AWARD NUMBER: W81XWH-19-1-0131

TITLE: Identifying Reversible Molecular Networks in Human Pulmonary Fibrosis Using Single Nuclear Transcriptomics and Systems Biology

PRINCIPAL INVESTIGATOR: Jonas Schupp

CONTRACTING ORGANIZATION: Yale University, Office of Sponsored Projects,
25 Science Park, 150 Munson Street, New Haven, CT 06520

REPORT DATE: MAY 2020

TYPE OF REPORT: Annual Report

PREPARED FOR: U.S. Army Medical Research and Development Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE		Form Approved OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAY 2020	Annual Report	5/1/2019 – 04/30/2020

4. TITLE AND SUBTITLE

Identifying Reversible Molecular Networks in Human Pulmonary Fibrosis Using Single Nuclear Transcriptomics and Systems Biology

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-19-1-0131

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Jonas Christian Schupp

E-Mail: Jonas.schupp@yale.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) AND ADDRESS(ES)
Yale University,
Office of Sponsored Projects,
25 Science Park, 150 Munson Street,
New Haven, CT 06520

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Development Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The goals of this study are to identify aberrant cell compositions and aberrant gene expression profiles in cellular subpopulations in differentially affected regions within the IPF lung, to establish cell-type-specific regulatory networks in the microenvironment of IPF and cell-type-specific pathways of disease progression and to discover cell-type-specific biomarkers of disease progression as well as targets for novel therapeutics. To this end, we identified the optimal nuclei isolation and purification method for single nuclei RNA sequencing and are in the progress of generating the final dataset. We developed an automated computational pipeline for data preprocessing of single nuclei RNA sequencing data. We applied this pipeline to a single transcriptome dataset of samples from cystic fibrosis and controls and published this as the manuscript ""Single Cell Transcriptional Archetypes of Airway Inflammation in Cystic Fibrosis".

15. SUBJECT TERMS

Idiopathic pulmonary fibrosis; single nuclei RNA sequencing; regulatory networks; disease progression; lung; biomarker; gene expression; spatial resolution.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
Unclassified	Unclassified	Unclassified	Unclassified	77	19b. TELEPHONE NUMBER *(include area code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

TABLE OF CONTENTS

1. INTRODUCTION

Pulmonary Fibrosis (PF) describes a chronic lung disease in which lung tissue becomes scarred over time in response to microinjuries leading to progressive shortness of breath and ultimately to death within 3-5 years. This condition can be idiopathic, as in idiopathic pulmonary fibrosis (IPF), or secondary to genetic or autoimmune disorders, or to exposure to environmental toxins, chemical warfare, or radiation. IPF is the most common idiopathic form of pulmonary fibrosis that affects approximately 120,000 patients in the US with a steady increase in both incidence and mortality.

Histologically, IPF is characterized by marked fibrosis with or without honeycombing in a predominantly subpleural and paraseptal location with central areas relatively spared. The fibrosis is distributed heterogeneously, with normal lung adjacent to established fibrosis. At the boundary between these regions, there are fibroblast foci, defined by accumulation of immature hyaluronic acid rich matrix underneath epithelial cells undergoing injury and cell death. In response to the cell death, there is an attempt at replacement with type II cell hypertrophy and hyperplasia. Temporal heterogeneity – the presence of acute or active disease (fibroblastic foci with or without epithelization) along with progressive disease (mature fibrotic scar) and non-diseased lung, as well as spatial heterogeneity – the presence of fibrotic lung adjacent to histologically normal lung are a molecular disease mechanism specific to IPF.

The application of high throughput transcript profiling approaches to pulmonary fibrosis discovered that the IPF lung exhibits dramatically different patterns of gene expression with over 2000 significant differentially expressed genes. However, conventional bulk RNA sequencing methods lack the ability to unravel the unique histopathologic features of IPF – temporal heterogeneity, alveolar cell hyperplasia, abundance of myofibroblast foci and aberrant remodeling – on a cellular level, and cell-type specific molecular networks that regulate disease progression are poorly understood. Recent technological advances led to the development of single cell and single nuclei sequencing. The overall objective of this proposal is to create a unique dataset of single nuclei transcriptomes of well-characterized, differentially affected regions within the IPF lung, so we can unravel the microenvironment in IPF by systems biology approaches. Based on these observations and technological innovations, we hypothesized that investigating the single nuclei transcriptomes of well-characterized, differentially affected regions within the IPF lung would allow us to investigate cell-type-specific regulatory networks associated with disease progression and to discover novel, more specific, drugable targets. We aim to identify aberrant cell compositions and aberrant gene

expression profiles in cellular subpopulations in differentially affected regions within the IPF lung. Furthermore, we plan to establish cell-type-specific regulatory networks in the microenvironment of IPF and cell-type-specific pathways of disease progression. Last, we aim to discover cell-type-specific biomarkers of disease progression as well as targets for novel therapeutics.

The successful completion of the specific aims of this application will substantially impact our understanding of pulmonary fibrosis and its disease progression, and to discover cell type-specific candidates for novel therapeutics for patients suffering from PF.

2. KEYWORDS

Idiopathic pulmonary fibrosis; single nuclei RNA sequencing; regulatory networks; disease progression; lung; biomarker; gene expression; spatial resolution.

3. ACCOMPLISHMENTS

What were the major goals of the project?

Goal 1: To identify aberrant cell compositions and aberrant gene expression profiles in cellular subpopulations in differentially affected regions within the IPF lung,

Goal 2: To establish cell-type-specific regulatory networks in the microenvironment of IPF and cell-type-specific pathways of disease progression.

Goal 3: To discover cell-type-specific biomarkers of disease progression as well as targets for novel therapeutics.

What was accomplished under these goals?

Major activities:

a) Single nuclei RNA sequencing experiments to compare isolation and clean-up of nuclei by either Fluorescence activated cell sorting (FACS) sorting or enrichment of high-quality nuclei using a OptiPrep-based density cushion centrifugation
b) Single cell RNA sequencing of all remaining samples after enrichment of high-quality nuclei using a OptiPrep-based density cushion centrifugation – in progress

c) Development of an automated computational pipeline for data preprocessing including identification of valid barcodes and removal of background contamination, and an analytical protocol. This computational pipeline was field-tested on an independently generated dataset of sputum cells from patients with cystic fibrosis and controls.

Specific objectives:

a) Identification of the optimal nuclei isolation and purification method
b) Generating the single nuclei RNA sequencing dataset based on the optimal method of a) – in progress
c) Development of an automated computational pipeline for data preprocessing

Significant results or key outcomes:

First, we performed a comparison of nuclei isolated and cleaned by either a) Fluorescence activated cell sorting (FACS) sorting (n=8) or b) enrichment of high-quality nuclei using an OptiPrep-based density cushion centrifugation (n=4) in two independent single nuclei RNA seq experiments. In both cases, nuclei were isolated using a hypotonic sucrose solution with an additional mechanical tissue disruption using the gentleMACS Dissociator. Regarding FACS sorting, isolated nuclei were stained with DAPI, then sorted at our FACS core facility. The OptiPrep-based density cushion centrifugation was performed such that the isolated nuclei were resuspended in media containing 25% OptiPrep, then overlaid over cushions of 35% and 30% OptiPrep-containing solutions. These layered solutions were centrifuged at 4696g for 20min at 4°C and the nuclei collected at the 35%-30%-interphase. With both methods, nuclei with little contaminant debris were obtained. Both were nuclei preparations were then subjected to single nuclei RNA barcoding, library preparation and sequencing using our standard

AVERAGE	Sorting (n=8)	Density Cushion (n=4)
Raw Reads [*10^6]	73.025	83.675
Reads With Valid Barcodes [%]	97.34%	98.63%
Sequencing Saturation [%]	73.78%	27.48%
Q30 Bases inCBandUMI [%]	98.16%	96.95%
Q30 Bases in RNA read [%]	67.55%	68.43%
TSO Trimmed [%]	6.83%	26.78%
PolyA Trimmed [%]	11.98%	5.88%
Pass Trimming [%]	98.50%	99.33%
Reads Pass Filter [*10^6]	71.9	83.1
Mapped Unique [%]	86.91%	81.49%
Mapped Multi [%]	4.00%	7.43%
Reads too Short [%]	8.46%	10.43%
Splice Junctions [*10^6]	3.375	4.75
Non-Canonical Splices Junctions [	4.10%	2.92%

Table 1: Data processing summary of the two single nuclei RNA seq runs. Fluorescence activated cell sorting (FACS) sorting ("Sorting", n=8) and enrichment of high-quality nuclei using a OptiPrep-based density cushion centrifugation ("Density cushion", n=4)

protocol. The data processing QC measurements in general are were similar and are summarized in Table 1. However, two crucial differences were observed: The median number of UMI of the nuclei purified with the density cushion centrifugation (median 1,328 UMI) was roughly 500 UMI higher compared to the FACS sorted nuclei. Furthermore, the sequencing saturation with a dramatically lower in the density cushion samples (27.48% vs. 73.78%), which means that if we increase the reads per sample the advantage of the density cushion samples with regards to the median number of UMI will further increase. Taken together, the amount of information per nuclear transcriptome is radically higher in the samples isolated by the density cushion and could be further improved by a higher sequencing depths. This clearly favors density-cushion- based nuclei isolation method, which we will use now on the whole cohort.

Second, were processed the data from our single nuclei RNA seq experiment with nuclei isolated using the density cushion method using our newly developed pipeline (see next paragraph). We could identify 35,784 valid single nuclei transcriptome, embedded them in “Uniform manifold approximation and projection” (UMAP) space, clustered them and assigned cell type annotations. The quality of this data was good enough, even with this low sequencing saturation, that we could identify all major cell populations of the human lung (see figure 1). In the single IPF sample from a severely affected lung, we clearly observed a loss of AT1 and AT2 cells, and a shift towards bronchial epithelial cells and fibroblasts, suggesting that our goal of analyzing differences in areas of mild and severe fibrosis is feasible.

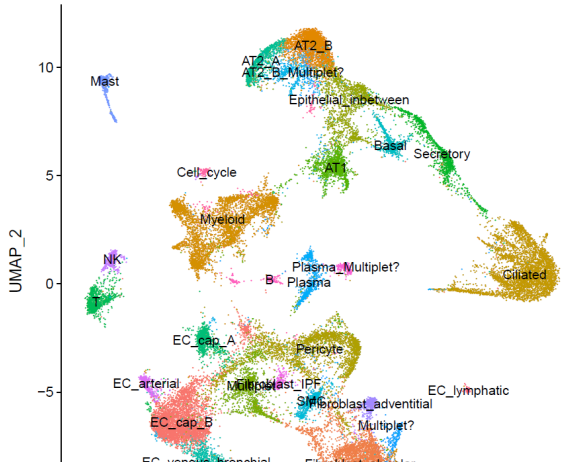


Figure 1: Preliminary embedding of 35,784 single nuclei transcriptomes of the density cushion experiment. Preliminary Uniform manifold approximation and projection (UMAP) embedding of 35,784 of the density cushion trial run without removal of multiplets. Each dot is a single nucleus transcriptome, colored by cell type identity

In the meantime, we developed a computational pipeline for processing of raw sequencing data and analysis protocol, which was field-tested on a dataset of cystic fibrosis samples (see publication “Single Cell Transcriptional Archetypes of Airway Inflammation in Cystic Fibrosis” under “6. Products”). As we automated all major steps of this computational pipeline, we will be able to perform this step on

the final dataset of IPF and control single nuclei RNA seq data within less than two weeks. Our computational pipeline consists of the following steps: Basecalls are converted to reads with the implementation mkfastq in the software “Cell Ranger”. Read2 files are subject to two passes of contaminant trimming with cutadapt: first for the template switch oligo sequence anchored on the 5' end; secondly for poly(A) sequences on the 3' end. Following trimming, read pairs are removed if the read 2 was trimmed below 20bp. Subsequent read processing is conducted with the software “STAR” and its single cell sequencing implementation “STARsolo”. Reads are aligned to the human genome reference GRCh38. Collapsed unique molecular identifiers (UMIs) with reads that span both exonic and intronic sequences are retained as both separate and combined gene expression assays. Cell barcodes representative of quality cells are delineated from barcodes of apoptotic cells or background RNA based on the following three thresholds: fraction of intron spanning UMI, i.e. unspliced reads indicative of nascent mRNA; total number of UMI; fraction of UMI of mitochondrial origin. It is important to mention that the identification of valid barcodes deviates from CellRanger’s standard workflow, but utilizing this threshold-based methods enables to adapt to the lower total UMI counts of nuclei, as alternatively, the majority of valid nuclei barcodes might get discarded. Raw UMI counts are normalized with a scale factor of 10,000 UMIs per cell and subsequently natural log transformed with a pseudocount of 1. Highly variable genes are identified using the method “vst” of the R package Seurat, then data is scaled and the total number of UMI and the percentage of UMI arising from mitochondrial genes are regressed out. The scaled are were then subject to principle component analysis (PCA) for linear dimension reduction. A shared nearest neighbor network is created based on Euclidean distances between cells in multidimensional PC space and a fixed number of neighbors per cell, which is used to generate a 2-dimensional Uniform Manifold Approximation and Projection (UMAP) for visualization. For cell type

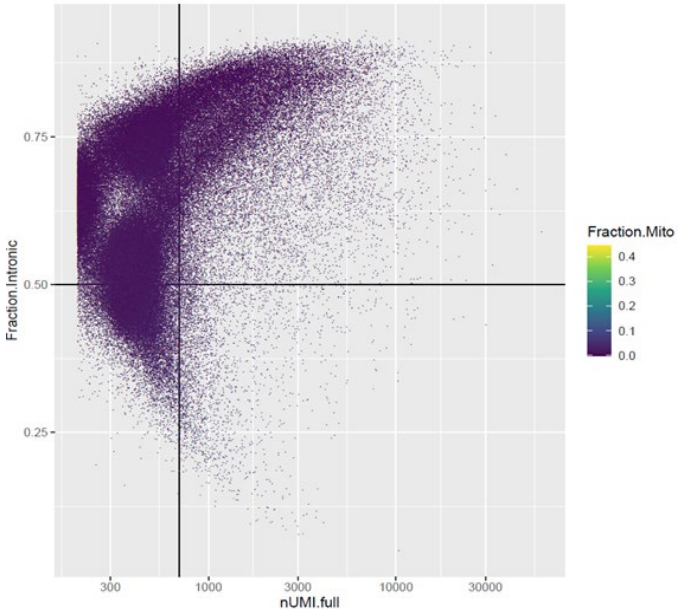


Figure 2: Quality characteristics of barcodes. Plotted are the number of Unique Molecule Identifiers (UMI) on the x-axis versus the fraction of intronic, i.e. unspliced, mRNA per barcode on the y-axis. Valid barcodes can be found in the top right quadrant and were identified using nUMI >700 and faction.intronic >50% as filters. Barcodes are colored by the fraction of UMIs originating from the mitochondrial genome. Valid barcodes have very low fraction of mitochondrial reads (data not shown) and are filtered with a maximum of 5% mitochondrial reads.

identification, scaled data is clustered using the Leiden algorithm. In addition to general filtering based on quality control variables, a curated multiplet removal based on prior literature knowledge is performed. In order to evaluated cell-type markers we use 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 to compare differential gene expression in specific cell types.

