Online Supplement Methods

Subjects and tissue preparation. Normal control lungs were obtained under a protocol approved by the University of Pittsburgh, Committee for Oversight of Research and Clinical Training Involving Decedents, following rejection as candidate donors for transplant. IPF lung tissue was obtained under a protocol approved by the University of Pittsburgh, Institutional Review Board, during transplantation surgery. Following the removal of the transplant recipient’s lungs, sections from the upper and lower lobe were cut, placed in Perfadex® (XVIVO), and arrived for processing/enzymatic digestion within twenty minutes of removal from the patient. Pieces were then allocated for formalin fixation, preservation in RNA Later, and immediate digestion for scRNA-seq. Tissue for scRNA-seq was diced and enzymatically digested in DMEM (Thermo Fisher Scientific) containing 0.7 mg/mL collagenase A (Roche) and 30 ug/mL DNAase I (Roche) for one hour and further mechanically dispersed using the Miltenyi gentleMACS Octo Dissociator (Miltenyi Biotec). The resulting cell suspension was washed with PBS, filtered through 70 micron cell strainers twice, underwent RBC lysis using BD Pharm Lyse, and cells were finally re-suspended in PBS (Thermo Fisher Scientific) containing 0.04% BSA (Sigma). For lower lobes of IPF lungs, a duplicate sample was run following a failed attempt to deplete CD16-expressing cells from a portion of the lower IPF lobe single cell suspension.

Single Cell RNA Library Preparation and Sequencing. Single cell RNA library preparation was performed utilizing the 10X Genomics Chromium instrument and its associated V2 single cell chemistry per the manufacturer’s protocol. In brief, 7000 cells were mixed with reverse transcription reagents, loaded into a Single Cell A chip. Single Cell 3’ gel beads were then loaded into the chip, followed by partitioning oil. Cells were separated into oil micro-droplet partitions by the Chromium instrument, containing a cell and a gel bead scaffold for an oligonucleotide composed of oligo-dT, and 10X and UMI barcodes, and reverse transcription reagents, as described (1). Reverse transcription was performed, the emulsion broken and pooled fractions obtained using a recovery agent. cDNAs were amplified by 11 cycles of PCR (C1000, Bio-Rad), enzymatically sheared and DNA fragment ends were repaired, A-tailed and adaptors ligated. The library was quantified using a KAPA Universal Library Quantification Kit KK4824 (KAPA Biosystems) and further characterized for cDNA length on a Bioanalyzer using a High Sensitivity DNA kit (Agilent). Single cell RNA-seq libraries were sequenced using the Illumina NextSeq-500 through the University of Pittsburgh Genomics Core, Sequencing Facility.

Data Analysis. Raw sequencing reads were examined by quality metrics and mapped to human reference genome GRCh38 using the Cell Ranger pipeline (10X Genomics), assigning reads to their individual cells of origin according to the cell-identifying 10x barcodes. To ensure that PCR amplified transcripts were counted only once, only single UMI’s were counted for gene expression level (2). In this way, cell x UMI count matrices were generated for downstream analyses. Seurat, an R package developed for single-cell analysis (3), was used to normalize gene expression, to identify distinct cell populations and to visualize clusters graphically (4). Specifically, the cell-UMI matrix was filtered such that only cells expressing at least 200 genes were utilized in downstream analysis. Cells containing greater than 35% mitochondrial genes were also filtered out of the analysis. Highly variable genes were identified per sample, based on their average expressions and dispersions, and these sets of genes were combined for the alignment procedure. For each sample, data was scaled and the number of UMI’s per cell as well as the percentage of mitochondrial gene content were regressed out. Initially data was analyzed by simply combining the datasets from different subjects, and using the standard Seurat pipeline, generating t-SNE plots based on PCA, and SLM algorithm to assign clusters. Data was also analyzed to remove batch effects utilizing Seurat’s canonical correlation analysis (CCA), which aligned subpopulations while allowing for shifts in subpopulation frequency among samples (5). A shared gene correlation structure was learned from the data and used as a scaffold for alignment. From this, shared clusters were identified and the datasets aligned using a non-linear “warping” algorithm to normalize for differences in feature scale while accounting for shifts in population density (5). Aligning the CCA subspaces returned a new dimensional reduction that was used for clustering with the smart local
moving algorithm (SLM) implemented in Seurat (3, 6). The resultant clusters were then visualized using t-distributed stochastic neighbor embedding (t-SNE), (7, 8). We initially identified cell populations by examining gene markers in the associated transcriptomes. These two analytical approaches gave similar results, but all figures shown are from the CCA analysis with the exception of differentially expressed genes, which are derived from PCA analysis without CCA alignment (online Table S8).

