DISEASES AND DISORDERS

Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis

Taylor S. Adams[1]*, Jonas C. Schupp[1]*, Sergio Poli[2]*, Ehab A. Ayaub[2], Nir Neumark[1,3], Farida Ahangari[1], Sarah G. Chu[2], Benjamin A. Raby[2,4,5], Giuseppe Deluliis[1], Michael Januszyk[6], Qiaonan Duan[6], Heather A. Arnett[6], Asim Siddiqui[6], George R. Washko[2], Robert Homer[7,8], Xiting Yan[1], Ivan O. Rosas[2]*†, Naftali Kaminski[1]*†

Copyright © 2020 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).

We provide a single-cell atlas of idiopathic pulmonary fibrosis (IPF), a fatal interstitial lung disease, by profiling 312,928 cells from 32 IPF, 28 smoker and nonsmoker controls, and 18 chronic obstructive pulmonary disease (COPD) lungs. Among epithelial cells enriched in IPF, we identify a previously unidentified population of aberrant basaloid cells that coexpress basal epithelial, mesenchymal, senescence, and developmental markers and are located at the edge of myofibroblast foci in the IPF lung. Among vascular endothelial cells, we identify an ectopically expanded cell population transcriptomically identical to bronchial restricted vascular endothelial cells in IPF. We confirm the presence of both populations by immunohistochemistry and independent datasets. Among stromal cells, we identify IPF myofibroblasts and invasive fibroblasts with partially overlapping cells in control and COPD lungs. Last, we confirm previous findings of profibrotic macrophage populations in the IPF lung. Our comprehensive catalog reveals the complexity and diversity of aberrant cellular populations in IPF.

INTRODUCTION

Idiopathic pulmonary fibrosis (IPF) is a progressive lung disease characterized by irreversible scarring of the distal lung, leading to respiratory failure and death (*1*). Despite substantial progress in our understanding of pulmonary fibrosis in laboratory animals, we have a limited perspective of the cellular and molecular processes that determine the IPF lung phenotype. IPF is still best described by its histopathological pattern of usual interstitial pneumonia (UIP), which includes the presence of fibroblast foci, hyperplastic alveolar epithelial cells that localize adjacent to fibroblastic foci, and a distortion of airway architecture combined with an accumulation of microscopic airway epithelial–lined cysts known as "honeycombs" in the distal parenchyma and the lack of evidence for other conditions (*1*).

Evidence for molecular aberrations in the IPF lung have mostly been obtained by following hypotheses derived from animal models of disease, from discovery of genetic associations in humans, or from genes differentially expressed in transcriptomic studies of bulk IPF tissue with limited cellular resolution (*2*). Recent studies have demonstrated the value of single-cell RNA sequencing (scRNA-seq) by identifying profibrotic macrophages in lungs of human and mice with pulmonary fibrosis (*3*). Here, we harness the cell-level resolution afforded by scRNA-seq to provide an atlas of the extent of complexity and diversity of aberrant cellular populations in the three major parenchymal compartments of the IPF lung: the epithelium, endothelium, and stroma.

RESULTS

We profiled 312,928 cells from distal lung parenchyma samples obtained from 32 IPF lungs (45 libraries yielding 147,169 cells), 18 chronic obstructive pulmonary disease (COPD) lungs (24 libraries yielding 69,456 cells), and 28 control donor lungs [38 libraries, 96,303 cells; 11 of these control lungs were included in another dataset we recently published (Fig. 1A and table S1) (*4*)], identifying 38 discrete cell types (Fig. 1B) based on distinct markers (Fig. 1C and data S2 to S6) with no discernible batch or cell cycle effects on data architecture (figs. S1 and S2). Manually curated cell classifications are consistent with automated annotations drawn from several independent databases (fig. S3). The detailed cellular repertoires of epithelial, endothelial, and mesenchymal cells are provided below. Our data have been deposited on Gene Expression Omnibus (GEO) (GSE136831), and the results can be explored through the IPF Cell Atlas Data Mining portal (www.ipfcellatlas.com) (*5*).

The epithelial cell repertoire of the fibrotic lung is markedly changed and contains aberrant basaloid cells

In nondiseased tissue, we identified all known lung epithelial cells populations, including alveolar type 1 (AT1) and type 2 (AT2) cells, ciliated cells, basal cells, goblet cells, club cells, pulmonary neuroendocrine cells, and ionocytes (Figs. 1, B and C, and 2A). The epithelial cell repertoire of IPF lungs is characterized by an increased proportion of airway epithelial cells (IPF versus control Wilcoxon false discovery rate (FDR) adjusted $P < 0.05$ for basal, ciliated, and goblet cells; data S7 and S12) and substantial decline in alveolar epithelial cells [Wilcoxon FDR $P < 5 \times 10^{-6}$ for AT1 and AT2; Fig. 2, A and B, and data S12], a pattern consistent with previous

[1]Section of Pulmonary, Critical Care and Sleep Medicine, Yale School of Medicine, New Haven, CT, USA. [2]Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [3]Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA. [4]Division of Pulmonary Medicine, Boston's Children Hospital, Harvard Medical School, Boston, MA, USA. [5]Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [6]NuMedii, Inc., San Mateo, CA, USA. [7]Department of Pathology, Yale School of Medicine, New Haven, CT, USA. [8]Pathology and Laboratory Medicine Service and VA CT HealthCare System, West Haven, CT, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: irosas@rics.bwh.harvard.edu (I.O.R.); naftali.kaminski@yale.edu (N.K.)

Downloaded from http://advances.sciencemag.org/ on July 10, 2020

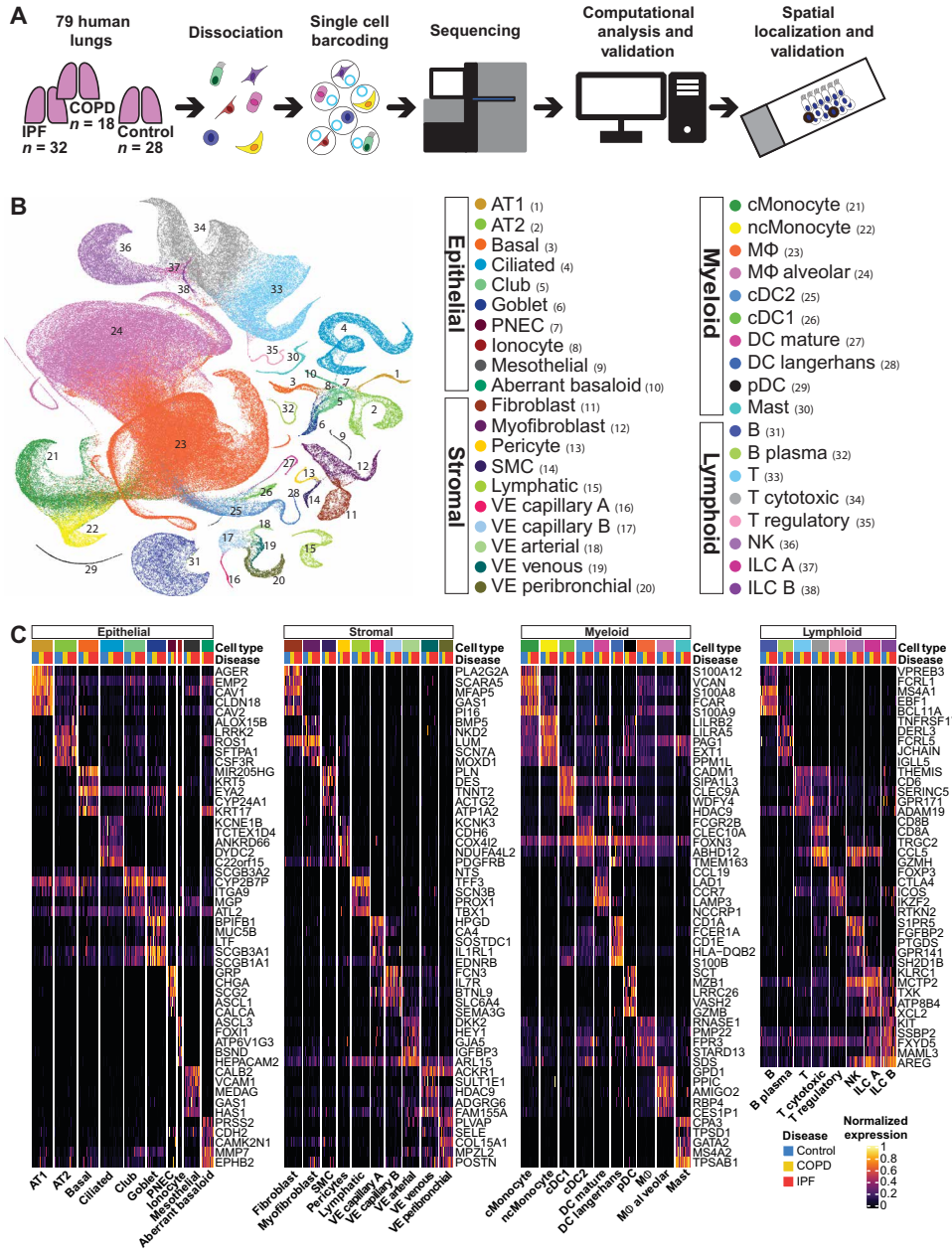

Fig. 1. Profiling human lung heterogeneity with scRNA-seq. (**A**) Overview of experimental design. (i) Disease lung explants and unused donor lungs collected. (ii) Lungs dissociated to single-cell suspension. (iii) Droplet-based scRNA-seq library preparation (iv) sequencing. (v) Exploratory analysis. (vi) Spatial localization with IHC. (**B**) Uniform Manifold Approximation and Projection (UMAP) representation of 312,928 cells from 32 IPF, 18 COPD, and 28 control donor lungs; each dot represents a single cell, and cells are labeled as one of 38 discrete cell varieties. AT, alveolar type; cDC, classical dendritic cell; pDC, plasmacytoid dendritic cell; M, macrophage; NK, natural killer; ILC, innate lymphoid cell; PNEC, pulmonary neuroendocrine cell; SMC, smooth muscle cell; VE, vascular endothelial. (**C**) Heat map of marker genes for all 38 identified cell types, categorized into four broad cell categories. Each cell type is represented by the top five genes ranked by false discovery rate (FDR) adjusted *P* value of a Wilcoxon rank sum test between the average expression per subject value for each cell type against the other average subject expression of the other cell types in their respective grouping. Each column represents the average expression value for one subject, hierarchically grouped by disease status and cell type. Gene expression values are unity normalized from 0 to 1 across rows within each categorical cell type group.

reports (*6*, *7*) and further supported by deconvolution of recently published (*8*) IPF bulk RNA-seq data that confirmed reduction in genes originating from alveolar epithelial cells in the distal lung in IPF (fig. S4). Impressively, nearly every epithelial cell type we identified exhibited profound changes in gene expression in the IPF lung compared to control or COPD [data S8 and S10; IPF cell Atlas Data Mining portal (*5*)]. Among epithelial cells, we identified a population of cells that is transcriptionally distinct from any epithelial cell type previously described in the lung that we termed aberrant basaloid cells (Fig. 2C). In addition to epithelial markers, these cells express the basal cell markers TP63, KRT17, LAMB3, and LAMC2 but do not express other established basal markers such as KRT5

