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 Bandopadhyay[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, Clinical & Translational Science Institute, University of Rochester, [5]Department of Pediatrics, Cincinnati Children's Hospital Medical Center and [6]Saban Research Institute, Children's Hospital 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: 5
No. of Supplemental Figures: 7
Word count (manuscript):4214
Word count (abstract): 316

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

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. Single cell RNA sequencing generated transcriptional profiles of 5500 cells obtained from two newborn human lungs from the LungMAP Human Tissue Core Biorepository. 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 an additional 32000 postnatal day 1 through 10 mouse lung cells generated by the LungMAP Cincinnati Research Center was integrated with the human data. Transcriptional interrogation of 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 population. Computational integration of newborn human and postnatal mouse lung development cellular transcriptomes facilitated the identification of distinct epithelial lineages including AT1, AT2 and ciliated epithelial cells. Integration of the newborn human and mouse cellular transcriptomes also demonstrated cell type-specific differences in maturation states of newborn human lung cells. In particular, newborn human lung matrix fibroblasts could be separated into those representative of younger cells (n=393), or older cells (n=158). Cells with each molecular profile were spatially resolved within newborn human lung tissue. This is the first comprehensive molecular map of the cellular landscape of neonatal human lung, including biomarkers for cells at distinct states of maturity.

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 intracellular 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.

Transition of the lung from fetal to neonatal states is highly complex, and has been characterized in murine models [10-16]. While early stages of human fetal lung development have been characterized at the molecular level [17, 18]), data describing the newborn human lung is lacking [19]. Although, molecular profiling has been applied to pre-viable human lung [18, 20], 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). Additionally, most high-throughput molecular studies of lung development have used whole lung tissue [10], limiting insights into the activities of and interactions among different cell types. Single cell RNA-seq enables transcriptomic mapping of individual cells to measure and understand cellular heterogeneity and responses in complex biological systems [21].

LungMAP was developed to generate detailed structural and molecular data regarding normal perinatal and postnatal lung development in the mouse and human [22]. 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 [20, 22-24]. 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 [21, 25]. We computationally integrate single-cell signatures of newborn human lungs with single cell transcriptomic profiles of developing perinatal mouse lung to generate a trans-species cellular impression of the developing lung.

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 suspected anencephaly, they differed at their gestational age at birth (GAB). One of them was a full term (GAB of 38 weeks), whereas the other hadwas 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 [23]. The LungMAP program and resulting studies are approved by the University of Rochester IRB (RSRB00047606).

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 [23]. 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 magnetic selection (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 "MeanVarPlot" function in Seurat [26]. 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 [27], were used for t-Distributed Stochastic Neighbor Embedding (tSNE) and clustering analysis. All single cell sequencing data analysis was performed using Seurat v2.4 [28, 29]. No significant differences in overall capture quality and data were observed between the two runs from each individual lung, therefore data from both captures were combined. Differential expression was defined using a parametric Wilcoxon rank sum test at a corrected significance level of $p<0.05$. Pathway analysis and cell type association was performed using ToppGene Functional Annotation tool (ToppFun) [30].

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. C57BL6/J mice were used in the production of the mouse single cell RNA-seq data set as previously described [21]. 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 [30].

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 [21]. 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 [31]. We integrated Human cells (n=5499; *15%*) and Mouse cells (n=32849) using canonical correlation analysis (CCA) implemented within Seurat [28] (https://satijalab.org/seurat/)(Figure S1). No batch effects were evident following implementation of CCA. Markers genes from individual clusters were used to determine the cellular identity of co-clustered human cells in ToppFun.

Estimation of Human Cell Age

We used the defined age of mouse cells to estimate the age of the human cells in the integrated data set. A schematic describing the approach used is provided in Supplemental Figure S1. First, we separated the data set based upon major cell type clusters (e.g., endothelial cells, matrix fibroblasts, etc.) using cell type-specific annotations. Next, we identified a surrogate of mouse age in each cell type independently, using principal components (PC) analysis. For each cell type, the most informative PC which was significantly correlated with age was used as the age surrogate (PC^{Age}). We then measured the linear (Euclidean) distance between each human cell and the 100 nearest mouse cells in PC^{Age}. The human cell age was defined as the average age of the nearest mouse cells, adjusted for the proportion of cells in the data set from each mouse age (Figure S1). Additional information is provided in the Supplemental Methods section.

Flow Cytometry

The presence of immune cell populations in newborn human lungs were validated by flow cytometry, essentially as previously published [32] . 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, Dead cells (7-AAD+) and erythrocytes (CD235a+) were excluded from analysis.

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 [33], 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 [34].

Results

Cellular Landscape of the 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 lung samples. While the two lungs were obtained from individuals born at different GAB, there were no observable differences in cellular composition among the two, and hence cells from both lungs were combined using Canonical Correlation Analysis (CCA) as implemented in Seurat [28]. When combined, the analytic dataset comprised of a total of 5499 cells (Table S1), with an average detection of 2000-3000 genes per cell (Figure S2). To exclude low quality events, cells having fewer than 500 UMIs detected, or with $\geq$ 12.5% mitochondrial genes, were excluded (Figure S2). This filtering resulted in an analytical dataset of 19,136 genes in 5499 cells; 3001 cells from two separate captures on lung 1 and 2498 cells from two separate captures on lung 2.

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, as defined by known selective markers, were identified on basis of expression of known markers (Figure 1D). 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*.

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 S4.

