Online Data Supplement

Single Cell Transcriptional Archetypes of Airway Inflammation in Cystic Fibrosis

Jonas C. Schupp, Sara Khanal, Jose L. Gomez, Maor Sauler, Taylor S. Adams, Geoffrey L. Chupp, Xiting Yan, Sergio Poli, Yujiao Zhao, Ruth R. Montgomery, Ivan O. Rosas, Charles S. Dela Cruz, Emanuela M. Bruscia, Marie E. Egan, Naftali Kaminski, Clemente J. Britto

Materials and Methods

Subject Cohort

A total of nine subjects with a confirmed diagnosis of CF from the Yale Adult CF Program provided sputum samples for this study, five during exacerbation and five during periods of stability. These subjects were recruited during a) Scheduled routine visits (n=5) and b) Unscheduled "sick" visits, in which they reported new respiratory symptoms and were diagnosed with a CF exacerbation (n=4). A CF exacerbation was defined by the emergence of four of twelve signs or respiratory symptoms, prompting a change in therapy and initiation of antimicrobial treatment (modified from Fuchs' criteria (E1)). These criteria included: change in sputum; change in hemoptysis; increased cough; increased dyspnea; malaise, fatigue or lethargy; fever; anorexia or weight loss; sinus congestion; change in sinus discharge; change in chest physical exam; or FEV_1 decrease >10% from a previous value (E1). Individuals without new symptoms and those that did not meet AE criteria were characterized as "CF Stable". Our recruitment period extended through 2019. We also recruited five healthy volunteers (Healthy Controls, HC) to undergo sputum induction according to previous protocols (E2). Since we did not identify significant differences in the gene expression profiles of stable and exacerbation subjects, all CF subjects were grouped as "CF" as compared to healthy control samples for analysis as a group. The study protocol was approved by the Yale University Institutional Review Board and informed consent was obtained from each subject.

Sputum Collection and Processing

CF subjects expectorated sputum spontaneously for our studies. Induced sputum samples were obtained from HC as previously described (E2, E3). Briefly, subjects inhaled nebulized 3%

hypertonic saline for five minutes on three cycles. To reduce squamous cell contamination, subjects were asked to rinse their mouth with water and clear their throat. Expectorated sputum samples were collected into specimen cups and placed on ice. Sputum plug material from HC and CF subjects were selected and weighed prior to washing with 9x their volume of PBS. Samples were incubated in Dulbecco's Phosphate-Buffered Saline (PBS) with agitation for 15 minutes and filtered through 40-micron strainers. Samples were centrifugated at 300 g for five minutes and supernatants were stored at -80°C. The pellets were suspended in RPMI/10%FBS medium with 10% DMSO. Aliquots of 1 ml were saved into cryogenic vials and placed in Nalgene Cryo 1° C Freezing Container (Sigma, St. Louis, MO) overnight at -80°C. Samples were stored in liquid nitrogen the next day. Frozen samples were thawed in a water bath at 37°C, resuspended with 20ml DMEM + 10% heat-inactivated FBS (Life Technologies, USA), then centrifuged at 300g, 5min, 4°C. Supernatant was discarded, cells were resuspended in 2ml DMEM + 10% FCS, passed through a 70µm cell strainer (Fisher Scientific, USA). Non-viable cells and debris were removed from the cell suspensions using a OptiPrep (Iodixanol) density gradient centrifugation according to the manufacturer's protocol (OptiPrep Application Sheet C13 – Strategy 2). In brief, 1.86ml of the cell suspensions were mixed with 40% OptiPrep in DMEM + 10% FCS by repeated gentle inversion, overlaid with a density barrier (density: 1.09g/ml, 780µl OptiPrep in 2.22ml DMEM + 10% FCS), then overlaid with 500µl DMEM + 10% FCS. After centrifugation at 800g, 20min, 4°C, viable cells were collected from the top interface and diluted with 2ml DMEM + 10% FCS, centrifuged at 400g, 5min, 4°C, then resuspended in 1ml PBS + 0.04% BSA (New England Biolabs, USA) and passed through a final 40µm cell strainer (Fisher Scientific, USA). For cell concentrations, cells were stained with Trypan blue and counted on a Countess Automated Cell Counter (Thermo Fisher, USA).

Single Cell Barcoding, Library Preparation, and Sequencing

Single cells were barcoded using the 10x Chromium Single Cell platform, and cDNA libraries were prepared according to the manufacturer's protocol (Single Cell 3' Reagent Kits v3, 10x Genomics, USA). In brief, cell suspensions, reverse transcription master mix and partitioning oil were loaded on a single cell "B" chip, then run on the Chromium Controller. mRNA was reverse transcribed within the droplets at 53°C for 45min. cDNA was amplified for a 12 cycles total on a BioRad C1000 Touch thermocycler. cDNA was size-selected using SpriSelect beads (Beckman Coulter, USA) with a ratio of SpriSelect reagent volume to sample volume of 0.6. For qualitative control purposes, cDNA was analyzed on an Agilent Bioanalyzer High Sensitivity DNA chip. cDNA was fragmented using the proprietary fragmentation enzyme blend for 5min at 32°C, followed by end repair and A-tailing at 65°C for 30min. cDNA were double-sided size selected using SpriSelect beads. Sequencing adaptors were ligated to the cDNA at 20°C for 15min. cDNA was amplified using a sample-specific index oligo as primer, followed by another round of double-sided size selection using SpriSelect beads. For qualitative control purposes, final libraries were analyzed on an Agilent Bioanalyzer High Sensitivity DNA chip. cDNA libraries were sequenced on a HiSeq 4000 Illumina platform aiming for 150 million reads per library. Full de-identified sequencing data for all subjects is available in the gene expression omnibus (GEO) under accession number GSE145360.

