A molecular cell atlas of the human lung from single-cell RNA sequencing

https://doi.org/10.1038/s41586-020-2922-4

Received: 24 August 2019

Accepted: 26 August 2020

Published online: 18 November 2020

Check for updates

Kyle J. Travaglini[1,2,16], Ahmad N. Nabhan[1,2,12,16], Lolita Penland[3,13], Rahul Sinha[4,5], Astrid Gillich[1,2], Rene V. Sit[3], Stephen Chang[1,2], Stephanie D. Conley[4,5], Yasuo Mori[4,5,14], Jun Seita[4,5,15], Gerald J. Berry[5], Joseph B. Shrager[6], Ross J. Metzger[2,7], Christin S. Kuo[8], Norma Neff[3], Irving L. Weissman[4,5,9,10], Stephen R. Quake[3,11] ✉ & Mark A. Krasnow[1,2] ✉

Although single-cell RNA sequencing studies have begun to provide compendia of cell expression profiles[1–9], it has been difficult to systematically identify and localize all molecular cell types in individual organs to create a full molecular cell atlas. Here, using droplet- and plate-based single-cell RNA sequencing of approximately 75,000 human cells across all lung tissue compartments and circulating blood, combined with a multi-pronged cell annotation approach, we create an extensive cell atlas of the human lung. We define the gene expression profiles and anatomical locations of 58 cell populations in the human lung, including 41 out of 45 previously known cell types and 14 previously unknown ones. This comprehensive molecular atlas identifies the biochemical functions of lung cells and the transcription factors and markers for making and monitoring them; defines the cell targets of circulating hormones and predicts local signalling interactions and immune cell homing; and identifies cell types that are directly affected by lung disease genes and respiratory viruses. By comparing human and mouse data, we identified 17 molecular cell types that have been gained or lost during lung evolution and others with substantially altered expression profiles, revealing extensive plasticity of cell types and cell-type-specific gene expression during organ evolution including expression switches between cell types. This atlas provides the molecular foundation for investigating how lung cell identities, functions and interactions are achieved in development and tissue engineering and altered in disease and evolution.

Since Malpighi[10], dozens of lung cell types have been discovered by microscopy[11], creating histological atlases that are the cellular foundation for pulmonary medicine. More recently, cell-type-specific markers[12,13] have been identified that provide molecular definitions and functions of the cell types[14], reaching its apex in genome-wide expression profiling by single-cell RNA sequencing (scRNA-seq)[15–19]. We sought to create a comprehensive molecular cell atlas of adult human lung using scRNA-seq analysis, a substantial challenge because the 45 histological cell types have diverse structures, locations, and abundances that vary over five orders of magnitude (Supplementary Table 1).

Fifty-eight molecular cell types of the human lung

We acquired histologically normal lung tissue intraoperatively from bronchi (proximal), bronchiole (medial), and alveolar (distal) regions along with peripheral blood (Extended Data Fig. 1a, d). Lung samples were dissociated into cell suspensions, and each suspension was sorted into epithelial (EPCAM$^+$), endothelial/immune (CD31$^+$CD45$^+$) and stromal (EPCAM$^-$CD31$^-$CD45$^-$) populations (Supplementary Fig. 1a). This allowed us to balance tissue compartment representation for sequencing. We also sorted blood cells to balance immune lineages (Supplementary Fig. 1b). Sequencing libraries were prepared using 10x Chromium (10x) or SmartSeq2 (SS2)[20]. Higher throughput of 10x enabled discovery of rare cell types, whereas SS2 gave deeper transcriptomic information; there were also platform-specific idiosyncrasies in cell capture. We sequenced thousands of cells from each compartment for each subject (Supplementary Table 2) to directly compare cell types without batch correction, and did so for three subjects to address individual differences. High-quality transcriptomes were obtained from approximately 75,000 cells (65,662 10x; 9,404 SS2).

[1]Department of Biochemistry, Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA, USA. [2]Vera Moulton Wall Center for Pulmonary Vascular Disease, Stanford University School of Medicine, Stanford, CA, USA. [3]Chan Zuckerberg Biohub, San Francisco, CA, USA. [4]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Stanford, CA, USA. [5]Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA. [6]Department of Cardiothoracic Surgery, Stanford University School of Medicine, Stanford, CA, USA. [7]Department of Pediatrics, Division of Cardiology, Stanford University School of Medicine, Stanford, CA, USA. [8]Department of Pediatrics, Pulmonary Medicine, Stanford University School of Medicine, Stanford, CA, USA. [9]Ludwig Center for Cancer Stem Cell Research and Medicine, Stanford University School of Medicine, Stanford, CA, USA. [10]Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA. [11]Department of Bioengineering, Stanford University, Stanford, CA, USA. [12]Present address: Genentech, South San Francisco, CA, USA. [13]Present address: Calico Life Sciences, South San Francisco, CA, USA. [14]Present address: Department of Medicine and Biosystemic Science, Kyushu University Graduate School of Medical Science, Fukuoka, Japan. [15]Present address: Medical Sciences Innovation Hub Program, RIKEN, Tokyo, Japan. [16]These authors contributed equally: Kyle J. Travaglini, Ahmad N. Nabhan. ✉e-mail: steve@czbiohub.org; krasnow@stanford.edu

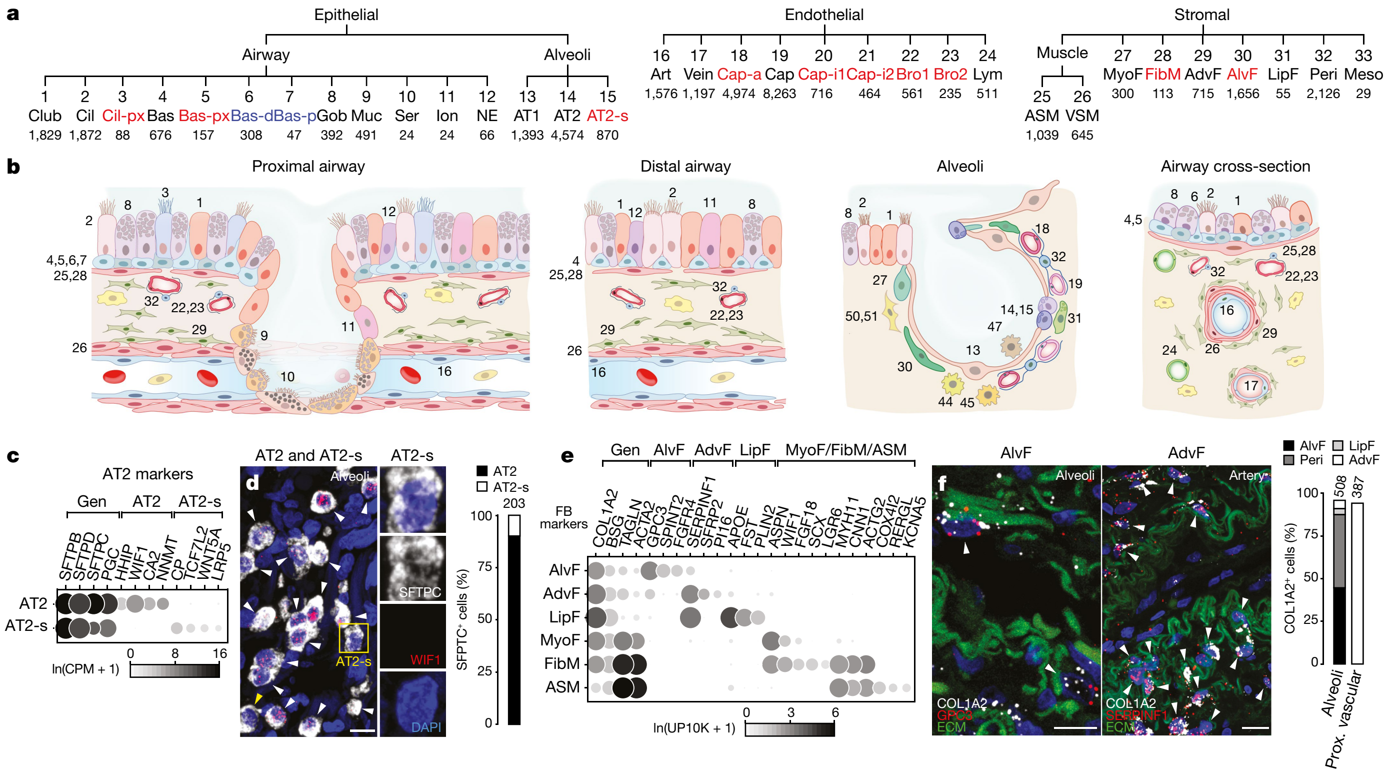

Fig. 1 | Identities and locations of lung epithelial, endothelial, and stromal cell types. a, Human lung molecular cell types identified after iterative clustering (each level of hierarchy is an iteration) of scRNA-seq profiles of cells in indicated tissue compartments. Black, canonical types; blue, proliferating or differentiating subpopulations; red, novel populations. Number of cells shown below cluster name. AdvF, adventitial fibroblast; AlvF, alveolar fibroblast; Art, artery; ASM, airway smooth muscle; AT2-s, AT2-signalling; Bas, basal; Bas-d, differentiating basal; Bas-p, proliferating basal; Bas-px, proximal basal; Bro1, bronchial vessel 1 cell; Bro2, bronchial vessel 2 cell; Cap, general capillary cell; Cap-a, capillary aerocyte; Cap-i1, capillary intermediate 1 cell; Cap-i2, capillary intermediate 2 cell; Cil, ciliated; Cil-px, proximal ciliated; FibM, fibromyocyte; Gob, goblet; Ion, ionocytes; LipF, lipofibroblast; Lym, lymphatic; Meso, mesothelial; MyoF, myofibroblast; Muc, mucous; NE, neuroendocrine; Peri, pericyte; Ser, serous; VSM, vascular smooth muscle. **b**, Diagrams showing localization and morphology of each type (cell type numbering or names in **a** and Fig. 2a). **c**, Dot plot of AT2 marker expression (10x dataset). UP10K, unique molecular identifiers (UMIs) per 10,000. CPM, counts per million mapped reads. **d**, smFISH and quantification ($n = 203$ cells scored, staining repeated in two different participants from those profiled) for shared AT2 and AT2-signalling marker *SFTPC* (white) and specific AT2 marker *WIF1* (red puncta). Scale bar, 10 μm. AT2-signalling cells (*SFTPC*$^+$ *WIF1*$^-$; box, enlarged at right, yellow arrowhead) are intermingled among AT2 cells (*SFTPC*$^+$ *WIF1*$^+$, white arrowheads). **e**, Dot plot of stromal markers (10x dataset). FB, fibroblast. **f**, smFISH and quantification for general fibroblast marker *COL1A2* (white), alveolar fibroblast marker *GPC3* (red, left) and adventitial fibroblast marker *SERPINF1* (red, right). Blue, DAPI; green, extracellular matrix (ECM; autofluorescence); prox., proximal. Adventitial fibroblasts (arrowheads, right) localize around vessels (ECM). Graph shows quantification of stromal cell type in alveolar and proximal vascular regions (*n* denotes number of cells scored in each region; staining repeated in two different participants from those profiled). Pericyte and lipofibroblast marker staining in Extended Data Fig. 4h, i. Scale bars, 10 μm. For more details on statistics and reproducibility, see Methods.

