Single cell transcriptomic profiling identifies molecular phenotypes of newborn human lung cells.

Soumyaroop Bhattacharya[1], Jacquelyn L. Myers[3], Cameron Baker[3], Minzhe Guo[5], Soula Danopoulos[6], Jason R. Myers[3], Gautam Bandyopadhyay[1], Stephen Romas[1], Heidie L. Huyck[1], Ravi S. Misra[1], Jennifer Dutra[4], Jeanne Holden-Wiltse[2,4], Andrew McDavid[2], John M. Ashton[3], Denise Al Alam[6], S. Steven Potter[5], Jeffrey A. Whitsett[5], Yan Xu[5], Gloria S. Pryhuber[1], Thomas J. Mariani[1]*

[1]Department of Pediatrics, [2]Department of Biostatistics and Computational Biology, and [3]Genomic Research Center, [4]Clinical & Translational Science Institute, University of Rochester, [5]Department of Pediatrics, Cincinnati Children's Hospital Medical Center and [6] Lundquist Institute for Biomedical Innovation, Harbor-UCLA Medical Center, University of California Los Angeles, Los Angeles, CA

* Corresponding Author

Address for Correspondence:

Thomas J Mariani, PhD
Division of Neonatology and
Pediatric Molecular and Personalized Medicine Program
University of Rochester Medical Center
601 Elmwood Ave, Box 850
Rochester, NY 14642, USA.
Phone: 585-276-4616;
Fax: 585-276-2643;
E-mail: Tom_Mariani@urmc.rochester.edu.

Running Title: Newborn human lung cellular diversity

Key Words: Single cell RNAseq, matrix fibroblast, lung development, newborn lung

No. of Tables: 0
No. of Figures: 6
No of Supplemental Tables: 11
No. of Supplemental Figures: 9
Word count (manuscript):6100
Word count (abstract): 340

This work was supported by the Human Developing Lung Molecular Atlas Program (LungMAP) funded by National Heart, Lung, and Blood Institute of National Institutes of Health (U01HL122700 to G.S.P. and U01HL122642 to J.A.W.).

Abstract

Rationale: While animal model studies have extensively defined mechanisms controlling cell diversity in the developing mammalian lung, the limited data available from late stage human lung development represents a significant knowledge gap. The NHLBI Molecular Atlas of Lung Development Program (LungMAP) seeks to fill this gap by creating a structural, cellular and molecular atlas of the human and mouse lung.

Methods: Single cell RNA sequencing generated transcriptional profiles of 5500 cells obtained from two one-day old human lungs (born at gestational ages of 39 and 31 weeks) from the LungMAP Human Tissue Core Biorepository at the University of Rochester. Frozen single cell isolates were captured, and library preparation was completed on the Chromium 10X system. Data was analyzed in Seurat, and cellular annotation was performed using the ToppGene functional analysis tool. Single cell sequence data from 32000 postnatal day 1, 3, 7 and 10 mouse lung (n = 2 at each time point) cells generated by the LungMAP Research Center at Cincinnati Children's Hospital and Medical Center, using Dropseq platform, was integrated with the human data. In situ hybridization was used to confirm the spatial location of cellular phenotypes.

Results: Transcriptional interrogation of donor newborn human lung cells identified distinct clusters representing multiple populations of epithelial, endothelial, fibroblasts, pericytes, smooth muscle, and immune cells and signature genes for each of these populations were identified. Computational integration of newborn human and postnatal mouse lung development cellular transcriptomes facilitated the identification of distinct cellular lineages among all the major cell types. Integration of the human and mouse cellular transcriptomes also demonstrated cell type-specific differences in developmental states of the newborn human lung cells. In particular, matrix fibroblasts could be separated into those representative of younger cells (n=393), or older cells (n=158).

This is the first comprehensive molecular map of the cellular landscape of neonatal human lung, including biomarkers for cells at distinct states of development. Our results indicate that integrated single cell RNA profiling of human and mouse lung will help identify common and species-specific mechanisms of lung development and respiratory disease.

INTRODUCTION

The lung is a complex organ comprised of over 40 different cell types [1, 2]. Despite recent advances in our understanding of lung development, the complex cellular function and intercellular interactions in the developing human lung are yet to be clearly understood. Development and maintenance of lung structure requires cross talk among multiple cell types to coordinate lineage specification, cell proliferation, differentiation, migration, morphogenesis, and injury repair. The diverse array of pulmonary cells can be categorized into four major cell populations, namely, epithelial cells, endothelial cells, mesenchymal cells, and lung resident and transient immune cells, with each group being relatively well distinguished by specific cell-surface proteins. Even though key signaling molecules, genes, and pathways driving lung development have been identified [3-9], significant knowledge gaps still exist in our understanding of this process especially in humans and interdisciplinary efforts will be required further our understanding the early life origin of lung disease [10].

Transition of the lung from fetal to neonatal states is highly complex, and has been characterized in murine models [11-17]. Cell lineages, and their relationships, during lung development and in diseased states have been extensively studied in rodent models, thanks in large part to use of transgenic technology. It has been more difficult to confirm independent cell types and their lineages in the relatively rare and non-experimental nature of human lung tissue analysis. While early stages of human fetal lung development have been characterized at the molecular level [18, 19], data describing the newborn human lung is lacking [20]. Although, molecular profiling has been applied to pre-viable human lung [19, 21], further understanding of later human lung development has been limited by lack of access to tissue of sufficient quality for molecular analysis. These limitations have been recently overcome by the NHLBI Molecular Atlas of Lung Development Program (LungMAP). The establishment of the LungMAP program, and its success in obtaining human tissues for structural, cellular and molecular analysis has,

and will continue to, facilitate this progress [22-25]. Additionally, most high-throughput molecular studies of lung development have used whole lung tissue [11], limiting insights into the activities of and interactions among different cell types. However, given recent advancements in high-throughput molecular profiling technologies, rapid progress is being made [26-28]. Single cell RNA-seq enables transcriptomic mapping of individual cells to measure and understand cellular heterogeneity and responses in complex biological systems [28]. It not only offers an opportunity to explore the whole transcriptome of individual cells, and thereby organizing cells into cell states in an unsupervised manner can be performed in multitude of organisms, allowing unbiased comparisons across species [29]. Recently, the Human Cell Atlas consortium has prioritized the characterization of cellular composition of lung tissue using single cell sequencing [30]. Single-cell RNA-Seq has also been applied for the identification of known or novel cell populations and to assess cellular heterogeneity and gene expression in changes in lung cell populations during health and diseases [22, 31-38]. Here, we report transcriptional analysis of donor newborn human lung cells, including all major cell types, and describe the molecular profile for matrix fibroblast subtypes that may represent cells at different states of development.

LungMAP was developed to generate detailed structural and molecular data regarding normal perinatal and postnatal lung development in the mouse and human [23, 39]. We have recently reported that high-throughput analysis (transcriptomics, proteomics, etc.) of sorted dissociated cells from human neonatal and pediatric lungs reveals retention of in vivo phenotypes [21, 23, 40, 41]. Here, we build on rapid advancement in single-cell transcriptomics that enables the identification of cell-type specific transcriptomes of neonatal and ageing murine lungs, serving as a comparative basis for understanding the transcriptomic landscape of the newborn human lung [26, 28]. We computationally integrated single-cell signatures of donor

newborn human lungs with single cell transcriptomic profiles of developing perinatal mouse lung to generate a trans-species cellular impression of the developing lung.

RESULTS

Cellular Landscape of the Donor Newborn Human Lung

To characterize cellular heterogeneity in the newborn human lung, we performed single cell RNA sequencing (scRNAseq) of protease-dissociated cells from two one-day old donor lung samples (Donor 1: GAB 38 weeks, and Donor 2: GAB 31 weeks). While these donors were deceased on day one of their life due to non-lung related complications (anencephaly), the lungs appeared normal in appearance in histopathological analyses, and were processed with warm ischemic time. To exclude low quality events, cells having fewer than 500 genes detected, or with $\geq$ 12.5% mitochondrial genes, were excluded (Figure S2). Prior studies have shown that high levels of mitochondrial gene expression is an indication of presence of multiplets [42], and hence by using the filtering threshold of greater than 12.5% mitochondrial genes would exclude potential doublets from the analytical dataset. This filtering resulted in an analytical dataset of 19,136 genes in 5499 cells (Table S1): 3001 cells from two separate captures on lung 1 and 2498 cells from two separate captures on lung 2. Since the two lungs were obtained from individuals born at different gestational age at birth (GAB), cells from each of the lungs were initially analyzed independently and their cellular composition was assessed at both major cell type level, as well as the level of cellular subtypes (Figure S3). Irrespective of the differences in terms of the estimated gestational age at birth, there were no observable differences in cellular composition of the two lungs, and hence cells from both lungs were combined using Canonical Correlation Analysis (CCA) as implemented in Seurat [43].

This data set was used for analysis and visualization by t-distributed stochastic neighbor embedding (t-SNE). We identified 15 separate clusters of cells, along with corresponding marker genes (Figure 1). Each cluster displayed relatively equivalent distribution of cells from both subjects (Figure 1A). Among these 15 clusters, four major cell types were identified on basis of expression of known markers (Figure 1B). Epithelial cells (n = 209) were defined by

expression of *EPCAM, SFTPB, SCGB1A1,* and *NKX2-1.* Endothelial cells (*n* = 1092) were defined by expression of *PECAM1, VWF, CLDN5,* and *CDH5*). Mesenchymal cells (*n* = 3553) were defined by expression of *ACTA2*, *ELN*, *COL1A1*, and *CYR61*. Immune cells (*n* = 618) were defined by expression of leukocyte and lymphocyte cell markers *PTPRC*, *CD8A*, *CD19*, and *CD3E* (Figure 1C).