A majority of the cells (>63%) appeared to be of mesenchymal origin. Distinct large populations of myofibroblasts (Cluster 0, n=820) expressing *ACTG2, DES, ACTA2, TAGLN, CNN1*, matrix fibroblasts (Cluster 1, n=814) expressing *CFD, ADH1B, LUM, GPC3, TCF21,* smooth muscle cells (Cluster 2, n=592) expressing *ADIRF, PI15, PTN, SOD3, PLN, NTRK3* were identified. Two distinct populations of pericytes were observed, one cluster expressed FAM162B, HIGD1B, NDUFA4L2, COX4I2, CHN1 (Cluster 5, n=419), while the other expressed *PRSS35, THY1, AGT, ID4, COL1A1*(Cluster 6, n=398).

We also identified four separate endothelial cells clusters; Cluster 3 (n=567) defined by expression of RGCC2, EDN1, RAMP23, CA4, IFI273 among others, Cluster 8 (n=321) defined by expression of HPGD3, HLA-E8, ITM2A1, EMCN1, CLDN58 among others, Cluster 10 (n=146) defined by expression of ACKR1, PTGDS, VWF, HYAL24, SLCO2A12, among others, and Cluster 12 (n=85) defined by expression of SERPINE2, GLUL7, ID19, GJA42, SLC9A3R29, among others. Interestingly, functional analysis of the cluster marker genes associated Cluster 3 cells with vascular development and Cluster 12 cells with integrin signaling.

A much smaller fraction of cells (<4%) were identified as epithelial cells. Epithelial cells were separated into AT1 cells (Cluster 11, n=131) expressing *CCL21, TFF3, SFTPB, AGR3, AGER*) and AT2 cells (Cluster 14, n=14) expressing *SFTPC, SFTPB, AGER, SFTPA1, KRT19*.

Interestingly, immune cells represented a sizeable fraction (11%) within the newborn human lung. Among immune cells, we were able to distinguish multiple discrete populations including macrophages (Cluster 7, n=349) expressing *S100A8, S100A9, LYZ, HLA-DRA, HLA-DRB1,* T

cells (Cluster 9, n=190) expressing GNLY, NKG7, KLRB1, GZMB, CCL4 among others, and B cells (Cluster 13, n=79) expressing IGHM, IGKC, IGLC2, IGLC3, CD79B, among others.

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). To 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. The percentage of leukocytes detected varied from donor to donor, and ranged from 3-14%, which was consistent with the frequency of immune cells observed in the single cell transcriptomics data set (Figure S5).

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 [21]. Cells with less than 500 detected genes, and greater than 10% of transcript counts mapped to mitochondrial genes were removed. Filtering the cells based on the aforementioned criteria 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, with and each cluster had relatively similar distribution of cells from individual time points (Figure 2A). Similar to the human cells, the four major cell types were readily identified based upon the expression of known selective marker gene expression (Figure S3, Table S3). Mesenchymal cells, again representing the largest fraction of the population (n=10678;33%), were defined by expression of *ACTA2*, *ELN*, *COL1A1*, and *CYR61.* The mesenchymal cell were 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 were defined by expression of *PECAM1, VWF, CLDN5,* and *CDH5.* In the mouse data set, epithelial cells were well

represented (*n* = 6133) and defined by the expression of *EPCAM, SFTPB, SCGB1A1,* and *NKX2-1*. Epithelial cells clusters 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 frequent in the neonatal mouse lung (*n* = 7244) as defined by expression of *PTPRC*, *CD8A*, *CD19*, and *CD3E*). The immune cells were further classified as B-cells, T-cells, macrophages, monocytes, and myeloid cells among others (Figure 2B). Compared to human newborn lungs, mouse lungs appear to have relatively greater proportion of epithelial cells. In addition, even among the mesenchymal cells, there appears to be a greater proportion of matrix fibroblasts.

Integration of Newborn Human and Mouse Lung Data Sets

We next combined the human and murine lung data sets (Figure 4). A total of 14,502 orthologous genes were identified using BioMart [31]. Canonical correlation analysis (CCA), implemented in Seurat, was used for data integration across the species. After performing data quality filtering similar to the human and mouse only datasets, the species-integrated analytical data set contained a total of 29762 cells: 2327 (15%) human cells and 27435 (85%) mouse cells. 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 and Table S4). We again identified four major cell types by known cell-type selective marker expression (Figure 4B-D); mesenchymal cells (*n* = 9980, 25% human), endothelial cells (*n* = 8292, 12.5% human), epithelial cells (*n* = 5146, 4% human) and immune cells (*n* = 6344, 7% 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 Maturity of Human Cells

There exists a degree of uncertainty regarding the state of maturity of the human and mouse lung at the time of birth, since rodents and humans are born at different histological stages, alveolar in humans, but saccular in mice. We sought to determine the “cellular maturity” of the newborn human lung in comparison to the postnatal mouse lung, using the integrated human-mouse data set. We performed this analysis separately for each distinct cell type (Figure S1), in order to identify age as a contributing variable. Cell types were assigned based upon annotations described above and separated into independent data sets. For each cell type data set, we independently performed PCA, and tested the relationship between each PC vector and the known age of the mouse cells. For simplicity, the surrogate for age was chosen as the single PC that was statistically correlated with age and explained the greatest variance in the data set. We calculated the age of every human cell using its linear distance to 100 mouse cells in space defined by the age-related PC (PC^{Age}). 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).

The estimated age of individual human cell types differed slightly, but primarily remained in the range of 5-9 mouse 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 age. 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 age consistent with other cell types (5-9 days), while a second set of cells appeared to be of much younger estimated age (1-4 days) (Figure 6B and Figure S7A). We identified marker genes for younger and more mature matrix fibroblast population using DESeq2 [35] leading to identification of 210 differentially expressed genes. Among the 23 genes that were over expressed in "the mature matrix fibroblasts" (those presenting with an older estimated age), included *B2M, CYBA, and CCBE1*. Among the 187 genes that were over expressed in "immature matrix fibroblasts" (those presenting with a younger estimated age), included *IGFBP7, HES1, RGS3, TAGLN, C11orf96 and EGFL6*. 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 affected in the immature matrix fibroblasts (Figure S6).

