

From the: *Comprehensive Pneumology Center/Institute of Lung Biology and Disease,  
Helmholtz Center Munich*



Dissertation

to acquire the Doctor of Philosophy (Ph.D) title

from the Medical Faculty of the

Ludwig-Maximilians-University of Munich

***Single cell transcriptomic mapping of cell state identities in  
lung aging, injury and repair***

Submitted By:

**Ilias Angelidis**

from:

Manhattan, NYC, USA

Year:

2021

---

With the approval of the Medical Faculty of the  
Ludwig-Maximilians-University of Munich

**First Supervisor:**      *Prof. Dr. Jürgen Behr*

**Second Supervisor:**    *Dr. Herbert Schiller*

**Third Supervisor:**      *Prof. Dr. Heiko Adler*

**Fourth Supervisor:**    *Prof. Dr. Silke Meiners*

**Dean:**                      **Prof. Dr. med. dent. Reinhard Hickel**

Date of Defence:

***20/05/2021***

---

# Affidavit



**Affidavit**

Angelidis, Ilias

\_\_\_\_\_  
Surname, first name

\_\_\_\_\_  
Street

\_\_\_\_\_  
Zip code, town, country

I hereby declare, that the submitted thesis entitled:

Single cell transcriptomic mapping of cell state identities in lung aging, injury and repair  
.....

is my own work. I have only used the sources indicated and have not made unauthorised use of services of a third party. Where the work of others has been quoted or reproduced, the source is always given.

I further declare that the submitted thesis or parts thereof have not been presented as part of an examination degree to any other university.

Munich, 20/05/2021

\_\_\_\_\_  
place, date

Ilias Angelidis,

\_\_\_\_\_  
Signature of doctoral candidate

---

## Confirmation of congruency



LUDWIG-  
MAXIMILIANS-  
UNIVERSITÄT  
MÜNCHEN

Promotionsbüro  
Medizinische Fakultät



**Confirmation of congruency between printed and electronic version of  
the doctoral thesis**

Angelidis, Ilias

\_\_\_\_\_  
Surname, first name

\_\_\_\_\_  
Street

\_\_\_\_\_  
Zip code, town, country

I hereby declare, that the submitted thesis entitled:

Single cell transcriptomic mapping of cell state identities in lung aging, injury and repair

.....

is congruent with the printed version both in content and format.

Munich, 20/05/2021  
\_\_\_\_\_  
place, date

Ilias Angelidis,  
\_\_\_\_\_  
Signature of doctoral candidate



---

## Table of content

<b>Affidavit .....</b>	<b>3</b>
<b>Confirmation of congruency .....</b>	<b>4</b>
<b>Table of content.....</b>	<b>5</b>
<b>List of abbreviations .....</b>	<b>6</b>
<b>List of publications .....</b>	<b>7</b>
<b>Contribution to the publications.....</b>	<b>8</b>
1.1 Contribution to paper I .....	8
1.2 Contribution to paper II .....	8
<b>2. Introductory summary .....</b>	<b>9</b>
2.1 Lung Aging and Fibrosis .....	9
2.2 Lung Function .....	9
2.3 Epithelial Cells of the lung.....	10
2.4 Establishing Lung Fibrosis .....	13
2.5 The Bleomycin mouse model of pulmonary fibrosis .....	16
2.6 Aims of the study .....	16
<b>3. Paper I .....</b>	<b>19</b>
<b>4. Paper II .....</b>	<b>47</b>
<b>References .....</b>	<b>82</b>
<b>Acknowledgements.....</b>	<b>86</b>

