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

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

[1]Department of Biochemistry and Howard Hughes Medical Institute,
[2]Institute for Stem Cell Biology and Regenerative Medicine,
[3]Department of Pathology,
[4]Department of Cardiothoracic Surgery,
[5]Vera Moulton Wall Center for Pulmonary Vascular Disease,
[6]Department of Pediatrics, Division of Cardiology,
[7]Department of Pediatrics, Pulmonary Medicine,
[8]Ludwig Center for Cancer Stem Cell Research and Medicine,
[9]Stanford Cancer Institute,
Stanford University School of Medicine, Stanford, CA, USA.

[10]Chan Zuckerberg Biohub, San Francisco, CA, USA.

[11]Department of Bioengineering, Stanford University, Stanford, CA, USA.

* These authors contributed equally to this work and will list themselves first on their CVs.
** Corresponding authors. Send correspondence to M.A.K. (krasnow@stanford.edu; 650-723-7191 (phone); 650-723-6783 (fax)) and S.R.Q. (steve@czbiohub.org)

†Present address: Calico Life Sciences, South San Francisco, CA USA (L.P.). Department of Medicine and Biosystemic Science, Kyushu University Graduate School of Medical Science, Fukuoka, Japan (Y.S.). Medical Sciences Innovation Hub Program, RIKEN, Japan (J.S.).

Abstract

Although single cell RNA sequencing studies have begun providing compendia of cell expression profiles, it has proven more difficult to systematically identify and localize all molecular cell types in individual organs to create a full molecular cell atlas. Here we describe droplet- and plate-based single cell RNA sequencing applied to ~75,000 human lung and blood cells, combined with a multi-pronged cell annotation approach, which have allowed us to define the gene expression profiles and anatomical locations of 58 cell populations in the human lung, including 41 of 45 previously known cell types or subtypes and 14 new ones. This comprehensive molecular atlas elucidates the biochemical functions of lung cell types and the cell-selective transcription factors and optimal markers for making and monitoring them; defines the cell targets of circulating hormones and predicts local signaling interactions including sources and targets of chemokines in immune cell trafficking and expression changes on lung homing; and identifies the cell types directly affected by lung disease genes and respiratory viruses. Comparison to mouse identified 17 molecular types that appear to have been gained or lost during lung evolution and others whose expression profiles have been substantially altered, 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.

Introduction

Over the past two centuries, hundreds of human cell types have been discovered, categorized, and studied by microscopy, creating the classical atlases that provide the cellular foundation for modern medicine[1-5]. In the past several decades, cell-type-specific marker genes have been identified that supplement the histological descriptions and provide molecular definitions and functions of the cell types[6-9]. This reached its apex in the systematic profiling studies that elucidate genome-wide expression profiles of purified cell populations and, more recently, of individual cells by single cell RNA sequencing (scRNAseq)[10-14].

Although initial scRNAseq studies focused on specific cell types, tissue compartments, and biological processes, large scale molecular cell atlases of organs and even whole organisms are now possible because of improvements in throughput and cost that allow expression profiling of thousands of individual cells, without cell purification or prior knowledge of cell identity to obtain expression profiles of at least the most abundant and easy to isolate cell types[15-27]. Perhaps the greatest current challenge is obtaining profiles of rare and fragile cell types, or even just assessing the “completeness” of a molecular cell atlas. Indeed, most scRNAseq efforts to date have left the identities and tissue locations of many transcriptionally-distinct cell populations (computationally-defined cell "clusters") uncertain, obscuring their relationship to the classical, histologically-defined cell types and leaving open the possibility that some represent previously unrecognized cell types, subtypes or cell states. A complete molecular cell atlas could identify new cell types and new biochemical functions and interactions of known cell types, and would provide the molecular foundation for investigating how cells are specified and how they are altered in disease and evolution.

We set out to create a comprehensive molecular cell atlas of the adult human lung, both for its basic science value and potential clinical applications since pulmonary diseases and

infections are among the leading causes of morbidity and mortality worldwide[28,29]. This is a substantial technical challenge because the lung is comprised of 45 histologically-defined cell types with diverse structures and functions that vary in abundance over five orders of magnitude[30] (Table S1). Here we describe a systematic single cell RNA sequencing approach toward capturing, annotating, and analyzing the expression profiles of all lung cell types. We do so by using fresh blood and lung tissue obtained intraoperatively, balancing the abundance across the major tissue compartments of the 75,000 cells analyzed, employing both broad cell capture (droplet-based) and deep expression (plate-based) scRNAseq profiling strategies, and using the distinguishing molecular features along with ascertained spatial information of the novel and ambiguous cell clusters to assign cellular identities. We identify 41 of 45 previously known lung cell types and subtypes including all but the exceedingly rare ones, plus 14 new ones. We show how this comprehensive lung cell atlas provides novel insights into the functions, regulation, and interactions of the known as well as the new cell types; into which cell types are targeted in disease and how each type can be specified in development and tissue engineering; and into organ evolution by revealing lung cell types that have been gained, lost, and altered from mouse to human. The atlas provides a benchmark for analysis of diseased lung tissue, and the approach can be readily applied to other organs to create a human molecular cell atlas.

Results

Single cell RNA sequencing of ~75,000 cells identifies 58 human lung cell populations

To create a comprehensive atlas of the human lung, we set out to capture the full diversity of cell types in its four major tissue compartments (epithelial, endothelial, stromal, immune), and across the proximal-to-distal axis of the lung. We acquired histologically normal lung tissue samples intraoperatively from bronchi (proximal), bronchiole (medial), and alveolar (distal) regions (Fig. S1) along with samples of peripheral blood and immediately began processing (Fig. 1a). Lung samples were independently dissociated into single cell suspensions, and each lung cell suspension was then separated into epithelial ($EPCAM^+$), endothelial/immune ($CD31^+/CD45^+$) and stromal ($EPCAM^-$, $CD31^-/CD45^-$) populations by fluorescence-activated cell sorting (FACS) or magnetic-assisted cell sorting (MACS) (Fig. S2a). This allowed us to balance compartmental representation for sequencing, which for lung is otherwise dominated by immune and endothelial cells. Index sorting further provided surface antigen levels and other cell parameters for the FACS-sorted cells. Some of the blood cells were also index sorted by FACS to balance representation of the major immune cell lineages (Figs. 1a, S2b). Sequencing libraries were prepared from MACS-sorted lung cell populations and unsorted blood cells using droplet-based 10x Chromium (10x) protocol, or from FACS-sorted individual lung and blood cells using SmartSeq2 (SS2) protocol[31]. The higher cell throughput and lower cost of 10x enabled discovery of rare cell types, whereas SS2 gave deeper transcriptomic information that aided cell classification and detection of genes expressed at low levels such as transcription factor and receptor genes. There were also platform-specific idiosyncrasies; for example alveolar macrophages were the most abundant cell type detected in 10x samples but rare in SS2 samples, whereas neutrophils were represented only in SS2 samples. We sequenced thousands of cells

from each compartment for each subject (Table S2) to directly compare as many cell types as possible without batch correction, and we did so for three subjects (two males, one female) to address individual differences (see below). High quality transcriptomes were obtained from ~75,000 cells, 65,662 using 10x and 9,404 using SS2 (Fig. 1a).

Iterative, graph-based clustering was used to identify transcriptionally distinct clusters among cells with high quality transcriptomes separately for each subject[32]. 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 (e.g., basal, goblet club, neuroendocrine and ionocyte) had a tendency to co-cluster. We therefore initially grouped cells by tissue compartment based on expression of canonical compartment-specific marker genes (Fig. 1b) and then separately clustered cells within each compartment for each subject. Homologous clusters between subjects were subsequently established based on cluster-specific marker genes then merged for downstream analyses; batch correction algorithms were not needed because of centralized and robotic library preparation. Our approach identified 58 transcriptionally distinct cell populations (mean 51 per subject, Fig. 1c, Table S2). A state-of-the-art scRNAseq analysis of human lung published while this paper was in preparation identified 21 (36%) of the 58 lung cell populations described here[33].

Transcriptomes of nearly all of the 45 previously known human lung cell types

The 58 identified cell populations included 15 epithelial (clusters 1-15), 9 endothelial (16-24), 9 stromal (25-33), and 25 immune (34-58) populations, greater than the number of classical cell types in each compartment (11 epithelial, 5 endothelial, 7 stromal, 20 immune, and 2 neuronal cell types) (Table S2). Using extant markers for the classical cell types and/or the homologous cell types in mice (Table S1) and single molecule fluorescence *in situ* hybridization,

we identified cell clusters representing nearly all of the classical lung cell types in the epithelial (club, ciliated, basal, goblet, mucous, serous, ionocyte, neuroendocrine, alveolar type 1, alveolar type 2), endothelial (pulmonary artery, vein, capillary, lymphatic), and stromal (airway smooth muscle, vascular smooth muscle, fibroblast, myofibroblast, lipofibroblast, pericyte, mesothelium) compartments (Fig. 2a,b). No markers were known for bronchial vessels, but *in situ* staining for cluster-specific markers subsequently identified bronchial endothelial clusters (see below). The only cell type not captured was epithelial tuft cells, which are rare or absent in the normal lung[34] so was not expected to be found (Table S1). However, seven of the 23 assigned classical cell types in these three compartments were surprisingly represented by more than one cell cluster (ciliated, basal, alveolar type 2, capillary, bronchial vessel, myofibroblast, and fibroblast), revealing molecular diversity beyond the established lung cell types, as detailed below.

Immune cells were the most heterogeneous compartment presumably because of the critical role of the lung as first responder to inhaled toxins and pathogens, and because the analyzed lung samples were not perfused so included circulating and egressed as well as lung resident immune cells. To aid assignment of identities to the immune cell clusters, we first defined the transcriptional profiles of circulating immune cells by sorting human blood cells using established surface markers followed by bulk RNA sequencing of the sorted cell populations. This defined the transcriptional profiles of 21 functionally-characterized classes of circulating immune cells (Fig. S5a, Table S3). We also obtained scRNAseq profiles of ~5000 circulating blood cells from two of the subjects whose lung cells we analyzed. Canonical immune cell markers along with the ascertained panels of differentially-expressed genes allowed us to assign identities to 25 molecularly distinct immune cell clusters obtained in our scRNAseq

analysis of the human lung samples, including all of the 20 previously known lung immune cell types except eosinophils (Fig. 3a, S5b, see below).

Thus, our approach provides genome-wide expression profiles for nearly all the classical lung cell types (41 of 45, 91%), from the most abundant cell types (e.g. capillary endothelial cells, ~23% of lung cells) down to exceedingly rare ones, with estimated abundances as low as 0.01% (e.g., neuroendocrine cells, ionocytes) (Table S1). One-quarter of the classical cell types (11 of 45, 24%) previously lacked high quality single cell expression data, either because they had never been profiled or correctly annotated (mucous, capillary, bronchial vessel, airway smooth muscle, lipofibroblast, myofibroblast) or had only been profiled after culturing (artery, pericyte, mesothelium, platelet/megakaryocyte, and intermediate monocyte) (Table S1). The only canonical cell types not captured by our approach are extremely rare and primarily found in disease settings (tuft cells) or have structures or features that require special isolation or enrichment methods (intrinsic neurons, Schwann cells, eosinophils). This comprehensive dataset of cell expression profiles suggests novel functions, signaling interactions, and contributions to disease for many canonical cell types, such as a role for lung pericytes in regulation of microvascular tone through circulating hormones and as a culprit in pulmonary hypertension (see below).

Discovery of new molecularly-defined human lung cell types, subtypes and states

Although specific clusters were identified for nearly all canonical cell types, many cell types were surprisingly represented by more than one cluster, so the specific identities of 25 clusters remained uncertain. Most of the extra clusters were unlikely to have arisen from batch effects or subject-specific pathology because all except three clusters were found in more than one subject by scRNAseq or in situ hybridization (Table S2). This suggested that the distinct

expression profiles uncovered for these cell types might represent discrete molecular states of the known cell types or previously unrecognized cell types or subtypes. To begin to distinguish these possibilities, we analyzed the differentially-expressed genes among clusters assigned to each canonical cell type to identify cluster-specific biochemical and biological functions and examined the cells' structure and spatial distribution in the lung.

We first identified clusters representing common cell states. Three of the clusters (cluster 7, Bas-p; 42, NK/T-p; 48, MP-p) were significantly enriched in expression of cell cycle genes relative to the other cluster(s) of the same cell type, indicating that these represent the proliferative states of basal cells, NK cells, T cells, and macrophages, respectively, and these are the most proliferative cell types in the adult lung (Fig. 2c, S3a). Approximately 4% of basal cells were in the proliferative cluster (cluster 7, 47 cells), most of which were isolated from proximal lung samples (Fig. S3b); this suggests that proximal basal cells are more proliferative than distal ones, consistent with immunostains for proliferation marker KI67. (Fig. S4a). Another cluster (cluster 6, Bas-d, 308 cells), comprising 28% of all basal cells, had reduced expression of *KRT5* and increased expression of *HES1*, *KRT7*, and *SCGB3A2*, indicating that they were differentiating to other epithelial fates[35,36], consistent with their transitional morphology (Fig. 2c, S4b). Differentiating basal cells also derived mostly from proximal lung samples (Fig. S3b, S4b). These expression patterns and immunostaining suggest that a surprisingly[37] large proportion of human basal cells are proliferating or differentiating, especially in the proximal airways where approximately one-third of isolated basal cells appear active.

The other basal cell clusters corresponded to quiescent basal cells and localized by smFISH and isolation site either to exclusively proximal (large, pseudostratified) airways (cluster 5, Bas-px), or to both proximal and distal (small, simple) airways (cluster 4, Bas) (Fig. 2d, S3b,

S4c). Despite their similar morphology, Bas-px and Bas basal cells are distinguished by hundreds of genes, suggesting they are molecularly distinct cell types or subtypes that differ in hormone production (*ALOX15*, *ADH7*, *SNCA*) and adhesion (*POSTN*, *ISLR*, *PCDH7*), and, based on morphology and immunostaining for proliferation and differentiation markers, perhaps their ability to function as stem cells (Fig. 2c). We also found distinct clusters (clusters 2, 3) and associated molecular signatures along the proximal-distal axis for ciliated cells, indicating additional heterogeneity among epithelial cells along this axis, consistent with a recent study[33] (Fig. S3b,c). smFISH of *DHRS9*, restricted in expression to proximal ciliated cells, localized this population to larger pseudostratified airways (Fig. S4d).

We uncovered two clusters (14, 15) of alveolar type 2 (AT2) cells (Fig. 2e), which produce surfactant that prevents alveolar collapse. These were intermingled throughout the alveolar epithelium, as shown by in situ localization of cluster-specific markers (Fig. 2f). One of the clusters (cluster 14, marked by *WIF1*, *HHIP*, *CA2*) expressed higher levels of some canonical AT2 markers (e.g., surfactant genes *SFTPA1* and *SFTPC* and transcription factor *ETV5*), and also selectively expressed inhibitors of Wnt (*WIF1*) and Hedgehog (*HHIP*) signaling as well as the cell cycle (*CDKN1A*), implying the cells are quiescent (Fig. S3d, left). The other, 10-fold less abundant cluster (cluster 15, marked by *CP* and lacking *WIF1* and *HHIP*) selectively expressed detoxification genes (*CP*, *GSTA1*, *CYP4B1*) as well as Wnt pathway genes including a ligand (*WNT5A*), co-receptor (*LRP5*), regulatory protein (*CTNNBIP1*), and canonical transcription factor (*TCF4/TCF7L2*) (Fig. S3d, right); we refer to these as AT2-signaling (AT2-s). AT2-s could be alveolar stem cells, homologous to the rare subpopulation of Wnt-active AT2 cells recently identified in mouse ($AT2^{stem}$)[38,39], whereas cluster 15 could be classical AT2 cells, homologous to "bulk" AT2 cells of mouse. However, the homology and possible functional

similarity between human AT2-s and mouse AT2^{stem} is provisional because although both show elevated Wnt signaling or components, the many other observed differences between human AT2-s and "bulk" AT2 are not shared by mouse AT2^{stem}.

We also found unexpected molecular diversity in the endothelial compartment beyond the five canonical endothelial cell types (artery, vein, capillary, bronchial vessel, lymphatic). Two populations (clusters 22, 23) were identified as bronchial endothelial cells (Bro1, Bro2) by their localization around bronchi by in situ hybridization for the cluster-specific marker genes *MYC* and *ACKR1* as well as panendothelial marker *CLDN5* (Fig. 2g,h; S4h). This indicates that bronchial endothelial cells are molecularly distinct cell types from their counterparts in the pulmonary circulation and lymphatics, distinguished by expression of extracellular matrix (*VWA1*, *HSPG2*), fenestrated morphology[40] (*PLVAP*) and cell cycle associated (*MYC*, *HBEGF*) genes (Fig. 2g). Four clusters of endothelial cells in the pulmonary circulation expressed capillary markers. Two of these (clusters 18, 19) are a pair of intermingled and intercommunicating but not interconverting alveolar capillary cell types (Cap-a, Cap), as shown by cluster-specific lineage-labeling and mapping in mice (Gillich et al); the other two (clusters 20, 21) are rare cell types showing mixed features of Cap-a and Cap cells that we call capillary-intermediate type 1 (Cap-i1) and 2 (Cap-i2).

We also identified new cell types or subtypes in the stroma, the least characterized lung compartment. Two cell clusters (29, 30) expressed classical fibroblast markers (*COL1A1*, *COL1A2)* (Fig. 2i) but RNA *in situ* hybridization showed one population (cluster 30, marked by *SPINT2*, *FGFR4*, *GPC3*) localized to the alveolus ("alveolar fibroblasts") in human and mouse (Fig. 2j,left; S4k,l) and the other (cluster 29, marked by *SFRP2, PI16*, *SERPINF1*) to vascular adventitia and nearby airways ("adventitial fibroblasts") in human and mouse (Fig. 2j,right;

S4m). Both express genes involved in canonical fibroblast functions including extracellular matrix biosynthesis, cell adhesion, and angiogenesis regulators and other intercellular signals and modulators. However the specific genes for these functions often differ, for example adventitial fibroblasts expressed different Wnt pathway modulators (*SFRP2*) than those of alveolar fibroblasts (*NKD1*, *RSPO1*) (Fig. S3e). Each cluster also appears to have distinct functions: specific expression of voltage-gated sodium channel *SCN7A* and glutamate receptor *GRIA1* suggest alveolar fibroblasts are excitable cells that receive glutamatergic input (Table S4). The expression profiles also suggest novel, shared immune functions including immune cell recruitment (*IL1RL1*, *IL32*, *CXCL2*, *MHCII*) and the complement system (*C2*, *C3*, *C7*, *CFI*, *CFD*, *CFH*, *CFB*) (Table S4).

Two other stromal clusters (clusters 27 and 28) were enriched for *ACTA2*, a canonical marker of myofibroblasts (Fig. 2i), cells that help form alveoli during development and can act inappropriately later in disease. One of these (cluster 27, marked by *WIF1*, *FGF18*, *ASPN*) are classical myofibroblasts, which were isolated from distal lung samples and localized to alveolar ducts (Fig. S4o). The other population (cluster 28) showed higher expression of contractile genes (*MYH11*, *CNN1*, *TAGLN*), was preferentially isolated from proximal lung samples, and was found both intermingled with airway smooth muscle and in alveoli (Fig. S3b,S4n), so we call them “fibromyocytes.” Both populations shared expression of genes for canonical fibroblast functions (e.g. ECM genes *VCAN*, *COL12A1*, *CLU*), though the specific genes differed from both alveolar and adventitial fibroblasts, and they also both expressed a rich set of genes associated with TGF-beta signaling (*LTBP1*, *LTBP2*, *ASPN*, *DPT*, *TGFBR3*, *TGFBI*, *SCX*, *MDFI*) (Table S4).

Profiles of lung immune cells and blood cells identify lung homing and residency signatures

To distinguish lung resident and itinerant immune cell clusters from intravascular cells circulating through the lung (Fig. 3a), we first compared the relative abundance of each immune cell population in lung samples and matched peripheral blood from the same subject (Fig. 3b). Eleven of the clusters were comprised of cells only from lung samples, with no or only rare exception (clusters 40, 44, 45, 47, 48, 50, 52, 53, 54, 56, 58). This indicated that not only are alveolar macrophages (clusters 47, 48), myeloid dendritic type 1 cells (50), and basophil/mast cells (44, 45) lung resident[41] or greatly enriched in the lung interstitium or lumen as expected, but surprisingly so are natural killer T cells (40) and intermediate monocytes (58). We also discovered four novel lung myeloid populations (52, 53, 54, 56) that are lung resident or greatly enriched. These appear to have specialized functions and localizations: IGSF21+ dendritic cells express genes implicated in asthma (*CCL2*, *CCL13, IGSF21*) and localize proximal vessels, rare EREG+ dendritic cells are enriched in developmental signals (*EREG*, *VEGFA*, *AREG*) and also localize to proximal vessels, and TREM2+ dendritic cells express lipid handling machinery (*APOC1*, *APOE*, *CYP27A1*) and localize to both vessels and alveoli (Fig. 3c, S5d-g).

The other immune cell types were found in both the lung and blood samples. For some cell types, every cell—whether from blood or lung— clustered together. This included B cells (cluster 34), plasma cells (35), platelets/megakaryocytes (46), plasmacytoid dendritic cells (49), and neutrophils (43). For other cell types (every T cell subset, natural killer cells, myeloid dendritic type 2 cells, and both classical and nonclassical monocytes), the cells from lung formed a separate cluster from those from the blood. Some of the differentially-expressed genes may be due to technical differences (e.g., collagenase treatment of lung but not blood[42] or circulating RNA in blood[43]), but other detected differences such as upregulation in lung cells of lymphocyte-residence gene *CD69* likely represent genes induced following egression[44]. We

identified a core transcriptional signature for all human lung resident lymphocytes (e.g. *CD69*, *LMNA*, *RGCC*), which partially overlaps with a residence signature recently defined by bulk RNAseq of CD8+ T cells in the mouse spleen, gut and liver[45] (Fig. 3d). We also discovered a residency signature for lung myeloid cells (e.g. *AREG*, *MPHOSPH6*, *HBEGF*). This signature partially overlaps with the one for lymphocytes, supporting a core residency program for immune cells as well as specific subprograms for myeloid cells and lymphocytes.

Lung cell type markers, transcription factors, hormone targets, and signaling interactions

Our nearly complete gene expression atlas for the adult human lung has important implications for medicine. We identified the most selective marker genes for each of the previously known and newly identified cell types (Fig. 4a, Table S4). A battery of ~200 markers can unambiguously distinguish virtually all lung cell types (Fig. 4b), so could be used with multiplex in situ hybridization methods[46-48] to simultaneously detect in clinical specimens any alterations in their numbers, structures, and spatial relationships, ushering in a new era of diagnostic pathology. A similar compendium of cell-type-selective membrane proteins (Table S4) could be used to purify or therapeutically target specific lung cell types. We also identified the transcription factors selectively enriched in each cell type (Fig. 4e, Table S4), putative "master regulators" that can be used to create each cell type by cellular reprogramming, and whose efficacy in reprogramming can be evaluated with the markers described above. We found ~400 such transcription factors (1 to 36 for nearly every cell type in our SS2 data), over half of which are novel (Fig. 4e, red text). This includes what may be the long-sought master regulators (such as *MYRF*) of alveolar epithelial type 1 (AT1) cells, which comprise nearly all of the lung's gas exchange surface and are impacted in many important lung diseases, and a regulator of pericyte identity (*TBX5*). We confirmed specific expression of *MYRF* in AT1 cells and *TBX5* in

pericytes by single molecule *in situ* hybridization (smFISH) (Fig. 4c,d). These ~400 cell-type selective transcription factors are excellent candidates for creating differentiation protocols for nearly all the cell types needed to engineer a human lung.

The expression atlas allowed us to map the direct cell targets in the lung of circulating hormones, based on expression patterns of their cognate receptors. Receptors for some hormones (e.g., glucocorticoid receptor *NR3C1*, insulin receptor *INSR*, IGF receptor *IGF1R*, and adiponectin receptors *ADIPOR1* and *ADIPOR2*; all related to cell growth) are broadly expressed, implying direct action of these hormones throughout the lung, with some notable exceptions such as lymphocytes (Fig. 5a). Other hormones have highly specific and unexpected targets, such as somatostatin (*SSTR1* expressed only in arteries), mineralocorticoids (*NR3C2*, goblet cells), parathyroid hormone (*PTH1R*, smooth muscle, pericytes), ghrelin (*GHSR*, neuroendocrine cells), melanocortin (ionocytes and lipofibroblasts), and gastric inhibitory peptide (*GIPR*) and oxytocin (*OXTR*) (ciliated cells). Pericytes are predicted targets of angiotensin, endothelin, parathyroid hormone, and prostacyclin, which could regulate their contractile machinery to tune alveolar perfusion (Fig. 5b) similar to vascular smooth muscle control of arterial flow. The receptor genes for half of all circulating hormones (e.g. opioids, human growth hormone, and glucagon) were not detectably expressed in any lung cell type so may not directly influence lung physiology.

We next used the expression map to predict other signaling interactions among lung cells by identifying complementary expression patterns of ligands and their receptors among cells in the airways and accompanying vessels and among alveolar cells using CellPhoneDB[25] (see Methods). This analysis predicts diverse sets of up to hundreds of interactions of each lung cell type and its neighbors, with stromal cells the leading interactors largely due to growth factor-mediated signaling interactions (Fig. S6).

Expression of chemokine receptors provides insight into immune cell homing in the lung (Fig. 5c). Our data confirmed canonical homing interactions such as expression of *CCR7* by CD4+ T cells guiding them to *CCL21*-producing lymphatic vessels (green in Fig. 5c), and provides specificity for others such as *CCR10*-mediated plasma cell homing to epithelial mucosa through *CCL28* from serous cells in submucosal glands (gold). It also predicts new interactions such as *CXCR3*-mediated homing of pDCs to *CXCL9*-expressing TREM2+ dendritic cells (blue), and *CX3CR1*-mediated homing of nonclassical monocytes to *CX3CL1*-expressing endothelial cells and airway epithelial cells (purple). The novel dendritic populations do not express *CCR6* and *CCR7*, which guide myeloid dendritic cells to *CCL21*-expressing lymphatic vessels, suggesting they do not egress via lymphatics. However, all three express *CCR1* (orange), which could mediate their attraction to veins that express ligand *CCL23*, as well as to bronchial endothelial cells (*CCL14*), ciliated cells (*CCL15*), and lymphocytes (*CCL5*). We also identified ionocytes (bold, red), the main source of *Cftr* expression in the mouse airway, as the only non-immune cell to express appreciable levels of any chemokine receptor *(CXCR4)*, whose ligand (CXCL12) is curiously expressed almost exclusively by non-immune cells.

Expression map of human lung disease genes

We used the lung atlas to map the cellular sites of expression of the 233 extant genes that cause or contribute to human lung disease, curated from the Online Mendelian Inheritance in Man (OMIM) database[49] (179 genes) and GWAS studies[50] (54 genes, -log(p-value) > 20, see Methods). Disease genes that showed cell-type-specific expression within the lung are of special interest (Fig. 6a), because they can pinpoint the cell type(s) in which the disease originates, especially significant for the many lung diseases whose cellular origins are poorly understood. This supported the known or suspected 'culprit' cells for 27 genes related to 12 diseases, such as

arteries in pulmonary hypertension (*SMAD9*, *BMPR2*, *CAV1*; Fig. 6b), ciliated cells in ciliary dyskinesia (*CCDC151*), and capillaries in alveolar capillary dysplasia (*FOXF1*). It also identified potential novel culprit and contributing cells for 21 genes implicated in 15 diseases, including pericytes in pulmonary hypertension (potassium channel *KCNK3*), capillaries in atrioventricular dysplasia (*ACVRL1*), and AT2 cells in COPD (*SERPINA1*, *HMOX1*, *HHIP*). We confirmed pericyte, capillary, and AT2 expression of disease genes by smFISH (Fig. 6c-e).

We also mapped lung expression of 80 genes encoding proteins implicated as cellular receptors for human viruses (Fig. S8a), including 26 used by airborne viruses that enter primarily through epithelial cells lining the respiratory tract (Fig. 6b). *NECTIN4*, a receptor for measles virus was enriched in club, ciliated, Bas-D, and goblet cells, and *CDHR3*, the receptor for common-cold Rhinovirus C was enriched in ciliated and neuroendocrine cells, implying viral tropism and pathogenesis through these bronchial cell types. By contrast, *ACE2*, used by SARS and COVID-19 coronaviruses, and *DPP4* used by MERS coronavirus, were both detected at low levels but enriched in alveolar AT2 cells (Fig S8b), consistent with clinical data indicating lung pathology and respiratory distress of these epidemic/pandemic-prone diseases initiates in the alveolar region[51].

Loss, gain and diversification of lung cell types and gene expression patterns in evolution

Our recent construction of a mouse lung cell atlas from young, middle-aged and elderly adult animals[19], supplemented with additional mouse lung cells and annotated as above for human lung (see Methods), allowed us to analyze the evolutionary conservation and divergence of lung cell types at single cell resolution and across the transcriptome. Homologous cell types were identified by conserved expression of cell-type-specific markers aligned by HomologyIDs

(Fig. 7a). A striking result was that mice appear to lack 17 (29%) of the 58 molecularly-defined human lung cell populations, including many of the newly identified human molecular types (proximal ciliated (3); proximal (5), differentiating (6), and proliferating (7) basal; AT2-s (15); Cap-i1 (20) and i2 (21) capillary cells; bronchial vessel 1 (22) and 2 (23); fibromyocytes (28), lipofibroblasts (31); IGSF21+ (52), EREG+ (53), TREM2+ (54) dendritic cells; and OLR1+ monocytes (56)). Some of the missing cell populations might be rare (e.g. Cap-i1 and i2 as well as Bro1 and 2), transient (e.g. Bas-d and Bas-p), unstable (e.g. LipF), or too diverged to relate transcriptionally (e.g. TREM2+ DC) in mice so may be uncovered by further scRNA sequencing and functional follow up. By contrast, only five mouse molecular types, all immune cells (Zbtb32+ B, Ly6g5b+ T, Alox5+ T, Ccr7+ DC, and interstitial macrophages), were not found in human. These results suggest that there has been substantial molecular diversification and specialization of lung cell types during human evolution (or reduction and streamlining during mouse evolution). This also has important implications for medicine because mice would not be expected to accurately model any human disease for which an implicated cell type was missing or substantially altered.