Based upon selective expression, we identified marker genes for individual clusters (Figure 2A). Functional enrichment analysis successfully identified lung cell sub-types for each of the 15 clusters (Figure 2B and Table S1). The markers for each individual cluster are presented in Table S2.

A majority of the cells (>63%) appeared to be of mesenchymal origin. Distinct large populations of myofibroblasts (Cluster 0, n=820), matrix fibroblasts (Cluster 1, n=814), and smooth muscle cells (Cluster 2, n=592) were identified. In addition, two distinct populations of pericytes were observed (Cluster 5, n=419 and Cluster 6, n=398).

We also identified four separate endothelial cells clusters; Cluster 3 (n=567), Cluster 8 (n=321), Cluster 10 (n=146), and Cluster 12 (n=85). Interestingly, pathway analysis performed independently using the cluster marker genes using ToppGene Functional Annotation tool (ToppFun) [44], associated Cluster 3 cells with vascular development and Cluster 12 cells with integrin signaling pathways.

A much smaller fraction of cells (<4%) were identified as epithelial cells. Epithelial cells were separated into AT1 cells (Cluster 11, n=131) and AT2 cells (Cluster 14, n=14).

Interestingly, immune cells represented a sizeable fraction (11%) within the newborn human lung cells obtained from the two donors. Among immune cells (n = 649), we were able to distinguish multiple discrete populations including macrophages (Cluster 7, n=349), T cells (Cluster 9, n=190), and B cells (Cluster 13, n=79).

When analyzed independently, the proportion of immune cells varied between the two donors ranging from 17% in Donor 1 (GAB 38 weeks), to 4% in Donor 2 (GAB 31 weeks). While there have been prior reports of presence of immune cells in neonatal lungs, [45], we have previously shown the proportion of mixed immune cells in newborn lungs to range from 4 – 15% [40]. Further validation of the presence of these immune cells in newborn human lungs was performed by flow cytometry of single cell dissociates from additional age-matched lungs (one day old) as described in the section on Flow Cytometry in the Supplemental Methods. The percentage of leukocytes detected varied from donor to donor, and ranged from 3-14%, which was similar to the observed frequency of immune cells from the donor lungs used for generating the single cell transcriptomics data set (Figure S6).

Cellular Landscape of the Postnatal Mouse Lung

Single cell RNA sequencing of murine lung tissue was performed using custom Drop-seq technology as previously described [28]. Cells with less than 500 detected genes, and greater than 12.5% of transcript counts mapped to mitochondrial genes were removed. Filtering resulted in an analytical data set of 17508 genes, from 32849 cells (PND1 n=8003, PND3 n=8090, PND7 n=6324 and PND10 n=10432). All mouse cells were grouped into 32 clusters, and each cluster had relatively similar distribution of cells from individual time points (Figure 3A). Similar to the human cells, the four major cell types were readily identified based upon the expression of known selective marker gene expression (Figure S4, Table S3 and S8). Mesenchymal cells, again represented the largest fraction of the population (n=10678; 33%) and were further identified into different sub-types including of multiple clusters of matrix fibroblasts, myo-fibroblasts, stromal cells, and mixed fibroblasts. Endothelial cells comprised a sizeable portion of the mouse cells (n = 8891), and compared to human newborn lungs, mouse lungs appear to have relatively greater proportion of epithelial cells (n = 6133), which

were further sub-classified into pulmonary alveolar type I (AT1), alveolar type II (AT2), and ciliated respiratory epithelial cells. As in the human, immune cells were detected in the neonatal mouse lung (n = 7244) and were further classified as B-cells, T-cells, macrophages, monocytes, and myeloid cells among others (Figure 3B).

Integration of Newborn Human and Mouse Lung Data Sets

The two donor newborn human lungs differed in terms of gestational age at birth leading to minor morphological differences, however the cellular compositions were very similar across the two lungs (as shown in Figure S3) when it comes cellular annotations associated with the clusters of cells from individual donors. Admittedly, there would be differences in maturity status in the two donor lungs, however due to limited sample size, it is difficult to determine whether the differences between the two are due to technical variation (batch effect) or biological variation (maturity). In order to expand our analytical horizons, we subsequently combined the human and murine lung data sets (Figure 4). A total of 14,502 orthologous genes were identified using BioMart [46]. Canonical correlation analysis (CCA), implemented in Seurat, was used for data integration across the species. The combined dataset was filtered using the same criteria as applied independently on the human and mouse datasets (excluding cells having fewer than 500 genes detected, or with $\geq$ 12.5% mitochondrial genes), the species-integrated analytical data set contained a total of 29762 cells: 2327 (15%) human cells and 27435 (85%) mouse cells. Loss of some of the human and mouse cells can be attributed to the use of a limited set of genes (14, 502) that were used for integrating the data across species, and thereby leading to those cells being filtered due to not passing the filtering thresholds. In this integrated data set, we identified 17 clusters of cells, along with corresponding cluster marker genes. Each cluster was composed of a combination of mouse and human cells (Figure 4A, Table S4 and S9). We again identified four major cell types by known cell-type selective marker expression (Figure 4B-

D); mesenchymal cells (*n* = 9980, 15% human), endothelial cells (*n* = 8292, 8% human), epithelial cells (*n* = 5146, 0.5% human) and immune cells (*n* = 3342, 6% human).

Based upon selective expression, we identified marker genes for individual clusters (Figure 5A). Functional enrichment analysis successfully identified lung cell sub-types for each of the 17 clusters in the integrated data set representing postnatal lung tissues from human and mouse (Figure 5B and Table S4). We observed multiple clusters of mesenchymal cells; myofibroblasts (n=4592; Clusters 3, 10, and 11), matrix fibroblasts (n=4743; Clusters 1 and 14) and pericytes (n=645, Cluster 12). We observed multiple clusters of endothelial cells (n=8292; Clusters 0, 5, 9, and 15). We observed multiple immune cell populations including macrophages (n=3002, Cluster 4), T cells (n=1330, Cluster 8), B cells (n=1388, Cluster 7), and myeloid cells (n=585, Cluster 13) as well.

Estimating Developmental State of Human Cells

Lung development and morphogenesis occurs both prenatally and postnatally, and is typically divided into five phases, with the final alveolar phase occurring principally after birth in humans and rodents. The final stage of lung development, termed alveolarization or alveogenesis, begins prior to birth in humans and extends through at least the first decade of life, while occurring entirely postnatally in mice [20]. There exists a degree of uncertainty regarding the state of development states of the human and mouse lungs at the time of birth, since rodents and humans are born at different histological stages, alveolar in humans, but saccular in mice. The process of lung morphogenesis is a continuous process which varies by cellular types, as shown in newborn mouse lung [28]. Even though the two donor human lungs were collected at day 1 of birth, we hypothesized that due to the continuous nature of cellular development and dissociation, individual cells would be at different stages of their development throughout the

process of lung morphogenesis. We sought to determine the “cellular developmental state” of the newborn human lung in comparison to the postnatal mouse lung, using the integrated human-mouse data set. Since a one-day old human lung is developmentally and morphologically similar to 5-7 day old mouse lung, we aimed to estimate the developmental stage or maturity of the human cells in terms of the corresponding murine maturation state by using the full spectrum of the available murine lung development data. Even if we consider the early gestational age of one of the donor lungs, the timespan covered by the mouse lung data complements the developmental state of the human lungs, and is therefore most appropriate background for integration and estimation of developmental state of the human cells. We performed this analysis separately for each distinct cell type independently (Figure S1), performed Principal Component Analysis (PCA), and tested the relationship between each PC vector and the known age of the mouse cells. We calculated an “estimated age” for every human cell using its linear distance to 100 mouse cells in space defined by the age-related PC (PC^{Age}) that was statistically correlated with age and explained the greatest variance in the data set. The PC associated with age differed for each cell type, and the correlation coefficients (r) values, which were used as metric for identifying the PC related to age (Table S5). Irrespective of the cell type, the largest influence on the variance within each of them was due to species differences, as demonstrated by cells separating by species on the first principal component (PC1).

The estimated maturity of individual human cell types differed slightly, but primarily remained in the range of 5-9 murine days, consistent with the known histological relationships between human and mouse (Figure 6A). Interestingly, epithelial cells, endothelial cells and matrix fibroblasts displayed a more diverse distribution in estimated maturity. Matrix fibroblasts, which represented a large proportion of all cells displayed a somewhat bi-phasic pattern, where 29 % of cells appeared to be an estimated maturity stage consistent with other cell types (5-9 days),

while a second set of cells appeared to be of much younger estimated maturity stage (1-4 days) (Figure 6B and Figure S8A). Even though the maturation stages have been defined in terms of murine lung cell maturity, it is significant that we observe two different groups of cells that have been classified as matrix fibroblast. Both the early, and late maturity matrix fibroblasts were evenly distributed across both samples and showed no statistical difference in distribution across the samples. We identified marker genes for younger and more mature matrix fibroblast population using DESeq2 [47] leading to identification of 210 differentially expressed genes, of which 23 genes were over expressed in “the older matrix fibroblasts” (those presenting with an older estimated age), while 187 genes were over expressed in “younger matrix fibroblasts” (those presenting with a younger estimated age). When queried for cell types associated with these genes, we mostly observed then to be associated with lung or matrix fibroblasts, however, a subset 44 (out of 187) genes were found to be associated with progenitor fibroblasts. Pathway analysis using these 187 genes revealed smooth muscle, matrix, and collagen related pathways, along with oxidative stress, stress response, degranulation and scavenging related pathways were upregulated in the immature matrix fibroblasts (Figure S7 and Table S10).