Data Processing and Computational Analyses

Basecalls were converted to reads with the implementation mkfastq in the software Cell Ranger (v3.0.2). Read2 files were subject to two passes of contaminant trimming with cutadapt (v2.7): first for the template switch oligo sequence

(AAGCAGTGGTATCAACGCAGAGTACATGGG) anchored on the 5' end; secondly for poly(A) sequences on the 3' end. Following trimming, read pairs were removed if the read 2 was trimmed below 20bp. Subsequent read processing was conducted with the STAR (v2.7.3a) (E4) and its single cell sequencing implementation STARsolo. Reads were aligned to the human genome reference GRCh38 release 31 (GRCh38.p12) from GENECODE (E5). Collapsed unique molecular identifiers (UMIs) with reads that span both exonic and intronic sequences were retained as both separate and combined gene expression assays. Cell barcodes representative of quality cells were delineated from barcodes of apoptotic cells or background RNA based on the following three thresholds: at least 10% of transcripts arising from intron spanning, i.e. unspliced reads indicative of nascent mRNA; more than 750 transcripts profiled; less than 15% of their transcriptome was of mitochondrial origin. Technical summaries related to sequencing and data processing can be found in Supplemental Data file E4.

Data Normalization and Cell Population Identification

UMIs from each cell barcode - irrespective of intron or exon coverage - were retained for all downstream analysis and analyzed using the R package Seurat (version 3.1.1) (E6). Raw UMI counts were normalized with a scale factor of 10,000 UMIs per cell and subsequently natural log transformed with a pseudocount of 1. More than double the cell barcodes were detected in two subjects compared to all other subjects, so cells were randomly downsampled to a maximum of 2,250 cells per subject to avoid predominance of those two subjects. 3000 highly variable genes were identified using the method “vst”, then data was scaled and the total number of UMI and the percentage of UMI arising from mitochondrial genes were regressed out. The scaled values were then subject to principle 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 12 PC were used) and a fixed number of neighbors per cell, which was used to generate a 2-dimensional Uniform Manifold Approximation and Projection UMAP for visualization. For cell type identification, scaled data was clustered using the Leiden algorithm. In addition to general filtering based on quality control variables, a curated multiplet removal based on prior literature knowledge was performed: Cell barcodes were identified as mulitplets if their expression level was higher than 1 in the following marker genes (outside the appropriate cluster): MS4A1 (B cells), CD2 (T cells), VCAN (monocytes), FCGR3B (neutrophil granulocytes), KRT19 (epithelial), and FABP4 (alveolar macrophages). Cell barcodes flagged as multiplets were not included in downstream analyses.

Generation of Cell Type Markers and Differential Expression Between Disease Conditions

In order to evaluated cell-type markers we used Seurat's FindAllMarkers (to calculate log fold changes, percentages of expression within and outside a group, and p-values of Wilcoxon-Rank Sum test comparing a group to all cells outside that specific group including adjustment for multiple testing) and additionally calculated binary classifier system based on diagnostic odd's ratios as described in our earlier work (E7) (Supplemental Data file E2). For each cell type in the data, we identified the genes whose expression was log fold change >= 0.25 greater than the other cells in the data. We then calculated the diagnostics odds ratio (DOR) for each of these genes, where we binarize the expression values by treating any detection of a gene (normalized expression value > 0) as a positive value, and zero expression detection as negative. We included a pseudocount of 0.5 to avoid undefined values, as:

DOR = ((TruePositives + 0.5) / (FalsePositives + 0.5)) / ((FalseNegatives + 0.5) / (TrueNegatives + 0.5)) where True Positives represents the number of cells within the group detected expressing the gene (value > 0), FalsePositives represents the number of cells outside of the group detected expressing the gene, FalseNegatives represents the number of cells within the group with no detected expression, and TrueNegatives represents the number of cells outside of the group with no detected expression of the gene. For differential expression testing between disease conditions, Seurat's implementation of a Wilcoxon-Rank Sum in FindMarkers was used, only testing genes whose expression was log fold change >= 0.25 greater between both disease conditions.

Scoring of regulon activity and pathways

A regulon is defined as a group of target genes regulated by a common transcription factor. To score the activity of each regulon in each cell, we utilized the package pySCENIC (E8) with default settings and the following database: cisTarget databases (hg38__refseq-r80__500bp_up_and_100bp_down_tss.mc9nr.feather, hg38__refseq-r80__10kb_up_and_down_tss.mc9nr.feather) and the transcription factor motif annotation database (motifs-v9-nr.hgnc-m0.001-o0.0.tbl) which were both downloaded from resources.aertslab.org/cistarget/, and the list of human transcription factors (hs_hgnc_tfs.txt) which was downloaded from github.com/aertslab/pySCENIC/tree/master/resources.

In order to calculate pathway activity scores, Gene Ontology (GO; geneontology.org) pathways related to monocyte/macrophage functions were downloaded, then scored using Seurat's AddModuleScore using default settings.

Pseudotime Analysis of PMN and monocytes/macrophages

We observed already in UMAP space that many features in the data were represented by a continuum of increasing phenotypic deviation, e.g. increase of maturation markers in neutrophil granulocyte, maturation from monocytes to macrophages, and gradual increase of classical markers of inflammation in monocytes. Consequently, we sought to implement pseudotime analysis of these continua to assess features rather than relying on traditional group-wise comparisons. Cell barcodes were subsetted to either only neutrophil granulocytes or monocytes/macrophages. Due to major differences in number of cells profiled per subject, PMN were randomly downsampled to a maximum of 200 cell barcodes per subject, and in the Mo/MΦ subgroup to a maximum of 250 cell barcodes per subject. As for the full dataset, data of the subgroups was normalized, variable features were extracted (200 for PMN, 500 for Mo/MΦ), scaled, then subject to PC analysis. PHATE (Potential of Heat-diffusion for Affinity-based Trajectory Embedding) (E9) embedding was performed which is specifically suitable to continua (50 nearest neighbors, 5 PCs, t=50 in Mo/MΦ and t=100 in PMN). Cell barcodes were clustered using the cluster_phate function (k=8) for PMN and the Leiden clustering for Mo/MΦ. Trajectories were identified using Slingshot (E10) on the PHATE embeddings with default settings, and a central starting cluster for the Mo/MΦ. Pseudotime analysis was used to distinguish gene expression trajectories, and in turn, the most extreme phenotypes of these trajectories defined transcriptional archetypes in sputum (E11-E13). Pearson's correlation coefficients and their p values, including Bonferroni adjustment for multiple testing, were calculated between the resulting pseudotime distances of these trajectories and gene expression and the regulon activity scores (Supplemental Data file E2). Gene expression and regulon activity scores correlating with pseudotime values were visualized by heatmaps.