One of the markers for younger matrix fibroblasts, HES1, was also expressed in other mesenchymal cells (pericytes and stromal cells) and endothelial cells as well (Figure S7B). We further identified cells expressing HES1 alone, and cells coexpressing it with matrix fibroblast markers (COL6A3 or TCF21; Figure S7C-D) or markers for other cell types (PECAM1 for endothelial cells or CDH1 for epithelial cells; Figure S7E-F). Finally, we tested whether we could identify the older/mature and younger/immature matrix fibroblasts in the newborn human lung. We performed combined immunohistochemistry and in situ hybridization 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). 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

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. However, given recent advancements in high-throughput molecular profiling technologies, rapid progress is being made [21, 25, 33]. 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, 36-38]. Here, we report transcriptional analysis of 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 maturity.

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. Although a novel and necessary study, it has to be acknowledged that some bias likely exists in the cells isolated and herein. We note a paucity in the capture of epithelial cells, consistent with our recent report of similar analyses of human fetal lung tissues [33]. Our prior studies using similar cell isolation protocols in older pediatric lung samples, demonstrated a higher proportion of epithelial cells [23]. Low capture/detection of epithelial cells in the fetal and newborn human lung may be attributed to the developmental age of the studied samples 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 removal of unhealthy cells. Importantly, we noted consistent recovery of all major cell populations regardless of capture (Tables S1 and S4). We did observe an absence in type II epithelial cell capture prior to selection, suggesting epithelial cell viability may contribute their diminished detection.

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 [39]. 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 S4). We took advantage of recent data from the LungMAP program, describing postnatal mouse lung development at the single cell level, to infer the “age” or “maturity” of newborn human lung cells. 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 [40]. We focused subsequent analysis on the matrix fibroblasts, as phenotypic diversity among this population is less well described.

Our data on newborn human lung matrix fibroblast diversity are consistent with a prior report of different types of murine lung matrix fibroblasts [41]. The 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. HES1 is a regulator of Notch signaling and appears to actively suppress differentiation [42]. Interestingly, regulation of collagen expression by Notch is achieved through a Hes1-dependent mechanism [43]. Furthermore, HES1 appears to play a

critical role in regulating lung fibroblast differentiation [44]. HES1 is known to be expressed in mucus cells from patients with chronic obstructive pulmonary disease, idiopathic pulmonary artery hypertension or IPF [45]. Another gene that displayed higher expression in the younger/immature matrix fibroblasts was IGFBP7, which has previously been associated with resistance to lung cancer [46]. 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.

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. We also successfully integrated the transcriptomes of newborn human cells with postnatal developing mouse lung cells, enabling the estimation of cell-type specific maturity 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.

Acknowledgements:

We thank Corey Poole, Chin-Yi Chu, Christopher Anderson, 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., 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.

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. 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.

11. 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.

12. 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.

13. 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.

14. 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.

15. 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.

16. 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.

17. 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.

18. Kho, A.T., et al., *Transcriptomic analysis of human lung development.* Am J Respir Crit Care Med, 2010. **181**(1): p. 54-63.

19. 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.

20. 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.

21. 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.

22. 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.

23. 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.

24. Kyle, J.E., et al., *Cell type-resolved human lung lipidome reveals cellular cooperation in lung function.* Sci Rep, 2018. **8**(1): p. 13455.

25. 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.

26. Satija, R., et al., *Spatial reconstruction of single-cell gene expression data.* Nat Biotechnol, 2015. **33**(5): p. 495-502.

27. Shekhar, K., et al., *Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics.* Cell, 2016. **166**(5): p. 1308-1323 e30.

28. Butler, A., et al., *Integrating single-cell transcriptomic data across different conditions, technologies, and species.* Nat Biotechnol, 2018. **36**(5): p. 411-420.

29. Stuart, T., et al., *Comprehensive Integration of Single-Cell Data.* Cell, 2019. **177**(7): p. 1888-1902 e21.

30. 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.

31. 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.

32. Misra, R.S., et al., *Flow-based sorting of neonatal lymphocyte populations for transcriptomics analysis.* J Immunol Methods, 2016. **437**: p. 13-20.

33. Danopoulos, S., et al., *Transcriptional characterisation of human lung cells identifies novel mesenchymal lineage markers.* Eur Respir J, 2020. **55**(1).

34. Danopoulos, S., et al., *Discordant roles for FGF ligands in lung branching morphogenesis between human and mouse.* J Pathol, 2018.

35. 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.

36. Raredon, M.S.B., et al., *Single-cell connectomic analysis of adult mammalian lungs.* Sci Adv, 2019. **5**(12): p. eaaw3851.

37. 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.

38. 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.

39. 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.

40. Niethamer, T.K., et al., *Defining the role of pulmonary endothelial cell heterogeneity in the response to acute lung injury.* Elife, 2020. **9**.

41. Xie, T., et al., *Single-Cell Deconvolution of Fibroblast Heterogeneity in Mouse Pulmonary Fibrosis.* Cell Rep, 2018. **22**(13): p. 3625-3640.

42. Bray, S.J., *Notch signalling: a simple pathway becomes complex.* Nat Rev Mol Cell Biol, 2006. **7**(9): p. 678-89.

43. Hu, M., et al., *Notch signaling regulates col1alpha1 and col1alpha2 expression in airway fibroblasts.* Exp Biol Med (Maywood), 2014. **239**(12): p. 1589-96.

