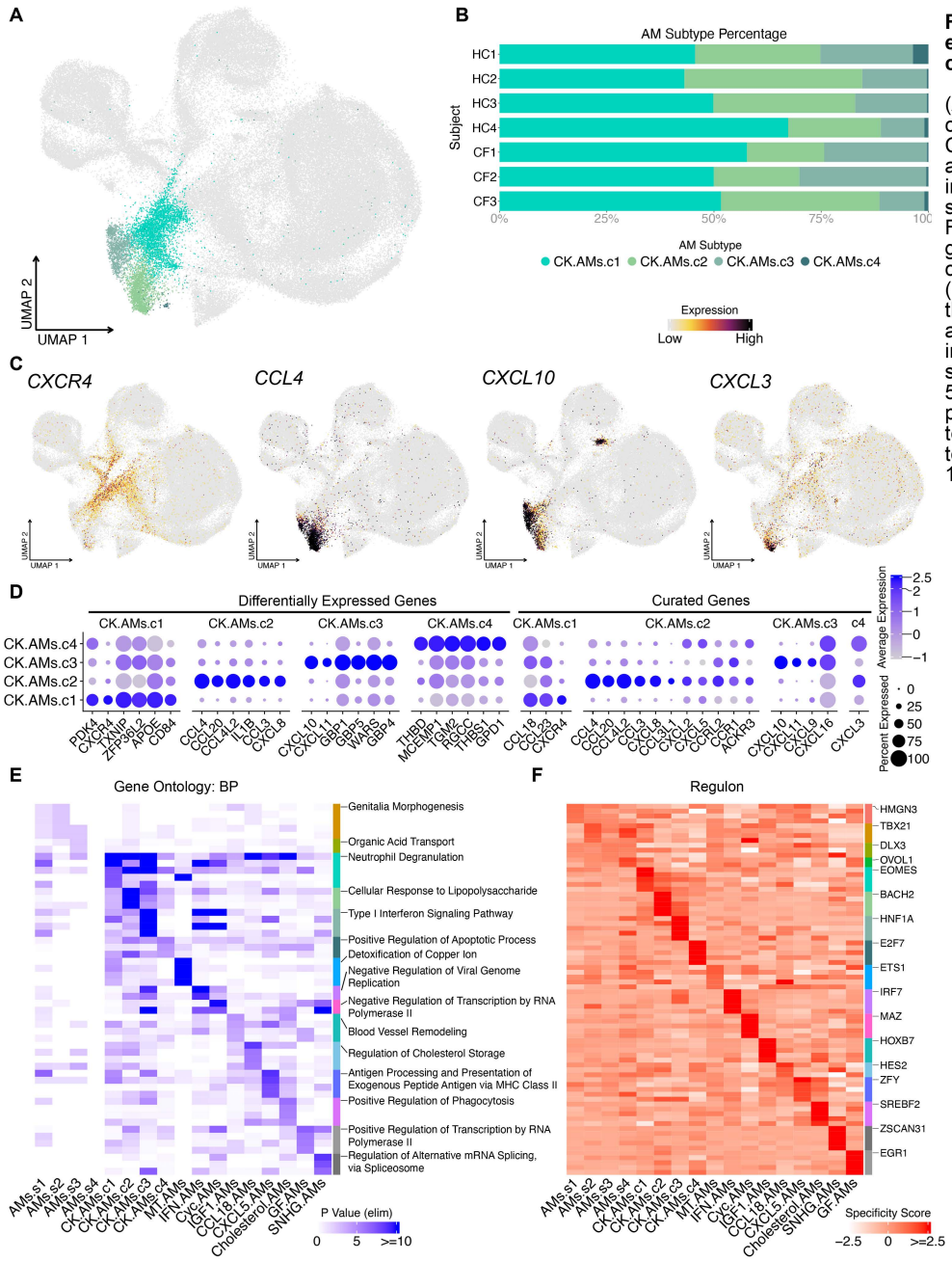

Figure.4 Chemokine AM subcluster express a unique combination of chemokines.