One of the markers for younger matrix fibroblasts which had over two-fold induction was HES1, which was also expressed in other mesenchymal cells (pericytes and stromal cells) and endothelial cells as well (Figure S8B). We identified human cells expressing HES1 alone, and cells co-expressing HES1 with matrix fibroblast markers (COL6A3 or TCF21; Figure S8C-D) or markers for other cell types (PECAM1 for endothelial cells or CDH1 for epithelial cells; Figure S8E-F). We further demarcated both human and mouse cells expressing HES1 alone (Figure S9A, and cells co-expressing HES1 with matrix fibroblast markers (COL6A3 or TCF21; Figure S9B-C) or markers for other cell types (PECAM1 for endothelial cells or CDH1 for epithelial cells; Figure S8D-E). In both instances, we observed HES1 being coexpressed mesenchymal markers in human cells in cluster(s) identified as matrix fibroblasts. Finally, we tested whether

we could spatially resolve the older and younger matrix fibroblasts in the newborn human lung. We performed combined immunohistochemistry and in situ hybridization in three independent donor lungs of same age (one day old), to identify the expression of general- (COL6A3 and TCF21) and immature population-specific (HES1) markers at the cellular level. We were able to identify individual matrix fibroblasts (as defined by expression of COL6A3 or TCF21, but not PECAM1 or CDH1) that expressed HES1, as well as matrix fibroblasts that did not express HES1 (Figure 6C and Figure S8G). These data indicate the presence of a distinct group of immature matrix fibroblasts in the newborn human lung that display high expression of HES1 transcript.

DISCUSSION

Applying single cell RNA sequencing to newborn human lungs, we identified a diversity of pulmonary cells, including epithelial, fibroblast, immune, endothelial, and other cell subtypes based upon distinct gene expression patterns. We note a paucity in the capture of epithelial cells, which different from previous reports [33-35], but is consistent with our recent report of similar analyses of human fetal lung tissues [27]. Our prior studies using similar cell isolation protocols in older pediatric lung samples, demonstrated a higher proportion of epithelial cells [40]. Similar to our prior studies applying similar methodologies, we observed an over-representation of mesenchymal cell, which were further classified into subtypes namely, fibroblasts, pericytes, and stromal cells [27]. We also observed a sizeable cluster of endothelial cells, however, the number of cells were not sufficient to identify rare capillary type endothelial cell types seem in mouse lung cells [48]. Our analysis also indicates the presence of immune cells in the newborn lungs, which has not be widely reported before. While there have been multiple recent studies involving single cell sequencing of the human lung [33, 36, 37, 49], this is one of the first studies reporting the cellular composition of human lungs at the time of birth.

Although a novel and necessary study, it has to be acknowledged that some bias likely exists in the cells isolated and described. While the sensitivity of single cell RNA sequencing allows the discovery of high resolution cellular topography, it also gives rise to susceptibility to technical and procedural biases which sometimes hide the true biological signals [50]. Cellular recovery can be impacted by the methods used for mechanical and enzymatic disaggregation, processing time and reagent concentrations. Developing enzyme mediated digestion conditions will depend on which assays will be completed downstream. Our group's initial goal was to optimally release all cell types from primary human lung, including fibroblasts which are held in place by the proteinaceous extracellular matrix [40, 51]. The process requires careful testing of how changing conditions affects the quality of sample generated [52]. While aggressive or

prolonged digestion protocols may lead to cell death or fragmentation [53], gentle dissociation protocols may lead to greater capture of easily dissociated cells, which might be a challenge in the lung diseases characterized by altered matrix structures [54]. We did observe an absence in type II epithelial cell capture prior to selection, suggesting epithelial cell viability may contribute their diminished detection. In fact in a review of published datasets, Alexander et al, have reported that in normal lung tissue cell types, such as broad, flat, sail-like alveolar type I (AT1) cells and matrix-embedded fibroblasts, are difficult to dissociate and liberate and as such are relatively underrepresented when compared with cuboidal alveolar type II (AT2) cells and alveolar macrophages [53]. Low capture/detection of epithelial cells in the fetal and newborn human lung may be attributed to the developmental age of the studied samples, cellular stress encountered due to the dissociation protocols or difficulty in capture of these cells with the Chromium 10X protocol. Cellular capture from the young lungs is further complicated by the lack of knowledge regarding cell dissociation, different protease sensitivity, and cell survival during digestion and capture procedures. We performed two independent captures on each sample, one involving selection for high-quality cells by removing dead or dying cells by magnetic selection. Importantly, we noted consistent recovery of all major cell populations regardless of capture (Tables S1 and S4). Furthermore, the use of dissociated cells has been used by our group to develop in vitro air liquid interface cell culture model system [6, 9]. This provides an opportunity for further refinement of dissociation protocols to improve recovery of epithelial cell populations [27].

It is clear that although stages of lung development, and their morphological correlates, are highly conserved across species, significant differences exist in their relative length and timing [55]. An example is that the mouse lung is in the saccular stage at birth, while the human lung at term birth is in the alveolar stage. The newborn human lung is histologically and developmentally similar to a 1week old mouse lung (Figure S5). We took advantage of recent

data from the LungMAP program, describing postnatal mouse lung development at the single cell level, to infer the “estimated maturity” or “developmental state” of newborn human lung cells [28].

We have analyzed and annotated the species-specific single-cell datasets separately and helped us identify major cell types, as well as cellular subtypes. While the separate analyses preserve the intra-dataset heterogeneity, the combined analyses across the two species (human and mouse) increases the number of cells used for clustering and annotation, potentially allowing for identification of additional heterogeneity and rare cell populations. However, the combinatorial analytical process is more complex and computationally intensive, and may undermine some of the species-specific cell types [56]. In the combined analyses species level differences can be minimized by applying “batch-correction” approaches on the underlying transcriptomic data, and we have incorporated that by using gene names matched across species [46], and performing subsequent integration using CCA [43]. In order to alleviate concerns regarding the retention of the lineage and cell type information of the human cells when integrating with the murine cells, we tested for the cell type identities and observed that over 96% of human cells identified as mesenchymal, endothelial, and immune cells retain their identities in the combined classification as well (Table S11). Even in case of epithelial cells, which were only 24 to begin with, we observed 75% accuracy.

Our single cell data was generated from two donated lungs that were of one day of age, one of which was born full time (at 38 weeks of gestation), while the other was born prematurely (at 31 weeks of gestation). The lung as we know is a heterogenous organ and is composed of over 40 cell types. We hypothesized that different cell types may have different stages of development during the process of development. In order to assess the differential growth of developmental states of different cell types, we have integrated the human cells with the single cell data from

mouse lungs harvested from animals at different ages of development (post-natal days 1, 3, 7, and 10).

Majority of the human cells, regardless of cell type/lineage, were estimated to be 4 to 9 days of mouse age, consistent with the histological comparisons. For some cell types (e.g., matrix fibroblasts, endothelial cells, epithelial cells), greater diversity in estimated age was noted. The epithelial cell population was not large enough to separate known distinct lineages. Among the other cellular populations, the extent of endothelial cell diversity (e.g., large vs. small vessel), has been well documented [57]. We focused subsequent analysis on the matrix fibroblasts, as phenotypic diversity among this population is less well described.

Our observations on newborn human lung matrix fibroblast diversity are possibly one of the first to be reported in human, but there has been a prior report of different types of murine lung matrix fibroblasts in a mouse model of pulmonary fibrosis [58]. While Xie et. al. [58] focused on mouse lung fibroblasts, they did observe a class of murine progenitor lung fibroblasts. Interestingly, a set of 44 (out of 187) genes unregulated in the early matrix fibroblasts in our newborn human lungs were found to be associated with progenitor lung fibroblasts. Majority of newborn human lung matrix fibroblasts appeared to be more similar to younger mouse matrix fibroblasts and displayed higher levels of expression of HES1. These less mature fibroblasts were evenly derived from both the newborn lungs, as there was no significance difference in distribution observed across the two donor lungs irrespective of the difference in gestational ages of the two. HES1 is a regulator of Notch signaling and appears to actively suppress differentiation [59]. Interestingly, regulation of collagen expression by Notch is achieved through a Hes1-dependent mechanism [60]. Furthermore, HES1 appears to play a critical role in regulating lung fibroblast differentiation [61], and is known to be expressed in mucus cells from patients with chronic obstructive pulmonary disease, idiopathic pulmonary artery hypertension or IPF [62]. Hes1-knockout mice have been reported to have a phenotype of premature

differentiation and rigorous defects in nervous tissue [63]. Another gene that displayed higher expression in the younger/immature matrix fibroblasts was IGFBP7, which has previously been associated with resistance to lung cancer by performing tumor suppression function, especially in epithelial cells [64]. The younger/immature matrix fibroblasts may represent cells in an immature state, with importance for normal development, and may hint at a developmental origin for some adult diseases such as lung fibrosis.