To characterize the conservation in the transcriptomes of homologous cell types, we compared expression levels of all active genes in each human cell type with the expression levels of the orthologous genes in the corresponding mouse cell type (Fig. S9). Most cell types showed highly conserved expression patterns as expected, with CD4+ memory/effector T cells showing the greatest conservation (correlation coefficient 0.76) and plasma cells the least (0.56). Surprisingly, one human cell type (goblet cells) showed greater correlation with another mouse cell type (club cells, 0.68) than with the homologous type (goblet cells, 0.63) (Fig. 7b), despite the conserved expression in goblet cells of canonical markers and the master transcription factor

SPDEF (Fig. S10a). Corresponding cell types in human and mouse diverged significantly in the expression (≥20-fold change, $p < 0.05$) of tens to hundreds of homologous genes, with CD4+ M/E T cells showing the fewest diverged genes (79) and basophils the most (343) (Fig. 7c). The lung as a whole had fewer divergently-expressed genes (60) than any one cell type, suggesting expression lost in one cell type is usually gained in another (Fig. 7c). Diverged genes varied above age-related expression changes in mice (Fig. S11) and included canonical cell type markers, transcription factors, signaling molecules, and disease genes (see below).

Evolutionary changes in gene expression occurred in several different cellular scenarios (Table S7). Type 0 ("conserved") genes are expressed in exactly the same cell types in mouse and human lungs (Fig. 7d, S10b). Type 1 ("expression gain/loss") genes show simple gain (or loss) of expression in the lung between species, which could involve a single cell type (Type 1a, e.g. *PGC* expressed in human, but not mouse, AT2 cells, Fig. 7d), multiple cell types including full tissue compartments (Type 1b, e.g. *RNASE1* expressed in human, but not mouse, epithelial and endothelial cells, Fig. S10c), or the entire lung (Type 1c, *TRIM38* expressed in all human, but not mouse, lung cell types, Fig. S10c). Type 2 ("expression expansion/contraction") changes involve gain (or loss) of expression in additional lung cell types such that expression of the gene expanded (or contracted) in the lung during evolution. For example, Hopx, the canonical AT1-selective transcription factor in mouse, is expressed in both AT1 and AT2 cells in human (Fig. 7d,f), implying the existence of other key AT1 transcription factors (such as *MYRF* (see above), which shows AT1-selective expression in both species (Fig. S9g)) and important changes in the AT2 expression program; and expansion in endothelial cell type expression of *RAMP3*, a co-receptor for vasodilators CGRP and adrenomedullin (ADM), that presumably alters the pulmonary vascular response to these hormones (Fig. S10d).

Type 3 ("expression switch") changes are the most surprising scenario, involving a switch in gene expression from one lung cell type to another during evolution. Two medically important examples are the COPD/emphysema disease genes *SERPINA1* and *HHIP*, both of which are selectively expressed in alveolar AT2 cells in human but in alveolar stromal cell types in mice (*Hhip*, myofibroblasts, Fig. 7d,g; *Serpina1a-e* (combined), pericytes, Fig. S10e). Other components of the hedgehog pathway were mostly conserved (Fig S10f). Extreme examples of expression switching occurred during evolution of species-specific lung cell types, such as the consolidation in expression of anti-bacterial enzymes *LTF*, *LYZ* and *BPIFB1* from multiple mouse airway epithelial cells into human-specific serous cells for airway defense, and consolidation of broadly expressed lipid-handling genes (e.g. *PLIN2*, *APOE*) from mouse alveolar fibroblasts (some of which contain lipid droplets) and myofibroblasts to the human-specific lipofibroblast to support surfactant production (Fig. S10h).

Despite the general conservation of cell type gene expression patterns noted above, only ~6% of expressed genes showed fully conserved cellular expression patterns in the lung (Type 0), and most of these were genes broadly expressed across the lung (Fig. 7e). Thus, the expression patterns of nearly all (94%) genes are labile during organ evolution, with most genes undergoing broadening (~55%, Type 2 change) or simple gain/loss (29%, Type 1) of expression, and more rarely cell type switching (10%, Type 3) (Table S8). It will be important to unravel the genetic mechanisms underlying such widespread evolutionary changes in gene expression and the selective forces operative for the small number of genes with conserved expression, and the consequences for cellular function and mouse models of human diseases.

Discussion

Using droplet and plate-based single cell RNA sequencing, we constructed a nearly complete molecular cell atlas of the human lung comprising 58 molecular types (Fig. 2b). This provides expression profiles for 41 of the 45 known human lung cell types, missing just the exceedingly rare intrinsic neurons and glia (estimated ~0.001% of lung cells) and tuft cells and eosinophils that are prominent only in diseased lungs. We identified 14 novel cell populations distributed across the four tissue compartments (epithelial, endothelial, stromal, and immune) that were as distinct molecularly as the canonical cell types. These include an abundant alveolar epithelial cell type (AT2-s) intermingled with the classical cell types (AT1 and AT2), which may have stem cell function; two molecularly distinct and intermingled alveolar capillary cell types (Cap-a, Cap); new fibroblast subtypes including alveolar and adventitial fibroblasts and ones we call "fibromyocytes" that share markers with airway smooth muscle; and three new resident dendritic cell populations (DC-IGSF21+, EREG+, TREM2+). Each of the newly identified molecular cell types must be thoroughly characterized to define their structures, functions, stabilities, development, and role in disease, as we have done for the two types of alveolar capillary cells (Gillich et al.). If there are other human lung cell types, subtypes, or cell states, they must be exceedingly rare, fragile, region- (lobe) or stage-specific (younger or older donors), or so similar to the 58 molecular types described here that they are not resolved by current profiling and clustering methods.

This comprehensive molecular cell atlas has important implications for human physiology and medicine. The cellular expression patterns predict tens to hundreds of signaling interactions between each of the cell types, many of them new and unexpected. The atlas suggests where hormones act, including 12 hormones with specific cell targeting inferred from

the expression pattern of their cognate receptors. It provides the relative abundance of each immune cell type isolated from the lung and circulation—and allowed us to identify lung resident and homing cell types and to infer the expression changes they undergo following egression as well as the cellular sources of the homing signals. The atlas provides expression maps of the hundreds of genes that have been implicated genetically in lung diseases and as receptors for human viruses, and uncovered dozens of genes with highly specific lung expression patterns that point to likely "culprit cells" or targets in 28 lung diseases and respiratory infections. It also identified the transcription factors selectively expressed in lung cell types that can be used to rationally design cellular reprogramming strategies for tissue engineering of an entire lung.

Our recent creation of a similar lung atlas in mouse provided an unprecedented opportunity in evolutionary biology to comprehensively compare lung cell types and their molecular functions across species[19]. This revealed that mice lack 17 of 58 (~29%) of the molecular cell types and subtypes in the human lung, including most (~86%, 12 of 14) of the newly discovered molecular types. This suggests a dramatic expansion of cell types in the human lineage (or streamlining in the mouse lineage), perhaps to provide new functions, durability, or regenerative capacity for our much larger (6000-fold) lungs and 30-times longer lifespan[52,53]. Even homologous cell types diverged in expression of tens to hundreds of genes. Indeed, we found that gene expression is remarkably labile during evolution, with only 7% of expressed genes showing conserved cellular expression patterns. Most genes (~55%) have undergone evolutionary broadening or consolidation in their expression pattern, and some (29%) show simple gain or loss of lung expression; the rest (10%) switched cell types. These results suggest widespread gain, loss, or conversion of cell-type-specific transcriptional enhancers during

mammalian evolution. Expression changes included important transcription factors, signaling molecules, and disease genes—highlighting the need for caution when using mouse models. Indeed, the atlases predict where mouse models will fail, such as for human diseases where a culprit cell is missing in mouse or does not express the homolog of the disease gene.

Our approach improves on state-of-the-art methods that recently uncovered just over a third (21) of the 58 lung molecular types described here[33], and provides a paradigm for the human cell atlas project that can be easily adapted for other organs. Best practices that served this project and should be useful for other organ atlases include: compiling from the literature a table of known cell types, their relative abundance, and extant markers; obtaining fresh samples intraoperatively (or immediately post-mortem) from multiple positions in the organ along with matched peripheral blood; balancing tissue compartments after dissociation to ensure adequate representation of all cell types including difficult to dissociate and fragile cells; and profiling large numbers of cells (tens of thousands) from each of several subjects with both broad cell coverage (90%, e.g. microfluidic droplet sequencing) and deep gene coverage (10%, e.g., plate-based SmartSeq2). Subject and tissue compartment data should be clustered separately and iteratively until the obtained subclusters are no longer distinguished by biologically meaningful genes. Clusters should then be identified using extant marker genes, characterized by the functions of selectively-expressed genes, and histologic structures and locations of the cells in the organ confirmed by in situ hybridization, immunostaining for cluster-specific markers, or sampling location.

The tissue location of each cell cluster, including most of the newly identified molecular types, provides an initial atlas of not only cell identities and expression profiles but their environment, which enhances the value of each transcriptome by suggesting local interactions

between cell types such as targets of the signals they express and sources of signals for the receptors they express. Perhaps the greatest utility of the atlas will come from comparisons to similarly rich profiles of additional lungs to parse normal variation from technical noise, from other life stages to elucidate how each cell type develops and ages, and to profiles of diseased lung tissue to define which and how cells are molecularly altered in disease[54,55]. Indeed, this atlas should speed the molecular transition in pathology because now almost every lung cell type can be distinguished by the optimal set of sensitive and specific markers. And, with the progress in spatial transcriptomics[46-48], it may soon be possible to obtain full expression profiles of all cells in situ, which would provide context for the individual cell variations seen in scRNAseq data and potentially reveal transient and specialized cell interactions and the dynamics of disease.

Methods

Human lung tissue and peripheral blood

Freshly resected lung tissue was procured intraoperatively from patients undergoing lobectomy for focal lung tumors. Normal lung tissues (~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 a 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 phosphate buffered saline (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 RNAseq of canonical immune populations, whole blood from healthy human donors was obtained from AllCells Inc 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 prior to surgery.

Mouse lung tissue

Lung tissue for Tabula Muris Senis[56] was obtained as previously described. We obtained additional tissue from two mice expressing Cre recombinase and two expressing estrogen-inducible Cre recombinase (Cre-ERT2) for conditional cell-specific labeling *in vivo* with the gene-targeted alleles FVB-*Tbx4-LME-Cre*[57,58] (lung stroma) and B6.129-*Axin2-Cre-ERT2*[57], respectively. Cre-dependent reporter alleles *Rosa26ZsGreen1*, which expresses cytosolic ZsGreen1 following Cre-mediated recombination, and *Rosa26mTmG*, which expresses membrane-targeted GFP (mGFP) following recombination and membrane-targeted tdTomato (mTomato) in all other tissues, were used to label cells expressing *Tbx4* and *Axin2*, respectively[59,60]. Induction of the *Axin2-Cre-ERT2* allele was done by intraperitoneal injection of tamoxifen (3 mg) once a day for three days as described[38]. All mouse experiments 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 Liberase DL (Sigma 5466202001) and 100 µg/ml 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 minutes, 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, Worthington LS006344) were then added and the samples rocked at 37°C for 5 minutes. Samples were then placed at 4°C for the remainder of the protocol. Cells were

filtered through a 100 µm filter, pelleted (300 x g, 5 minutes, 4°C), and resuspended in ACK red blood cell lysis buffer (Gibco A1049201) for 3 minutes, 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 x g, 5 minutes, 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[61]. Briefly, peripheral blood was diluted 10-fold with FACS buffer (2% FBS in PBS), carefully layered on an RT Ficoll gradient (Sigma HISTOPAQUE®-1119), and centrifuged at 400 x g for 30 minutes at room temperature. The buffy coat was carefully removed, diluted 5-fold with FACS buffer, pelleted (300 x g, 5 minutes, 4°C), and incubated in ice cold FACS buffer containing DNAse I (Worthington LS006344) for 10 minutes 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[19]. Briefly, each lung was dissected, minced, and placed in gentleMACS c-tubes (Miltenyi 130-096-334) with digestion buffer (400 µg/ml 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 minutes, completely dissociated on a gentleMACS by running 'm_lung_02', and kept at 4ºC or on ice for the remainder of the protocol. Cells were washed with 5% FBS in PBS, centrifuged at 300 x g

for 5 minutes, 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 flow through 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 labeled "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[19]. 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.

Following negative selection against immune and endothelial cells by MACS, the remaining human lung cells were incubated with FcR Block (Becton Dickinson (BD) 564219) for 5 minutes and stained with directly conjugated anti-human CD45 (Biolegend 304006) and EPCAM

(eBioscience 25-9326-42) antibodies on a Nutator for 30 minutes. Cells were then pelleted (300 x g, 5 minutes, 4°C), washed with FACS buffer three times, then incubated with cell viability marker Sytox blue (1:3000, 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 minutes 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 minutes. Cells were pelleted (300 x g, 5 minutes, 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 ($CD4^{+}$), NK cells ($CD19^{-}CD3^{-}CD56^{+}CD14^{-}$), classical monocytes ($CD19^{-}CD3^{-}CD56^{-}CD14^{+}$). After sorting, plates were quickly sealed, vortexed, spun down for 1 minute at 1000 x 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 Fc Block for 20 minutes and then stained with one of six panels of directly conjugated antibodies for 30 minutes: 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, $CD16^+$ DCs"); 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) ("NK 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. 40,000 cells from 21 canonical immune populations (Table S6) were sorted in duplicate into Trizol LS (Invitrogen 10296010).

After sorting, all plates and samples were quickly sealed, vortexed, spun down for 1 minute at 1000 x g and then snap frozen on dry ice and stored at -80 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 complementary DNA (cDNA) and amplified as previously described[19]. Illumina sequencing libraries for cDNA from single cells were prepared as previously described[19]. Briefly, 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 minutes. The reaction was neutralized by adding "Neutralize Tagment Buffer" (Illumina) and centrifuging at room temperature at 3,220 x g for 5 minutes. 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)[19].

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 analyzer (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[19]. 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[61]. Libraries were sequenced on a NextSeq 500 (Illumina).

Immunohistochemistry

Mouse and human lungs were collected as previously described[38,62]. 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: pro-SftpC (rabbit, Chemicon AB3786, 1:250 dilution), E-cadherin (rat, Life Technologies 131900 clone ECCD-2, 1:100), Mucin 1 (Muc1, hamster, Thermo Scientific HM1630, clone MH1, 1:250), Ki67 (rat,

DAKO M7249, clone TEC-3, 1:100), Carbonic Anhydrase 2 (rabbit, Abcam EPR19839, 1:100). Primary antibodies were detected with Alexa Fluor-conjugated secondary antibodies (Jackson ImmunoResearch) unless otherwise noted, then mounted in Vectashield containing DAPI (5 ug/ml, Vector labs). Images were acquired with a laser scanning confocal fluorescence microscope (Zeiss LSM780) and processed with ImageJ and Imaris (version 9.2.0, Oxford Instruments). All immunostains were performed on at least 2 human subjects 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). All smFISH experiments were performed on at least 2 human subjects distinct from the donors used for sequencing, and quantifications were based on at least 10 fields of view for each. Fields of view were scored manually, calling a cell positive for each gene probed if its nucleus had >3 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).

Bioinformatic Methods

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 1000 unique molecular identifiers (UMIs) were excluded from further analyses.

Reads from single cells isolated by flow cytometry were demultiplexed using bcl2fastq (version 2.19.0.316, Illumina), pruned for low nucleotide quality scores and adapter sequences using skewer (version 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 (version 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 (version 2.3)[63]. Briefly, 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 neighbor map and an empirically set resolution, as implemented in the 'FindClusters' function.

Clusters were grouped and separated based on expression of tissue compartment markers (e.g. *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[64] implemented in the 'FindMarkers' command, were no longer biologically meaningful (e.g.

clusters distinguished by dissociation-induced genes[42], ribosomal genes, mitochondrial genes, or ambient RNA released by abundant cells such as RBCs[43]). 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) (Table S1). RNA markers for canonical immune populations were obtained using the 'MAST' statistical framework (Table S4). Clusters were assigned a canonical identity based on enriched expression of these marker genes. 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 (e.g. proximal versus distal, alveolar versus adventitial) and prepended these locations when applicable. When both clusters localized to the same tissue region (e.g. 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 (e.g. Signaling AT2 versus AT2, Proliferative Basal versus Basal). If the clusters could not be resolved by location or function, we prepended a representative marker gene to their "canonical" identity (e.g. 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[19] 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 Reyfman et al, GSE130148 for Braga et al), clustered, and annotated using the strategy described above. New annotations for each cell are available on GitHub (see below).

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.

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 artifact 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.

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. 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 Mathews Correlation Coefficients (MCCs). 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[65], 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). 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, as previously described[25]. 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 median expression profiles for each cell type 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).

To compare the expression pattern of each gene across species we binarized genes as expressed (1) or not expressed (0) in each cell type. 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/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/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 ($V = (a – b) + 2ab$, where 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.

Data and code availability

Counts/UMI tables, cellular metadata, Seurat objects, and scanpy objects are available on Synapse (accession syn21041850). The data can be explored in a browser using cellxgene at https://hlca.ds.czbiohub.org/. Code for demultiplexing counts/UMI tables, clustering, annotation, and other downstream analyses are available on GitHub (https://github.com/krasnowlab/HLCA). Human sequencing data will be available by data access agreement on the European Genome-phenome Archive (EGA) upon publication of this manuscript. Mouse sequencing data will be available on the National Institute of Health's Sequence Read Archive (SRA) also upon publication of this manuscript.

References

1. Schwann, T. *Mikroskopische Untersuchungen über die Uebereinstimmung in der Struktur und dem Wachsthum der Thiere und Pflanzen*. (1839).
2. Weigert, C. Ueber die pathologischen Gerinnungsvorgänge. *Archiv f. pathol. Anat.* **79,** 87–123 (1880).
3. Porter, K. R. *et al.* A study of tissue culture cells by electron microscopy: methods and preliminary observations. *Journal of Experimental Medicine* **81,** 233–246 (1945).
4. Gehr, P. *et al*. The normal human lung: ultrastructure and morphometric estimation of diffusion capacity. *Respiration Physiology* **32,** 121–140 (1978).
5. Ross, M. H. & Pawlina, W. *Histology*. (Lippincott Williams & Wilkins, 2006).
6. Bogoch, S. *et al.* Separation of cerebroproteins of human brain. *Nature* **204,** 73–75 (1964).
7. King, R. J. *et al.* Isolation of apoproteins from canine surface active material. *Am. J. Physiol.* **224,** 788–795 (1973).
8. Balis, J. U. *et al.* Distribution and subcellular localization of surfactant-associated glycoproteins in human lung. *Lab. Invest.* **52,** 657–669 (1985).
9. Hermans, C. & Bernard, A. Lung epithelium-specific proteins: characteristics and potential applications as markers. *Am. J. Respir. Crit. Care Med.* **159,** 646–678 (1999).
10. Schena, M. *et al.* Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270,** 467–470 (1995).
11. Wang, Z. *et al.* RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).
12. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6,** 377–382 (2009).
13. Wu, A. R. *et al.* Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* **11,** 41–46 (2014).
14. Gawad, C. *et al.* Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17,** 175–188 (2016).
15. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509,** 371–375 (2014).
16. Montoro, D. T. *et al.* A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560,** 319–324 (2018).
17. Plasschaert, L. W. *et al.* A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560,** 377–381 (2018).
18. Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **173,** 1307 (2018).
19. Tabula Muris Consortium *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562,** 367–372 (2018).
20. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161,** 1202–1214 (2015).
21. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566,** 496–502 (2019).
22. Pijuan-Sala, B. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566,** 490–495 (2019).
23. Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* **174,** 999–1014.e22 (2018).

24. Saunders, A. *et al.* Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell* **174,** 1015–1030.e16 (2018).
25. Vento-Tormo, R. *et al.* Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* **563,** 347–353 (2018).
26. 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).
27. Angelidis, I. *et al.* An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nature Communications 2017 8* **10,** 1–17 (2019).
28. Xu, J. *et al. Deaths: final data for 2016.* (2018).
29. World Health Organization. *World health statistics 2019: monitoring health for the SDGs, sustainable development goals.* (2019).
30. Franks, T. J. *et al.* Resident Cellular Components of the Human Lung. *Proceedings of the American Thoracic Society* **5,** 763–766 (2012).
31. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols 2014 9:1* **9,** 171–181 (2014).
32. Blondel, V. D. *et al.* Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008,** P10008 (2008).
33. Braga, F. A. V. *et al.* A cellular census of human lungs identifies novel cell states in health and in asthma. *Nature Medicine 2018 24:8* **25,** 1153–1163 (2019).
34. Howitt, M. R. *et al.* Tuft cells, taste-chemosensory cells, orchestrate parasite type 2 immunity in the gut. *Science* **351,** 1329–1333 (2016).
35. Rock, J. R. *et al.* Notch-dependent differentiation of adult airway basal stem cells. *Cell Stem Cell* **8,** 639–648 (2011).
36. Garcia, S. R. *et al.* Single-cell RNA sequencing reveals novel cell differentiation dynamics during human airway epithelium regeneration. *bioRxiv* **140,** 451807 (2018).
37. Watson, J. K. *et al.* Clonal Dynamics Reveal Two Distinct Populations of Basal Cells in Slow-Turnover Airway Epithelium. *Cell Reports* **12,** 90–101 (2015).
38. Nabhan, A. N. *et al.* Single-cell Wnt signaling niches maintain stemness of alveolar type 2 cells. *Science* **359,** 1118–1123 (2018).
39. Zacharias, W. J. *et al.* Regeneration of the lung alveolus by an evolutionarily conserved epithelial progenitor. *Nature* **555,** 251–255 (2018).
40. Stan, R. V. *et al.* The diaphragms of fenestrated endothelia: gatekeepers of vascular permeability and blood composition. *Developmental Cell* **23,** 1203–1218 (2012).
41. Tan, S. Y. S. & Krasnow, M. A. Developmental origin of lung macrophage diversity. *Development* **143,** 1318–1327 (2016).
42. van den Brink, S. C. *et al.* Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14,** 935–936 (2017).
43. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature Communications 2017 8* **8,** 14049 (2017).
44. Shiow, L. R. *et al.* CD69 acts downstream of interferon-α/β to inhibit S1P 1 and lymphocyte egress from lymphoid organs. *Nature* **440,** 540–544 (2006).
45. Mackay, L. K. *et al.* Hobit and Blimp1 instruct a universal transcriptional program of tissue residency in lymphocytes. *Science* **352,** 459–463 (2016).
46. Moffitt, J. R. & Zhuang, X. RNA Imaging with Multiplexed Error-Robust Fluorescence In Situ Hybridization (MERFISH). *Methods in Enzymology* **572,** 1–49 (2016).

47. Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361,** eaat5691 (2018).
48. Eng, C.-H. L. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568,** 235–239 (2019).
49. Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), (2018). https://omim.org/
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,** D1005–D1012 (2019).
51. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet* **395,** 497–506 (2020).
52. Limjunyawong, N. *et al.* Measurement of the pressure-volume curve in mouse lungs. *JoVE (Journal of Visualized Experiments)* e52376 (2015).
53. Seeley, R. R. *et al. Essentials of anatomy and physiology*. (2005).
54. Miller, A. J. *et al.* Basal stem cell fate specification is mediated by SMAD signaling in the developing human lung. *bioRxiv* 461103 (2018).
55. Lambrechts, D. *et al.* Phenotype molding of stromal cells in the lung tumor microenvironment. *Nature Medicine 2018 24:8* **24,** 1277–1289 (2018).
56. Tabula Muris Consortium *et al.* A Single Cell Transcriptomic Atlas Characterizes Aging Tissues in the Mouse. *bioRxiv* 661728 (2019).
57. van Amerongen, R. *et al.* 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).
58. Greif, D. M. *et al.* Radial Construction of an Arterial Wall. *Developmental Cell* **23,** 482–493 (2012).
59. Muzumdar, M. D. *et al.* A global double-fluorescent Cre reporter mouse. *Genesis* **45,** 593–605 (2007).
60. Madisen, L. *et al.* A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. *Nature Neuroscience 2009 13:1* **13,** 133–140 (2010).
61. Moraga, I. *et al.* Tuning Cytokine Receptor Signaling by Re-orienting Dimer Geometry with Surrogate Ligands. *Cell* **160,** 1196–1208 (2015).
62. Desai, T. J. *et al.* Alveolar progenitor and stem cells in lung development, renewal and cancer. *Nature* **507,** 190–194 (2014).
63. Butler, A. *et al.* Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology 2018 36:5* **36,** 411–420 (2018).
64. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. **16,** 278 (2015).
65. Hu, H. *et al.* AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. http://bioinfo.life.hust.edu.cn/AnimalTFDB/

Acknowledgements

We are grateful to the tissue donors and the clinical staff at Stanford Medical Center who made tissue collection possible, especially Jalen Benson and Emily 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 Dr. Lisa Nichols and Meredith Weglarz; members of Chan Zuckerberg Biohub and Quake Lab who supported this work, particularly Aaron McGeever, Dr. Brian Yu, Bob Jones, and Saroja Kolluru; Dr. Maya Kumar for discussions on annotation of stromal cells; and Maria Petersen for illustrating the lung schematic (Fig. 2b) and Dr. Camilla 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 Jim Spudich and members of the Krasnow lab for valuable discussions and comments on the manuscript, and Alexander Lozano for discussions on bioinformatic analyses. This work was supported by funding from the Chan Zuckerberg Biohub (S.R.Q.) and the Howard Hughes Medical Institute (M.A.K.). 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.,

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.

Figure Legends

Figure 1. Strategy for single cell RNA sequencing and annotation of human lung and blood cells. a, Workflow for capture and mRNA sequencing of single cells from the healthy unaffected regions indicated (D, distal; M, medial; P, proximal lung tissue) of fresh, surgically resected lungs with focal tumors from three subjects (1, 2, 3) and their matched peripheral blood. Cell representation was balanced among the major tissue compartments (Endo, endothelial; Immune; Epi, epithelial; Stroma) by magnetic and fluorescence activated cell sorting (MACS and FACS) using antibodies for the indicated surface markers (CD31, CD45, EPCAM; +, marker-positive; -, marker-negative). Cell capture and single cell RNA sequencing (scRNAseq) was done using 10x droplet technology or SmartSeq2 (SS2) analysis of plate-sorted cells. Number of profiled cells from each compartment are shown in parentheses. For blood, immune cells were isolated on a high density Ficoll gradient, and unsorted cells captured and profiled by 10x and sorted cells (using canonical markers for the indicated immune populations) by SS2. Total cell number (all 3 subjects) and median expressed genes per cell are indicated for each method. **b,** Cell clustering and annotation pipeline. Cell expression profiles were computationally clustered by nearest-neighbor relationships and clusters were then separated into tissue compartments based on expression of compartment-specific markers (*EPCAM* (blue), *CLDN5* (red), *COL1A2* (green), and *PTPRC* (purple)), as shown for tSNE plot of lung and blood cell expression profiles obtained by 10x from Patient 3. Cells from each tissue compartment were then iteratively re-clustered until differentially-expressed genes driving clustering were no longer biologically meaningful. Cell cluster annotation was based on expression of canonical marker genes from the literature, markers found through RNA sequencing of purified cell populations (Bulk RNA markers), ascertained tissue location, and inferred molecular function from differentially-expressed genes.

c, Heatmap of pairwise Pearson correlations of the average expression profile of each cluster in the combined 10x dataset plus SS2 analysis of neutrophils. Tissue compartment and identification number of each of the 58 clusters are indicated.

Figure 2. Identity and location of human lung epithelial, endothelial, and stromal cell types. **a,** Hierarchical tree showing human lung molecular cell types and their annotations in the indicated tissue compartments following iterative clustering (each level of hierarchy is an iteration) of scRNAseq profiles of cells in each compartment. Black, canonical cell types; blue, proliferative subpopulations or differentiating states; red, novel populations (light red, found only in one subject). Number of cells in each cluster and canonical marker genes are shown below. **b,** Diagrams showing localization and morphology of each lung cell type (numbering scheme as in (a) and Figure 3a). **c,** Dot plot of mean level of expression (dot intensity, gray scale) of indicated basal cell markers and percent of cells in population with detected expression (dot size). **d,** RNAscope single molecule fluorescence in situ hybridization (smFISH) and quantification for general basal marker *KRT5* (red) and Bas-px marker *SERPINB3* (white) with DAPI (nuclear) counter stain (blue) and autofluorescence (ECM, green). Scale bar, 20 µm. Note Bas-px cell (*KRT5 SERPINB3* double positive, yellow arrowhead and box) enrichment at base of pseudostratified airways. **e,** Dot plot of expression of AT2 markers. **f,** smFISH and quantification for general AT2 marker (*SFTPC*, white) and AT2-s marker *WIF1* (red puncta, arrowheads). Scale bar, 10 µm. Note AT2-s cells ($SFTPC^{pos}$ $WIF1^{neg}$, yellow arrowhead and box) intermingled with AT2 cells ($SFTPC^{pos}$ $WIF1^{pos}$, white arrowheads). **g,** Dot plot of expression of endothelial markers. **h,** smFISH for general endothelial marker *CLDN5* (red, left panel), bronchial vessel-specific markers *MYC* (green, left panels) and Bro1-specific marker *ACKR1* (red, right) on serial

sections of bronchial vessel cells (arrowheads) adjacent to an airway (see Fig S4h), co-stained for DAPI (blue). Scale bar, 10 µm. Quantification shows abundance of Bro1 and Bro2 cells. **i,** Dot plot of stromal markers. **j**, smFISH and quantification for the general fibroblast marker *COL1A2* (white), alveolar fibroblast (AlvF) marker *GPC3* (red, left) and adventitial fibroblast (AdvF) marker *SERPINF1* (red, right) with DAPI counterstain (blue) and autofluorescence (ECM, green). Adventitial fibroblasts (arrowheads, k) localize around blood vessels (ECM, green). Scale bars, 10 µm. Quantification shows abundance of stromal cell types in alveolar and proximal regions, pericyte (Peri) and lipofibroblast (LipF) staining in Fig S4i-j. All stains were done on lungs from at least two human subjects distinct from those used for profiling; quantifications represent ≥10 fields of view.