We grouped cells based on the expression of compartment-specific markers (Extended Data Fig. 1b), then iteratively clustered[21] them for each subject to identify transcriptionally distinct cell populations. Populations between subjects were merged using cluster-specific marker genes for downstream analyses. Our approach identified 58 transcriptionally distinct cell populations (mean 51 per subject) (Extended Data Fig. 1c, Supplementary Table 2), 37 more than a recent state-of-the-art study[19].

Transcriptomes of canonical cell types

The 58 molecular types included 15 epithelial, 9 endothelial, 9 stromal and 25 immune populations, greater than the number of classical cell types in each compartment (Supplementary Table 2). Using extant and newly identified (bronchial vessel) markers (Supplementary Table 1) and single-molecule fluorescence in situ hybridization (smFISH), we found clusters that represent all but one classical lung cell type in epithelial, endothelial and stromal compartments (Fig. 1a, b).

Immune cells were the most heterogeneous and included circulating, egressed and lung-resident cells. To aid identity assignment, we defined transcriptional profiles of circulating immune cells by bulk RNA sequencing of 21 sorted, functionally characterized classes of human blood cells (Extended Data Fig. 2a, Supplementary Table 3). We also obtained scRNA-seq profiles of around 5,000 blood cells from two patients whose lung cells we analysed. Canonical immune markers and the ascertained panels of differentially expressed genes were used to assign the identities to 25 immune clusters from our lung and blood scRNA-seq analysis, including all but one previously known lung immune cell type (Fig. 2a, Extended Data Fig. 2b).

Our approach defined genome-wide expression profiles for nearly all classical lung cell types (41 out of 45, 91%), from the most abundant (capillaries, approximately 23% of lung cells) to exceedingly rare (ionocytes, 0.01%) (Supplementary Table 1). One-quarter (11 out of 45) previously lacked high-quality single-cell transcriptomes. The only classical types not captured are extremely rare (neurons, glia), primarily found in disease (tuft cells)[22], or require special isolation methods (eosinophils).

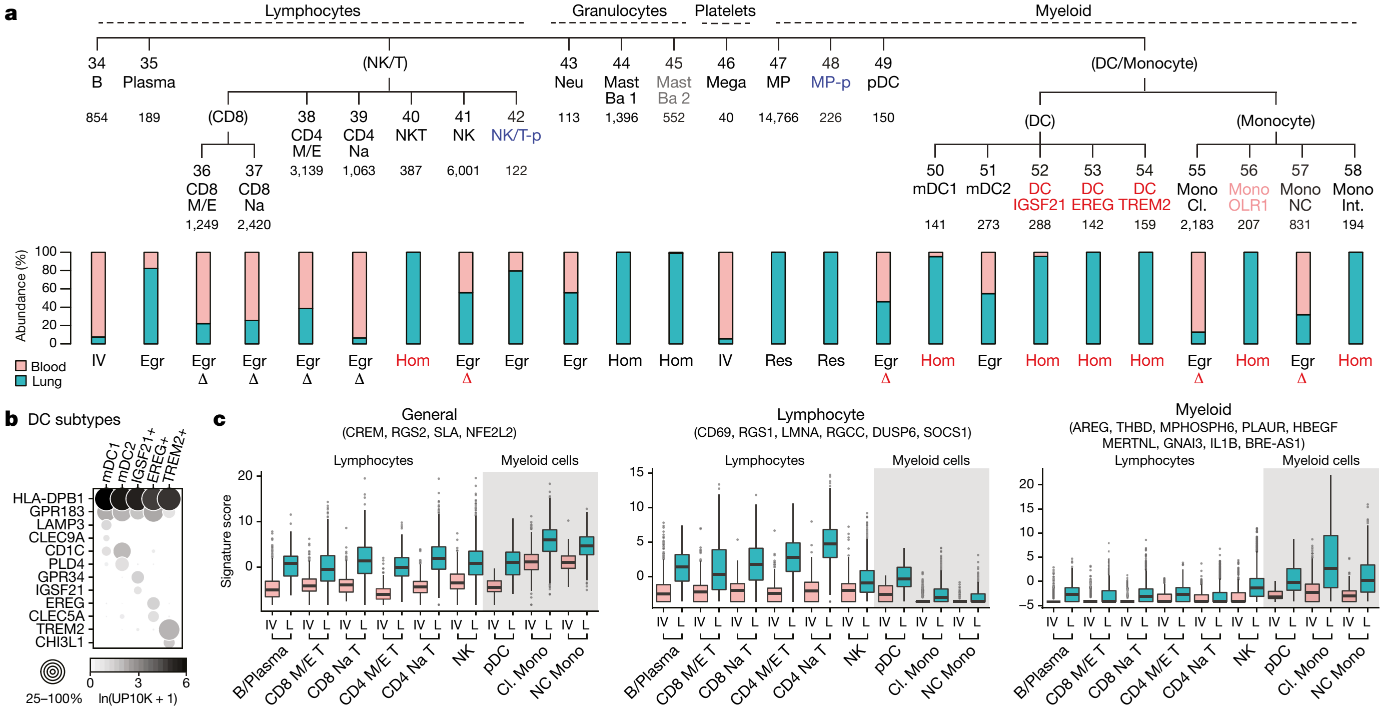

Fig. 2 | Identity and residency of lung immune cells. a, Human lung immune molecular types clustered and annotated as in Fig. 1a. Clusters 45 (grey) and 56 (light red) were found only in one subject. Bar graphs show relative abundance of each immune type in lung (blue) and blood (red) samples. Lung 'resident' (Res) or 'homing' (Hom) immune types, more than 90% enrichment in lung samples; 'intravascular' (IV), more than 90% enrichment in blood; 'egressed' (Egr), all other types (assignments are provisional because cell harvesting influences enrichment values). Red lettering denotes cells not previously known to home to (be enriched in) lung or change expression (Δ) after egression from blood. Mono Cl., classical monocyte; CD4 M/E, CD8$^+$ memory/effector T cell; CD4 Na, CD4$^+$ naive T; CD8 M/E, CD8$^+$ memory/effector T; CD8 Na, CD8$^+$ naive T; DC, dendritic cell; Mono Int., intermediate monocyte; mDC, myeloid dendritic; pDC, plasmacytoid dendritic cell; MP, macrophage; MP-p, proliferating macrophage; Mono NC, non-classical monocyte; NK, natural killer cell; NKT, natural killer T cell; NK/T-p, proliferating natural killer/T cells. **b**, Dot plot showing expression (10x dataset) in dendritic cell clusters 50–54 of, from top row to bottom: two canonical dendritic markers, four myeloid dendritic (mDC1, mDC2) markers, and six markers for three novel dendritic populations (IGSF21$^+$, EREG$^+$ and TREM2$^+$). **c**, Box-and-whisker plots of general, lymphocyte-specific, and myeloid-specific lung residency (egression) signature scores (of cells in **a**) based on expression of indicated genes in 10x profiles of indicated immune types isolated from blood (intravascular, IV) or lung (L). Many previously known lymphocyte residency genes (for example, *S1PR1*, *RUNX3*, *RBPJ* and *HOBIT*) were lowly expressed and only uncovered in SS2 profiles. Grey shading denotes myeloid cells. *n* cells in each box-and-whisker (from left to right): 725; 187; 419; 771; 631; 1,411; 594; 2,419; 644; 288; 519; 4,250; 21; 116; 1,064; 1,013; 200; and 604. For more details on statistics and reproducibility, see Methods.

New lung cell types, subtypes and states

Many canonical types were represented by more than one cluster, so the specific identities of 25 clusters remained uncertain. All but one were found in samples from several participants so were unlikely to be subject-specific (Supplementary Table 2). This suggested that the distinct expression profiles uncovered represented discrete molecular states or novel cell types or subtypes. To distinguish these possibilities, we analysed the differentially expressed genes and examined cell structure and location.

We first identified clusters representing common cell states. Three clusters (proliferating basal cells, proliferating natural killer/T cells, and proliferating macrophages) were enriched in the expression of cell cycle genes, which indicates that they represent the proliferative states of basal cells, natural killer cells, T cells and macrophages, respectively, and are the most proliferative lung cell types (Extended Data Fig. 3a). Another cluster (differentiating basal cells) had reduced expression of *KRT5* and increased expression of *HES1*, *KRT7* and *SCGB3A2*, indicating active differentiation to other epithelial fates[23,24], consistent with their transitional morphology (Extended Data Fig. b, c). Proliferating and differentiating basal cells derived mostly from proximal lung samples (Extended Data Fig. 3d, e), suggesting one-third of proximal basal cells are active.

The other basal cell clusters were quiescent and localized to proximal (large, pseudostratified) airways, or both proximal and distal (small, simple) airways (Extended Data Fig. 3e, f). The basal cell clusters are distinguished by hundreds of genes, which suggests that they are molecularly distinct cell types that differ in hormone production (*ALOX15*, *ADH7*, *SNCA*) and adhesion (*POSTN*, *ISLR*, *PCDH7*) (Extended Data Fig. 3b). There were also distinct clusters of ciliated cells along the proximal–distal axis (Extended Data Fig. 3g, h).