We understand that the study is limited to the data from only two one day old human lung samples, which are difficult to obtain. However when integrated with the mouse lung cells, they provide sufficient information to develop a molecular map of the human neonatal lung, and potentially overcome the limitation of low sample numbers. Also, while the two samples differ in their gestational age, independent analysis of the two samples indicated similar cellular composition (Figure S3), thereby alleviating any concerns regarding age specific differences. Additionally, there is an underrepresentation of epithelial cells among the two subjects, which can be attributed to the dissociation protocol, which has previously been documented [40], and hence leaves the room for further refinement of the protocols leading to improved recovery of the pulmonary epithelial cells.

To summarize, here we report a dataset describing the transcriptome of newborn human lung cells defined using single cell RNA sequencing. Our results include markers for all major lung cell types including multiple populations of mesenchymal, endothelial, epithelial and immune cells. These lineage markers have been validated in the major pulmonary cells types through bulk RNA-seq [40]. We also successfully integrated the transcriptomes of newborn human cells with postnatal developing mouse lung cells, enabling the estimation of cell-type specific developmental states of human cells. The data show that maturation states, even though largely in the expected range of 4 to 9 murine postnatal days, differ by cell type. Integrated single cell

RNA profiling of human and mouse lung will help identify common and species-specific mechanisms of lung development and respiratory disease.

MATERIALS AND METHODS

Study Population

Two newborn (one-day old) lungs were donated for research and provided, with de-identified clinical data, through the federal United Network of Organ Sharing via the National Disease Research Interchange (NDRI) and International Institute for Advancement of Medicine (IIAM). While both the lungs were from individuals who were deceased at one day of life due to diagnosed anencephaly, they differed at their gestational age at birth (GAB). While both the donors were female, one of them (Donor 1) was a full term (GAB of 38 weeks), whereas the other (Donor 2) had a pre-term birth (GAB of 31 weeks). The organs were received by the LungMAP Human Tissue Core at the University of Rochester, and subjected to processing as previously described [40]. The LungMAP program and resulting studies are approved by the University of Rochester IRB (RSRB00047606). Sample demographic information for the donor lungs used in the study have been presented in Table S6.

Single cell suspension preparation

The right upper and middle lung lobes were digested to single cell suspensions using a four enzyme cocktail (collagenase A, DNase, dispase and elastase) according to LungMAP protocol, as described previously [40]. Isolated cells were resuspended in freezing media (90%FBS, 10%DMSO) at a concentration no more than $60x10^6$cells/mL, slow cooled to -80 C overnight and stored in liquid nitrogen until use.

Human Single Cell Sequencing

Unfractionated dissociated cells from each subject were rapidly thawed and, without resting, used for two separate captures of single cells for RNAseq. For each lung, one capture was preceded by removal of dead or dying cells through magnetic selection using Dead Cell Removal Kit (Miltenyi Biotech, Santa Barbara, CA). Cell capture and library production was

performed on the Chromium 10X Genomics system. Sequencing was performed on a HiSeq4000, with read alignment to GRCh38. Cells filtered to exclude low quality cells, and potential doublets, were used to create analytical dataset. Highly variable genes were identified using "FindVariableGenes" function in Seurat [65]. Principal component analysis (PCA) was used for dimension reduction based on only the highly variable genes. Top principal components (PCs) identified by JackStrawPlot(),graph based Louvain-Jaccard methods [66], were used for t-Distributed Stochastic Neighbor Embedding (tSNE) and clustering analysis. All single cell sequencing data analysis was performed using Seurat v2.4 [43, 67]. Cluster markers were defined by differential expression using a parametric Wilcoxon rank sum test at a corrected significance level of $p<0.05$ and upregulated the cluster compared to all other cells using the FindAllMarkers() command implemented in Seurat. Pathway analysis and cell type association was performed using ToppGene Functional Annotation tool (ToppFun) [44]. The human cell sequencing counts data will be made available for visualization and downloading through the LungMAP portal (https://lungmap.net/) and is currently available by request to the corresponding author.

Mouse Single Cell Sequencing

Animal protocols were approved by the Institutional Animal Care and Use Committee at Cincinnati Children's Medical Center (CCMC) in accordance with NIH guidelines. RNA from cells isolated from C57BL6/J mice (n = 2, at each time point) were used in the production of the mouse single cell RNA-seq data set using Drop-Seq platform, as previously described [28]. Filtered data were log transformed, scaled, clustered and represented by t-Distributed Stochastic Neighbor Embedding (t-SNE), similar to the analysis of human cells. Cell clusters were assigned to putative cell types based on inspecting the expression of known cell type markers, and the individual cluster markers analyzed using ToppFun. Single cell gene

expression data from mouse lungs from all four time points have been made available on Lung Gene Expression in Single Cell (LungGENS) web portal [21].

Integrating Human and Mouse Data

We integrated our newborn human lung data set with single cell sequencing data (Drop-Seq) from longitudinal postnatal (post-natal days 1, 3, 7, and 10) mouse lung samples [28] (Figure S5). These data characterized mouse datasets, hosted by LungMAP (https://lungmap.net/), were used as a reference to help define the human lung cell populations. The useMart and getLDS functions, within the biomaRt package, were used to identify human and mouse orthologues (https://bioconductor.org/packages/release/bioc/html/biomaRt.html). A list of 21,608 mouse MGI symbols were queried against HGNC symbols to identify 14,647 non-redundant orthologues. For all downstream analyses, HGNC symbols replaced MGI symbols within the mouse data set based on the biomaRt query [46]. We integrated Human cells (n=5499; *15%*) and Mouse cells (n=32849) using canonical correlation analysis (CCA) implemented within Seurat [43] (https://satijalab.org/seurat/)(Figure S1). No batch effects were evident following implementation of CCA. Marker genes from individual clusters were used to determine the cellular identity of co-clustered human cells in ToppFun.

Flow Cytometry

The presence of immune cell populations in newborn human lungs were validated by flow cytometry, essentially as previously published [68] . Frozen lung cells were thawed, blocked (2% serum in 1% BSA/DPBS) and stained for anti-hCD45 (APC-R700, clone HI30), anti-hCD235a (PE-Cy5.5, clone GA-R2), anti-hCD3 (PE-Cy7, clone SK7) (all from BD Biosciences, San Jose, CA) and anti-hHLA-DR (BV785, clone L243, Biolegend, San Diego, CA) and 7-AAD

(viability marker, BD Biosciences). Staining was assessed on a 4-laser 18-color FACSAria flow cytometer (Becton Dickenson, San Jose, CA). Single antibody stained Simply Cellular® compensation beads (Bangs Lab, Fishers, IN) were used for fluorescence overlap compensation. Fluorescence minus one (FMO) controls and heat-killed 7AAD stained cells were used to set expression gates for each antibody and for live/dead gating. Data were analyzed using FlowJo software (version 10; FlowJo LLC, Ashland, OR). Cell multiplets, DTo ensure high viability and to exclude lysis-resistant nucleated RBCs found in neonates, 7-AAD+ dead cells, and CD235a+ erythrocytes were detected and excluded from FACS analysis. From viable, RBC-depleted cells, mixed immune cells (MICs) were identified by CD45.

Estimation of Maturation State of Human Cells (In Terms of Murine Maturity)

For the human cells, we estimated their individual developmental state in terms of post-natal days (PND) age of mouse lung cells. In order to look at different cell types independently, we isolated the human and mouse cells of same type from the integrated human and mouse cell Seurat object, and created new objects of each of those cell types. As a first step, within the Seurat object of each individual cell types, we applied Principal Component Analysis (PCA) as a means of dimension reduction. Once we identified the loadings for each of the cells for first 10 principal components, we correlated the ages of the cells with the PC loadings, restricted to the mouse cells only. Based on the r values, we identified the PC with highest significant correlation with age and termed that as the age-associated PC or PC^{Age}.

Now, within this PC^{Age} space, for each human cell, we calculated the absolute distance between that human cell and every mouse cells of the same type. We subsequently identified the 100 nearest neighboring mouse cells from every human cell, in terms of sample loadings. We further adjusted the estimated age based on the proportion of mouse cells belonging to each of the time points (post-natal days 1, 3, 7, ad 10). The estimated maturation stages of the individual human cells was calculated according to the equation (1) listed below.

$$\text{Estimated Maturity of Human Cells (In Terms of Murine Maturation)} = \text{Average}\{(\text{Frequency}*\text{Age}*\text{Weight})_{PND1} + (\text{Frequency}*\text{Age}*\text{Weight})_{PND3} + (\text{Frequency}*\text{Age}*\text{Weight})_{PND7} + (\text{Frequency}*\text{Age}*\text{Weight})_{PND10}\} \quad \text{…Equation (1)}$$

In situ hybridization and Immunostaining

Fluorescence in situ hybridization (FISH), combined with immunofluorescence staining, were performed on formalin fixed, paraffin embedded native human postnatal lung sections (6 µm). FISH was completed using the RNAscope Fluorescent Multiplex Assay (Advanced Cell Diagnostics, Newark, CA, cat. # 323110) as previously described [27, 69], with minor adjustments. Treatment time with Protease Plus was reduced to 22 minutes. Tissues were incubated with the following probes: *HES1* (cat. # 311191-C4), *TCF21* (cat. # 470371) or *COL6A3* (cat. # 482631) (Advanced Cell Diagnostics). Following washing and signal development, tissues were blocked (3% bovine serum albumin in 5% normal goat sera and 0.1% Triton) and incubated overnight at 4°C with primary antibodies: CD31 (Neomarkers, RB-10333-P0) or CDH1 (BD Biosciences, 6315829). Slides were washed and incubated with Cy3-goat-anti-mouse or anti-rabbit-conjugated secondary antibodies (Jackson Immunoresearch Laboratories, Inc., West Grove, PA). Slides were counter-stained with DAPI (LifeTechnologies, Carlsbad, CA, cat. # DE571) and mounted using ProLong Diamond Antifade Mountant (LifeTechnologies). Images were acquired on an LSM710 confocal system with a 20x/0.8 Plan-APOCHROMAT objective lens [69].