Cell cycle analysis. Cell cycle scoring was completed using canonical markers to calculate and assign cell cycle phase scores. A list of gene markers was derived as described (9), and then separated into markers of G2/M phase and markers of S phase. From this, a cell cycle score was assigned with cells lacking expression of genes from either marker set assumed to be non-cycling and in phase G1.

Differential gene expression. Average expression of a gene per cluster in IPF versus normal control lungs after PCA T-SNE clustering (excluding normal lung SC155/SC156 due to high ambient RNA) was calculated utilizing Seurat’s AverageExpression function. For each cell, the UMI count of a gene was divided by the total number of UMIs. This was then averaged for each condition (normal control vs. IPF) for each of the clusters. The Seurat implementation of the Wilcoxon rank sum and MAST (10) statistical tests were employed to determine genes differentially expressed between normal controls and IPF lungs, producing a p-value ranked matrix of differentially expressed genes. A Bonferroni correction was made to correct for multiple comparisons of Wilcoxon p-values. For comparisons of altered gene expression (Table S8), statistically significant genes (p<0.05 using Bonferroni corrected Wilcoxon test) and relatively highly expressed genes (log normalized average expression in IPF cells > 0.5) were filtered to exclude ambient gene expression.

We did see evidence of doublet cell captures, expected at a rate of 1-2% using this technology. These doublet cells expressed makers of more than one cell type, but were generally proportional to the frequency of the cell population of origin. For example, we saw groupings of macrophages as well as ciliated cells, co-expressing markers of goblet cells. These highly likely represent doublet cell captures.

In some cases unexpected genes appeared upregulated in SPP1 macrophages, for example SCGB3A1 and IGKC. These did not appear to represent doublet captures since they were seen expressed at low levels in many of the cells in the cluster (online supplement Table S8). Instead these appear to be ambient RNAs possibly derived from dying cells expressing very high levels of these mRNAs.

Statistical Analysis of Population Shifts. Significant changes in the percent composition of samples by cell type were evaluated using the nonparametric Kruskal–Wallis with Dunn’s multiple comparisons test and Benjamini-Hochberg correction.

Immunofluorescent staining. Single and dual antibody staining using tyramide signal amplification (Tyramide SuperBoost Kits with Alexa Fluor Tyramides; ThermoFisher Scientific, Waltham, MA) were performed on formalin fixed paraffin embedded human lung explants obtained from either IPF transplant recipients or healthy tissue donors. Tissue sections (5µm thick) were deparaffinized and rehydrated followed by heat induced antigen retrieval in citrate buffer pH6.0 (Vector Labs, Burlingame, CA) for 10 minutes then allowed to cool for 10 minutes. Blocking was achieved by using 3% H2O2 followed by 10% goat serum (ThermoFisher Scientific, Waltham, MA) for one hour each. Single antibody stains were performed using rabbit anti-SPP1 (1:100; Abcam, Cambridge, MA); rabbit anti-FABP4 (1:1000; Abcam, Cambridge, MA); rabbit anti-MERTK (1:500; Abcam, Cambridge, MA); rabbit anti-CD163 (1:100; Abcam, Cambridge, MA); mouse anti-SMA (1:400; Dako/Agilent, Santa Clara, CA); mouse anti-KIAA0101 (1:100; Abcam, Cambridge, MA); mouse anti-FCN1 (1:10; LifeSpan Bio., Seattle, WA). All primary antibodies were incubated overnight at 4°C. Enzymatic development was performed using appropriate goat anti-rabbit or goat anti-mouse Poly-HRP conjugated secondary antibodies (ThermoFisher Scientific, Waltham, MA) for one hour followed by Alexa Fluor 488, 594 or 647 labeled tyramide solution and completed with reaction stop solution (ThermoFisher Scientific, Waltham, MA). Dual antibody stained samples underwent the same staining process; after the first antibody
development was complete tissue again underwent an additional heat induced antigen retrieval, blocking, incubation with compatible primary antibody, poly-HRP secondary antibody and development with spectrally compatible tryamide Alexa Fluor. Slides underwent nuclear staining with Hoeschst stain (1:2000; ThermoFisher Scientific, Waltham, MA). All wash steps consisted of PBS washes 3 times 10 minutes each. Finally, slides were cover slipped, using ProLong Diamond Antifade Mountant (ThermoFisher Scientific, Waltham, MA). Images were taken using an Olympus FLUOVIEE FV1000 (Olympus, Waltham, MA) confocal laser-scanning microscope.