We uncovered two clusters of alveolar type 2 (AT2) cells (Fig. 1c), which produce surfactant that prevents alveolar collapse. These are intermingled throughout the alveolar epithelium (Fig. 1d). One cluster (*WIF1*$^+$*HHIP*$^+$*CA2*$^+$) expressed higher levels of some canonical AT2 markers (*SFTPA1*, *SFTPC* and *ETV5*) and selectively expressed inhibitors of Wnt (*WIF1*) and Hedgehog (*HHIP*) signalling and the cell cycle (*CDKN1A*), indicating that they are quiescent (Extended Data Fig. 3i, left). The other, tenfold less-abundant cluster (AT2-signalling cells) selectively expressed genes involved in Wnt signalling (*WNT5A*, *LRP5*, *CTNNBIP* and *TCF7L2* (also known as *TCF4*)) and detoxification (*CP*, *GSTA1* and *CYP4B1*) (Extended Data Fig. 3i, right). AT2-signalling cells could be alveolar stem cells, homologous to the rare, Wnt-active subpopulation of mouse AT2 cells (AT2$^{\text{stem}}$)[25,26]. However, homology between human AT2-signalling and mouse AT2$^{\text{stem}}$ cells is provisional, because although both show increased Wnt signalling or components, the many other expression differences between human AT2-signalling and 'bulk' AT2 cells are not shared by mouse AT2$^{\text{stem}}$ cells.

We found unexpected molecular diversity in the endothelial compartment (Extended Data Fig. 3j). Two populations were identified as bronchial by their localization around bronchi (Extended Data Fig. 3k). Thus, bronchial endothelial cells are distinct from their counterparts in the pulmonary circulation, distinguished by matrix (*VWA1* and *HSPG2*), fenestrated morphology[27] (*PLVAP*) and cell cycle-associated (*MYC* and *HBEGF*) genes. Four clusters of endothelial cells in the pulmonary circulation expressed capillary markers. Two (capillary aerocytes and general capillary cells) are intermingled alveolar capillary cell types[28]; the others are rare capillary types showing features of both (capillary 'intermediates' 1 and 2).

We identified new types in the stroma, the least characterized compartment. Two clusters expressed classical fibroblast markers (*BSG* and *COL1A2)* (Fig. 1e) but one (*SPINT2*$^+$*FGFR4*$^+$*GPC3*$^+$) localized to alveoli ('alveolar fibroblasts') and the other (*SFRP2*$^+$*PI16*$^+$*SERPINF1*$^+$) to vascular adventitia and nearby airways ('adventitial fibroblasts') (Fig. 1f, Extended Data Fig. 4a–d). Both expressed genes involved in canonical fibroblast functions (matrix biosynthesis, adhesion and signalling regulators) but the specific genes often differed (Extended Data Fig. 4e). Each cluster also has distinct functions: expression of voltage-gated sodium channel *SCN7A* and glutamate receptor *GRIA1* suggest alveolar fibroblasts are excitable cells with glutamatergic input (Supplementary Table 4). Their profiles also suggest novel, shared functions including the recruitment of immune cells (*IL1RL1*, *IL32*, *CXCL2* and genes in the class II major histocompatibility complex locus) and the complement system (*C2*, *C3*, *C7*, *CFI*, *CFD*, *CFH* and *CFB*).

Two stromal clusters were enriched for *ACTA2*, a canonical marker of myofibroblasts (Fig. 1e), which help form and stabilize alveoli. One cluster (*WIF1*$^+$*FGF18*$^+$*ASPN*$^+$) is classical myofibroblasts and localized to alveolar ducts (Extended Data Fig. 4f). The other ('fibromyocytes') showed higher expression of contractile genes (*MYH11*, *CNN1* and *TAGLN*), was preferentially isolated from samples of proximal lungs, and was found both intermingled with airway smooth muscle and in alveoli (Extended Data Figs. 3e, 4g). Both populations shared expression of genes for canonical fibroblast functions, although the specific genes differed from alveolar and adventitial fibroblasts (Supplementary Table 4).

Lung immune cell residency signatures

To distinguish between lung-resident, egressed and circulating immune cells, we compared the relative abundance of each immune population in lung and peripheral blood samples from the same subject (Fig. 2a). Eleven clusters (including alveolar macrophages, as expected[29]) consisted of cells only from lung samples, with no or rare exception, which indicates that they are lung-resident or greatly enriched. This included three novel lung dendritic populations: IGSF21$^+$ and rare EREG$^+$ dendritic cells express asthma genes (*CCL2*, *CCL13* and *IGSF21*) and developmental signals (*EREG*, *VEGFA*, *AREG*), respectively, and both localize to proximal vessels; TREM2$^+$ dendritic cells localize to vessels and alveoli and express lipid machinery (*APOC1*, *APOE* and *CYP27A1*) (Fig. 2b, Extended Data Fig. 4k–n).

The other immune cell types were found in both lung and blood samples. For some types, every cell—whether from lung or blood—clustered together. However, for other types, cells from lung formed a separate cluster (Extended Data Fig. 4o). Some of the differentially expressed genes may be due to technical differences (for example, collagenase treatment of lung[30], circulating RNA in blood[31]), but others such as upregulation in lung cells of lymphocyte-residence gene *CD69* probably represent genes induced after egression[32]. We identified a core transcriptional signature for all human lung-resident lymphocytes (Fig. 2c), which overlaps a residence signature found by bulk RNA sequencing of CD8$^+$ T cells in mouse spleen, gut and liver[33]. We also found a residency signature for lung myeloid cells that overlaps the lymphocyte signature, supporting a core residency program for immune cells plus specific subprograms for myeloid cells and lymphocytes.

Cell markers, regulators and interactions

We identified optimal markers for each previously known and newly identified lung cell type (Extended Data Fig. 5a, Supplementary Table 4). Approximately 200 markers can distinguish virtually all types (Extended Data Fig. 5b), so could be used with multiplexed smFISH[34–36] to simultaneously detect in clinical specimens alterations in their numbers and relationships. A similar compendium of membrane protein markers (Supplementary Table 4) could be used to purify or therapeutically target specific lung cell types. We also identified around 400 cell type-selective transcription factors (Extended Data Fig. 5e, Supplementary Table 4), putative 'master regulators' that could help create all lung cell types by cellular reprogramming. These include what may be long-sought master regulators of AT1 cells (for example, *MYRF*), which comprise the gas-exchange surface, and of pericytes (*TBX5*) (Extended Data Fig. 5c, d).

The atlas allowed us to map the cell targets of circulating hormones, based on expression of their cognate receptors. Receptors for some hormones are broadly expressed, indicating direct action throughout the lung (Extended Data Fig. 6a). Other hormones have specific and unexpected targets, such as somatostatin (*SSTR1*, arteries), melanocortin (*MC1R*, ionocytes), and oxytocin (*OXTR*, ciliated cells). Pericytes are predicted targets of several hormones, which could affect their contractile machinery to regulate alveolar perfusion (Extended Data Fig. 6b). Receptors for half the hormones were not detectably expressed so these hormones may not directly influence lung physiology. We also mapped local signalling interactions by examining expression of ligands and receptors, which predicts up to hundreds of interactions among neighbouring cell types (Extended Data Fig. 6c, Supplementary Table 5).

The expression of chemokine receptors illuminated immune cell homing (Fig. 3). Our data confirmed canonical homing interactions such as CD4$^+$ T cells to lymphatic vessels, and provides specificity for others such as plasma cell homing to epithelial mucosa through *CCL28* from serous cells. It also predicts new interactions such as *CX3CR1*-mediated homing of nonclassical monocytes to *CX3CL1*-expressing endothelial and airway epithelial cells. All three new dendritic populations express *CCR1*, which could mediate their attraction to veins (*CCL23*), bronchial vessels (*CCL14*), ciliated cells (*CCL15*), and lymphocytes (*CCL5*). Ionocytes are the only non-immune cell to express appreciable levels of any chemokine receptor (*CXCR4*).

Mapping cellular focus of lung diseases

We determined the expression of 233 extant lung disease genes (Extended Data Fig. 7). Disease genes with cell-type-specific expression (Extended Data Fig. 8a) and cell types expressing many genes associated with a specific disease (Extended Data Fig. 8b) are of special interest because they can pinpoint the cellular origin of disease. This supported known or suspected 'culprit' cells for 27 genes involved in 12 diseases, and identified potential culprits for 21 genes implicated in 15 diseases including pericytes in pulmonary hypertension, capillaries in atrioventricular dysplasia, and AT2 cells in chronic obstructive pulmonary disease (COPD). We confirmed pericyte, capillary and AT2 expression of disease genes by smFISH (Extended Data Fig. 8c–e).

We mapped expression of 80 genes encoding virus receptors, including 26 used by respiratory viruses (Extended Data Figs. 9a, 10). *NECTIN4* (measles virus receptor) was enriched in club, ciliated, differentiating basal, and goblet cells, and *CDHR3* ('common cold' rhinovirus C) was enriched in ciliated and neuroendocrine cells, indicating that infections initiate in those bronchial types. By contrast, *ACE2* (SARS, COVID-19 coronaviruses) and *DPP4* (MERS coronavirus) were both detected in AT2 cells (Extended Data Fig. 9b), consistent with severe alveolar pathology[37].

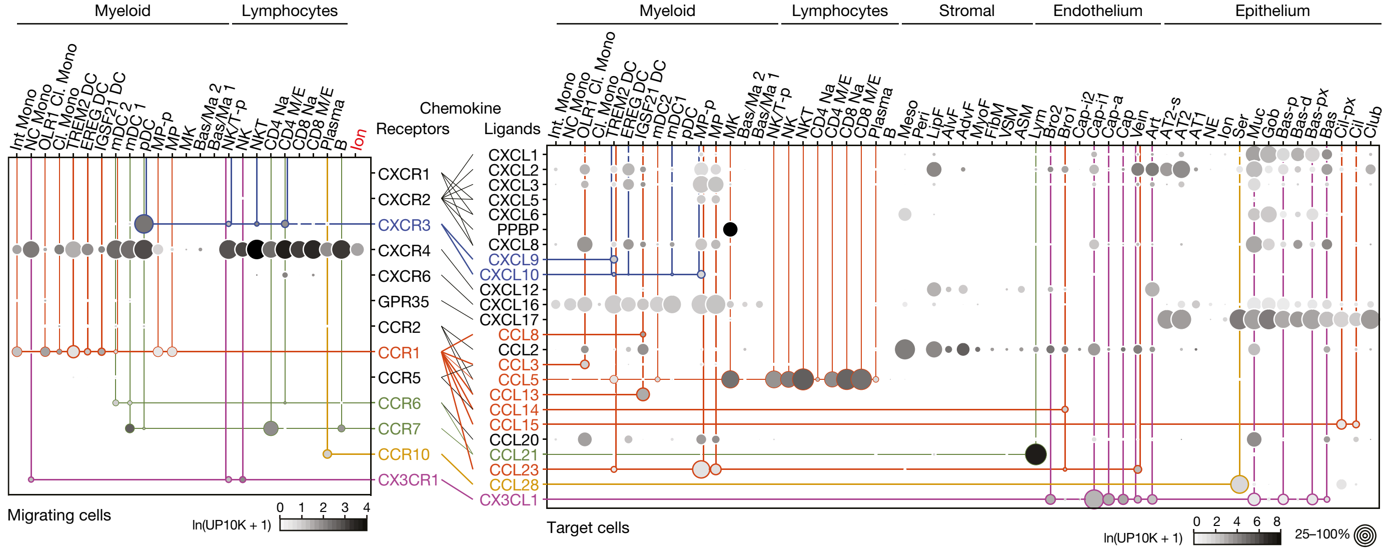

Fig. 3 | Chemokine signalling predicts immune cell homing in lung. Dot plots showing expression of chemokine receptors (left) and ligands (right) in human lung cells (10x dataset); only cell types and chemokines with detected expression are shown. Colored lines connect ligand sources (target cells) with migrating immune cell types and ionocytes (ion, red) expressing cognate receptor; thicker lines indicate previously unknown interactions. For more details on statistics and reproducibility, see Methods. Bas/Ma, basophil/mast cell; MK, megakaryocyte.