Downloaded from http://advances.sciencemag.org/ on July 10, 2020

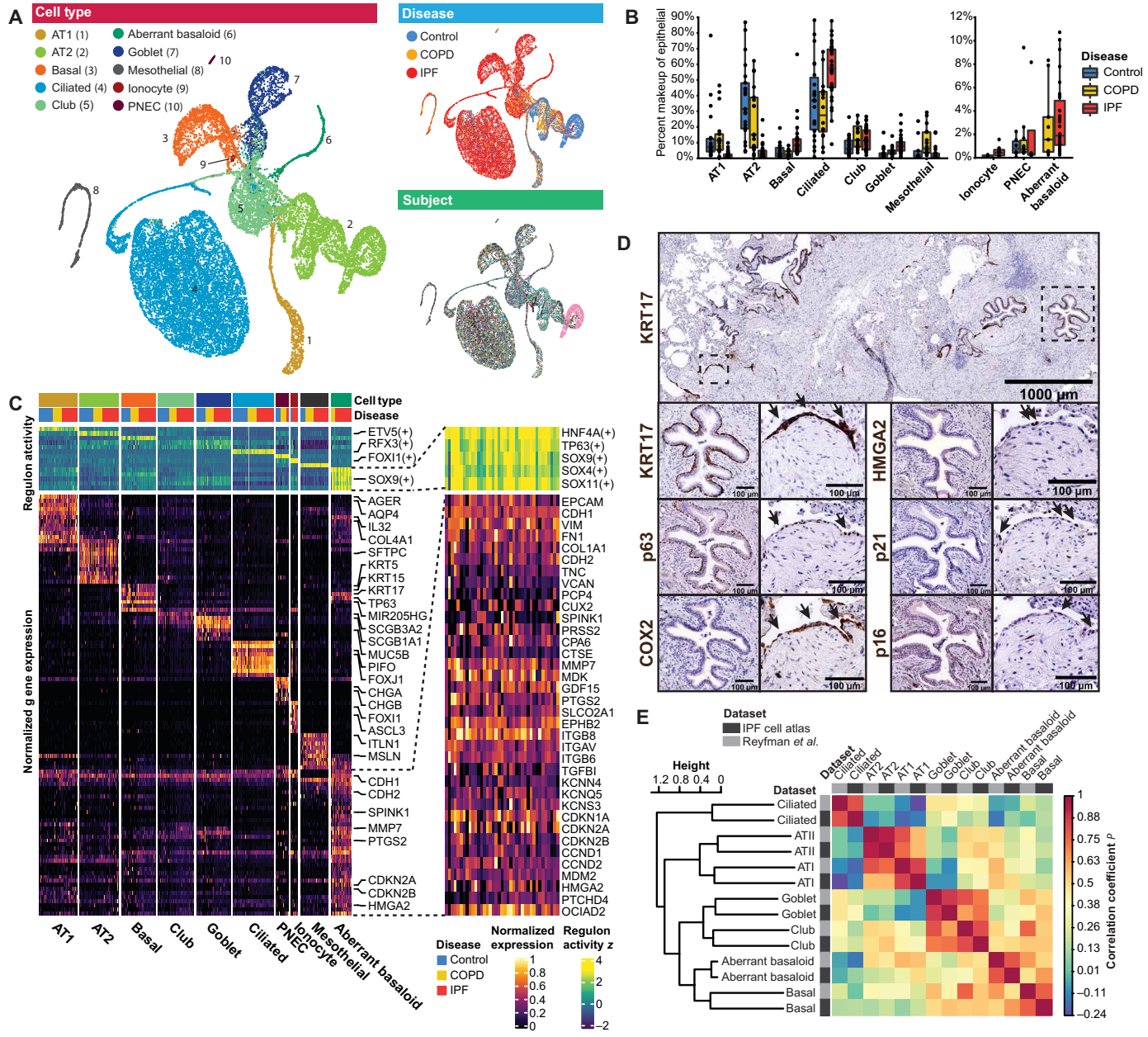

Fig. 2. Identification of aberrant basaloid cells in IPF and COPD lungs. (**A**) UMAPs of 21,184 epithelial cells from 32 IPF, 18 COPD, and 28 control lungs labeled by cell type (top left), disease status (bottom right), and subject (bottom right). In the subject plot, each color depicts a distinct subject. (**B**) Boxplots representing the nonzero percent makeup distributions of epithelial cell types as a proportion of all sampled epithelial cells per subject within each disease group. Each dot represents a single subject, and whiskers represent 1.5 × interquartile range (IQR). FDR-adjusted Wilcoxon rank sum test results comparing IPF and control proportions are reported in data S12. (**C**) Heat map of average gene expression and predicted transcription factor activity per subject across each of the identified epithelial cell types. Columns are hierarchically ordered by disease status and cell type. The average gene expression per subject per cell type is unity normalized between 0 and 1 across samples. Top (green): Transcription factor signatures predicted by analysis with pySCENIC (*43*), and *z* scores are calculated across samples. Right: Zoom annotation of distinguishing markers for aberrant basaloid cells. (**D**) IHC staining of aberrant basaloid cells in IPF lungs: epithelial cells covering fibroblast foci are $p63^{+}$ $KRT17^{+}$ basaloid cells staining COX2-, p21-, and HMGA2-positive, while basal cells in bronchi do not. (**E**) Correlation matrix of epithelial cell populations were identified and reannotated in an independent dataset (*3*) with analogous cell types from our data. Matrix cells are colored by Spearman's rho, and cell populations are ordered with hierarchical clustering. The origin dataset for each cell population is denoted by in the annotation bars.

and KRT15 (*9*). These cells express markers of epithelial-mesenchymal transition (EMT) such as VIM, CDH2, FN1, COL1A1, TNC, and HMGA2 (*10*), senescence-related genes including CDKN1A, CDKN2A, CCND1, CCND2, MDM2, and GDF15 (Fig. 2C, right) (*11*, *12*), and the highest normalized expression levels across all profiled cell types of established IPF-related molecules, such as matrix metallopeptidase 7 (MMP7) (*13*), combined with $\alpha_V\beta_6$ integrin subunits (*14*) and EPHB2 (*15*). When assessed by transcription factor motif enrichment, the

Downloaded from http://advances.sciencemag.org/ on July 10, 2020

cells exhibit high levels of expression and evidence of downstream activation of SOX9 (Fig. 2C, top), a transcription factor critical to distal airway development (*16*) and repair (*17*). No aberrant basaloid cells could be detected in control lungs (Fig. 2, A and B). Of the 483 aberrant basaloid cells identified, 448 were from IPF lungs compared to just 33 cells from COPD lungs, representing 3.5% of all epithelial cells in IPF compared with 1.1% in COPD (χ^2 $P < 2.2 \times 10^{-16}$). We confirmed the presence and localized these cells in the IPF lung by immunohistochemistry (IHC) using p63, KRT17, HMGA2, COX2, and p21 as markers (Fig. 2D; control stains can be found in fig. S5). In IPF lungs, these cells consistently localize to the epithelial layer covering myofibroblast foci. To independently validate our results, we reanalyzed the IPF single-cell data published by Reyfman *et al.* (*3*) identifying 80 aberrant basaloid cells, distinguished by a similar marker profile to the cells described herein (fig. S6). We observe greater correlation in aberrant basaloid cells between datasets (Spearman's rho, 0.81) than across epithelial cell types within each dataset (Fig. 2E), confirming our results.

The endothelial cell repertoire of the IPF lung contains ectopic peribronchial endothelial cells

While changes in vascular endothelium have been long noted in the IPF lung (*18*), little is known about their cellular and molecular characteristics. Cluster analysis of vascular endothelial (VE) cells reveal four populations readily characterized as capillary, arterial, or venous VE cells (Fig. 3, A to C). A fifth VE population is best distinguished by its expression of COL15A1. IHC assessment of COL15A1 in healthy lungs obtained from the Human Protein Atlas (*19*) suggests that COL15A1$^+$ VE cells are restricted to vasculature adjacent to major airways (fig. S7); we consequently refer to these cells as peribronchial VE (pVE). While pVE cells are found in subjects from all disease states, they compose a substantially higher portion of the overall VE makeup within IPF samples compared to controls or COPD (IPF versus control Wilcoxon FDR $P = 8.06 \times 10^{-5}$, data S12; median percentage among all VE: 54, 8.9, and 7.1%, respectively; Fig. 3C). Localization of pVE cells using the pan-endothelial marker CD31 alongside COL15A1 confirms that, within control lungs, they are confined to bronchial vasculature surrounding large proximal airways (Fig. 3D). In IPF lungs, COL15A1$^+$ CD31$^+$ VE cells are observed abundantly in areas of bronchiolization and fibrosis (Fig. 3D). Reanalysis of a recently published scRNA-seq dataset that contained normal airway and lung parenchyma samples (*20*) confirmed that genes specific to the pVE were not observed in distal lung VE (Fig. 3E). Collectively, these observations indicate that COL15A1$^+$ VE cells represent an ectopic pVE population in the distal lung in IPF.

The IPF lung exhibits disease-related archetypes among fibroblasts and myofibroblasts

To characterize myofibroblasts and fibroblasts in the IPF lung, we first focused on cell populations characterized by PDGFRB expression and then removed cells characterized as smooth muscle cells (DES, ACTG2, and PLN) or pericytes (RGS5 and COX4I2). This strategy allowed us to identify two stromal populations: fibroblasts, characterized by expression of CD34, FBN1, FBLN2, and VIT and myofibroblasts that consistently express MYLK, NEBL, MYO10, MYO1D, RYR2, and ITGA8, as described in the murine lung (Fig. 4A) (*21*). While these features are consistent across both cell populations in IPF and control lungs (Fig. 4, A and B), the IPF lung contains a myofibroblast phenotype enriched with fibrillar collagens, and ACTA2 (Fig. 4, B to D) and a fibroblast phenotype that exhibits increased expression of HAS1, HAS2, FBN1, and the SHH-induced chemokine, CXCL14 (Fig. 4, B to D) (*22*). Application of lineage reconstruction technique partition-based graph abstraction (PAGA) (*23*) to sub-clustered population of both cell types (Fig. 4B) reveals that the likelihood of connective structures in phenotype space between fibroblast and myofibroblast is relatively weak (edge confidence < 0.2), when compared to connectivity within either fibroblast or myofibroblast subpopulations, irrespective of disease state enrichment (edge confidence between 0.3 and 0.6) (Fig. 4B, bottom). Unsupervised pseudotime ordering of cells within each cell type results in an enrichment of cells originating from IPF lungs at the far end of each manifold (Fig. 4, C to E). Cross-subject IPF enrichment analysis of these archetypes demonstrates a high level of significance in myofibroblasts (Wilcoxon rank-sum, $P = 9.45 \times 10^{-9}$; linear mixed model, $P < 1 \times 10^{-4}$) but only a marginal level of significance among fibroblasts (Wilcoxon rank-sum, $P = 0.0153$; linear mixed model, $P = 0.0645$). Spearman correlations between diffusion pseudotime (DPT) distances and gene expression reveal both gradual and stepwise increases in expression of genes commonly associated with activated myofibroblasts (Fig. 4D) or invasive fibroblasts (Fig. 4E) at the IPF edge of each manifold. Together, these results suggest a continuous trajectory toward IPF archetypes co-occurring within fibroblasts and myofibroblasts, but not necessarily across them.

A profibrotic macrophage archetype overshadows the IPF lung immune landscape

We identify 18 distinct varieties of immune cells (Fig. 5A) represented in samples from multiple subjects across all disease conditions. The proportion of Langerhans dendritic cell and regulatory T cells is elevated in IPF compared to control (2.3- and 1.9-fold change, respectively; Fig. 5B); however, these differences do not reach significance at the multicomparison adjusted *P* value threshold (FDR $P = 0.078$ and 0.054, respectively). We observed IPF-specific macrophages similar to those observed by two previous reports (*3*, *7*). To further elaborate on the features of these cells, we applied archetype analysis in a ternary fashion to classical monocytes, the profibrotic IPF macrophage archetype, and a control-enriched, largely subject-specific, inflammatory macrophage archetype (Fig. 5, C to E). As previously reported (*7*), we observe both a gradual and sequential shift in associated features along the IPF macrophage archetype as it approached its most extreme terminus (Fig. 5, D and E), SPP1 and cholesterol esterases lipoprotein lipase and LIPA expression steadily increasing in expression relatively early along the trajectory, ECM remodeling genes SPARC, GPC4, PALLD, CTSK, and MMP9 ramping up in expression further along the manifold. At the terminus of the IPF archetype, macrophage start expressing CSF1, suggesting the possibility of an autocrine feedback loop for recruitment and activation.