Validation of major cell types by Cytometry Time of Flight (CyTOF)

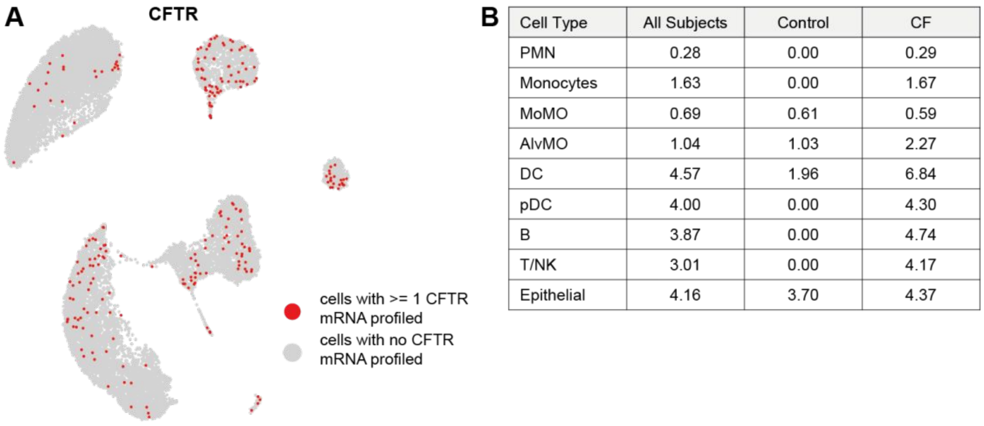
CyTOF-derived fcs files from the study by Yao et al. (E14) were processed using the bead-based Normalizer Release R2013a (E15). Normalized files were then processed in Cytobank (https://premium.cytobank.org/) using gates to select singlets, remove beads and identify live cells. Events identified using this workflow were exported and processed further using the R package cytofkit version 1.12.0 (E16). The Rphenograph function in cytofkit was implemented to cluster cells using cytofAsinh method, with the tsne dimensionality reduction method applied on 80000 events, using k=40. Files were merged using the fixed method and the HLA-DR, CD11b, CD8a, CD20, CD16, MIP-1β, TNF, CD45, CD4, IL-6, CD11c, CD14, Cytokeratin, CD80, CD15, CD163, IFNγ, EGFR, CD66b, IL-8, CD62L and CD56 markers were used in this model. Resulting clusters were manually curated and merged after review of surface marker profiles.

Correlation matrix of immune cell populations comparing sputum and lung cell populations

To identify classifier genes, differential gene expression of immune cell types of this study and analogue cell types from an independent scRNAseq, a dataset of 28 healthy distal lung samples (E7) was established using Seurat's FindAllMarkers with an absolute log fold change threshold of 1 (the lung dataset was downsampled within the FindAllMarkers function using the settings: max.cells.per.ident=1000, seed=7). Classifier genes were filtered such that all genes had a Bonferroni adjusted p-value < 1E-5. For each cell type and each dataset, the top 50 marker genes, ordered by fold change, were selected. We took the intersection of the genes from both datasets as top classifiers (n=154). The average gene expression of these 154 genes were calculated for each cell type per dataset. Spearman correlation matrix was calculated using base

R’s function “cor”. The R package “corrplot” was used to visualize the Spearman correlation matrix. Unsupervised hierarchical complete clustering was performed to order the cell types in the heatmap.

Fig. E1.



Cell Type	All Subjects	Control	CF
PMN	0.28	0.00	0.29
Monocytes	1.63	0.00	1.67
MoMO	0.69	0.61	0.59
AlvMO	1.04	1.03	2.27
DC	4.57	1.96	6.84
pDC	4.00	0.00	4.30
B	3.87	0.00	4.74
T/NK	3.01	0.00	4.17
Epithelial	4.16	3.70	4.37

Fig. E1. *CFTR* expression in CF and healthy control sputum cells. **(A)** UMAP colored by cells in which at least one *CFTR* mRNA molecule was profiled (red). **(B)** Percentages of cells in which at least one *CFTR* mRNA molecule was profiled, separated by cell type; second column ("all subjects") represents the full dataset, which was divided in the third and fourth column by disease state.

Fig. E2.