Evolution of cell types and expression

Construction of a mouse lung atlas[2] plus additional cells annotated as above for human (Supplementary Table 6) allowed analysis of evolutionary conservation of lung cell types and their transcriptomes. Homologous cell types were assigned by conserved expression of cell-type markers (Fig. 4a). Notably, mice seem to lack 17 (29%) of the 58 human lung cell types including 12 of the 14 (86%) newly identified types. Some missing mouse populations might be rare, transient, unstable, or too diverged to relate transcriptionally so may be uncovered by further studies. By contrast, just five mouse cell populations, all immune, were not found in human. This suggests substantial diversification of lung cell types during mammalian evolution.

We compared expression levels of all active genes in each human cell type with those of the orthologous genes in the corresponding mouse type (Extended Data Fig. 11a, Supplementary Table 7). Most cell types correlated best with their counterparts across species, but surprisingly one human type (goblet) showed greater correlation with another mouse type (club, $R = 0.68$ versus 0.63) (Extended Data Fig. 11b)—despite conserved expression of canonical markers and master regulator *SPDEF* (Extended Data Fig. 11c). Corresponding cell types in human and mouse diverged in expression (a greater than 20-fold change, $P < 0.05$) of hundreds of genes, such as *SERPINA1*, *PGC*, *WIF1* and *LYZ* in AT2 cells (Fig. 4b). Lung as a whole had fewer diverged genes than any cell type, which suggests that expression lost in one type is gained in another (Extended Data Fig. 11d). Diverged genes varied above age-related expression changes in mice (Extended Data Fig. 11e) and included canonical cell-type markers, transcription factors, signalling molecules and disease genes.

Evolutionary changes in expression grouped into four types (Supplementary Table 7). Type 0 ('conserved') genes are expressed in the same cell types in mouse and human (Fig. 4e, Extended Data Fig. 12a). Type 1 ('expression gain/loss') genes show simple gain (or loss) of expression between species, which involved a single cell type (type 1a, *PGC*) (Fig. 4e), several types (type 1b, *RNASE1*) (Extended Data Fig. 12b), or entire lung (type 1c, *TRIM38*) (Extended Data Fig. 12b). Type 2 ('expression expansion/contraction') changes involved gain (or loss) of expression in additional lung cell types, expanding (or contracting) expression of the gene during evolution. For example, HOPX, the canonical AT1 transcription factor in mouse, is expressed in both AT1 and AT2 cells in human (Fig. 4c, e), which indicates the existence of other AT1 transcription factors such as *MYRF*, which is AT1-selective in both species (Extended Data Fig. 12c). Expanded expression of *RAMP3*, co-receptor for vasodilators CGRP and adrenomedullin, presumably alters pulmonary vascular response to these hormones (Extended Data Fig. 12d).

Type 3 ('expression switch') changes involve a switch in expression from one cell type to another. Two medically important examples are COPD/emphysema genes *SERPINA1* and *HHIP*, both selectively expressed in AT2 cells in human but alveolar stromal cells in mice (Fig. 4d, e, Extended Data Fig. 12e); other hedgehog pathway components were mostly conserved (Extended Data Fig. 12f). Extreme examples occurred during evolution of species-specific cell types, such as consolidation in the expression of anti-bacterial enzymes (*LTF*, *LYZ* and *BPIFB1*) from several mouse airway cells into human-specific serous cells, and consolidation of broadly expressed lipid-handling genes (*PLIN2* and *APOE*) from mouse alveolar fibroblasts (which can contain lipid droplets) and myofibroblasts to human-specific lipofibroblasts (Extended Data Fig. 12g).

Despite general conservation of cell type expression patterns noted above, only 6% of expressed genes showed fully conserved patterns (type 0), most extremely specific or broadly expressed (Extended Data Fig. 12h, Supplementary Table 8). Thus, expression patterns of nearly all genes are evolutionarily labile, most undergoing broadening (55%, type 2) or simple gain or loss (29%, type 1) and rarely cell type switching (10%, type 3) (Supplementary Table 9).

Discussion

We constructed a comprehensive expression atlas of human lung comprising 58 molecular types and their locations (Fig. 1b) including 41 out of 45 previously known cell types, all but the exceedingly rare. We identified 14 novel populations across all four compartments that are as distinct molecularly as the canonical cell types; each must be thoroughly characterized, as done for new capillary types[28]. If there are other lung cell types, they must be exceedingly rare, fragile, region- or stage-specific, or so similar to the 58 that they are not resolved by current methods.

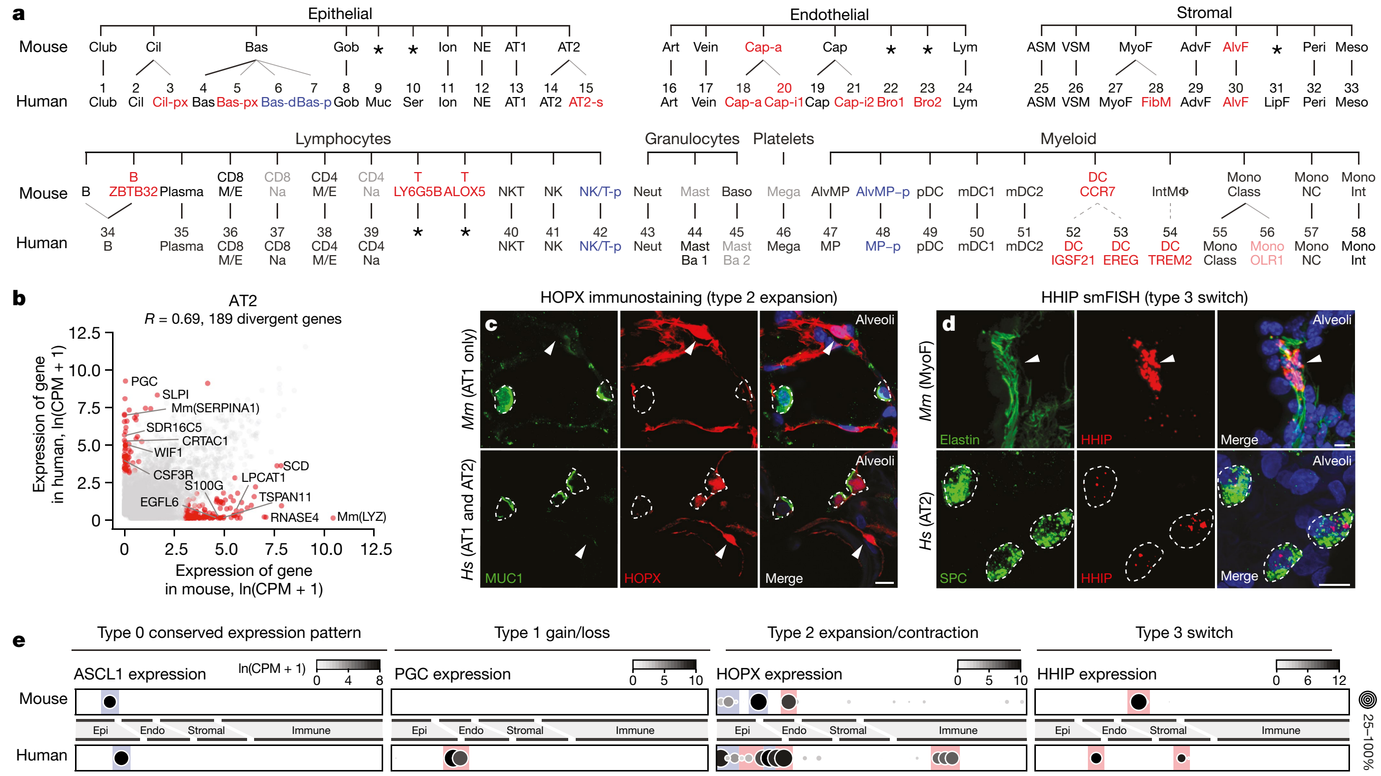

Fig. 4 | Evolutionary divergence of lung cell types and expression patterns. **a**, Mouse (top) lung molecular cell types (profiled and identified as for human, see Methods) aligned with homologous human types (bottom, Figs. 1a, 2a) by expression of classical markers in Supplementary Table 6. Thin lines, evolutionary expansions; dashed lines, potential expansions of functionally-related types. Red text, newly identified populations (light red, identified in only one subject); blue, cell states more abundant in human; grey, extant mouse cell types not captured in our data or found in only one patient in human. Asterisk denotes missing cell types. AlvMP, alveolar macrophages; AlvMP-p, proliferating alveolar macrophages. **b**, Scatter plot comparing average expression levels (dots) in AT2 cells of each expressed human gene and mouse orthologue (SS2 datasets; $n = 3{,}404$ human and 318 mouse AT2 cells). *R* denotes Pearson correlation coefficient. Red dots denote divergent genes (selected ones indicated) expressed 20-fold higher in either species. $P < 0.05$, 'MAST' differential gene expression test. Scale, ln(CPM + 1). **c**, Alveolar sections from mouse (top, *Mm*) and human (bottom, *Hs*) immunostained for HOPX (red) and AT2 marker MUC1 (green), and DAPI (blue). HOPX is expressed selectively in AT1 cells (arrowheads) in mouse but in human expression has expanded to AT2 and AT2-signalling cells (dashed circles). Scale bars, 10 μm. Staining repeated on three participants and mice. **d**, Alveolar sections from mouse (top) and human (bottom) probed by smFISH for *Hhip* and *HHIP* (red) and hydrazide staining for myofibroblast marker elastin (green) in mouse and smFISH for AT2 marker *SFTPC* (green) in human. Note *HHIP* expression switch from myofibroblast (mouse, arrowhead) to AT2 cells (human, dashed circles). Scale bars, 10 μm. Staining repeated on three human partipicants and mice. **e**, Dot plots of expression (SS2 data sets) of homologous genes indicated in mouse and human lung cell types (ordered as in **a**) exemplifying the four observed scenarios (type 0, 1, 2, 3) for evolution of cellular expression pattern. Colours highlight cell types with conserved (blue) and diverged (red) expression. Endo, endothelial; Epi, epithelial. For more details on statistics and reproducibility, see Methods.