Statistics

Statistical analysis and calculations involving estimation of maturity state of human cells and additional analyses were performed Minitab 17 (Minitab Inc. State College, PA). Additional images were generated using Graphpad Prism and Microsoft-Excel.

Acknowledgements:

We thank Corey Poole, Chin-Yi Chu, Christopher Anderson, the UR Genomics Research Center, and the URMC Flow Cytometry Shared Resource for assistance in generating and analyzing the 10X human lung single cell sequencing data. We also thank Mike Adam and Parvati Sudha for assistance in generating and analyzing the Dropseq mouse lung single cell sequencing data.

This work was supported by the Human Developing Lung Molecular Atlas Program (LungMAP) funded by National Heart, Lung, and Blood Institute of National Institutes of Health (U01HL122700 to G.S.P. and U01HL122642 to J.A.W.).

Author Contributions

T.J.M., J.A.W., Y.X., and G.S.P. conceived and designed the experiments. G.B., R.S.M., J.M.A., S.R., and S.S.P. conducted single-cell RNA-seq experiments. S.B., J.L.M, C.B., J.R.M., A.M., M.G., and Y.X. performed computational analysis of single-cell RNA-seq. S.D., D.A., and S.R. performed staining and immunohistochemistry experiments. G.B., R.S.M., and S.R. performed flow cytometry experiments. H.L.H., J.D., and J.H-W., performed sample acquisition, record management and data archiving. S.B., G.S.P., and T.J.M. wrote the manuscript. All authors contributed to the data interpretation, troubleshooting, manuscript writing, and editing. All the authors have read and approved the final manuscript and none of the authors have any competing interests.

References

1. Rannels, D.E., S.E. Dunsmore, and R.N. Grove, *Extracellular matrix synthesis and turnover by type II pulmonary epithelial cells.* Am J Physiol, 1992. **262**(5 Pt 1): p. L582-9.
2. Franks, T.J., et al., *Resident cellular components of the human lung: current knowledge and goals for research on cell phenotyping and function.* Proc Am Thorac Soc, 2008. **5**(7): p. 763-6.
3. Besnard, V., et al., *Maternal synchronization of gestational length and lung maturation.* PLoS One, 2011. **6**(11): p. e26682.
4. Beauchemin, K.J., et al., *Temporal dynamics of the developing lung transcriptome in three common inbred strains of laboratory mice reveals multiple stages of postnatal alveolar development.* PeerJ, 2016. **4**: p. e2318.
5. Xu, Y., et al., *Transcriptional programs controlling perinatal lung maturation.* PLoS One, 2012. **7**(8): p. e37046.
6. Anderson, C.S., et al., *CX3CR1 as a respiratory syncytial virus receptor in pediatric human lung.* Pediatr Res, 2020. **87**(5): p. 862-867.
7. Bhattacharya, S., et al., *Genome-wide transcriptional profiling reveals connective tissue mast cell accumulation in bronchopulmonary dysplasia.* Am J Respir Crit Care Med, 2012. **186**(4): p. 349-58.
8. Steiner, L.A., et al., *Disruption of normal patterns of FOXF1 expression in a lethal disorder of lung development.* J Med Genet, 2019.
9. Wang, Q., et al., *A novel in vitro model of primary human pediatric lung epithelial cells.* Pediatr Res, 2020. **87**(3): p. 511-517.
10. Ubags, N.D.J., et al., *Early origins of lung disease: towards an interdisciplinary approach.* Eur Respir Rev, 2020. **29**(157).
11. Whitsett, J.A., S.E. Wert, and T.E. Weaver, *Alveolar surfactant homeostasis and the pathogenesis of pulmonary disease.* Annu Rev Med, 2010. **61**: p. 105-19.
12. Besnard, V., et al., *Conditional deletion of Abca3 in alveolar type II cells alters surfactant homeostasis in newborn and adult mice.* Am J Physiol Lung Cell Mol Physiol, 2010. **298**(5): p. L646-59.
13. Whitsett, J.A., *Review: The intersection of surfactant homeostasis and innate host defense of the lung: lessons from newborn infants.* Innate Immun, 2010. **16**(3): p. 138-42.
14. Mariani, T.J., J.J. Reed, and S.D. Shapiro, *Expression profiling of the developing mouse lung: insights into the establishment of the extracellular matrix.* Am J Respir Cell Mol Biol, 2002. **26**(5): p. 541-8.
15. Mariani, T.J. and S.D. Shapiro, *Thomas A. Neff lecture. Application of expression profiling to the developing lung: identification of putative regulatory networks controlling matrix production.* Chest, 2002. **121**(3 Suppl): p. 42S-44S.
16. Kho, A.T., et al., *Expression profiles of the mouse lung identify a molecular signature of time-to-birth.* Am J Respir Cell Mol Biol, 2009. **40**(1): p. 47-57.
17. Mereness, J.A., et al., *Collagen VI Deficiency Results in Structural Abnormalities in the Mouse Lung.* Am J Pathol, 2020. **190**(2): p. 426-441.
18. Du, R., et al., *Platform dependence of inference on gene-wise and gene-set involvement in human lung development.* BMC Bioinformatics, 2009. **10**: p. 189.
19. Kho, A.T., et al., *Transcriptomic analysis of human lung development.* Am J Respir Crit Care Med, 2010. **181**(1): p. 54-63.
20. Bhattacharya, S. and T.J. Mariani, *Systems biology approaches to identify developmental bases for lung diseases.* Pediatr Res, 2013. **73**(4 Pt 2): p. 514-22.

21. Du, Y., et al., *Integration of transcriptomic and proteomic data identifies biological functions in cell populations from human infant lung.* Am J Physiol Lung Cell Mol Physiol, 2019. **317**(3): p. L347-L360.
22. Raredon, M.S.B., et al., *Single-cell connectomic analysis of adult mammalian lungs.* Sci Adv, 2019. **5**(12): p. eaaw3851.
23. Ardini-Poleske, M.E., et al., *LungMAP: The Molecular Atlas of Lung Development Program.* Am J Physiol Lung Cell Mol Physiol, 2017. **313**(5): p. L733-L740.
24. Pan, H., et al., *Comprehensive anatomic ontologies for lung development: A comparison of alveolar formation and maturation within mouse and human lung.* J Biomed Semantics, 2019. **10**(1): p. 18.
25. Dou, M., et al., *High-Throughput Single Cell Proteomics Enabled by Multiplex Isobaric Labeling in a Nanodroplet Sample Preparation Platform.* Anal Chem, 2019. **91**(20): p. 13119-13127.
26. Angelidis, I., et al., *An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics.* Nat Commun, 2019. **10**(1): p. 963.
27. Danopoulos, S., et al., *Transcriptional characterisation of human lung cells identifies novel mesenchymal lineage markers.* Eur Respir J, 2020. **55**(1).
28. Guo, M., et al., *Single cell RNA analysis identifies cellular heterogeneity and adaptive responses of the lung at birth.* Nat Commun, 2019. **10**(1): p. 37.
29. Zilionis, R., et al., *Single-Cell Transcriptomics of Human and Mouse Lung Cancers Reveals Conserved Myeloid Populations across Individuals and Species.* Immunity, 2019. **50**(5): p. 1317-1334 e10.
30. Schiller, H.B., et al., *The Human Lung Cell Atlas: A High-Resolution Reference Map of the Human Lung in Health and Disease.* Am J Respir Cell Mol Biol, 2019. **61**(1): p. 31-41.
31. Reyfman, P.A., et al., *Single-Cell Transcriptomic Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis.* Am J Respir Crit Care Med, 2019. **199**(12): p. 1517-1536.
32. Xu, Y., et al., *Single-cell RNA sequencing identifies diverse roles of epithelial cells in idiopathic pulmonary fibrosis.* JCI Insight, 2016. **1**(20): p. e90558.
33. Travaglini, K.J., et al., *A molecular cell atlas of the human lung from single-cell RNA sequencing.* Nature, 2020. **587**(7835): p. 619-625.
34. Habermann, A.C., et al., *Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis.* Science Advances, 2020. **6**(28): p. eaba1972.
35. Plasschaert, L.W., et al., *A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte.* Nature, 2018. **560**(7718): p. 377-381.
36. Cohen, M., et al., *Lung Single-Cell Signaling Interaction Map Reveals Basophil Role in Macrophage Imprinting.* Cell, 2018. **175**(4): p. 1031-1044.e18.
37. Treutlein, B., et al., *Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq.* Nature, 2014. **509**(7500): p. 371-375.
38. Muus, C., et al., *Single-cell meta-analysis of SARS-CoV-2 entry genes across tissues and demographics.* Nat Med, 2021. **27**(3): p. 546-559.
39. Wang, A., et al., *Single-cell multiomic profiling of human lungs reveals cell-type-specific and age-dynamic control of SARS-CoV2 host genes.* Elife, 2020. **9**.
40. Bandyopadhyay, G., et al., *Dissociation, cellular isolation, and initial molecular characterization of neonatal and pediatric human lung tissues.* Am J Physiol Lung Cell Mol Physiol, 2018. **315**(4): p. L576-L583.
41. Kyle, J.E., et al., *Cell type-resolved human lung lipidome reveals cellular cooperation in lung function.* Sci Rep, 2018. **8**(1): p. 13455.
42. Ilicic, T., et al., *Classification of low quality cells from single-cell RNA-seq data.* Genome Biology, 2016. **17**(1): p. 29.