Other achievements:

Supported by this grant, we published an editorial "Towards Early Detection of IPF" (see "6. Products") in which we discuss steps necessary for an early identification of patients with IPF or of subjects with increased risk for developing IPF in the context of interstitial lung abnormalities. An early diagnosis will enable treatment of minimal fibrotic lesions, before extensive remodeling and bronchiolization have occurred, which is a critically important mission. We argue for a paradigm shift from focusing on developing cohorts of patients already diagnosed with IPF toward cohorts of individuals highly likely to develop the disease.

What opportunities for training and professional development has the project provided?

The PI, Jonas Schupp, mentored by Naftali Kaminski, has been trained in developing automated computational pipelines as well as expanded his expertise in experimental methods with regards to sample processing for single nuclei RNA sequencing. In addition, his scientific writing skills have been developed, as highlighted by the editorial "Towards Early Detection of IPF". The "Professional development" activities of Jonas Schupp included the participation in the annual conference of the American Thoracic Society.

How were the results disseminated to communities of interest?

Nothing to report.

What do you plan to do during the next reporting period to accomplish the goals?

Having established the best method to isolate and purify nuclei and confirmed the usefulness of the generated data for cell type identification and already observed basic differences in severely affected IPF samples, we will now finalize generating the single nuclei RNA seq dataset samples from

differentially affected regions within 10 IPF lungs (3 tissue cores per lung) and 10 controls lungs. As outlined in the paragraph “5. CHANGES/PROBLEMS”, our automated computational pipeline will allow preprocessing of the data within less than 2 weeks, once the whole raw data is ready. We will then continue as outlined in the SOW and identify aberrant gene expression profiles in cellular subpopulations, establish cell-type-specific regulatory networks in the microenvironment of IPF and cell-type-specific pathways of disease progression and discover cell-type-specific biomarkers of disease progression as well as targets for novel therapeutics.

4. IMPACT

What was the impact on the development of the principal discipline(s) of the project?

Nothing to report.

What was the impact on other disciplines? Nothing to report.

What was the impact on technology transfer? Nothing to report.

What was the impact on society beyond science and technology? Nothing to report.

5. CHANGES/PROBLEMS

Changes in approach and reasons for change:

Nothing to report.

Actual or anticipated problems or delays and actions or plans to resolve them:

Two major problems/issues caused a delay of milestones of this project. First, the secondary ethics review by the DoD approved the use of the tissue samples on 12/26/2019. Following the DoD’s regulations, we therefore were not allowed to perform any research on those samples before that date. Second, due to the outbreak of the Covid-19 pandemic, Yale shut down all non-Covid-19-related research at the beginning of March 2020, which included all core facilities and our lab. Any bench work for this project including single nuclei RNA sequencing was therefore not allowed. Furthermore, also the sequencing core facility was shut down for non-Covid19 related projects. Our lab was only partially reopened on June 12, 2020. Both issues, the longer than expected secondary ethics review by the DoD and the shutdown of all labs at Yale, caused a major delay of goals and milestones, in particular, we

could only generate a small part of the single nuclei RNAseq dataset which is the foundation of all downstream goals. As our lab has partially reopened now, we are working full steam to finalize the dataset. To speed up the completion of the downstream goals, significant parts of the computational pipeline have been automated (see “significant results” under “achievements”) such that we will be able to perform task 2 of Goal 1 and Goal 2 in a fraction of the time stated in the original SOW (presumably in less than 2 weeks). Goal 3 is highly dependent on the complete dataset and the results of Goal 2 and will be carried out as originally outlined.

Changes that had a significant impact on expenditures:

Nothing to report.

Significant changes in use or care of human subjects, vertebrate animals, biohazards, and/or select agents:

Significant changes in use or care of human subjects

Nothing to report.

Significant changes in use or care of vertebrate animals.

Not applicable. No research on vertebrate animals.

Significant changes in use of biohazards and/or select agents

Nothing to report.

6. PRODUCTS

Publications, conference papers, and presentations

Journal publications:

Schupp JC, Khanal S, Gomez JL, Sauler M, Adams TS, Chupp GL, Yan X, Poli S, Montgomery RR, Rosas IO, Dela Cruz CS, Bruscia EM, Egan ME, Kaminski N, Britto CJ. Single Cell

Transcriptional Archetypes of Airway Inflammation in Cystic Fibrosis. Am J Respir Crit Care Med. 2020 Jun 30. DOI: *10.1164/rccm.202004-0991OC*.

Status of publication: published

Acknowledgement of federal support: yes

Schupp JC, Kaminski N. Towards Early Detection of IPF. Am J Respir Crit Care Med. 2019 Aug 14. DOI: 10.1164/rccm.201908-1530ED

Status of publication: published

Acknowledgement of federal support: yes

Books or other non-periodical, one-time publications: Nothing to report

Other publications, conference papers, and presentations:

Schupp JC, Adams T, Ahangari F, Poli De Frias S, Deluliis G, Yan Y, Rosas IO, Homer R, Kaminski N. Single Cell Transcriptomics Reveals Novel COL15A1+ Endothelial Population in Pulmonary Fibrosis and Lung Cancer. Conference abstract. Annual conference of the American Thoracic Society 2020.

Status of publication: published

Acknowledgement of federal support: during submission: yes; funding is however not visible on the congress homepage

Website(s) or other Internet site(s): Nothing to report

Technologies or techniques: Nothing to report

Inventions, patent applications, and/or licenses: Nothing to report

Other Products: Nothing to report

7. PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS

What individuals have worked on the project?

Name:	*Jonas Christian Schupp*
Project Role:	*PI*

Researcher Identifier (e.g. ORCID ID):	*ORCID iD: 0000-0002-7714-8076*
Nearest person month worked:	*4*
Contribution to Project:	*Jonas Schupp developed the data processing pipeline and established analytical algorithms to be used on the final dataset. He performed two preliminary single nuclei RNA experiments and analyzed them and is generating the final dataset at the moment.*

Name:	*Naftali Kaminski*
Project Role:	*Mentor*
Researcher Identifier (e.g. ORCID ID):	*ORCID iD: 0000-0001-5917-4601*
Nearest person month worked:	*<1*
Contribution to Project:	*Supervision of this project and mentoring of Jonas Schupp.*

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?

Nothing to report.

What other organizations were involved as partners?

Nothing to report.

8. SPECIAL REPORTING REQUIREMENTS

Nothing to report.

9. APPENDICES

The appendix includes the two publications (details see above in “6. Products”) supported by this grant:

a) “Single Cell Transcriptional Archetypes of Airway Inflammation in Cystic Fibrosis”

b) “Towards Early Detection of IPF”

Single Cell Transcriptional Archetypes of Airway Inflammation in Cystic Fibrosis

Jonas C. Schupp[1], Sara Khanal[1], Jose L. Gomez[1], Maor Sauler[1], Taylor S. Adams[1], Geoffrey L. Chupp[1], Xiting Yan[1], Sergio Poli[3,4], Yujiao Zhao[5], Ruth R. Montgomery[5], Ivan O. Rosas[3], Charles S. Dela Cruz[1], Emanuela M. Bruscia[2], Marie E. Egan[2], Naftali Kaminski[1], Clemente J. Britto[1]*

Affiliations:

[1]Section of Pulmonary, Critical Care, and Sleep Medicine, Yale University School of Medicine, New Haven, CT, USA.

[2]Division of Pediatric Pulmonology, Allergy, Immunology, and Sleep Medicine, Yale University School of Medicine, New Haven, CT, USA.

[3]Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.

[4]Division of Internal Medicine, Mount Sinai Medical Center, Miami, FL, USA.

[5]Department of Internal Medicine, Yale University School of Medicine, New Haven, CT, USA.

* Corresponding author: Clemente J. Britto, MD; Yale University School of Medicine; Section of Pulmonary, Critical Care, and Sleep Medicine; 333 Cedar Street, PO Box 208057; New Haven, CT 06519; Phone: 203-785-3627; FAX: 203-785-6094; Clemente.britto@yale.edu

Author contributions: CJB and NK conceptualized, acquired funding and supervised the study. CJB, SK, and MEE performed sample collection, phenotyping, and sputum processing. GLC facilitated sputum collection infrastructure and processing protocols. JCS and TSA performed single cell barcoding library construction. Data was processed, curated and visualized by JCS

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

under the supervision of XY, CJB, and NK, and analyzed by JCS, EMB, MS, CSD, and CJB. CyTOF data were reanalyzed by JLG, RRM, and EMB. JCS, TSA, PS, IOR, and NK created and provided scRNAseq data of control distal lungs, TSA calculated the correlation matrix. YZ and RRM performed the sample processing comparison experiments. The manuscript was drafted by JCS and CJB, and was reviewed and edited by all other authors.

Funding: This work was supported by The National Institutes of Health & National Heart, Lung, and Blood Institute (USA) through grants NIH T32-HL007778 and K01-HL125514-01 (CB); the Cystic Fibrosis Foundation through its Fifth Year Clinical Fellowship Award (CB); the American Thoracic Society Foundation's Unrestricted Research Award (CB); NIH U01 HL145567 and R01 HL127349 (NK); and DoD W81XWH-19-1-0131 (JCS).

Short running title: Single cell RNA sequencing of cystic fibrosis sputum

Descriptor number: 9.16 Cystic Fibrosis: Basic Studies

Total word count - manuscript: 3791 words

Total word count - abstract: 242 words

Some of the results of these studies have been previously reported in the form of a preprint (medRxiv, 10 March 2020 https://10.1101/2020.03.06.20032292v1).

This article has an online data supplement, which is accessible from this issue's table of content online at www.atsjournals.org.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

At a Glance

Scientific Knowledge:

Functionally different subsets of neutrophils and mononuclear phagocytes with defective bacterial killing, impaired phagocytic function, and enhanced cytokine production have been described in CF. However, the broad spectrum of transcriptional alterations underlying immune dysfunction in individual CF airway cells has not been characterized.

Add to the Field:

This is the first single-cell RNA sequencing characterization of airway immune cells from CF and healthy control subjects. We observed a shift in the airway immune cell repertoire of CF subjects from alveolar macrophages to a predominance of recruited monocytes and neutrophils. We identified a novel population of recruited lung mononuclear phagocytes in CF, with three distinct transcriptional archetypes: activated monocytes, monocyte-derived macrophages, and heat-shock activated monocytes, and characterized neutrophil subpopulations, highlighting a dominant immature proinflammatory archetype. Our findings offer an opportunity to understand subject-specific immune dysfunction and its potential contribution to CF pathogenesis.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

Abstract:

Rationale: Cystic fibrosis (CF) is a life-shortening multisystem hereditary disease caused by abnormal chloride transport. CF lung disease is driven by innate immune dysfunction and exaggerated inflammatory responses that contribute to tissue injury. In order to define the transcriptional profile of this airway immune dysfunction, we performed the first single-cell transcriptome characterization of CF sputum.

Objectives: To define the transcriptional profile of sputum cells and its implication in the pathogenesis of immune function and the development of CF lung disease.

Methods: We performed single-cell RNA sequencing of sputum cells of nine subjects with CF and five healthy controls. We applied novel computational approaches to define expression-based cell function and maturity profiles, here called transcriptional archetypes.

Measurements and Main Results: The airway immune cell repertoire shifted from alveolar macrophages in healthy controls to a predominance of recruited monocytes and neutrophils in CF. Recruited lung mononuclear phagocytes were abundant in CF, separated into three archetypes: activated monocytes, monocyte-derived macrophages, and heat-shock activated monocytes. Neutrophils were most prevalent in CF, with a dominant immature pro-inflammatory archetype. While CF monocytes exhibited pro-inflammatory features, both monocytes and neutrophils showed transcriptional evidence of abnormal phagocytic and cell-survival programs.

Conclusions: Our findings offer an opportunity to understand subject-specific immune dysfunction and its contribution to divergent clinical courses in CF. As we progress towards personalized applications of therapeutic and genomic developments, we hope this inflammation

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

profiling approach will enable further discoveries that change the natural history of CF lung disease.

Total word count - abstract: 242 words

MeSH key words: Neutrophils, RNA-Seq, Gene Expression Profiling, Macrophages, Monocytes, Cystic Fibrosis

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

Introduction

Cystic Fibrosis (CF) is a life-shortening, multiorgan hereditary disease affecting over 33,000 individuals in the United States (1, 2). Clinical manifestations of CF are caused by mutations in the *CFTR* gene that cause abnormal chloride and bicarbonate transport on epithelial surfaces (3, 4). The disruption of epithelial and innate immune functions is a key contributor to CF lung disease, the primary cause of morbidity and mortality in CF (5, 6). Non-*CFTR* disease-modifying genes also contribute to immune dysfunction, clinical phenotype, and disease progression in CF(7, 8).

Airway inflammation is crucial in the development of CF lung disease, where recruited cells cause tissue damage (9-11). Inflammatory cell populations are heterogeneous, with increasingly recognized CF-specific polymorphonuclear neutrophil (PMN) and macrophage (MΦ) subclasses (10). CF Immune cells from blood and lung biopsies have been profiled using bulk RNA sequencing to characterize transcriptional profiles associated with disease progression and clinical outcomes(12-15). Flow-cytometry studies, including our group's mass cytometry characterization of CF immune cells, also shed light on functional defects of CF immune cell subsets and distinct patterns of immune activation across subpopulations (10, 16-18). These studies have been constrained by the limited number of protein or genetic markers available per assay to define population clusters and assess immune responses. A study providing individualized cellular data on sputum cell types with the granularity afforded by single-cell RNA sequencing (scRNAseq) has not been reported in CF or any other lung disease.

Airway PMN in CF have been characterized in the past (19-24). However, progress in high-throughput single-cell immune profiling has been slow relative to other immune cells like peripheral blood mononuclear cells (PBMC). This may be in part due to the overall limited

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

viability and increased fragility of airway PMN *ex vivo*. CF PMNs generally have a proinflammatory profile, yet some studies reveal functionally different subsets, including populations with abnormal immune function and defective bacterial killing (10). Airway MΦ and other mononuclear phagocytes are also present in CF airway secretions (25-27). Specifically, CF airway monocytes have impaired phagocytic function and enhanced cytokine production (28, 29), playing an important role in driving exaggerated airway inflammation in CF (9, 25).

Single-cell transcriptome profiling is a powerful tool to study innate immune defects and define cell subpopulations that contribute to pathogenesis (30). The use of immune cells from sputum instead of circulating cells or cells differentiated *in vitro* allows us to investigate gene expression profiles that reflect airway transmigration, response to the airway microenvironment, and cell-cell and cell-pathogen interactions key to CF pathogenesis.