<table>
<thead>
<tr>
<th>Antibody</th>
<th>Source</th>
<th>Company</th>
<th>Cat #</th>
<th>Dilution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anti- MERTK</td>
<td>Rabbit monoclonal</td>
<td>Abcam</td>
<td>Ab52968</td>
<td>1:500</td>
</tr>
<tr>
<td>Anti- FABP4</td>
<td>Rabbit monoclonal</td>
<td>Abcam</td>
<td>Ab92501</td>
<td>1:1000</td>
</tr>
<tr>
<td>FCN1</td>
<td>Mouse monoclonal</td>
<td>LifeSpan BioSciences</td>
<td>LS-C490736</td>
<td>1:10</td>
</tr>
<tr>
<td>SMA</td>
<td>Mouse monoclonal</td>
<td>Dako/ Agilent</td>
<td>M0851</td>
<td>1:400</td>
</tr>
<tr>
<td>SPP1</td>
<td>Rabbit polyclonal</td>
<td>Abcam</td>
<td>Ab8448</td>
<td>1:100</td>
</tr>
<tr>
<td>CD163</td>
<td>Rabbit monoclonal</td>
<td>Abcam</td>
<td>Ab182422</td>
<td>1:500</td>
</tr>
<tr>
<td>KIAA0101</td>
<td>Mouse monoclonal</td>
<td>Abcam</td>
<td>Ab56773</td>
<td>1:100</td>
</tr>
</tbody>
</table>

**Explant culture with EDU labeling and fluorescent detection.** In a 24 well plate approximately 1mm tissue explants of IPF lower lobe lung tissue was cultured for 3 hours in macrophage medium (Cell Biologics, Chicago, IL) with 10μM EdU solution from CLICK-it Plus Edu Imaging Kit(ThermoFisher Scientific, Waltham, MA). Immediately following a 3-hour incubation the tissue was fixed with 10% neutral buffered formalin, then paraffin embedded and sectioned. Microscope slides with 5-micron tissue sections were permeabilized for 15 minutes with 0.5% Triton X-100, washed with 3% BSA. EdU detection was performed using EdU reaction buffer according to manufacturer’s recommendation. Immediately following antibody staining was performed using the same tyramide immunofluorescence protocol previously described with rabbit anti-MERTK (1:500; Abcam, Cambridge, MA). Nuclei were stained with Hoeschst stain (1:2000; ThermoFisher Scientific, Waltham, MA).

**Graphical models for connectivity maps.** In order to identify direct (“causal”) relationships among the expression levels of the selected differentially expressed genes in different cell types, we used the fast greedy equivalence search (FGES) algorithm (11-13). FGES performs a greedy search in which edges are added between nodes until no additional edge increases the Bayesian Information Criterion (BIC) score, and then edges are removed until no additional removal increases the score. The analysis included the top 100 differentially expressed genes for each of the following cell type groups: cluster 0 SPP1 macrophage (4,489 cells), cluster 2 fibroblast (2,270 cells), cluster 6 AT1 (87 cells), cluster 6 club and goblet (1,781 cells), cluster 12 KRT5 basal epithelial (797 cells), and cluster 14 AT2 (733 cells). Genes that were differentially expressed in more than one of these groups were excluded, resulting in 394 gene expression variables measured over 10,157 cells. The variables were transformed using nonparanormal (14, 15) to relax the normality assumption. The penalty discount parameter, c, for the BIC scoring function was chosen by a method analogous to Stability Approach to Regularization Selection, StARS (16) as we describe in (17). For a dataset with n samples by d variables and a given c, we drew 40 random subsamples of size \( \frac{n}{2} \) according to complementary pairs stability selection (18). A network was constructed for each subsample using FGES as implemented in Tetrad (19). For a given c, \( \hat{\theta}_{ij}(c) \) was defined as the fraction of subsample networks in which an edge between node i and node j appeared.
The edge instability was then calculated according to $\xi_{ij}(c) = 2 \theta_{ij}(c) \left(1 - \hat{\theta}_{ij}(c)\right)$. The total instability of the network for a given penalty discount was calculated as the average instability over all edges, $\bar{D}(c) = \frac{\sum_{ij} \xi_{ij}(c)}{\binom{d}{2}}$. A penalty discount value of 6 was chosen based on the total instability and sparsity in the resulting network.