43. Butler, A., et al., *Integrating single-cell transcriptomic data across different conditions, technologies, and species.* Nat Biotechnol, 2018. **36**(5): p. 411-420.
44. Chen, J., et al., *ToppGene Suite for gene list enrichment analysis and candidate gene prioritization.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W305-11.
45. Restori, K.H., et al., *Neonatal Immunity, Respiratory Virus Infections, and the Development of Asthma.* Frontiers in Immunology, 2018. **9**(1249).
46. Durinck, S., et al., *Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.* Nat Protoc, 2009. **4**(8): p. 1184-91.
47. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.
48. Gillich, A., et al., *Capillary cell-type specialization in the alveolus.* Nature, 2020. **586**(7831): p. 785-789.
49. Miller, A.J., et al., *In Vitro and In Vivo Development of the Human Airway at Single-Cell Resolution.* Developmental Cell, 2020. **53**(1): p. 117-128.e6.
50. Adam, M., A.S. Potter, and S.S. Potter, *Psychrophilic proteases dramatically reduce single-cell RNA-seq artifacts: a molecular atlas of kidney development.* Development, 2017. **144**(19): p. 3625-3632.
51. Quatromoni, J.G., et al., *An optimized disaggregation method for human lung tumors that preserves the phenotype and function of the immune cells.* J Leukoc Biol, 2015. **97**(1): p. 201-9.
52. Madissoon, E., et al., *scRNA-seq assessment of the human lung, spleen, and esophagus tissue stability after cold preservation.* Genome Biol, 2019. **21**(1): p. 1.
53. Alexander, M.J., G.R.S. Budinger, and P.A. Reyfman, *Breathing fresh air into respiratory research with single-cell RNA sequencing.* European Respiratory Review, 2020. **29**(156): p. 200060.
54. van den Brink, S.C., et al., *Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations.* Nature Methods, 2017. **14**(10): p. 935-936.
55. Danopoulos, S., et al., *Discordant roles for FGF ligands in lung branching morphogenesis between human and mouse.* J Pathol, 2019. **247**(2): p. 254-265.
56. Haghverdi, L., et al., *Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors.* Nat Biotechnol, 2018. **36**(5): p. 421-427.
57. Niethamer, T.K., et al., *Defining the role of pulmonary endothelial cell heterogeneity in the response to acute lung injury.* Elife, 2020. **9**.
58. Xie, T., et al., *Single-Cell Deconvolution of Fibroblast Heterogeneity in Mouse Pulmonary Fibrosis.* Cell Rep, 2018. **22**(13): p. 3625-3640.
59. Bray, S.J., *Notch signalling: a simple pathway becomes complex.* Nat Rev Mol Cell Biol, 2006. **7**(9): p. 678-89.
60. Hu, M., et al., *Notch signaling regulates col1alpha1 and col1alpha2 expression in airway fibroblasts.* Exp Biol Med (Maywood), 2014. **239**(12): p. 1589-96.
61. Liu, T., et al., *Notch1 signaling in FIZZ1 induction of myofibroblast differentiation.* Am J Pathol, 2009. **174**(5): p. 1745-55.
62. Plantier, L., et al., *Ectopic respiratory epithelial cell differentiation in bronchiolised distal airspaces in idiopathic pulmonary fibrosis.* Thorax, 2011. **66**(8): p. 651-7.
63. Kageyama, R., et al., *Roles of bHLH genes in neural stem cell differentiation.* Exp Cell Res, 2005. **306**(2): p. 343-8.
64. Wu, Q., et al., *Linkage disequilibrium and functional analysis of PRE1 insertion together with SNPs in the promoter region of IGFBP7 gene in different pig breeds.* J Appl Genet, 2018. **59**(2): p. 231-241.
65. Satija, R., et al., *Spatial reconstruction of single-cell gene expression data.* Nat Biotechnol, 2015. **33**(5): p. 495-502.

66. Shekhar, K., et al., *Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics.* Cell, 2016. **166**(5): p. 1308-1323 e30.
67. Stuart, T., et al., *Comprehensive Integration of Single-Cell Data.* Cell, 2019. **177**(7): p. 1888-1902 e21.
68. Misra, R.S., et al., *Flow-based sorting of neonatal lymphocyte populations for transcriptomics analysis.* J Immunol Methods, 2016. **437**: p. 13-20.
69. Danopoulos, S., et al., *Discordant roles for FGF ligands in lung branching morphogenesis between human and mouse.* J Pathol, 2018.

Figure 1 Identification of lung major cell types using single cell RNA sequencing of newborn human lung. (a) t-distributed Stochastic Neighbor Embedding (tSNE) analysis of cells. Cells are indicated by donor (Lung 1: GAB 39 Weeks; Lung 2: GAB 31 Weeks; GAB = Gestational Age at Birth). (b) The assignment of cell clusters to four major cell types, including endothelial cells, mesenchymal cells, immune cells, and epithelial cells. (c) Expression of some known cell type markers.

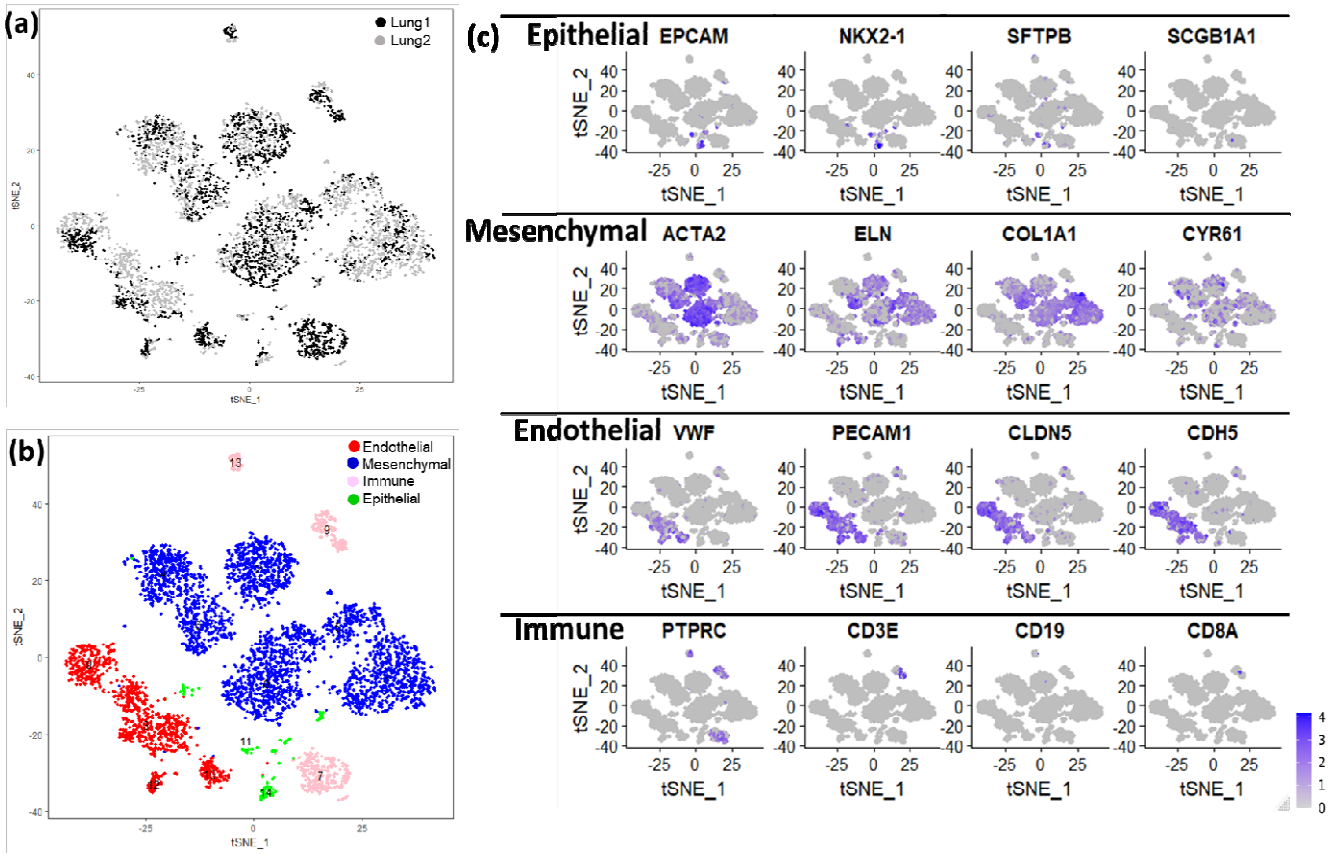


Figure 2: Identification of cell sub-type markers in newborn human lungs. (a) Gene expression patterns of select markers for corresponding cell clusters (5499 cells obtained from two newborn human lungs. (b) Assignment of cell types to 15 distinct cell clusters.