(A) UMAP illustrates distinct chemokine-expressing AM subclusters: CK.AMs.c1, CK.AMs.c2, CK.AMs.c3, and CK.AMs.c4. (B) Percentage of individual chemokine-expressing AM subclusters for each subject. (C) Featured plots show four representative gene expression for individual chemokine-expressing AM subcluster. (D) Dot plot illustrates the expression of the top 6 DEGs (duplicates removed) and curated chemokine genes in individual chemokine-expressing AM subcluster. (E) Heat map shows the top 5 gene ontology terms (biological process) for each AM subtype. Top 1 term is labeled. (F) Heat map shows the top 5 regulons for each AM subtype. Top 1 regulon is labeled.

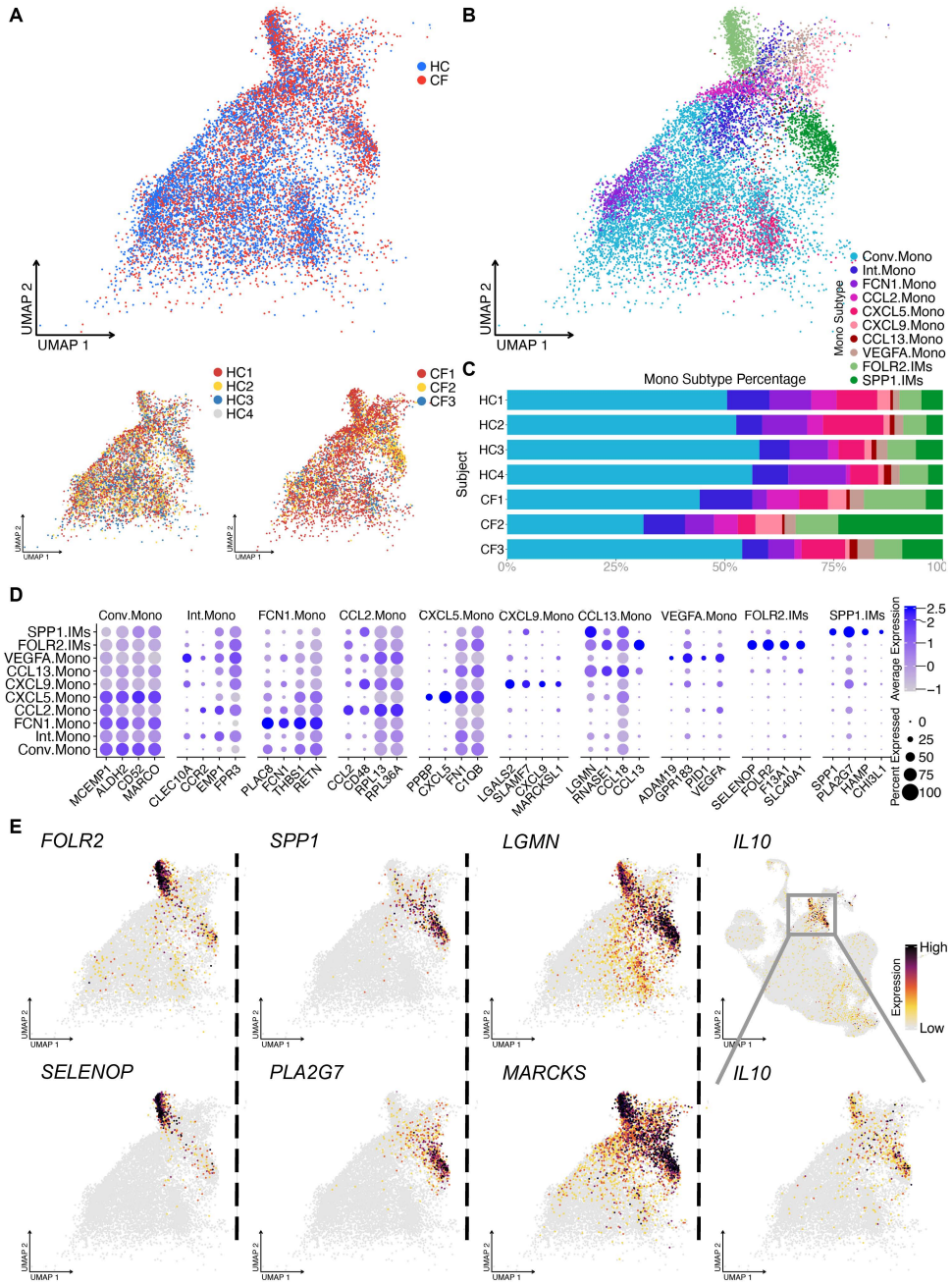

Figure.5 BAL monocytes contain 8 distinct clusters.

(A) UMAP of monocyte shows sample source distribution. (B) UMAP shows distinct monocyte clusters: conventional monocyte cluster (Conv.Mono) and intermediate monocyte cluster (Int.Mono) and others. (C) Percentage bar graph outlines the distribution of individual monocyte clusters in each sample. (E) Dot plot shows the expression of the top 6 DEGs in individual monocytes clusters, duplicates removed. (F) Heat map shows the top 5 gene ontology terms (biological process) for each monocyte subtype. Top 2 terms are labeled. (G) Heat map shows the top 5 regulons for each monocyte subtype. Top 2 regulons are labeled.