---

## List of abbreviations

- **IPF:** Idiopathic Pulmonary Fibrosis
- **ECM:** Extracellular Matrix
- **TLC:** Total Lung Capacity
- **FEV1:** Forced Expiratory Volume in 1 second
- **FVC:** Forced Vital Capacity
- **SCGB1A1:** Secretoglobin Family 1A Member 1
- **AEC1:** Alveolar Epithelial Type 1 Cells
- **AEC2:** Alveolar Epithelial Type 2 Cells
- **PI3K:** AKT–phosphoinositide 3-kinase
- **TNF:** Tumor Necrosis Factor
- **PDGF:** Platelet-derived Growth Factor
- **CXCL12:** CXC Chemokine Ligand 12
- **CTGF:** Connective Tissue Growth Factor
- **TF:** Tissue Factor
- **PAI1:** Plasminogen Activator Inhibitor 1
- **FX:** Coagulation Factor X
- **$\alpha$ SMA:**  $\alpha$ -smooth Muscle Actin
- **EMT:** Epithelial-mesenchymal Transition
- **MET:** Mesenchymal-epithelial Transition
- **ILD:** Interstitial Lung Diseases
- **KRT8:** Keratin 8
- **ADI:** Alveolar Differentiation Intermediate (ADI)

---

## List of publications

1. Angelidis, I., Simon, L.M., Fernandez, I.E., Strunz, M., Mayr, C.H., Greiffo, F.R., Tsitsiridis, G., Ansari, M., Graf, E., Strom, T.M. and Nagendran, M., 2019. An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nature communications*, 10(1), pp.1-17.
2. Strunz, M., Simon, L.M., Ansari, M., Kathiriya, J.J., Angelidis, I., Mayr, C.H., Tsidiridis, G., Lange, M., Mattner, L.F., Yee, M. and Ogar, P., 2020. Alveolar regeneration through a Krt8+ transitional stem cell state that persists in human lung fibrosis. *Nature communications*, 11(1), pp.1-20.

---

# Contribution to the publications

## 1.1 Contribution to paper I

In the presenting work my contribution included:

Performing all single-cell transcriptomic experiments preparing, NGS libraries and analyzing data, performing histology and immunofluorescence microscopy as well as performing FACS experiments. Additionally I prepared along with my supervisor the manuscript and figures for this paper.

In particular:

- a) Fig.1: tSNE plot generation and cell type labeling.
- b) Fig.3: I performed all bulk RNA experiments and generated the data. Panel (e) 2D annotation and correlation of proteome and transcriptome. Panel (f and g) IF staining.
- c) Fig.4: Subclustering of airway epithelial cells and IF staining and quantifications.
- d) Fig.5: Volcano plot and dot plot for Col14a1 and Dcn
- e) Fig.7: Volcano plots and FACS experiments for sorting epithelium, endothelium and stromal cells including NGS library prep for flow sorted bulk samples.
- f) Fig. 8: IPA analysis, staining and quantification of lipids.
- g) Supp. Fig.1: Violin plots for nGenes and n UMIs per sample tSNE plots e and f
- h) Supp. Fig.4: Volcano plots for panels a, b, c

## 1.2 Contribution to paper II

In the presenting work my contribution included:

Instilling bleomycin and scoring experimental mice, performing lung harvest and single-cell transcriptomic experiments for all mice used for this paper including daily sampling and scRNAseq experiments on mice for the high resolution epithelial cell subset analysis. Assisting in preparing most NGS libraries and finally giving feedback and assistance throughout the development of the project and the preparation of the manuscript.

Figures and panels that have been a result of the experimental procedures where I was heavily involved include:

- ❖ Fig.1, Fig.3 (a-d), Fig.4 (a-f), Fig.5 (a-g), Fig.9 (c-e),
- ❖ Supp.Fig.1, Supp.Fig.3, Supp.Fig.4, Supp.Fig.7

---

## **2. Introductory summary**

### **2.1 Lung Aging and Fibrosis**

The overall increase in life expectancy in the last decades has greatly affected the proportions of elderly over younger individuals, with the former group being represented by over 700 million people worldwide<sup>1,2</sup>. As a consequence, the incidence of age related diseases such as cancer, atherosclerosis and idiopathic pulmonary fibrosis (IPF), have also dramatically increased. Although we still lack the knowledge and understanding of most pathogenic mechanisms that lead to IPF, there is clear evidence that multiple structural changes of the lung components that develop over time facilitate its occurrence. In fact, most of the hallmarks that govern lung aging have also been identified in IPF, such as altered intercellular communication, dysregulated extracellular matrix (ECM) deposition and cellular exhaustion<sup>3</sup>.