Previously identified CF inflammatory cell subpopulations from other studies suggested to us that these cells exist as a continuum of immune maturation and function, rather than isolated, clearly defined, subpopulations. To define this spectrum, we applied scRNAseq followed by pseudotime analysis, and novel approaches to visualize high-dimensional data. In the continuum of sputum inflammatory cells, those with most extreme gene expression features defined functional and maturity trajectories, here called transcriptional archetypes (31). These archetypes constitute a dynamic, more inclusive way to understand transcriptional differences within immune cells. Our approach also allowed us to investigate the relationship between transcription factors and genes involved in immune activation and cell maturation, not previously possible due to an inability to sequence the full cellular transcriptome.

This work is the first to characterize the spectrum of maturation and immune activation states of inflammatory cell populations in CF airways at an unprecedented resolution enabled by

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

scRNAseq. Transcriptional profiling of inflammatory cell archetypes could open the door for highly-targeted therapeutic interventions in subjects with similar CF-causing mutations who experience divergent clinical courses.

Methods

Detailed methods are provided in the online data supplement.

Results

Disease-Specific Cell Distributions of CF Airway Inflammatory and Epithelial Cells

The primary objective of this study was to characterize sputum cell subpopulations in CF using unbiased transcriptome analysis of single cells obtained from CF and healthy control (HC) subjects. Our recruitment period extended from December 2018 through December 2019. Nine subjects with a confirmed CF diagnosis from the Yale Adult CF Program provided sputum samples. We also recruited five HC to undergo sputum induction according to previous protocols (16).

Study subjects were closely age-matched, with a higher inclusion of female subjects in the CF group (67% CF, n=6; 40% HC, n=2). The CF cohort was comprised primarily of *F508del* homozygous subjects (78%, n=7) with only two *F508del* heterozygotes harboring either one deletion or one frameshift mutation in one *CFTR* allele and an *F508del* in the other. The CF cohort's degree of lung function impairment, as determined by Forced Expiratory Volume in the first second (FEV_1), ranged from mild to severe (FEV_1 19-84% of predicted), with a mean FEV_1 of 57%. All CF subjects had pancreatic exocrine insufficiency and 44% (n=4) carried a diagnosis of CF-related diabetes. *Pseudomonas aeruginosa* was isolated in the sputum of 56% of CF subjects (n=5). The majority of CF subjects were receiving CFTR-modulator therapy (89%, n=8)

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

with a combination of either Ivacaftor/Tezacaftor (67%, n=6) or Ivacaftor/Lumacaftor (22%, n=2). For further demographic and clinical details see Table 1.

We developed a standardized scRNAseq workflow for sputum sample analysis (Fig. 1A) and profiled a total of 20,095 sputum cells (12,494 CF, 7,601 HC). We identified nine distinct sputum cell populations based on known transcriptomic markers (Fig. 1C, Supplemental Data file E1): mononuclear phagocytes (recruited lung monocytes, monocyte-derived MΦ (MoMΦ), and alveolar MΦ (alvMΦ)); classical and plasmacytoid dendritic cells (cDC, pDC); PMN; lymphocytes (B, T, and NK cells); and airway epithelial cells from buccal and tracheobronchial mucosa (Fig. 1B-D). The expression of *CFTR* in sputum cells was overall very low and *CFTR* was detected in most cell types in frequencies ranging from 0 to 6.84% (Supplemental Fig. E1).

The Inflammatory Cell Repertoire of CF Sputum Displays a Shift from alvMΦ to Airway Monocytes and PMN

The dominant cell populations in CF and HC samples were strikingly different. PMNs contributed 64% of all CF cells, with minimal numbers of alvMΦ (0.4%). In contrast HC samples were composed of 80.2% alvMΦ with almost no detectable PMN (<2%, both $p < 0.002$). Further, CF subjects also exhibited increased numbers of airway monocytes (19% CF, 1% HC, $p=0.001$) and B cells (4% CF, 1% HC, p = ns), and lower numbers of MoMΦ (1% CF, 6% HC, $p=0.007$) (Fig. 1B-D). Disease-associated PMN, MΦ, and monocyte cellular distributions were confirmed on mass cytometry data from a previously published study by our group, comparing surface markers of inflammatory sputum cells in CF and HC (Supplemental Fig. E2) (16). Furthermore, correlation of cell type gene classifiers in this study and analogous cell types in the largest scRNAseq dataset of the distal lung available to date (n=28) revealed a greater correlation between HC cell types from each dataset than within other cell types from the same dataset,

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

confirming our cell annotations (Supplemental Fig. E3)(31). Our findings indicate that immune cell populations in CF sputum are distinguishable from HC through scRNAseq, and that our cell annotations and shifts in major cell distributions in CF are consistent with other mass cytometry (CyTOF) and scRNAseq studies.

Recruited CF Lung Mononuclear Phagocytes Display Distinct Maturation and Immune Activation Archetypes

AlvMΦ were rare in CF sputum; however, we identified a distinct subpopulation of Recruited Lung mononuclear Phagocytes (RLPs, Fig. 1B) that included recruited lung monocytes and MoMΦ. These RLPs were defined by high expression of mononuclear phagocyte-associated genes (*LYZ, CTSB, CTSH, CTSL, CTSS, CTSZ, HLA-DRA, HLA-DRB1, LGALS1, FTL, CD74*). RLPs were relatively abundant in CF (20% of CF cells) and were rarely identified in HC sputum (7% of HC cells, p=0.06). RLPs were a heterogeneous group, with pronounced and notably different plasticity in CF. This suggested that RLPs would differ not only in abundance, but also in transcriptional profiles between HC and CF.

To characterize the spectrum of immune activation and maturation of monocytes and MoMΦ contained within CF and HC RLPs, we performed a Pseudotime analysis using "Potential of Heat diffusion for Affinity-based Transition Embedding" (PHATE). Pseudotime analysis is a computational technique that allows the distribution of single-cell expression profiles along the continuum of a biologic process marked by gene expression changes (in this case cell maturation, immune activation, and heat-shock response gene expression). Pseudotime analysis demonstrated three distinct gene expression trajectories, and in turn, the most extreme phenotypes of these trajectories defined three RLP transcriptional archetypes in sputum (Fig. 2A)(31, 32). Two of these archetypes were CF-predominant archetypes: activated pro-

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

inflammatory monocytes and heat-shock activated monocytes. The third RLP archetype, mature resting MoMΦ, was more prevalent in HC.

Next, we examined the sequence of gene expression changes leading to the mature resting MoMΦ and activated pro-inflammatory monocyte archetypes, correlating gene expression changes with Pseudotime distance values. The trajectory towards activated pro-inflammatory monocytes was characterized by a gradual and steady increase of pro-inflammatory chemokine and cytokine gene expression. This trajectory was characterized by increasing expression of *IL1B, CXCL2, CCL3, CCL4, CCL20, VEGFA* and *EREG*, Calprotectin (*S100A8, S100A9*)(33), anti-apoptotic proteins *MCL1* and *BCL2L1*, the inflammasome subunit *NLRP3(34)*, inducible cyclooxygenase 2 (*PTGS2*), and transcription factor *NFKB1* (Fig. 2B, Supplemental Fig. E4, E5, Supplemental Data file E2). In the activated monocyte archetype, imputed regulating factors of common activator/repressor genes (i.e. regulons), suggested increased expression of *NFKB1* and pro-inflammatory transcription factors *NFKB2, ETS* and *IRF1*. Pro-inflammatory cytokines *TNF* and *IL1A* were expressed only towards the extreme end of the trajectory, in the most activated monocytes. In contrast to CF RLPs, we did not observe similar immune activation archetypes in MoMΦ, or in alvMΦ from HC. Remarkably, although pro-inflammatory CF monocytes exhibited increased overall cytokine expression, they also showed impaired expression of key phagocytic and cytolytic components of the immune response (complement C1Q), markers of maturation towards a MΦ phenotype (*APOC1, APOE*), and phagocytic function (*MARCO*) compared to other RLP archetypes (Fig. 2B, D).

The mature resting MoMΦ archetype was enriched in HC, and none of the CF MΦ reached the distal end of this archetype (Fig. 2C). Key regulons involved in monocyte to MΦ maturation were active, and increasingly expressed towards the distal end of the archetype

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

trajectory, including canonical *SPI1* (*PU.1*), as well as *MITF* and *USF2*. Maturation of MoMΦ was accompanied by a gradual transcriptional increase of scavenger and pattern-recognition receptors *MSR1* and *MRC1*, surface markers *CD9* and *CD81*, apolipoproteins *APOC1* and *APOE*, and *FABP5*.

MoMΦ were overall rare in sputum, but more evenly distributed between CF and HC subjects, these were distinguished by expression of *PLA2G7,* an enzyme that inactivates platelet-activating factor, monocyte chemokine *CCL2, LGMN* a cysteine-protease involved in MHC-II presentation and differentiation towards DC, and activated-leukocyte cell adhesion molecule *ALCAM.* The majority of sputum cells in HC were alvMΦ. These highly abundant HC alvMΦ expressed the expected levels of phagocytosis-associated genes, underscoring the transcriptional readiness of healthy immune cells to participate in phagocytic functions and coordinate inflammatory cell recruitment, without the basal pro-inflammatory activity noted in the CF-predominant monocytes. Taken together, these findings show that CF RLPs have high pro-inflammatory gene expression but limited phagocytosis-associated transcriptional responses, consistent with excessive inflammation and impaired host defense responses known to occur on CF airways.

An Immature Pro-Inflammatory Archetype Prevails among CF Airway PMN

CF Sputum contained 64% PMN, in contrast with HC where PMN constituted 2% of sputum cells (Fig. 1D). PHATE of the PMN spectrum of gene expression (PMN manifold) enabled us to identify three PMN archetypes based on canonical markers of PMN immaturity (*CXCR4, IGF2R*) and maturity (*FCGR3B, ALPL, CXCR2*), as well as a heat-shock response archetype (Fig. 3A, 3B, Supplemental Fig. E6). To analyze gradual changes within the PMN manifold, we applied trajectory inference and correlated the resulting pseudotime distances with

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

gene expression and regulon activity. When tracing PMN maturation, we observed that expression of calprotectin (*S100A8, S100A9*), *S100A11, CSF3R* and antiapoptotic factor *BL2A1* are gained relatively early, in contrast to classical maturation markers *FCGR3B, ALPL, CXCR2* and *CD14* which ramp up in expression relatively late (Fig. 3B, Supplemental Data file E2)(35, 36). In immature PMN, we observed a gradual increase of transcription factors *TFEC, MITF, STAT3*, and maturation-associated transcription factors *CEBPB* and *NFIL3*. The CF-predominant immature PMN archetype was further defined by increased expression of PMN-activating chemokine MIP (*CCL3, CCL4*) and downstream transcription factor and adapter molecules *IRAK3* and *TRAF3*. These findings suggest that CF airway PMNs have an overall pro-inflammatory phenotype, with a large subpopulation of PMNs exhibiting a functional and maturity transcriptional shift, consistent with an immature PMN gene expression profile.

CF PMN Archetypes Have Decreased Phagocytic Marker and Tyrosine Kinase Expression

We compared the gene expression profiles of CF and HC PMN to understand transcriptomic differences associated with their immune function (Supplemental Data file E3). We categorized the top gene expression differences between CF and HC accordingly into: 1) Cell adhesion and maturation markers, 2) MHC class I molecules; 3) Pattern and IgG recognition, 4) Transcription factors and adaptor molecules; 5) Tyrosine Kinase expression; and 6) Survival and apoptosis genes (Fig. 3C). In CF PMN, cell adhesion and maturation markers were overall lower than in HC (*CSF2RB, CSF3R, CXCR2, ICAM3, PECAM1*), except for *ITGAX*. The decreased expression of these markers in CF reflects a higher prevalence of the immature PMN archetype described above. In addition to decreased CXCR- and CSF-receptor expression, CF PMN also expressed lower *CXCR1, IL1RN*, and *IL1B* that could condition further defects in phagocytosis and inflammatory cell recruitment. We identified striking differences in antigen presentation,

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

pathogen recognition, and phagocytosis-associated genes between CF and HC PMN. CF PMN showed decreased expression of numerous members of the MHC-I molecules (*HLA-A/B/C/E*), immunoglobin receptors (*FCGR3B, FCGR2A, FCGRT*), decreased pathogen recognition receptors *CD14, TLR2*, and *NLRP1*, and decreased expression of lysozyme (*LYZ*). Interestingly, two genes involved in the assembly of lipid rafts and primary neutrophil granule release were increased (*SYK*, *CD63*) suggesting that although PMN may suffer from defective phagocytic activity, the transcriptional infrastructure needed to express tissue proteases and inflammatory mediators into the airways is preserved. CF PMN demonstrated increased transcriptomic activation characterized by expression of transcription factors and pro-inflammatory adapter molecules (increased *PI3, IRAK2/3, TRAF3, TANK*), yet this activation did not translate into increased expression of inflammatory cytokines. Interestingly, the downstream response to cytokine activation appeared to be blunted, as shown by decreased overall tyrosine kinase gene expression (*ITPK1, MAP3K5, MAP2K4, CAMK1D, PIK3CD, HIPK3*). Finally, we observed the induction of genes involved in the hypoxic response (*HIF1A, VEGFA, FGF13, PTGS2*) and diverging proapoptotic signals with lower expression of *CASP4, RPS6KA5, CREB5, BCL2A*, and increased expression of *HES4, KRAS,* and *CREM* in CF. These observations underscore the presence of a hypoxic airway environment in CF and a dysfunctional cell death program that enhances the survival of functionally ineffective PMN. Taken together, these findings indicate that CF PMN do not carry out an effective transcriptional response to inflammatory stimuli and lack essential components for pathogen recognition and removal.

Discussion

This is the first single-cell transcriptome characterization of immune cells in CF sputum. We identified CF-specific differences in cell subpopulations including alvMΦ, RLPs, and PMN.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

Furthermore, these cells had markedly different transcriptional profiles when compared to their HC counterparts. Previous CF studies have used transcriptomic analysis to determine the likelihood of adverse outcomes in CF lung disease, however they have not focused on establishing differences between healthy and CF airway inflammatory cells, or characterizing their immune activation profiles (12-15). The most remarkable finding from this study is the discovery of novel archetypes of RLPs, enabled by an unprecedented depth of gene expression profiling. These inflammatory cell subpopulations exhibit a wide spectrum of maturity and immune activation in CF. Airway MΦ and other mononuclear cells have been described in human CF airway secretions (25, 26) and their role in driving exaggerated airway inflammation in CF has been well characterized in animal models (9, 25). However, a broader genomics approach to define sputum RLPs, their potential functional impairments, and pathogenic role has not been reported.