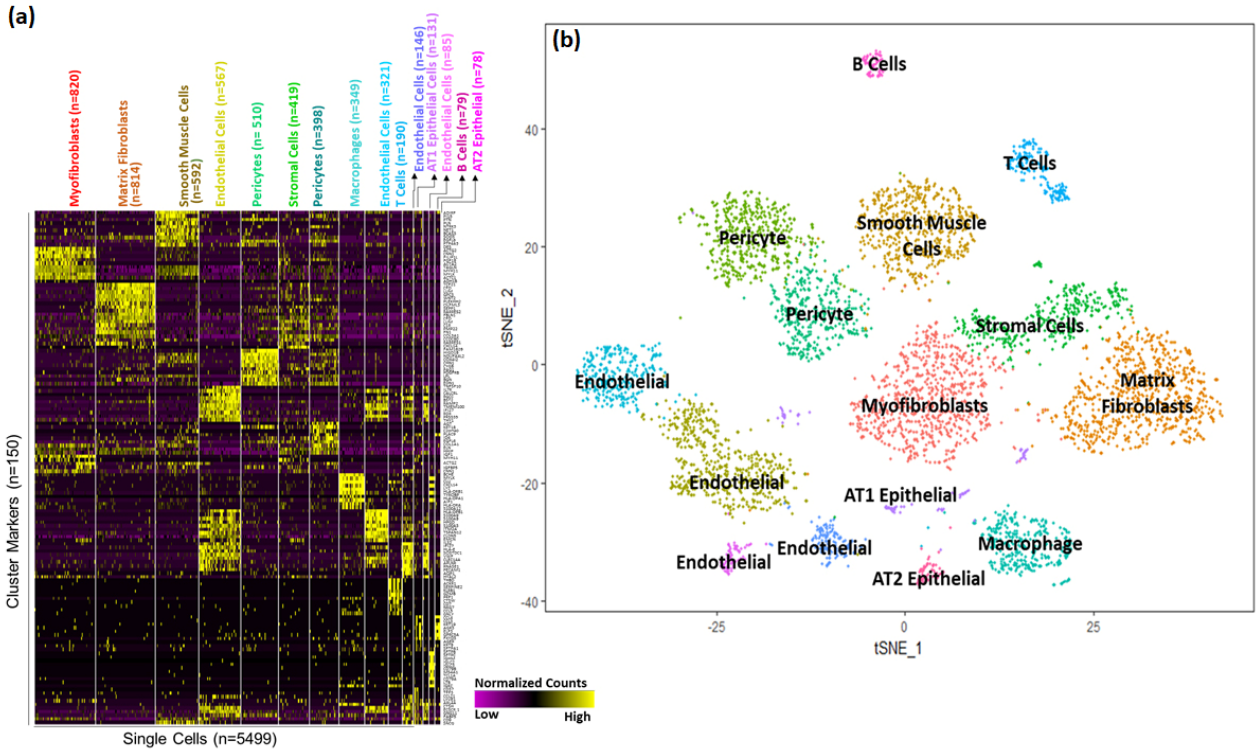


Figure 3: Identification of cell types in mouse lung. (a) t-distributed Stochastic Neighbor Embedding (tSNE) analysis of cells. Cells are colored by mouse age. (b) Visualization of cell clusters in tSNE plot of cells, with assignment of cell types to 32 distinct tSNE clusters. Mouse lung transcriptomic profile data was generated at Cincinnati Children's Hospital and Medical Center (CCHMC) [28].

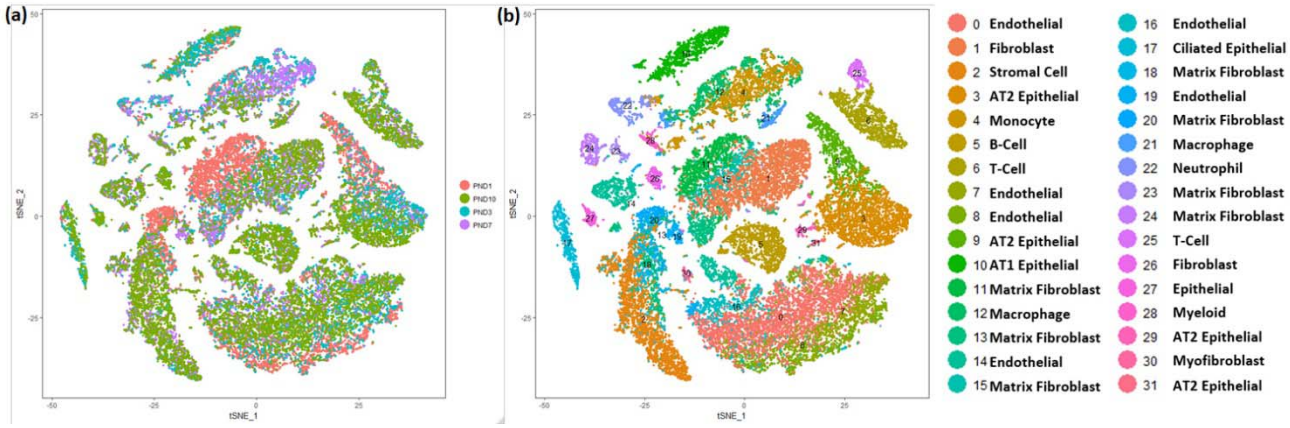


Figure 4: Integration of human and mouse lung cell data sets. (a) t-distributed Stochastic Neighbor Embedding (tSNE) analysis of cells. Cells are indicated by species. (b) Expression of known cell type markers in tSNE plot of cells in the integrated data set. (c) Visualization of mouse cell clusters in tSNE plot of cells grouped by major cell types. In the integrated object created from combining both human and mouse lung cells, the assignment of cell clusters to four major cell types, including endothelial cells, mesenchymal cells, immune cells, and epithelial cells. (d) Proportion of cells derived from human and mouse data. In the integrated dataset 9% of cells are human; human mesenchymal cells (15%) are over-represented, but endothelial (8.0%), immune (6.0 %) and epithelial cells (0.5%) are under-represented.

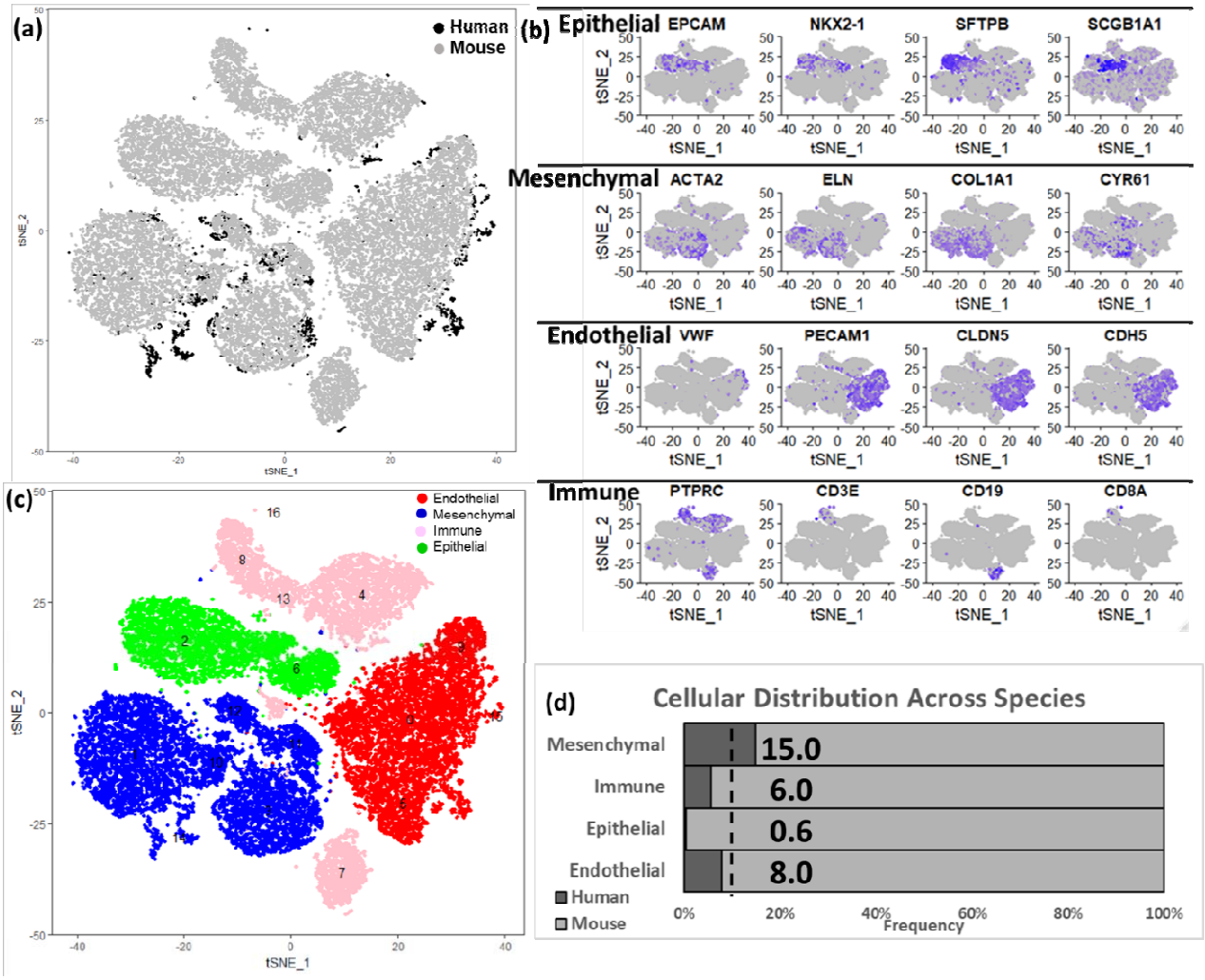


Figure 5: Identification of cellular sub-types in combined human and mouse lungs. (a) Gene expression patterns of select markers for corresponding cell types is shown in the heatmap. (b) The *t*-distributed stochastic neighbor embedding (tSNE) visualization shows unsupervised transcriptomic clustering, revealing 17 distinct cellular identities.