Figure 3. Identity and residency of human lung immune cells. a, Hierarchical tree showing human lung immune cell molecular types (grouped by lineage) and their annotations based on expression of marker genes in cell clusters of scRNAseq profiles as in Fig. 2a. Black, canonical cell types; blue, proliferative subpopulations; red, novel populations (light red, found only in one subject). Number of cells in each cluster (combined 10x and SS2 datasets) is indicated. Cell types showing significant expression differences based on whether cell was obtained from lung (L) vs. blood (IV, intravascular) samples are shown as additional level of hierarchy. **b,** Relative abundance of each immune cell type (shown in panel a) in lung (blue) and blood (red) samples. Cell types with >90% enrichment in lung samples are annotated “lung resident” (Res, alveolar macrophages) or “lung homing” (Hom), those with >90% enrichment in blood as “intravascular” (IV), and all others as “egressed” (Egr). (Note these assignments should be considered provisional because enrichment values can be influenced by efficiency of harvesting during cell

dissociation and isolation.) Red lettering/symbols, immune cells not previously known to home to (be enriched in) lung or have differential expression (delta symbol) following egression from blood vessel into tissue. **c,** Dot plot showing expression of canonical dendritic markers (*HLA-DPB1*, *GPR183*), myeloid dendritic cell type 1 and 2 (mDC1 and mDC2) markers (*LAMP3A*, *CLEC9A*, *CD1C*, *PLD4*), and six markers for three novel dendritic populations (IGSF21+, EREG+, and TREM2+) in the dendritic immune cell clusters indicated (clusters 50-54). UP10K, UMIs per 10,000. **d,** Box-and-whisker plots of general, lymphocyte-specific, and myeloid-specific lung residency (egression) gene expression signature scores based on expression of indicated genes in 10x profiles of the indicated immune cells types isolated from the indicated sample sites (IV, from blood; L, from lung). Many previously known lymphocyte residence genes (e.g. *S1PR1*, *RUNX3*, *RBPJ, HOBIT*) were expressed at low levels and only uncovered in SS2 profiles with better transcript capture. Gray shading, myeloid immune cells.

Figure 4. Markers and transcription factors that distinguish human lung cell types.

a, Violin plots of expression levels (ln(UP10K + 1)) in 10x profiles of the most sensitive and specific markers (gene symbols) for each human lung cell type in its tissue compartment. **b,** Scheme for selecting the most sensitive and specific marker genes for each cell type using, Matthews Correlation Coefficient (MCC). The box-and-whisker plots below show MCCs, True Positive Rates (TPR), and False Discovery Rates (FDR) for each cell type using the indicated number (nGene) of the most sensitive and specific markers. Note all measures saturate at 2-4 genes, hence simultaneous in situ probing of a human lung for the ~100-200 optimal markers would assign identity to nearly every cell. **c,** Alveolar region of human lung probed by smFISH for AT1 marker *AGER* and transcription factor *MYRF* mRNA. *MYRF* is selectively expressed in

AT1 cells (arrowheads; 97% of *MYRF*+ cells were *AGER*+, n=250 scored cells). Scale bar, 10 µm. **d,** Alveolar region of human lung probed by smFISH for pericyte marker *COX4I2* and transcription factor *TBX5* mRNA. *TBX5* is enriched in pericytes (arrowheads, 92% of *TBX5*+ were *COX4I2*+, n=250). Scale bar, 5 µm. **e,** Dot plot of expression of enriched transcription factors in each lung cell type from SS2 profiles. Red, genes not previously associated with the cell type. Red shading, transcription factors including *MYRF* that are highly enriched in AT1 cells, and *TBX5* and others highly enriched in pericytes.

Figure 5. Lung cell targets of circulating hormones and immune cells. a, Dot plot of hormone receptor gene expression in lung cells from combined SS2 profiles. Hormones for each receptor are shown at top. Teal, broadly-expressed receptors in lung; other colors, selectively-expressed receptors (<3 lung cell types). Small colored dots near cell types show selectively targeted cell types. AA, amino acid; CGRP, Calcitonin gene-related peptide; AM, adrenomedullin; SST, somatostatin; EPO, erythropoietin; GIP, gastric inhibitory peptide; GH, growth hormone; IGF, insulin-like growth factor; MCCT, mineralocorticoid; GCCT, glucocorticoid; RA, retinoic acid. **b,** Schematic of inferred pericyte cell contractility pathway and its regulation by circulating hormones (AGT, PTH) and capillary-expressed signals (EDN, NO). Dots show expression of indicated pathway genes: values at left (outlined red) in each pair of dots in capillary diagram (top) show expression in Cap-a cells (aerocytes) and at right (outlined blue) show expression in general Cap cells. Note most signal genes are preferentially expressed in Cap relative to Cap-a cells. **c,** Dot plots of expression of chemokine receptor genes (left) and their cognate ligand genes (right) in human lung cells from combined 10x expression profiles. Only cell types and chemokines with detected expression are shown; ionocytes (red lettering) are

the only non-immune cell with chemokine receptor expression. Colored lines highlight predicted target cells (chemokine-expressing cells, right) of migrating immune cell types and ionocytes (left) that express a cognate chemokine receptor gene; heavier weight lines indicate previously unknown interactions.

Figure 6. Mapping the cellular origins of human lung diseases by cell-selective expression of disease genes. a, Dot plots of expression of lung disease genes (numbered with the associated disease indicated above) that are enriched in specific lung cell types from combined SS2 expression profiles. Red, novel cell type association of gene/disease; gray, diseases with developmental phenotype. BBS, Bardet-Biedl syndrome; Dys, dysplasia; IPF, idiopathic pulmonary fibrosis; SMD, surfactant metabolism dysfunction; PH, pulmonary hypertension; SM, smooth muscle; SGB, Simpson-Golabi-Behmel; TB, tuberculosis; AWS, Alagille-Watson syndrome; VDES, Van den Ende-Gupta syndrome; EDS, Ehlers-Danlos syndrome; CF, Cystic fibrosis; Fam Med, Familial Mediterranean; COPD, Chronic Obstructive Pulmonary disease. **b,** Dot plots of expression of all genes implicated (OMIM, Mendelian disease gene from OMIM database; GWAS, genome-wide association at 10^{-20} significance) in the indicated disease (PH, pulmonary hypertension, first panel; tuberculosis susceptibility, second; COPD/emphysema, third). Red shading, cell type strongly implicated in disease; blue shading, weaker but still notable association. Note canonical AT2 cells (red) express every major COPD/emphysema disease gene whereas AT2-s cells (blue) express only some. Fourth panel, dot plot showing expression of genes used by viruses (listed on the right) to enter the lung. Fourth panel, dot plot showing expression of genes used by viruses indicated at right to enter lung. Red shading, cell types inhaled viruses could directly access (epithelial cells and macrophages); darker red shading

shows expression values for measles receptor *NECTIN4* (measles receptor) and rhinovirus C receptor *CDHR3*. **c,** smFISH of an alveolar region of adult human lung for PH disease gene *KCNK3* (red), pericyte marker *COX4I2* (white), and DAPI counterstain (blue). Note pericyte-specific expression (arrowheads, 91% of pericytes were *KCNK3*+, n=77). Scale bar, 5 µm. **d,** smFISH of alveolar region of an adult human lung for atrioventricular (AV) dysplasia gene *ACVRL1* (red), endothelial marker *CLDN5* (white), and DAPI. Note *ACVRL1* and *CLDN5* double positive capillaries (white arrowheads, 70% of capillaries were *ACVRL1*+, n=102) with some *CLDN5* single positive capillaries (yellow arrowheads). Scale bar, 5 µm. **e,** smFISH and quantification of alveolar region of adult human lung for COPD/emphysema gene *SERPINA1* and AT2 marker *SFTPC*, and DAPI. Note AT2-specific expression (arrowheads, 93% of AT2 cells were *SERPINA1*+, n=176). Scale bar, 5 µm.

Figure 7. Evolutionary divergence of lung cell types and gene expression patterns.

a, Alignment of mouse (top) lung cell types isolated, profiled by scRNAseq and clustered from each tissue compartment (see Methods) as for human (bottom, from Figs. 2a, 3a). Lines connect homologous cell types expressing classical markers shown at bottom. Thin lines indicate expansions; dashed lines, potential expansion of functionally related cell types. Note some cell types (34 total) have a unique match whereas others (8 total) have multiple matches (indicating cell type diversification during human evolution) or no identified match in mice (5 total, suggesting cell type gain in human lineage). Red text, newly identified populations (light red, identified in only 1 subject so could be subject- or disease-specific); blue, cell states more abundant in human lung; gray, extant mouse cell types not captured in our data. **b,** Heatmap showing global transcriptome correlation between indicated human and mouse epithelial cells

from SS2 dataset. Red outline, matched cell types based on classical markers described in (a). White dot, human to mouse correlation. **c,** Scatter plots showing the median expression levels (dots) in the indicated cell types of each expressed human gene and mouse ortholog in the mouse and human SS2 datasets. Scale, ln(CPM+1). Note Bas/Ma 1 cells have the most differentially-expressed genes (343), and $CD4^+$ M/E T cells have the least (79), but both have more than whole lung (60)—indicating genes lost in cell types are usually gained in others. Correlation scores (R values) between the average mouse and human gene expression profiles for each cell type are indicated. **d,** Dot plots of expression of homologous genes in mouse and human lung cells that exemplify the four scenarios observed for evolution of cellular expression pattern. Perfectly conserved cellular expression pattern for *ASCL1* ("Type 0" evolutionary change, NE-specific), evolutionary gain in expression in human lineage for *PGC* (acquired expression in AT2 and cells, Type 1 change), evolutionary expansion in expression from AT1, club, and ciliated cells for *HOPX* into AT2, AT2-s and other cell types indicated in human lineage (Type 2), and evolutionary switch in expression for *HHIP* from mouse myofibroblast to human AT2 cells and fibromyocytes (Type 3). Blue highlight, conserved cell expression; red, divergent cell expression. **e,** Pie chart of fraction of expressed genes in lung showing each of the four types of evolutionary changes in cellular expression patterns from mouse to human. Histogram below shows number of lung cell types that the 602 genes with perfectly conserved cellular expression patterns (Type 0) are expressed in. Note nearly all conserved genes are expressed in either a single cell type (67%) or nearly all cell types (33%). **f**, Alveolar sections from mouse (top) or human (bottom) immunostained for transcription factor HOPX (red) and AT2 marker MUC1 (green), and DAPI (blue). Note *HOPX* is expressed selectively in AT1 cells (arrowheads) in mouse but expression in humans has expanded to AT2 and AT2-s cells (dashed circles). Scale

bars, 10 µm. **g**, 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* in human. Note evolutionary switch in *HHIP* expression from myofibroblast (mouse, arrowhead) to AT2 cells (human, dashed circles). Scale bars, 10 µm.

Extended Data Figure Legends

Extended Data Figure S1. Histology of anatomical landmarks in donor lungs. Representative micrographs of donor lungs from formalin-fixed, paraffin-embedded (FFPE) sections stained with haematoxylin and eosin showing bronchi, bronchioles, submucosal glands, arteries, veins, and alveoli near regions used for single cell RNA sequencing. Scale bar, 100 µm.

Extended Data Figure S2. FACS gating strategies for human lung and peripheral blood cells. a, Sequential FACS data and sorting gates (red) for dissociated human lung cells from subject sample D1b (plate B001223) following MACS depletion of highly abundant immune ($CD45^+$) and endothelial ($CD31^+$) cells. The final sort (right) was of viable single cells from the lung epithelial ($EPCAM^+CD45^-$), immune ($CD45^+EPCAM^-$), and stromal/endothelial ($EPCAM^-CD45^-$) compartments into 384-well plates for SS2 scRNAseq. **b,** Sequential FACS data and sorting gates (red) for white blood cells isolated on a Ficoll gradient of matched subject peripheral blood (subject 1, plate BP1). Viable, single $CD235a^-$ (non-RBC) cells were captured without additional gating (panel 4), or further sorted as CD8 T ($CD8^+$; panel 8), CD4 T ($CD4^+$; panel 7), B ($CD19^+CD3^-$; panel 6), NK ($CD19^-CD3^-CD56^+CD14^-$; panel 9), or $CD14^+$ monocytes ($CD19^-CD3^-CD56^-CD14^+$; panel 9) for SS2 scRNAseq. Contours, 5% increments in cell density.

Extended Data Figure S3. Expression and positional differences among identified subtypes of canonical lung populations. a, Proliferative signature (based on expression of indicated genes) of each cluster of basal cells, NK/T cells, and macrophages. Note three clusters with a high signature: basal-proliferative (Bas-p), NK/T-proliferative (NK/T-p), and macrophage-proliferative (MP-p) **b,** Relative abundance of epithelial and stromal cell types in scRNAseq

analysis of human lung samples obtained from proximal (blue; samples P3) and distal (red; samples D1a, D1b, D2, D3) sites in the lung. In addition to the expected proximal enrichment of some airway cell types (goblet, ionocytes, neuroendocrine cells) and distal enrichment of alveolar cell types (AT1, AT2, AT2-s, myofibroblasts), note the three cell types (ciliated, cil; basal, bas; myofibroblasts, MyoF) with transcriptionally-distinct subsets that differ in their proximal-distal enrichment: ciliated (cil) vs. ciliated-proximal (cil-px), basal (bas) vs. basal-proximal (bas-px), myofibroblasts (MyoF) vs. fibromyocyte (FibM). (Note relative enrichment values should be considered provisional because enrichment values can be influenced by efficiency of harvesting during cell dissociation and isolation.) **c,** Dot plot of expression in ciliated cells (Cil) and proximal ciliated cells (Cil-px) of canonical (General) ciliated cell markers and Cil-px-specific (Proximal) markers. **d,** Heatmap of expression of representative general AT2 marker genes, AT2-selective and AT2-s selective marker genes indicated in AT2 (left panel) and (right panel) human lung cell clusters from the SS2 expression profiles. Note AT2 selective marker genes include negative regulators of the Hedgehog and Wnt signaling pathways (e.g., *HHIP* and *WIF1*, highlighted in red) and AT2-s selective markers include Wnt ligands, receptors, and transcription factors (e.g., *WNT5A*, *LRP5*, and *TFC7L2* highlighted in green). Values shown are ln(CPM+1) for 50 randomly-selected cells in each cluster from the SS2 analysis. **e,** Heatmap of expression of representative general, adventitial-selective, and alveolar-selective fibroblast markers in 50 randomly-selected cells in the adventitial fibroblast (left) and alveolar fibroblast (right) clusters from SS2 expression profiles. Note specialization in growth factors (AdvF: *PDGFRL*, *IGFBP4*, AlvF: *FGFR4*, *VEGFD*) and morphogen (AdvF: *SFRP2*, AlvF: *NKD1*, *DKK3*) signaling/regulation in red.

Extended Data Figure S4. smFISH and immunostaining of cell markers to localize epithelial, endothelial, stromal clusters in the lung. **a,** Immunostaining of adult human pseudostratified airway for proliferation marker MKI67 (green) in basal cells (marked by KRT5, red) with DAPI (nuclear) counter stain (blue). Scale bars, 5 µm. Quantification shows abundance of proliferating (MKI67-expresing) basal cells (Bas-p) in pseudostratified and simple epithelial airways; n, KRT5+ cells scored in sections of two human lungs. **b,** Immunostaining of adult human pseudostratified airway for differentiation marker HES1 (green) in basal cells (marked by KRT5, red) with DAPI (nuclear) counter stain (blue). Scale bars, 10 µm. Note apical processes extending from HES1+ basal cells (arrowheads) indicating migration away from basal lamina as they differentiate. Other HES1^{+} cells have turned off basal marker KRT5. Dashed outlines, basal cell nuclei. Quantification shows fraction of basal cells (Bas, cuboidal KRT5+ cells on basement membrane) and Bas-d cells (KRT5+ cells with apical processes) that were HES1+. **c,** smFISH of adult human simple epithelial airway for general basal cell marker *KRT5* (red) and proximal basal cell marker *SERPINB3* with DAPI counter stain (blue) and ECM autofluorescence (green). *SERPINB3* is not detected, indicating only Bas cells but not Bas-px cells are present in simple airways (compare with proximal, pseudostratified airway in Fig. 2d). Scale bar, 10 µm. **d,** smFISH and quantification of human pseudostratified epithelial (left panel) and simple epithelial (right panel) airways for general ciliated marker *C20orf85* (white) and proximal ciliated marker *DHRS9* (red) with DAPI counterstain (blue) and ECM autofluorescence (green). Note restriction of Cil-px ciliated cells (expressing proximal marker) to pseudostratified airways. Scale bars, 10 µm. **e-g,** smFISH and quantification of vessel types indicated (dotted outlines) showing expression of vein marker *ACKR1* (red, panel e), artery marker *GJA5* (red, panel f), lymphatic marker *CCL21* (red, panel g), and general endothelial marker *CLDN5* with DAPI counter stain

(blue) and ECM autofluorescence (green). Scale bars, 50 µm, 30 µm, and 40 µm. **h,** Micrograph (low magnification) of bronchial vessel (boxed region) in Fig. 2h showing vessel location near airway (dotted outline). **i-k,** smFISH and quantification of human alveolar region showing expression of pericyte marker *COX4I2* (red, panel i), lipofibroblast marker *APOE* (red, panel j), alveolar fibroblast marker *GPC3* (red, panel k), and general stromal marker *COL1A2* with DAPI counter stain (blue) and ECM autofluorescence (green). Note (panel j) the boxed *COL1A2* and *APOE* double-positive lipofibroblast (LipF) among single-positive *COL1A2* cells (other stromal cells) and *APOE* cells (macrophages); for quantification, see Fig. 2j. Scale bars, 20 µm. **l,** smFISH of alveolar region of adult mouse lung for alveolar fibroblast markers *Slc7a10* (white) and *Frfr4* (red) co-stained for alveolar entrance ring marker Elastin (green) with DAPI counterstain (blue). Scale bar, 5 µm. **m**, smFISH of mouse pulmonary artery for adventitial fibroblast markers *Pi16* (white) and *Serpinf1* (red) and co-stained for arterial marker Elastin (green) and counterstained with DAPI (blue). Note localization of the double-positive cells (adventitial fibroblasts, arrowheads) around artery. Scale bar, 10 µm. **n, o,** smFISH and quantification of human pseudostratified epithelial airway (n) and alveolar region (o) for airway smooth muscle (ASM) and fibrobmyocyte (FibM) markers *ACTG2* (white, n) and COX4I2 (white, o), fibromyocyte and myofibroblast (MyoF) marker *ASPN* (red), with DAPI counter stain (blue) and ECM autofluorescence (green). Note intermingled airway smooth muscle and fibromyocytes near a pseudostratified airway (dotted outline) in n, and *ASPN*$^{+}$ *COX4I2*$^{-}$ myofibroblasts (boxed and arrowhead) enriched in alveoli in o.

Extended Data Figure S5. Selectively-expressed RNA markers of human immune cell types from bulk mRNA sequencing of FACS-purified immune cells. a, Heatmap of RNA

expression of the most selectively-expressed genes from bulk mRNA sequencing of the indicated FACS-sorted immune populations (see Table S3). This dataset provided RNA markers for human immune cell populations that have been classically defined by their cell surface markers. **b,** Heatmap of pairwise correlation scores between the average expression profiles of the immune cell types indicated that were obtained from bulk mRNA sequencing (BulkSeq, panel a) to the average scRNAseq profiles of human blood immune cells in the SS2 dataset annotated by canonical markers and enriched RNA markers from the bulk RNA-seq analysis. The highest correlation in overall gene expression (white dot) of each annotated immune cell cluster in the SS2 dataset (columns) was to the bulk RNA-seq of the same FACS-purified immune population (rows), supporting the scRNAseq immune cluster annotations (red squares). **c,** tSNE of monocytes and B, T, and NK cells from patient 1 10x dataset. Note separate clusters of each immune cell type isolated from lung (no outline) and blood (dashed outline). Asterisk, small number of B cells isolated from the lung that cluster next to blood B cells. **d,** Heatmap of expression of dendritic cell marker genes in the scRNAseq profiles of the indicated dendritic cell clusters from human blood and lung 10x datasets. Note all clusters express general dendritic markers including antigen presentation machinery, but each has its own set of selectively-expressed markers. The red-highlighted markers that distinguish the three novel dendritic cell clusters we identified (IGSF21+, EREG+, TREM2+) suggest differential roles in asthma (IGSF21+), growth factor regulation (EREG+), and lipid handling (TREM2+). **e-g,** smFISH of adult human lung proximal and alveolar regions as indicated showing expression of IGSF21+ DC markers *GPR34* (white) and *IGSF21* (red) in e, EREG+ DC marker *EREG* (red) and general DC marker *GPR183* (white) in f, and TREM2+ DC markers *CHI3L1* (white) and *TREM2* (red) in g, all with DAPI (nuclear) counter stain (blue). (The non-punctate signal in red channel in panels

e and g is erythrocyte autofluorescence.) Scale bars, 20 µm. Arrowheads, double positive cells. Quantification shows distribution of each dendritic type; note strong proximal enrichment of IGSF21+ and EREG+ dendritic cells.

Extended Data Figure S6. **Global signaling interactions among human lung cell types inferred from expression patterns of ligands and cognate receptor genes.** Heatmaps showing number of interactions predicted by CellPhoneDB software between human lung cell types located in proximal lung regions (left panel in each pair) and distal regions (right panel) based on expression patterns of ligand genes ("Sending cell") and their cognate receptor genes ("Receiving cell") in the SS2 dataset. The pair of heatmaps at upper left show values for all predicted signaling interactions ("All interactions"), and other pairs show values for the indicated types of signals (growth factors, cytokines, integrins, WNT, Notch, Bmp, FGF, and TFGB). Predicted interactions between cell types range from 0 (lymphocyte signaling to neutrophils) to 136 (AdvF signaling to Cap-i1). Note expected relationships such as immune cells expressing integrins to interact with endothelial cells and having higher levels of cytokine signaling relative to their global signaling, and unexpected relationships such as fibroblasts expressing majority of growth factors and lack of Notch signaling originating from immune cells.

Extended Data Figure S7. Lung cell expression patterns of genes implicated in lung disease. Dot plots of expression of 233 lung disease genes curated from Genomewide Association Studies (GWAS) and Online Mendelian Inheritance in Man (OMIM).

Extended Data Figure S8. Lung cell expression patterns of viral entry receptors. a, Dot plot showing gene expression of entry receptors for non-respiratory viruses (compare with Fig. 6b showing expression values of receptors for respiratory viruses). **b,** Violin plots showing low but detectable gene expression of coronavirus receptors *ACE2*, *DPP4*, and *ANPEP* (left). Grey shading, cell types inhaled viruses can directly access. Right, Donut plots showing the relative number of expressing cells of cell types viruses can directly access (shaded grey in panel a), normalized by their abundance values from Table S1 (and refined by the relative abundance values in Figures 2 and S4). Note prevalence of AT2 and AT1 alveolar cells for *ACE2*, the receptor for SARS-CoV and SARS-CoV-2, and for *DPP4,* the receptor for MERS-CoV, in contrast to prevalence of macrophages for *ANPEP*, the receptor for the common cold causing coronavirus 229E.

Extended Data Figure S9. Comparison of mouse and human gene expression profiles in the homologous lung cell types. Scatter plots showing the median expression levels (ln(CPM+1)) in the indicated cell types of each expressed human gene and mouse ortholog in the mouse and human SS2 datasets. Note there are tens to hundreds of genes that show a 20-fold or greater expression difference (and p-value < 0.05) between species (red dots, with gene names indicated for some and total number given above). Bas/Ma 1 cells have the most differentially-expressed genes (343), and $CD4^+$ M/E T cells have the least (79). Correlation scores (R values) between the average mouse and human gene expression profiles for each cell type are indicated. “Mm()” and “Hs()”, genes where duplications between mouse and human were collapsed to HomologyID.

Extended Data Figure S10. Comparison of gene expression in homologous cell types in human and mouse by scRNAseq. a, Dot plot of expression of canonical goblet cell markers *MUC5B* and *MUC5AC* and transcription factor *SPDEF* in mouse and human goblet cells. **b-f**, Additional examples of the four scenarios (Type 0, 1, 2, 3) for evolution of cellular expression pattern as in Fig. 7d. **b,** Dot plots of expression of the indicated genes in mouse and human lung cell types showing conserved (Type 0) expression pattern **c,** Dot plots showing gain of expression (Type 1 change) in multiple human cell types of *RNASE1* (left panel) and all human cell types of *TRIM38* (right panel). **d,** Dot plots of **e**xpression of CGRP and ADM hormone receptor genes showing expansion of expression (Type 2 change) in human endothelial cells. **e,** Dot plots of expression of emphysema-associated gene *SERPINA1* showing switched expression (Type 3 change) from mouse pericytes (top) to human AT2 cells (bottom). **f,** Dot plots comparing expression and conservation of HHIP with those of other Hedgehog pathway genes including ligands (SHH, DHH, IHH), receptors (SMO, PTCH1, PTCH2), and transducers (GLI1, GLI2, GLI3). **g**, Alveolar region of adult mouse lung probed by smFISH for general alveolar epithelial marker *Nkx2-1*, AT2 marker *Sftpc*, and transcription factor *Myrf*. Note *Myrf* is selectively expressed in mouse AT1 cells (*Nkx2-1+ Sftpc-* cells), as it is in humans (Fig. 5c). Scale bar, 5 µm. **h,** Dot plots of expression of the serous cell markers *LTF*, *LYZ*, *BPIFBP1*, and *HP* showing switched expression (Type 3 change) from mouse airway epithelial cells to human serous cells, which mice lack. Dot plots of expression of lipid handling genes *APOE*, *PLIN2*, and *FST* show switched expression (Type 3 change) from mouse alveolar stromal cells to human lipofibroblasts, which mice lack. “Mm()” or “Hs()”, genes where duplications between mouse and human were collapsed to HomologyIDs. Red shading, diverged expression; blue shading, conserved expression.

Extended Data Figure S11. Few genes show >20-fold changes in expression in mouse lung cell types during aging. Scatter plots comparing median expression levels (ln(CPM+1)) in indicated lung cell types of each expressed gene in 3 month (x-axis) and 24 month (y-axis) mouse SS2 datasets from Tabula Muris Senis[56]. Correlation scores between average gene expression profile for each cell type at each age are indicated (R values), along with number of genes (red dots) showing 20-fold or greater expression difference (and p-value < 0.05) between ages. Names of some genes are given next to the corresponding red dot.

Supplemental Tables

Table S1. Canonical cell types (45) in the human lung and their abundances, markers, and available expression data.

Table S2. Human lung cell cluster identities and their abundances in each dataset.

Table S3. Surface markers used for canonical immune cell types in bulk mRNA sequencing.

Table S4. Enriched markers found in each cluster, with transcription factors, receptors/ligands, and disease associated genes annotated.

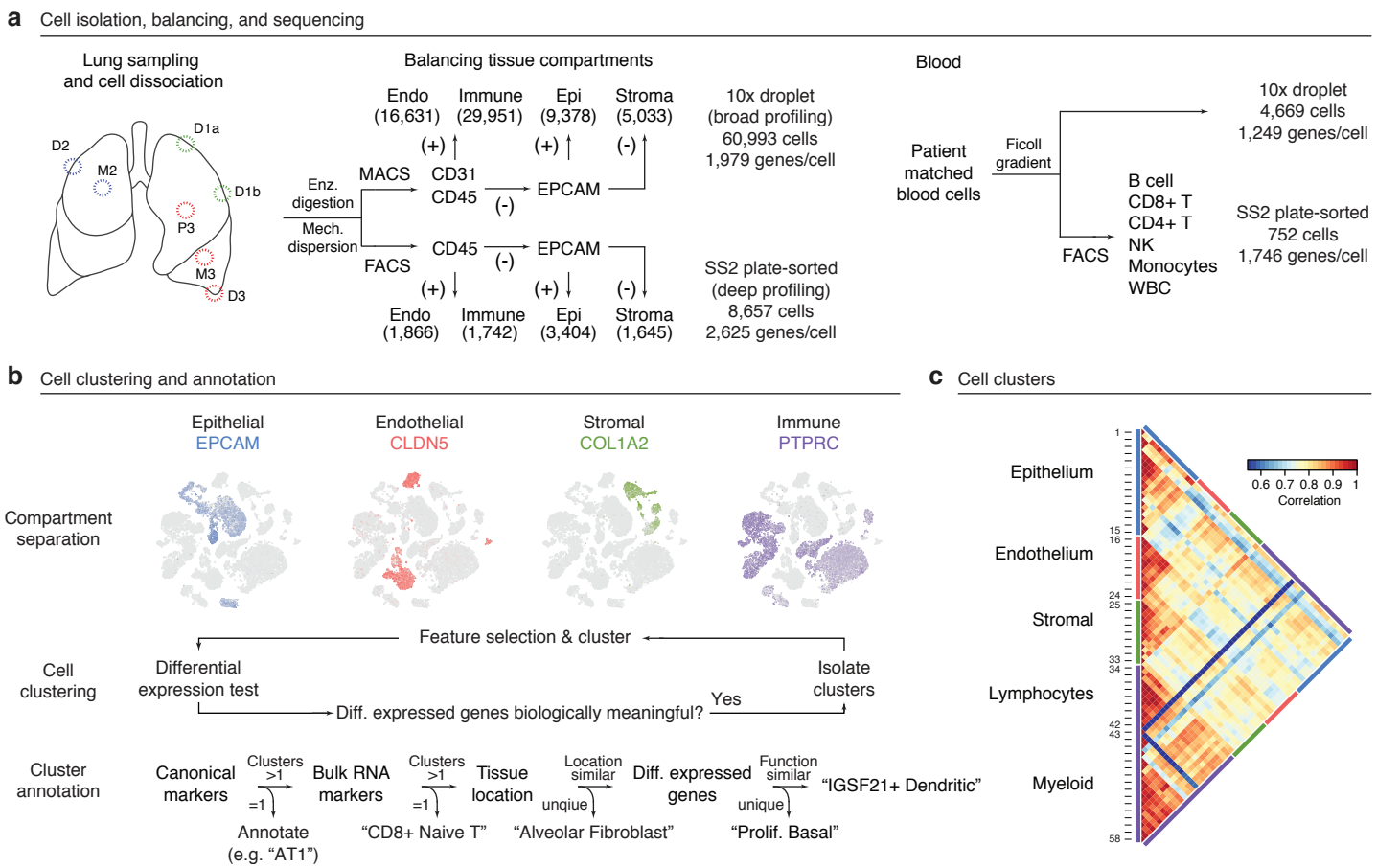
Table S5. P-value and scores of each CellPhoneDB Receptor-Ligand interaction from each cluster.