44. Liu, T., et al., *Notch1 signaling in FIZZ1 induction of myofibroblast differentiation.* Am J Pathol, 2009. **174**(5): p. 1745-55.

45. Plantier, L., et al., *Ectopic respiratory epithelial cell differentiation in bronchiolised distal airspaces in idiopathic pulmonary fibrosis.* Thorax, 2011. **66**(8): p. 651-7.

46. 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.

Figure Legends:

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. (b) Visualization of distinct cell clusters in tSNE plot. (c) Expression of some known cell type markers. (d) The assignment of cell clusters to four major cell types, including endothelial cells, mesenchymal cells, immune cells, and epithelial cells.

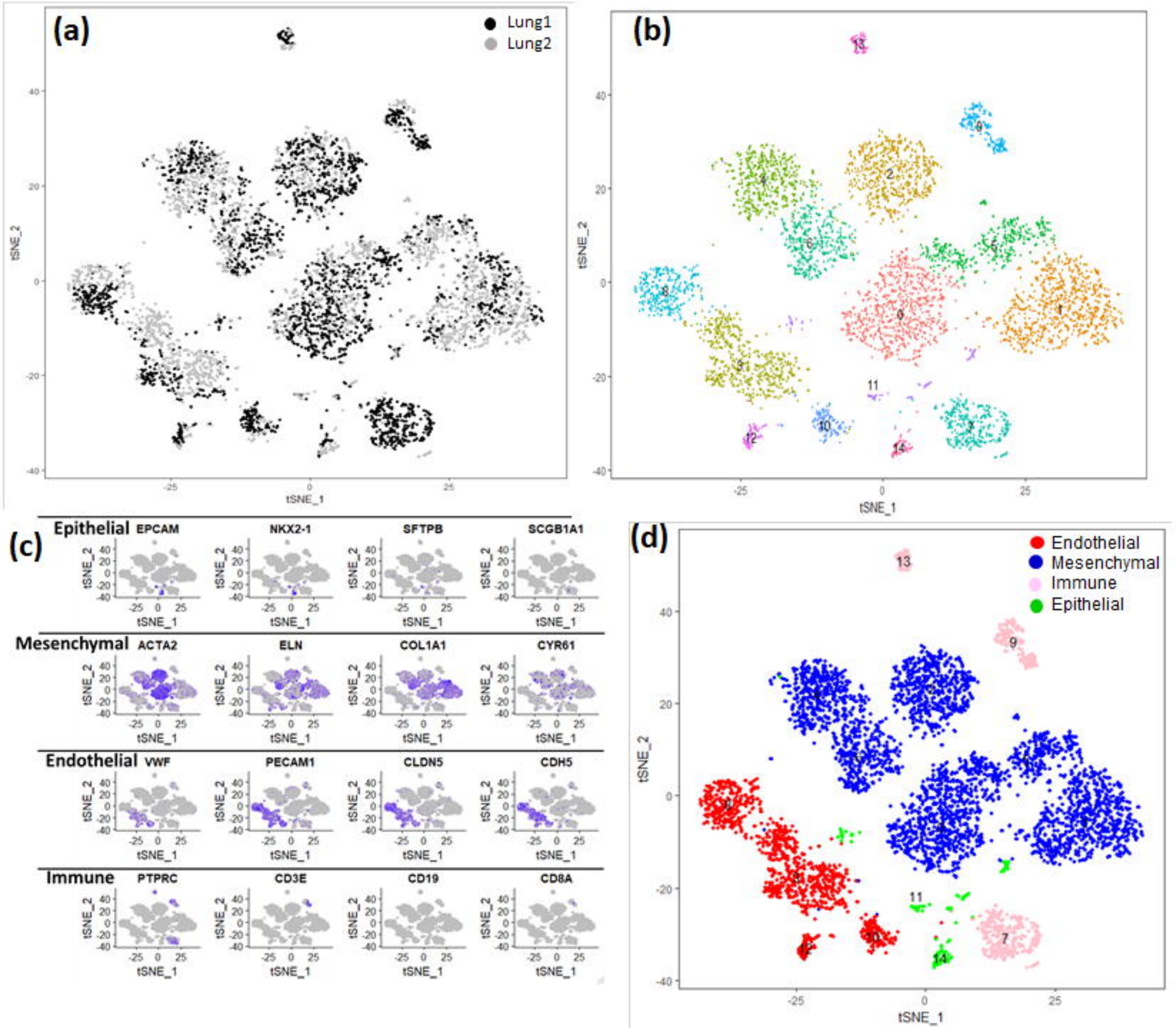
Figure 2: Identification of cell sub-type markers in newborn human lungs. (a)Genes expression patterns of select markers of corresponding cell clusters. (b) Assignment of cell types to 15 distinct tSNE clusters.

Figure 3: Identification of lung cell types using 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.

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 15% of cells are human; human mesenchymal cells (25%) are over-represented, but endothelial (12.5%), immune (7%) and epithelial cells (4%) are under-represented.

Figure 5: Identification of cellular sub-types in combined human and mouse lungs. (a) Gene expression patterns of select markers of corresponding cell types is shown in the heatmap. (b) The *t*-distributed stochastic neighbor embedding (tSNE) visualization shows unsupervised transcriptomic clustering, revealing 18 distinct cellular identities.

Figure 6: Estimating maturity 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 maturity 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 3 donor lungs 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).