The IPF lung gene regulatory network deviates markedly from controls

To better understand how lung global regulatory networks were altered in IPF, we implemented the gene regulatory network (GRN) inference approach bigSCale (*24*) to control and IPF cells (Fig. 6A). The topology of the IPF GRN exhibited a relatively higher density, as reflected by gene-gene relationships and modularity and as reflected by more within-community connections than cross-community

Downloaded from http://advances.sciencemag.org/ on July 10, 2020

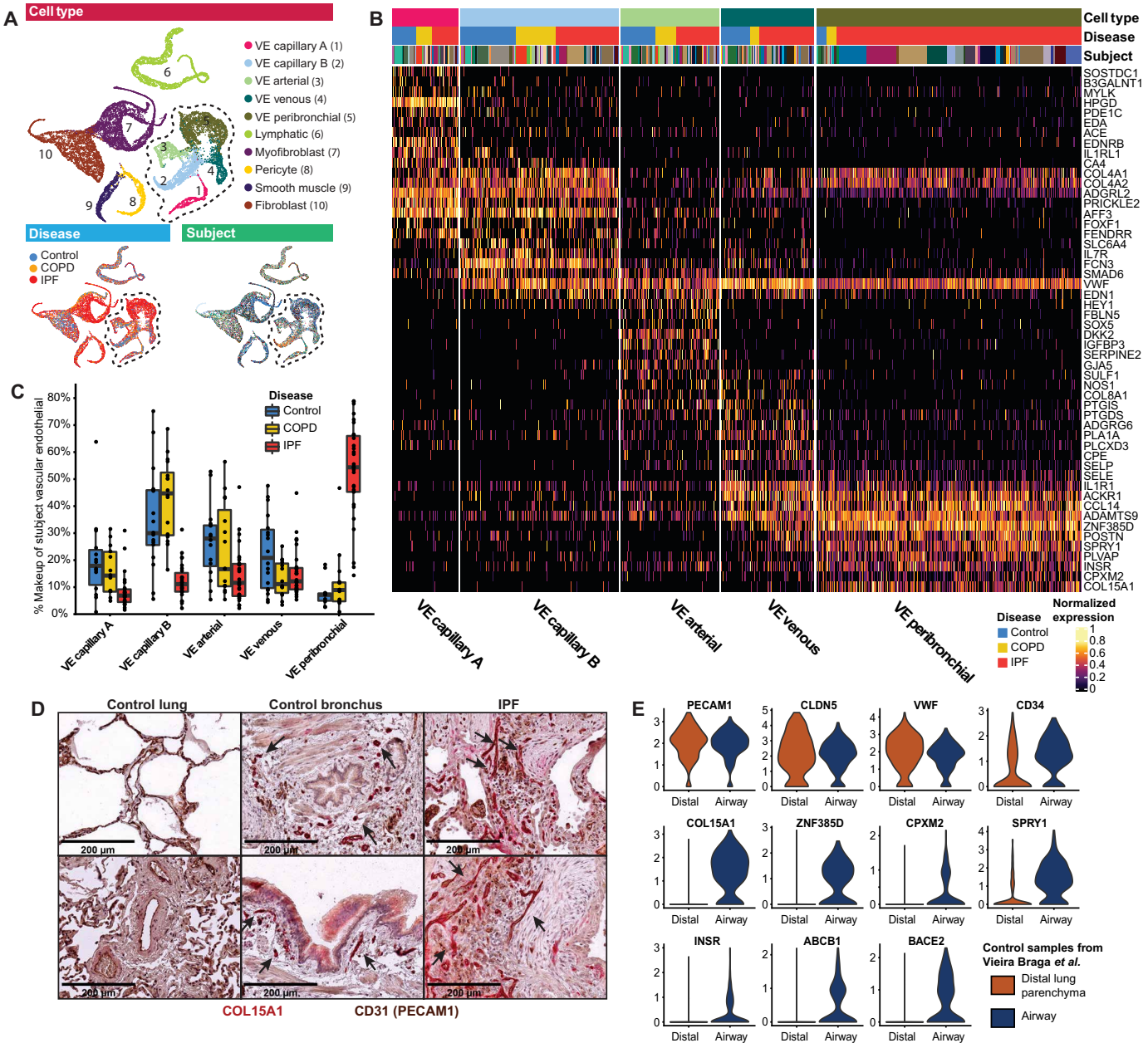

Fig. 3. Identification of disease-enriched VE cell population. (**A**) UMAPs of 14,985 endothelial and mesenchymal cells from 32 IPF, 18 COPD, and 27 control lungs labeled by cell type, disease status, and subject. In the subject plot, each color was represented by a unique color. The dotted line delineates the VE cells from the other stromal and endothelial cell types. (**B**) Heat map representing characteristics of five subtypes of VE cells. Gene expression is unity normalized between 0 and 1 across VE cells. Each column represents an individual cell information regarding subject and disease state, and then VE type is represented in the colored annotation bars above. Each subject is represented by a unique color. (**C**) Boxplots representing the nonzero percent makeup distributions of each VE cell type among all VE cells per subject organized by disease group. Each dot represents a single subject, and whiskers represent 1.5× IQR. FDR-adjusted Wilcoxon rank sum test results comparing IPF and control proportions are reported in data S12. (**D**) IHC staining for CD31 (PECAM1) and COL15A1 in control distal lung, control proximal lung, and affected regions of distal IPF lungs. Arrows indicate vessels with positive COL15A1 staining. (**E**) Violin plots of expression of pan-VE markers and pVE-specific markers across VE cells from distal and airway lung samples from independent dataset (*20*).

connections compared to control GRN (Fig. 6B). Comparing the array of cellular contributions to communities that compose each GRN, we found that control GRN communities show a relatively diverse array of cellular contributions both within communities and across them, consistent with organ function in the nonfibrotic lung (*4*). In IPF GRN, the major epithelial-associated communities remain largely isolated from the rest of the network whose community with the highest density is predominantly driven by aberrant basaloid cells (Fig. 6C). To assess the biological relevance of the differences between the two networks, we investigated the major distinguishing variables delineating their topology. Using PageRank centrality as a proxy for a gene's influence on the network, we identified

Downloaded from http://advances.sciencemag.org/ on July 10, 2020

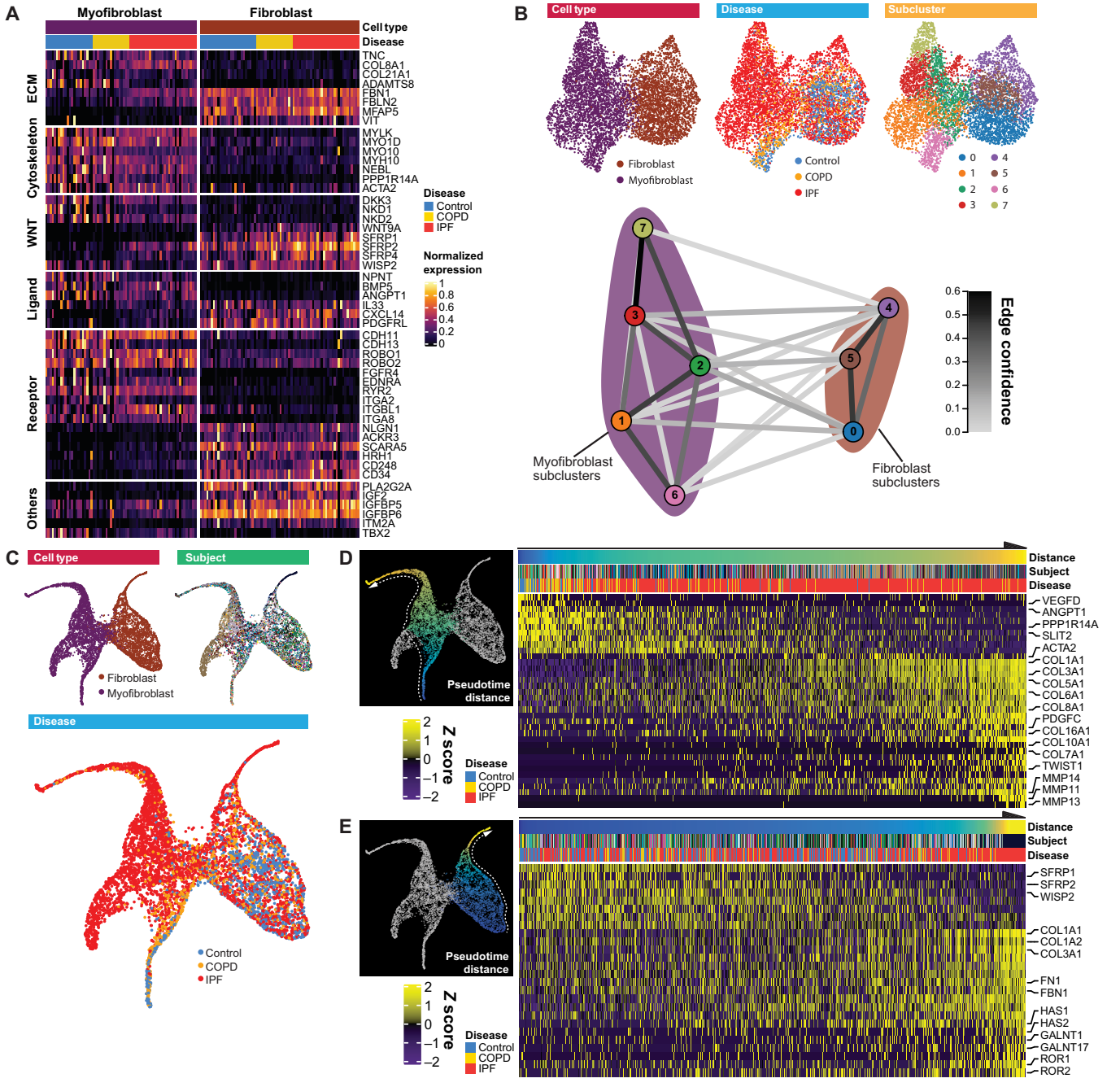

Fig. 4. IPF fibroblast and myofibroblast archetype analysis. (**A**) Heat map of unity-normalized gene expression of curated markers observed to delineate myofibroblast and fibroblast; each column is representative of the average expression value per cell type for one subject. (**B**) Top: UMAPs of 6166 myofibroblast and fibroblast cells from 32 IPF, 18 COPD, and 26 control lungs labeled by cell type, disease, and unsupervised Louvain subclusters. Bottom: Partition graph abstraction (PAGA) analysis. Nodes represent subclusters, and edges represent the probability of internode overlap based on the underlying network of cell neighborhoods. (**C**) UMAPs of myofibroblast and fibroblast cells following diffusion map implementation labeled by cell type, disease status, and subject. In the subject plot, each color represents a unique subject. (**D** and **E**) Heat maps of myofibroblast and fibroblast, respectively, with cells ordered by diffusion pseudotime (DPT) distances along UMAP manifolds representing the continuum of observed cellular phenotypes spanning from control-enriched phenotypes toward IPF-enriched archetypes (left to right). The arrows on each UMAP indicate the orientation of DPT cell ordering, matching the arrows above each heat map's annotation bars. The heat map's annotation bars represent the DPT distance, subject, and disease status for each cell. Expression values are centered and scaled across cells within each cell type.

the top 300 influencer genes ranked by PageRank differential compared to the control GRN (Fig. 6D). Gene set enrichment of these genes returned biological features including cellular aging, senescence, response to TGFB1, epithelial tube formation, and smooth muscle cell differentiation (Fig. 6E) consistent with current mechanistic paradigms. These findings underscore the heterotypic cellular contributions to the many key processes of molecular aberrations in IPF.