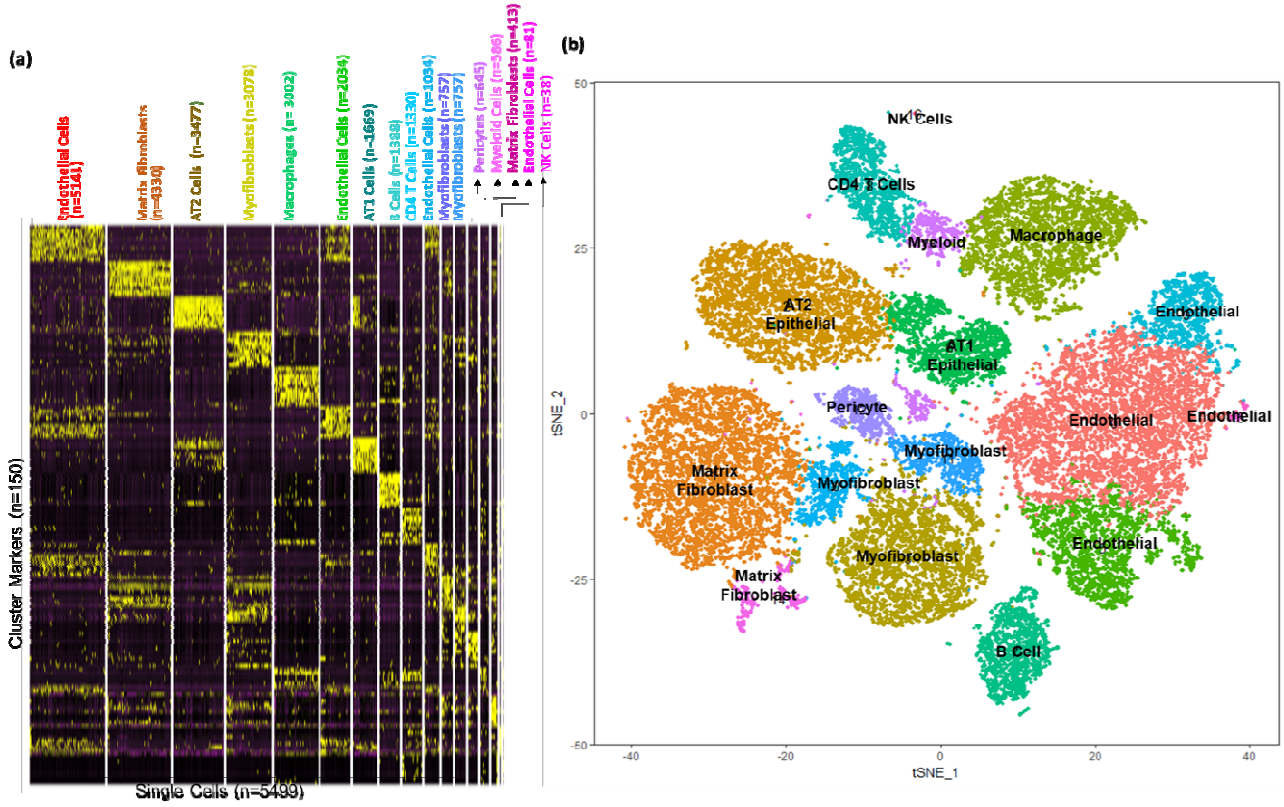
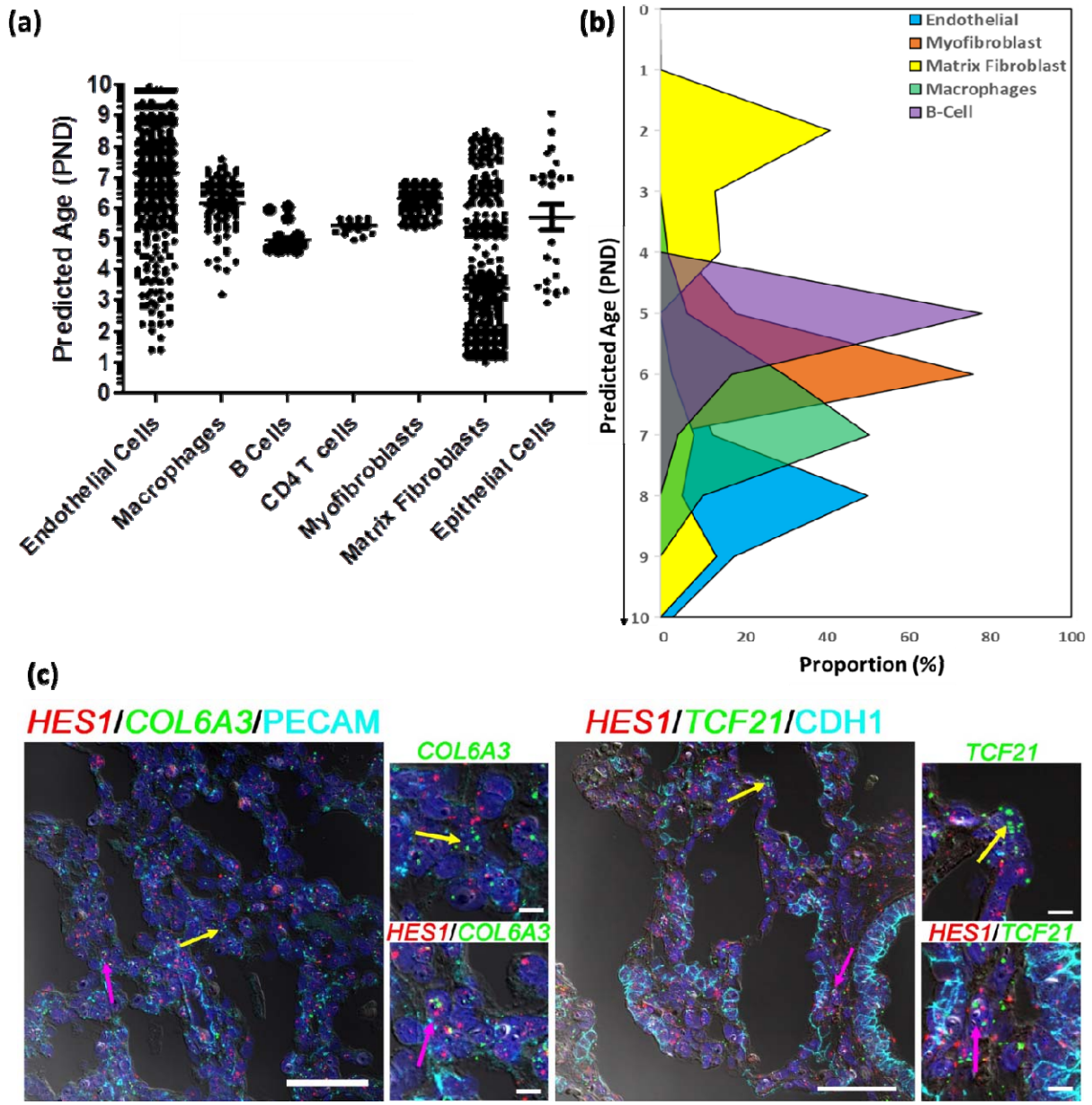


Figure 6: Estimating developmental states of human cells. (a) Distribution of the estimated ages of the human cells derived from post -natal age (PND) of 100 nearest mouse cells to each of the human cells. (b) Proportion of cells of individual human cell types at each stage of development defined in terms of estimated post-natal day age of mouse. (c) Fluorescent in situ hybridization (FISH) combined with immunofluorescence of Immature Matrix Fibroblast marker *HES1* (red), Non-Mesenchymal Cell Markers, *PECAM1* or *CDH1* cyan), and Mesenchymal Cell Markers *COL6A3 or TCF21* (green) on newborn human lung sections from a donor lung of 1 day of age. Pink arrows indicated the presence of immature matrix fibroblasts shown by co-localization of *HES1* (red) and Mesenchymal Cell Markers *COL6A3 or TCF21* (green), while yellow arrows indicate non HES1 expressing cells. The scale bar is 50um. PND = Post-Natal Days.


(a)
Predicted Age (PND)
Endothelial Cells
Macrophages
B Cells
CD4 T cells
Myofibroblasts
Matrix Fibroblasts
Epithelial Cells
(b)
Endothelial
Myofibroblast
Matrix Fibroblast
Macrophages
B-Cell
Predicted Age (PND)
Proportion (%)
(c)
HES1/COL6A3/PECAM
COL6A3
HES1/COL6A3
HES1/TCF21/CDH1
TCF21
HES1/TCF21