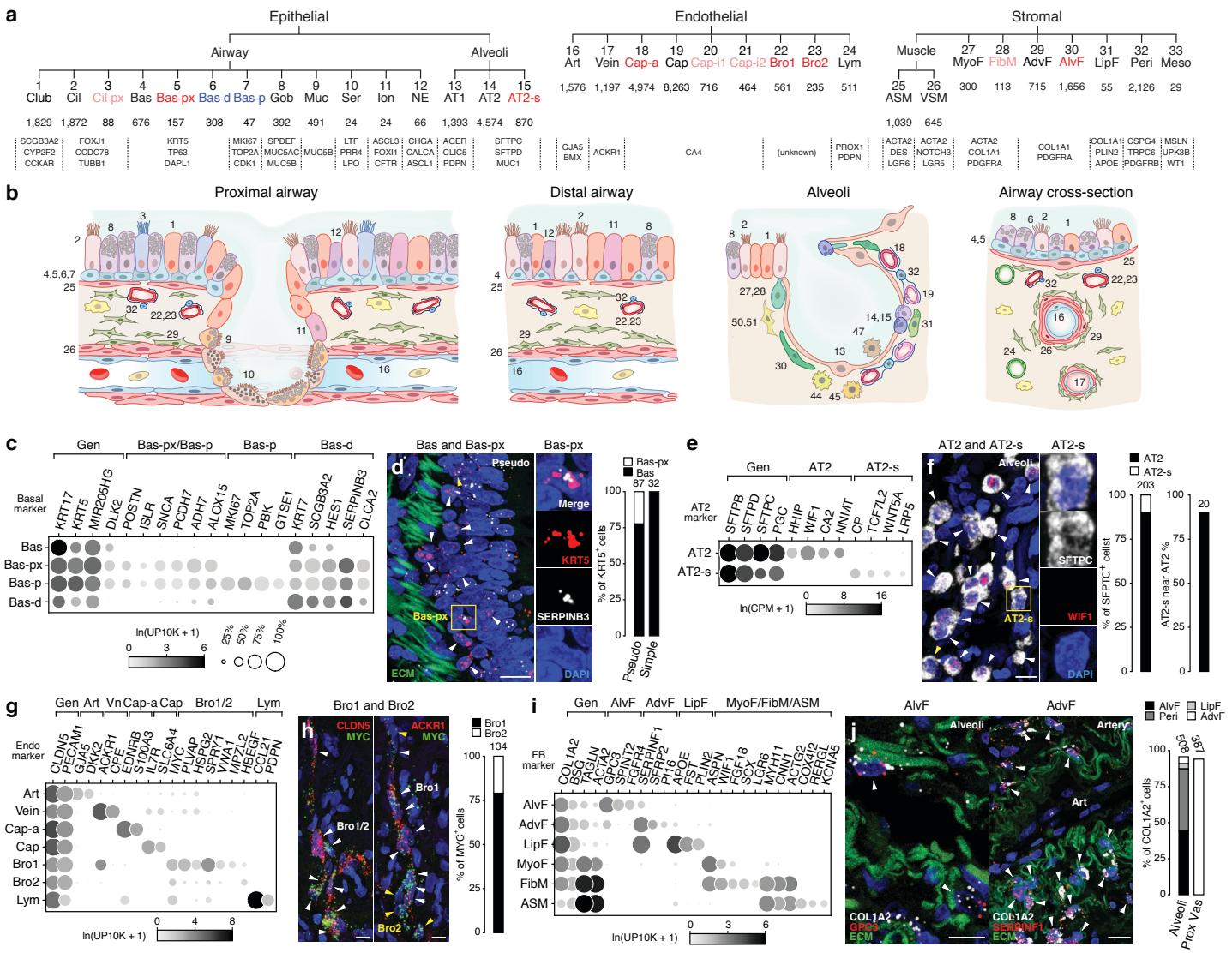
Table S6. Genes specific to mouse and human in each cluster and lung wide.

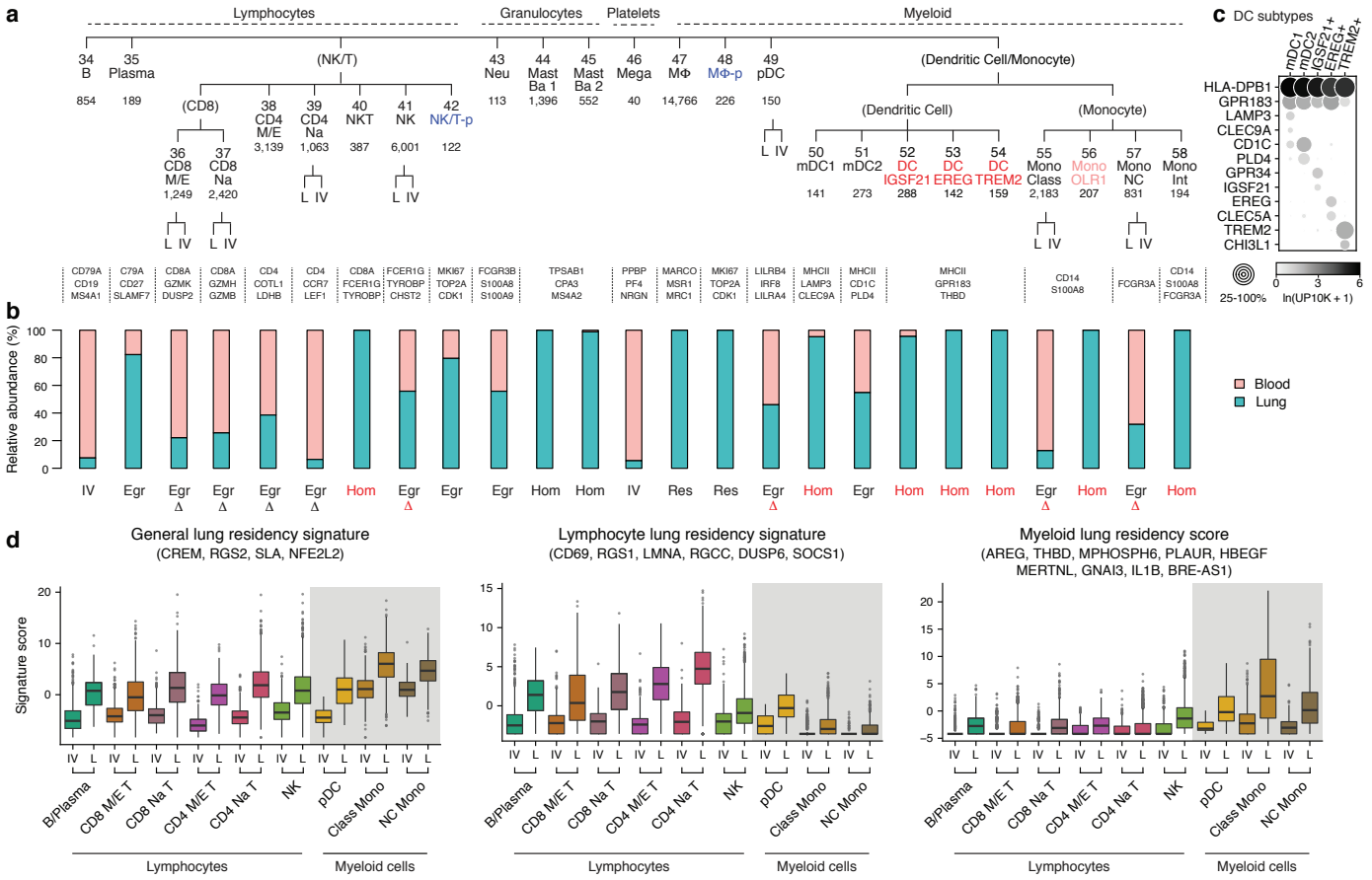
Table S7. Evolutionary changes in cellular patterns of lung gene expression between mouse and human.

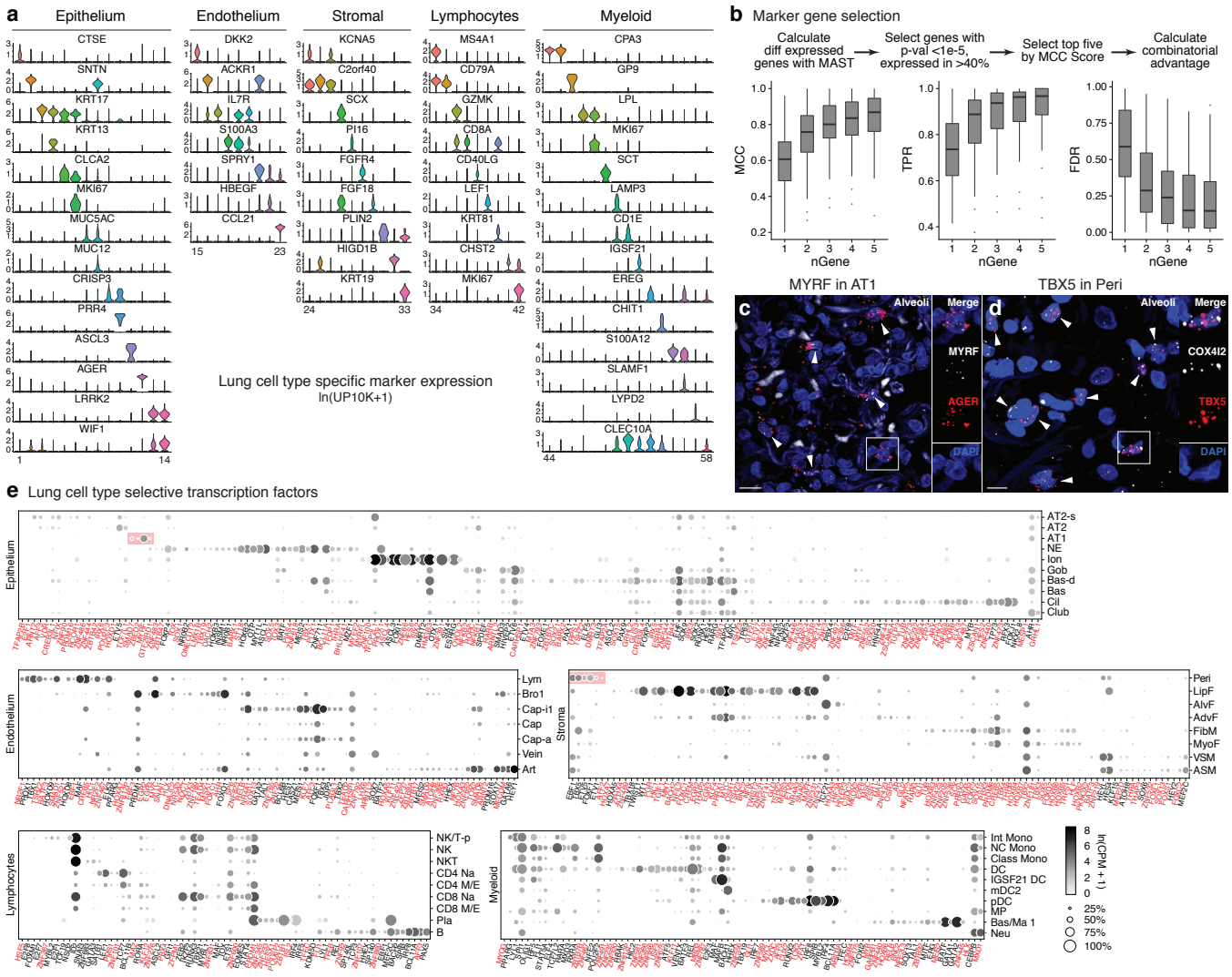
Table S8. Evolutionary and functional classes of each gene.

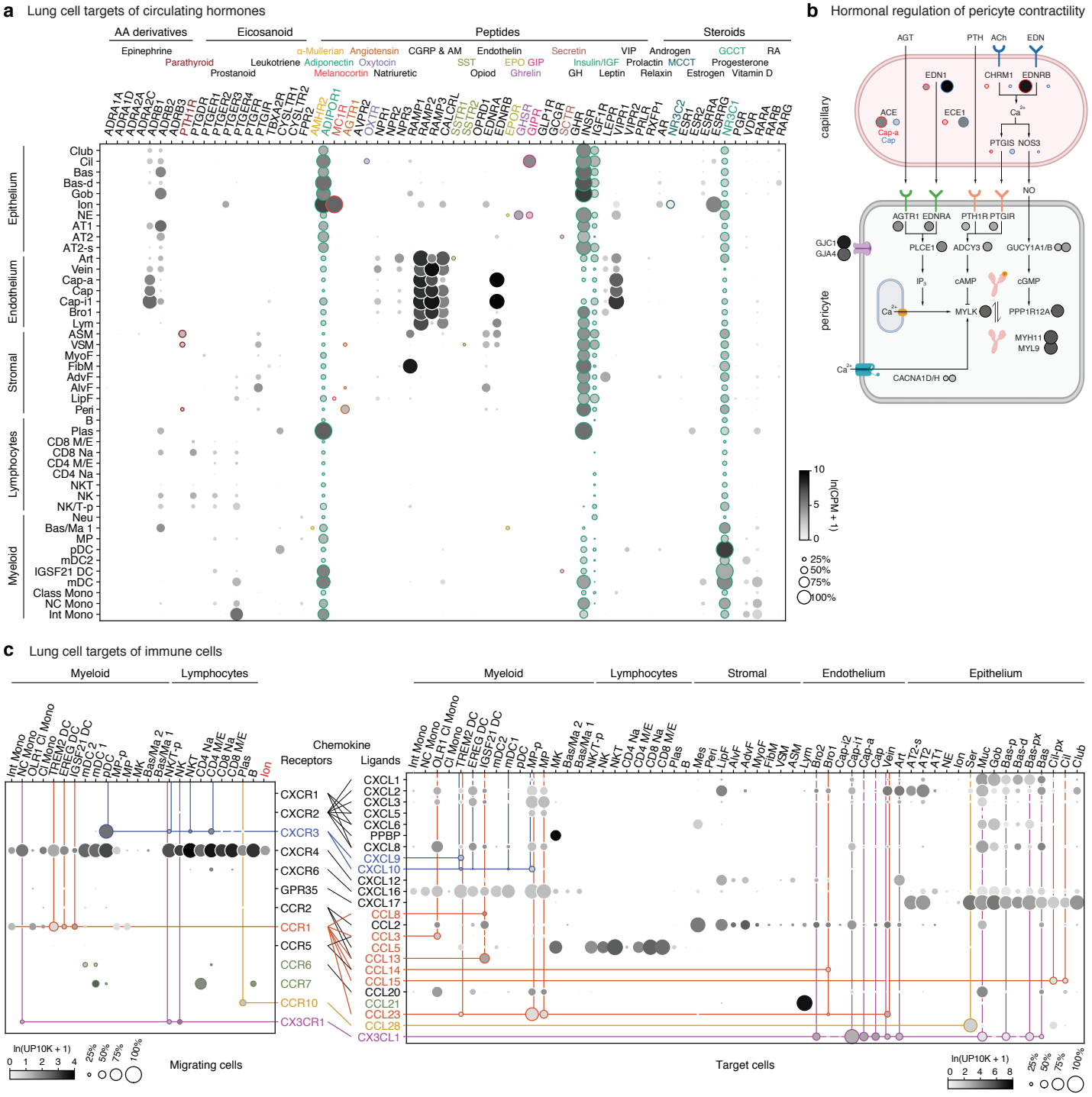
Figure 1, Travaglini et al

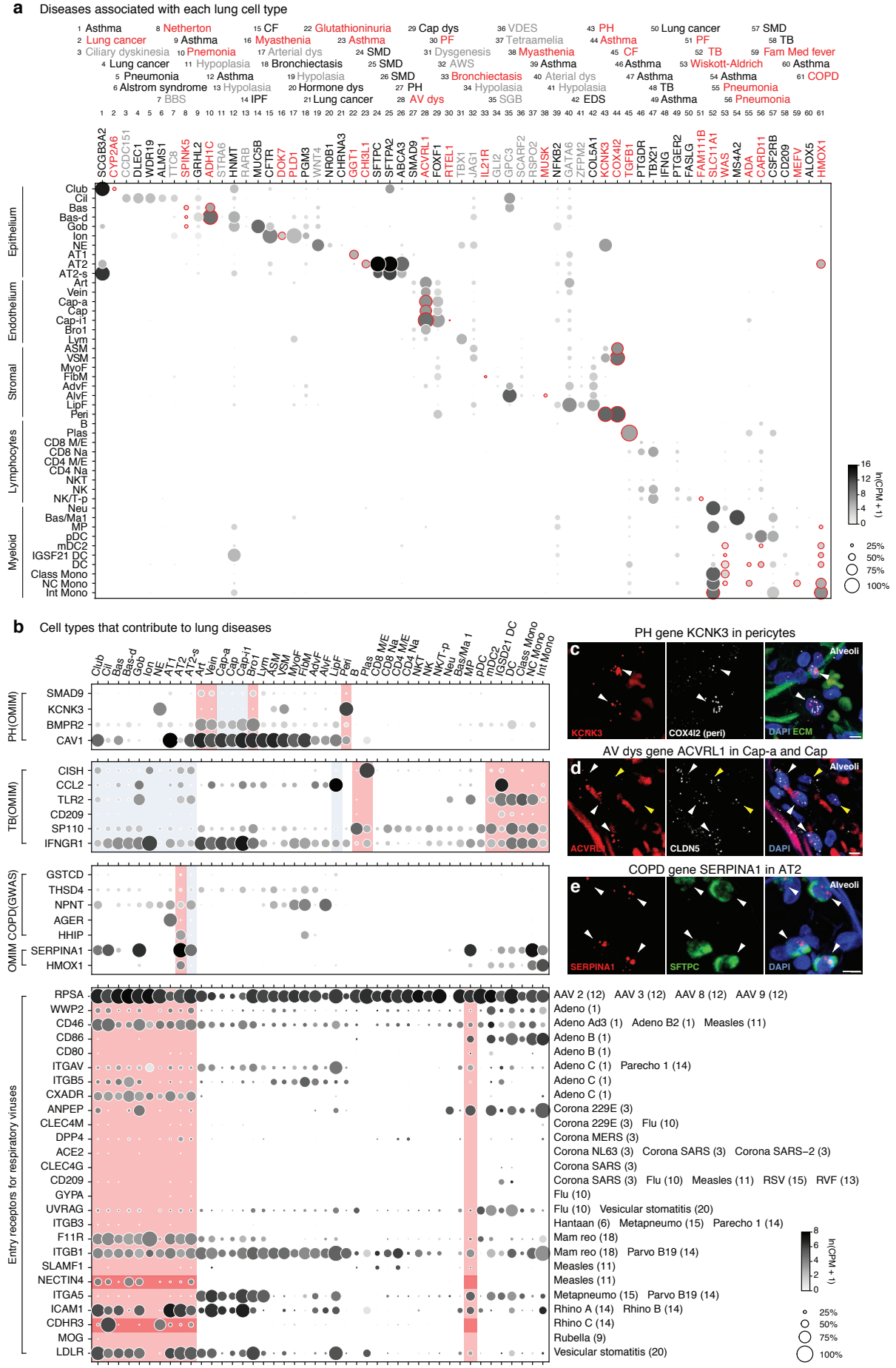


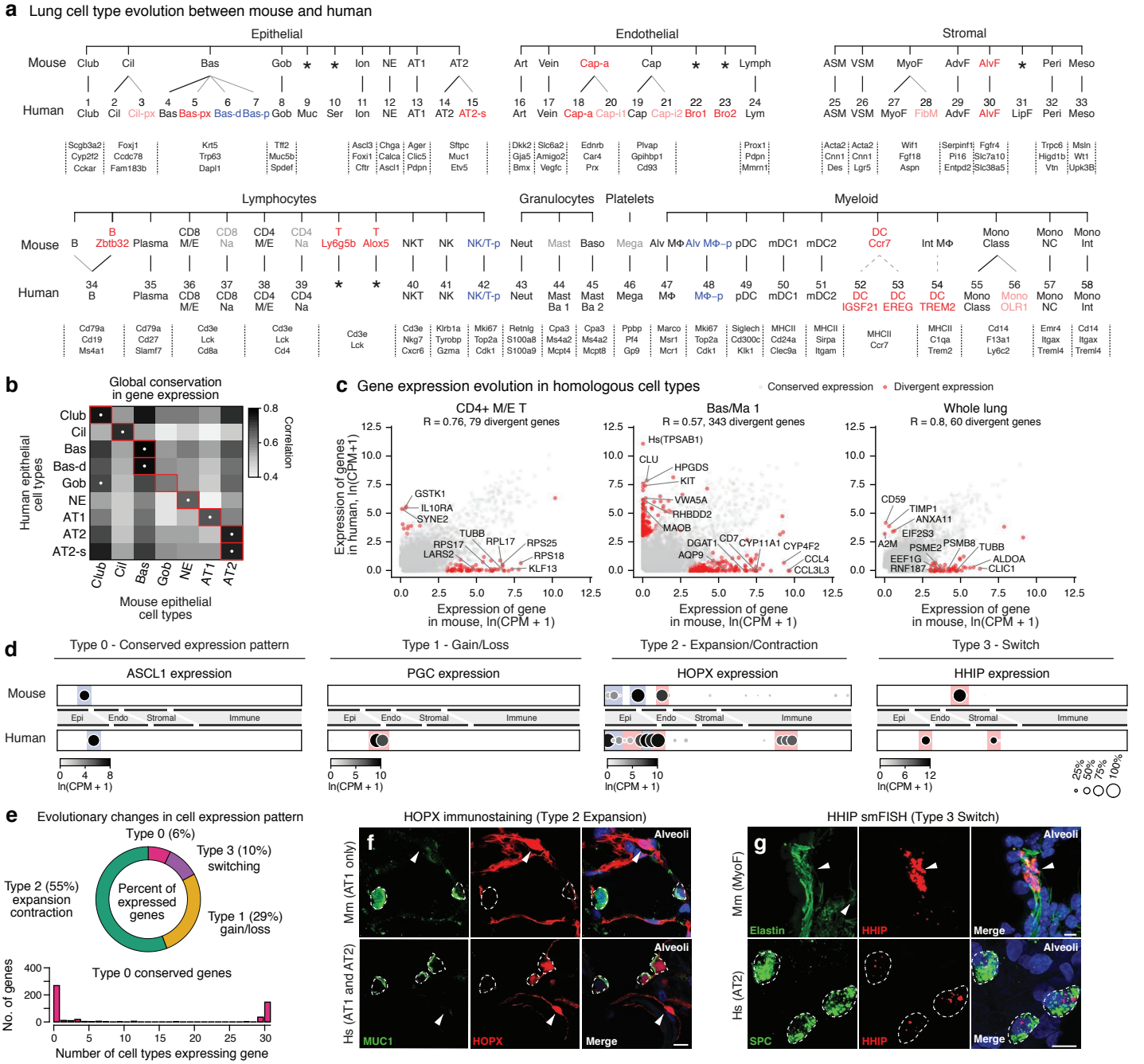

a
Epithelial
Airway
Alveoli
1 Club 1,829
2 Cil 1,872
3 Cil-px 88
4 Bas 676
5 Bas-px 157
6 Bas-d 308
7 Bas-p 47
8 Gob 392
9 Muc 491
10 Ser 24
11 Ion 24
12 NE 66
13 AT1 1,393
14 AT2 4,574
15 AT2-s 870
SCGB3A2 CYP2F2 CCKAR
FOXJ1 CCDC78 TUBB1
KRT5 TP63 DAPL1
MKI67 TOP2A CDK1
SPDEF MUC5AC MUC5B
MUC5B
LTF PRR4 LPO
ASCL3 FOXI1 CFTR
CHGA CALCA ASCL1
AGER CLIC5 PDPN
SFTPC SFTPD MUC1
Endothelial
16 Art 1,576
17 Vein 1,197
18 Cap-a 4,974
19 Cap 8,263
20 Cap-i1 716
21 Cap-i2 464
22 Bro1 561
23 Bro2 235
24 Lym 511
GJA5 BMX
ACKR1
CA4
(unknown)
PROX1 PDPN
Stromal
Muscle
25 ASM 1,039
26 VSM 645
27 MyoF 300
28 FibM 113
29 AdvF 715
30 AlvF 1,656
31 LipF 55
32 Peri 2,126
33 Meso 29
ACTA2 DES LGR6
ACTA2 NOTCH3 LGR5
ACTA2 COL1A1 PDGFRA
COL1A1 PDGFRA
COL1A1 PLIN2 APOE
CSPG4 TRPC6 PDGFRB
MSLN UPK3B WT1
b
Proximal airway
Distal airway
Alveoli
Airway cross-section
c
Gen
Bas-px/Bas-p
Bas-p
Bas-d
Basal marker
KRT17 KRT5 MIR205HG DLK2 POSTN ISLR SNCA PCDH7 ADH7 ALOX15 MKI67 TOP2A PBK GTSE1 KRT7 SCGB3A2 HES1 SERPINB3 CLCA2
Bas
Bas-px
Bas-p
Bas-d
ln(UP10K + 1)
25% 50% 75% 100%
d
Bas and Bas-px
Bas-px
Pseudo
Merge
KRT5
SERPINB3
DAPI
ECM
Bas-px
Bas
87 32
% of KRT5+ cells
Pseudo
Simple
e
Gen
AT2
AT2-s
AT2 marker
SFTPB SFTPD SFTPC PGC HHIP WIF1 CA2 NNMT CP TCF7L2 WNT5A LRP5
AT2
AT2-s
ln(CPM + 1)
f
AT2 and AT2-s
AT2-s
Alveoli
SFTPC
AT2-s
WIF1
DAPI
AT2
AT2-s
203
20
% of SFPTC+ cells
AT2-s near AT2 %
g
Gen Art Vn Cap-a Cap Bro1/2 Lym
Endo marker
CLDN5 PECAM1 GJA5 DKK2 ACKR1 CPE EDNRB S100A3 IL7R SLC6A4 MYC HLVAP HSPG2 SPRY1 WWA1 MPZL2 HBEGF CCL21 PDPN
Art
Vein
Cap-a
Cap
Bro1
Bro2
Lym
ln(UP10K + 1)
h
Bro1 and Bro2
CLDN5 MYC
ACKR1 MYC
Bro1
Bro1/2
Bro2
Bro1
Bro2
134
% of MYC+ cells
i
Gen
AlvF
AdvF
LipF
MyoF/FibM/ASM
FB marker
COL1A2 BGN TAGLN ACTA2 GPC3 SPINT2 FGFR4 SERPINF1 SFRP2 AR F3 APOE FS PLAUR ASPN WIF1 FGF18 SCX LGR6 MYH11 CNN1 ACTG2 COX4I2 RERGL KCNA5
AlvF
AdvF
LipF
MyoF
FibM
ASM
ln(UP10K + 1)
j
AlvF
AdvF
Alveoli
Artery
Art
COL1A2
GPC3
ECM
COL1A2
SERPINF1
ECM
AlvF
Peri
LipF
AdvF
605
387
% of COL1A2+ cells
Alveoli
Prox Vas


a
Lymphocytes
Granulocytes
Platelets
Myeloid
34 B 854
35 Plasma 189
(NK/T)
43 Neu 113
44 Mast Ba 1 1,396
45 Mast Ba 2 552
46 Mega 40
47 MΦ 14,766
48 MΦ-p 226
49 pDC 150
(CD8)
38 CD4 M/E 3,139
39 CD4 Na 1,063
40 NKT 387
41 NK 6,001
42 NK/T-p 122
36 CD8 M/E 1,249
37 CD8 Na 2,420
L IV
(Dendritic Cell/Monocyte)
(Dendritic Cell)
(Monocyte)
50 mDC1 141
51 mDC2 273
52 DC IGSF21 288
53 DC EREG 142
54 DC TREM2 159
55 Mono Class 2,183
56 Mono OLR1 207
57 Mono NC 831
58 Mono Int 194
CD79A CD19 MS4A1
C79A CD27 SLAMF7
CD8A GZMK DUSP2
CD8A GZMH GZMB
CD4 COTL1 LDHB
CD4 CCR7 LEF1
CD8A FCER1G TYROBP
FCER1G TYROBP CHST2
MKI67 TOP2A CDK1
FCGR3B S100A8 S100A9
TPSAB1 CPA3 MS4A2
PPBP PF4 NRGN
MARCO MSR1 MRC1
MKI67 TOP2A CDK1
LILRB4 IRF8 LILRA4
MHCII LAMP3 CLEC9A
MHCII CD1C PLD4
MHCII GPR183 THBD
CD14 S100A8
FCGR3A
CD14 S100A8 FCGR3A
c DC subtypes
mDC1
mDC2
IGSF21+
EREG+
TREM2+
HLA-DPB1
GPR183
LAMP3
CLEC9A
CD1C
PLD4
GPR34
IGSF21
EREG
CLEC5A
TREM2
CHI3L1
25-100%
0 3 6
ln(UP10K + 1)
b
Relative abundance (%)
100 80 60 40 20 0
Blood
Lung
IV Egr Egr Δ Egr Δ Egr Δ Egr Δ Hom Egr Δ Egr Egr Hom Hom IV Res Res Egr Δ Hom Egr Hom Hom Hom Egr Δ Hom Egr Δ Hom
d
General lung residency signature
(CREM, RGS2, SLA, NFE2L2)
Lymphocyte lung residency signature
(CD69, RGS1, LMNA, RGCC, DUSP6, SOCS1)
Myeloid lung residency score
(AREG, THBD, MPHOSPH6, PLAUR, HBEGF
MERTNL, GNAI3, IL1B, BRE-AS1)
Signature score
IV L
B/Plasma
CD8 M/E T
CD8 Na T
CD4 M/E T
CD4 Na T
NK
pDC
Class Mono
NC Mono
Lymphocytes
Myeloid cells