Indeed stem cell exhaustion that results in improper epithelial regeneration is a key trait that defines both aging and pulmonary fibrosis<sup>3</sup>. The cells of the respiratory tract gradually lose their primary function while the repair mechanisms that should be in place, to assure the resolution of these issues, are dysfunctioning. The investigation of the molecular alterations that occur during fibrosis and regeneration is of utmost importance in order to define novel therapeutics for severe non-reversible lung diseases such as IPF. The research described in this thesis aims to uncover novel hallmarks and previously undescribed features of aging and impaired regeneration. Current specific research questions in the field include: What are the molecular signals that drive fibrogenesis and repair? Why is healthy regeneration defected in aged individuals? How can we develop strategies to reverse these processes in an attempt to treat tomorrow's patients? The research described in this thesis was conducted with these questions in mind and have been addressed in variant degrees in this study.

### **2.2 Lung Function**

The respiratory tract's primary function is to deliver oxygen to the alveoli in order to reassure proper exchange of environmental oxygen with carbon dioxide from the blood circulation. Impaired oxygen exchange is thus the main characteristic of any lung related disorder and is a feature of the aged lung as well. Normal aging is defined by

---

homogeneous alveolar enlargement resulting in reduced lung elastic recoils. In addition to this, aging is associated with a decrease of the total lung capacity (TLC), forced expiratory volume in 1 s (FEV1) and forced vital capacity (FVC)<sup>4,5</sup>. Unsurprisingly, these alterations have been shown to be accelerated in smokers<sup>6</sup>. Susceptibility to disease can be a direct consequence of chronic physiological changes that occur during aging. In this regard and with respect to aging, age-related processes such as increased immunosenescence and inflamaging, a form of chronic, sterile, low-grade inflammation that develops with age, gradually diminish any remaining regenerative capacity of the lung and seem to play a key role in the development of lung fibrosis<sup>7,8</sup>. Pulmonary fibrosis also severely affects the lungs physiology, altering all parameters of respiratory equilibrium such as gas exchange and diffusion capacity<sup>9</sup>. The ability of the lung to expand is also greatly reduced in IPF as a result of pulmonary surfactant deregulation and ECM deposition. Studies have shown that in patients with IPF the lipid profile of the pulmonary surfactant is dramatically deregulated, resulting in a defective surface activity that is far from the one found in healthy individuals<sup>10</sup>.

### **2.3 Epithelial Cells of the lung**

There are two anatomical divisions of the respiratory tract. The nasal cavity, pharynx and larynx are all part of the upper respiratory tract. Meanwhile all other anatomic regions are part of the lower respiratory tract and include the conducting airways (trachea and bronchi), the small airways (bronchioles) and the respiratory zone (the alveoli). In the adult human the airways have a surface that can reach 70 m<sup>2</sup><sup>11</sup>. The groups and subsets of epithelial cells that span all the different anatomical sites of the respiratory tract vary both in composition and in structure and reflect their unique functions in each region. The epithelial cells of the lower respiratory tract constitute a highly effective barrier between the vital lung organ and potentially harmful environmental substances and function as the first line of defense against them<sup>12</sup>. Multiciliated cells, mucus-secreting goblet cells, club cells, neuroendocrine cells and basal cells, which secrete surfactants are the main cell types that build up the pseudostratified epithelium of the conducting airways. These cells are the initial cell types that interact with any externally inhaled particles and have developed abilities and traits that facilitate the effective mucociliary clearance of particles and microbes<sup>13</sup>. Multi-ciliated cells are equipped with cilia that reside on their apical surface and assist in transporting any intruding particles and mucus from the bronchi towards the