(a)
Lung1
Lung2
tSNE_2
tSNE_1
(b)
tSNE_2
tSNE_1
(c)
Epithelial
EPCAM
NKX2-1
SFTPB
SCGB1A1
Mesenchymal
ACTA2
ELN
COL1A1
CYR61
Endothelial
VWF
PECAM1
CLDN5
CDH5
Immune
PTPRC
CD3E
CD19
CD8A
tSNE_2
tSNE_1
(d)
Endothelial
Mesenchymal
Immune
Epithelial
tSNE_2
tSNE_1

(a)

Myofibroblasts (n=820)
Matrix Fibroblasts (n=814)
Smooth Muscle Cells (n=592)
Endothelial Cells (n=567)
Pericytes (n= 510)
Stromal Cells (n=419)
Pericytes (n=398)
Macrophages (n=349)
Endothelial Cells (n=321)
T Cells (n=190)
Endothelial Cells (n=146)
AT1 Epithelial Cells (n=131)
Endothelial Cells (n=85)
B Cells (n=79)
AT2 Epithelial (n=78)

Cluster Markers (n=150)

Single Cells (n=5499)

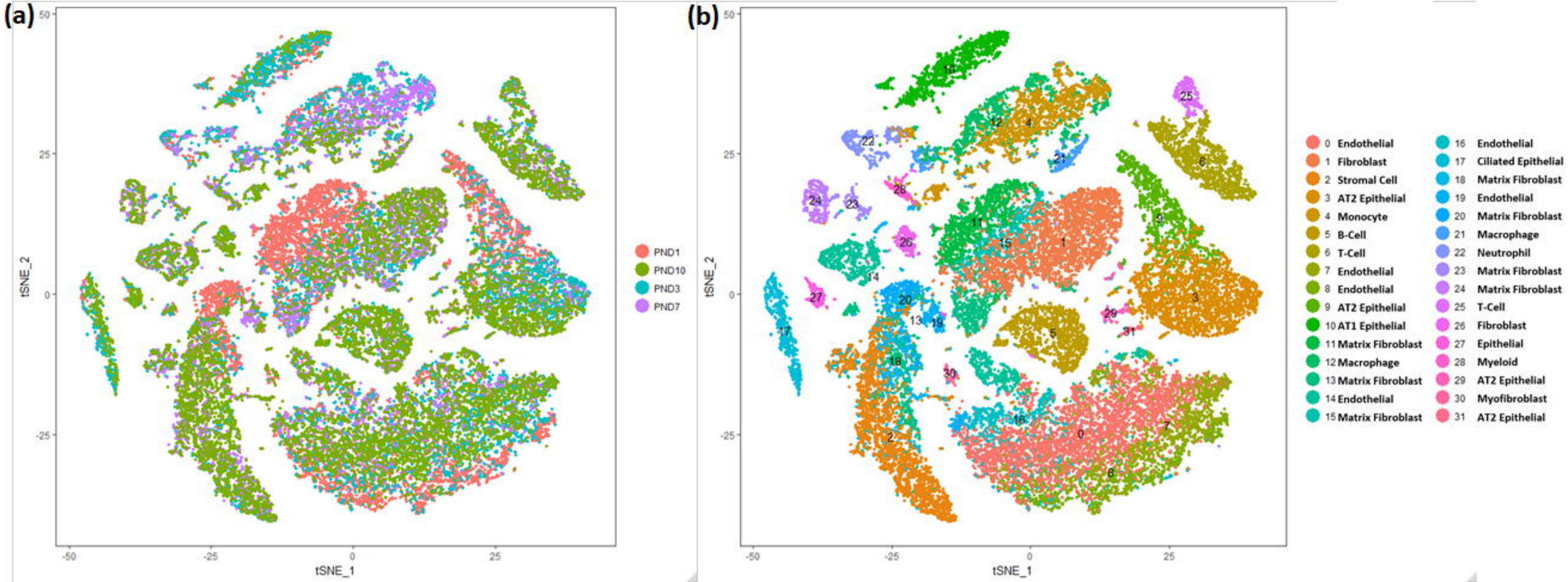
Normalized Counts
Low
High

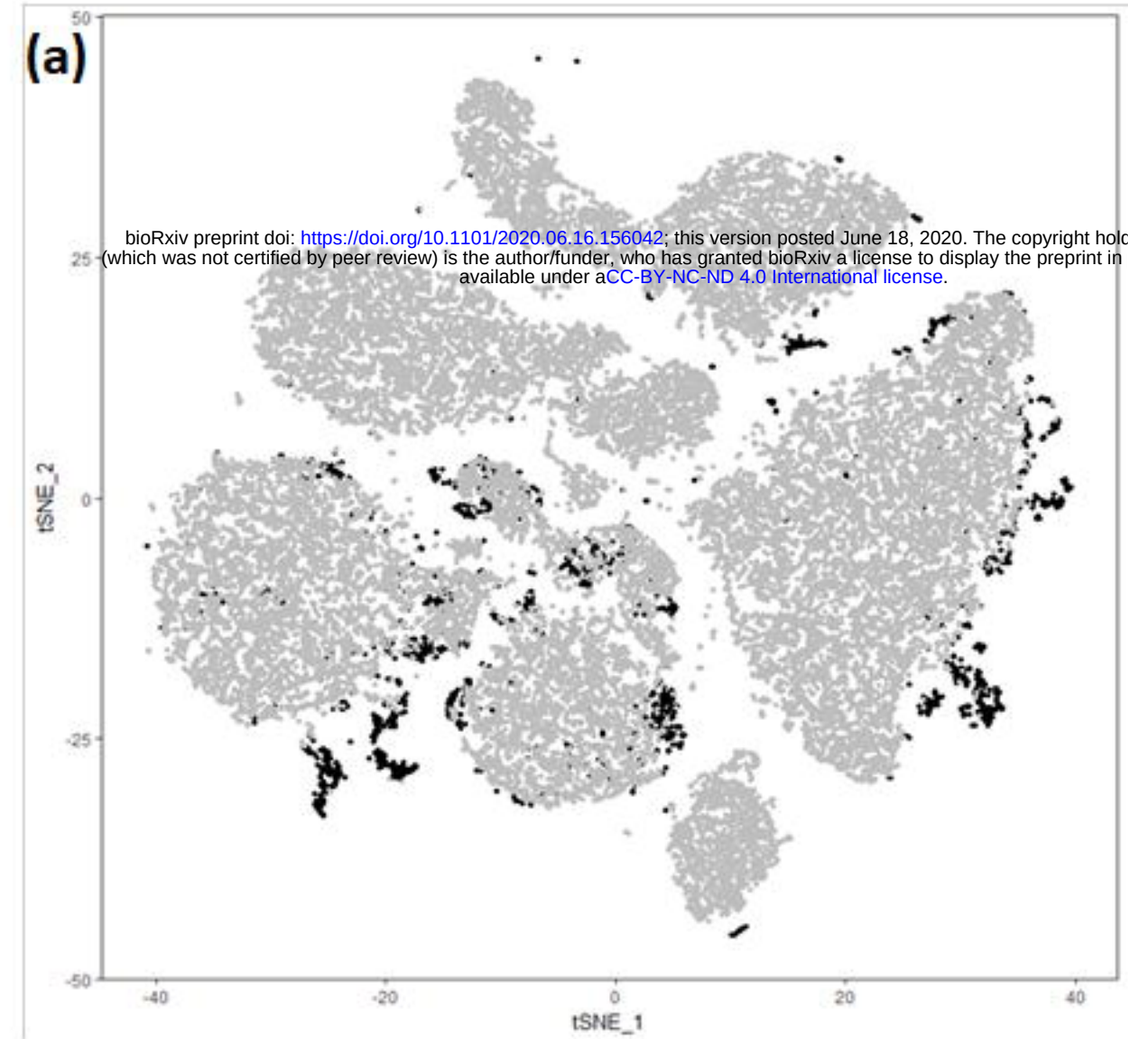
(b)

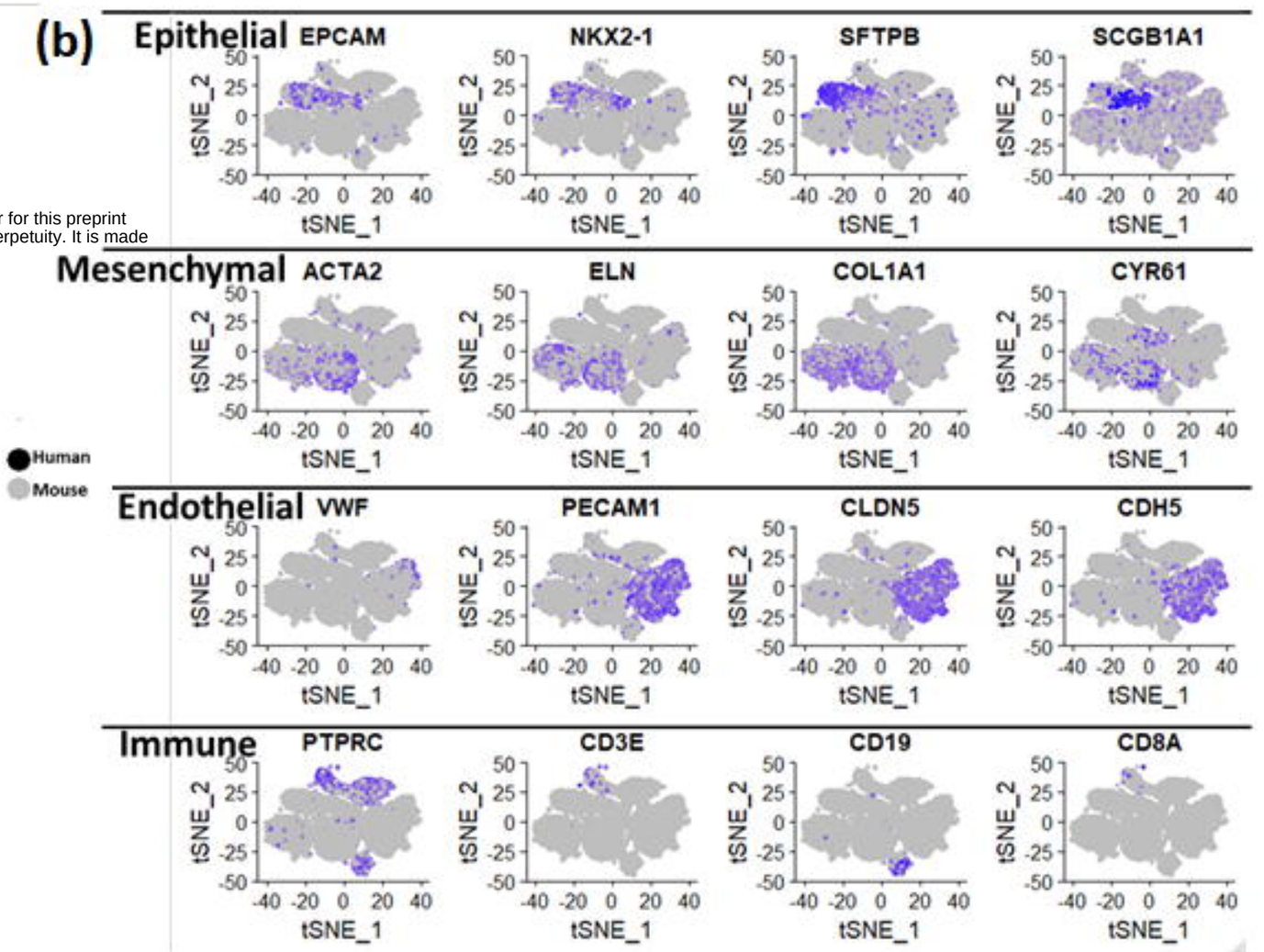
B Cells
T Cells
Pericyte
Smooth Muscle Cells
Pericyte
Stromal Cells
Endothelial
Matrix Fibroblasts
Myofibroblasts
Endothelial
AT1 Epithelial
Endothelial
Endothelial
Macrophage
AT2 Epithelial