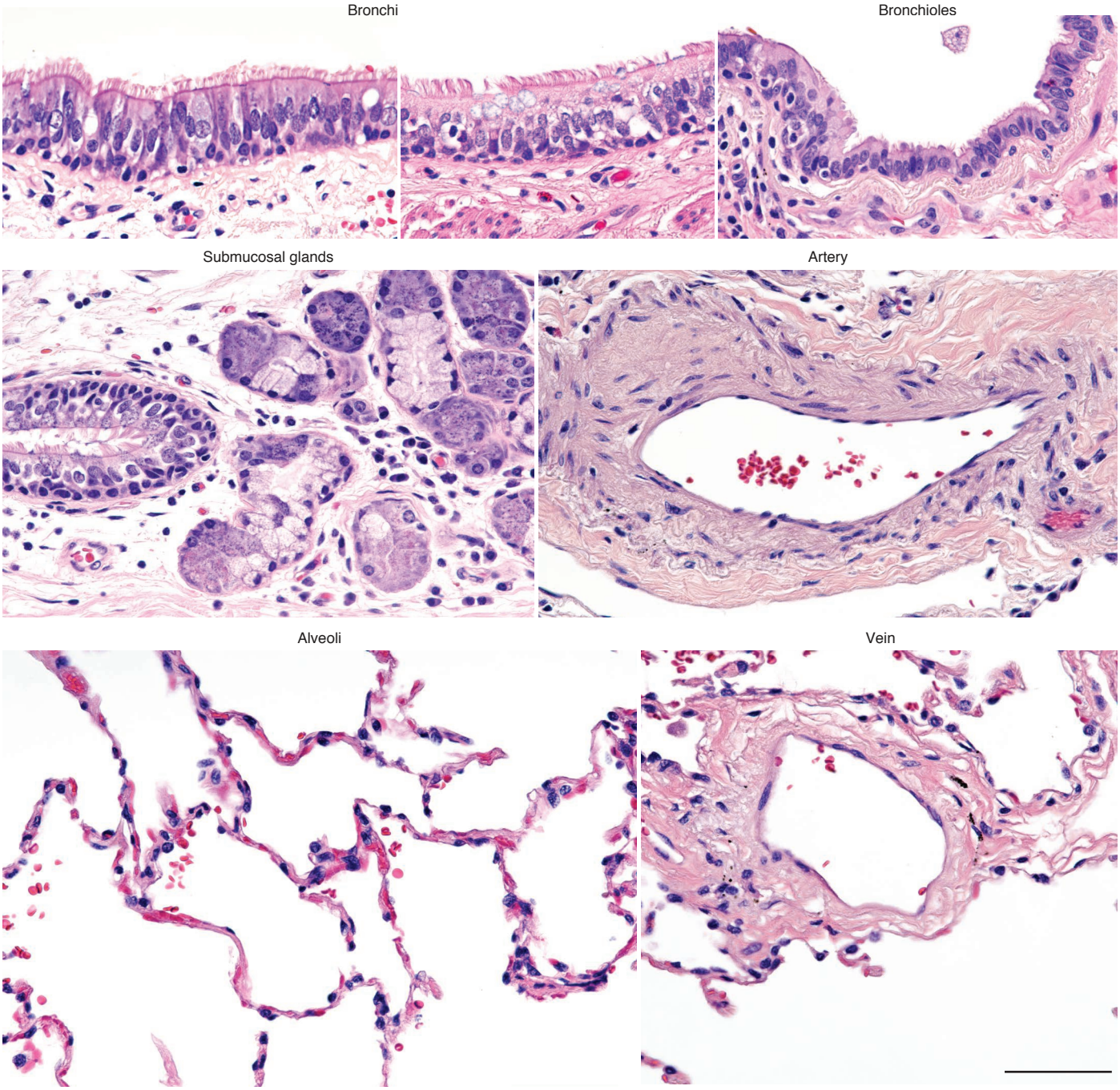

a
Epithelium
Endothelium
Stromal
Lymphocytes
Myeloid
CTSE
SNTN
KRT17
KRT13
CLCA2
MKI67
MUC5AC
MUC12
CRISP3
PRR4
ASCL3
AGER
LRRK2
WIF1
DKK2
ACKR1
IL7R
S100A3
SPRY1
HBEGF
CCL21
KCNA5
C2orf40
SCX
PI16
FGFR4
FGF18
PLIN2
HIGD1B
KRT19
MS4A1
CD79A
GZMK
CD8A
CD40LG
LEF1
KRT81
CHST2
MKI67
CPA3
GP9
LPL
MKI67
SCT
LAMP3
CD1E
IGSF21
EREG
CHIT1
S100A12
SLAMF1
LYPD2
CLEC10A
Lung cell type specific marker expression
ln(UP10K+1)
b
Marker gene selection
Calculate diff expressed genes with MAST
Select genes with p-val <1e-5, expressed in >40%
Select top five by MCC Score
Calculate combinatorial advantage
MCC
TPR
FDR
nGene
MYRF in AT1
TBX5 in Peri
c
d
Alveoli
Merge
MYRF
AGER
DAPI
COX4I2
TBX5
e
Lung cell type selective transcription factors
Epithelium
AT2-s
AT2
AT1
NE
Ion
Gob
Bas-d
Bas
Cil
Club
Endothelium
Lym
Bro1
Cap-i1
Cap
Cap-a
Vein
Art
Stroma
Peri
LipF
AlvF
AdvF
FibM
MyoF
vSM
ASM
Lymphocytes
NK/T-p
NK
NKT
CD4 Na
CD4 M/E
CD8 Na
CD8 M/E
Pla
B
Myeloid
Int Mono
NC Mono
Class Mono
DC
IGSF21 DC
mDC2
pDC
MP
Bas/Ma 1
Neu
ln(CPM + 1)
25%
50%
75%
100%


a
Lung cell targets of circulating hormones
AA derivatives
Eicosanoid
Peptides
Steroids
Epinephrine
Parathyroid
Prostanoid
Leukotriene
α-Mullerian
Angiotensin
Adiponectin
Oxytocin
Melanocortin
Natriuretic
CGRP & AM
Endothelin
SST
Opiod
EPO GIP
Ghrelin
Secretin
Insulin/IGF
GH
VIP
Prolactin
Leptin
Relaxin
Androgen
MCCT
Estrogen
GCCT
Progesterone
Vitamin D
RA
Epithelium
Endothelium
Stromal
Lymphocytes
Myeloid
Club
Cil
Bas
Bas-d
Gob
Ion
NE
AT1
AT2
AT2-s
Art
Vein
Cap-a
Cap
Cap-i1
Bro1
Lym
ASM
VSM
MyoF
FibM
AdvF
AlvF
LipF
Peri
B
Plas
CD8 M/E
CD8 Na
CD4 M/E
CD4 Na
NKT
NK
NK/T-p
Neu
Bas/Ma 1
MP
pDC
mDC2
IGSF21 DC
mDC
Class Mono
NC Mono
Int Mono
ln(CPM + 1)
25%
50%
75%
100%
b
Hormonal regulation of pericyte contractility
AGT
PTH
ACh
EDN
EDN1
CHRM1
EDNRB
capillary
ACE
Cap-a
Cap
ECE1
PTGIS
NOS3
NO
AGTR1
EDNRA
PTH1R
PTGIR
GJC1
GJA4
PLCE1
ADCY3
GUCY1A1/B
pericyte
cAMP
cGMP
MYLK
PPP1R12A
MYH11
MYL9
CACNA1D/H
c
Lung cell targets of immune cells
Myeloid
Lymphocytes
Stromal
Endothelium
Epithelium
Chemokine
Receptors
Ligands
CXCR1
CXCR2
CXCR3
CXCR4
CXCR6
GPR35
CCR2
CCR1
CCR5
CCR6
CCR7
CCR10
CX3CR1
CXCL1
CXCL2
CXCL3
CXCL5
CXCL6
PPBP
CXCL8
CXCL9
CXCL10
CXCL12
CXCL16
CXCL17
CCL8
CCL2
CCL3
CCL5
CCL13
CCL14
CCL15
CCL20
CCL21
CCL23
CCL28
CX3CL1
Migrating cells
Target cells
ln(UP10K + 1)

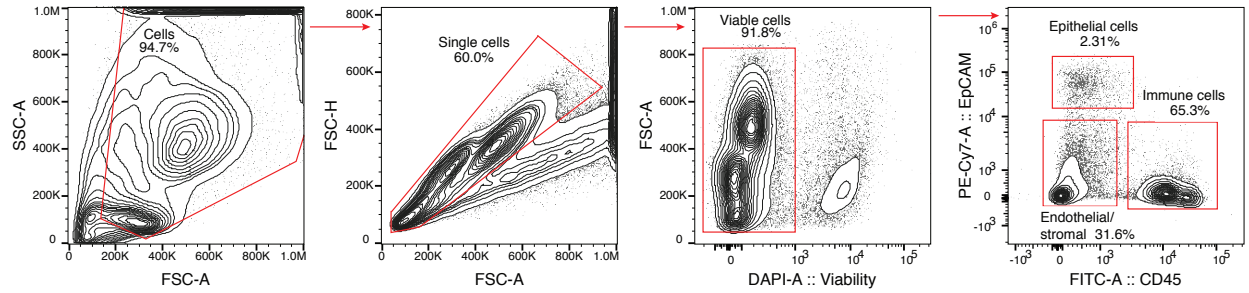

a
Diseases associated with each lung cell type
1 Asthma
2 Lung cancer
3 Ciliary dyskinesia
4 Lung cancer
5 Pneumonia
6 Alstrom syndrome
7 BBS
8 Netherton
9 Asthma
10 Pnemonia
11 Hypoplasia
12 Asthma
13 Hypolasia
14 IPF
15 CF
16 Myasthenia
17 Arterial dys
18 Bronchiectasis
19 Hypolasia
20 Hormone dys
21 Lung cancer
22 Glutathioninuria
23 Asthma
24 SMD
25 SMD
26 SMD
27 PH
28 AV dys
29 Cap dys
30 PF
31 Dysgenesis
32 AWS
33 Bronchiectasis
34 Hypolasia
35 SGB
36 VDES
37 Tetraamelia
38 Myasthenia
39 Asthma
40 Aterial dys
41 Hypolasia
42 EDS
43 PH
44 Asthma
45 CF
46 Asthma
47 Asthma
48 TB
49 Asthma
50 Lung cancer
51 PF
52 TB
53 Wiskott-Aldrich
54 Asthma
55 Pneumonia
56 Pneumonia
57 SMD
58 TB
59 Fam Med fever
60 Asthma
61 COPD
SCGB3A2
CYP2A6
CCDC151
DLEC1
WDR19
ALMS1
TTC8
SPINK5
GRHL2
ADH1C
STRA6
HNMT
RARB
MUC5B
CFTR
DOK7
PLD1
PGM3
WNT4
NR0B1
CHRNA3
GGT1
CHI3L1
SFTPC
SFTPA2
ABCA3
SMAD9
ACVRL1
FOXF1
RTEL1
TBX1
JAG1
IL21R
GLI2
GPC3
SCARF2
RSPO2
MUSK
NFKB2
GATA6
ZFPM2
COL5A1
KCNK3
COX4I2
TGFB1
PTGDR
TBX21
IFNG
PTGER2
FASLG
FAM111B
SLC11A1
WAS
MS4A2
ADA
CARD11
CSF2RB
CD209
MEFV
ALOX5
HMOX1
Epithelium
Endothelium
Stromal
Lymphocytes
Myeloid
Club
Cil
Bas
Bas-d
Gob
Ion
NE
AT1
AT2
AT2-s
Art
Vein
Cap-a
Cap
Cap-i1
Bro1
Lym
ASM
VSM
MyoF
FibM
AdvF
AlvF
LipF
Peri
B
Plas
CD8 M/E
CD8 Na
CD4 M/E
CD4 Na
NKT
NK
NK/T-p
Neu
Bas/Ma1
MP
pDC
mDC2
IGSF21 DC
DC
Class Mono
NC Mono
Int Mono
ln(CPM + 1)
16
12
8
4
0
25%
50%
75%
100%
b
Cell types that contribute to lung diseases
PH(OMIM)
SMAD9
KCNK3
BMPR2
CAV1
TB(OMIM)
CISH
CCL2
TLR2
CD209
SP110
IFNGR1
OMIM COPD(GWAS)
GSTCD
THSD4
NPNT
AGER
HHIP
SERPINA1
HMOX1
Entry receptors for respiratory viruses
RPSA
WWP2
CD46
CD86
CD80
ITGAV
ITGB5
CXADR
ANPEP
CLEC4M
DPP4
ACE2
CLEC4G
CD209
GYPA
UVRAG
ITGB3
F11R
ITGB1
SLAMF1
NECTIN4
ITGA5
ICAM1
CDHR3
MOG
LDLR
AAV 2 (12) AAV 3 (12) AAV 8 (12) AAV 9 (12)
Adeno (1)
Adeno Ad3 (1) Adeno B2 (1) Measles (11)
Adeno B (1)
Adeno B (1)
Adeno C (1) Parecho 1 (14)
Adeno C (1)
Adeno C (1)
Corona 229E (3)
Corona 229E (3) Flu (10)
Corona MERS (3)
Corona NL63 (3) Corona SARS (3) Corona SARS–2 (3)
Corona SARS (3)
Corona SARS (3) Flu (10) Measles (11) RSV (15) RVF (13)
Flu (10)
Flu (10) Vesicular stomatitis (20)
Hantaan (6) Metapneumo (15) Parecho 1 (14)
Mam reo (18)
Mam reo (18) Parvo B19 (14)
Measles (11)
Measles (11)
Metapneumo (15) Parvo B19 (14)
Rhino A (14) Rhino B (14)
Rhino C (14)
Rubella (9)
Vesicular stomatitis (20)
ln(CPM + 1)
8
6
4
2
0
25%
50%
75%
100%
PH gene KCNK3 in pericytes
c
KCNK3
COX4I2 (peri)
Alveoli
DAPI ECM
AV dys gene ACVRL1 in Cap-a and Cap
d
ACVRL1
CLDN5
Alveoli
DAPI
COPD gene SERPINA1 in AT2
e
SERPINA1
SFTPC
Alveoli
DAPI


a Lung cell type evolution between mouse and human
Epithelial
Endothelial
Stromal
Mouse
Human
Lymphocytes
Granulocytes
Platelets
Myeloid
b Global conservation in gene expression
Human epithelial cell types
Mouse epithelial cell types
Correlation
c Gene expression evolution in homologous cell types
Conserved expression
Divergent expression
CD4+ M/E T
R = 0.76, 79 divergent genes
Bas/Ma 1
R = 0.57, 343 divergent genes
Whole lung
R = 0.8, 60 divergent genes
Expression of genes in human, ln(CPM+1)
Expression of gene in mouse, ln(CPM + 1)
d
Type 0 - Conserved expression pattern
Type 1 - Gain/Loss
Type 2 - Expansion/Contraction
Type 3 - Switch
ASCL1 expression
PGC expression
HOPX expression
HHIP expression
Epi
Endo
Stromal
Immune
ln(CPM + 1)
e Evolutionary changes in cell expression pattern
Type 0 (6%)
Type 3 (10%) switching
Type 2 (55%) expansion contraction
Type 1 (29%) gain/loss
Percent of expressed genes
Type 0 conserved genes
No. of genes
Number of cell types expressing gene
HOPX immunostaining (Type 2 Expansion)
HHIP smFISH (Type 3 Switch)
Mm (AT1 only)
Hs (AT1 and AT2)
Mm (MyoF)
Hs (AT2)
Alveoli
MUC1
HOPX
Merge
Elastin
HHIP
SPC

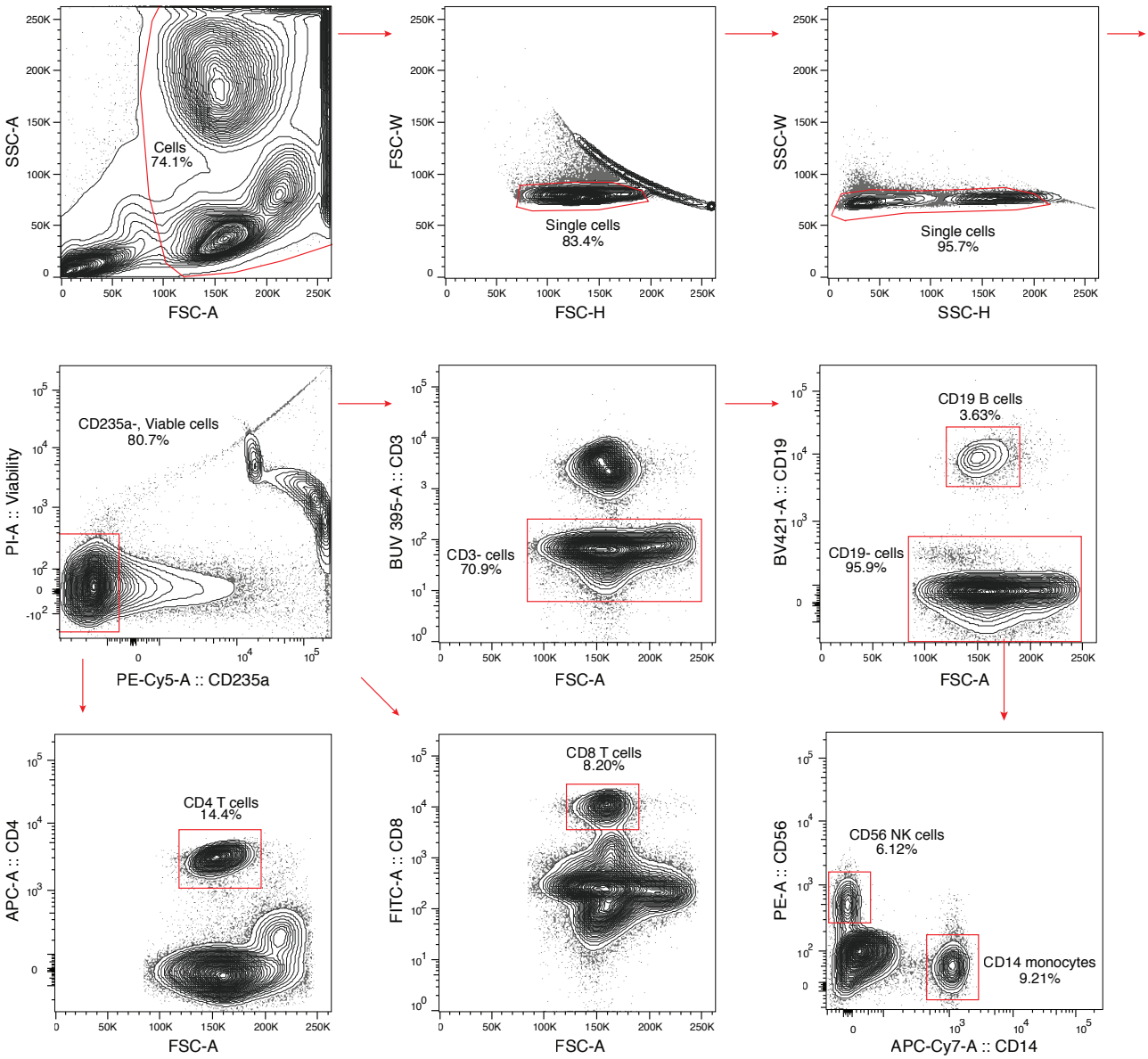
Extended Data Figure S1, Travaglini et al