Downloaded from http://advances.sciencemag.org/ on July 10, 2020

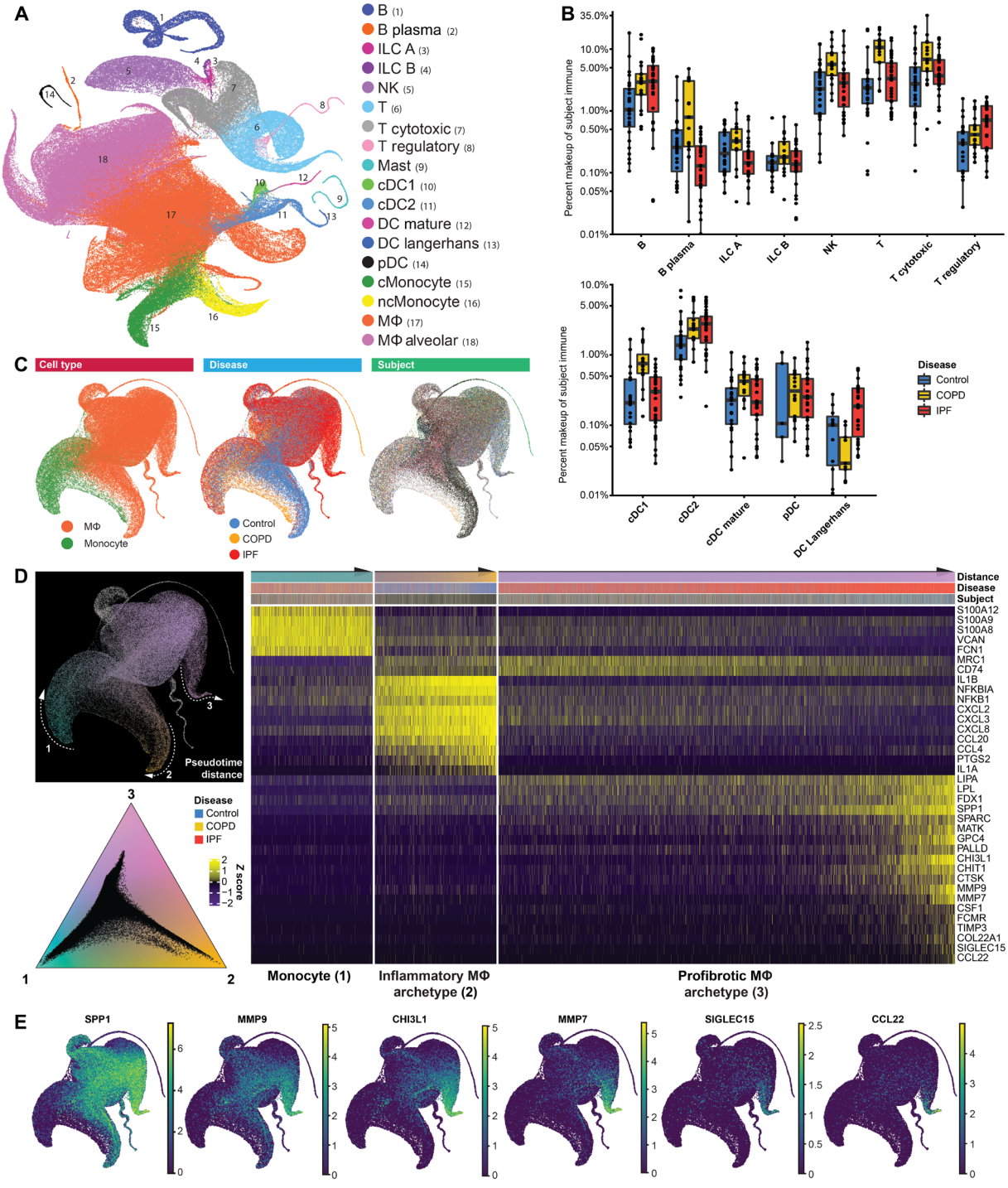

Fig. 5. Immune analysis confirms profibrotic macrophage archetype in IPF. (**A**) UMAP of 271,481 immune cells from 32 IPF, 18 COPD, and 28 control lungs labeled by cell type. (**B**) Boxplots representing the nonzero percent makeup distributions of individual varieties of immune cell as a proportion of all immune cells per subject within each disease group. Each dot represents a single subject, and whiskers represent 1.5 × IQR. FDR-adjusted Wilcoxon rank sum test results comparing IPF and control proportions are reported in data S12. (**C**) UMAPs of 124,470 classical monocytes and monocyte-derived macrophage cells from 32 IPF, 18 COPD, and 28 control cells labeled by cell type, disease, and subject. Each color represents a unique subject. (**D**) Archetype analysis of classical monocytes and two macrophage archetypes. Cells are assigned colors along three gradients with a ternary plot based on each cell's relative DPT distance from three reference points in the UMAP: a monocyte terminus (1, cyan), an inflammatory macrophage archetype (2, yellow), and a profibrotic, IPF-enriched macrophage (3, magenta). Distance color assignments are also projected onto cells for UMAP visualization. UMAP arrows represent the orientation of DPT cell ordering at each terminus and match the arrows above the heat map annotation bar. In the heat map, each column represents a single cell whose respective subject, disease, and DPT color assignment are in the annotation bar above. Macrophage cells belonging to two separate, single-subject–enriched archetypes are removed from analysis. (**E**) UMAPs of monocyte and macrophage cells colored by gene expression values of features that are associated with increasing aberrancy along the IPF archetype manifold.

Downloaded from http://advances.sciencemag.org/ on July 10, 2020

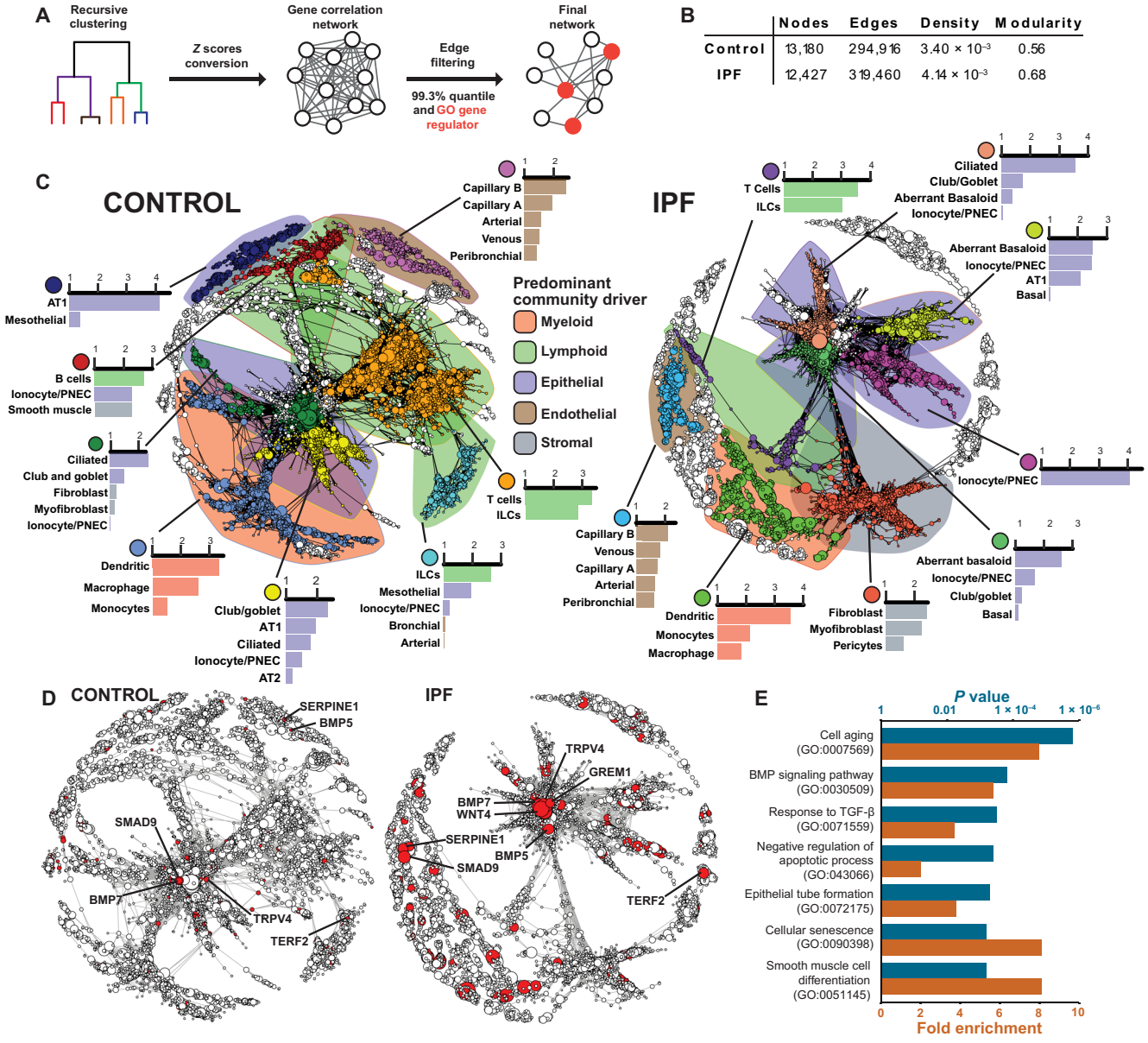

Fig. 6. GRN analysis of IPF and control lungs. (**A**) Overview of bigSCale method for computing **GRN**. (i) Cells are recursively clustered down to subclusters. (ii) *Z* scores are calculated on the basis of differential expression between subclusters (iii) correlations between all genes via Pearson and cosine. (iv) Correlation edges are thresholded using the top 99.3% quantile of correlation coefficients; only edges where at least one node has a Gene Ontology (GO) annotation as a gene regulator. (**B**) Summary of network structure for control and IPF GRNs. (**C**) GRNs of control and IPF lung cells. Nodes represent genes, and edges represent correlations of putative regulatory relationships. Nodes sizes correspond to PageRank centralities, and the largest clusters are assigned colors to their nodes, with each color representing a distinct cluster. The top cell types relevant to each highlighted cluster are shown. Behind each highlighted cluster is a polygon shape covering the domain of the cluster colored by the category of cell type that is predominantly relevant to the community. (**D**) The same GRNs with the top 300 nodes ranked by differential PageRank centrality between IPF and control highlighted in red. Node sizes correspond to PageRank centralities. (**E**) Selected results from GO gene set enrichment of the top 300 differential PageRank nodes between IPF and controls, with all nodes used as a reference. TGF-β, transforming growth factor–β; BMP, bone morphogenetic protein.

DISCUSSION

In this study, we provide a single-cell atlas of the IPF lung with a focus on aberrant epithelial, fibroblast, and endothelial cell populations. Among epithelial cells, we discover a shift in epithelial cell population composition in the distal lung, from alveolar epithelial cells toward enrichment in cell populations that typically reside in the airway. We identify the existence of aberrant basaloid cells, a rare, disease-enriched and previously unidentified epithelial cell population, that coexpress basal epithelial markers, mesenchymal markers, senescence markers, developmental transcription factors, and known markers of IPF and are located at the edge of myofibroblast foci in the IPF lung. Analysis of VE cells reveals an expanded VE population expressing markers usually characteristic of VE cells restricted to the bronchial circulation. This population localizes to areas of remodeling and aberrant angiogenesis in the distal lung parenchyma in IPF is restricted to the airways in the normal lung and represents a previously

Downloaded from http://advances.sciencemag.org/ on July 10, 2020

unknown cellular aberration in IPF. Among stromal cells, we identify two independent fibrotic archetypes associated with stromal populations present in the normal lung: invasive fibroblasts likely related to resident lung fibroblasts and IPF myofibroblasts related to resident normal lung myofibroblasts. The extent of the impact of aberrant cell populations on the phenotype of the IPF lung is highlighted by our GRN analysis that reveals a shift from a balanced multicellular GRN in the homeostatic lung to a fragmented network dominated by the aberrant basaloid cells, the IPF fibroblasts, and myofibroblasts.

The involvement and extent of aberrations of epithelial cells in the human IPF lung have been under substantial scrutiny and debate since the introduction of the epithelial paradigm of IPF (*25*). In our paper, we provide the most detailed and extensive profile of epithelial cells in IPF. This profile allows us to provide cellular detail to "bronchiolization," a term often used by pathologists with regard to the IPF lung, and it is impressive. AT1 and AT2 that constitute a substantial fraction of epithelial cells in the distal normal lung seem to be replaced with cells usually restricted to the proximal lung and to the airways. Of particular interest is the increase in airway basal cells, a population of airway epithelium progenitor cells critically important in lung development and capable of differentiating into any type of airway cells (*9*). We and others have previously found that airway basal cell were abundant in the areas of remodeling in the IPF lung and potentially indicative of a graver outcome (*26*, *27*). Our results support these findings but point to a population of cells that could potentially be confused with airway basal cells if a limited combination of markers is used. These are the aberrant basaloid cells, a rare population of cells that are unique in several ways. First, these cells never appear in control lungs and are detectable in some COPD lungs but are observed in nearly every IPF lung we tested, suggesting that they are a feature of the disease. Second, while they express some of the most typical airway basal cell markers (TP63, KRT17, and LAMB3 LAMC2), they seem to have lost others (KRT5 and KRT15) (*9*) and express SOX9, a transcription factor critical to distal airway development and repair (*16*, *17*). Third, these cells seem to be undergoing EMT, they express mesenchymal markers (COL1A1, CDH2, and HMGA2) and have lost the expression MIR205HG, a gene containing mir-205 and a gatekeeper against EMT (*28*). Fourth, they express multiple markers of senescence (CDKN1A, CDKN2A, MDM2, and GDF15) (*11*, *12*). Last, these cells are the highest expressers of genes implicated in the pathogenesis of IPF such as AVB6 (*14*), MMP7 (*13*), GDF15 (*12*), and EPHB2 (*15*). Anatomically, these cells localize to a highly enigmatic region in the IPF lung at the active edge of the myofibroblast foci as we have demonstrated using a combination of markers including p16 (CDKN2A), p21 (CDKN1A), HMGA2, COX2 (PTGS2), p63(TP63), and KRT17. Thus, the identification of these cells may provide a critical answer to the question of the monolayer of cuboidal epithelial cells found on the surface of IPF myofibroblastic foci. Sometimes referred to as hyperplastic alveolar epithelial cells (*25*), these cells have been a source for controversy because they seemed to have acquired mesenchymal features and lost some epithelial features, suggesting active EMT (*29*), a finding not supported by lineage tracing experiments in animal models of disease (*30*). Our results suggest that myofibroblast foci are lined by a layer of aberrant basaloid cells that exhibit EMT features, are capable of activating TGFB1 locally through their integrin repertoire, and are partially senescent but still maintain progenitor features. Our findings concur with *Chilosi et al.* that described the epithelium of active fibroblastic foci as decidedly bronchiolar (*26*) and subsequently showed WNT signaling targets CCND1 and the IPF biomarker MMP7 associated with atypical $p63^+$ cells in affected regions of the IPF lung (*31*). While it is impossible to fully ascertain the origin of this population, it seems that it is distinct from resident alveolar epithelial cell populations and perhaps derived from a rare progenitor niche with the potential to serve as secondary progenitor for depleted AT1 and AT2 cells in normal human lungs, much like bronchoalveolar stem cells are known to do in the murine lung (*30*, *32*). In IPF, repeated damage and potential genetic predisposition to replicative exhaustion could perhaps lead to the conflicting state of proliferation, differentiation, and senescence apparent in these aberrant basaloid cells. Speculation aside, the discovery that the cells lining myofibroblast foci are fundamentally basaloid in nature, as well as the expansion of regular airway basal cells in the IPF lung serves to couple the two most distinct histopathological features of IPF—fibroblastic foci and honeycomb cysts—to a singular commonality: the manifestation of cells with proximal airway epithelial features in the distal lung.