We identified three novel archetypes of CF RLP including activated monocytes, mature MoMΦ, and heat-hock activated monocytes. Airway monocytes in CF have impaired ion transport and phagocytic function, however their role in CF lung disease remains undefined (28, 37). Others have described dramatic changes in monocyte cell adhesion and chemotaxis that perpetuate inflammation in CF lungs, along with enhanced chemokine production that sustains PMN recruitment and injury (38). In agreement with these studies, we observe that monocytes are rather abundant in CF sputum, but are deficient in monocyte maturation gene expression markers (*MITF, SPI1*). Furthermore, CF monocytes were not only abundant, but also highly active from the immune perspective, expressing high levels of inflammation-related genes (*CXCL8, IL1B, CCL3*, and Calprotectin). These observations underscore a defect in CF

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

monocyte maturation that preserves a highly pro-inflammatory phenotype and contributes to airway damage and aberrant inflammatory cell recruitment (39).

MΦs recovered from CF lungs are relatively smaller in size and express minimal levels of mannose receptor MRC1 or MARCO typically detected on alvMΦ (Supplemental Fig E7) (26, 27). This has been interpreted as an indication that CF airway MΦs are recruited from the circulation, as opposed to tissue-resident alvMΦ which are of embryonic origin. Here, we show that most CF airway MΦ originate from recruited monocytes, while the majority of healthy control airway cells were *bona fide* tissue-resident alvMΦ.

In contrast to CF airway monocytes, more mature CF phagocytes (MoMΦ, alvMΦ) showed low levels of immune activation markers observed in CF monocytes, and of key phagocytic and cytolytic components of the immune response (complement *C1Q*s, *MARCO*). This underscores that in CF, RLPs that reach maturity exhibit transcriptomic evidence of impaired or limited phagocytic function, accounting for the known impaired phagocytic abilities of these cells in CF.

We did not detect a distinct acute exacerbation signature in CF samples. This may reflect our stringent gene expression analysis strategy, a lack of paired sputum samples, and sample size limitations to perform this subgroup analysis. This is an important question to pursue in the future, as paired samples in stable and exacerbation states from the same individual may reveal critical genetic modifiers of a patient's clinical course.

PMN were the most abundant immune cells in the sputum of patients with CF, which is consistent with reports in the CF literature, similar to the predominance of alvMΦ in HC sputum (19-24). Here, we report the discovery of new archetypes of CF PMN based on inflammatory and maturity gene expression markers; one, characterized by high maturity and limited pro-

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

inflammatory transcriptional state, and another with higher pro-inflammatory activity and delayed expression of maturity markers. Overall, the increased expression of pro-inflammatory genes in immature PMN highlights a highly activated and pro-inflammatory state, clearly distinguishable from the transcriptional profile of HC PMN. The immature airway PMN archetype shares features of a previously described subpopulation of transmigrated PMN with increased granule release, immunoregulatory and metabolic activity, and defective bacterial killing in *in vitro* studies, referred to as "GRIM" neutrophils (10, 40). We identified cells with similar characteristics, but as part of a spectrum of granulocyte maturation that encompasses vigorously activated PMN on one extreme and PMN with decreased expression of maturity markers & evidence of recent airway migration on the other extreme. Adding to the complexity of these PMN subpopulations, counterproductive pro- and anti-apoptotic signals were present across the CF PMN when compared to HC (increased *UVRAG, PLPP3, ATG7*, decreased *CASP4, RPS6KA5, CREB5, BCL2A*). Taken together, these findings underscore an aberrant pro-inflammatory state in CF PMN, exacerbated by disruption of immunomodulatory and anti-inflammatory mechanisms like apoptosis and transcription factor suppression.

The presence of B cells in CF sputum was an intriguing finding. Single nucleotide polymorphisms (SNPs) in class II major histocompatibility complex (MHCII) of the *F508del* population are associated with delayed *Pseudomonas aeruginosa (*PA) colonization and slower lung function decline (41-44). Although we observed no differences in MHCII gene expression in B cells of CF subjects (Supplemental Fig. E8), a focused study on MHCII SNPs could identify B cell subpopulations with a protective role against PA and its associated impact on pulmonary health.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

This work includes two technical advances. First, this is the only reported scRNAseq study of CF sputum, a notoriously complex biological sample with high variability in cell viability and in cellularity between subjects. Second, our sputum processing protocol avoids the use of reducing agents to solubilize sputum and instead minimizes immune cell activation and injury by using mechanical disruption and filtering. Importantly, ours is the first report of a sputum cryopreservation protocol allowing the retrieval of live cells for scRNAseq analysis while avoiding sputum solubilizing agents typically used in sputum sample processing (Supplementary materials, Supplementary Fig. E9)(45-49). The ability to use cryopreserved cells overcomes a major limitation of previous single-cell studies that required fresh samples (13, 16), this is particularly important for the recovery of PMN, known for their short life-span ex-vivo and susceptibility to immune activation. Our study has several limitations: 1) Large differences in predominant cell types between CF and HC subjects make it difficult to generalize gene expression changes between disease and control groups. Although we present these comparisons, our focus is on understanding CF-specific cell distributions and their spectrum of maturity and activation markers; 2) Since HC express minimal sputum if any at all, we used a standardized approach for sputum induction in these subjects, while CF cells were obtained from spontaneously expectorated sputum. As single cell suspensions are standardized for number of cells before any analysis, these sampling differences likely have a minor impact on our observations; 3) There was an uneven sex distribution across the study groups. This may be of particular importance in CF, as female sex in CF is associated with disparities in life expectancy, frequency of exacerbation, and early acquisition of respiratory pathogens(50). However, of the differentially expressed genes between CF and controls, we did not observe divergent differential gene expression changes in females or males; and finally, 4) Our study has a small sample size;

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

however, we sought to match subjects according to age and sex, and HC were compared to a relatively homogeneous CF cohort in terms of *CFTR* mutation background, CF comorbidities, and ongoing therapy. Although a small number of patients were recruited for this study, we believe they are representative of patients with CF based on the *F508del* allele frequency in our cohort and the identification of nine distinct cell types representative of airway cells in CF. Despite these limitations, our findings are robust and representative of the CF airway compartment.

CF research is progressing rapidly towards clinical, molecular, and functional characterization based on individualized high-throughput diagnosis and functional profiling. Our application of scRNAseq enabled the discovery of transcriptional archetypes in CF-specific cell subpopulations that may underlie subject-specific differences in disease progression and response to therapy. As we advance towards early applications of therapeutic and genomic technologies, we hope this approach to individualized airway inflammation profiling will serve as a foundation for further discoveries that transform the natural history of CF lung disease.

Acknowledgments

We thank our patients, the medical staff at the Yale Adult Cystic Fibrosis Program, Dr. Farida Ahangari, and Dr. Jonathan Koff, Director of the Yale Adult CF Program, for their support and contributions to this project. Sequencing was conducted by Mei Zhong at Yale Stem Cell Center Genomics Core facility which was supported by the Connecticut Regenerative Medicine Research Fund and the Li Ka Shing Foundation.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

References

1. Foundation CF. Cystic Fibrosis Foundation Patient Registry 2017 Annual Data Report; 2018.

2. Cutting GR. Cystic fibrosis genetics: from molecular understanding to clinical application. *Nat Rev Genet* 2015; 16: 45-56.

3. Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou JL, et al. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* 1989; 245: 1066-1073.

4. Welsh MJ, Ramsey BW, Accurso F, Cutting GR. Cystic fibrosis. In: Scriver CR BA, Sly WS,Valle D,Childs B,Vogelstein B, editor. The metabolic and molecular basis of inherited disease. New York: McGraw-Hill; 2001. p. pp. 5121–5189.

5. McCague AF, Raraigh KS, Pellicore MJ, Davis-Marcisak EF, Evans TA, Han ST, Lu Z, Joynt AT, Sharma N, Castellani C, Collaco JM, Corey M, Lewis MH, Penland CM, Rommens JM, Stephenson AL, Sosnay PR, Cutting GR. Correlating Cystic Fibrosis Transmembrane Conductance Regulator Function with Clinical Features to Inform Precision Treatment of Cystic Fibrosis. *American journal of respiratory and critical care medicine* 2019; 199: 1116-1126.

6. Stoltz DA, Meyerholz DK, Welsh MJ. Origins of cystic fibrosis lung disease. *The New England journal of medicine* 2015; 372: 351-362.

7. Corvol H, Blackman SM, Boelle PY, Gallins PJ, Pace RG, Stonebraker JR, Accurso FJ, Clement A, Collaco JM, Dang H, Dang AT, Franca A, Gong J, Guillot L, Keenan K, Li W, Lin F, Patrone MV, Raraigh KS, Sun L, Zhou YH, O'Neal WK, Sontag MK, Levy H, Durie PR, Rommens JM, Drumm ML, Wright FA, Strug LJ, Cutting GR, Knowles MR.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis. *Nature communications* 2015; 6: 8382.

8. Polineni D, Dang H, Gallins PJ, Jones LC, Pace RG, Stonebraker JR, Commander LA, Krenicky JE, Zhou YH, Corvol H, Cutting GR, Drumm ML, Strug LJ, Boyle MP, Durie PR, Chmiel JF, Zou F, Wright FA, O'Neal WK, Knowles MR. Airway Mucosal Host Defense Is Key to Genomic Regulation of Cystic Fibrosis Lung Disease Severity. *American journal of respiratory and critical care medicine* 2018; 197: 79-93.

9. Bruscia EM, Zhang PX, Ferreira E, Caputo C, Emerson JW, Tuck D, Krause DS, Egan ME. Macrophages directly contribute to the exaggerated inflammatory response in cystic fibrosis transmembrane conductance regulator-/- mice. *American journal of respiratory cell and molecular biology* 2009; 40: 295-304.

10. Forrest OA, Ingersoll SA, Preininger MK, Laval J, Limoli DH, Brown MR, Lee FE, Bedi B, Sadikot RT, Goldberg JB, Tangpricha V, Gaggar A, Tirouvanziam R. Frontline Science: Pathological conditioning of human neutrophils recruited to the airway milieu in cystic fibrosis. *Journal of leukocyte biology* 2018; 104: 665-675.

11. Margaroli C, Garratt LW, Horati H, Dittrich AS, Rosenow T, Montgomery ST, Frey DL, Brown MR, Schultz C, Guglani L, Kicic A, Peng L, Scholte BJ, Mall MA, Janssens HM, Stick SM, Tirouvanziam R, Arest CF, Impede CF. Elastase Exocytosis by Airway Neutrophils Associates with Early Lung Damage in Cystic Fibrosis Children. *American journal of respiratory and critical care medicine* 2018.

12. Chesne J, Danger R, Botturi K, Reynaud-Gaubert M, Mussot S, Stern M, Danner-Boucher I, Mornex JF, Pison C, Dromer C, Kessler R, Dahan M, Brugiere O, Le Pavec J, Perros F,

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

Humbert M, Gomez C, Brouard S, Magnan A, Consortium C. Systematic analysis of blood cell transcriptome in end-stage chronic respiratory diseases. *PloS one* 2014; 9: e109291.

13. Jiang K, Poppenberg KE, Wong L, Chen Y, Borowitz D, Goetz D, Sheehan D, Frederick C, Tutino VM, Meng H, Jarvis JN. RNA sequencing data from neutrophils of patients with cystic fibrosis reveals potential for developing biomarkers for pulmonary exacerbations. *Journal of cystic fibrosis : official journal of the European Cystic Fibrosis Society* 2019; 18: 194-202.

14. Kormann MSD, Dewerth A, Eichner F, Baskaran P, Hector A, Regamey N, Hartl D, Handgretinger R, Antony JS. Transcriptomic profile of cystic fibrosis patients identifies type I interferon response and ribosomal stalk proteins as potential modifiers of disease severity. *PloS one* 2017; 12: e0183526.

15. Levy H, Jia S, Pan A, Zhang X, Kaldunski M, Nugent ML, Reske M, Feliciano RA, Quintero D, Renda MM, Woods KJ, Murkowski K, Johnson K, Verbsky J, Dasu T, Ideozu JE, McColley S, Quasney MW, Dahmer MK, Avner E, Farrell PM, Cannon CL, Jacob H, Simpson PM, Hessner MJ. Identification of molecular signatures of cystic fibrosis disease status with plasma-based functional genomics. *Physiological genomics* 2019; 51: 27-41.

16. 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.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

17. Zhang S, Shrestha CL, Kopp BT. Cystic fibrosis transmembrane conductance regulator (CFTR) modulators have differential effects on cystic fibrosis macrophage function. *Sci Rep* 2018; 8: 17066.

18. Sorio C, Montresor A, Bolomini-Vittori M, Caldrer S, Rossi B, Dusi S, Angiari S, Johansson JE, Vezzalini M, Leal T, Calcaterra E, Assael BM, Melotti P, Laudanna C. Mutations of Cystic Fibrosis Transmembrane Conductance Regulator Gene Cause a Monocyte-Selective Adhesion Deficiency. *American journal of respiratory and critical care medicine* 2016; 193: 1123-1133.

19. Ramsey BW, Downey GP, Goss CH. Update in Cystic Fibrosis 2018. *American journal of respiratory and critical care medicine* 2019; 199: 1188-1194.

20. Alexis NE, Muhlebach MS, Peden DB, Noah TL. Attenuation of host defense function of lung phagocytes in young cystic fibrosis patients. *Journal of cystic fibrosis : official journal of the European Cystic Fibrosis Society* 2006; 5: 17-25.

21. Konstan MW, Hilliard KA, Norvell TM, Berger M. Bronchoalveolar lavage findings in cystic fibrosis patients with stable, clinically mild lung disease suggest ongoing infection and inflammation. *American journal of respiratory and critical care medicine* 1994; 150: 448-454.

22. Pohl K, Hayes E, Keenan J, Henry M, Meleady P, Molloy K, Jundi B, Bergin DA, McCarthy C, McElvaney OJ, White MM, Clynes M, Reeves EP, McElvaney NG. A neutrophil intrinsic impairment affecting Rab27a and degranulation in cystic fibrosis is corrected by CFTR potentiator therapy. *Blood* 2014; 124: 999-1009.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

23. Rosenfeld M, Gibson RL, McNamara S, Emerson J, Burns JL, Castile R, Hiatt P, McCoy K, Wilson CB, Inglis A, Smith A, Martin TR, Ramsey BW. Early pulmonary infection, inflammation, and clinical outcomes in infants with cystic fibrosis. *Pediatr Pulmonol* 2001; 32: 356-366.

24. Witko-Sarsat V, Allen RC, Paulais M, Nguyen AT, Bessou G, Lenoir G, Descamps-Latscha B. Disturbed myeloperoxidase-dependent activity of neutrophils in cystic fibrosis homozygotes and heterozygotes, and its correction by amiloride. *J Immunol* 1996; 157: 2728-2735.

25. Bruscia EM, Bonfield TL. Cystic Fibrosis Lung Immunity: The Role of the Macrophage. *J Innate Immun* 2016; 8: 550-563.

26. Garratt LW, Wright AK, Ranganathan SC, Grigg J, Sly PD, behalf of AC. Small macrophages are present in early childhood respiratory disease. *Journal of cystic fibrosis : official journal of the European Cystic Fibrosis Society* 2012; 11: 201-208.