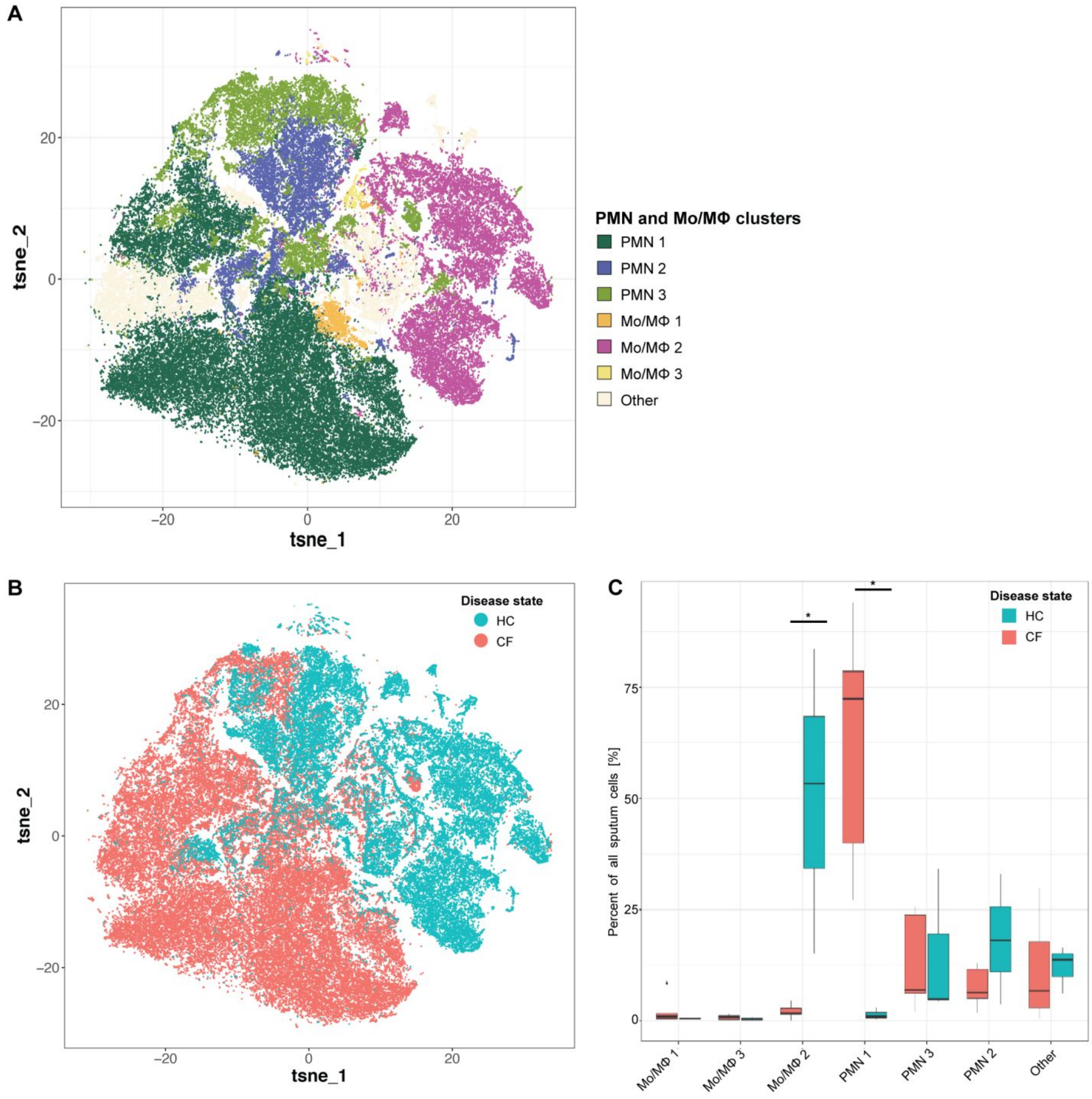
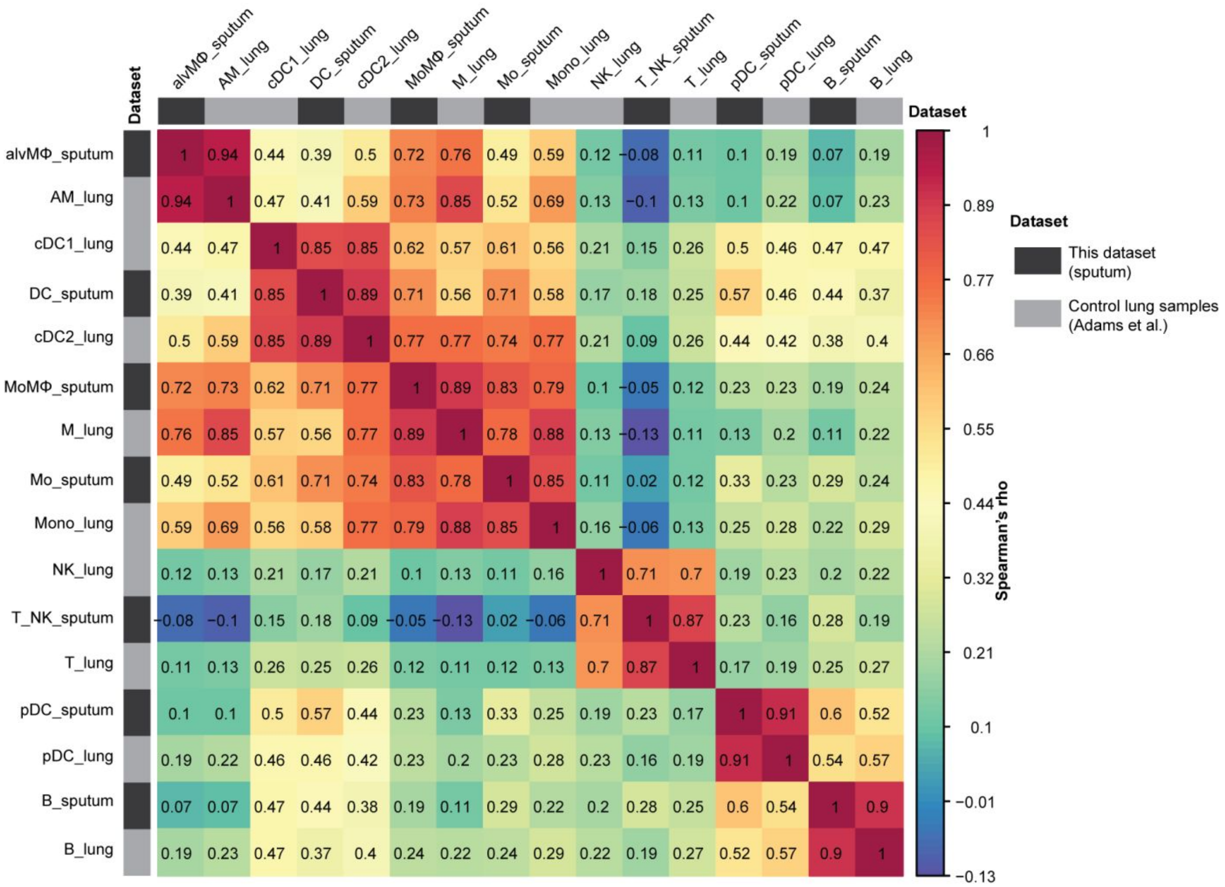