direction of the trachea and toward the respiratory exit. As for goblet and other secretory cells, their roles include trapping inhaled particulates and microorganisms to reassure their early neutralization. Club cells, in particular, account for one fifth of all airway cells and specifically secrete the anti-inflammatory protein secretoglobulin family 1A member 1 (SCGB1A1)<sup>14</sup>. Basal cells are able to self-renew in addition to differentiating into the cell populations that comprise the pseudostratified epithelium during homeostasis and after injury<sup>15</sup>. Stromal cells such as interstitial fibroblasts, that play a crucial role in the tissue's repair after injury, reside mainly in the tracheal region of the respiratory tract and in the larger airways<sup>16</sup>.

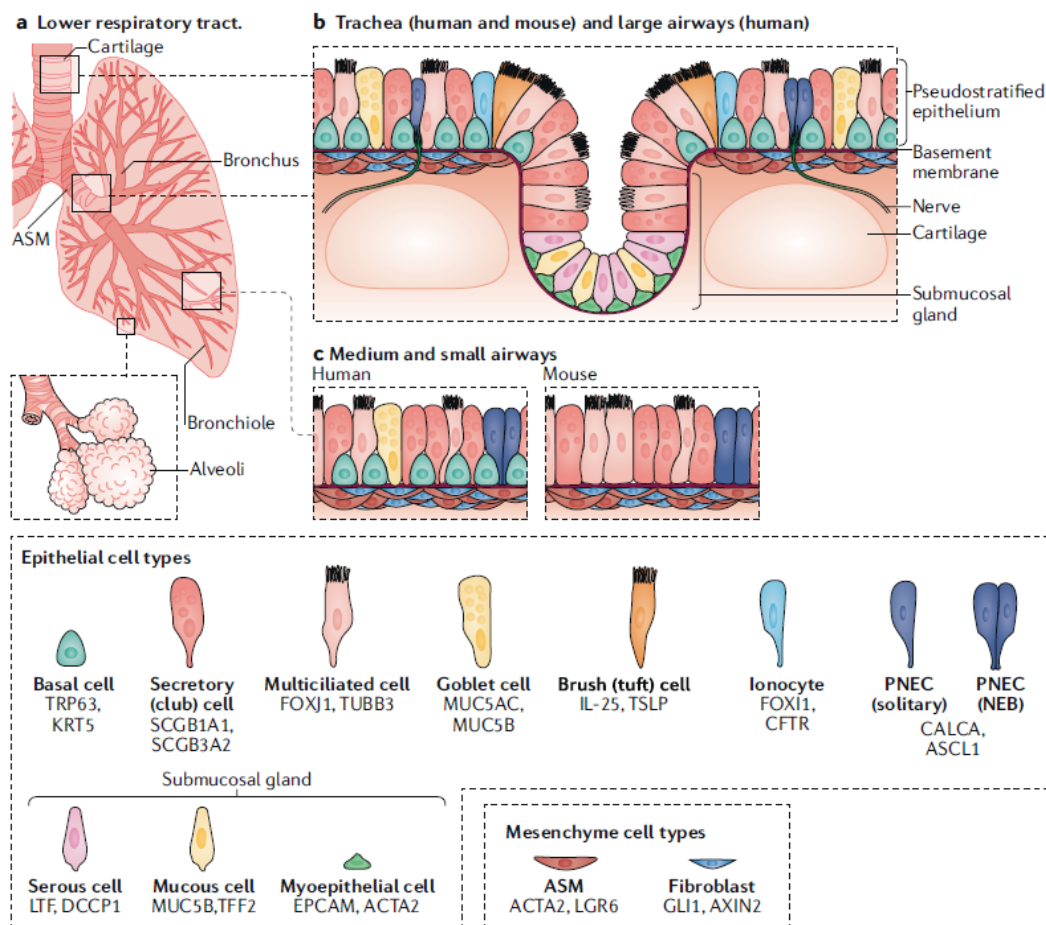


Figure 1 | **Cellular composition of the airways.** a) Illustration of the lower respiratory tract. The trachea is followed by a continuous branching of the larger and smaller airways that terminate in the alveoli regions. b,c) Both murine and human epithelium are composed of multiple cell types with distinct functions and markers. Illustration and text have been adapted from Zepp, J. A. & Morrissey, E. E, 2019<sup>17</sup>