27. Wright AK, Rao S, Range S, Eder C, Hofer TP, Frankenberger M, Kobzik L, Brightling C, Grigg J, Ziegler-Heitbrock L. Pivotal Advance: Expansion of small sputum macrophages in CF: failure to express MARCO and mannose receptors. *Journal of leukocyte biology* 2009; 86: 479-489.

28. Riquelme SA, Lozano C, Moustafa AM, Liimatta K, Tomlinson KL, Britto C, Khanal S, Gill SK, Narechania A, Azcona-Gutierrez JM, DiMango E, Saenz Y, Planet P, Prince A. CFTR-PTEN-dependent mitochondrial metabolic dysfunction promotes Pseudomonas aeruginosa airway infection. *Sci Transl Med* 2019; 11.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

29. Riquelme SA, Liimatta K, Wong Fok Lung T, Fields B, Ahn D, Chen D, Lozano C, Saenz Y, Uhlemann AC, Kahl BC, Britto CJ, DiMango E, Prince A. Pseudomonas aeruginosa Utilizes Host-Derived Itaconate to Redirect Its Metabolism to Promote Biofilm Formation. *Cell Metab* 2020.

30. Reyfman PA, Walter JM, Joshi N, Anekalla KR, McQuattie-Pimentel AC, Chiu S, Fernandez R, Akbarpour M, Chen CI, Ren Z, Verma R, Abdala-Valencia H, Nam K, Chi M, Han S, Gonzalez-Gonzalez FJ, Soberanes S, Watanabe S, Williams KJN, Flozak AS, Nicholson TT, Morgan VK, Winter DR, Hinchcliff M, Hrusch CL, Guzy RD, Bonham CA, Sperling AI, Bag R, Hamanaka RB, Mutlu GM, Yeldandi AV, Marshall SA, Shilatifard A, Amaral LAN, Perlman H, Sznajder JI, Argento AC, Gillespie CT, Dematte J, Jain M, Singer BD, Ridge KM, Lam AP, Bharat A, Bhorade SM, Gottardi CJ, Budinger GRS, Misharin AV. Single-Cell Transcriptomic Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis. *American journal of respiratory and critical care medicine* 2019; 199: 1517-1536.

31. Mohammadi S, Ravindra V, Gleich DF, Grama A. A geometric approach to characterize the functional identity of single cells. *Nature communications* 2018; 9: 1-10.

32. Mohammadi S, Davila-Velderrain J, Kellis M. Multi-resolution single-cell state characterization via joint archetypal/network analysis. *bioRxiv* 2019: 746339.

33. Reid PA, McAllister DA, Boyd AC, Innes JA, Porteous D, Greening AP, Gray RD. Measurement of serum calprotectin in stable patients predicts exacerbation and lung function decline in cystic fibrosis. *American journal of respiratory and critical care medicine* 2015; 191: 233-236.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

34. McElvaney OJ, Zaslona Z, Becker-Flegler K, Palsson-McDermott EM, Boland F, Gunaratnam C, Gulbins E, O'Neill LA, Reeves EP, McElvaney NG. Specific Inhibition of the NLRP3 Inflammasome as an Antiinflammatory Strategy in Cystic Fibrosis. *American journal of respiratory and critical care medicine* 2019; 200: 1381-1391.

35. Evrard M, Kwok IWH, Chong SZ, Teng KWW, Becht E, Chen J, Sieow JL, Penny HL, Ching GC, Devi S, Adrover JM, Li JLY, Liong KH, Tan L, Poon Z, Foo S, Chua JW, Su IH, Balabanian K, Bachelerie F, Biswas SK, Larbi A, Hwang WYK, Madan V, Koeffler HP, Wong SC, Newell EW, Hidalgo A, Ginhoux F, Ng LG. Developmental Analysis of Bone Marrow Neutrophils Reveals Populations Specialized in Expansion, Trafficking, and Effector Functions. *Immunity* 2018; 48: 364-379 e368.

36. Grassi L, Pourfarzad F, Ullrich S, Merkel A, Were F, Carrillo-de-Santa-Pau E, Yi G, Hiemstra IH, Tool ATJ, Mul E, Perner J, Janssen-Megens E, Berentsen K, Kerstens H, Habibi E, Gut M, Yaspo ML, Linser M, Lowy E, Datta A, Clarke L, Flicek P, Vingron M, Roos D, van den Berg TK, Heath S, Rico D, Frontini M, Kostadima M, Gut I, Valencia A, Ouwehand WH, Stunnenberg HG, Martens JHA, Kuijpers TW. Dynamics of Transcription Regulation in Human Bone Marrow Myeloid Differentiation to Mature Blood Neutrophils. *Cell Rep* 2018; 24: 2784-2794.

37. Van de Weert-van Leeuwen PB, Van Meegen MA, Speirs JJ, Pals DJ, Rooijakkers SH, Van der Ent CK, Terheggen-Lagro SW, Arets HG, Beekman JM. Optimal complement-mediated phagocytosis of Pseudomonas aeruginosa by monocytes is cystic fibrosis transmembrane conductance regulator-dependent. *American journal of respiratory cell and molecular biology* 2013; 49: 463-470.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

38. Kreisel D, Nava RG, Li W, Zinselmeyer BH, Wang B, Lai J, Pless R, Gelman AE, Krupnick AS, Miller MJ. In vivo two-photon imaging reveals monocyte-dependent neutrophil extravasation during pulmonary inflammation. *Proceedings of the National Academy of Sciences of the United States of America* 2010; 107: 18073-18078.

39. Bruscia EM, Zhang PX, Satoh A, Caputo C, Medzhitov R, Shenoy A, Egan ME, Krause DS. Abnormal trafficking and degradation of TLR4 underlie the elevated inflammatory response in cystic fibrosis. *J Immunol* 2011; 186: 6990-6998.

40. Ingersoll SA, Laval J, Forrest OA, Preininger M, Brown MR, Arafat D, Gibson G, Tangpricha V, Tirouvanziam R. Mature cystic fibrosis airway neutrophils suppress T cell function: evidence for a role of arginase 1 but not programmed death-ligand 1. *J Immunol* 2015; 194: 5520-5528.

41. Aron Y, Polla BS, Bienvenu T, Dall'ava J, Dusser D, Hubert D. HLA class II polymorphism in cystic fibrosis. A possible modifier of pulmonary phenotype. *American journal of respiratory and critical care medicine* 1999; 159: 1464-1468.

42. Laki J, Laki I, Nemeth K, Ujhelyi R, Bede O, Endreffy E, Bolbas K, Gyurkovits K, Csiszer E, Solyom E, Dobra G, Halasz A, Pozsonyi E, Rajczy K, Prohaszka Z, Fekete G, Fust G. The 8.1 ancestral MHC haplotype is associated with delayed onset of colonization in cystic fibrosis. *Int Immunol* 2006; 18: 1585-1590.

43. O'Neal WK, Gallins P, Pace RG, Dang H, Wolf WE, Jones LC, Guo X, Zhou YH, Madar V, Huang J, Liang L, Moffatt MF, Cutting GR, Drumm ML, Rommens JM, Strug LJ, Sun W, Stonebraker JR, Wright FA, Knowles MR. Gene expression in transformed

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

lymphocytes reveals variation in endomembrane and HLA pathways modifying cystic fibrosis pulmonary phenotypes. *Am J Hum Genet* 2015; 96: 318-328.

44. Lu S, Song K, Bomberger J, Kolls JK. Functional studies to understand immune modifiers in cystic fibrosis. Am Assoc Immnol; 2019.

45. Hector A, Jonas F, Kappler M, Feilcke M, Hartl D, Griese M. Novel method to process cystic fibrosis sputum for determination of oxidative state. *Respiration* 2010; 80: 393-400.

46. Sagel SD, Kapsner R, Osberg I, Sontag MK, Accurso FJ. Airway Inflammation in Children with Cystic Fibrosis and Healthy Children Assessed by Sputum Induction. *American journal of respiratory and critical care medicine* 2001; 164: 1425-1431.

47. Hisert KB, Liles WC, Manicone AM. A Flow Cytometric Method for Isolating Cystic Fibrosis Airway Macrophages from Expectorated Sputum. *American journal of respiratory cell and molecular biology* 2019; 61: 42-50.

48. Mayer-Hamblett N, Aitken ML, Accurso FJ, Kronmal RA, Konstan MW, Burns JL, Sagel SD, Ramsey BW. Association between pulmonary function and sputum biomarkers in cystic fibrosis. *American journal of respiratory and critical care medicine* 2007; 175: 822-828.

49. Ordonez CL, Stulbarg M, Grundland H, Liu JT, Boushey HA. Effect of clarithromycin on airway obstruction and inflammatory markers in induced sputum in cystic fibrosis: a pilot study. *Pediatr Pulmonol* 2001; 32: 29-37.

50. Han MK, Arteaga-Solis E, Blenis J, Bourjeily G, Clegg DJ, DeMeo D, Duffy J, Gaston B, Heller NM, Hemnes A, Henske EP, Jain R, Lahm T, Lancaster LH, Lee J, Legato MJ, McKee S, Mehra R, Morris A, Prakash YS, Stampfli MR, Gopal-Srivastava R, Laposky

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

AD, Punturieri A, Reineck L, Tigno X, Clayton J. Female Sex and Gender in Lung/Sleep Health and Disease. Increased Understanding of Basic Biological, Pathophysiological, and Behavioral Mechanisms Leading to Better Health for Female Patients with Lung Disease. *American journal of respiratory and critical care medicine* 2018; 198: 850-858.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

Tables

Table 1.

Number of Patients (n)	*HC (5)*	*CF (9)*
Age		
Age (Mean)	35.4 ± 5.9	30.6 ± 6.5
Age (Range)	26-42	24-43
Sex		
Female (n)	2 (40%)	6 (67%)
Male (n)	3 (60%)	3 (33%)
Mutation Background		
F508del/F508del (n)	*NA*	7 (77.8%)
F508del/other (n)	*NA*	2 (22.2%)
No *F508del* mutations (n)	*NA*	0 (0%)
FEV_1 (L)		
FEV_1 (Mean)	*NA*	1.9 ± 0.7
FEV_1 (Range)	*NA*	0.68 - 2.85
FEV_1 (%)		
FEV_1 (Mean)	*NA*	57 ± 21.5
FEV_1 (Range)	*NA*	19 - 84
BMI (Kg/m^2)		
BMI (Mean)	*NA*	22.2 ± 2.1
BMI (Range)	*NA*	19.11 - 25.73
CF Comorbidities		
Pancreatic Exocrine Insufficiency (n)	*NA*	9 (100%)
CF-related Diabetes (n)	*NA*	4 (44.4%)
Liver disease (n)	*NA*	1 (11.1%)
Microbiology		
Pseudomonas aeruginosa Colonization (n)	*NA*	5 (55.6%)
CFTR Modulators		
Ivacaftor/Tezacaftor (n)	*NA*	6 (66.7%)
Ivacaftor/Lumacaftor (n)	*NA*	2 (22.2%)
No modulator (n)	*NA*	1 (11.1%)

Table 1. Demographic characteristics of study subjects from the Yale Adult Cystic Fibrosis Program and healthy controls. HC: Healthy controls; CF: CF subjects; FEV_1 Forced expiratory

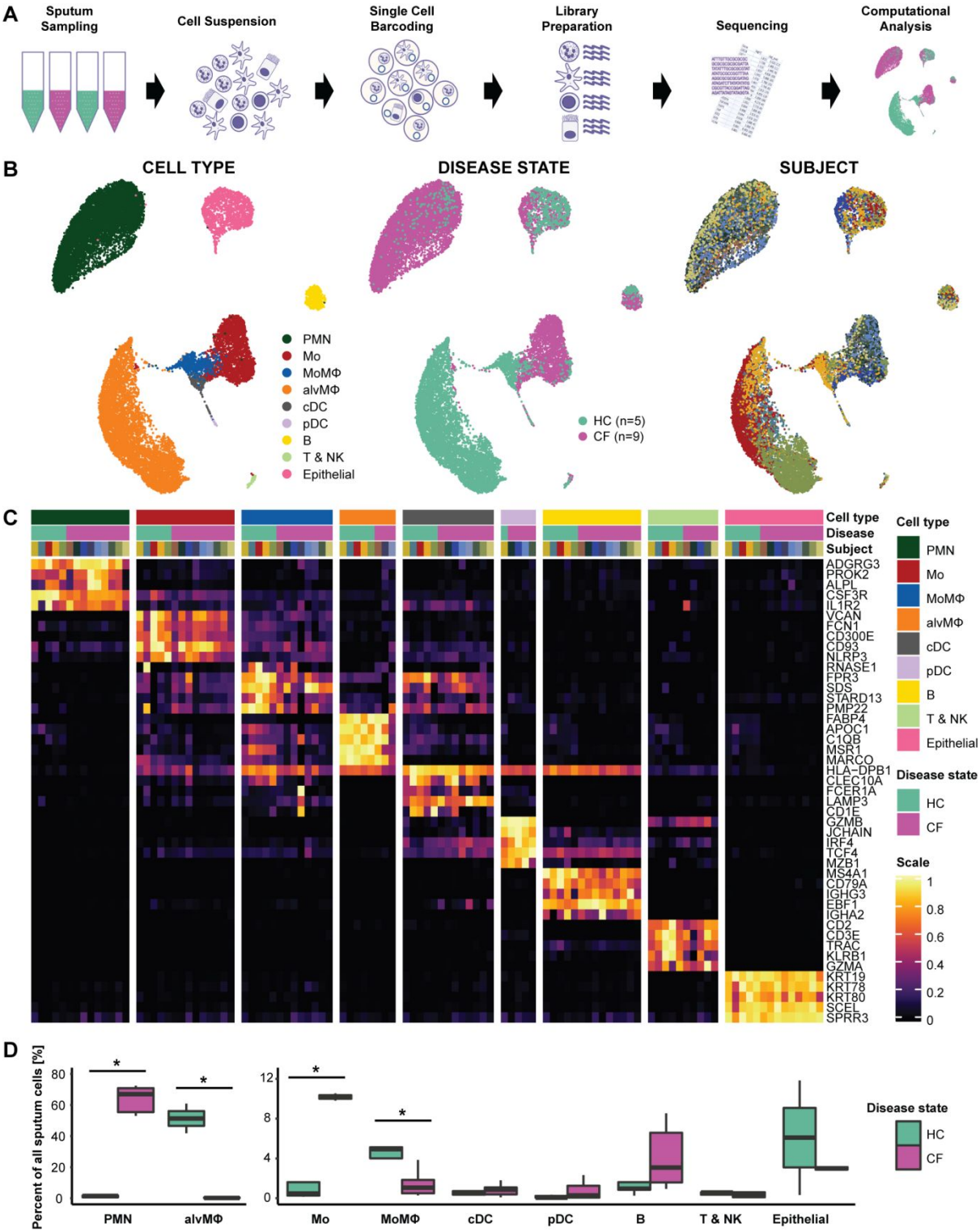
AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

volume in the first second; BMI: Body Mass Index; CFTR: Cystic Fibrosis Transmembrane conductance Regulator.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

Figures

Fig. 1.



AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

Fig.1. ScRNAseq Reveals an Immune Cell Repertoire Shift from Alveolar MΦ to Recruited Monocytes and PMN in CF. **(A)** Schematic of the experimental design. (i) Spontaneously expectorated sputum from patients with cystic fibrosis (CF) and induced sputum from healthy controls (HC) was collected. (ii) Sputum was processed into a single-cell suspension. (iii) Droplet-based scRNAseq barcoding (iii) library preparation (iv) sequencing (v) and computational analysis. **(B)** Uniform Manifold Approximation and Projection (UMAP) visualization of 20,095 sputum cells from nine patients with CF and five controls. Each dot represents a single cell, and cells are labelled by (i) cell type, (ii) disease status, and (iii) subject. **(C)** Heatmap of marker genes for all cell types identified. Each column represents the average expression value of one subject, grouped by disease status and cell type. Gene expression values are unity-normalized from 0 to 1. **(D)** Boxplots showing percentages of all identified cell types 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.

Mo: monocyte; MoMΦ: monocyte-derived macrophage; alvMΦ: alveolar macrophage; cDC: classical dendritic cell, pDC: plasmacytoid dendritic cell; B: B-lymphocyte; T & NK: T-lymphocytes and NK-cells; PMN: polymorphonuclear neutrophil.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

Fig. 2.

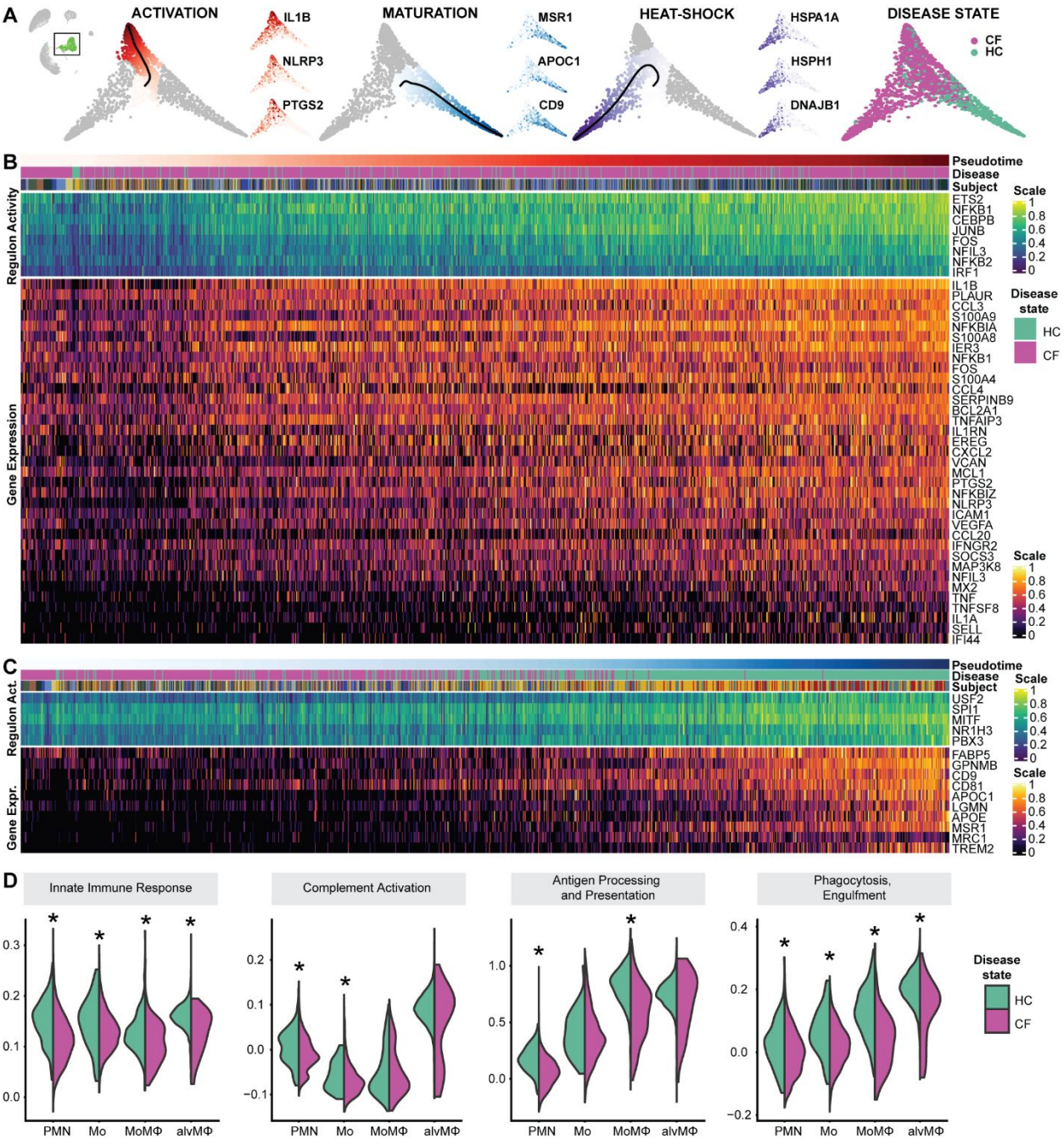


Fig.2. Recruited Lung Mononuclear Phagocytes are a Distinct Cell Population with a Broad Spectrum of Maturity and Immune Activation in CF Airways. **(A)** Potential of Heat diffusion for Affinity-based Transition Embedding (PHATE) of monocytes and monocyte-derived

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

macrophages, colored by pseudotime, all starting from quiescent monocytes towards (i) activated monocytes, (ii) mature monocyte-derived macrophages, (iii) monocytes expressing a heat-shock response. (iv) monocytes and monocyte-derived macrophages, colored by disease state. All three archetypes are accompanied by three PHATE plots colored by the gene expression of typical genes ramping up along a specific pseudotime. For corresponding UMAP embedding colored by gene expressions of the same genes, see Supplemental Fig. E4. For corresponding PHATE embedding colored by cell type and subjects, see Supplemental Fig. E5. **(B)** Heatmap of gene expression and regulon activity in monocytes undergoing activation, ordered by pseudotime distances along PHATE manifolds that transition from quiescent monocytes towards an activated monocyte archetype. **(C)** Heatmap of gene expression and regulon activity in monocytes undergoing maturation, ordered by pseudotime distances along PHATE manifolds that transition from quiescent monocytes towards a control-enriched mature monocyte-derived macrophage archetype. In both heatmaps: annotation bars represent the pseudotime distance, disease status, and subject for each cell; expression values are centered and scaled. **(D)** Violin plots of pathway activity scores, grouped by cell type, separated by disease state. Depicted pathway scores from left to right are: GO:0045087 - innate immune response, GO:0006958 - complement activation, classical pathway, GO:0019882 - antigen processing and presentation, GO:0006911 - phagocytosis, engulfment. * represents FDR-adjusted p-values < 0.05, calculated using the Wilcoxon signed-rank test. Mo: monocyte; MoMΦ: monocyte-derived macrophage; alvMΦ: alveolar macrophage; PMN: polymorphonuclear neutrophil.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

Fig. 3.

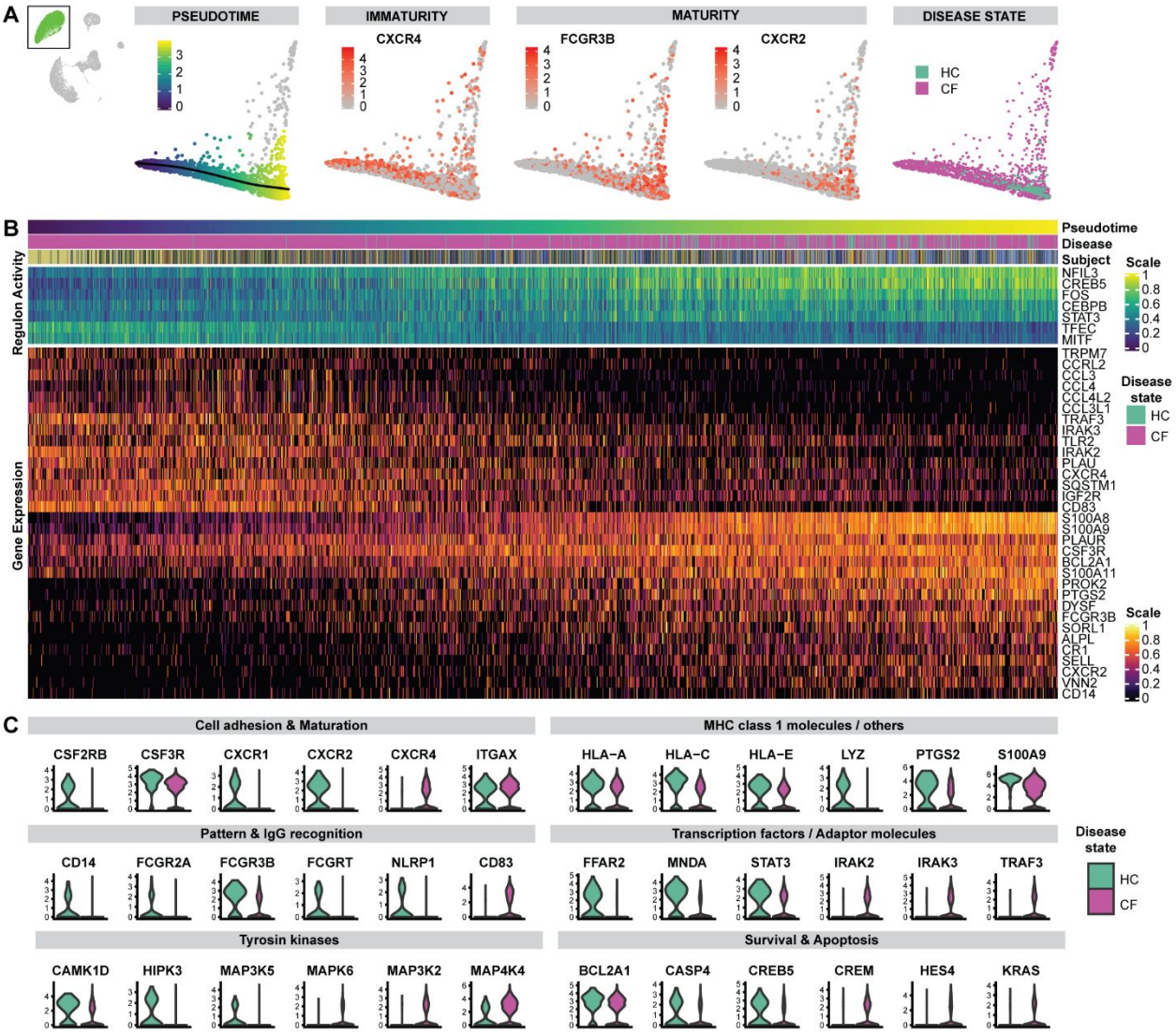


Fig.3. An Immature Pro-inflammatory Archetype Prevails Among CF Airway PMN. **(A)** PHATEs of PMN, colored by: (i) pseudo time from immature to mature PMNs, (ii) examples of canonical marker features of immaturity (CXCR4) and maturity (FCGR3B, CXCR2) in peripheral PMN, (iii) disease state. The cells deviating upward are PMN expressing heat-shock response genes, for PHATE embedding colored by gene expression of HSPA1A, HSPH1, and DNAJB1, see Supplemental Fig. E6A). For corresponding PHATE embedding colored by disease state and subjects, see Supplemental Fig. E6B. **(B)** Heatmap of gene expression and

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

regulon activity in PMNs, ordered by pseudotime distances along PHATE manifolds that transition from CF-enriched regions of immature and activated PMN archetype towards control-enriched mature PMN archetype. Annotation bars represent the pseudotime distance, disease status, and subject for each cell; expression values are centered and scaled. **(C)** Violin plots of differentially expressed genes comparing CF and control PMN populations (for p-values see Supplemental Data file E3), grouped by disease state, and sorted thematically.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

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

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

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%

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

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).

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

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

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

(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

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

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))

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

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.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

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.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

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

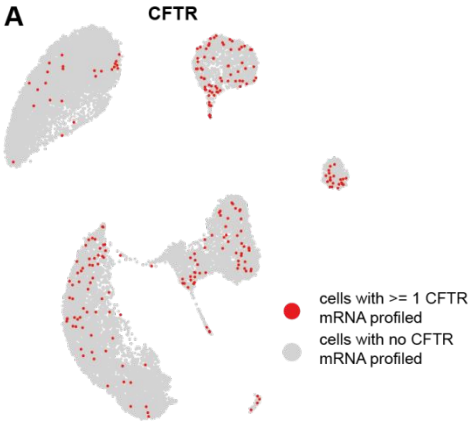
To identify classifier genes, differential gene expression of immune cell types of this study and analogue cell types from an independent scRNAseq, a dataset of 28 healthy distal lung samples (E7) was established using Seurat's FindAllMarkers with an absolute log fold change threshold of 1 (the lung 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

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

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.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

Fig. E1.



B

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.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

Fig. E2.