Fig. E2. Validation of the shift in major immune cell types in sputum of CF compared to HC. **(A)** RPhenograph clustering of Sputum CyTOF in patients with cystic fibrosis (CF) and healthy controls (HC) demonstrates differences in the populations of immune cells. The sputum of patients with CF is characterized by high percentages of neutrophils, while sputum from HC is characterized by high percentages of macrophages. **(B)** RPhenograph clustering of Sputum CyTOF according to Healthy Control (HC) and Cystic Fibrosis (CF) status. **(C)** Boxplots showing percentages of Mo/MΦ, PMN, and other to all cells profiled per subject, separated by disease state. Whiskers represent 1.5 x interquartile range (IQR). * $p < 0.05$ determined by a Wilcoxon rank sum test comparing cell percentages of CF patients and controls.

Fig. E3.



	alvMΦ_sputum	AM_lung	cDC1_lung	DC_sputum	cDC2_lung	MoMΦ_sputum	M_lung	Mo_sputum	Mono_lung	NK_lung	T_NK_sputum	T_lung	pDC_sputum	pDC_lung	B_sputum	B_lung
alvMΦ_sputum	1	0.94	0.44	0.39	0.5	0.72	0.76	0.49	0.59	0.12	−0.08	0.11	0.1	0.19	0.07	0.19
AM_lung	0.94	1	0.47	0.41	0.59	0.73	0.85	0.52	0.69	0.13	−0.1	0.13	0.1	0.22	0.07	0.23
cDC1_lung	0.44	0.47	1	0.85	0.85	0.62	0.57	0.61	0.56	0.21	0.15	0.26	0.5	0.46	0.47	0.47
DC_sputum	0.39	0.41	0.85	1	0.89	0.71	0.56	0.71	0.58	0.17	0.18	0.25	0.57	0.46	0.44	0.37
cDC2_lung	0.5	0.59	0.85	0.89	1	0.77	0.77	0.74	0.77	0.21	0.09	0.26	0.44	0.42	0.38	0.4
MoMΦ_sputum	0.72	0.73	0.62	0.71	0.77	1	0.89	0.83	0.79	0.1	−0.05	0.12	0.23	0.23	0.19	0.24
M_lung	0.76	0.85	0.57	0.56	0.77	0.89	1	0.78	0.88	0.13	−0.13	0.11	0.13	0.2	0.11	0.22
Mo_sputum	0.49	0.52	0.61	0.71	0.74	0.83	0.78	1	0.85	0.11	0.02	0.12	0.33	0.23	0.29	0.24
Mono_lung	0.59	0.69	0.56	0.58	0.77	0.79	0.88	0.85	1	0.16	−0.06	0.13	0.25	0.28	0.22	0.29
NK_lung	0.12	0.13	0.21	0.17	0.21	0.1	0.13	0.11	0.16	1	0.71	0.7	0.19	0.23	0.2	0.22
T_NK_sputum	−0.08	−0.1	0.15	0.18	0.09	−0.05	−0.13	0.02	−0.06	0.71	1	0.87	0.23	0.16	0.28	0.19
T_lung	0.11	0.13	0.26	0.25	0.26	0.12	0.11	0.12	0.13	0.7	0.87	1	0.17	0.19	0.25	0.27
pDC_sputum	0.1	0.1	0.5	0.57	0.44	0.23	0.13	0.33	0.25	0.19	0.23	0.17	1	0.91	0.6	0.52
pDC_lung	0.19	0.22	0.46	0.46	0.42	0.23	0.2	0.23	0.28	0.23	0.16	0.19	0.91	1	0.54	0.57
B_sputum	0.07	0.07	0.47	0.44	0.38	0.19	0.11	0.29	0.22	0.2	0.28	0.25	0.6	0.54	1	0.9
B_lung	0.19	0.23	0.47	0.37	0.4	0.24	0.22	0.24	0.29	0.22	0.19	0.27	0.52	0.57	0.9	1

Fig. E3. Concordance of cell type annotations. Correlation matrix of immune cell populations of this study and analogous cell types from an independent scRNA sequencing dataset of distal lung samples, subsetting to the 28 healthy controls. Matrix fields are colored by Spearman's rho, cell types are ordered by unsupervised hierarchical clustering. Annotation bars are highlighting the two different datasets (dark grey: this dataset, light grey: lung samples from healthy controls only from Adams, et al. (7)).

Fig. E4.