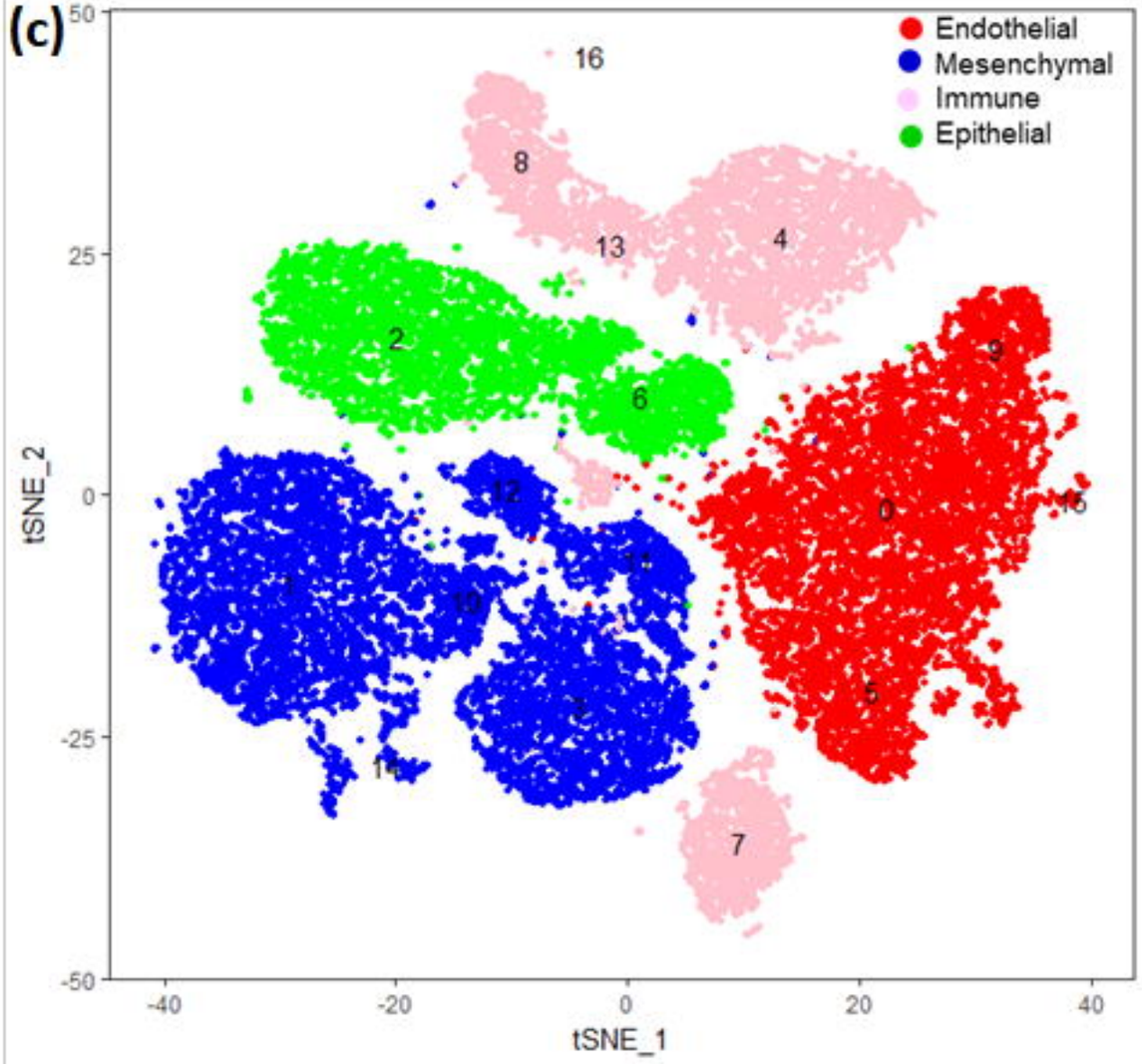
tSNE_2
40
20
0
-20
-40

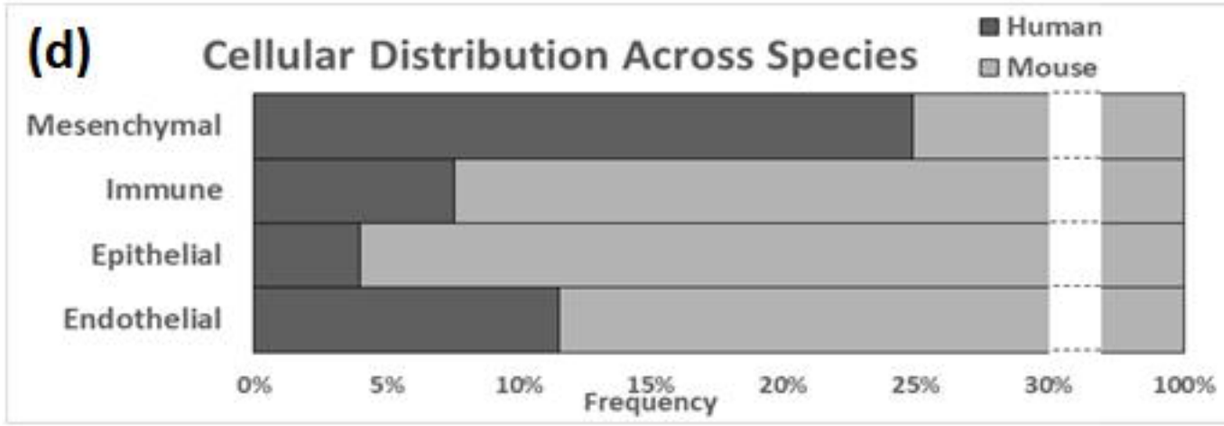
-25
0
25
tSNE_1


(a)
(b)
tSNE_1
tSNE_2
PND1
PND10
PND3
PND7
0 Endothelial
1 Fibroblast
2 Stromal Cell
3 AT2 Epithelial
4 Monocyte
5 B-Cell
6 T-Cell
7 Endothelial
8 Endothelial
9 AT2 Epithelial
10 AT1 Epithelial
11 Matrix Fibroblast
12 Macrophage
13 Matrix Fibroblast