In contrast to the aforementioned anatomical regions, the alveolar surfaces in the peripheral lung are lined by flat alveolar epithelial type 1 cells (AEC1). These cells are

specialized in gas exchange and form a continuous cell layer along the alveolar tract. Unlike the AEC1s, alveolar epithelial type 2 cells (AEC2) are morphologically cuboidal cells that have two main functions. They primarily secrete pulmonary surfactant reducing the surface tension of the alveoli during respiration, a function of paramount importance to avoid alveolar collapse. Additionally they act as the resident stem cells that differentiate into AEC1s during lung regeneration<sup>17</sup>.

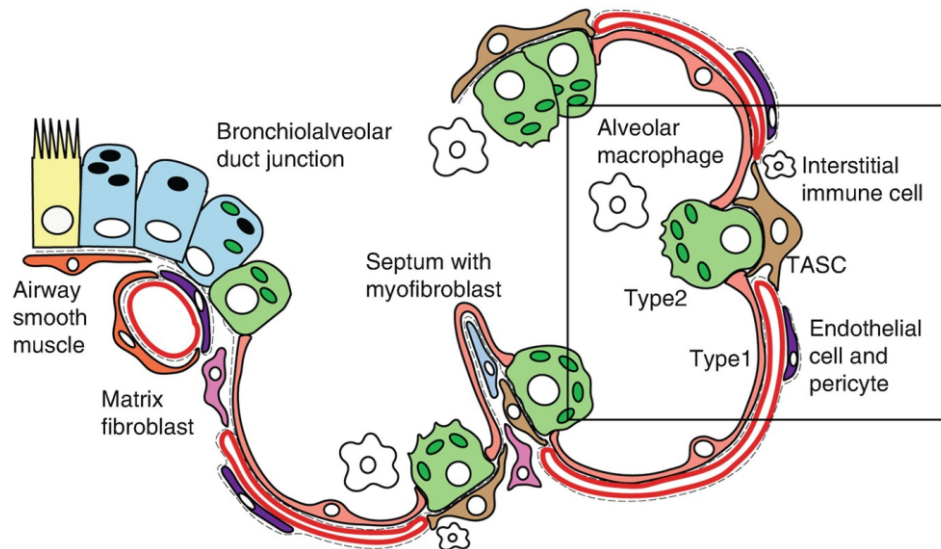


Figure 2 | **Illustration of the alveolar epithelium.** The stem cell niche of the alveoli include type 1 and type 2 alveolar epithelial cells as well as fibroblasts, pericytes, alveolar macrophages and various interstitial immune cells. Illustration and text have been adapted from Hogan B.L.M. 2020<sup>18</sup>

All the cells of the epithelial barrier have characteristic intercellular tight junctions that play a crucial role in maintaining the barrier's main trait of impermeability<sup>19</sup>. These tight junctions play a pivotal role in ensuring that the cells adhere together to form a regulated impermeable barrier. A vast number of interconnection proteins and receptors constitute the intracellular adhesion complex. Paracellular permeability is controlled by the tight junctions, while beneath those, the adherens junctions reside and play a crucial role in initiating differentiation and proliferation by mechanically connecting and disconnecting neighboring cells<sup>20</sup>. All these cellular junctions could malfunction as a result of improper cellular differentiation and contribute to the development of lung disorders that eventually may lead to immunopathology<sup>21-24</sup>.