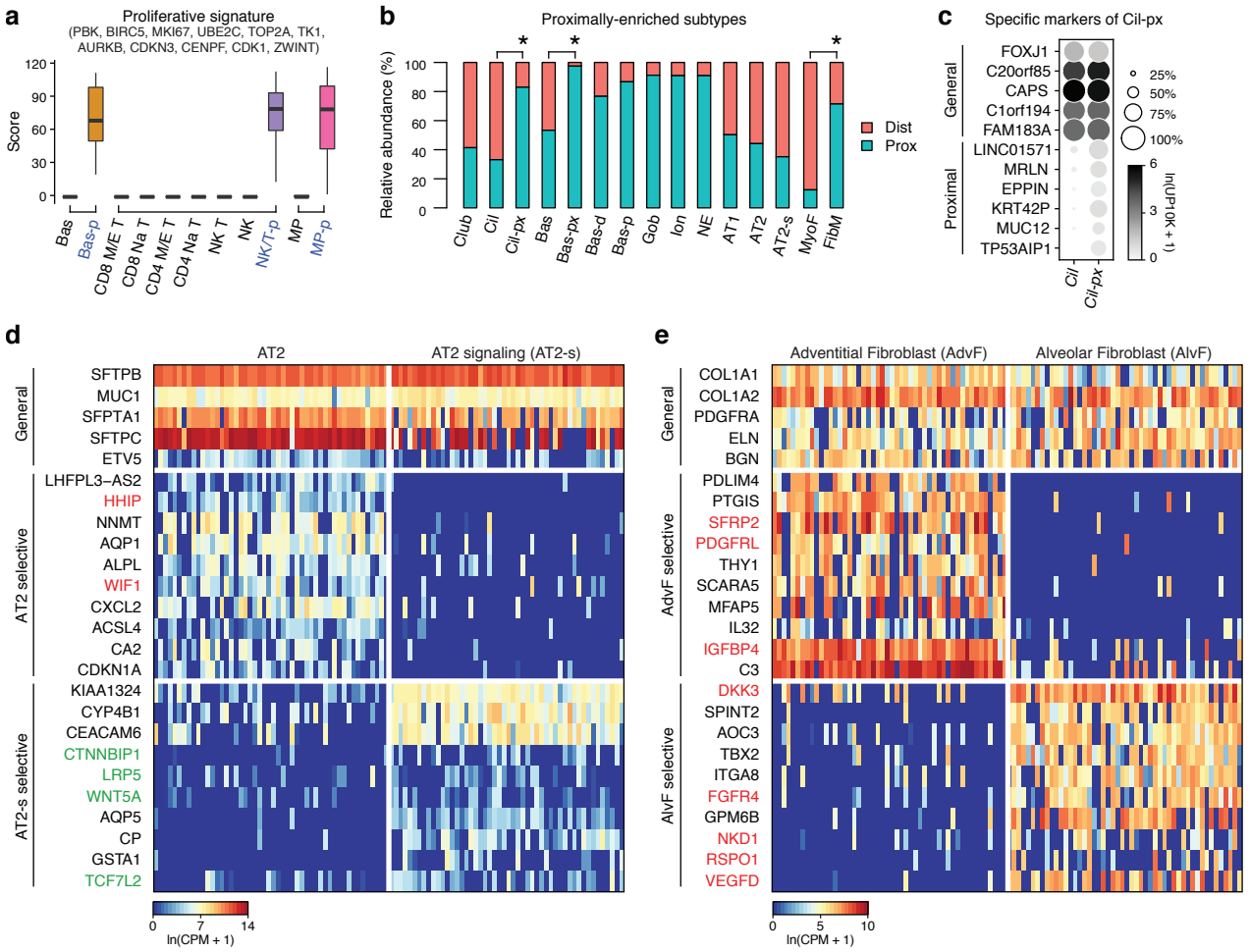


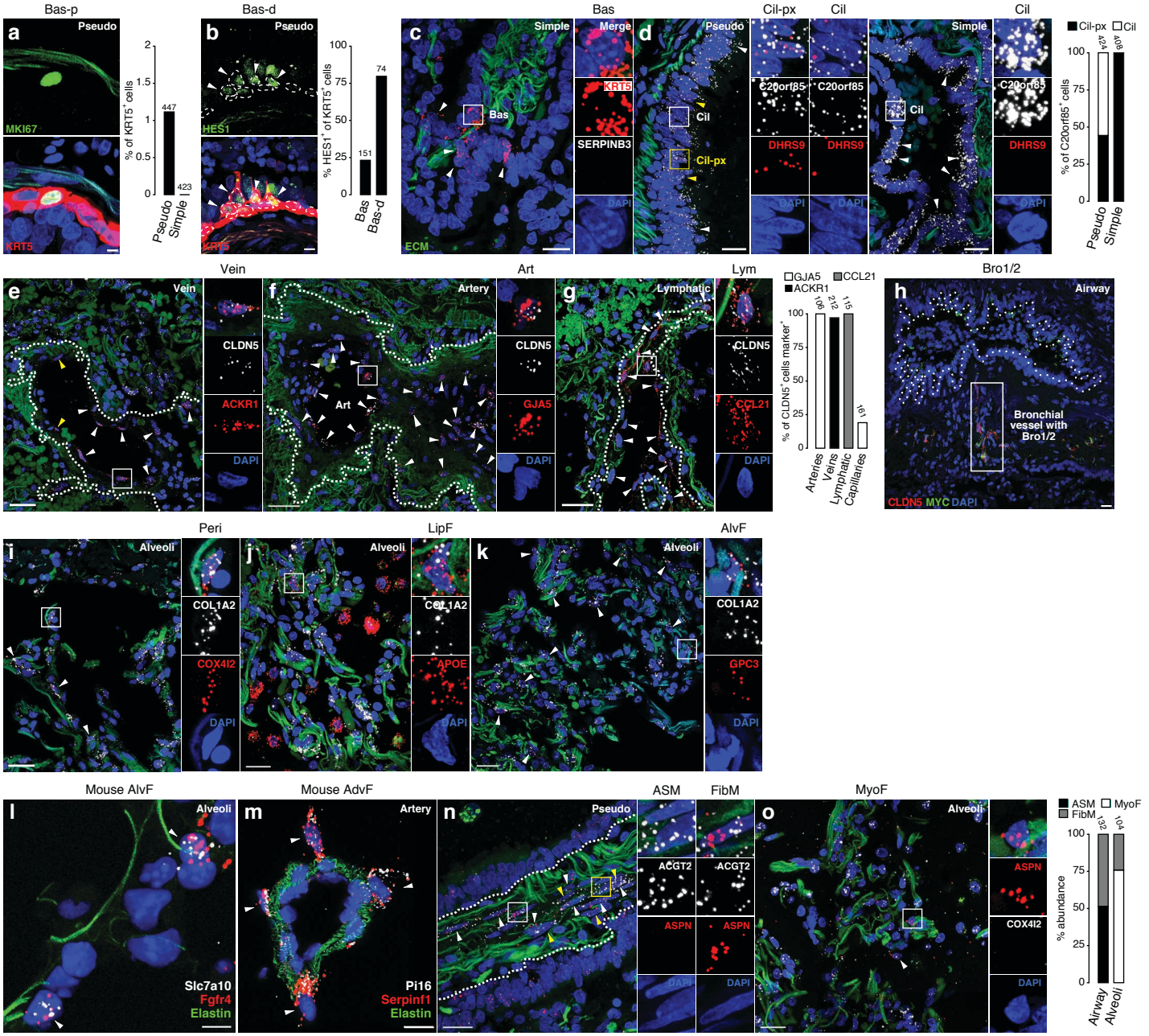
a Human lung

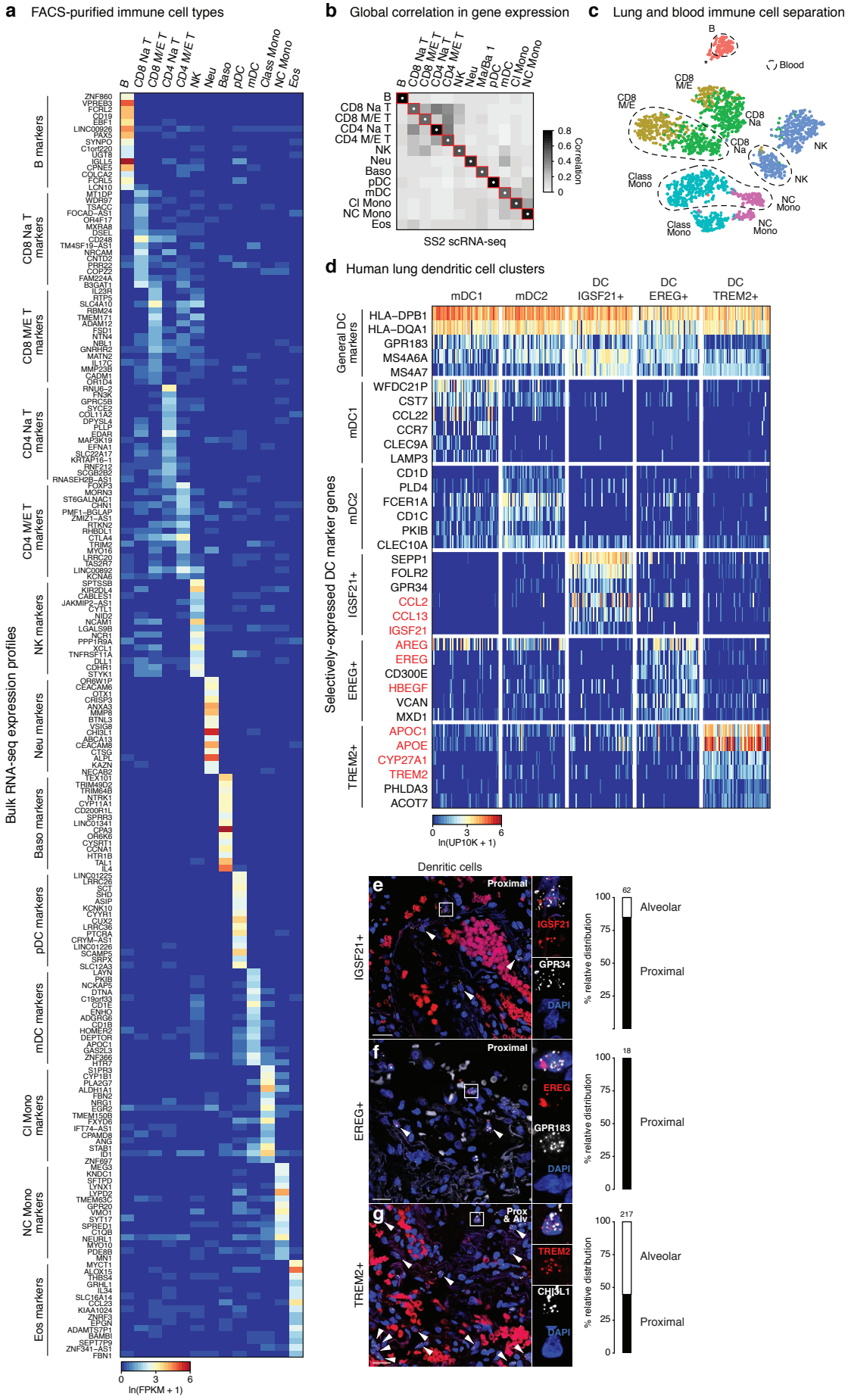


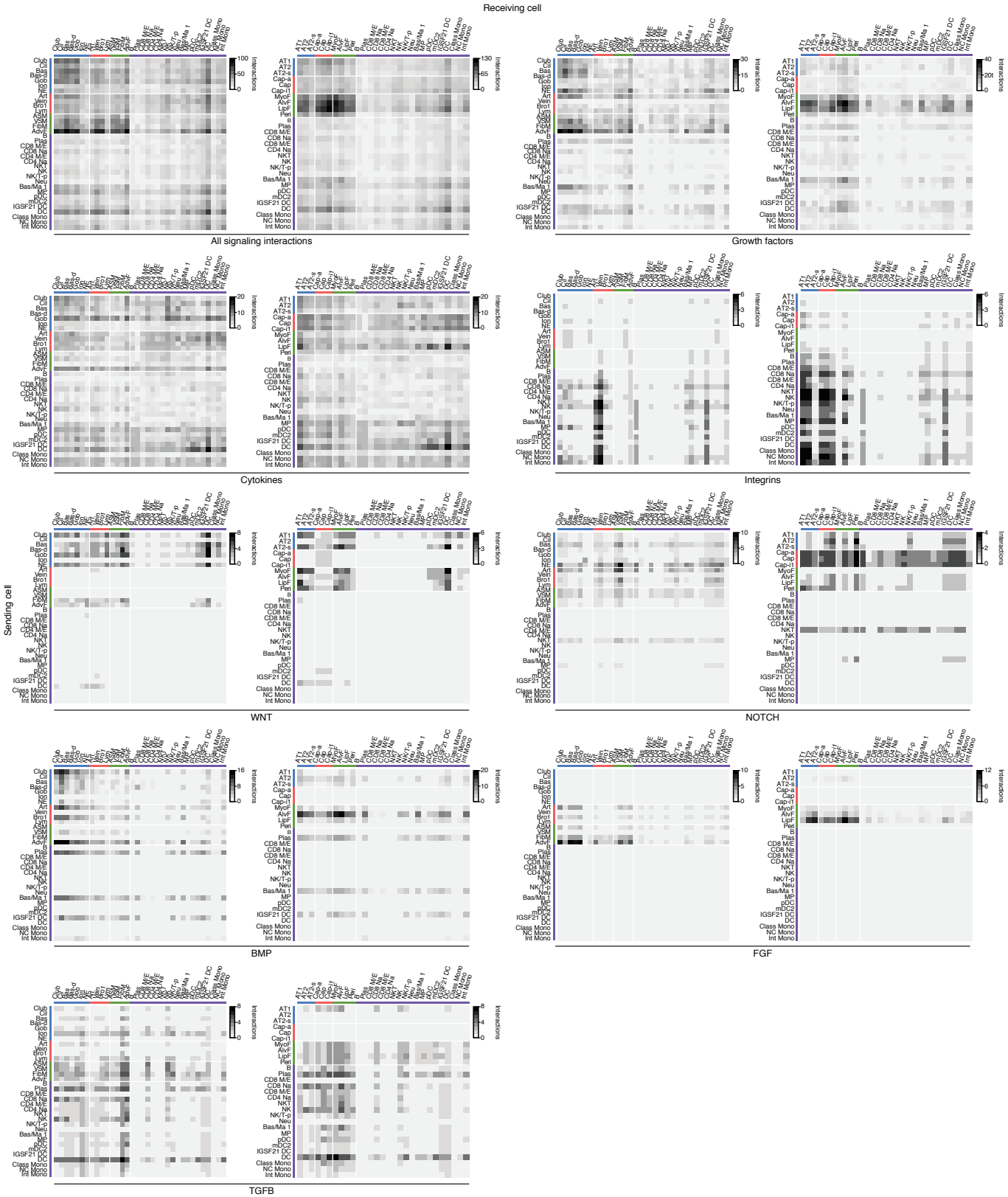
b Human blood

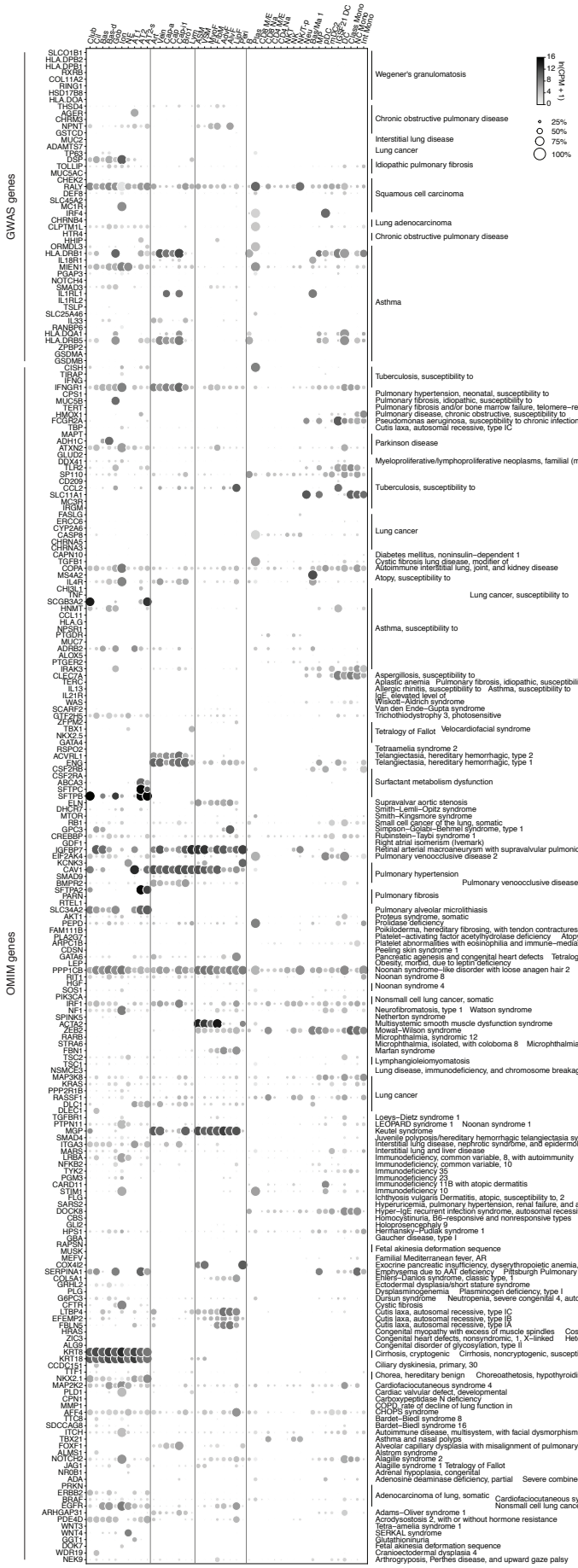



a
Proliferative signature
(PBK, BIRC5, MKI67, UBE2C, TOP2A, TK1, AURKB, CDKN3, CENPF, CDK1, ZWINT)
Score
Bas
Bas-p
CD8 M/E T
CD8 Na T
CD4 M/E T
CD4 Na T
NK T
NK
NK/T-p
MP
MP-p
b
Proximally-enriched subtypes
Relative abundance (%)
Dist
Prox
Club
Cil
Cil-px
Bas
Bas-px
Bas-d
Bas-p
Gob
Ion
NE
AT1
AT2
AT2-s
MyoF
FibM
c
Specific markers of Cil-px
General
Proximal
FOXJ1
C20orf85
CAPS
C1orf194
FAM183A
LINC01571
MRLN
EPPIN
KRT42P
MUC12
TP53AIP1
25%
50%
75%
100%
ln(UP10K + 1)
d
AT2
AT2 signaling (AT2-s)
General
AT2 selective
AT2-s selective
SFTPB
MUC1
SFPTA1
SFTPC
ETV5
LHFPL3–AS2
HHIP
NNMT
AQP1
ALPL
WIF1
CXCL2
ACSL4
CA2
CDKN1A
KIAA1324
CYP4B1
CEACAM6
CTNNBIP1
LRP5
WNT5A
AQP5
CP
GSTA1
TCF7L2
ln(CPM + 1)
e
Adventitial Fibroblast (AdvF)
Alveolar Fibroblast (AlvF)
General
AdvF selective
AlvF selective
COL1A1
COL1A2
PDGFRA
ELN
BGN
PDLIM4
PTGIS
SFRP2
PDGFRL
THY1
SCARA5
MFAP5
IL32
IGFBP4
C3
DKK3
SPINT2
AOC3
TBX2
ITGA8
FGFR4
GPM6B
NKD1
RSPO1
VEGFD
ln(CPM + 1)

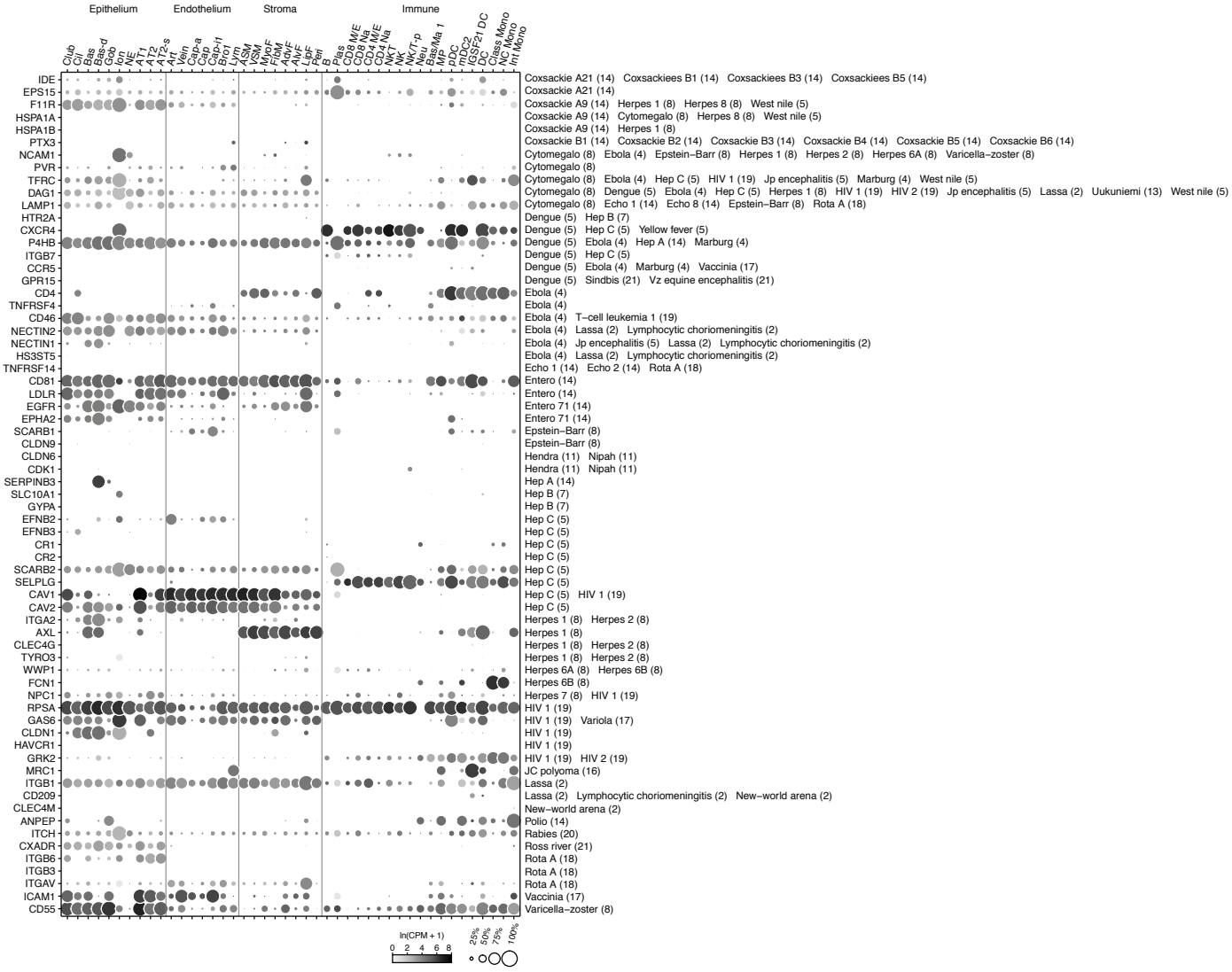

Bas-p
Bas-d
Bas
Cil-px
Cil
Vein
Art
Lym
Bro1/2
Peri
LipF
AlvF
Mouse AlvF
Mouse AdvF
ASM
FibM
MyoF
Pseudo
Simple
Merge
Artery
Lymphatic
Airway
Alveoli
MKI67
KRT5
HES1
SERPINB3
ECM
C20orf85
DHRS9
DAPI
CLDN5
ACKR1
GJA5
CCL21
COL1A2
COX4I2
APOE
GPC3
ACGT2
ASPN
Bronchial vessel with Bro1/2
CLDN5 MYC DAPI
Slc7a10
Fgfr4
Elastin
Pi16
Serpinf1
% of KRT5+ cells
% HES1+ of KRT5+ cells
% of C20orf85+ cells
% of CLDN5+ cells marker+
% abundance
Arteries
Veins
Lymphatic Capillaries
Pseudo
Simple
Bas
Bas-d
Airway
Alveoli


a
FACS-purified immune cell types
B
CD8 Na T
CD8 M/E T
CD4 Na T
CD4 M/E T
NK
Neu
Baso
pDC
mDC
Class Mono
NC Mono
Eos
Bulk RNA-seq expression profiles
B markers
CD8 Na T markers
CD8 M/E T markers
CD4 Na T markers
CD4 M/E T markers
NK markers
Neu markers
Baso markers
pDC markers
mDC markers
Cl Mono markers
NC Mono markers
Eos markers
0
3
6
ln(FPKM + 1)
b
Global correlation in gene expression
B
CD8 Na T
CD8 M/E T
CD4 Na T
CD4 M/E T
NK
Neu
Baso
pDC
mDC
Cl Mono
NC Mono
Eos
Ma/Ba 1
0.8
0.6
0.4
0.2
0
Correlation
SS2 scRNA-seq
c
Lung and blood immune cell separation
B
Blood
CD8 M/E
CD8 Na
NK
Class Mono
NC Mono
d
Human lung dendritic cell markers
mDC1
mDC2
DC IGSF21+
DC EREG+
DC TREM2+
General DC markers
HLA–DPB1
HLA–DQA1
GPR183
MS4A6A
MS4A7
mDC1
WFDC21P
CST7
CCL22
CCR7
CLEC9A
LAMP3
mDC2
CD1D
PLD4
FCER1A
CD1C
PKIB
CLEC10A
IGSF21+
SEPP1
FOLR2
GPR34
CCL2
CCL13
IGSF21
EREG+
AREG
EREG
CD300E
HBEGF
VCAN
MXD1
TREM2+
APOC1
APOE
CYP27A1
TREM2
PHLDA3
ACOT7
Selectively-expressed DC marker genes
0
3
6
ln(UP10K + 1)
Denritic cells
e
Proximal
IGSF21
GPR34
DAPI
IGSF21+
f
Proximal
EREG
GPR183
DAPI
EREG+
g
Prox & Alv
TREM2
CHI3L1
DAPI
TREM2+
% relative distribution
100
75
50
25
0
62
Alveolar
Proximal
18
Proximal
217
Alveolar
Proximal

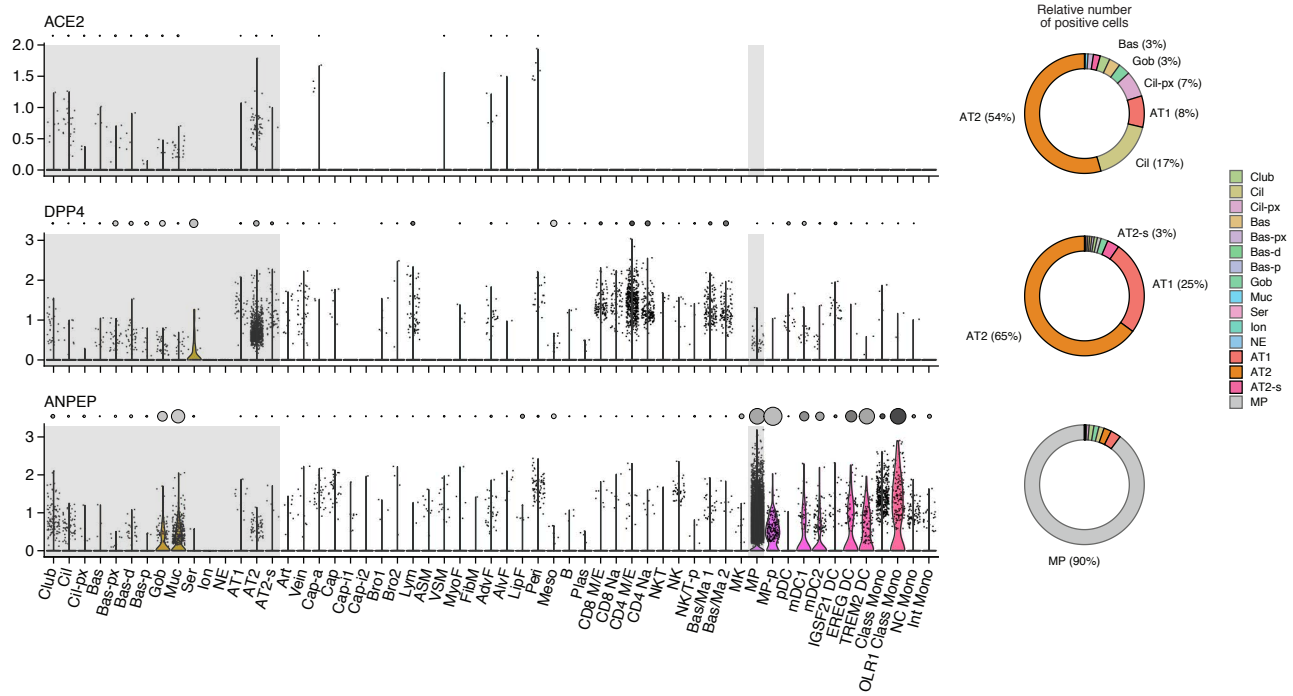

Receiving cell
Sending cell
All signaling interactions
Growth factors
Cytokines
Integrins
WNT
NOTCH
BMP
FGF
TGFB
Interactions

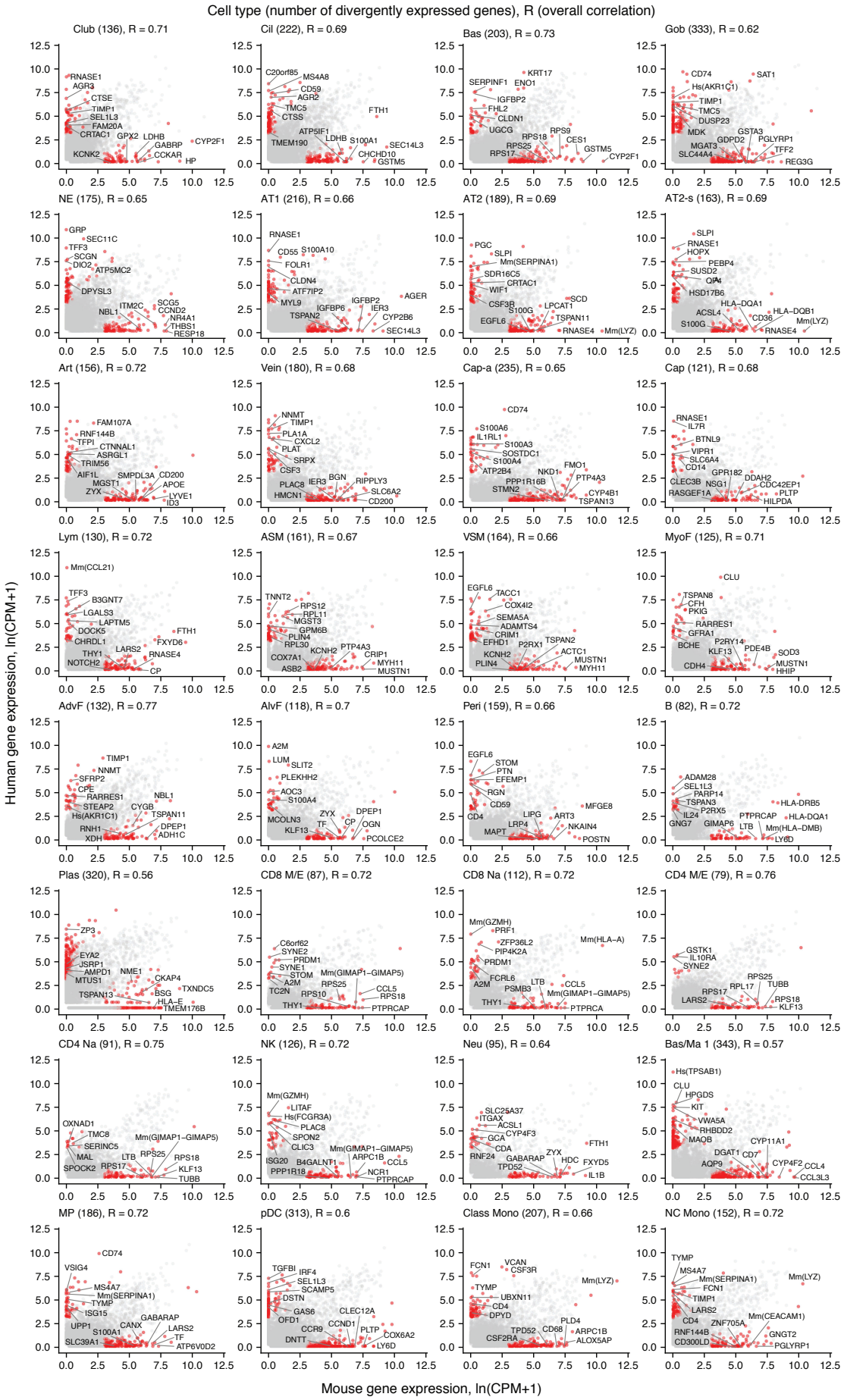

Ciliated
Club
Basal
Basal cycl
Goblet
NE
AT1
AT2
AT2-s
Stem
Mesothe
Cap
Cap-11
B
Plasma
PGC
MyofF
MyofP
AdvF
AdvP
Fibr
Fibr MYE
Fibr NE
CD4 M
CD4 N
CD8 M
CD8 N
NK
Neu
Neu-p
Mast
Mac Ma 1
Mac Ma 2
mTOC
pDC
CD1C
CD141
CD16
Fibr Mono
Int Mono
Inf Mono
GWAS genes
OMIM genes
SLCO1B1
HLA.DPB2
HLA.DPB1
RXRB
COL11A2
RIN01
HSD17B8
HLA.DOA
THSD4
AGER
CHRM3
NPNT
GSTCD
MUC2
ADAMTS7
TP63
DSP
TOLLIP
MUC5AC
CHEK2
RALY
DEF8
SLC45A2
MC1R
IRF4
CHRNB4
CLPTM1L
HTR4
HHIP
ORMDL3
HLA.DRB1
IL18R1
MIEN1
PGAP3
NOTCH4
SMAD3
IL1RL1
IL1RL2
TSLP
SLC25A46
IL33
RANBP6
HLA.DQA1
HLA.DRB5
ZPBP2
GSDMA
GSDMB
CISH
TIRAP
IFNG
IFNGR1
CPS1
MUC5B
TERT
HMOX1
FCGR2A
TBP
MAPT
ADH1C
ATXN2
GLID2
DDX41
TLR2
SP110
CD209
CCL2
SLC11A1
MC3R1
IRGM
FASLG
ERCC6
CYP2A6
CASP8
CHRNA5
CHRNA3
CAPN10
TGFB1
COPA
MS4A2
IL4R
CHI3L1
TNF
SCGB3A2
HNMT
CCL11
HLA.G
NPSR1
PTGDR
MUC7
ADRB2
ALOX5
PTGER2
IRAK3
CLEC7A
TERC
IL13
IL21R
WAS
SCARF2
GJC2/P15
ZFPM2
TBX1
NKX2.5
GATA4
RSPO2
ACVRL1
ENG
CSF2RB
CSF2RA
ABCA3
SFTPC
SFTPB
ELN
DHCR7
MTOR
RB1
GPC3
CREBBP
GDF1
IGFBP7
EIF2AK4
KCNK3
CAV1
SMAD9
BMPR2
SFTPA2
PARN
RTEL1
SLC34A2
AKT1
PEPD
FAM111B
PLA2G7
ARPC1B
CDSN
GATA6
LEP
PPP1CB
RIJ
TGF
SOS1
PIK3CA
IRF1
NF1
SPINK5
ACTA2
ZEB2
RARB
STRA6
FBN1
TSC2
TSC1
NSMCE3
MAP3K8
KRAS
PPP2R1B
RASSF1
DLC1
DLEC1
TGFBR1
PTPN11
MGP
SMAD4
ITGA3
MARS
LHBA
NFKB2
TYK2
PGM3
CARD11
STIM1
FLG
SARS2
DOCK8
CBS
GLI2
HPS1
GBA
RAPSN
MUSK
MEFV
COX4I2
SERPINA1
COL5A1
GRPL2
PLG
G6PC3
CFTR
LTBP4
EFEMP2
FBLN5
HRAS
ZIC3
ALG9
KRT8
KRT18
CCDC181
TTF1
NKX2.1
MAP2K2
PLD1
CPN1
MMP1
RFC1
TTC8
SDCCAG8
ITCH
TBX21
FOXF1
ALMS1
NOTCH2
JAG1
NR0B1
ADA
PRKN
ERBB2
BRAF
EGFR
ARHGAP31
PDE4D
WNT3
WNT4
GGT1
DOK7
WDR19
NEK9
Wegener's granulomatosis
Chronic obstructive pulmonary disease
Interstitial lung disease
Lung cancer
Idiopathic pulmonary fibrosis
Squamous cell carcinoma
Lung adenocarcinoma
Chronic obstructive pulmonary disease
Asthma
Tuberculosis, susceptibility to
Pulmonary hypertension, neonatal, susceptibility to
Pulmonary fibrosis, idiopathic, susceptibility to
Pulmonary fibrosis and/or bone marrow failure, telomere–related, 1
Pulmonary disease, chronic obstructive, susceptibility to
Pseudomonas aeruginosa, susceptibility to chronic infection by, in cystic fibrosis
Cutis laxa, autosomal recessive, type IC
Parkinson disease
Myeloproliferative/lymphoproliferative neoplasms, familial (multiple types), susceptibility to
Tuberculosis, susceptibility to
Lung cancer
Diabetes mellitus, noninsulin–dependent 1
Cystic fibrosis lung disease, modifier of
Autoimmune interstitial lung, joint, and kidney disease
Atopy, susceptibility to
Lung cancer, susceptibility to
Asthma, susceptibility to
Aspergillosis, susceptibility to
Aplastic anemia Pulmonary fibrosis, idiopathic, susceptibility to
Allergic rhinitis, susceptibility to Asthma, susceptibility to
IgE, elevated level of
Wiskott–Aldrich syndrome
Van den Ende–Gupta syndrome
Trichothiodystrophy 3, photosensitive
Tetralogy of Fallot Velocardiofacial syndrome
Tetraamelia syndrome 2
Telangiectasia, hereditary hemorrhagic, type 2
Telangiectasia, hereditary hemorrhagic, type 1
Surfactant metabolism dysfunction
Supravalvar aortic stenosis
Smith–Lemli–Opitz syndrome
Smith–Kingsmore syndrome
Small cell cancer of the lung, somatic
Simpson–Golabi–Behmel syndrome, type 1
Rubinstein–Taybi syndrome 1
Right atrial isomerism (Ivemark)
Retinal arterial macroaneurysm with supravalvular pulmonic stenosis
Pulmonary venoocclusive disease 2
Pulmonary hypertension
Pulmonary venoocclusive disease 1
Pulmonary fibrosis
Pulmonary alveolar microlithiasis
Proteus syndrome, somatic
Prolidase deficiency
Pokiloderma, hereditary fibrosing, with tendon contractures, myopathy, and pulmonary fibrosis
Platelet–activating factor acetylhydrolase deficiency Atopy, susceptibility to Asthma, susceptibility to
Platelet abnormalities with eosinophilia and immune–mediated inflammatory disease
Peeling skin syndrome 1
Pancreatic agenesis and congenital heart defects Tetralogy of Fallot
Obesity, morbid, due to leptin deficiency
Noonan syndrome–like disorder with loose anagen hair 2
Noonan syndrome 8
Noonan syndrome 4
Nonsmall cell lung cancer, somatic
Neurofibromatosis, type 1 Watson syndrome
Netherton syndrome
Multisystemic smooth muscle dysfunction syndrome
Mowat–Wilson syndrome
Microphthalmia, syndromic 12
Microphthalmia, isolated, with coloboma 8 Microphthalmia, syndromic 9
Marfan syndrome
Lymphangioleiomyomatosis
Lung disease, immunodeficiency, and chromosome breakage syndrome
Lung cancer
Loeys–Dietz syndrome 1
LEOPARD syndrome 1 Noonan syndrome 1
Keutel syndrome
Juvenile polyposis/hereditary hemorrhagic telangiectasia syndrome
Interstitial lung disease, nephrotic syndrome, and epidermolysis bullosa, congenital
Interstitial lung and liver disease
Immunodeficiency, common variable, 8, with autoimmunity
Immunodeficiency, common variable, 10
Immunodeficiency 35
Immunodeficiency 23
Immunodeficiency 11B with atopic dermatitis
Immunodeficiency 10
Ichthyosis vulgaris Dermatitis, atopic, susceptibility to, 2
Hyperuricemia, pulmonary hypertension, renal failure, and alkalosis
Hyper–IgE recurrent infection syndrome, autosomal recessive
Homocystinuria, B6–responsive and nonresponsive types Thrombosis, hyperhomocysteinemic
Holoprosencephaly 9
Hermansky–Pudlak syndrome 1
Gaucher disease, type I
Fetal akinesia deformation sequence
Familial Mediterranean fever, AR
Exocrine pancreatic insufficiency, dyserythropoietic anemia, and calvarial hyperostosis
Emphysema due to AAT deficiency Pittsburgh Pulmonary disease, chronic obstructive, susceptibility to
Ehlers–Danlos syndrome, classic type, 1
Ectodermal dysplasia/short stature syndrome
Dysplasminogenemia Plasminogen deficiency, type I
Dursun syndrome Neutropenia, severe congenital 4, autosomal recessive
Cystic fibrosis
Cutis laxa, autosomal recessive, type IC
Cutis laxa, autosomal recessive, type IB
Cutis laxa, autosomal recessive, type IA
Congenital myopathy with excess of muscle spindles Costello syndrome
Congenital heart defects, nonsyndromic, 1, X–linked Heterotaxy, visceral, 1, X–linked
Congenital disorder of glycosylation, type II
Cirrhosis, cryptogenic Cirrhosis, noncryptogenic, susceptibility to
Ciliary dyskinesia, primary, 30
Chorea, hereditary benign Choreoathetosis, hypothyroidism, and neonatal respiratory distress
Cardiofaciocutaneous syndrome 4
Cardiac valvular defect, developmental
Carboxypeptidase N deficiency
COPD, rate of decline of lung function in
CHOPS syndrome
Bardet–Biedl syndrome 8
Bardet–Biedl syndrome 16
Autoimmune disease, multisystem, with facial dysmorphism
Asthma and nasal polyps
Alveolar capillary dysplasia with misalignment of pulmonary veins
Alstrom syndrome
Alagille syndrome 2
Alagille syndrome 1 Tetralogy of Fallot
Adrenal hypoplasia, congenital
Adenosine deaminase deficiency, partial Severe combined immunodeficiency due to ADA deficiency
Adenocarcinoma of lung, somatic
Cardiofaciocutaneous syndrome
Nonsmall cell lung cancer, susceptibility to
Adams–Oliver syndrome 1
Acrodysostosis 2, with or without hormone resistance
Tetra–amelia syndrome 1
SERKAL syndrome
Glutathionuria
Fetal akinesia deformation sequence
Craniectodermal dysplasia 4
Arthrogryposis, Perthes disease, and upward gaze palsy
16
12
8
4
0
ln(CPM + 1)
25%
50%
75%
100%