The atlas has broad implications for physiology and medicine, providing insight into the functions, regulation and interactions of the known and new cell types. It identifies those directly affected by hormones, viruses and extant lung disease genes, and distinguishes lung-resident and homing immune cell types and infers their expression changes after egression from circulation and the cellular sources of homing signals. The atlas defines type-selective transcription factors for creating cells to engineer a lung, and provides optimal markers and a benchmark for monitoring all types and how they change during development, ageing, disease and evolution.

Mice appear to lack 17 out of the 58 human molecular lung cell types, including most (12 out of 14) of the newly discovered ones. This suggests a considerable expansion of cell types in the human lineage, perhaps for new functions, durability, or regenerative capacity of our 6,000-fold larger lungs and 30-times longer lifespan[38,39]. Even homologous cell types diverged in expression of hundreds of genes. Indeed, just 6% of expressed genes had fully conserved expression patterns across the lung, indicating widespread gain, loss or conversion of cell-type-specific transcriptional enhancers during mammalian evolution. It will be important to unravel the genetic mechanisms and functional consequences of these changes, and to determine the selective forces operative for genes with fully conserved expression. The evolutionary cell type and expression changes predict where mouse will fail to model human lung physiology and disease.

The success of our atlas relied on: procuring fresh tissue across the organ plus matched blood; balancing tissue compartments to ensure broad cell representation; extensive profiling of each subject using broad cell capture and deep gene coverage scRNA-seq strategies; clustering subject and compartment data separately and iteratively; assigning cell identies using extant markers, functions of selectively expressed genes, and tissue localization. Applying the approach to other organs could create a comprehensive human molecular cell atlas.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2922-4.

1. Enge, M. et al. Single-cell analysis of human pancreas reveals transcriptional signatures of aging and somatic mutation patterns. *Cell* **171**, 321–330.e14 (2017).
2. Tabula Muris Consortium. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
3. Han, X. et al. Mapping the mouse cell atlas by microwell-seq. *Cell* **173**, 1307 (2018).
4. Zeisel, A. et al. Molecular architecture of the mouse nervous system. *Cell* **174**, 999–1014.e22 (2018).
5. Saunders, A. et al. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* **174**, 1015–1030.e16 (2018).
6. Vento-Tormo, R. et al. Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* **563**, 347–353 (2018).
7. Young, M. D. et al. Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science* **361**, 594–599 (2018).
8. Aizarani, N. et al. A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature* **572**, 199–204 (2019).
9. Han, X. et al. Construction of a human cell landscape at single-cell level. *Nature* **581**, 303–309 (2020).
10. Young, J. Malpighi's "De pulmonibus.". *Proc. R. Soc. Med*. **23**, 1–11 (1929).
11. Gehr, P., Bachofen, M. & Weibel, E. R. The normal human lung: ultrastructure and morphometric estimation of diffusion capacity. *Respir. Physiol*. **32**, 121–140 (1978).
12. Balis, J. U., Paterson, J. F., Paciga, J. E., Haller, E. M. & Shelley, S. A. Distribution and subcellular localization of surfactant-associated glycoproteins in human lung. *Lab. Invest*. **52**, 657–669 (1985).
13. Hermans, C. & Bernard, A. Lung epithelium-specific proteins: characteristics and potential applications as markers. *Am. J. Respir. Crit. Care Med*. **159**, 646–678 (1999).
14. Franks, T. J. et al. Resident cellular components of the human lung: current knowledge and goals for research on cell phenotyping and function. *Proc. Am. Thorac. Soc*. **5**, 763–766 (2008).
15. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
16. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet*. **17**, 175–188 (2016).
17. Treutlein, B. et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014).
18. Reyfman, P. A. et al. Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. *Am. J. Respir. Crit. Care Med*. **199**, 1517–1536 (2019).
19. Braga, F. A. V. et al. A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat. Med*. **25**, 1153–1163 (2019).
20. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc*. **9**, 171–181 (2014).
21. Blondel, V. D. et al. Fast unfolding of communities in large networks. *J. Stat. Mech*. **2008**, P10008 (2008).
22. Howitt, M. R. et al. Tuft cells, taste-chemosensory cells, orchestrate parasite type 2 immunity in the gut. *Science* **351**, 1329–1333 (2016).
23. Rock, J. R. et al. Notch-dependent differentiation of adult airway basal stem cells. *Cell Stem Cell* **8**, 639–648 (2011).
24. Garcia, S. R. et al. Single-cell RNA sequencing reveals novel cell differentiation dynamics during human airway epithelium regeneration. Preprint at https://doi.org/10.1101/451807 (2018).
25. Nabhan, A. N., Brownfield, D. G., Harbury, P. B., Krasnow, M. A. & Desai, T. J. Single-cell Wnt signaling niches maintain stemness of alveolar type 2 cells. *Science* **359**, 1118–1123 (2018).
26. Zacharias, W. J. et al. Regeneration of the lung alveolus by an evolutionarily conserved epithelial progenitor. *Nature* **555**, 251–255 (2018).
27. Stan, R. V. et al. The diaphragms of fenestrated endothelia: gatekeepers of vascular permeability and blood composition. *Dev. Cell* **23**, 1203–1218 (2012).
28. Gillich, A. et al. Capillary cell-type specialization in the alveolus. *Nature* **586**, 785–789 (2020).
29. Tan, S. Y. S. & Krasnow, M. A. Developmental origin of lung macrophage diversity. *Development* **143**, 1318–1327 (2016).
30. van den Brink, S. C. et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14**, 935–936 (2017).
31. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun*. **8**, 14049 (2017).
32. Shiow, L. R. et al. CD69 acts downstream of interferon-α/β to inhibit S1P1 and lymphocyte egress from lymphoid organs. *Nature* **440**, 540–544 (2006).
33. Mackay, L. K. et al. Hobit and Blimp1 instruct a universal transcriptional program of tissue residency in lymphocytes. *Science* **352**, 459–463 (2016).
34. Moffitt, J. R. & Zhuang, X. RNA imaging with multiplexed error-robust fluorescence in situ hybridization (MERFISH). *Methods Enzymol*. **572**, 1–49 (2016).
35. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).
36. Eng, C. L. et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**, 235–239 (2019).
37. Huang, C. et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020).
38. Limjunyawong, N., Fallica, J., Horton, M. R. & Mitzner, W. Measurement of the pressure-volume curve in mouse lungs. *J. Vis. Exp*. **52376**, 52376 (2015).
39. Seeley, R. R., Stephens, T. D. & Tate, P. *Essentials of Anatomy and Physiology* 7th edn (2005).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Human lung tissue and peripheral blood

Freshly resected lung tissue was procured intraoperatively from patients undergoing lobectomy for focal lung tumours. Normal lung tissues (approximately 5 cm^3) were obtained from uninvolved regions and annotated for the specific lung lobe and location along the airway or periphery. Pathological evaluation (by G.B.) confirmed normal histology of the profiled regions, except for areas of very mild emphysema in patient 1. Patient 1 was a 75-year-old male with a remote history of smoking, diagnosed with early stage adenocarcinoma who underwent left upper lobe (LUL) lobectomy; two blocks of normal tissue were obtained from lung periphery (distal 1a and 1b). Patient 2 was a 46-year-old male, non-smoker with a right middle lobe (RML) endobronchial carcinoid, who underwent surgical resection of the right upper and middle lobes; two blocks of tissue were selected from mid-bronchial region (medial 2) and periphery (distal 2) of right upper lobe (RUL). Patient 3 was a 51-year-old female, non-smoker with mild adult-onset asthma and a left lower lobe (LLL) endobronchial typical carcinoid, who underwent LLL lobectomy; three tissue blocks were resected from the bronchus (proximal 3), mid-bronchial (medial 2), and periphery (distal 3) of the LLL. All tissues were received and immediately placed in cold PBS and transported on ice directly to the research lab for single cell dissociation procedures. Peripheral blood was collected from patients 1 and 3 in EDTA tubes. For bulk RNA-seq of canonical immune populations, whole blood from healthy human donors was obtained commericially (AllCells) in EDTA tubes. Patient tissues were obtained under a protocol approved by Stanford University's Human Subjects Research Compliance Office (IRB 15166) and informed consent was obtained from each patient before surgery. All experiments followed applicable regulations and guidelines.

Mouse lung tissue

Lung tissue for Tabula Muris Senis[40] was obtained as previously described. We obtained additional tissue from two mice expressing Cre recombinase and two expressing oestrogen-inducible Cre recombinase (Cre-ERT2) for conditional cell-specific labelling in vivo with the gene-targeted alleles FVB-*Tbx4-LME-cre*[41,42] (lung stroma) and B6.129-*Axin2-cre-ERT2*[41], respectively. Cre-dependent reporter alleles *Rosa26ZsGreen1*, which expresses cytosolic ZsGreen1 following Cre-mediated recombination, and *Rosa26mTmG*, which expresses membrane-targeted green fluorescent protein (mGFP) after recombination and membrane-targeted tdTomato (mTomato) in all other tissues, were used to label cells expressing *Tbx4* and *Axin2*, respectively[43,44]. Induction of the *Axin2-cre-ERT2* allele was done by intraperitoneal injection of tamoxifen (3 mg) once a day for three days as described[25]. All mouse experiments followed applicable regulations and guidelines and were approved by the Institutional Animal Care and Use Committee at Stanford University (Protocol 9780).