14 Endothelial
15 Matrix Fibroblast
16 Endothelial
17 Ciliated Epithelial
18 Matrix Fibroblast
19 Endothelial
20 Matrix Fibroblast
21 Macrophage
22 Neutrophil
23 Matrix Fibroblast
24 Matrix Fibroblast
25 T-Cell
26 Fibroblast
27 Epithelial
28 Myeloid
29 AT2 Epithelial
30 Myofibroblast
31 AT2 Epithelial

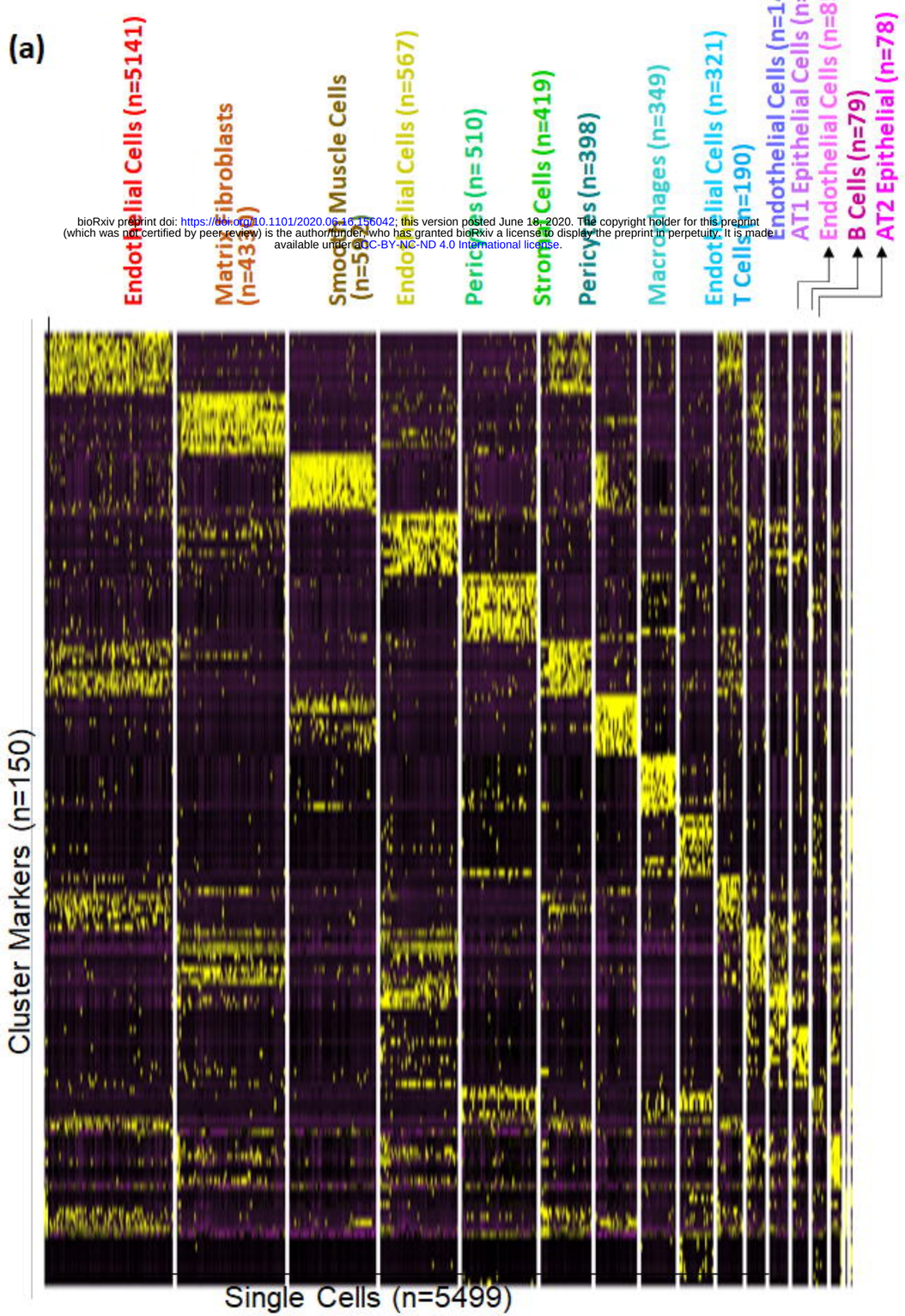








(a)



(b)