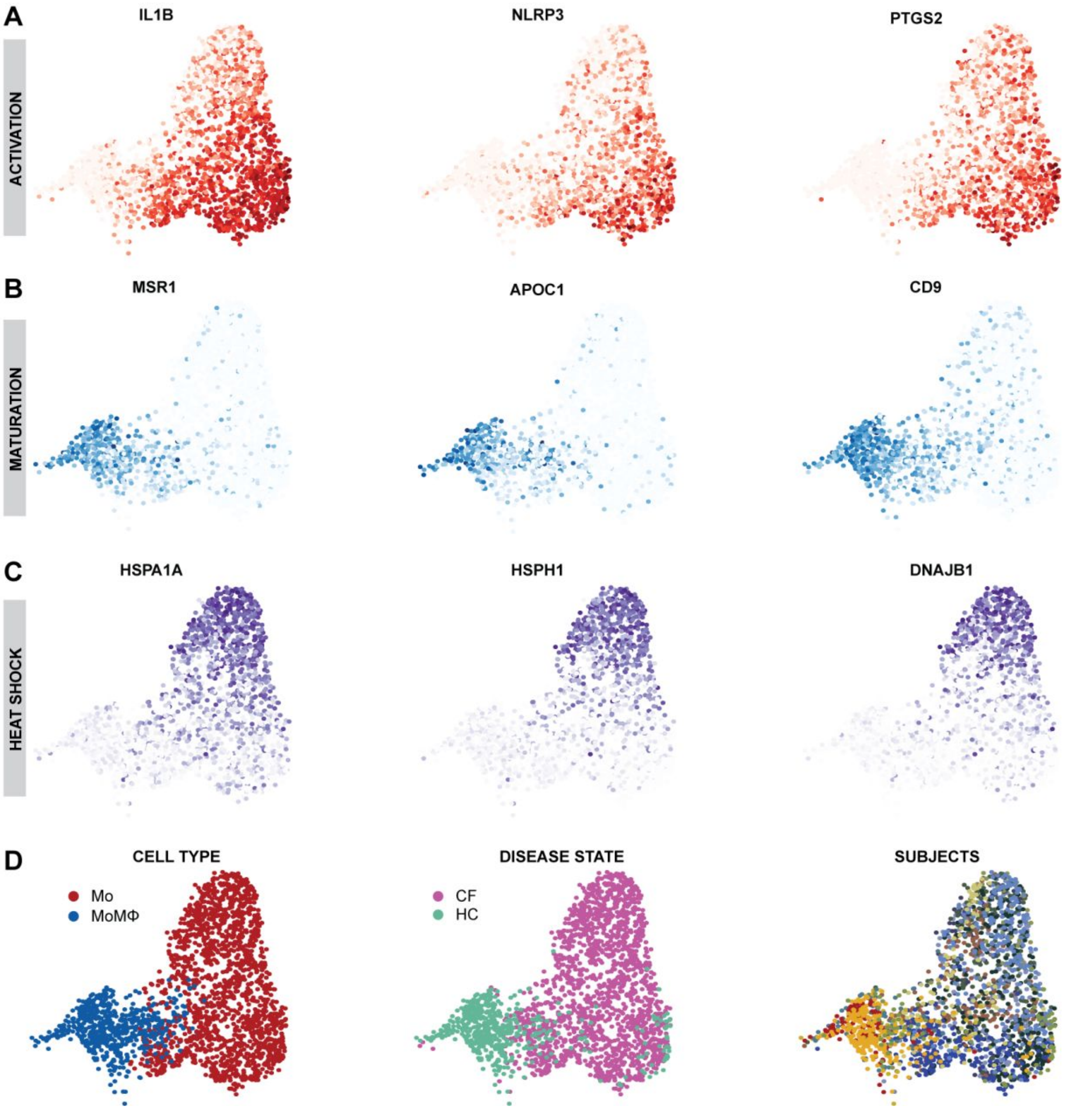


Fig. E4. Expression of selected marker genes of Mo/MoMΦ trajectories on UMAPs. **(A)** UMAP, zoomed in on Mo and MoMΦ, colored by expression of inflammatory genes IL1B, NLRP3, PTGS2. **(B)** UMAP, zoomed in on Mo and MoMΦ, colored by expression of mature macrophage genes MSR1, APOC1, CD9. **(C)** UMAP, zoomed in on Mo and MoMΦ, colored by expression of heat shock genes HSPA1A, HSPH1, DNAJB1. **(D)** UMAP, zoomed in on Mo and MoMΦ, colored by (i) cell type, (ii) disease state, (iii) subjects. CF: Cystic Fibrosis, HC: Healthy Control, Mo: Monocyte; MoMΦ: monocyte-derived macrophage.

Fig. E5.

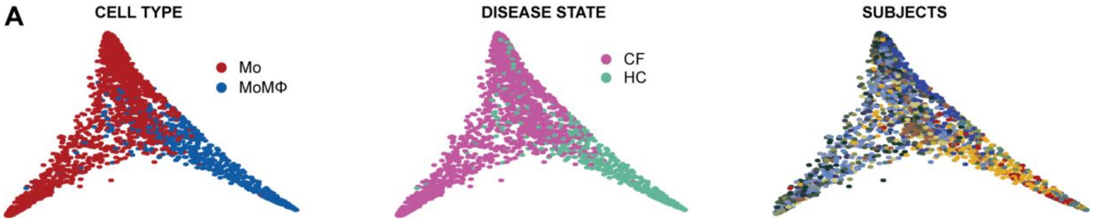


Fig. E5. Additional annotations of Mo/MoMΦ on PHATE embedding. **(A)** UMAP of Mo and MoMΦ colored by (i) Cell type, (ii) Disease state, (iii) Subjects.

CF: Cystic Fibrosis, HC: Healthy Control, Mo: Monocyte; MoMΦ: monocyte-derived macrophage

Fig. E6.

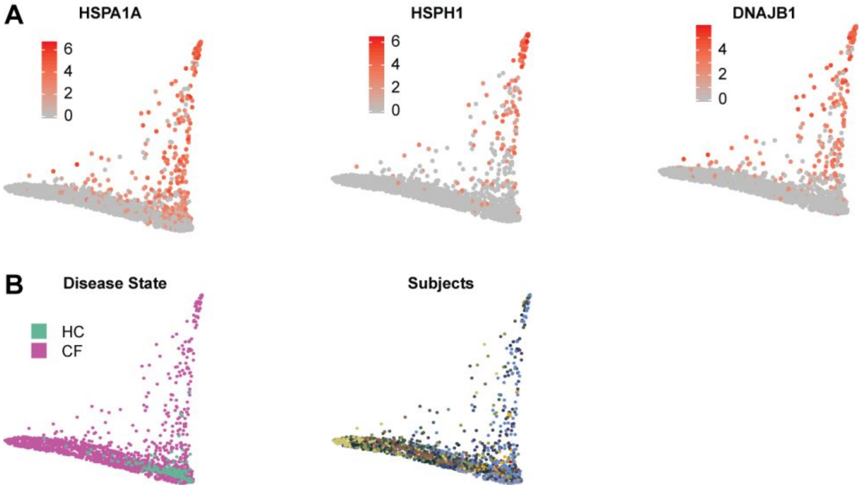


Fig. E6. Additional annotations of PMN on PHATE embedding. **(A)** PHATE of PMN colored by expression of heat shock genes HSPA1A, HSPH1 and DNAJB1. **(B)** PHATE of PMN colored by disease state (HC: Healthy Control, CF: Cystic Fibrosis) and subjects.