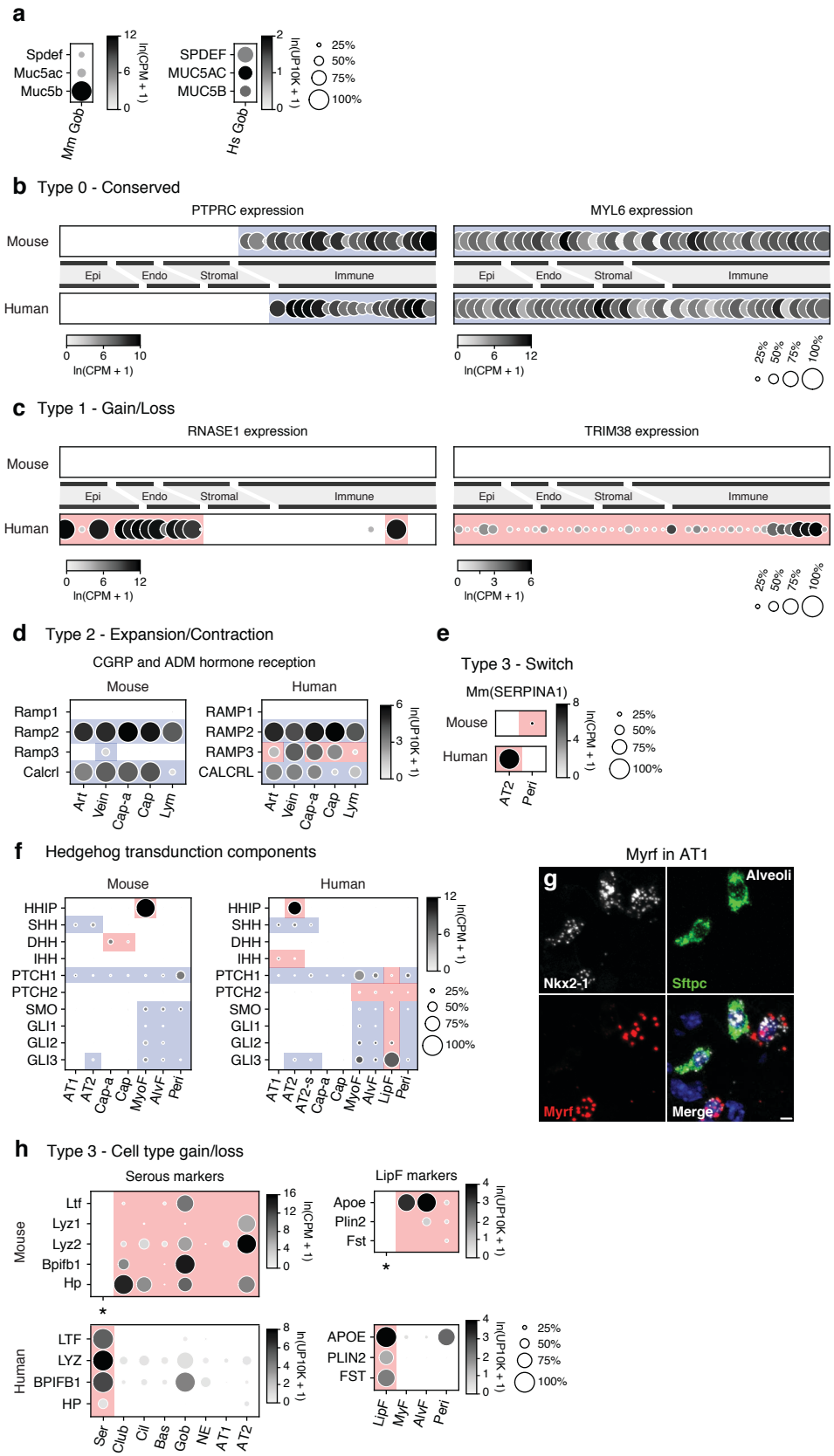
a Entry receptors for non-respiratory viruses



b Expression of SARS, SARS-2, MERS, NL63, and 229E coronavirus receptors




Cell type (number of divergently expressed genes), R (overall correlation)
Club (136), R = 0.71
Cil (222), R = 0.69
Bas (203), R = 0.73
Gob (333), R = 0.62
NE (175), R = 0.65
AT1 (216), R = 0.66
AT2 (189), R = 0.69
AT2-s (163), R = 0.69
Art (156), R = 0.72
Vein (180), R = 0.68
Cap-a (235), R = 0.65
Cap (121), R = 0.68
Lym (130), R = 0.72
ASM (161), R = 0.67
VSM (164), R = 0.66
MyoF (125), R = 0.71
AdvF (132), R = 0.77
AlvF (118), R = 0.7
Peri (159), R = 0.66
B (82), R = 0.72
Plas (320), R = 0.56
CD8 M/E (87), R = 0.72
CD8 Na (112), R = 0.72
CD4 M/E (79), R = 0.76
CD4 Na (91), R = 0.75
NK (126), R = 0.72
Neu (95), R = 0.64
Bas/Ma 1 (343), R = 0.57
MP (186), R = 0.72
pDC (313), R = 0.6
Class Mono (207), R = 0.66
NC Mono (152), R = 0.72
Human gene expression, ln(CPM+1)
Mouse gene expression, ln(CPM+1)


a
Spdef
Muc5ac
Muc5b
Mm Gob
ln(CPM + 1)
SPDEF
MUC5AC
MUC5B
Hs Gob
ln(UP10K + 1)
25%
50%
75%
100%
b Type 0 - Conserved
PTPRC expression
MYL6 expression
Mouse
Human
Epi
Endo
Stromal
Immune
ln(CPM + 1)
c Type 1 - Gain/Loss
RNASE1 expression
TRIM38 expression
d Type 2 - Expansion/Contraction
CGRP and ADM hormone reception
Ramp1
Ramp2
Ramp3
Calcrl
RAMP1
RAMP2
RAMP3
CALCRL
Art
Vein
Cap-a
Cap
Lym
e
Type 3 - Switch
Mm(SERPINA1)
AT2
Peri
f Hedgehog transdunction components
HHIP
SHH
DHH
IHH
PTCH1
PTCH2
SMO
GLI1
GLI2
GLI3
AT1
AT2
AT2-s
MyoF
AlvF
LipF
Myrf in AT1
g
Alveoli
Nkx2-1
Sftpc
Myrf
Merge
h Type 3 - Cell type gain/loss
Serous markers
LipF markers
Ltf
Lyz1
Lyz2
Bpifb1
Hp
Apoe
Plin2
Fst
LTF
LYZ
BPIFB1
HP
APOE
PLIN2
FST
Ser
Club
Cil
Bas
Gob
NE
ln(UP10K + 1)

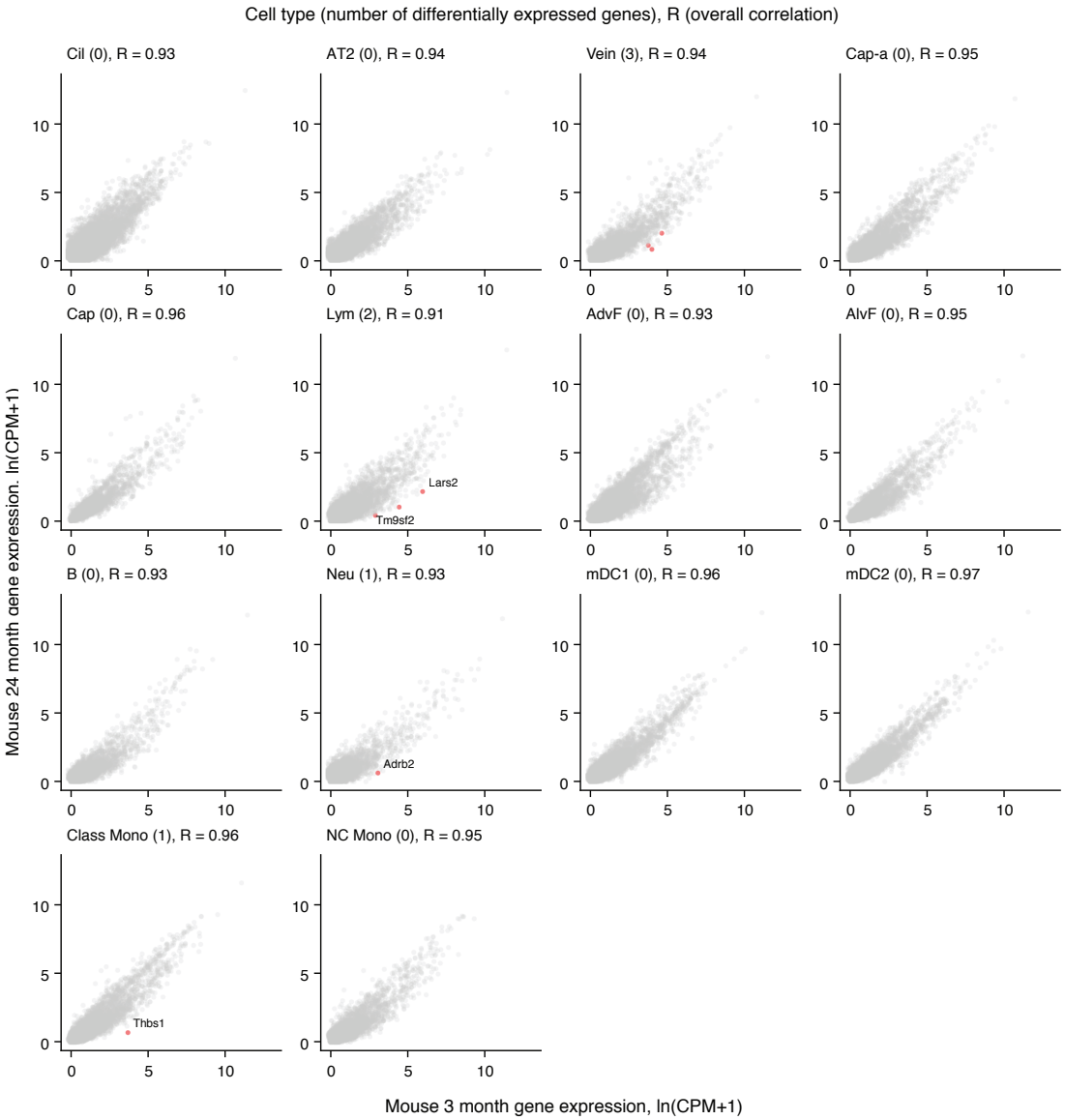

Cell type (number of differentially expressed genes), R (overall correlation)
Cil (0), R = 0.93
AT2 (0), R = 0.94
Vein (3), R = 0.94
Cap-a (0), R = 0.95
Cap (0), R = 0.96
Lym (2), R = 0.91
Lars2
Tm9sf2
AdvF (0), R = 0.93
AlvF (0), R = 0.95
B (0), R = 0.93
Neu (1), R = 0.93
Adrb2
mDC1 (0), R = 0.96
mDC2 (0), R = 0.97
Class Mono (1), R = 0.96
Thbs1
NC Mono (0), R = 0.95
Mouse 24 month gene expression, ln(CPM+1)
Mouse 3 month gene expression, ln(CPM+1)

Table S1. Canonical cell types (45) in the human lung and their abundances, markers, and available expression data.

Cell type	Relative abundance (%)	Number (millions)[a]	Canonical markers[b]	Extant expression profiles Single cell	Primary[c]	Expression accession codes	Abundance reference (method)[d]
Epithelium							
Club Cell	0.5	1,500	CYP2F2, SCGB3A2, CCKAR	Yes	Yes	MTAB-6149, E-MTAB-6653	Boers et al. 1999 (e)
Ciliated Cell	2	6,000	FOXJ1, TUBB1, TP73, CCDC78	Yes	Yes	GSE122960	Raman et al. 2009 (e)
Basal Cell	0.5	1,500	KRT5, KRT14, TP63, DAPL1	Yes	Yes	MTAB-6149, E-MTAB-6653	Boers et al. 1998 (e)
Goblet Cell	0.2	500	MUC5B, MUC5AC, SPDEF	Yes	Yes	EGAS00001001755	Boers et al. 1999 (e)
Mucous Cell	0.03	80	MUC5B				Widdicombe and Wine 2015 (e)
Serous Cell	0.03	80	PRR4, LPO, LTF	Yes	Yes	EGAS00001001755	Basbaum et al. 1990 (e)
Ionocyte	0.03	100	CFTR, FOXI1, ASCL3	Yes	Yes	EGAS00001001755	Montoro et al. 2018 (e)
Neuroendocrine Cell	0.01	40	CALCA, CHGA, ASCL1	Yes	Yes	EGAS00001001755	Boers et al. 1996 (e)
Tuft Cell	0.1	200	DCLK1, ASCL2	Yes		GSE102580	Chang et al. 1986; Montoro et al. 2018 (e)
Alveolar Epithelial Type 1 Cell	13	40,000	AGER, PDPN, CLIC5	Yes	Yes	MTAB-6149, E-MTAB-6653	Crapo et al. 1982 (f)
Alveolar Epithelial Type 2 Cell	7	20,000	SFTPB, SFTPC, SFTPD, MUC1, ETV5	Yes	Yes	GSE122960	Crapo et al. 1982; Fehrenbach et al. 1994 (f)
Total	23	70,000					
Endothelium							
Artery Cell	1	3,000	GJA5, BMX	(bulk)	(cultured)	phs000998.v1.p1	Townsley et al. 2012; The Lung, Chapter 74 (g)
Vein Cell	1	3,000	ACKR1				Townsley et al. 2012; The Lung, Chapter 74 (g)
Capillary Cell	23	70,000	CA4				Crapo et al. 1982 (f)
Bronchial Vessel	0.7	2,000					Deffebach et al. 1987 (g)
Lymphatic Cell	0.7	2,000	PROX1, PDPN	Yes	Yes	MTAB-6149, E-MTAB-6653	Kambouchner et al. 2009; Sozio et al. 2012 (g)
Total	27	80,000		(unannotated)		MTAB-6149, E-MTAB-6653, EGAS00001001755	
Stroma							
Vascular Smooth Muscle	2	5,000	CNN1, ACTA2, TAGLN, RGS5	Yes	Yes	GSE75990	Townsley et al. 2012; The Lung, Chapter 74 (h)
Airway Smooth Muscle	1	4,000	CNN1, ACTA2, TAGLN, DES, LGR6			GSE75990	Elliot et al. 1999; The Lung, Chapter 74 (h)
Fibroblast	7	20,000	COL1A1, PDGFRA	Yes	Yes	EGAS00001001755	Crapo et al. 1982 (f,i)
Myofibroblast	7	20,000	COL1A1, PDGFRA, ELN, ACTA2	Yes	Yes	EGAS00001001755	Crapo et al. 1982 (f,i)
Lipofibroblast	7	20,000	COL1A1, PDGFRA, PLIN2, APOE				Crapo et al. 1982 (f,i)
Pericyte	7	20,000	CSPG4, TRPC6, PDGFRB	(bulk)	(cultured)	GSE75990	Crapo et al. 1982 (f,i)
Mesothelial Cell	0.3	1,000	MSLN, UPK3B, WT1	(bulk)	(cultured)	GSE63966	Michailova et al. 1997 (j)
Total	30	90,000		(unannotated)		MTAB-6149, E-MTAB-6653, EGAS00001001755	
PNS							
Intrinsic Neuron	0.0003	1	SNAP25				Fox et al. 1980; Sparrow et al. 1999 (j)
Glial Cell	0.0002	0.5					Sparrow et al. 1999 (j)
Total	0.0005	1.5					
Immune							
B Cell	0.5	1,500	CD79A, CD24, MS4A1, CD19	Yes	Yes	E-MTAB-6701, E-MTAB-6678	Finkelstein et al. 1995; Banat et al. 2015 (k)
Plasma Cell	0.7	2,000	CD79A, CD27, SLAMF7	Yes	Yes	E-MTAB-6701, E-MTAB-6678	Banat et al. 2015 (k)
CD8+ Mem/Eff T Cell	1	3,000	CD3E, CD8A, GZMK, DUSP2	Yes	Yes	E-MTAB-6701, E-MTAB-6678	Finkelstein et al. 1995; Banat et al. 2015 (k,i)
CD8+ Naive T Cell	1	3,000	CD3E, CD8, GZMH, GZMB	Yes	Yes	MTAB-6149, E-MTAB-6653	Finkelstein et al. 1995; Banat et al. 2015 (k,i)
CD4+ Mem/Eff Cell	0.7	2,000	CD3E, CD8, COTL1, LDHB	Yes	Yes	E-MTAB-6701, E-MTAB-6678	Finkelstein et al. 1995; Banat et al. 2015 (k,i)
CD4+ Naive T Cell	0.7	2,000	CD3E, CD4, CCR7, LEF1	Yes	Yes	MTAB-6149, E-MTAB-6653	Finkelstein et al. 1995; Banat et al. 2015 (k,i)
Natural Killer Cell	1	3,000	KLRD1, NKG7, TYROBP	Yes	Yes	E-MTAB-6701, E-MTAB-6678	Marquardt et al. 2017 (l)
Natural Killer T Cell	0.7	2,000	CD3E, CD8A, FCER1G, TYROBP	Yes	Yes	MTAB-6149, E-MTAB-6653	Marquardt et al. 2017 (k,i)
Neutrophil	0.8	2,500	S100A8, S100A9, IFITM2, FCGR3B	Yes	Yes	EGAS00001001755	Finkelstein et al. 1995; Banat et al. 2015 (k)
Basophil	0.3	1,000	MS4A2, CPA3, TPSAB1	Yes	Yes	MTAB-6149, E-MTAB-6653	Finkelstein et al. 1995 (k,i)
Mast Cell	1	3,000	MS4A2, CPA3, TPSAB1	Yes	Yes	MTAB-6149, E-MTAB-6653	Finkelstein et al. 1995; Banat et al. 2015 (k)
Eosinophil	0.3	1,000	SIGLEC8	(bulk)	(cultured)		Finkelstein et al. 1995 (k,i)
Megakaryocyte	0.3[i]	1,000	NRGN, PPBP, PF4, OST4	(bulk)	Yes		Dejima et al. 2018; Skoczynski et al. 2019 (m)
Macrophage	7	20,000	MARCO, MSR1, MRC1	Yes	Yes	MTAB-6149, E-MTAB-6653	Crapo et al. 1982; Fehrenbach et al. 1994 (f)
Plasmacytoid Dendritic Cell	0.3	800	LILRB4, IRF8, LILRA4	Yes	Yes	GSE94820	Banat et al. 2015 (k,i)
Myeloid Dendritic Cell 1	0.3	1,000	MHCII CLEC9A, LAMP3	Yes	Yes	GSE94820	Banat et al. 2015 (k)
Myeloid Dendritic Cell 2	0.1	200	MHCII, CD1C, PLD4	Yes	Yes	GSE94820	Banat et al. 2015 (k)
Classical Monocyte	2	4,000	CD14, S100A8	Yes	Yes	E-MTAB-6701, E-MTAB-6678	Hance et al. 1985; Hoogsteden et al. 1989 (k,i)
Intermediate Monocyte	2	4,000	CD14, S100A8, CD16	(bulk)	Yes	GSE80095	Hance et al. 1985; Hoogsteden et al. 1989 (k,i)
Nonclassical Monocyte	1	3,000	CD16	Yes	Yes	GSE94820	Hance et al. 1985; Hoogsteden et al. 1989 (k)
Total	20	60,000					
Total (all compartments)	100	300,000					

a, numbers of each type were calculated with their abundances and the total number of lung cells (estimated by comparing volume of lungs to the whole body). **b,** Canonical markers were obtained from referenced expression data or commonly used markers in the literature. **c,** Expression profiles captured immediately following tissue dissociation are considered primary. **d,** Alveoli were assumed to occupy ~90% of the total lung volume for all estimations. **e,** Inferred from mean relative abundance in proximal, medial and distal airway epithelium. **f,** Calculated by stereology **g,** Resin casts showed similar surface area of arteries and veins. **h,** Vascular smooth muscle is estimated to be slightly more abundant than airway smooth muscle. **i,** abundance of a more general cell type was split evenly. **j,** inferred from impression of light or electron microscopy. **k,** inferred from histological abundance in non-perfused healthy tissue. **l,** inferred from abundance among immune cells with FACS. **m,** Calculated using microfluidic capture

Table S2. Human lung cell cluster identities, abundances, and locations

	Annotation	Short name	Patient 1 (SS2)	Patient 1 (10x)	Patient 2 (SS2)	Patient 2 (10x)	Patient 3 (SS2)	Patient 3 (10x)	Total	Lung location	Basis of annotation
	Epithelial										
1	Club	Club	640	176	93	109	165	646	1,829	Bronchioles	Boers et al (1999)
2	Ciliated	Cil	196	86	331	747	24	488	1,872	Airways	smFISH (S4d)
3	Proximal Ciliated	Cil-px	-	-	-	-	-	88	88	PS airway	smFISH (S4d)
4	Basal	Bas	24	24	86	181	130	231	676	Airway	smFISH (S4c)
5	Proximal Basal	Bas-px	-	-	-	-	-	157	157	PS airway	smFISH (2d)
6	Differentiating Basal	Bas-d	-	63	-	-	43	202	308	Airway	IHC (S4b)
7	Proliferating Basal	Bas-p	-	8	-	-	-	39	47	Airway	IHC (S4a)
8	Goblet	Gob	11	-	213	-	9	159	392	Airway	Buisine et al (1999)
9	Mucous	Muc	-	-	-	358	-	133	491	SMG	Buisine et al (1999)
10	Serous	Ser	-	-	-	-	-	24	24	SMG	Fischer et al (2009)
11	Ionocyte	Ion	-	-	-	-	2	22	24	Airway and SMG	Montoro et al (2018)
12	Neuroendocrine	NE	50	-	-	-	5	11	66	Airway	Boers et al (1996)
13	Alveolar Epithelial Type 1	AT1	101	26	101	290	219	656	1,393	Alveoli	smFISH (4c)
14	Alveolar Epithelial Type 2	AT2	137	494	369	535	254	2,785	4,574	Alveoli	smFISH (2f)
15	Signaling Alveolar Epithelial Type 2	AT2-s	175	-	-	-	26	669	870	Alveoli	smFISH (2f)
	Total		1,334	877	1,193	2,220	877	6,310	12,811		
	Endothelial										
16	Artery	Art	61	208	24	968	7	308	1,576	Proximal vessels	smFISH (S4f)
17	Vein	Vein	-	154	42	802	-	199	1,197	Proximal vessels	smFISH (S4e)
18	Capillary Aerocyte	Cap-a	154	300	443	3,237	25	815	4,974	Alveoli	Gillich et al
19	Capillary (general)	Cap	259	572	533	5,385	80	1,434	8,263	Alveoli	Gillich et al
20	Capillary Intermediate 1	Cap-i1	-	-	88	628	-	-	716	Alveoli	Gillich et al
21	Capillary Intermediate 2	Cap-i2	-	-	-	464	-	-	464	Alveoli	Gillich et al
22	Bronchial Vessel 1	Bro1	106	127	-	-	-	328	561	Peribronchial	smFISH (2h,S4h)
23	Bronchial Vessel 2	Bro2	-	103	-	-	-	132	235	Peribronchial	smFISH (2h,S4h)
24	Lymphatics	Lym	18	33	19	254	7	180	511	Proximal vessels	smFISH (S4g)
	Total		598	1,497	1,149	11,738	119	3,396	18,497		
	Stroma										
25	Airway Smooth Muscle	ASM	272	108	21	258	53	327	1,039	Peribronchial	smFISH (S4n)
26	Vascular Smooth Muscle	VSM	114	-	30	86	32	383	645	Perivascular	Daniel et al (2016)
27	Myofibroblast	MyoF	-	34	10	52	42	162	300	Alveolar ducts and alveoli	smFISH (S4o)
28	Fibromyocyte	FibM	-	-	-	-	15	98	113	Peribronchial	smFISH (S4n)
29	Adventitial Fibroblast	AdvF	54	133	12	71	92	353	715	Perivascular and peribronchial	smFISH (4j)
30	Alveolar Fibroblast	AlvF	80	45	85	402	195	849	1,656	Alveoli	smFISH (4j,S4k)
31	Lipofibroblast	LipF	20	35	-	-	-	-	55	Alveoli	smFISH (S4j)
32	Pericyte	Peri	73	74	342	1,135	103	399	2,126	Alveoli	smFISH (S4i)
33	Mesothelial	Meso	-	29	-	-	-	-	29	Pleura	Kanamori-Katayama et al (2011)
	Total		613	458	500	2,004	532	2,571	5,646		
	Immune										
34	B	B	118	162	2	30	9	533	854	LN	Sampling
35	Plasma	Plasma	-	22	-	-	2	165	189	Throughout	Pereira et al (2009) Banat et al (2015)
36	CD8+ Memory/Effector T	CD8 M/E T	59	462	-	414	-	314	1,249	Throughout	Synder et al (2019)
37	CD8+ Naive T	CD8 Na T	365	644	13	513	-	885	2,420	Throughout	Synder et al (2019)
38	CD4+ Memory/Effector T	CD4 M/E T	126	359	-	505	-	2,149	3,139	Throughout	Synder et al (2019)
39	CD4+ Naive T	CD4 Na T	122	338	9	78	-	516	1,063	Throughout	Synder et al (2019)
40	Natural Killer T	NKT	-	-	9	92	43	243	387	Throughout	Uhlén et al (2015)
41	Natural Killer	NK	379	553	345	2,678	508	1,538	6,001	Throughout	Uhlén et al (2015)
42	Proliferating NK/T	NK/T-p	-	-	7	59	10	46	122	-	-
43	Neutrophil	Neu	101	-	4	-	8	-	113	Thoughout	Cohen et al (1983)
44	Mast Cell/Basophil Type 1	Mast/Ba 1	25	117	7	765	14	468	1,396	Alveoli	Fox et al (1981)
45	Mast Cell/Basophil Type 2	Mast/Ba 2	-	-	-	-	0	552	552	Alveoli	Fox et al (1981)
46	Platelet/Megakaryocyte	MK	-	18	-	-	-	22	40	Circulation	Sampling (3b)
47	Macrophage	MP	-	3,167	22	6,262	10	5,305	14,766	Alveoli	smFISH (S4j)
48	Proliferating Macrophage	MP-p	-	31	-	107	-	88	226	-	-
49	Plasmacytoid Dendritic	pDC	7	23	4	53	2	61	150	Throughout	Masten et al (2006)
50	Myeloid Dendritic Type 1	mDC1	10	15	-	65	-	51	141	Throughout	Masten et al (2006)
51	Myeloid Dendritic Type 2	mDC2	-	66	-	27	7	173	273	Throughout	Masten et al (2006)
52	IGSF21+ Dendritic	IGSF21+ DC	-	48	3	112	5	120	288	Perivascular	smFISH (S5e)
53	EREG+ Dendritic	EREG+ DC	-	27	-	115	-	-	142	Throughout	smFISH (S5f)
54	TREM2+ Dendritic	TREM2+ DC	-	59	-	-	-	100	159	Throughout	smFISH (S5g)
55	Classical Monocyte	Cl Mono	106	607	-	292	-	1,178	2,183	Throughout	Heyl et al (2014)
56	OLR1+ Classical Monocyte	OLR1+ Cl Mono	-	-	-	207	-	-	207	-	-
57	Nonclassical Monocyte	NC Mono	24	194	3	269	-	341	831	Throughout	Desch et al (2015)
58	Intermediate Monocyte	Int Mono	-	-	3	188	3	-	194	Throughout	Desch et al (2015)
	Total		1,442	6,912	431	12,831	621	14,848	37,085		
	Total (all compartments)		4,352	9,744	3,273	28,793	2,149	27,125	75,066		

Table S3. Surface markers used for canonical immune cell types in bulk mRNA sequencing.

Immune cell types and populations	Surface markers used for sorting
Neutrophil	SSC^{hi}CD16+CD123-CCR3-b7 integrin-
Eosinophil	SSC^{hi}CD16-CD123-CCR3+b7 integrin+
Basophil	SSC^{lo}CD16-CD123^{hi}CCR3+b7 integrin+
Classical monocyte	FSC^{hi}CD4^{lo}CD14^{hi}CD16^{lo}
Non-classical monocyte	FSC^{hi}CD4^{lo}CD14$^{-/lo}$CD16^{hi}
Myeloid dendritic cell	CD11c^{hi}CD123^{lo}CD1c+HLA-DRhiCD16-
CD16+ dendritic cell	CD11c^{hi}CD123^{lo}CD1c-HLA-DR+CD16+
Plasmacytoid dendritic cell	CD11c-CD123^{hi}CCR3-HLA-DR+b7 integrin-
Naïve B cell	CD19+CD20+CD27-IgM/D+
Unswitched memory B cell	CD19+CD20+CD27+IgM/D+
Switched memory B cell	CD19+CD20+CD27+IgM/D-
Immature NK cell	CD16-CD56^{hi}CD57-
Mature NK cell	CD16+CD56+CD57-
More mature NK cell	CD16+CD56+CD57+
Naïve CD4-T cell	CD4+CD45RA+CD45RO-CCR7+CD62L+
Central memory CD4-T cell	CD4+CD45RA-CD45RO+CCR7+CD62L+
Effector memory CD4-T cell	CD4+CD45RA-CD45RO+CCR7-CD62L-
Naïve CD8-T cell	CD8+CD45RA+CD45RO-CCR7+CD62L+
Central memory CD8-T cell	CD8+CD45RA-CD45RO+CCR7+CD62L+
Effector memory CD8-T cell	CD8+CD45RA-CD45RO+CCR7-CD62L-
Effector memory CD45RA+ CD8-T cell	CD8+CD45RA+CD45RO-CCR7-CD62L-

Table S4. Enriched markers found in each cluster, with transcription factors, receptors/ligands, and disease associated genes annotated.