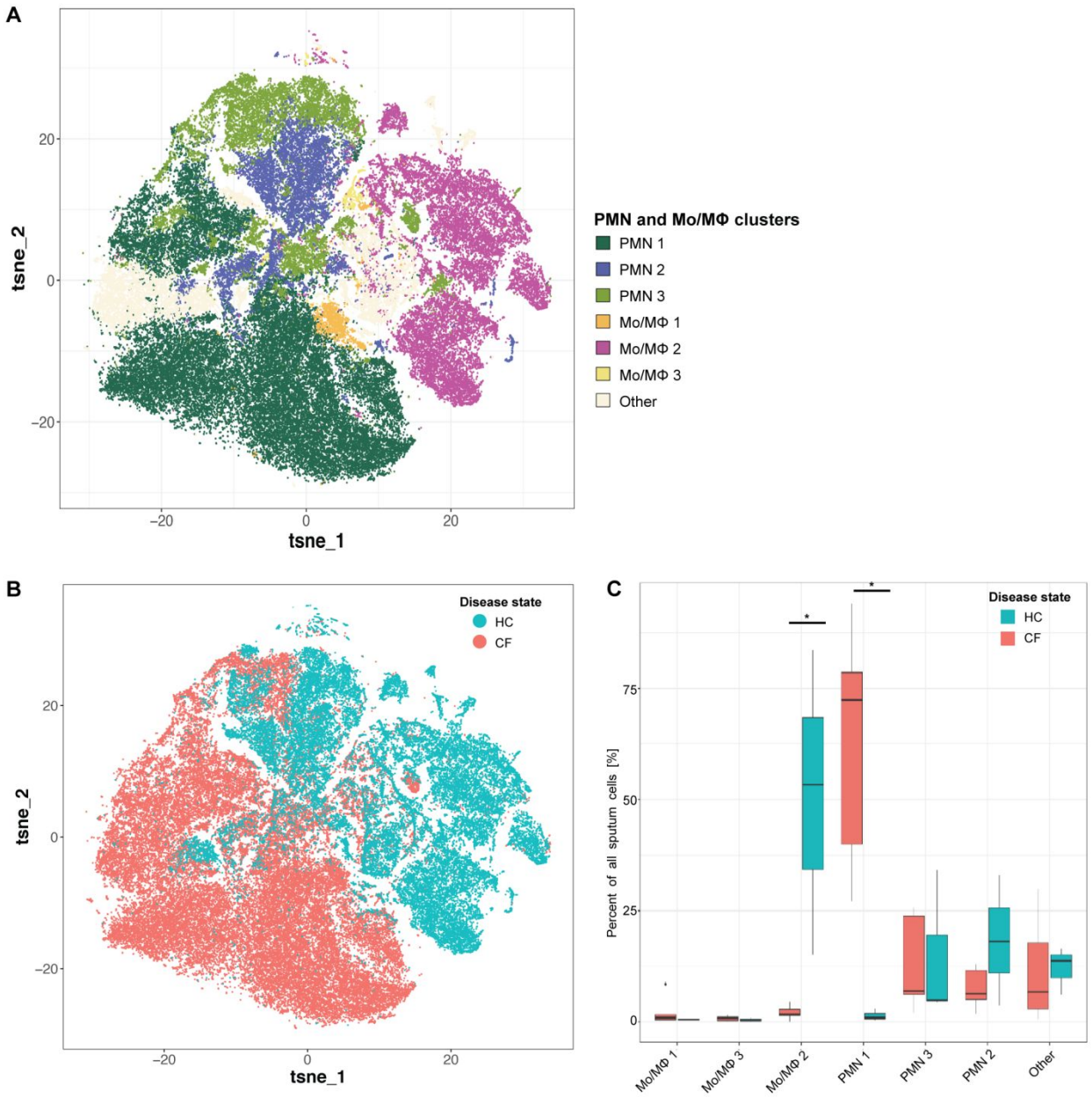


Fig. E2. Validation of the shift in major immune cell types in sputum of CF compared to HC. **(A)** RPhenograph clustering of Sputum CyTOF in patients with cystic fibrosis (CF) and healthy controls (HC) demonstrates differences in the populations of immune cells. The sputum of patients with CF is characterized by high percentages of neutrophils, while sputum from HC is

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

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.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

Fig. E3.

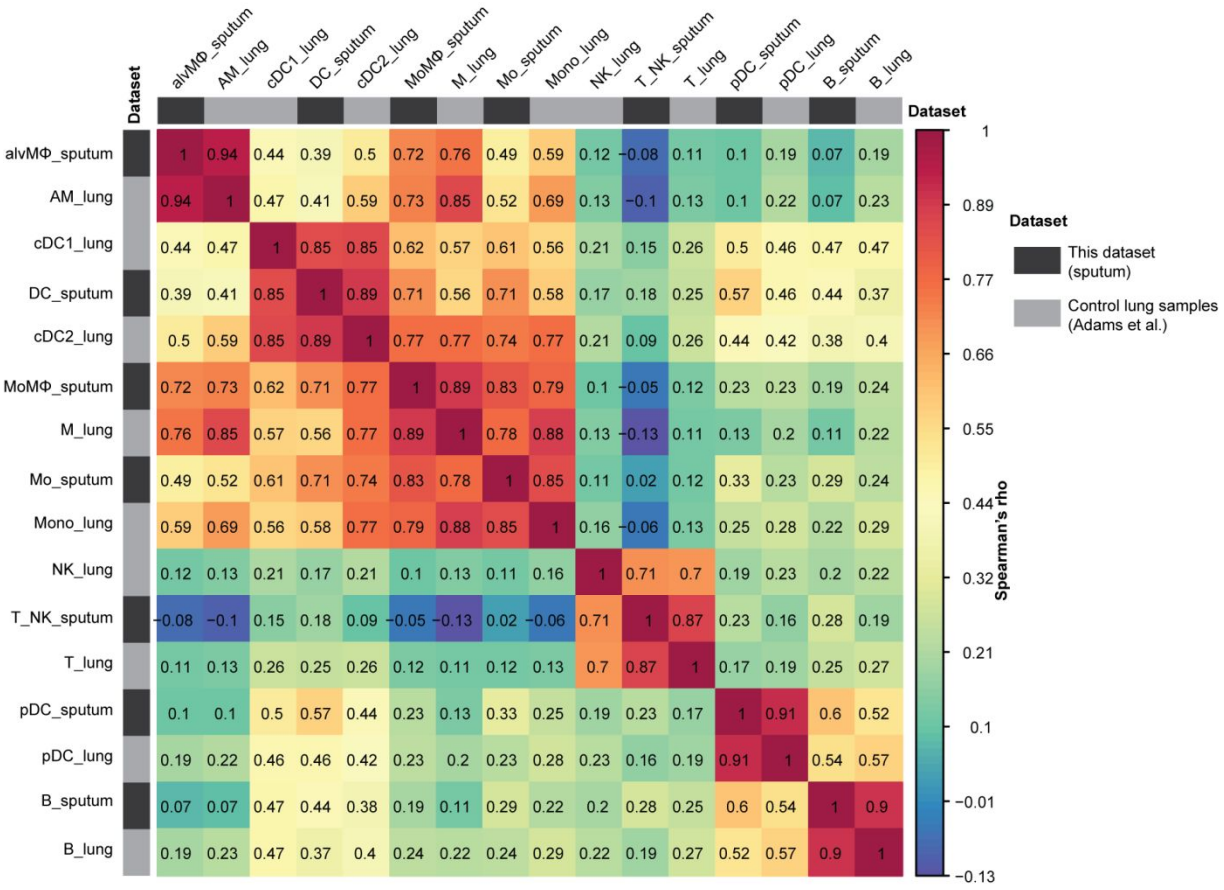


Fig. E3. Concordance of cell type annotations. Correlation matrix of immune cell populations of this study and analogous cell types from an independent scRNA sequencing dataset of distal lung samples, subsetting to the 28 healthy controls. Matrix fields are colored by Spearman's rho, cell types are ordered 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)).

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

Fig. E4.

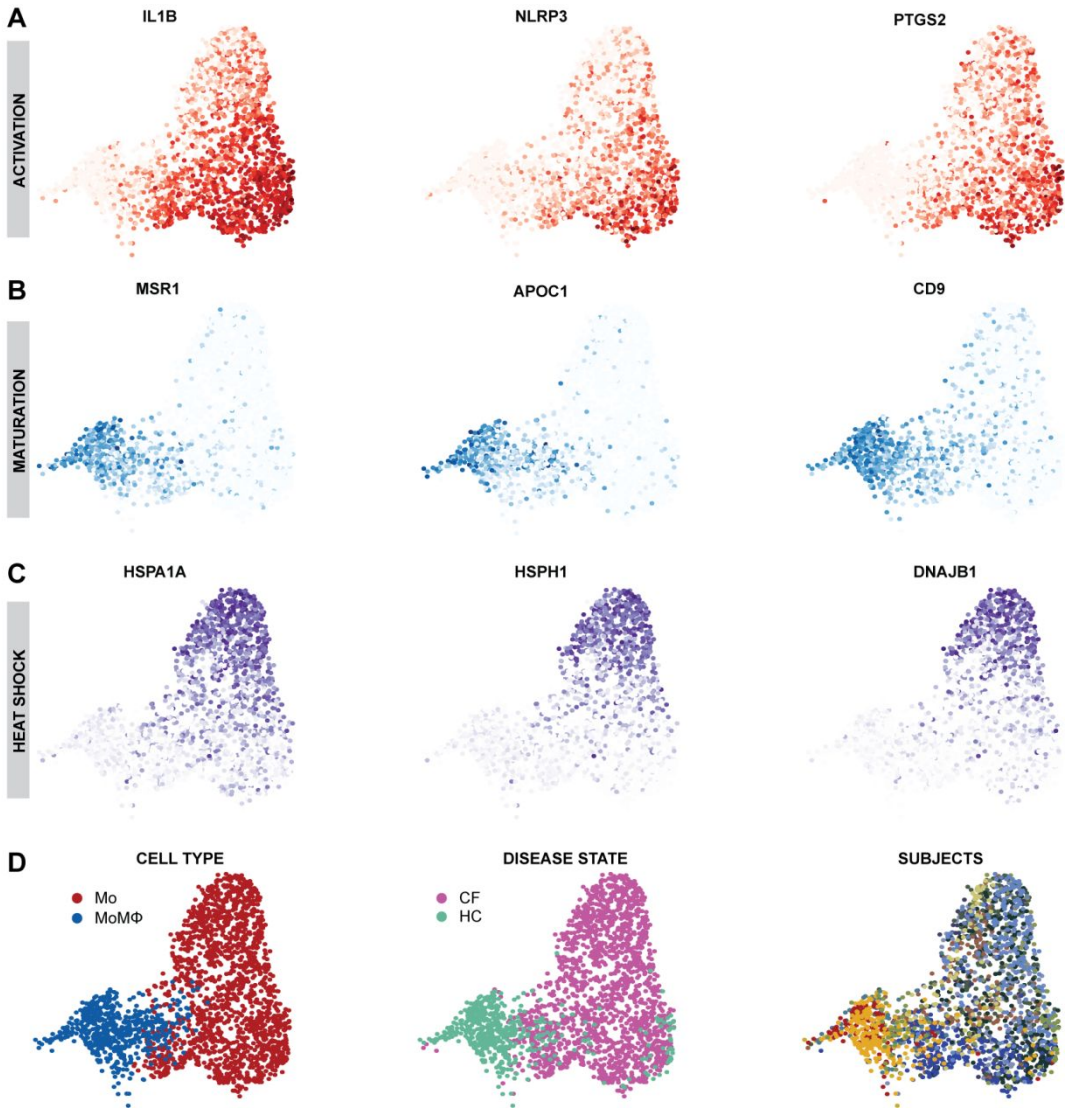


Fig. E4. Expression of selected marker genes of Mo/MoMΦ trajectories on UMAPs. **(A)** UMAP, zoomed in on Mo and MoMΦ, colored by expression of inflammatory genes IL1B, NLRP3, PTGS2. **(B)** UMAP, zoomed in on Mo and MoMΦ, colored by expression of mature macrophage genes MSR1, APOC1, CD9. **(C)** UMAP, zoomed in on Mo and MoMΦ, colored by expression of heat shock genes HSPA1A, HSPH1, DNAJB1. **(D)** UMAP, zoomed in on Mo and MoMΦ,

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

colored by (i) cell type, (ii) disease state, (iii) subjects. CF: Cystic Fibrosis, HC: Healthy Control, Mo: Monocyte; MoMΦ: monocyte-derived macrophage.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

Fig. E5.

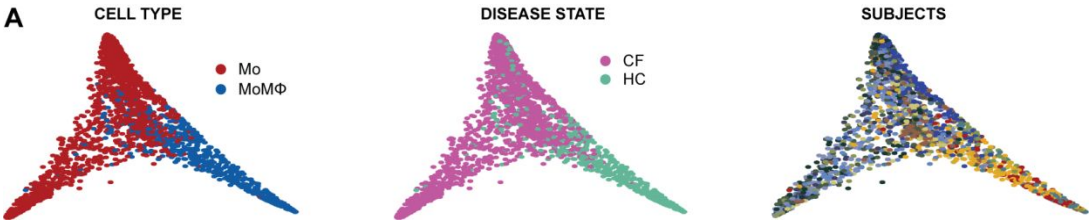


Fig. E5. Additional annotations of Mo/MoMΦ on PHATE embedding. **(A)** UMAP of Mo and MoMΦ colored by (i) Cell type, (ii) Disease state, (iii) Subjects.

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

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

Fig. E6.

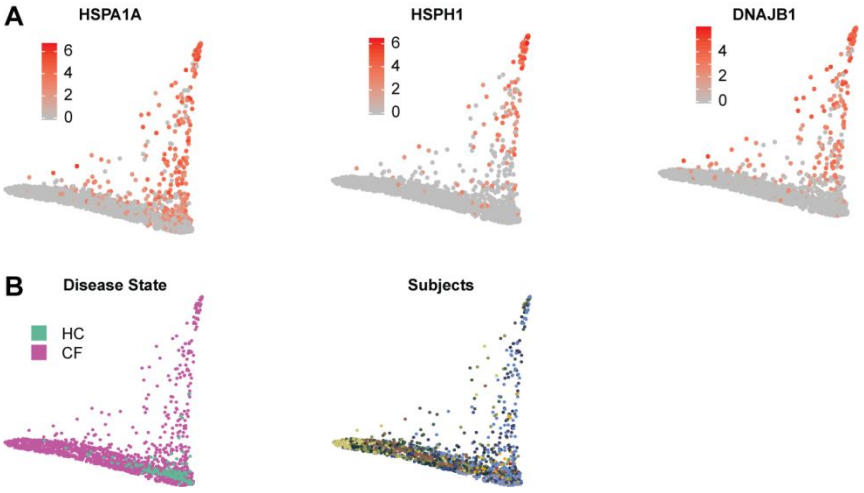


Fig. E6. Additional annotations of PMN on PHATE embedding. **(A)** PHATE of PMN colored by expression of heat shock genes HSPA1A, HSPH1 and DNAJB1. **(B)** PHATE of PMN colored by disease state (HC: Healthy Control, CF: Cystic Fibrosis) and subjects.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

Fig. E7.

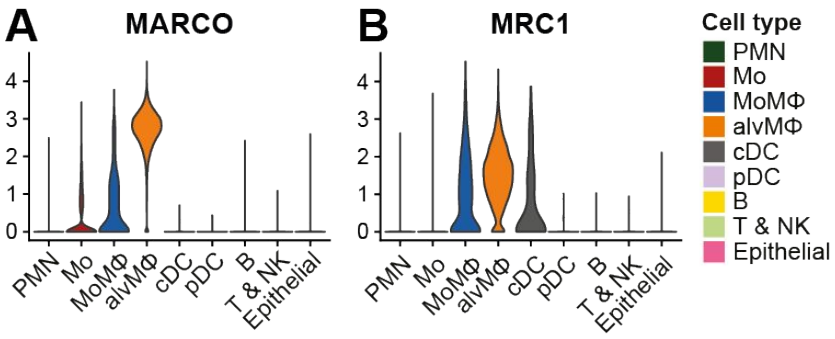


Fig. E7. Violin plots of **(A)** *MARCO* and **(B)** *MRC1*, grouped by cell type.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

Fig. E8.