Fig. E7.

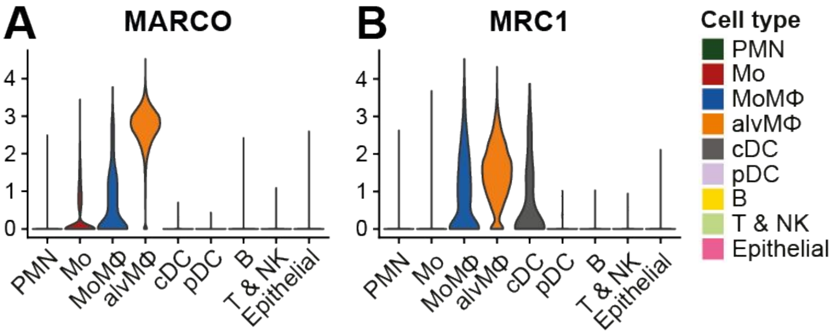


Fig. E7. Violin plots of **(A)** *MARCO* and **(B)** *MRC1*, grouped by cell type.

Fig. E8.

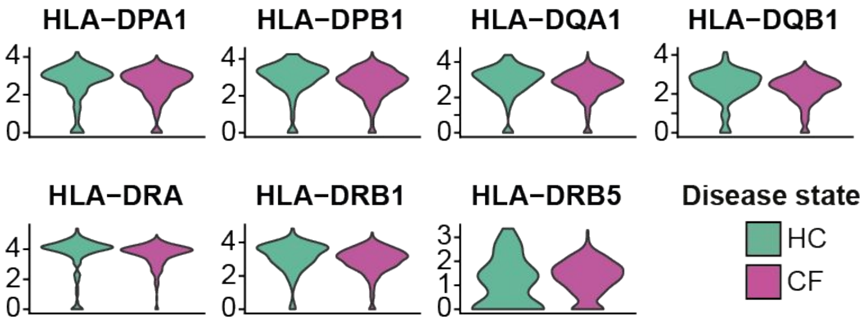


Fig. E8. Violin plots of major histocompatibility complex class 2 genes in B cells, grouped by disease state. For all: p>0.05, i.e. not significantly different.

Fig. E9.

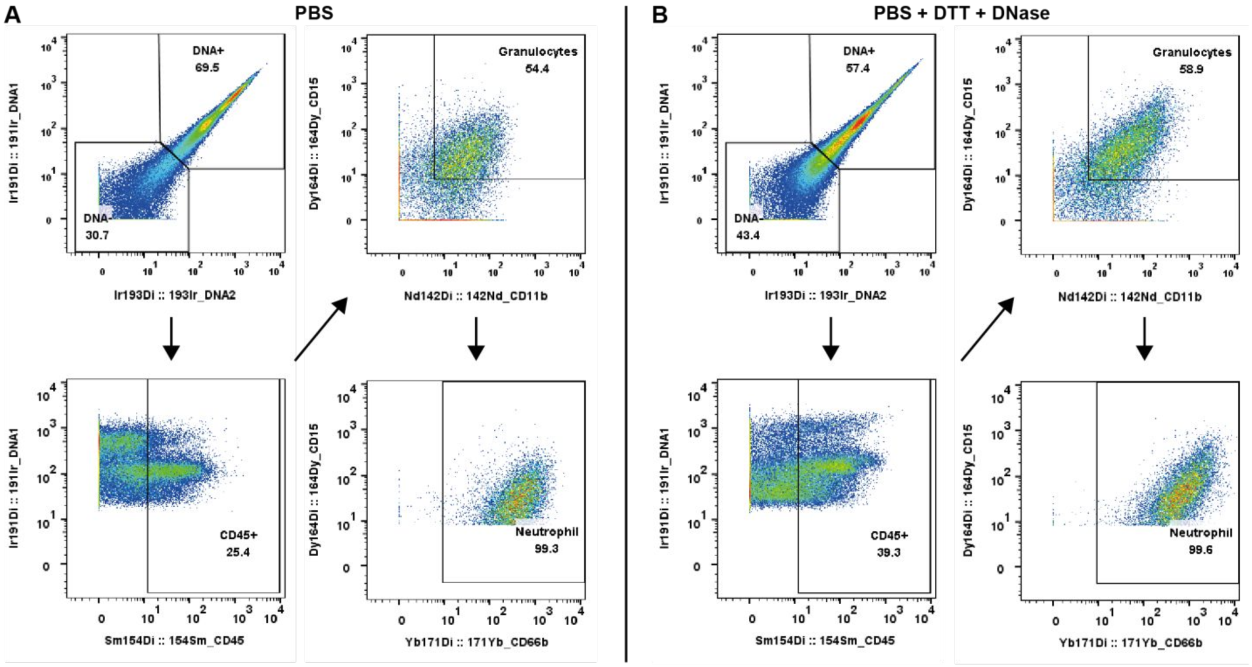


Fig. E9. Viable cell yield using our sputum processing protocol is comparable to previously established approaches for sputum processing (proof-of-principle). Aliquots from the same sample were processed using **(A)** our PBS-only protocol or **(B)** treated sequentially with DNAse (0.56kU/ml, D4527-500KU, Sigma) with gentle agitation for 10 min at room temperature followed by DTT (final concentration 1.5-2uM) with gentle agitation for 10 min at room temperature. Airway cells were incubated with iridium intercalator (125 nM, Fluidigm) to label DNA and analyzed by mass cytometry as previously reported (E14). Representative gating strategy for live cells determined following exclusion of DNA^{lo} cellular debris reflecting enrichment for CD45^{+} (Fluidigm, clone # HI30) CD15^{+} (Fluidigm, clone # W6D3) PMN lineages (CD11b, Clone# M1/7, Longwood and CD66b, self-labeled, Clone# 913542, R&D).