Our analysis of VE cell populations in the IPF lung has led to a completely unexpected observation. With few exceptions, assessments of vascular remodeling in IPF have primarily focused on pulmonary capillary displacement by nonvascularized patches of fibroblastic lesions with inconclusive observation about potential changes in angiogenesis (*33*, *34*). The absence of discriminating markers for lung VE cells and the difficulty in culturing these cells have limited investigations into vascular remodeling in IPF. We identified an expanded VE cell population that expresses COL15A1 in the IPF lung, which are transcriptomically indistinguishable from systemically supplied bronchial VE cells detected in control lungs. IHC reveals that these cells can readily be found in affected regions in the distal IPF parenchyma. Reanalysis of a publicly available dataset (*20*) revealed that these cells in the normal lung are restricted to the peribronchial vasculature and are never seen in the lung parenchyma. It is possible that our discovery provides the cellular molecular correlate of Turner-Warwick's 1963 observation that the bronchial vascular network is expanded throughout the IPF lung (*35*). While it is impossible at this stage to tell whether the ectopic presence of pVE cells is involved in the pathogenesis of clinical manifestations of IPF, this novel finding fits a larger pattern in the IPF lung described previously, the cellular proximalization of distal lung, and highlights the importance of studying the vascular endothelium in the IPF lung.

Integral to the discussion of the lung phenotype in IPF are the notions of fibroblast activation and myofibroblast accumulation. Unexpectedly, despite notable progress in description of the molecular and metabolic changes required to drive a fibrotic profile of fibroblasts in vitro or in animal models of disease (*36*), little is known about the diversity of cellular populations in the IPF lung. This is because of the lack of specific markers for the distinct populations, as well as the phenotypic changes fibroblasts undergo during the common culturing methods (*37*). While it is well recognized that lung fibroblast populations demonstrate considerable plasticity (*36*–*38*), cell types are traditionally defined by the presence of a singular molecular feature. One relevant example is the use of ACTA2 to define the IPF myofibroblast that comprise the disease's lesions. The cellular source of this pathological cell population remains a matter of considerable debate. scRNA-seq analysis provides a more comprehensive view of cellular identity, where global transcriptomic features, rather than singular cell markers, are exploited

Downloaded from http://advances.sciencemag.org/ on July 10, 2020

to define cells and assess the extent of population variance. Our analysis suggests that pathological, ACTA2-expressing IPF myofibroblasts are not a discrete cell type, but rather one extreme pole of a continuum connected to a quiescent ACTA2-negative stromal population represented in control lungs. This population consistently expresses other genes associated with muscle like MYLK, NEBL, MYO10, and MYO1D, suggesting that the pathological phenotype observed in IPF is an extended, disease-related feature of this population, rather than the result of transdifferentiation from a discrete cellular population. The coexistence of normal and disease-relevant cell populations concomitantly provided us with an opportunity to use computational tools to try and determine what is the source of these populations within the human IPF lung, and we did not observe any evidence suggesting that resident lung fibroblasts or activated fibroblasts were the source for IPF myofibroblasts.

Our study is not without limitations. The dissociation bias intrinsic to all single-cell tissue experiments may lead to spurious changes in cell distribution and limits our ability to provide an exact numeric description of the changes in cell populations across diseased conditions. We addressed that using several strategies. The first and most important was that all samples were handled exactly the same way, obtained from the distal lung, and dissociated following a protocol specifically designed to allow storage of dissociated tissues, enhance viability, and reproducibility (*39*). The use of stored tissues also allowed us to run multiple samples at the same time and minimize batch effects. The second was the use of multiple lines of evidence, including detailed immunohistochemical validation of our findings and reanalysis of different datasets. Thus, in the case of the aberrant basaloid cells and ectopic pVE, we demonstrated their presence in IPF tissues by immunohistochemical analysis and by demonstrating their presence in independent single-cell datasets. We also confirmed the changes in cell subpopulations within epithelial or stromal cell populations by applying deconvolution techniques. Despite these reassuring confirmatory findings, we believe that the changes in cellular proportions reported by us need to be taken as estimates and not as accurate numeric estimations. Another important limitation is again common to studies similar to ours, which is the difficulty in distinguishing transient cellular states from distinct populations. We applied state of the art computational and immunohistochemical approach to ensure that the results we report represent cellular entities that are in existence in the IPF lung, but at this stage, we do not have any way to determine whether those cellular entities are involved in the pathogenesis or clinical manifestations of IPF. However, we believe that the availability of our data will allow follow up studies that will directly assess the role of these cell populations in pulmonary fibrosis. Another important limitation was the use of explanted tissues from patients who underwent lung transplant. Naturally, these tissues represent the end stage of the disease and not early stages. To control for the effects of end stage lung disease, we used explanted lungs from patients with COPD; this allowed us to rule out that the changes we observed were not simply due to the nonspecific effects of the end stage lung. This was extremely helpful in the case of the IPF fibroblast cell populations, the aberrant pVE cells, and the bronchiolization of the epithelial cell repertoire in IPF. It was not helpful in the case of the aberrant basaloid cells, as we could find 33 such cells in COPD lungs. The aberrant basaloid cells were significantly rarer in COPD lungs but could represent common lung pathology to IPF and COPD. COPD and IPF are both associated with age, smoking, accelerated cellular senescence, and a progressive loss of alveolar epithelium. While the pathological manifestations of these diseases are exceedingly different, a common niche of aberrant senescent cells representing a failed repair process is a possibility. Further studies are needed to assess whether metaplastic basaloid cells can be implicated in COPD pathogenesis.

Our results provide a comprehensive portrait of the fibrotic niche in IPF, where aberrant basaloid cells interface with myofibroblast foci, forming a lesion vascularized by ectopic bronchial vessels in the presence of profibrotic monocyte-derived macrophage. The identification and detailed description of aberrant cell populations in the IPF lung may lead to identification of novel, cell type–specific therapies and biomarkers. Last, our IPF cell atlas (*5*) provides an interactive and highly accessible resource to allow the exploration of cell specific changes in gene expression in lung health and disease and thus accelerate discovery and translation.

MATERIALS AND METHODS

Sample preparation for single-cell sequencing

All lung explants from the three conditions (control, COPD, and IPF) were processed following the same protocol. Explanted organs were longitudinally sliced and washed with cold sterile phosphate-buffered saline (PBS). Three to four longitudinal biopsies, which contained tissue representation from apical to most basal segments of the lung, were selected for further processing. Total tissue collected from each explant was on average 150 g. Visible airway structures, vessels, blood clots, and mucin were removed. Lung specimens were minced mechanically into small pieces (<1 mm^3) and then incubated for 45 min in 37°C in modified medium containing Dulbecco's modified Eagle's medium (DMEM)/F12K with added digestion enzymes [elastase (30 U/ml; Elastin Products Company, Owensville, MO), deoxyribonuclease I (0.2 mg/ml; Sigma-Aldrich, St. Louis, MO), liberase (0.3 mg/ml; Roche, Basel, Switzerland), and 1% penicillin/streptomycin]. Digested tissue was filtered using a metal strainer. Unfiltered tissue was incubated a second time in digestion medium for 30 min, followed by repeat filtration. Fetal bovine serum (FBS) (10%) was added to the flow through to stop the enzymatic reaction. Product from filtration was centrifuged at 300*g* at 4°C for 10 min to collect cells dissociated from the tissue. Cell pellets were resuspended in red blood cell lysis buffer (VWR International, Radnor, PA) for 3 min in 37°C and centrifuged again. The pellet was resuspended in DMEM/F12 medium and filtered using a 100-μm strainer. Cells were resuspended in freezing medium (10% FBS and 10% dimethyl sulfoxide in DMEM/F12), aliquoted, and stored in liquid nitrogen.

Samples were thawed in a water bath at 37°C, filtered through a 100-μm cell strainer (Thermo Fisher Scientific, USA), and rinsed with 20 ml of cold (4°C) PBS and 10% heat-inactivated FBS (both Life Technologies, USA). Cell suspensions were centrifuged at 300*g* for 5 min at 4°C. Supernatant was discarded, and cells were resuspended in 200 μl of dead cell removal microbead solution and incubated at 4°C for 15 min. Two milliliter 1× binding buffer was added, and cell suspensions were passed through a 70-μm cell strainer (Thermo Fisher Scientific, USA) and loaded with 500 μl of 1× binding buffer onto a prewashed MS column (Dead Cell Removal Kit and MS columns, Miltenyi Biotec, USA). Cell suspensions passed through the MS columns, and then the columns were rinsed with 2 ml of 1× binding buffer. Cell suspensions were centrifuged at 300*g* for 5 min at 4°C, the supernatant was discarded, and the cells were resuspended in 1 ml of PBS and 0.04 % BSA (New England Biolabs, USA). Cell suspensions

Downloaded from http://advances.sciencemag.org/ on July 10, 2020

were passed through a final 40-μm cell strainer (Thermo Fisher Scientific, USA). For cell concentrations and viability, cells were stained with Trypan blue and then counted on a Countess Automated Cell Counter (Thermo Fisher Scientific, USA). The average recorded cell viability was 80.86 ± 8.54%.

Single-cell barcoding, library preparation, and sequencing

Single cells were barcoded using the 10× chromium single-cell platform, and complementary DNA (cDNA) libraries were prepared according to the manufacturer's protocol (Single Cell 3′ Reagent Kits v2, 10× Genomics, USA). Cell suspensions, reverse transcription master mix, and partitioning oil were loaded on a single-cell "A" chip with a targeted cell output of 10,000 cells per library, assuming 100% viability, and then run on the Chromium Controller. Reverse transcription was performed within the droplets at 53°C for 45 min. cDNA was amplified for 12 cycles total on a Bio-Rad C1000 Touch thermocycler. cDNA size selection was performed using SPRIselect beads (Beckman Coulter, USA) and a ratio of SPRIselect reagent volume to a sample volume of 0.6. cDNA was analyzed on an Agilent Bioanalyzer High Sensitivity DNA chip for qualitative control purposes. cDNA was fragmented using the proprietary fragmentation enzyme blend for 5 min at 32°C, followed by end-repair and A-tailing at 65°C for 30 min. cDNA were double-sided size selected using SPRIselect beads. Sequencing adaptors were ligated to the cDNA at 20°C for 15 min. cDNA was amplified using a sample-specific index oligo as a primer, followed by another round of double-sided size selection using SPRIselect beads. Final libraries were analyzed on an Agilent Bioanalyzer High Sensitivity DNA chip for qualitative control purposes. cDNA libraries were sequenced on a HiSeq 4000 Illumina platform aiming for 150 million reads per library and a sequencing configuration of 26 base pair (bp) on read1 and 98 bp on read2.

Fastq generation and read trimming

Basecalls were converted to reads with the software Cell Ranger's (v2.2) implementation mkfastq. Multiple fastq files from the same library and strand were catenated to single files. Read2 files were subject to two passes of contaminant trimming with cutadapt (v1.17): (i) for the template switch oligo sequence (AAGCAGTGGTATCAACGCAGAGTACATGGG) anchored on the 5′ end and (ii) for poly(A) sequences on the 3′ end. Following trimming, read pairs were removed if the read2 was trimmed below 30 bp.

Cell barcode and unique molecular identifier demultiplexing, alignment, and transcript counting

Subsequent read processing was conducted with the zUMIs pipeline (v2.0). Paired reads were filtered if either the cell barcode or unique molecular identifier (UMI) sequence had more than 1 bp with a phred of <20. Reads were aligned with STAR (v2.6.0c) to the human genome reference GRCh38 release 91 from ensemble. Collapsed UMIs with reads that span both exonic and intronic sequences were retained as both separate and combined gene expression assays.