**Online Supplement Results**

**Identification of healthy lung cells on t-SNE analysis.** Normal lung cells were identified on t-SNE plots by expression of marker genes. Club cells, expressed the marker SCGB1A1 (cluster 12), alveolar type I (AT1) cells expressed AGER (cluster 8), alveolar type II cells (AT2) expressed SFTPC (cluster 6), and ciliated cells expressed FOXJ1 (cluster 9; 20). KRT5 expressing basal airway cells, and MUC5B expressing goblet cells were not found in discrete clusters (21). Fibroblasts expressing COL1A1 and smooth muscle cells expressing DES were both found in cluster 7. Endothelial cells, expressing VWF, and lymphatic endothelial cells, expressing LYVE1, clustered in two distinct groups (clusters 4 and 11, respectively). Cells with pericyte markers RGS5 were also distributed in cluster 4. Macrophages, identified by AIF1 and CD163 expression, were found in six different clusters, (clusters 0, 1, 3 and 14). A small subset of CD1C-expressing dendritic cells were found in cluster 10. TPSAB1, a marker of mast cells, was expressed by cells in cluster 13. T lymphocytes, including CD3 and CD8A positive cells, clustered together (cluster 2), while B lymphocyte lineage were marked with M54A1/CD20, (cluster 15), including high expression of immunoglobulin genes (IGKC), and MZB1, a plasma cell marker. NK cells, expressing GNLY, were found in cluster 5.

**Identification of healthy and IPF lung cells on combined t-SNE analysis.** Similarly as for the normal lung analyses, cell types in the combined IPF, healthy lung t-SNE plots were identified by marker genes. Alveolar epithelial cells were identified using markers described previously (20). Club cells expressed the markers SCGB1A1 and SCGB3A2, alveolar type I (AT1) cells, expressed AGER, and goblet cells expressed MUC5B. Alveolar type II cells (AT2) were marked by SFTPC, FOXJ1 marked ciliated cells and KRT5 marked basal airway cells (21). Mesenchymal, perivascular and vascular cells were identified in several clusters. Fibroblasts were identified by expression of COL1A1, COL1A2 and PDGFRA and smooth muscle cells by expression of DES and ACTG2. The markers on these cells are highly conserved with markers we recently described in the skin (22). Endothelial cells, expressing VWF, clustered in cluster 3. RGS5 positive pericytes clustered with DES-expressing smooth muscle cells (cluster 10). Lymphatic endothelial cells, marked by LYVE1 (23), formed a distinct cluster.

Cells in other clusters showed markers of different inflammatory cell types. Macrophages, were identified by AIF1 and CD163 expression. A small subset of CD1C-expressing dendritic cells were found in cluster 0. T lymphocytes, including CD3 and CD8A positive cells, clustered together, while B lymphocyte lineage cells grouped in two distinct clusters, one showing markers of both immature and mature B cells: CD79A, part of the B-cell receptor and M54A1/CD20; the other showing high expression of immunoglobulin genes (IGKC, IGHA1 and IGHG3), and MZB1, recently described as a marker upregulated in plasma cells in IPF lungs (24). Adjacent to T cells, were cells showing markers most consistent with NK cells, expressing GNLY, NKG7, GZMB, PRF1, and CST7 but not CD3D or TRAC, T cell receptor genes (25-27). GNLY/Granulysin a protein found in cytolytic granules was the most discrete marker of these cells. Transcriptomes of innate lymphoid cells have been recently characterized in detail. The cells in this cluster also express other markers of ILC1/NK cells, CCL3, CCL4 and CCL5 (28). Finally cells with markers of mast cells: TPSAB1, CPA3, M54A2 were found in cluster 13.

Differentially expressed genes associated with each cell type were analyzed excluding healthy lung SC155/156 due to high ambient RNA in this sample and are shown online. Differential gene expression was performed with and without CCA, showing similar results. Differential gene expression and statistical analysis without CCA is presented online: https://www.dept-med.pitt.edu/rheum/Centers%20and%20Institutes/CORT/Database.html
Connectivity maps of cell subtypes. We built a graphical model to observe the direct relationships between differentially expressed genes for SPP1/MERTK macrophages, fibroblasts, and various epithelial cell types. The resulting network included 394 differentially expressed genes and 3,951 edges (connections between these genes, full connectivity map is shown in Table S7). SPP1 macrophages and fibroblasts were the most densely connected gene groups with 944 edges between them (Figure S26), suggesting a causal relationship between these two cell types.

We also explored the first order neighbors of SPP1 and MERTK, which were first order neighbors of each other, and COMP, the most highly upregulated gene in fibroblasts (Figure S26). First order neighbors of SPP1 and MERTK included genes downregulated in IPF fibroblasts: CFD/adipsin/complement factor D, and SEPP1/Selenoprotein P, a gene also downregulated in prostatic fibroblasts by TGFβ (25). The only fibroblast first-order neighbor of MERTK/SPP1 macrophages was VCAN/versican, a protein associated with early fibroblastic foci (26), upregulated in IPF fibroblasts. FGG and SERPINA1, first order neighbors expressed by AT2 cells, may represent genes involved in AT2 cell injury.

Online Supplement References
landscape of intestinal innate lymphoid cells are shaped by the microbiome.