Supplemental Data file E1. Results of Wilcoxon rank-sum test and log transformed diagnostics odds ratio of genes for cell types, subsetting to genes with log transformed fold change > 0.25 for each cell population compared to all other cell populations.

Supplemental Data file E2. Results of Pearson correlation between gene expression and pseudotime distance values within each trajectory.

Supplemental Data file E3. Results of Wilcoxon rank-sum test on gene expression within each cell type comparing CF to HC.

Supplemental Data file E4. Technical summary of all sequenced and processed libraries of this dataset. TSO: template switch oligo.

References:

E1. Fuchs HJ, Borowitz DS, Christiansen DH, Morris EM, Nash ML, Ramsey BW, Rosenstein BJ, Smith AL, Wohl ME. Effect of aerosolized recombinant human DNase on exacerbations of respiratory symptoms and on pulmonary function in patients with cystic fibrosis. The Pulmozyme Study Group. The New England journal of medicine 1994; 331: 637-642.

E2. Yan X, Chu JH, Gomez J, Koenigs M, Holm C, He X, Perez MF, Zhao H, Mane S, Martinez FD, Ober C, Nicolae DL, Barnes KC, London SJ, Gilliland F, Weiss ST, Raby BA, Cohn L, Chupp GL. Noninvasive analysis of the sputum transcriptome discriminates clinical phenotypes of asthma. American journal of respiratory and critical care medicine 2015; 191: 1116-1125.

E3. Esther CR, Jr., Peden Db Fau - Alexis NE, Alexis Ne Fau - Hernandez ML, Hernandez ML. Airway purinergic responses in healthy, atopic nonasthmatic, and atopic asthmatic subjects exposed to ozone. 2011.

E4. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013; 29: 15-21.

E5. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, Barnes I, Berry A, Bignell A, Carbonell Sala S, Chrast J, Cunningham F, Di Domenico T, Donaldson S, Fiddes IT, Garcia Giron C, Gonzalez JM, Grego T, Hardy M, Hourlier T, Hunt T, Izuogu OG, Lagarde J, Martin FJ, Martinez L, Mohanan S, Muir P, Navarro FCP, Parker A, Pei B, Pozo F, Ruffier M, Schmitt BM, Stapleton E, Suner MM, Sycheva I, Uszczynska-Ratajczak B, Xu J, Yates A, Zerbino D, Zhang Y, Aken B, Choudhary JS, Gerstein M, Guigo R, Hubbard TJP, Kellis M, Paten B, Reymond A, Tress ML, Flicek P. GENCODE reference annotation for the human and mouse genomes. Nucleic acids research 2019; 47: D766-D773.

E6. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive Integration of Single-Cell Data. Cell 2019; 177: 1888-1902 e1821.

E7. Adams TS, Schupp JC, Poli S, Ayaub EA, Neumark N, Ahangari F, Chu SG, Raby B, DeIuliis G, Januszyk M, Duan Q, Arnett HA, Siddiqui A, Washko GR, Homer R, Yan X, Rosas IO, Kaminski N. Single Cell RNA-seq reveals ectopic and aberrant lung resident cell populations in Idiopathic Pulmonary Fibrosis. bioRxiv 2019: 759902.

E8. Aibar S, Gonzalez-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine JC, Geurts P, Aerts J, van den Oord J, Atak ZK, Wouters J, Aerts S. SCENIC: single-cell regulatory network inference and clustering. Nat Methods 2017; 14: 1083-1086.

E9. Moon KR, van Dijk D, Wang Z, Gigante S, Burkhardt DB, Chen WS, Yim K, Elzen AVD, Hirn MJ, Coifman RR, Ivanova NB, Wolf G, Krishnaswamy S. Visualizing structure and transitions in high-dimensional biological data. Nat Biotechnol 2019; 37: 1482-1492.

E10. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, Purdom E, Dudoit S. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics 2018; 19: 477.

E11. Mohammadi S, Davila-Velderrain J, Kellis M. A multiresolution framework to characterize single-cell state landscapes. bioRxiv 2019: 746339.

E12. Cutler A, Breiman L. Archetypal analysis. Technometrics 1994; 36: 338-347.

E13. Korem Y, Szekely P, Hart Y, Sheftel H, Hausser J, Mayo A, Rothenberg ME, Kalisky T, Alon U. Geometry of the Gene Expression Space of Individual Cells. PLOS Computational Biology 2015; 11: e1004224.

E14. Yao Y, Welp T, Liu Q, Niu N, Wang X, Britto CJ, Krishnaswamy S, Chupp GL, Montgomery RR. Multiparameter Single Cell Profiling of Airway Inflammatory Cells. Cytometry B Clin Cytom 2017; 92: 12-20.

E15. Finck R, Simonds EF, Jager A, Krishnaswamy S, Sachs K, Fantl W, Pe'er D, Nolan GP, Bendall SC. Normalization of mass cytometry data with bead standards. Cytometry A 2013; 83: 483-494.

E16. Chen H, Lau MC, Wong MT, Newell EW, Poidinger M, Chen J. Cytofkit: A Bioconductor Package for an Integrated Mass Cytometry Data Analysis Pipeline. PLoS Comput Biol 2016; 12: e1005112.