Filtering valid cell barcodes and quality control

Cell barcodes representative of quality cells were delineated from barcodes of apoptotic cells or background RNA based on a threshold of having at least 12% of transcripts arising from intron spanning, or unspliced reads, indicative of nascent mRNA (fig. S8). Cells with less than 1000 transcripts profiled or more than 20% of their transcriptome of mitochondrial origin were then removed. The average number of cells returned per library is 2925 (±1811.852). Sequencing depth comparisons across samples categorized by associated disease state were statistically assessed for disease-associated biases via Kruskal-Wallis test, where differences were observed to be statistically insignificant (*P* value of 0.29; fig. S9A). A detailed summary of the average UMI and gene per cell type per library distribution per disease state is presented in fig. S9 (B and C, respectively).

Gene name conversion

Genes were originally output in ensemble gene ID format. To improve the interpretability of the variables without making compromises to sensitivity, we converted ensemble gene IDs to Hugo Gene Nomenclature Committee (HGNC) format using the R package BioMart only when an exact one-to-one translation was available.

Data normalization and cell population identification

UMIs from each cell barcode, irrespective of intron or exon coverage, were retained for all downstream analysis. 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. Aggregated data were subject to Louvain cluster analysis for cell type identification using the R package Seurat (version 2.3.1). Recursive clustering analysis of subpopulations of pure immune ($PTPRC^+$), epithelial and the remaining mesenchymal populations were conducted to improve the granularity of our cell annotations.

Multiplet cell populations were clusters, which were identified as having a transcriptomic signature that resembled the resulting combination of two or more disparate cell type signatures that already existed in the dataset. For each cell population flagged as potential multiplets, a differential expression test was performed comparing the population against their suspected combination of cell populations; a confident assessment of multiplet status was made when there were no genes uniquely expressed in the suspected population and no genes represented in the cell combination group that could not be detected in the suspected multiplet population. The average number of multiplets identified per library is 49.33% (±63.8) or 1.29% (±0.95%) of all cells identified in each library. Cell barcodes identified as multiplets were not included in downstream analyses.

Generation of cell type markers

Cell type marker lists were generated with two separate approaches. In the first approach, we sought to adjust for unbalanced representation of different cell types across subjects and reduce the zero inflation in the data to generate a more reliable *P* value from a differential expression test. We collapsed the gene expression matrix values to one where each column represents average gene expression for a given subject and for a given cell type. To highlight gene expression among relatively similar cell types, we then grouped each cell type into one of four groups: epithelial, stromal/mesenchymal, myeloid, and lymphoid. We then applied the Seurat FindAllMarkers implementation of the Wilcoxon rank sum test for each of these four subsets of cell types to generate separate marker lists for each (data S2 to S5).

The second approach to generating cell type markers uses a binary classifier system to assess the utility of detecting a given gene, irrespective of its intensity of expression, for classifying a cell. For each cell type in the data, we identified the genes whose expression

Downloaded from http://advances.sciencemag.org/ on July 10, 2020

was log fold change ≥ 0.3 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

$$\mathrm{DOR} = \frac{(TruePositives + 0.5)/(FalsePositives + 0.5)}{(FalseNegatives + 0.5)/(TrueNegatives + 0.5)}$$

where TruePositives 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. The log transformed DOR marker values are contained in data S6.

Cell population composition comparisons

All figures and statistical tests for proportion makeups were performed exclusively with subjects who had nonzero proportion makeups of the given cell type. This choice was made in an effort to be conservative in our estimates, as we typically considered the lack of detection of a given cell type as a technical limitation. Test results reported in data S12 are Wilcoxon rank sum tests of nonzero proportions between IPF and control subjects. We advise caution in the interpretation of these statistics as the model does not account for unbalanced sampling rates of cells across subjects, and the proportion values of any given cell type are not independent of the proportions for any other cell types that belong to the same category.

Differential expression between disease conditions

A major limitation that persists in scRNA-seq analysis is the lack of a differential expression model that can account for cellular population differences while ultimately testing differences at the level of the subject, rather than treating each individual cell as an independent sample. While we were unable to design a model capable of handling the complex sample hierarchy scRNA-seq, we nonetheless applied a crude method of differential expression.

For each subject and each cell type, the gene expression for each gene was averaged to create a single "sample" representative of an individual. We then applied a Wilcoxon rank sum test between within each cell type and average values from subjects of each disease condition for a total of three comparisons (IPF versus control, COPD versus control, and IPF versus COPD; data S8, S9, and S10, respectively). Cell types with less than five subjects representative of a particular disease condition were not included its analysis.

UMAP visualizations

For similarity-based cell network analysis and visualization, we used tools from the Python (version 3.6.8) library Scanpy (version 1.3.7). Uniform Manifold Approximation and Projection (UMAP) figures for all subsets of cells were generated using the same sequence of implementations. Feature selection of the top genes ranked by dispersion (scaled variance/mean) across 20 bins of the expression distribution are identified. The expression values for these genes are then adjusted for differences in total UMI and the fraction of mitochondrial reads across cells during *z* normalization with a maximum absolute *z* score of 10. The scaled values are then subject to principal components analysis (PCA) for linear dimension reduction.

The top principal components are subject to exploratory analysis to identify contributions to variance at the level of library batch, subject, cell type, and disease. In addition, the residuals of each principal component were explored to ascertain functional relevance of the signatures. Feature selection of principal components was conducted on the basis of these analyses, and a shared nearest neighbor network was then created on the basis of Euclidean distances between cells in multidimensional PC space and a fixed number of neighbors per cell, which was used to generate an intermediate two-dimensional (2D) UMAP.

The resulting network was then subject to PAGA using the cell type categories as abstraction nodes. A confidence threshold of cell type interconnectivity was implemented to avoid spurious manifold connections between cell types of disparate lineages. Diffusion maps is then applied for nonlinear dimension reduction on a more limited subset of principal component dimensions, and a new neighborhood graph is then computed on the basis of representations of diffusion map distances. A final 2D UMAP is then generated using the new neighborhood graph distances, initialized by the positions from the thresholded PAGA.

The values for all nondefault parameters used during the generation of each UMAP are represented in data S11; if a parameter is unspecified, then the default Scanpy implementation parameter was used.

Evaluation for the potential influence of outlying subject-specific variation

To ensure that our single-cell analysis was not unduly driven by the outlying characteristics of one or few patient transcriptional profiles, we ran a parallel analysis using deep generative modeling to "correct" for any potentially aberrant subject-specific variational signatures in accordance with the single-cell variational inference (scVI) approach described by Lopez *et al.* (*40*). This was implemented through a dedicated python library downloaded from GitHub (https://github.com/YosefLab/scVI). We found that, even after applying this "batch effect" normalization, our cluster groups were preserved with high fidelity as evaluated by multiple metrics (fig. S1), reinforcing that outlying subject-specific effects were not driving the results of our analysis.

Evaluation for the potential influence of cell cycle state

To further demonstrate the robustness of our approach, we applied a similar method to evaluate for possible data artifacts specific to cell cycle state. In another parallel analysis, we used the approach described by Scialdone *et al.* (*41*) to extract G_1, G_2-M, and S components from each individual cell before normalization using a set of 94 known cycle-associated genes. We then regressed out for these elements using Scanpy, similar to our approach for mitochondrial RNA, and proceeded with downstream analysis. We found that the G_1, G_2-M, and S states were well distributed at both a cluster level and patient level. Furthermore, regressing out the effects of cell cycle state did not meaningfully alter our resulting cluster configurations (fig S4).

Comparison of manual cell annotations with automated methods

We compared our manual annotations to those produced through automated classification using SingleR. Strong correlation with manual

Downloaded from http://advances.sciencemag.org/ on July 10, 2020

annotations was found using both the Human Primary Cell Atlas (HPCA) and ENCODE human reference sets across our dataset comparable to the correlation between the HPCA and ENCODE annotations themselves for these cells (fig. S3).

PAGA connectivity analysis of fibroblast and myofibroblast

To assess the most likely trajectories of cell progression toward IPF-enriched fibrosis among fibroblast and myofibroblast, we used unsupervised Louvain clustering to generate eight subpopulations, which were then subjected to PAGA analysis to ascertain the most likely intersubcluster trajectories. The edge confidences between each subcluster node for all edges is visualized using the R package igraph (v1.2.4.1).

Scoring of regulon activity

A regulon is a group of target genes regulated by a common transcription factor. To score the activity of each regulon in each nonimmune cell, we used the package pySCENIC 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) were downloaded from resources.aertslab.org/cistarget/, and the list of human transcription factors (hs_hgnc_tfs.txt) was downloaded from github.com/aertslab/pySCENIC/tree/master/resources.

Archetype analysis of fibroblast and myofibroblast

We observed that many IPF-enriched features in the data were represented by a continuum of increasing phenotypic deviation from controls, rather than discrete features readily amenable to delineation (e.g., cluster analysis) from control-enriched signals. Consequently, we sought to implement archetype analysis of these continua to assess disease-enriched features rather than relying on traditional group-wise comparisons.

We first assessed the most likely trajectories toward fibrotic archetypes among fibroblast and myofibroblast using the DPT implementation from Scanpy to plot the distances along the UMAP manifold toward each archetype's terminus. The same diffusion map component structure used to generate the 2D UMAP visualizations was used for calculating DPT distances. For fibroblast, we took the cell with the highest expression of ITGB1, which lied at the terminus of fibroblast manifold's tendril as the root cell and calculated the relative distances of all other fibroblast cells. We used 1-DPT value for the final distance value ordering and assigned a gradient of colors to cells for the range of 0 to 0.7 to represent each cells distance on the UMAP legend and in the accompanying heat map.

For myofibroblast, we calculated DPT ordering from all three termini in the myofibroblast manifold using the highest MMP11-expressing cell to represent the IPF-enriched terminus, the furthest control cell to represent the control, and COPD-enriched terminus and the furthest cell from 225I to represent the single-subject–enriched terminus. To mitigate the impact on our analysis from the single-subject–enriched archetype, we removed the nearest 500 cells to the subject "225I-" enriched terminus that did not overlap with the nearest 399 cells to the control and COPD-enriched myofibroblast archetype. Among the remaining myofibroblast cells, we used the difference between both 1-DPT distance values from the remaining two nonsubject-specific archetypes as a singular distance vector. A color gradient for the final 1-DPT range of −1 to 1 was then applied to each unfiltered cell in the UMAP legend and accompanying heat map.

For both archetype heat maps, cells were plotted in order of the final 1-DPT values from lowest to highest. A spearman correlation test was conducted between each cell's 1-DPT value and gene expression to ascertain which genes increased or decreased in expression along the manifold and would thus serve as candidate features in the heat map. In addition to each cell's DPT distance value associated color, each cell's respective disease and subject color identity were included in the annotation bar to represent the extent of disease enrichment and subject-level contribution to the feature.

Pseudotime orderings were each subject to statistical assessment to determine whether cells from IPF subjects were enriched at the far end of the DPT ordering results (the rightmost portion of Fig. 4, D and E, heat maps). The first approach was a Wilcoxon rank sum test (performed with the base R "wilcox.test" implementation) between IPF and control for the average DPT distance per subject, with the alternative hypothesis that IPF DPT distribution was right skewed. The second approach was a generalized linear mixed model (performed with the R package nlme v3.1's implementation) applied to the DPT distance of the cells, where disease was a fixed variable and subject was treated as a random variable nested within each disease.

Archetype analysis of classical monocytes and macrophage

Among classical monocytes and macrophage, we identified one monocyte archetype connected to distinct four macrophage archetypes: an inflammatory archetype, an IPF-enriched archetype, an outlier mitochondrial transfer RNA (MT-tRNA)–enriched archetype driven by two subjects, and another outlier archetype driven by two separate subjects characterized by heat shock protein expression. 1-DPT distance values for all monocytes and macrophage were calculated from each archetype terminus as described for fibroblast and myofibroblast. We removed 1700 and 5800 most proximal cells to the MT-tRNA and heat shock protein archetypes, respectively, to avoid contributions from outlier signals.

Following outlier removal, the 1-DPT distances values from the monocyte, inflamed macrophage and IPF-enriched macrophage were each independently unity normalized to values between 0 and 1. Distances along the three normalized trajectories were then used plot the cells in a ternary plot using the R gglot2 extension package tricolore (v1.2.0) to assign a color to each cell based on its relative proximity to each of the three archetype termini, which is visualized in both the UMAP color legend and ternary plot legend itself. The 15,000, 20,000, and 25,000, closest cells to the monocyte, inflamed macrophage and IPF-enriched macrophage archetype, respectively, were then selected for correlation analysis between 1-DPT values from each respective trajectory and gene expression to assess candidate genes for plotting the heat map.