Isolation of lung and blood cells

Individual human lung samples were dissected, minced, and placed in digestion media (400 μg ml^{-1} liberase DL (Sigma 5466202001) and 100 μg ml^{-1} elastase (Worthington LS006365) in RPMI (Gibco 72400120) in a gentleMACS c-tube (Miltenyi 130-096-334). Samples were partially dissociated by running 'm_lung_01' on a gentleMACS Dissociator (Miltenyi 130-093-235), incubated on a Nutator at 37 °C for 30 min, and then dispersed to a single cell suspension by running 'm_lung_02'. Processing buffer (5% fetal bovine serum in PBS) and DNase I (100 μg ml^{-1}, Worthington LS006344) were then added and the samples rocked at 37 °C for 5 min. Samples were then placed at 4 °C for the remainder of the protocol. Cells were filtered through a 100-μm filter, pelleted (300*g*, 5 min, 4 °C), and resuspended in ACK red blood cell lysis buffer (Gibco A1049201) for 3 min, after which the buffer was inactivated by adding excess processing buffer. Cells were then filtered through a 70-μm strainer (Fisherbrand 22363548), pelleted again (300*g*, 5 min, 4 °C), and resuspended in magnetic activated cell sorting (MACS) buffer (0.5% BSA, 2 mM EDTA in PBS) with Human FcR Blocking Reagent (Miltenyi 130-059-901) to block non-specific binding of antibodies (see below).

Immune cells, including granulocytes, were isolated from peripheral blood using a high density ficoll gradient[45]. In brief, peripheral blood was diluted tenfold with FACS buffer (2% FBS in PBS), carefully layered on an RT Ficoll gradient (Sigma HISTOPAQUE-1119), and centrifuged at 400*g* for 30 min at room temperature. The buffy coat was carefully removed, diluted fivefold with FACS buffer, pelleted (300*g*, 5 min, 4 °C), and incubated in ice cold FACS buffer containing DNase I (Worthington LS006344) for 10 min at 4 °C. Clumps were separated by gentle pipetting to create a single-cell suspension.

Mouse lung samples were processed into single cell suspensions as previously described[2]. In brief, each lung was dissected, minced, and placed in gentleMACS c-tubes (Miltenyi 130-096-334) with digestion buffer (400 μg ml^{-1} liberase DL (Sigma 5466202001) in RPMI (Gibco 72400120)). The minced tissue was partially dissociated by running 'm_lung_01' on a gentleMACS Dissociator (Miltenyi 130-093-235), incubated at 37 °C on a nutator for 30 min, completely dissociated on a gentleMACS by running 'm_lung_02', and kept at 4 °C or on ice for the remainder of the protocol. Cells were washed with 5% FBS in PBS, centrifuged at 300*g* for 5 min, resuspended in 5% FBS in PBS, filtered through a 70-μm strainer (Fisherbrand 22363548), and centrifuged again and resuspended in FACS buffer (2% FBS in PBS).

Magnetic separation of lung tissue compartments

Immune and endothelial cells were overrepresented in our previous mouse single-cell suspensions. To partially deplete these populations in our human samples, we stained cells isolated from lung with MACS microbeads conjugated to CD31 and CD45 (Miltenyi 130-045-801, 130-091-935) then passed them through an LS MACS column (Miltenyi, 130-042-401) on a MidiMACS Separator magnet (Miltenyi, 130-042-302). Cells retained on the column were designated 'immune and endothelial enriched'. The flowthrough cells were then split, with 80% immunostained for FACS (see below) and the remaining 20% stained with EPCAM microbeads (Miltenyi 130-061-101). EPCAM stained cells were passed through another LS column. Cells retained on the column were labelled 'epithelial enriched', and cells that flowed through were designated 'stromal'.

Flow cytometry and cell sorting

Lysis plates for single-cell mRNA sequencing were prepared as previous described[2]. 96-well lysis plates were used for cells from the blood and mouse samples and contained 4 μl of lysis buffer instead of 0.4 μl.

After negative selection against immune and endothelial cells by MACS, the remaining human lung cells were incubated with FcR Block (Becton Dickinson 564219) for 5 min and stained with directly conjugated anti-human CD45 (Biolegend 304006) and EPCAM (eBioscience 25-9326-42) antibodies on a Nutator for 30 min at the manufacturer's recommended concentration. Cells were then pelleted (300*g*, 5 min, 4 °C), washed with FACS buffer three times, then incubated with cell viability marker Sytox blue (1:3,000, ThermoFisher S34857) and loaded onto a Sony SH800S cell sorter. Living single cells (Sytox blue-negative) were sorted into lysis plates based on three gates: EPCAM$^+$CD45$^-$ (designated epithelial), EPCAM$^-$CD45$^+$ (designated immune), and EPCAM$^-$CD45$^-$ (designated endothelial or stromal).

Immune cells from subject matched blood were incubated with FcR Block and Brilliant Violet buffer (BD 563794) for 20 min and then stained with directly conjugated anti-human CD3 (BD 563548), CD4 (BD 340443), CD8 (BD 340692), CD14 (BD 557831), CD19 (Biolegend 302234), CD47 (BD 563761), CD56 (BD 555516), and CD235a (BD 559944) antibodies for 30 min at the manufacturer's recommended concentration. Cells were pelleted (300*g*, 5 min, 4 °C), washed with FACS buffer twice, and then incubated with the viability marker propidium iodide and loaded onto a BD FACSAria II cell sorter. Living (propidium iodide-negative) single, non-red blood (CD235a$^-$) cells were sorted into lysis plates along with specific immune populations: B cells (CD19$^+$CD3$^-$), CD8+ T cells (CD8$^+$), CD4$^+$ T cells (CD4$^+$), natural killer cells (CD19$^-$CD3$^-$CD56$^+$CD14$^-$), classical monocytes (CD19$^-$CD3$^-$CD56$^-$CD14$^+$). After sorting, plates were quickly sealed, vortexed, spun down for 1 min at 1,000*g*, snap frozen on dry ice, and stored at −80 until cDNA synthesis.

Mouse cells were incubated with the viability marker DAPI and loaded onto a BD Influx cell sorter. Living (DAPI-negative) single cells were sorted into lysis plates based on presence or absence of the fluorescent lineage label (mEGFP for *Axin2-cre-ERT2*, ZsGreen1 for *Tbx4-LME-cre*).

Immune cells for bulk mRNA sequencing were incubated with FcR Block for 20 min and then stained with one of six panels of directly conjugated antibodies for 30 min at the manufacturers recommended concentration: anti-human CD16 (BD 558122), CD123 (BD 560826), CCR3 (R&D FAB155F), ITGB7 (BD 551082), CD3 (BD 555341), CD14 (Invitrogen MHCD1406), CD19 (BD 555414), and CD56 (BD 555517) (basophils, neutrophils and eosinophils); anti-human CD16 (BD 558122), CD14 (BD 347497), CD4 (BD 340443), CD3 (BD 555341), CD8 (BD 555368), CD19 (BD 555414), and CD56 (BD 555517) (classical and nonclassical monocytes); anti-human CD16 (BD 558122), CD1c (Miltenyi Biotec 130-098-007), CD11c (BD 340544), CCR3 (R&D FAB155F), CD123 (BD 560826), HLA-DR (BD 335796), CD3 (BD 555341), CD4 (BD 555348), CD8 (BD 555368), CD14 (Invitrogen MHCD1406), CD19 (BD 555414), and CD56 (BD 555517) (pDCs, mDCs, CD16$^+$ dendritic cells); anti-human IgM/IgD (BD 555778), CD19 (BD 557835), CD27 (BD 558664), CD20 (BD 335794), CD3 (BD 555341), CD4 (BD 555348), CD14 (Invitrogen MHCD1406), and CD56 (BD 555517) (B cells); anti-human CD16 (BD 558122), CD57 (BD 347393), CD56 (BD 557747), CD3 (BD 555341), CD4 (BD 555348), CD14 (Invitrogen MHCD1406), and CD19 (BD 555414) (natural killer cells); and anti-human CD45RA (Biolegend 304118), CCR7 (R&D FAB197F), CD62L (BD 555544), CD45RO (BD Pharmingen 560608), CD4 (BD 340443), CD8 (BD 340584), CD11b (BD 555389), CD14 (Invitrogen MHCD1406), CD19 (BD 555414), CD56 (BD 555517) (T cells). Cells were washed with FACS buffer twice, incubated with the viability marker propidium iodide and loaded onto a BD FACSAria II cell sorter. Approximately 40,000 cells from 21 canonical immune populations (Supplementary Table 3) were sorted in duplicate into Trizol LS (Invitrogen 10296010).

After sorting, all plates and samples were quickly sealed, vortexed, spun down for 1 min at 1,000*g* and then snap frozen on dry ice and stored at −80 °C until cDNA synthesis.

Single-cell mRNA sequencing

mRNA from single cells sorted from human and mouse lungs and human blood into lysis plates was reverse transcribed to cDNA and amplified as previously described[2]. Illumina sequencing libraries for cDNA from single cells were prepared as previously described[2]. In brief, cDNA libraries were prepared using the Nextera XT Library Sample Preparation kit (Illumina, FC-131-1096). Nextera tagmentation DNA buffer (Illumina) and Tn5 enzyme (Illumina) were added, and the sample was incubated at 55 °C for 10 min. The reaction was neutralized by adding Neutralize Tagment Buffer (Illumina) and centrifuging at room temperature at 3,220*g* for 5 min. Mouse samples were then indexed via PCR by adding i5 indexing primer, i7 indexing primer, and Nextera NPM mix (Illumina). Human samples were similarly indexed via PCR using custom, dual-unique indexing primers (IDT)[2].

Following library preparation, wells of each library plate were pooled using a Mosquito liquid handler (TTP Labtech), then purified twice using 0.7x AMPure beads (Fisher A63881). Library pool quality was assessed by capillary electrophoresis on a Tapestation system (Agilent) with either a high sensitivity or normal D5000 ScreenTape assay kit (Agilent) or Fragment analyser (AATI), and library cDNA concentrations were quantified by qPCR (Kapa Biosystems KK4923) on a CFX96 Touch Real-Time PCR Detection System (Biorad). Plate pools were normalized and combined equally to make each sequencing sample pool. A PhiX control library was spiked in at 1% before sequencing. Human libraries were sequenced on a NovaSeq 6000 (Illumina) and mouse libraries on a NextSeq 500 (Illumina).