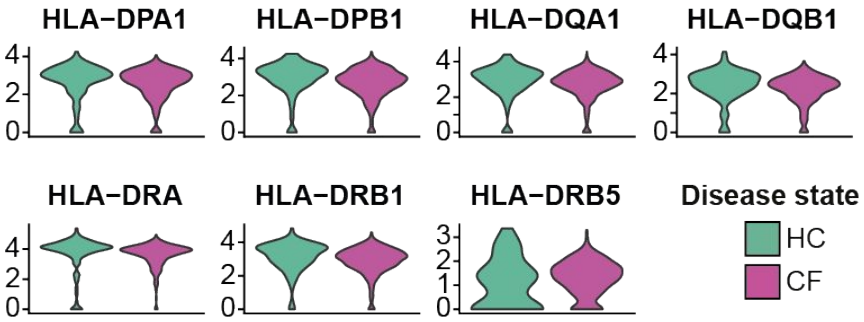


Fig. E8. Violin plots of major histocompatibility complex class 2 genes in B cells, grouped by disease state. For all: $p>0.05$, i.e. not significantly different.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

Fig. E9.

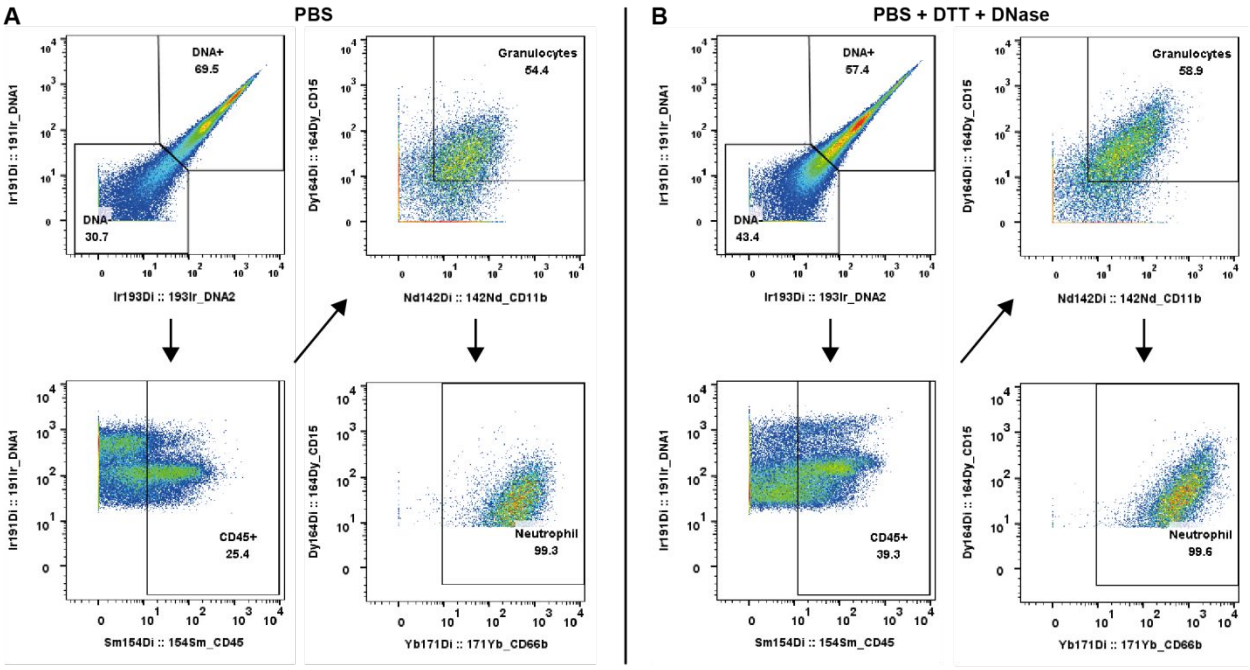


Fig. E9. Viable cell yield using our sputum processing protocol is comparable to previously established approaches for sputum processing (proof-of-principle). Aliquots from the same sample were processed using **(A)** our PBS-only protocol or **(B)** treated sequentially with DNAse (0.56kU/ml, D4527-500KU, 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).

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

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.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

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,

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

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.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

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.

AJRCCM Articles in Press. Published June 30, 2020 as 10.1164/rccm.202004-0991OC

14. Sassano MF, Ghosh A, Tarran R. Tobacco smoke constituents trigger cytoplasmic calcium release. *Appl In Vitro Toxicol* 2017;3: 193–198.
15. Czoli CD, Goniewicz ML, Palumbo M, Leigh N, White CM, Hammond D. Identification of flavouring chemicals and potential toxicants in e-cigarette products in Ontario, Canada. *Can J Public Health* [online ahead of print] 25 Apr 2019; DOI: 10.17269/s41997-019-00208-1.

Toward Early Detection of Idiopathic Pulmonary Fibrosis

Since their emergence as a frequent and potentially clinically meaningful finding in computed tomography (CT) screenings of smokers a decade ago (1), interstitial lung abnormalities (ILAs) have drawn significant interest and controversy. A specific set of radiologic abnormalities on chest CT scans, ILAs are relatively common and can be found in up to 10% of lung cancer screenings and older smokers (2). ILAs have traditionally been taken lightly by physicians and affected individuals alike, as symptoms in subjects with ILA are often lacking or very mild, and the prognostic significance of ILA was unknown. This has changed in recent years with the increased recognition that individuals with ILAs are at higher risk of death and exhibit higher rates of lung restriction (3–5) and that on tissue histology they often exhibit fibrosis (6). The possibility that individuals with ILAs may represent a population at risk for subsequent development of idiopathic pulmonary fibrosis (IPF) or other interstitial lung disease (ILD) is of particular importance, because of the potential for more effective interventions when the disease is diagnosed early. The connection between ILAs and pulmonary fibrosis has been supported by radiologic progression of ILAs, the presence of ILAs in asymptomatic family members of individuals with familial pulmonary fibrosis, and the significant association of ILAs with rs35705920 in the promotor region of MUC5B (Mucin 5B, oligomeric mucus/gel-forming) (4), the same gene variant that accounts for approximately 30% of cases of IPF (7). However, so far, the genetic overlap between patients with ILAs and IPF has not been studied in detail.

In this issue of the *Journal*, Hobbs and colleagues (pp. 1402–1413) performed a meta-analysis using available genome-wide data of 1,699 subjects with ILA and 10,274 control subjects from six cohorts and compared the results with genetic associations in patients with IPF (8). Because subpleural ILAs are believed to be more clinically relevant, they performed the analysis of ILAs in general and subpleural ILAs separately. In the ILA analysis, they identified three genome-wide significant associations that included the known MUC5B promoter polymorphism rs35705950 and two novel loci: rs6886640 at 5q12 near IPO11 (importin 11) and rs73199442 at 3q13 near the long noncoding RNA FCF1P3 (FCF1 pseudogene 3). In the subpleural ILA analysis—in addition to MUC5B—they identified a genetic association at the 6q15 locus with rs7744971 near HTR1E (5-hydroxytryptamine receptor 1E). None of the novel ILA loci replicated in IPF genome-wide association studies. Of the 12 reported genome-wide association study loci for IPF, only the MUC5B variant reached genome-wide significance, whereas the genetic variants near DPP9 (dipeptidyl peptidase 9), DSP (desmoplakin), FAM13A (family with sequence similarity 13 member A), and IVD (isovaleryl-CoA dehydrogenase) were nominally associated with ILA.

The findings of this study have several major implications. The most important is that although individuals with ILAs represent a population at risk for IPF, they are not synonymous with the IPF population. Only a subset of individuals with ILA exhibit a genetic risk profile that is similar to individuals with IPF, whereas others exhibit genetic associations that do not occur in IPF: the reported odds ratio is 1.97 for rs35705950 for all ILAs, and 2.22 when subsetting to subpleural ILAs, but 4.84 for IPF. None of the other IPF risk loci were significant on a genome-wide level, and all of them had a lower odds ratio in ILA. This could suggest an ILA subpopulation that is at risk of developing IPF but is being diluted by a larger fraction of subjects with ILA who do not share the same genetic risk. The finding of three novel ILA genetic associations not observed in IPF also indicates a potentially distinct entity, possibly a predisposition to other non-IPF ILDs or even the presence of gene variants that reduce the probability of progression of ILAs to fibrosis and may be protective. Regardless of their potential functional relevance, the finding of variants associated with ILA but not IPF, if replicated, could be useful developing a polygenic genetic risk profile. This is important because, currently, chest CT screenings to detect early IPF are not clinically feasible or justified. The results of this study should encourage investigators to design further studies assessing whether genetic risk profiling, potentially combined with other noninvasive biomarkers, could be used to prioritize individuals for CT screening.

Although exciting and intriguing, this study has some limitations that should be highlighted. Of course, the most obvious limitation of the discovered novel ILA associations is the lack of an independent replication cohort, but the limitations regarding the negative results should not go unnoticed. Indeed, only MUC5B reached genome-wide significance in this study, but the main study population consisted of data obtained from several cohorts that were not designed to capture early ILD. These populations differed in the definitions of ILA, the depth of phenotyping, and the original aims of the studies. Thus, it is highly possible that although the strongest association (MUC5B) was able to emerge, other valid associations simply were drowned by the sea of differences and may emerge again if comparably sized future studies are designed to detect ILAs using standard definitions, adjudicated radiological reading, and patient phenotyping.

In summary, the study by Hobbs and colleagues (8) represents a major step toward better understanding ILAs as tools for defining

Supported by Department of Defense Discovery Award W81XWH-19-1-0131 (J.C.S.) and NIH NHLBI grants R01HL127349, R01 HL141852, and UH2 HL123886 (N.K.).

Originally Published in Press as DOI: 10.1164/rccm.201908-1530ED on August 14, 2019

populations that should be targeted for early detection of IPF. This is a critically important mission. Although there has been considerable progress in the development of novel therapeutic options for IPF, it is highly unlikely that any of the drugs currently in the pipeline will be able to reverse the extensive lung remodeling that is often observed when patients initially present. On the other hand, it is possible that therapeutic targeting of minimal fibrotic lesions—before extensive remodeling and bronchiolization have occurred—will allow complete eradication of the disease. Thus, to truly eradicate IPF, we need a paradigm shift from focusing on developing cohorts of patients already diagnosed with IPF toward cohorts of individuals highly likely to develop the disease. We could use these cohorts to develop and test algorithms for early detection. Then we could implement a multistep strategy to eradicate IPF: identification of a population with high risk for ILA and performing chest CT screenings when appropriate; in subjects with ILA, identification of patients who will develop IPF; and last, systematic study of interventions aimed at preventing progression to IPF. In an editorial in 2012 (9) discussing an early report on ILAs (10), Dr. David Lederer compared our traditional symptom-linked diagnosis of IPF to diagnosing coronary artery disease only after the patient presented with a myocardial infarction and called for new ways for risk prediction and early detection of IPF. Seven years later, the article by Hobbs and colleagues (8) suggests that we can move forward—that we can diagnose IPF while the horse is still in the barn. ■

Author disclosures are available with the text of this article at www.atsjournals.org.

Jonas Christian Schupp, M.D.
Naftali Kaminski, M.D.
Section of Pulmonary, Critical Care, and Sleep Medicine
Yale University School of Medicine
New Haven, Connecticut

ORCID IDs: 0000-0002-7714-8076 (J.C.S.); 0000-0001-5917-4601 (N.K.).

References

1. Washko GR, Hunninghake GM, Fernandez IE, Nishino M, Okajima Y, Yamashiro T, *et al*.; COPDGene Investigators. Lung volumes and emphysema in smokers with interstitial lung abnormalities. *N Engl J Med* 2011;364:897–906.
2. Jin GY, Lynch D, Chawla A, Garg K, Tammemagi MC, Sahin H, *et al*. Interstitial lung abnormalities in a CT lung cancer screening population: prevalence and progression rate. *Radiology* 2013;268: 563–571.
3. Putman RK, Hatabu H, Araki T, Gudmundsson G, Gao W, Nishino M, *et al*.; Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints (ECLIPSE) Investigators; COPDGene Investigators. Association between interstitial lung abnormalities and all-cause mortality. *JAMA* 2016;315:672–681.
4. Putman RK, Gudmundsson G, Araki T, Nishino M, Sigurdsson S, Gudmundsson EF, *et al*. The *MUC5B* promoter polymorphism is associated with specific interstitial lung abnormality subtypes. *Eur Respir J* 2017;50:1700537.
5. Araki T, Putman RK, Hatabu H, Gao W, Dupuis J, Latourelle JC, *et al*. Development and progression of interstitial lung abnormalities in the Framingham Heart Study. *Am J Respir Crit Care Med* 2016;194: 1514–1522.
6. Miller ER, Putman RK, Vivero M, Hung Y, Araki T, Nishino M, *et al*. Histopathology of interstitial lung abnormalities in the context of lung nodule resections. *Am J Respir Crit Care Med* 2018;197: 955–958.
7. Kaur A, Mathai SK, Schwartz DA. Genetics in idiopathic pulmonary fibrosis pathogenesis, prognosis, and treatment. *Front Med (Lausanne)* 2017;4:154.
8. Hobbs BD, Putman RK, Araki T, Nishino M, Gudmundsson G, Gudnason V, *et al*.; COPDGene Investigators, ECLIPSE Investigators, SPIROMICS Research Group, and UK ILD Consortium. Overlap of genetic risk between interstitial lung abnormalities and idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2019;200:1402–1413.
9. Lederer DJ. Secondary prevention of idiopathic pulmonary fibrosis: catching the horse still in the barn. *Am J Respir Crit Care Med* 2012; 185:697–699.
10. Doyle TJ, Washko GR, Fernandez IE, Nishino M, Okajima Y, Yamashiro T, *et al*.; COPDGene Investigators. Interstitial lung abnormalities and reduced exercise capacity. *Am J Respir Crit Care Med* 2012;185: 756–762.

The Respiratory Mucosa: Front and Center in Respiratory Syncytial Virus Disease

Infantile bronchiolitis is a major scourge of early childhood, and winter outbreaks fill the pediatric wards with wearisome regularity. Most cases are caused by respiratory syncytial virus (RSV), which was first isolated in 1956. Despite a vast amount of research in both human and animal models, a deep understanding of the inefficiency of protective immunity and, indeed, of the pathogenesis of RSV disease has been frustratingly slow to come by.

Most infants will be infected by RSV before their second birthday, with the risk of severe disease peaking at just 2 months of age. Despite the relative antigenic stability of the virus, reinfections with RSV occur throughout life. Studying disease in infants with primary disease presents considerable technical and logistical challenges; therefore, animal models (especially cotton rats, mice, and cows) have been widely used to enhance our understanding of primary infection and vaccine-enhanced disease. These models have been central in our efforts to understand the host immune response to RSV and the role of these responses in causing inflammatory bronchiolitis, but they do not recapitulate human disease in every detail.

Although animal models have advanced our understanding of the pathogenesis of bronchiolitis, a role for the type 2 immune

Originally Published in Press as DOI: 10.1164/rccm.201907-1306ED on August 14, 2019