GRN construction

GRNs for both control and IPF cell populations were generated using the R package bigSCale2 (*24*). Control and IPF cells were split, and for each disease state, cells were randomly down-sampled using Seurat's SubsetData implementation (seed = 7) to a maximum of 500 cells per cell type, resulting in 10,560 cells from control and 15,388 cells from IPF. The resulting matrices were then filtered to remove genes with ensembl identifiers and passed to bigSCale2, where both networks were constructed under the "normal" clustering parameter,

Downloaded from http://advances.sciencemag.org/ on July 10, 2020

with an edge cutoff of the top 0.993 quantile for correlation coefficient. Networks were visualized with the R package igraph, and each network's layout is derived from 10,000 iterations of the Fruchterman-Reingold algorithm and "nogrid" parameter (seed = 7).

Assigning cell relevance scores to GRN communities

Each GRN was clustered with igraph's "cluster_fast_greedy" function (seed = 7). Similar cell types were collapsed into the groups to avoid excessive phenotypic overlap: all dendritic cells, both macrophage varieties, both monocyte varieties, all T cells, natural killer and innate lymphocytes, both B cell varieties, both secretory, and pulmonary neuroendocrine cell and ionocytes. For each cell type group, we then took the top 500 genes ranked by log(DOR) markers as representative features. The relevance score for each cell type group in a network's community is a z score of the cumulative log(DOR) markers that intersect with members of each GRN community, weighted by the gene's normalized PageRank centrality within its community as follows

$$PR^{*}_{ij} = \frac{PR_{ij} - PR_{min.j}}{PR_{max.j} - PR_{min.j}}$$

$$CumulativeScore_{kj} = \sum_{i \in kj} \log(DOR)_{ik} \times PR^{*}_{ij}$$

where PR denotes the PageRank centrality for a node in the network and DOR is the diagnostic odds ratio classifier calculated earlier in the cell type marker list. Given i genes, j communities, and k cell type groups, let PR^{*}_{ij} denote the normalized PageRank value for each gene i ($i = 1, \ldots, I$) within its respective network community j ($j = 1, \ldots, J$), let $\log(DOR)_{ik}$ be the natural log transformed DOR for gene i that belongs to cell type group k ($k = 1, \ldots, K$), let $i \in kj$ represent the intersection of genes from cell type group k and GRN community j, and let $CumulativeScore_{kj}$ represents the cumulative relevance score for cell type group k in GRN community j. Last, the cumulative scores are then converted to z scores by centering and scaling across cell type groups within each community, only cell type groups with a relevance score greater than or equal to 1 were retained for annotation of the community. The aberrant basaloid cell type community was excluded from the control GRN analysis because this population of cells was not detected in any control samples.

Immunohistochemistry

The FFPE (Formalin fixed paraffin embedded) blocks were cut at 5 μm and then rehydrated using standard xylene/ethanol deparaffinization. For heat-induced antigen retrieval, specimens were boiled at 95°C for 20 min in 1× tris-based Antigen Unmasking Solution (Vector Laboratories, USA). Tissue slides were incubated for 10 min in BLOXALL Endogenous Peroxidase and Alkaline Phosphatase Blocking Solution (Vector Laboratories, USA) to block endogenous peroxidase and alkaline phosphatase activity. Tissue slides were blocked using 2.5% Normal Horse Serum Blocking Solution (Vector Laboratories, USA) for 20 min, then incubated with the primary antibody (data S13), and diluted in 2.5% Normal Horse Serum Blocking Solution for 30 min at room temperature. Specimen were incubated for 30 min with secondary antibodies (anti-mouse or anti-rabbit ImmPRESS reagent, conjugated with horseradish peroxidase or alkaline phosphatase, as appropriate; all Vector Laboratories, USA). Slides were incubated for 10 min in DAB working solution or for 30 min in Vector Reds working solution, as appropriate (both Vector Laboratories, USA). For sequential double stainings, this protocol was repeated from the protein-blocking step on using the alternative enzyme and substrate reagents. Tissue slides were counterstained in Hematoxilin Solution Gill no. 1 (Sigma-Aldrich, USA) for 3 min and then washed with tap water. For permanent mounting, specimens were dehydrated in ethanol/xylene and then mounted with VectaMount permanent mounting solution (Vector Laboratories, USA). Stained slides were digitalized on a Aperio Scanner (Leica) and then analyzed using the softwares QuPath and ImageJ.

Independent validation of COL15A1 protein expression in pVE cells

Images of COL15A1 immunohistochemical stainings of lung parenchyma and bronchi specimens (fig. S7) were downloaded from the Human Protein Atlas [(*19*); www.proteinatlas.org/ENSG00000204291-COL15A1/tissue). All specimens had been stained using the polyclonal antibody HPA017915 (Sigma-Aldrich).

Deconvolution of publicly available bulk RNA-seq data

Twenty of the 38 cell varieties identified were used on the basis of their anticipated abundance in the lung, relevance to IPF, and low propensity for colinearity spillover problems. A maximum of 250 cells per subject randomly downsampled in R (seed 7) from each of these cell types across IPF and control samples. Cell signature markers were then curated on the basis of their cell specificity and consistency in expression across control and IPF samples alike. Data from GEO accession number GSE134692, representing bulk RNA-seq data from 26 control lung homogenates and 46 transplant stage IPF lungs (*8*), were used. Deconvolution was performed with the R package MuSiC (*42*, *43*).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/6/28/eaba1983/DC1

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. D. J. Lederer, F. J. Martinez, Idiopathic pulmonary fibrosis. *N. Engl. J. Med.* **378**, 1811–1823 (2018).
2. G. Y. Yu, G. H. Ibarra, N. Kaminski, Fibrosis: Lessons from OMICS analyses of the human lung. *Matrix Biol.* **68-69**, 422–434 (2018).
3. P. A. Reyfman, J. M. Walter, N. Joshi, K. R. Anekalla, A. C. McQuattie-Pimentel, S. Chiu, R. Fernandez, M. Akbarpour, C.-I. Chen, Z. Ren, R. Verma, H. Abdala-Valencia, K. Nam, M. Chi, S. H. Han, F. J. Gonzalez-Gonzalez, S. Soberanes, S. Watanabe, K. J. N. Williams, A. S. Flozak, T. T. Nicholson, V. K. Morgan, D. R. Winter, M. Hinchcliff, C. L. Hrusch, R. D. Guzy, C. A. Bonham, A. I. Sperling, R. Bag, R. B. Hamanaka, G. M. Mutlu, A. V. Yeldandi, S. A. Marshall, A. Shilatifard, L. A. N. Amaral, H. Perlman, J. I. Sznajder, A. C. Argento, C. T. Gillespie, J. Dematte, M. Jain, B. D. Singer, K. M. Ridge, A. P. Lam, A. Bharat, S. M. Bhorade, C. J. Gottardi, G. R. S. Budinger, A. V. Misharin, Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* **199**, 1517–1536 (2019).
4. M. S. B. Raredon, T. S. Adams, Y. Suhail, J. C. Schupp, S. Poli, N. Neumark, K. L. Leiby, A. M. Greaney, Y. Yuan, C. Horien, G. Linderman, A. J. Engler, D. J. Boffa, Y. Kluger, I. O. Rosas, A. Levchenko, N. Kaminski, L. E. Niklason, Single-cell connectomic analysis of adult mammalian lungs. *Sci. Adv.* **5**, eaaw3851 (2019).
5. N. Neumark *et al.*, in *IPF Cell Atlas* (2019); www.ipfcellatlas.com.
6. Y. Xu, T. Mizuno, A. Sridharan, Y. du, M. Guo, J. Tang, K. A. Wikenheiser-Brokamp, A. K. T. Perl, V. A. Funari, J. J. Gokey, B. R. Stripp, J. A. Whitsett, Single-cell RNA sequencing identifies diverse roles of epithelial cells in idiopathic pulmonary fibrosis. *Jci Insight* **1**, e90558 (2016).
7. C. Morse, T. Tabib, J. Sembrat, K. L. Buschur, H. T. Bittar, E. Valenzi, Y. Jiang, D. J. Kass, K. Gibson, W. Chen, A. Mora, P. V. Benos, M. Rojas, R. Lafyatis, Proliferating