---

## 2.4 Establishing Lung Fibrosis

Any damage on the epithelium results in the activation of cells that reside both in the vascular and interstitial compartments of the lungs, along with existing resident macrophages. As a consequence of this local activation mesenchymal cells start accumulating in site and reprogram to differentiate into myofibroblasts effectively establishing fibrotic lesions<sup>25</sup>. The accumulating myofibroblasts secrete ECM proteins in abnormal quantities that help in activating mesenchymal cells via mechanotransduction, escalating even further the fibrotic features of the tissue<sup>26,27</sup>. Fibrotic disorders such as IPF, have also been related to rare genetic disorders that increase the risk of the disease<sup>28</sup>. Genetic mutations in key regulatory genes such as SFTPC and SFTPA2<sup>29</sup> or transport genes such as ABCA<sup>30</sup>, lead to the development of chronic lung diseases and increase the risk of IPF.

Extensive research on IPF has proven that a dysfunctional alveolar epithelium, which is related to senescence, mutations and stress, plays a significant role in the injury and repair process that occurs both in sporadic and familial IPF<sup>31</sup>. Fibrotic scarring is believed to be driven by the increase of fibroblasts and myofibroblast populations within the region of the injury, further enhancing the destructive nature of the disease<sup>31,32</sup>. From recent single-cell transcriptomic studies, epithelial cells during the development and progression of IPF have been shown to develop specific phenotypes that are governed by the activation of many canonical pathways. These include AKT–phosphoinositide 3-kinase (PI3K), p53, HIPPO–YAP, TGF $\beta$ 1, and WNT<sup>33</sup>. Activated lung epithelial cells are capable of producing almost all the mediators that promote migration of many mesenchymal cells of diverse origins (such as resident fibroblasts and fibrocytes) in addition to enhancing myofibroblast differentiation. Myofibroblasts thereafter secrete increased amounts of ECM components, which mainly include fibrillar collagens, leading eventually to the destruction of the lungs architecture and function. In addition AEC2s, in IPF, express many proteins that promote profibrotic response. These include tumor necrosis factor (TNF), osteopontin, endothelin-1, platelet-derived growth factor (PDGF), CXC chemokine ligand 12 (CXCL12), connective tissue growth factor (CTGF) and TGF $\beta$ 1<sup>32,34–40</sup>. Interestingly, these features have been shown to gradually accumulate in the aging lung and seem to be developing also as a natural consequence of aging, giving multiple profibrotic characteristics to the aged lung as well<sup>41</sup>.

Of all these factors, ACE2-driven TGF $\beta$ 1 is most likely the strongest profibrotic mediator. It has been shown that the transduction of AECs with retrovirus that encode for activated TGF $\beta$ 1 resulted in the remodeling of lung explants with interstitial fibrosis accompanied by an increase in fibroblasts populations along with AEC2 hyperplasia and enlarged air space<sup>42</sup>. AEC2s are also able to activate the latent form of TGF $\beta$ 1 through surface expression of integrin  $\alpha\beta$ 6<sup>43,44</sup>. It has been observed that angiogenesis along with the expression of signaling pathways critical for maintaining wound healing procedures, can be inhibited by activated AEC2s. Indeed the expression of pigment epithelium-derived factor, inhibits angiogenesis and could be the reason for the characteristic lack of capillaries in fibrotic foci. Moreover, via secreting tissue factor (TF) and plasminogen activator inhibitor 1 (PAI1), AEC2s also influence the fibrin turnover<sup>45,46</sup>.