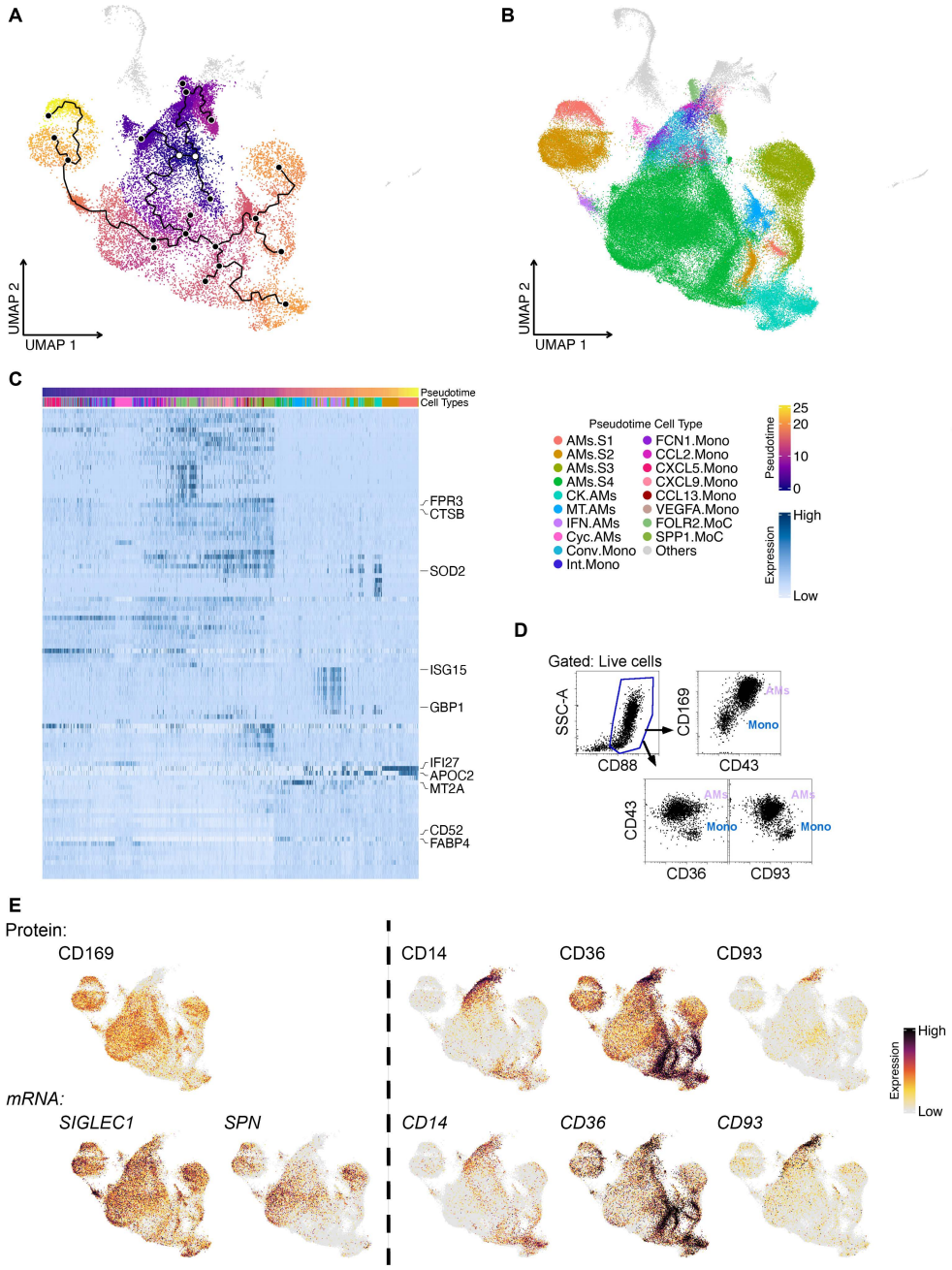

Figure 6. Pseudotime analysis of monocytes and macrophages in BALF.

(A) Pseudotime plot shows the evolvement of monocytes to macrophages with the root as white dots and the principal points as solid black dots. (B) UMAP shows distinct monocyte and macrophage clusters included for Pseudotime analysis. (C) Heat map shows the top 100 pseudotime DEGs across the monocyte-to-macrophage branch. Top 10 DEGs are labelled. (D) Flow cytometry plot illustrates the high expression of CD43 and CD169 on AMs and CD93, CD36 and CD14 on monocytes. (E) Featured plots show representative gene expression and TotalSeq detection in AMs and monocytes.