Downloaded from http://advances.sciencemag.org/ on July 10, 2020

SPP1/MERTK-expressing macrophages in idiopathic pulmonary fibrosis. *Eur. Respir. J.* **54**, 1802441 (2019).
8. P. Sivakumar, J. R. Thompson, R. Ammar, M. Porteous, C. McCoubrey, E. Cantu III, K. Ravi, Y. Zhang, Y. Luo, D. Streltsov, M. F. Beers, G. Jarai, J. D. Christie, RNA sequencing of transplant-stage idiopathic pulmonary fibrosis lung reveals unique pathway regulation. *ERJ Open Res.* **5**, 00117–02019 (2019).
9. J. Ordovas-Montanes, D. F. Dwyer, S. K. Nyquist, K. M. Buchheit, M. Vukovic, C. Deb, M. H. Wadsworth II, T. K. Hughes, S. W. Kazer, E. Yoshimoto, K. N. Cahill, N. Bhattacharyya, H. R. Katz, B. Berger, T. M. Laidlaw, J. A. Boyce, N. A. Barrett, A. K. Shalek, Allergic inflammatory memory in human respiratory epithelial progenitor cells. *Nature* **560**, 649–654 (2018).
10. M. Zeisberg, E. G. Neilson, Biomarkers for epithelial-mesenchymal transitions. *J. Clin. Invest.* **119**, 1429–1437 (2009).
11. P. J. Barnes, J. Baker, L. E. Donnelly, Cellular senescence as a mechanism and target in chronic lung diseases. *Am. J. Respir. Crit. Care Med.* **200**, 556–564 (2019).
12. Y. Z. Zhang, M. Jiang, M. Nouraie, M. G. Roth, T. Tabib, S. Winters, X. Chen, J. Sembrat, Y. Chu, N. Cardenes, R. M. Tuder, E. L. Herzog, C. Ryu, M. Rojas, R. Lafyatis, K. F. Gibson, J. F. M. Dyer, D. J. Kass, J. K. Alder, GDF15 is an epithelial-derived biomarker of idiopathic pulmonary fibrosis. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **317**, L510–L521 (2019).
13. F. R. Zuo, N. Kaminski, E. Eugui, J. Allard, Z. Yakhini, A. Ben-Dor, L. Lollini, D. Morris, Y. Kim, B. De Lustro, D. Sheppard, A. Pardo, M. Selman, R. A. Heller, Gene expression analysis reveals matrilysin as a key regulator of pulmonary fibrosis in mice and humans. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 6292–6297 (2002).
14. J. S. Munger, X. Huang, H. Kawakatsu, M. J. D. Griffiths, S. L. Dalton, J. Wu, J. F. Pittet, N. Kaminski, C. Garat, M. A. Matthay, D. B. Rifkin, D. Sheppard, The integrin αvβ 6 binds and activates latent TGF β1: A mechanism for regulating pulmonary inflammation and fibrosis. *Cell* **96**, 319–328 (1999).
15. D. Lagares, P. Ghassemi-Kakroodi, C. Tremblay, A. Santos, C. K. Probst, A. Franklin, D. M. Santos, P. Grasberger, N. Ahluwalia, S. B. Montesi, B. S. Shea, K. E. Black, R. Knipe, M. Blati, M. Baron, B. Wu, H. Fahmi, R. Gandhi, A. Pardo, M. Selman, J. Wu, J.-P. Pelletier, J. Martel-Pelletier, A. M. Tager, M. Kapoor, ADAM10-mediated ephrin-B2 shedding promotes myofibroblast activation and organ fibrosis. *Nat. Med.* **23**, 1405–1415 (2017).
16. S. Danopoulos, I. Alonso, M. E. Thornton, B. H. Grubbs, S. Bellusci, D. Warburton, D. Al Alam, Human lung branching morphogenesis is orchestrated by the spatiotemporal distribution of ACTA2, SOX2, and SOX9. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **314**, L144–L149 (2018).
17. T. Volckaert, T. Yuan, J. Yuan, E. Boateng, S. Hopkins, J. S. Zhang, V. J. Thannickal, R. Fässler, S. P. de Langhe, Hippo signaling promotes lung epithelial lineage commitment by curbing Fgf10 and β-catenin signaling. *Development* **146**, dev166454 (2019).
18. M. Ebina, Pathognomonic remodeling of blood and lymphatic capillaries in idiopathic pulmonary fibrosis. *Respir. Investig.* **55**, 2–9 (2017).
19. M. Uhlen, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. M. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, F. Pontén, Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
20. F. A. Vieira Braga, G. Kar, M. Berg, O. A. Carpaij, K. Polanski, L. M. Simon, S. Brouwer, T. Gomes, L. Hesse, J. Jiang, E. S. Fasouli, M. Efremova, R. Vento-Tormo, C. Talavera-López, M. R. Jonker, K. Affleck, S. Palit, P. M. Strzelecka, H. V. Firth, K. T. Mahbubani, A. Cvejic, K. B. Meyer, K. Saeb-Parsy, M. Luinge, C. A. Brandsma, W. Timens, I. Angelidis, M. Strunz, G. H. Koppelman, A. J. van Oosterhout, H. B. Schiller, F. J. Theis, M. van den Berge, M. C. Nawijn, S. A. Teichmann, A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat. Med.* **25**, 1153–1163 (2019).
21. J. Green, M. Endale, H. Auer, A. K. T. Perl, Diversity of interstitial lung fibroblasts is regulated by platelet-derived growth factor receptor α kinase activity. *Am. J. Respir. Cell Mol. Biol.* **54**, 532–545 (2016).
22. G. Jia, S. Chandriani, A. R. Abbas, D. J. DePianto, E. N. N'Diaye, M. B. Yaylaoglu, H. M. Moore, I. Peng, J. DeVoss, H. R. Collard, P. J. Wolters, J. G. Egen, J. R. Arron, CXCL14 is a candidate biomarker for Hedgehog signalling in idiopathic pulmonary fibrosis. *Thorax* **72**, 780–787 (2017).
23. F. A. Wolf, F. K. Hamey, M. Plass, J. Solana, J. S. Dahlin, B. Göttgens, N. Rajewsky, L. Simon, F. J. Theis, PAGA: Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).
24. G. Iacono, R. Massoni-Badosa, H. Heyn, Single-cell transcriptomics unveils gene regulatory network plasticity. *Genome Biol.* **20**, 110 (2019).
25. M. Selman, T. E. King, A. Pardo; American Thoracic Society; European Respiratory Society; American College of Chest Physicians, Idiopathic pulmonary fibrosis: Prevailing and evolving hypotheses about its pathogenesis and implications for therapy. *Ann. Intern. Med.* **134**, 136–151 (2001).
26. M. Chilosi, V. Poletti, B. Murer, M. Lestani, A. Cancellieri, L. Montagna, P. Piccoli, G. Cangi, G. Semenzato, C. Doglioni, Abnormal re-epithelialization and lung remodeling in idiopathic pulmonary fibrosis: The role of ΔN-p63. *Lab. Invest.* **82**, 1335–1345 (2002).
27. A. Prasse, H. Binder, J. C. Schupp, G. Kayser, E. Bargagli, B. Jaeger, M. Hess, S. Rittinghausen, L. Vuga, H. Lynn, S. Violette, B. Jung, K. Quast, B. Vanaudenaerde, Y. Xu, J. M. Hohlfeld, N. Krug, J. D. Herazo-Maya, P. Rottoli, W. A. Wuyts, N. Kaminski, BAL cell gene expression is indicative of outcome and airway basal cell involvement in IPF. *Am. J. Respir. Crit. Care Med.* **199**, 622–630 (2019).
28. Y. Zeng, J. Zhu, D. Shen, H. Qin, Z. Lei, W. Li, J.-A. Huang, Z. Liu, Repression of Smad4 by miR-205 moderates TGF-β-induced epithelial-mesenchymal transition in A549 cell lines. *Int. J. Oncol.* **49**, 700–708 (2016).
29. B. C. Willis, J. M. Liebler, K. Luby-Phelps, A. G. Nicholson, E. D. Crandall, R. M. du Bois, Z. Borok, Induction of epithelial-mesenchymal transition in alveolar epithelial cells by transforming growth factor-β1. *Am. J. Pathol.* **166**, 1321–1332 (2005).
30. J. R. Rock, C. E. Barkauskas, M. J. Cronce, Y. Xue, J. R. Harris, J. Liang, P. W. Noble, B. L. M. Hogan, Multiple stromal populations contribute to pulmonary fibrosis without evidence for epithelial to mesenchymal transition. *Proc. Natl. Acad. Sci. U.S.A.* **108**, E1475–E1483 (2011).
31. M. Chilosi, V. Poletti, A. Zamò, M. Lestani, L. Montagna, P. Piccoli, S. Pedron, M. Bertaso, A. Scarpa, B. Murer, A. Cancellieri, R. Maestro, G. Semenzato, C. Doglioni, Aberrant Wnt/beta-catenin pathway activation in idiopathic pulmonary fibrosis. *Am. J. Pathol.* **162**, 1495–1502 (2003).
32. C. F. Kim, E. L. Jackson, A. E. Woolfenden, S. Lawrence, I. Babar, S. Vogel, D. Crowley, R. T. Bronson, T. Jacks, Identification of bronchioalveolar stem cells in normal lung and lung cancer. *Cell* **121**, 823–835 (2005).
33. M. Ebina, M. Shimizukawa, N. Shibata, Y. Kimura, T. Suzuki, M. Endo, H. Sasano, T. Kondo, T. Nukiwa, Heterogeneous increase in CD34-positive alveolar capillaries in idiopathic pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* **169**, 1203–1208 (2004).
34. M. P. Keane, D. A. Arenberg, J. P. Lynch III, R. I. Whyte, M. D. Iannettoni, M. D. Burdick, C. A. Wilke, S. B. Morris, M. C. Glass, B. Di Giovine, S. L. Kunkel, R. M. Strieter, The CXC chemokines, IL-8 and IP-10, regulate angiogenic activity in idiopathic pulmonary fibrosis. *J. Immunol.* **159**, 1437–1443 (1997).
35. M. Turner-Warwick, Precapillary systemic-pulmonary anastomoses. *Thorax* **18**, 225–237 (1963).
36. D. W. Waters, K. E. C. Blokland, P. S. Pathinayake, J. K. Burgess, S. E. Mutsaers, C. M. Prele, M. Schuliga, C. L. Grainge, D. A. Knight, Fibroblast senescence in the pathology of idiopathic pulmonary fibrosis. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **315**, L162–L172 (2018).
37. L. R. Rodriguez, M. Emblom-Callahan, M. Chhina, S. Bui, B. Aljeburry, L. H. Tran, R. Novak, M. Lemma, S. D. Nathan, G. M. Grant, Global gene expression analysis in an in vitro fibroblast model of idiopathic pulmonary fibrosis reveals potential role for CXCL14/CXCR4. *Sci. Rep.* **8**, 3983 (2018).
38. T. Xie, Y. Wang, N. Deng, G. Huang, F. Taghavifar, Y. Geng, N. Liu, V. Kulur, C. Yao, P. Chen, Z. Liu, B. Stripp, J. Tang, J. Liang, P. W. Noble, D. Jiang, Single-cell deconvolution of fibroblast heterogeneity in mouse pulmonary fibrosis. *Cell Rep.* **22**, 3625–3640 (2018).
39. S. G. Chu, S. P. De Frias, Y. Sakairi, R. S. Kelly, R. Chase, K. Konishi, A. Blau, E. Tsai, K. Tsoyi, R. F. Padera, L. M. Sholl, H. J. Goldberg, H. R. Mallidi, P. C. Camp, S. Y. El-Chemaly, M. A. Perrella, A. M. K. Choi, G. R. Washko, B. A. Raby, I. O. Rosas, Biobanking and cryopreservation of human lung explants for omic analysis. *Eur. Respir. J.* **55**, 1801635 (2019).
40. R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, N. Yosef, Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
41. A. Scialdone, A. Scialdone, K. N. Natarajan, L. R. Saraiva, V. Proserpio, S. A. Teichmann, O. Stegle, J. C. Marioni, F. Buettner, Computational assignment Of cell-cycle stage from single-cell transcriptome data. *Methods* **85**, 54–61 (2015).
42. X. Wang, J. Park, K. Susztak, N. R. Zhang, M. Li, Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10**, 380 (2019).
43. S. Aibar, C. B. González-Blas, T. Moerman, V. A. Huynh-Thu, H. Imrichova, G. Hulselmans, F. Rambow, J.-C. Marine, P. Geurts, J. Aerts, J. van den Oord, Z. K. Atak, J. Wouters, S. Aerts, SCENIC: Single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
44. A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, M. Milacic, C. D. Roca, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, G. Viteri, J. Weiser, G. Wu, L. Stein, H. Hermjakob, P. D'Eustachio, The reactome pathway knowledgebase. *Nucleic Acids Res.* **46**, D649–D655 (2018).

Acknowledgments: We are indebted to all patients and control subjects who participated in this study. The sequencing was conducted by M. Zhong at Yale Stem Cell Center Genomics Core facility. We thank A. Brooks and the staff of Yale Pathology Tissue Services for tissue processing. We also thank M.C. Nawijn for providing access to group's data. **Funding:** This work was supported by NIH grants R01HL127349, U01HL145567, U01HL122626, and U54HG008540 to N.K., NHLBI P01 HL114501 and support from the Pulmonary Fibrosis Fund to

Downloaded from http://advances.sciencemag.org/ on July 10, 2020

I.O.R., an unrestricted gift from Three Lake Partners to I.O.R. and N.K., and by the German Research Foundation (SCHU 3147/1) to J.C.S. **Author contributions:** N.K. and I.O.R. conceptualized, acquired funding, and supervised the study. B.A.R. and G.R.W. performed sample collection and phenotyping. S.P., E.A.A, and S.G.C. procured and dissociated the lungs. T.S.A., J.C.S., S.P., F.A., and G.D. performed library construction. Data were processed, curated, and visualized by T.S.A., J.C.S., N.N., and X.Y. and analyzed by T.S.A., J.C.S., N.K., X.Y., N.N., M.J., Q.D., H.A.A., and A.S. The online tool "IPFCellAtlas.com" was developed by N.N. IHC was performed by J.C.S. and T.S.A and evaluated by R.H., N.K., and J.C.S. The manuscript was drafted by T.S.A, J.C.S, and N.K. and was reviewed and edited by all other authors. **Competing interests:** T.S.A., J.C.S., S.P., E.A.A, H.A.A., A.S., I.O.R., and N.K. are inventors on a provisional patent application (62/849,644) submitted by NuMedii Inc., Yale University, and Brigham and Women's Hospital Inc. that covers methods related to IPF-associated cell subsets. N.K. served as a consultant to Biogen Idec, Boehringer Ingelheim, Third Rock, Pliant, Samumed, NuMedii, Indaloo, Theravance, LifeMax, Three Lake Partners, and Optikira over the past 3 years and received nonfinancial support from MiRagen. H.A.A., A.S., M.J., and Q.D. are employees of NuMedii Inc. The authors declare that they have no other competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors. Transcriptomic data was deposited to the GEO under accession number GSE136831.

Submitted 13 November 2019
Accepted 1 June 2020
Published 8 July 2020
10.1126/sciadv.aba1983

Citation: T. S. Adams, J. C. Schupp, S. Poli, E. A. Ayaub, N. Neumark, F. Ahangari, S. G. Chu, B. A. Raby, G. Deluliis, M. Januszyk, Q. Duan, H. A. Arnett, A. Siddiqui, G. R. Washko, R. Homer, X. Yan, I. O. Rosas, N. Kaminski, Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Sci. Adv.* **6**, eaba1983 (2020).

Downloaded from http://advances.sciencemag.org/ on July 10, 2020

Science Advances

Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis

Taylor S. Adams, Jonas C. Schupp, Sergio Poli, Ehab A. Ayaub, Nir Neumark, Farida Ahangari, Sarah G. Chu, Benjamin A. Raby, Giuseppe DeIuliis, Michael Januszyk, Qiaonan Duan, Heather A. Arnett, Asim Siddiqui, George R. Washko, Robert Homer, Xiting Yan, Ivan O. Rosas and Naftali Kaminski

Sci Adv **6** (28), eaba1983.
DOI: 10.1126/sciadv.aba1983

ARTICLE TOOLS http://advances.sciencemag.org/content/6/28/eaba1983

SUPPLEMENTARY MATERIALS http://advances.sciencemag.org/content/suppl/2020/07/06/6.28.eaba1983.DC1

REFERENCES This article cites 43 articles, 9 of which you can access for free
http://advances.sciencemag.org/content/6/28/eaba1983#BIBL

PERMISSIONS http://www.sciencemag.org/help/reprints-and-permissions

Use of this article is subject to the Terms of Service

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.

Copyright © 2020 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).

Downloaded from http://advances.sciencemag.org/ on July 10, 2020