Club

Gene	avg_logFC	pct_in_cluster	pct_out_cluster	p_val	p_val_adj	TF	Receptor	Ligand	Enzymes	OMIM	GWAS
SCGB3A2	2.63	0.98	0.47	4.38E-51	2.67E-46			TRUE		TRUE	
MGP	3.87	0.68	0.15	2.71E-41	1.65E-36					TRUE	
PIGR	1.66	0.95	0.58	4.85E-34	2.96E-29						
VIM	1.8	0.92	0.42	5.5E-33	3.35E-28			TRUE			
SCGB3A1	2.91	0.91	0.38	9.5E-33	5.79E-28			TRUE			
TIMP1	1.57	0.92	0.37	7.32E-32	4.46E-27			TRUE			
KLK11	1.41	0.87	0.28	1.75E-30	1.07E-25						
FAM129A	1.45	0.76	0.17	7.28E-30	4.44E-25						
FMO2	1.57	0.84	0.25	4.89E-29	2.98E-24				TRUE		
SCGB1A1	3.23	0.5	0.08	3.27E-28	1.99E-23			TRUE			
KLK10	1.94	0.75	0.19	4.29E-28	2.62E-23						
SCNN1B	1.41	0.9	0.38	2.32E-27	1.42E-22		TRUE				
CTSE	1.51	0.87	0.36	3.87E-27	2.36E-22						
CXCL17	1.05	0.98	0.55	5.09E-26	3.1E-21						
CX3CL1	2.61	0.49	0.05	2.19E-25	1.33E-20			TRUE			
STEAP4	1.44	0.89	0.45	2.07E-24	1.26E-19						
CRACR2B	1.47	0.8	0.29	5.04E-24	3.07E-19						
CP	1.46	0.8	0.27	5.86E-24	3.57E-19			TRUE			
KLK13	2.84	0.37	0.02	2.53E-23	1.55E-18						
C16orf89	0.82	0.98	0.53	7.54E-23	4.6E-18						
CLU	1.21	0.97	0.67	1.41E-22	8.62E-18						TRUE
CYP2F1	3.35	0.38	0.02	1.43E-22	8.73E-18						
ERP27	1.41	0.76	0.23	2.01E-22	1.23E-17						
CD82	1.63	0.76	0.26	3.59E-22	2.19E-17		TRUE				
KDR	2.36	0.36	0.02	8.78E-22	5.35E-17		TRUE				
SFTPB	0.86	1	0.68	1.46E-21	8.91E-17					TRUE	
SOX4	1.73	0.88	0.63	3.58E-21	2.18E-16	TRUE					
CYP2B7P	0.85	0.98	0.7	4.01E-21	2.44E-16						
MET	1.27	0.88	0.47	9.63E-21	5.87E-16		TRUE				
CFH	1.86	0.58	0.16	2.67E-19	1.63E-14			TRUE			TRUE
ANPEP	1.93	0.45	0.07	8.76E-19	5.34E-14						
MFSD4A	1.38	0.54	0.11	1.04E-18	6.36E-14						
CYB5A	0.93	0.97	0.9	1.38E-18	8.44E-14						
WFDC2	1.02	0.97	0.87	2.42E-18	1.48E-13						
CYP2A6	3.28	0.36	0.04	2.86E-18	1.74E-13					TRUE	TRUE
NOTCH3	1.73	0.53	0.12	4.29E-18	2.61E-13		TRUE				
RNASE1	0.85	0.99	0.68	1.23E-17	7.52E-13						
CTSC	1.49	0.77	0.39	3.04E-17	1.85E-12						
SLPI	0.39	0.98	0.69	3.41E-17	2.08E-12			TRUE			
TMEM45A	3.48	0.28	0.03	7.6E-17	4.64E-12						
TSPAN8	2.33	0.38	0.06	2.41E-16	1.47E-11						TRUE
NT5E	2.13	0.44	0.08	2.7E-16	1.65E-11		TRUE				
HNMT	1.43	0.73	0.29	5.02E-16	3.06E-11				TRUE	TRUE	
PAG1	1.4	0.45	0.08	6.54E-16	3.99E-11						
AARD	2.41	0.51	0.15	7.64E-16	4.66E-11						
HSD17B13	2.33	0.44	0.09	8.27E-16	5.04E-11						
CEACAM6	0.66	0.91	0.53	1.03E-15	6.27E-11						
MDK	1.11	0.75	0.34	1.52E-15	9.25E-11			TRUE			
SEL1L3	1.17	0.83	0.42	2.57E-15	1.56E-10						
ASS1	1.28	0.71	0.28	3.11E-15	1.9E-10						
RHOV	2.02	0.27	0.02	3.17E-15	1.94E-10						
GGT5	2.29	0.48	0.14	5.04E-15	3.07E-10				TRUE		
S100A13	0.87	0.85	0.59	1.47E-14	8.99E-10						
RPS6	0.69	0.99	0.95	1.92E-14	1.17E-09						
B3GNT3	2.8	0.21	0.01	4.39E-14	2.68E-09						
ARHGDIB	1.55	0.72	0.37	5.1E-14	3.11E-09						
DHRS3	0.93	0.95	0.74	9.2E-14	5.61E-09				TRUE		
MMP7	2.05	0.43	0.11	9.28E-14	5.66E-09			TRUE			
ST6GALNAC	1.2	0.63	0.23	9.75E-14	5.95E-09						
MRPS25	1.05	0.9	0.56	1.21E-13	7.36E-09						
SLC39A6	1.26	0.72	0.33	1.37E-13	8.37E-09		TRUE				TRUE
PAEP	2.79	0.2	0.01	2.88E-13	1.76E-08						
TCIM	0.85	0.74	0.33	3.02E-13	1.84E-08						
RPL13AP5	0.63	0.97	0.91	5.64E-13	3.44E-08						
RPL13P12	0.65	0.95	0.81	7.35E-13	4.48E-08						
TENT5C	1.33	0.72	0.36	8.5E-13	5.18E-08						
RPLP0	0.71	0.99	0.91	8.91E-13	5.43E-08						
KIAA1324	0.73	0.85	0.53	1.01E-12	6.13E-08						TRUE
SCUBE2	1.64	0.44	0.11	1.19E-12	7.23E-08						
RPL12	0.7	0.99	0.92	1.38E-12	8.44E-08						
CLIC6	0.83	0.79	0.4	1.66E-12	1.01E-07						
INPP4B	2.14	0.25	0.02	1.95E-12	1.19E-07						
TSPAN11	1.83	0.4	0.09	2.29E-12	1.4E-07						TRUE

...

75 enriched genes are shown for first cluster (Club), table continues to include remaining enriched genes (p-val > 0.05) with remaining cell types in separate sheets. Abbreviations: avg_logFC, the natural log of the average fold change between the cell type and other cell types in its tissue compartment; pct_in_cluster, percentage of cells within the cluster that express the gene; pct_out_cluster, percentage of cells outside cluster that express the gene; p_val_adj, p-value with Bonferroni correction applied; TF, transcription factor; OMIM, Online Mendelian Inheritance in Man; GWAS, genome wide association study.

Table S5. P-value and scores of each CellPhoneDB Receptor-Ligand interaction from each cluster.

interacting_pair	partner_a	partner_b	source	secreted	is_integrin	Adventitial Fibroblast_Adventitial Fibroblast
CXCL12_ACKR3	simple:P48061	simple:P25106	guidetopharmacology.org	TRUE	FALSE	1.8 | 0.0
CXCL11_ACKR3	simple:O14625	simple:P25106	guidetopharmacology.org	TRUE	FALSE	0.0 | 1.0
CXCL12_CXCR3	simple:P48061	simple:P49682	guidetopharmacology.org	TRUE	FALSE	0.0 | 1.0
CXCL11_CXCR3	simple:O14625	simple:P49682	curated	TRUE	FALSE	0.0 | 1.0
CCL19_CXCR3	simple:Q99731	simple:P49682	guidetopharmacology.org	TRUE	FALSE	0.0 | 1.0
CCL20_CXCR3	simple:P78556	simple:P49682	guidetopharmacology.org	TRUE	FALSE	0.0 | 1.0
CXCL10_CXCR3	simple:P02778	simple:P49682	curated	TRUE	FALSE	0.0 | 1.0
CXCL9_CXCR3	simple:Q07325	simple:P49682	curated	TRUE	FALSE	0.0 | 1.0
CXCL12_CXCR4	simple:P48061	simple:P61073	curated	TRUE	FALSE	0.0 | 1.0
CXCL12_DPP4	simple:P48061	simple:P27487	curated	TRUE	FALSE	1.3 | 1.0
CXCL11_DPP4	simple:O14625	simple:P27487	curated	TRUE	FALSE	0.0 | 1.0
CXCL10_DPP4	simple:P02778	simple:P27487	curated	TRUE	FALSE	0.0 | 1.0
CXCL9_DPP4	simple:Q07325	simple:P27487	curated	TRUE	FALSE	0.2 | 1.0
CCL11_DPP4	simple:P51671	simple:P27487	curated	TRUE	FALSE	0.1 | 1.0
GCG_DPP4	simple:P01275	simple:P27487	MINT	TRUE	FALSE	0.1 | 1.0
NPPB_DPP4	simple:P16860	simple:P27487	MINT	TRUE	FALSE	0.0 | 1.0
VIP_DPP4	simple:P01282	simple:P27487	MINT	TRUE	FALSE	0.0 | 1.0
GIP_DPP4	simple:P09681	simple:P27487	MINT	TRUE	FALSE	0.0 | 1.0
CCL22_DPP4	simple:O00626	simple:P27487	curated	TRUE	FALSE	0.0 | 1.0
ADCYAP1_DPP4	simple:P18509	simple:P27487	MINT	TRUE	FALSE	0.0 | 1.0
CXCL2_DPP4	simple:P19875	simple:P27487	curated	TRUE	FALSE	0.7 | 1.0
CCL19_CCRL2	simple:Q99731	simple:O00421	guidetopharmacology.org	TRUE	FALSE	0.0 | 1.0
CCL19_ACKR4	simple:Q99731	simple:Q9NPB9	curated	TRUE	FALSE	0.0 | 1.0
CCL21_ACKR4	simple:O00585	simple:Q9NPB9	curated	TRUE	FALSE	0.3 | 1.0
CCL25_ACKR4	simple:O15444	simple:Q9NPB9	curated	TRUE	FALSE	0.0 | 1.0
CXCL13_ACKR4	simple:O43927	simple:Q9NPB9	curated	TRUE	FALSE	0.0 | 1.0
CCL19_CCR7	simple:Q99731	simple:P32248	curated	TRUE	FALSE	0.0 | 1.0
CCL21_CCR7	simple:O00585	simple:P32248	curated	TRUE	FALSE	0.0 | 1.0
CCL20_CCR6	simple:P78556	simple:P51684	curated	TRUE	FALSE	0.0 | 1.0
CXCR3_PF4	simple:P49682	simple:P02776	curated	TRUE	FALSE	0.0 | 1.0
CCL11_CCR3	simple:P51671	simple:P51677	curated	TRUE	FALSE	0.0 | 1.0
CCL5_CCR3	simple:P13501	simple:P51677	curated	TRUE	FALSE	0.0 | 1.0
CCL13_CCR3	simple:Q99616	simple:P51677	curated	TRUE	FALSE	0.0 | 1.0
CCL7_CCR3	simple:P80098	simple:P51677	curated	TRUE	FALSE	0.0 | 1.0
CCL8_CCR3	simple:P80075	simple:P51677	curated	TRUE	FALSE	0.0 | 1.0
CCL26_CCR3	simple:Q9Y258	simple:P51677	curated	TRUE	FALSE	0.0 | 1.0
CCL24_CCR3	simple:O00175	simple:P51677	curated	TRUE	FALSE	0.0 | 1.0
CCL28_CCR3	simple:Q9NRJ3	simple:P51677	guidetopharmacology.org	TRUE	FALSE	0.0 | 1.0
CCL11_CCR2	simple:P51671	simple:P41597	guidetopharmacology.org	TRUE	FALSE	0.0 | 1.0
CCL13_CCR2	simple:Q99616	simple:P41597	curated	TRUE	FALSE	0.0 | 1.0
CCL7_CCR2	simple:P80098	simple:P41597	curated	TRUE	FALSE	0.0 | 1.0
CCL8_CCR2	simple:P80075	simple:P41597	curated	TRUE	FALSE	0.0 | 1.0
CCL26_CCR2	simple:Q9Y258	simple:P41597	curated	TRUE	FALSE	0.0 | 1.0
CCL24_CCR2	simple:O00175	simple:P41597	guidetopharmacology.org	TRUE	FALSE	0.0 | 1.0
CCL16_CCR2	simple:O15467	simple:P41597	curated	TRUE	FALSE	0.0 | 1.0
CCL2_CCR2	simple:P13500	simple:P41597	curated	TRUE	FALSE	0.0 | 1.0
GCG_GLP2R	simple:P01275	simple:O95838	guidetopharmacology.org	TRUE	FALSE	0.1 | 1.0
GCG_IDE	simple:P01275	simple:P14735	InnateDB-All	TRUE	FALSE	0.3 | 1.0
IGF2_IDE	simple:P01344	simple:P14735	InnateDB-All	TRUE	FALSE	0.3 | 1.0
INS_IDE	simple:P01308	simple:P14735	InnateDB-All	TRUE	FALSE	0.0 | 1.0
NPPB_NPR1	simple:P16860	simple:P16066	guidetopharmacology.org	TRUE	FALSE	0.0 | 1.0
NPPA_NPR1	simple:P01160	simple:P16066	guidetopharmacology.org	TRUE	FALSE	0.0 | 1.0
NPPB_NPR3	simple:P16860	simple:P17342	InnateDB-All	TRUE	FALSE	0.0 | 1.0
NPPC_NPR3	simple:P23582	simple:P17342	InnateDB-All	TRUE	FALSE	0.0 | 1.0
OSTN_NPR3	simple:P61366	simple:P17342	guidetopharmacology.org	TRUE	FALSE	0.0 | 1.0
NPPB_NPR2	simple:P16860	simple:P20594	InnateDB-All	TRUE	FALSE	0.0 | 1.0
NPPC_NPR2	simple:P23582	simple:P20594	guidetopharmacology.org	TRUE	FALSE	0.0 | 1.0
VIP_VIPR1	simple:P01282	simple:P32241	guidetopharmacology.org	TRUE	FALSE	0.0 | 1.0
GHRH_VIPR1	simple:P01286	simple:P32241	guidetopharmacology.org	TRUE	FALSE	0.0 | 1.0
CCL22_CCR4	simple:O00626	simple:P51679	curated	TRUE	FALSE	0.0 | 1.0
CCL5_CCR4	simple:P13501	simple:P51679	curated	TRUE	FALSE	0.0 | 1.0
CCL17_CCR4	simple:Q92583	simple:P51679	curated	TRUE	FALSE	0.0 | 1.0
ADCYAP1_SCTR	simple:P18509	simple:P47872	InnateDB-All	TRUE	FALSE	0.0 | 1.0
ADCYAP1_ADCYAP1R1	simple:P18509	simple:P41586	InnateDB-All	TRUE	FALSE	0.0 | 1.0
CXCL2_CXCR2	simple:P19875	simple:P25025	curated	TRUE	FALSE	0.0 | 1.0
CXCL8_CXCR2	simple:P10145	simple:P25025	curated	TRUE	FALSE	0.0 | 1.0
CXCL5_CXCR2	simple:P42830	simple:P25025	curated	TRUE	FALSE	0.0 | 1.0
CXCL1_CXCR2	simple:P09341	simple:P25025	curated	TRUE	FALSE	0.0 | 1.0
CXCL6_CXCR2	simple:P80162	simple:P25025	curated	TRUE	FALSE	0.0 | 1.0
PPBP_CXCR2	simple:P02775	simple:P25025	curated	TRUE	FALSE	0.0 | 1.0
CXCL3_CXCR2	simple:P19876	simple:P25025	curated	TRUE	FALSE	0.0 | 1.0
CCL25_CCR9	simple:O15444	simple:P51686	curated	TRUE	FALSE	0.0 | 1.0
CXCL13_CXCR5	simple:O43927	simple:P32302	curated	TRUE	FALSE	0.0 | 1.0
TNFRSF1A_FASLG	simple:P19438	simple:P48023	InnateDB-All	TRUE	FALSE	0.0 | 1.0

⋮

⋮

75 receptor-ligand interactions and their expression score and p-value (score | p-value) in the first pairwise cell-cell interaction (AdvF and AdvF). Table continues to include all 1085 receptor-ligand pairs and all pairwise cell type combinations. Abbreviations: partner_a, cellphonedb ID for the first interaction partner protein; partner_b, cellphonedb ID for the second interaction partner protein; source, reference from cellphonedb; secreted, whether the ligand is secreted or membrane bound; is_integrin, whether the interaction includes an integrin complex.

Table S6. Genes specific to mouse and human in each cluster and lung wide.

human_Club vs mouse_Club - R = 0.74

gene	mouse_avg_exp	human_avg_exp	avg_logFC	pct_mouse	pct_human	p_val	p_val_adj	log_p_val_adj	enriched
RNASE1	0.02	8.86	-9.26	0.02	0.99	2.44E-48	3.67E-44	43.43	TRUE
HP	9.29	0.05	8.41	1	0.01	8.68E-44	1.31E-39	38.88	TRUE
CYP2F1	10.39	2.16	4.95	1	0.38	3.67E-43	5.52E-39	38.26	TRUE
GSTM5	7.18	0.03	7.66	0.96	0.01	2.14E-40	3.22E-36	35.49	
SLPI	0.2	9.1	-6.9	0.04	0.98	4.58E-40	6.89E-36	35.16	TRUE
AGR3	0	6.77	-7.37	0	0.96	1.15E-39	1.72E-35	34.76	TRUE
CCKAR	6.24	0	7.15	0.93	0	1.04E-38	1.57E-34	33.8	TRUE
FOLR1	0	7.21	-7.93	0	0.95	1.51E-38	2.27E-34	33.64	TRUE
RPL5	4.06	7.87	-3.09	0.88	0.98	2.16E-38	3.25E-34	33.49	
LARS2	6.9	0.4	4.59	1	0.09	1.36E-37	2.04E-33	32.69	TRUE
RPL26	1.35	6.09	-3.58	0.4	0.97	2.21E-37	3.32E-33	32.48	
CD59	0.04	6.13	-6.7	0.02	0.95	4.09E-37	6.15E-33	32.21	
GABRP	6.03	0.03	6.75	0.93	0.01	2.85E-36	4.29E-32	31.37	TRUE
HOPX	2.12	6.93	-4.06	0.6	0.95	4.14E-35	6.23E-31	30.21	TRUE
HLA-DRA	0	6.13	-7.17	0	0.9	7.21E-34	1.08E-29	28.96	
RPL30	1.02	6.22	-3.33	0.26	0.97	1.07E-32	1.61E-28	27.79	
SUSD2	0	5.34	-6.6	0	0.88	2.74E-32	4.11E-28	27.39	TRUE
TMC5	0.14	5.17	-4.1	0.04	0.93	5.46E-32	8.2E-28	27.09	TRUE
GPX2	4.78	0	6.31	0.82	0	1.46E-31	2.2E-27	26.66	TRUE
KLK11	0	5.76	-6.76	0	0.87	1.52E-31	2.28E-27	26.64	TRUE
CTSE	0	5.72	-7.09	0	0.87	1.52E-31	2.28E-27	26.64	TRUE
CES1	7.09	0.7	4.62	0.93	0.16	3.53E-31	5.3E-27	26.28	TRUE
VIM	0.27	6.26	-4	0.07	0.93	3.79E-31	5.7E-27	26.24	
SEC14L3	6.52	0.12	6.82	0.88	0.03	5.39E-31	8.11E-27	26.09	TRUE
TIMP1	0.19	6.1	-4.52	0.04	0.92	6.09E-31	9.16E-27	26.04	
S100A6	2.36	7.93	-3.01	0.44	1	4.32E-30	6.5E-26	25.19	TRUE
SERPING1	0.3	5.46	-3.49	0.1	0.92	5.69E-30	8.55E-26	25.07	
CEBPD	1.17	6.51	-3.1	0.28	0.98	8.02E-30	1.21E-25	24.92	
CAV1	0.24	5.95	-4.48	0.05	0.91	5.26E-29	7.91E-25	24.1	
RBP4	4.51	0	6.13	0.77	0	1.34E-28	2.02E-24	23.7	TRUE
ATP11A	6.75	2	3.85	0.93	0.52	3.6E-28	5.42E-24	23.27	TRUE
KCNK2	3.79	0.03	5.27	0.81	0.01	8.5E-28	1.28E-23	22.89	TRUE
NUPR1	6.79	1.55	3.17	0.98	0.35	1.02E-27	1.54E-23	22.81	TRUE
CD46	0	3.69	-5.09	0	0.81	1.41E-27	2.13E-23	22.67	
FAM20A	0	3.57	-5.22	0	0.81	1.41E-27	2.13E-23	22.67	TRUE
FABP5	0.09	5.17	-5.1	0.02	0.85	4.05E-27	6.08E-23	22.22	TRUE
EID1	0.42	5.68	-3.91	0.1	0.9	9.97E-27	1.5E-22	21.82	
ENO1	1.82	6.2	-3.65	0.53	0.94	4.98E-26	7.49E-22	21.13	
CD36	4.77	0	7.06	0.72	0	6.69E-26	1.01E-21	21	
MIF	3.36	0	4.84	0.7	0	4.75E-25	7.14E-21	20.15	
GSTA3	4.06	0	5.85	0.7	0	4.75E-25	7.14E-21	20.15	TRUE
ERP27	0	4.11	-5.88	0	0.76	9.39E-25	1.41E-20	19.85	TRUE
SEL1L3	0	3.42	-5.08	0	0.75	3.13E-24	4.71E-20	19.33	TRUE
AQP3	0	4.1	-6.3	0	0.75	3.13E-24	4.71E-20	19.33	TRUE
EIF4A1	3.57	0	5.27	0.68	0	3.21E-24	4.82E-20	19.32	
RPS6KA2	0	3.26	-5.03	0	0.74	1.02E-23	1.53E-19	18.82	
Mm(SERPIN	0	4.77	-7.13	0	0.74	1.02E-23	1.53E-19	18.82	TRUE
SOX4	1.1	5.83	-3.99	0.3	0.88	4.5E-23	6.76E-19	18.17	TRUE
CHCHD10	4.52	0.53	3.31	0.82	0.24	6.54E-23	9.83E-19	18.01	
CRTAC1	0	3.93	-5.82	0	0.72	9.93E-23	1.49E-18	17.83	TRUE
CD200	3.56	0	5.8	0.65	0	1.27E-22	1.92E-18	17.72	
KLK10	0.2	5.01	-5.09	0.05	0.81	1.95E-22	2.93E-18	17.53	TRUE
PARP14	0.39	4.21	-3.3	0.1	0.87	3.49E-22	5.25E-18	17.28	
UQCRHL	3.84	0.26	3.76	0.75	0.17	4.19E-22	6.3E-18	17.2	
TMEM176E	6.6	1.65	3.36	0.91	0.37	6.41E-22	9.64E-18	17.02	
ASS1	0	3.68	-5.59	0	0.7	8.82E-22	1.33E-17	16.88	TRUE
HIST1H4C	0	3.16	-5.17	0	0.69	2.54E-21	3.82E-17	16.42	
EPHX1	4.55	0.59	4.51	0.74	0.25	3.03E-21	4.56E-17	16.34	TRUE
HLA-E	1.66	5.63	-3.05	0.51	0.94	3.09E-21	4.64E-17	16.33	
EIF3C	3.2	0.02	5.27	0.68	0.04	3.86E-21	5.81E-17	16.24	
Mm(DYNLT	1.02	4.62	-3.53	0.39	0.83	7.81E-21	1.17E-16	15.93	
C12orf49	0.08	3.78	-4.46	0.02	0.74	8.92E-21	1.34E-16	15.87	TRUE
SPARC	3.58	0.05	5.32	0.67	0.01	1.14E-20	1.71E-16	15.77	
CLDN10	4.44	0.21	3.21	0.74	0.05	1.98E-20	2.98E-16	15.53	TRUE
B3GNT7	0.02	3.73	-5.6	0.02	0.71	2.15E-20	3.24E-16	15.49	TRUE
SERPINB1	0.11	3.8	-3.48	0.02	0.73	2.52E-20	3.78E-16	15.42	
TSTD1	0.82	5.23	-3.04	0.18	0.88	3.57E-20	5.37E-16	15.27	TRUE
DUOX1	0	3.06	-5.24	0	0.66	5.39E-20	8.11E-16	15.09	TRUE
AGER	6	0.8	3.06	0.84	0.17	9.23E-20	1.39E-15	14.86	TRUE
SUMO2	1.7	4.99	-3.1	0.53	0.87	1.62E-19	2.44E-15	14.61	
DMKN	0.03	3.12	-5.04	0.02	0.7	2.4E-19	3.6E-15	14.44	TRUE
Hs(AKR1C1	0	3.33	-5.38	0	0.64	3.77E-19	5.66E-15	14.25	TRUE
GDF15	0	4.31	-7.22	0	0.64	3.77E-19	5.66E-15	14.25	TRUE

...

75 genes that vary in expression between mouse and human for the first cluster (Club), table continues to include remaining differentially expressed genes (p-val > 0.05) with remaining comparisons in separate sheets. Abbreviations: avg_logFC, the natural log of the average fold change between the mouse and human cell type indicated; pct_mouse, percentage of mouse cells within the cluster that express the gene; pct_human, percentage of human cells within the cluster that express the gene; p_val_adj, p-value with Bonferroni correction applied; enriched, gene is enriched in cluster in mouse or human.

Table S7. Evolutionary changes in cellular patterns of lung gene expression between mouse and human.

Evolutionary scenarios	Number of genes	Percent of genes	Percent of transcriptome
Type 0 (Conserved)	602 ± 27	4 ± 0.2%	6 ± 0.33%
Type 1	3,020 ± 69	20 ± 0.5%	29 ± 0.7%
Gain in mouse	1,600 ± 33	11 ± 0.2%	15 ± 0.3%
Gain in human	1,420 ± 36	10 ± 0.2%	15 ± 0.3%
Type 2	5,815 ±249	39 ± 1.7%	55 ± 2.3%
Expansion in mouse	1,095 ± 36	7 ± 0.2%	10 ± 0.3%
Expansion in human	1,110 ± 47	7 ± 0.3%	11 ± 0.4%
Expansion in both	3,610 ±166	24 ± 1.1%	34 ± 1.6%
Type 3 (Cell type switch)	1,062 ± 22	7 ± 0.1%	10 ± 0.3%
Not expressed	4,437 ±209	30 ± 1.4%	

Numbers of genes and percentages are mean (±SEM) from using a median expression cutoff minus 0 to 2 standard deviations (0.25 increments) and a median percent cutoff plus 0 to 2 standard deviations (0.25 increments).

Table S8. Evolutionary and functional classes of each gene.

Gene	Evo type	Gene class	Conserved clusters	Human specific clusters	Mouse specific clusters
KIAA1841	2	Other	1	1	0
C17orf49	2	Other	3	20	0
C2orf68	2	Other	5	19	0
C4orf19	2	Other	1	5	0
C11orf58	2	Other	27	1	0
C9orf16	2	Other	14	10	2
C9orf24	0	Other	1	0	0
C11orf1	3	Other	0	1	2
C3orf80	1	Other	0	0	1
C8orf33	2	Other	4	8	5
C11orf49	2	Other	4	4	6
C9orf85	1	Other	0	0	7
KIAA1143	3	Other	0	8	5
C16orf74	3	Other	0	2	2
C12orf73	1	Other	0	0	1
C19orf24	3	Other	0	11	1
C6orf120	1	Other	0	0	4
C19orf12	3	Other	0	4	2
C2orf70	0	Other	1	0	0
C5orf49	0	Other	1	0	0
C9orf50	NA	Other	0	0	0
C17orf98	NA	Other	0	0	0
C2orf81	0	Other	1	0	0
C9orf116	0	Other	1	0	0
C19orf81	NA	Other	0	0	0
C6orf118	1	Other	0	1	0
C4orf51	NA	Other	0	0	0
C11orf97	0	Other	1	0	0
C1orf158	1	Other	0	1	0
C1orf194	0	Ligand	1	0	0
C1orf100	NA	Other	0	0	0
C4orf36	NA	Other	0	0	0
C17orf97	1	Other	0	0	1
C15orf39	2	Other	1	3	0
C12orf50	NA	Other	0	0	0
C16orf95	NA	Other	0	0	0
C2orf88	NA	Other	0	0	0
C10orf82	NA	Other	0	0	0
C20orf141	NA	Other	0	0	0
C17orf50	NA	Other	0	0	0
C6orf203	1	Other	0	0	1
C20orf85	1	Other	0	1	0
C9orf131	NA	Other	0	0	0
C10orf53	1	Other	0	0	1
C1orf21	2	Other	2	8	2
C1orf146	NA	Other	0	0	0
C9orf135	2	Other	1	1	0
C11orf94	NA	Other	0	0	0
C4orf47	0	Other	1	0	0
C16orf46	0	Other	1	0	0
C19orf44	NA	Other	0	0	0
LOC1005058	NA	Other	0	0	0
C12orf71	NA	Other	0	0	0
C16orf90	1	Other	0	0	1
C20orf27	3	Other	0	2	7
C2orf69	1	Other	0	1	0
C19orf67	NA	Other	0	0	0
C22orf23	0	Other	1	0	0
C2orf74	1	Other	0	2	0
C1orf189	0	Other	1	0	0
C3orf62	1	Other	0	3	0
C12orf42	NA	Other	0	0	0
C14orf119	2	Other	10	12	2
C17orf64	NA	Other	0	0	0
C11orf98	2	Other	6	6	8
C17orf58	1	Other	0	0	5
C16orf72	3	Other	0	1	12
CXorf38	2	Other	1	6	3
C11orf53	NA	Other	0	0	0
C11orf24	2	Other	2	3	0
C11orf86	NA	Other	0	0	0
KIAA1211L	2	Other	1	3	1
C9orf64	1	Other	0	0	16
C12orf43	3	Other	0	2	5

...

75 genes are shown, ordered by HomologID. Table continues to include remaining genes conserved between mouse and human. Abbreviations: Evo type, evolutionary scenario gene falls into, with NA for genes not expressed; Gene class, type of gene (receptor, ligand, enzyme, transcription factor, etc); conserved clusters, number of homologous clusters gene where gene is expressed in both mouse and human; Human specific clusters, number of homologous clusters where the gene is only expressed in human; Mouse specific clusters, above for mouse.