Cells isolated from each compartment (immune and endothelial enriched, epithelial enriched, stromal) and subject blood were captured in droplet emulsions using a Chromium Single-Cell instrument (10x Genomics) and libraries were prepared using the 10x Genomics 3′ Single Cell V2 protocol as previously described[2]. All 10x libraries were pooled and sequenced on a NovaSeq 6000 (Illumina).

Immune cell bulk mRNA sequencing

Total RNA from bulk-sorted canonical immune populations was reverse transcribed to cDNA, amplified, and prepared as sequencing libraries as previously described[45]. Libraries were sequenced on a NextSeq 500 (Illumina).

Immunohistochemistry

Mouse and human lungs were collected as previously described[25,46]. After inflation, lungs were removed en bloc, fixed in 4% paraformaldehyde (PFA) overnight at 4 °C with gentle rocking, then cryo-embedded in Optimal Cutting Temperature compound (OCT, Sakura) and sectioned using a cryostat (Leica) onto Superfrost Plus Microscope Slides (Fisherbrand). Immunohistochemistry was performed using primary antibodies raised against the following antigens and used at the indicated dilutions to stain slides overnight at 4 °C: anti-proSP-C (rabbit, Chemicon AB3786, 1:250 dilution), HES1 (rabbit, Cell Signaling 11988S clone D6P2U, 1:100), MUC-1 (hamster, Thermo Scientific HM1630, clone MH1, 1:250), Ki67 (rat, DAKO M7249 clone MIB-1, 1:100), and keratin-5 (chicken, Biolegend 905901, 1:100). Primary antibodies were detected with Alexa Fluor-conjugated secondary antibodies (Jackson ImmunoResearch) unless otherwise noted, then mounted in Vectashield containing DAPI (5 µg ml^{-1}, Vector labs). Images were acquired with a laser scanning confocal fluorescence microscope (Zeiss LSM780) and processed with Fiji (v.2.0) and Imaris (v.9.2.0, Oxford Instruments). Immunostaining experiments were performed on at least two human or mouse participants distinct from the donors used for sequencing, and quantifications were based on at least 10 fields of view in each.

Single molecule in situ hybridization

Samples were fixed in either 10% neutral buffered formalin, dehydrated with ethanol and embedded in paraffin wax or fixed in 4% paraformaldehyde and embedded in OCT compound. Sections from paraffin (5 µm) and OCT (20 µm) blocks were processed using standard pre-treatment conditions for each per the RNAscope multiplex fluorescent reagent kit version 2 (Advanced Cell Diagnostics) assay protocol. TSA-plus fluorescein, Cy3 and Cy5 fluorophores were used at 1:500 dilution. Micrographs were acquired with a laser scanning confocal fluorescence microscope (Zeiss LSM780) and processed with ImageJ and Imaris (version 9.2.0, Oxford Instruments). smFISH experiments were performed on at least two human or mouse participants distinct from the donors used for sequencing, and quantifications were based on at least 10 fields of view in each. For smFISH, fields of view were scored manually, calling a cell positive for each gene probed if its nucleus had at least three associated expression puncta. Proprietary (Advanced Cell Diagnostics) probes used were: KRT5 (547901-C2), SERPINB3 (828601-C3), SFTPC (452561-C2), WIF1 (429391), CLDN5 (517141-C2, 517141-C3), MYC (311761-C3), ACKR1 (525131, 525131-C2), COL1A2 (432721), GPC3 (418091-C2), SERPINF1 (564391-C3), C20rf85 (560841-C3), DHRS9 (467261), GJA5 (471431), CCL21 (474371-C2), COX4I2 (570351-C3), APOE (433091-C2), ACGT2 (828611-C2), ASPN (404481), IGSF21 (572181-C3), GPR34 (521021), EREG (313081), GPR183 (458801-C2), TREM2 (420491-C3), CHI3L1 (408121), MYRF (499261), AGER (470121-C3), TBX5 (564041), KCNK3 (536851), ACVRL1 (559221), SERPINA1 (435441), HHIP (464811), SLC7A10 (497081-C2), FGFR4 (443511), PI16 (451311-C2), SERPINF1 (310731), HHIP (448441-C3), SFTPC (314101-C2), NKX2-1 (434721-C3), and MYRF (524061).

Sequencing read alignments and quality control

Reads from single cells isolated using 10x chromium were demultiplexed and then aligned to the GRCh38.p12 human reference (from 10x Genomics) using Cell Ranger (version 2.0, 10x Genomics). Cells with fewer than 500 genes detected or 1,000 UMIs were excluded from further analyses.

Reads from single cells isolated by flow cytometry were demultiplexed using bcl2fastq (v.2.19.0.316, Illumina), pruned for low nucleotide quality scores and adaptor sequences using skewer (v.0.2.2), and aligned to either (depending on organism) the GRCh38.p12 human reference genome with both the gencode-vH29 and NCBI-108 annotations or the GRCm38.p6 mouse reference genome with the NCBI-106 annotation (with fluorescent genes mEGFP, tdTomato, and ZsGreen1 supplemented) using STAR (v.2.6.1d) in two-pass mapping mode, in which the first pass identifies novel splice junctions and the second pass aligns reads after rebuilding the genome index with the novel junctions. The number of reads mapping to each annotated gene were calculated by STAR during the second pass alignment, and cells with fewer than 500 genes detected or 50,000 mapped reads were excluded from later analyses. Reads from mRNA sequencing of canonical immune populations were demultiplexed, aligned and quantified using the same pipeline.

Cell clustering, doublet calling, and annotation

Expression profiles of cells from different subjects and different capture approaches (10x and SS2) were clustered separately using the R software package Seurat (v.2.3)[47]. In brief, counts (SS2) and UMIs (10x) were normalized across cells, scaled per million (SS2) or per 10,000 (10x), and converted to log scale using the 'NormalizeData' function. These values were converted to z-scores using the 'ScaleData' command and highly variable genes were selected with the 'FindVariableGenes' function with a dispersion cutoff of 0.5. Principle components were calculated for these selected genes and then projected onto all other genes with the 'RunPCA' and 'ProjectPCA' commands. Clusters of similar cells were detected using the Louvain method for community detection including only biologically meaningful principle components (see below) to construct the shared nearest neighbour map and an empirically set resolution, as implemented in the 'FindClusters' function.

When clustering all cells from a single subject at once, we found that the first principal components defining heterogeneity represented differences in tissue compartment, but some cell types within a compartment (for example, basal, goblet club, neuroendocrine and ionocyte) had a tendency to co-cluster. Clusters were therefore grouped based on expression of tissue compartment markers (for example, *EPCAM*, *CLDN5*, *COL1A2* and *PTPRC*) using the 'SubsetData' command and the same procedure (from 'ScaleData' onwards) was applied iteratively to each tissue compartment until the markers enriched in identified clusters, identified using the 'MAST' statistical framework[48] implemented in the 'FindMarkers' command, were no longer biologically meaningful (for example, clusters distinguished by dissociation-induced genes[30], ribosomal genes, mitochondrial genes, or ambient RNA released by abundant cells such as RBCs[31]). Doublets were identified by searching for cells with substantial and coherent expression profiles from two or more tissue compartments and/or cell types.

To assign clusters identities, we first compiled a list of all established lung cell types, their abundances, their classical markers, and any RNA markers (when available) (Supplementary Table 1). RNA markers for canonical immune populations (Supplementary Table 3) were obtained from bulk mRNA sequencing by correlating the average expression (each captured in duplicate) with a test vector where the target population position equaled 10 and all others equaled 0 (see GitHub for details). Clusters were assigned a canonical identity based on enriched expression of these marker genes. Pearson correlations were calculated between the average expression profiles from each immune cluster for all cells in the SS2 with the average bulk profiles using the 'cor' function in R. There were no clusters that lacked expression of canonical marker genes. When two or more clusters were assigned the same identity, we first determined whether their tissue locations differed substantially (for example, proximal versus distal, alveolar versus adventitial) and prepended these locations when applicable. When both clusters localized to the same tissue region (for example, capillary endothelial cells or AT2 cells), we next compared their differentially expressed genes head-to-head to identify differences in molecular functions. These functional differences were also prepended, when applicable (for example, signalling AT2 versus AT2, proliferating basal versus basal). If the clusters could not be resolved by location or function, we prepended a representative marker gene to their 'canonical' identity (for example, IGSF21$^+$ dendritic, EREG$^+$ dendritic, and TREM2$^+$ dendritic). Cells from different subjects with the same annotation were merged into a single group for all downstream analyses.

Approximately 35,000 mouse lung and blood cell expression profiles by SS2 and 10x from Tabula Muris Senis[2] were combined with 522 cells isolated from *Axin2-Cre-ERT2>Rosa26mTmG* (A.N.N.) and *Tbx4-LME-Cre >Rosa26ZsGreen1* (K.J.T.) mice and amplified by SS2. Cells were stratified by technology (10x versus SS2), re-clustered and re-annotated using the strategy described above for human lung cells.

Re-annotation of existing human lung single cell RNA sequencing datasets

UMI tables were obtained from the Gene Expression Omnibus (GSE122960 for ref. [18], GSE130148 for ref. [19]), clustered, and annotated using the strategy described above. New annotations for each cell are available on GitHub (see below).

Cell type pairwise correlations

We obtained average expression profiles for each cell type from all cells in the 10x dataset, supplemented with the average expression profile from neutrophils in the SS2 dataset, and calculated pairwise Pearson correlation coefficients using the 'cor' function in R.

Identification of proliferation signature

Expression profiles from matched proliferating and quiescent cell types were compared head-to-head using the 'MAST' statistical framework implemented in the 'FindMarkers' command in Seurat. Differentially-expressed genes common in each proliferating cell type were converted to *z*-scores using the 'ScaleData' command in Seurat, and summed to create a proliferation score for each cell in the 10x dataset.

Identification of immune egression signatures

Blood and tissue expression profiles for each immune cell type were compared head-to-head using the 'MAST' statistical framework implemented in the 'FindMarkers' command in Seurat. Differentially-expressed genes common in each subject were screened for dissociation artefact and contamination by red blood cells. Genes specific to tissue immune cells were binned based on their breadth of expression (lymphocyte, myeloid or both), converted to *z*-scores using the 'ScaleData' command in Seurat, and summed to create an egression score for each cell in the 10x dataset.