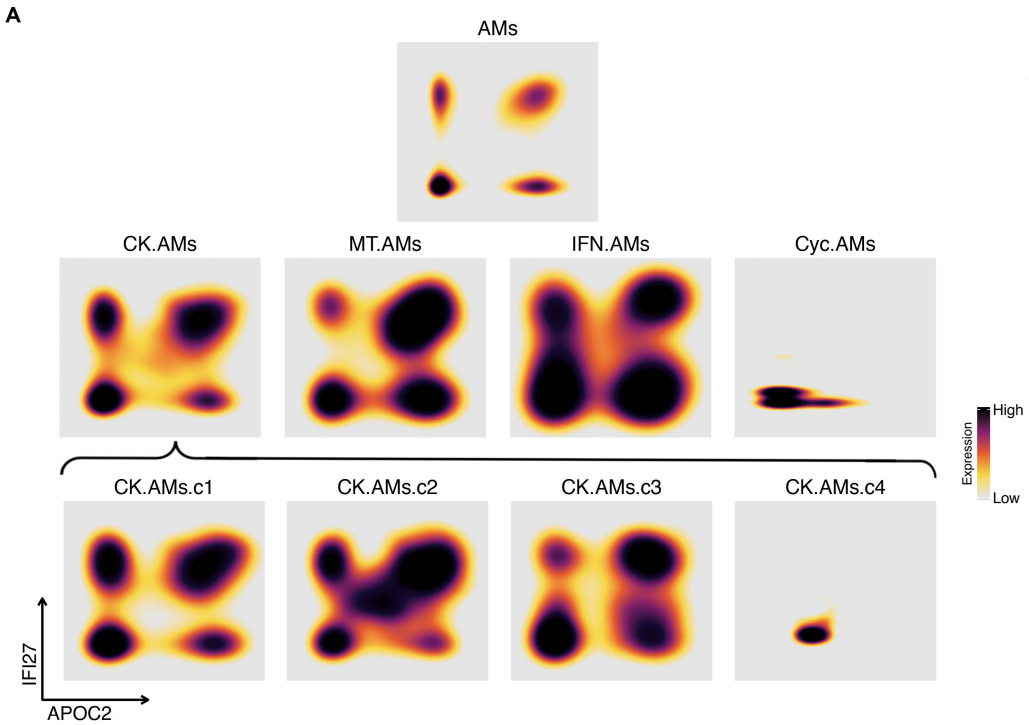

Figure 7. Quadrant distribution of AM subtypes for the bivariate expression of AM subtypes for the of IFI27 and APOC2 and summary of suggested AM subtype classification.

(A) Density plot showing the expression of IFI27 and APOC2. Most AM subtypes show supercluster-like expression profile of IFI27 and APOC2. (B) AM subtype model supported by current data. (C) Alternative AM subtype model suggested in the discussion.

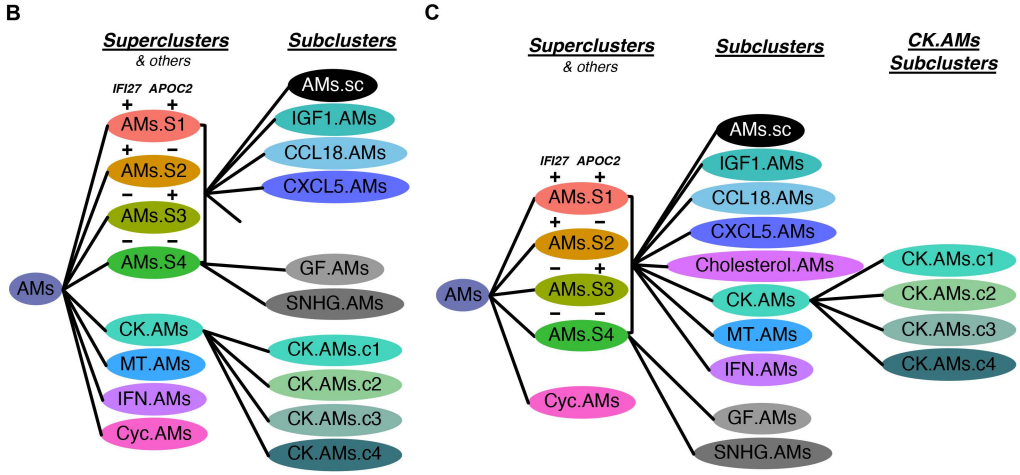

Table 1. Demographic characteristics for recruited subjects

Characteristic	CF (n = 3)	HC (n = 4)
Sex, female % (n)	33% (1)	50% (2)
Average age, years (SD)	28 ± 4.6	29 ± 6.1
FEV1, percent predicted (SD)	89.8 ± 11	
***Pa* colonization % (n)**	67% (2)	
Staph aureus colonization % (n)	33% (1)	
F508del/F508del % (n)	100% (3)	
CFTR Modulator Use (within 12 months) % (n)	0% (0)	
Dornase alpha use % (n)	100% (3)	
Hypertonic saline use % (n)	67% (2)	
Azithromycin use % (n)	67% (2)	
Inhaled antibiotic use % (n)	0% (0)	
Oral/IV antibiotics within 6 mos % (n)	0% (0)	
Pancreatic insufficiency % (n)	100% (3)	
CFRD % (n)	0% (0)	
CF Liver Disease	0% (0)	

Values are means ± standard deviation (SD); n = number of subjects; FEV1 = forced expiratory volume in 1 second; Pa = *Pseudomonas aeruginosa*; CFRD = CF related diabetes; CF=cystic fibrosis; HC=healthy control.