Identification of enriched marker genes, transcription factors, and disease genes

Differentially expressed genes for each annotated cell type relative to the other cells within its tissue compartment were identified using the 'FindMarkers' command in Seurat with the 'MAST' statistical framework after downsampling each cell type to 100 (SS2) or 500 (10x) cells. To obtain the most sensitive and specific markers for each cell type, we ranked enriched genes, with a P value less than 10^{-5} and a sensitivity greater than 0.4, by their Matthews correlation coefficients (MCCs) calculated for each cell type from all cells in the 10x data set (numbers available in Supplementary Table 2). To measure the utility of using multiple markers in assigning cell identities, we calculated MCC scores for all possible combinations of each cell type's top five marker genes.

Enriched genes were annotated as transcription factors or genes associated with pulmonary pathology based on lists compiled from The Animal Transcription Factor Database (http://bioinfo.life.hust.edu.cn/AnimalTFDB), The Online Mendelian Inheritance in Man Catalog (OMIM)[49], and Genome Wide Association Studies (GWAS) obtained from the EMBL-EBI Catalog[50] (EFO IDs 0000270, 0000341, 0000464, 0000571, 0000702, 0000707, 0000708, 0000768, 0001071, 0003060, 0003106, 0004244, 0004312, 0004313, 0004314, 0004647, 0004713, 0004806, 0004829, 0005220, 0005297, 0006505, 0006953, 0007627, 0007744, 0007944, 0008431, 0009369, 0009370; GO IDs 0031427, 0097366; Orphanet IDs 586 182098; log(p-value) < -20, statistical tests vary in indicated studies). Viral entry genes were obtained from Gene Ontology (GO:0046718) and then curated and associated with their cognate virus(es) based on literature citations available in our GitHub repository.

Cellular interaction and hormone target mapping

Interactions between cell types were predicted using CellPhoneDB ('statistical_analysis' method) with all cells in the SS2 dataset, as previously described[6]. For our targeted analyses, we curated the chemokine receptor-ligand interaction map and list of hormone receptors from an extensive literature search (available on GitHub, see below).

Human and mouse gene alignment, cell type correlation, and gene expression comparisons

The gene expression matrices from our human SS2 cells and the Tabula Muris Senis SS2 cells, supplemented with the 522 mouse cells from *Axin2-creER > mTmG* and *Tbx4-Cre > ZsGreen1* described above, were collapsed to HomologyIDs obtained from the Mouse Genome Informatics database to enable direct comparison. We obtained mean expression profiles for each cell type from all cells in the SS2 dataset and calculated pairwise Pearson correlation coefficients using the 'cor' function in R. We defined species-specific gene expression as those enriched 20-fold in either direction (mouse > human or human > mouse) with a P value less than 10^{-5} (calculated by 'MAST' as above) from all cells for the indicated types in the SS2 dataset. Correlations and age-specific genes were obtained the same manner using all cells from 3-month and 24-month in the combined SS2 mouse dat set.

To compare the expression pattern of each gene across species we binarized genes as expressed (1) or not expressed (0) in each cell type's average expression profile calculated from all mouse and human SS2 cells of the types compared above. A cell type 'expressed' a gene if the median of that gene's non-zero expression values across the constituent cells was greater than the median of every non-zero expression value for all other genes plus or minus two standard deviations (varied in 0.25 increments) and if the percentage of cells within the cell type with non-zero expression values was greater than the median percent of non-zero expression values for all other genes plus or minus two standard decisions (varied in 0.25 increments). These cutoffs were varied independently to ensure genes were robustly categorized. We then ordered these gene vectors to match homologous cell types between species with at least five cells and combined them to a single vector for each gene ($V = (a - b) + 2ab$, in which a is the ordered human vector and b is the ordered mouse vector) that indicated for each cell type whether: Both mouse and human expressed the gene (2), only human (1), only mouse (−1), or neither (0). We then classified genes by the following: conserved if any element of V equaled 2 and all other elements equaled 0; type 2 if any element equaled 2 and any other equaled 1 or −1; not expressed if all elements equaled 0; type 3 if elements were both positive and negative; and type 1 if elements were either positive or negative and 0.

Statistics and reproducibility

All heat maps and plots with single cell expression data include every cell from indicated types (numbers available in Supplementary Table 2 for human and Supplementary Table 6 for mouse) for sequencing technology specified (SS2 or 10x), unless otherwise stated. Scatter plots were generated with ggplot2's 'geom_point' function. Dot plots were generated using a modified version of Seurat's 'DotPlot' function (available on GitHub). Violin plots were created with Seurat's 'VlnPlot' function and show proportion of single cells at indicated expression levels. Box-and-whisker plots were generated with ggplot2's 'geom_boxplot' function; lower and upper hinges correspond to first and third quartiles, whiskers extend from hinge to the largest or smallest value no further than 1.5 times the interquartile range. Data beyond whiskers are shown as outlying points. Correlations use Pearson's coefficient. Differentially expressed genes were identified using the 'MAST' statistical framework[48] implemented in Seurat's 'FindMarkers' function. Immunostaining and smFISH experiments were performed on at least 2 human or mouse subjects distinct from the donors used for sequencing, and quantifications were based on at least 10 fields of view in each. For smFISH, fields of view were scored manually, calling a cell positive for each gene probed if its nucleus had at least three associated expression puncta. No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Counts/UMI tables, cellular metadata, Seurat objects, and scanpy objects are available on Synapse (https://www.synapse.org/#!Synapse:syn21041850). The data can be explored in a browser using cellxgene at https://hlca.ds.czbiohub.org/. Human sequencing data are available by data access agreement on the European Genome-phenome Archive (EGA) under accession EGAS00001004344. Use of human sequencing data are restricted to not for profit research only and requires approval or a waiver from requesting investigator's institutional review board. Mouse sequencing data are available on the National Institute of Health's Sequence Read Archive (SRA) under BioProject accession PRJNA632939. Source data are provided with this paper.

Code availability

The code for demultiplexing counts/UMI tables, clustering, annotation, downstream analyses, and obtaining source data/generating figures that include single-cell expression data are available on GitHub (https://github.com/krasnowlab/HLCA).

40. Tabula Muris Consortium. A single cell transcriptomic atlas characterizes aging tissues in the mouse. *Nature* **583**, 590–595 (2019).
41. van Amerongen, R., Bowman, A. N. & Nusse, R. Developmental stage and time dictate the fate of Wnt/β-catenin-responsive stem cells in the mammary gland. *Cell Stem Cell* **11**, 387–400 (2012).
42. Greif, D. M. et al. Radial construction of an arterial wall. *Dev. Cell* **23**, 482–493 (2012).
43. Muzumdar, M. D., Tasic, B., Miyamichi, K., Li, L. & Luo, L. A global double-fluorescent Cre reporter mouse. *Genesis* **45**, 593–605 (2007).
44. Madisen, L. et al. A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. *Nat. Neurosci.* **13**, 133–140 (2010).
45. Moraga, I. et al. Tuning cytokine receptor signaling by re-orienting dimer geometry with surrogate ligands. *Cell* **160**, 1196–1208 (2015).
46. Desai, T. J., Brownfield, D. G. & Krasnow, M. A. Alveolar progenitor and stem cells in lung development, renewal and cancer. *Nature* **507**, 190–194 (2014).

47. Butler, A. et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol*. **36**, 411–420 (2018).

48. Finak, G. et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol*. **16**, 278 (2015).

49. Amberger, J. S. et al. OMIM.org: Online Mendelian Inheritance in Man (OMIM), an online catalog of human genes and genetic disorders. *Nucleic Acids Res*. **43**, D789–D798 (2014).

50. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. **47** (D1), D1005–D1012 (2019).

Acknowledgements We are grateful to the tissue donors and the clinical staff at Stanford Medical Center who made tissue collection possible, especially J. Benson and E. Chen. We are especially grateful to Jim Spudich who spurred this study. We also thank the Stanford Shared FACS Facility for their expertise and sorting services, especially L. Nichols and M. Weglarz; members of Chan Zuckerberg Biohub and Quake laboratory who supported this work, particularly A. McGeever, B.Yu, B. Jones and S. Kolluru; M. Kumar for discussions on annotation of stromal cells; and M. Petersen for illustrating the lung schematic (Fig. 1b) and C. Kao for help with figure formatting. Some computing for this project was performed on the Sherlock cluster; we thank Stanford University and the Stanford Research Computing Center for providing computational resources and support that contributed to the results. We thank J. Spudich and members of the Krasnow laboratory for discussions and comments on the manuscript, and A. Lozano for discussions on bioinformatic analyses. This work was supported by funding from the Chan Zuckerberg Biohub (S.R.Q.), the Howard Hughes Medical Institute, National Institutes of Health, and the Vera Moulton Wall Center for Pulmonary Vascular Disease (M.A.K.), and the Ludwig Cancer Center at Stanford (I.L.W.). K.J.T was supported by a Paul and Mildred Berg Stanford Graduate Fellowship. M.A.K. is an investigator of the Howard Hughes Medical Institute.

Author contributions K.J.T., A.N.N., L.P., R.S., A.G., C.S.K., R.J.M. and M.A.K. conceived the project and designed the lung and blood cell isolation strategy, J.B.S. and C.S.K. designed clinical protocols, reviewed clinical histories and coordinated patient care teams to obtain profiled tissues, G.B. provided expert clinical evaluation and micrographs of donor tissue histology, K.J.T., A.N.N., R.S. and A.G. processed tissue to single-cell suspensions, K.J.T., A.N.N., L.P. A.G., R.S. and S.D.C. sorted cells for SS2, A.N.N., L.P., S.C. and R.V.S. prepared sequencing libraries, and K.J.T., R.V.S. and L.P. processed and aligned sequencing data. R.S., J.S. and Y.M. performed and supervised bulk mRNA sequencing on defined immune populations. K.J.T., A.N.N., R.S. A.G. and R.J.M. provided tissue expertise and annotated cell types. K.J.T., A.N.N. and M.A.K. designed and implemented bioinformatic methods and interpreted results. K.J.T., A.N.N. and A.G. performed follow up stains. M.A.K., S.R.Q., N.F.N., I.L.W., C.S.K. and R.J.M. supervised and supported the work. K.J.T., A.N.N. and M.A.K. wrote the manuscript, and all authors reviewed and edited the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2922-4.

Correspondence and requests for materials should be addressed to S.R.Q. or M.A.K.

Peer review information *Nature* thanks Shalev Itzkovitz and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.