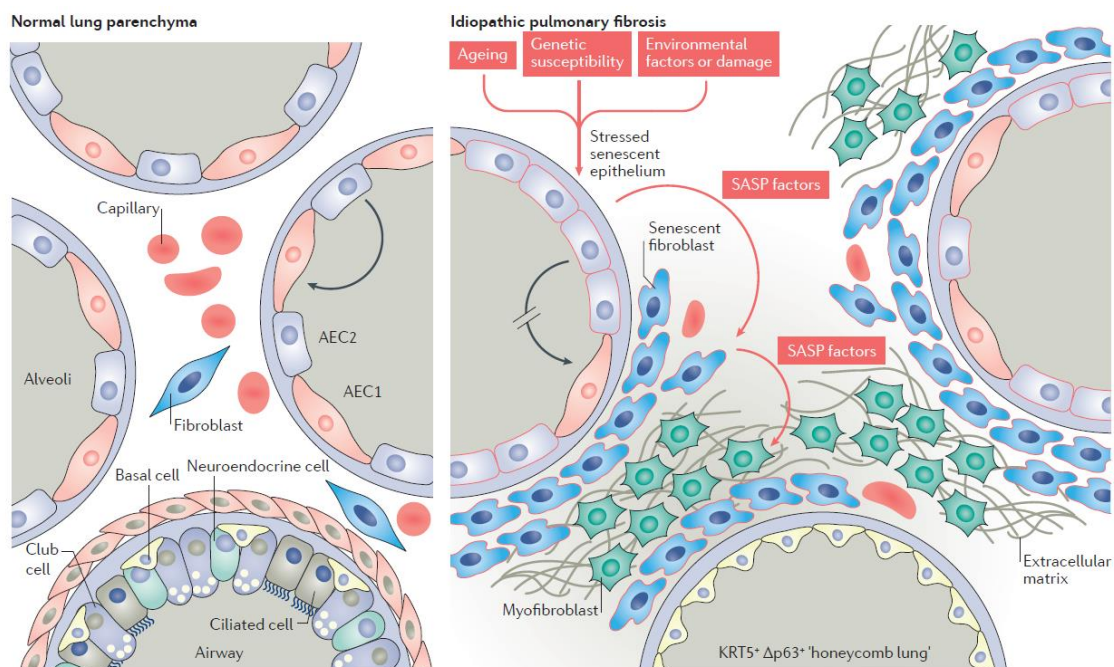


Figure 3 | **Model of idiopathic pulmonary fibrosis pathogenesis.** The lungs capacity to regenerate and respond to stress is linked to age-related perturbations and exposures along with genetic predispositions. In IPF, epithelial cells express multiple senescence and stress markers disrupting their normal function. At the same time, KRT5+  $\Delta$ p63+ epithelial cells are incapable of properly regenerating ACE2s, promoting the characteristic in IPF, ‘honeycomb’ formations. Illustration and text have been adapted from Mora et al. 2017<sup>47</sup>

In addition to the aforementioned functions, the alveolar epithelium also expresses coagulation factor X (FX), a key proteinase of the coagulation pathway<sup>48</sup>. FXa enhances

---

fibroblasts differentiation into myofibroblasts via a TGF $\beta$ 1-dependent mechanism of  $\alpha$ -smooth muscle actin ( $\alpha$ SMA) induction. AEC2s from fibrotic lung regions have been also shown to express markers of apoptosis and senescence effectively neutralizing these cells throughout the disease. It is thus clear that the approach of replacing AEC2s with functional counterparts is essential for tissue repair. Although the source of the cells that promote re-epithelization and repair is uncertain, novel experimental procedures have emphasized on the emergence of a KRT5+ KRT14+  $\Delta$ p63+ (a p63 splice variant) epithelial population in disease that is usually not present in healthy human lungs, as a potential source.

KRT5+  $\Delta$ p63+ cells have notch signaling pathways constantly activated which adds to the overall increase in honeycomb structures and simultaneous decrease in the regenerative capacity of the alveoli<sup>49</sup>. These findings indicate that stressed AEC2s in IPF are incapable of lung regeneration and that an additional independent pool of progenitor cells moderate the general repair response and facilitates fibrosis and honeycomb structure generations<sup>33,49</sup>. Immunohistochemical evidence has shown that some epithelial cells can undergo epithelial-mesenchymal transition (EMT)<sup>49,50</sup> and might add this way to the increased mesenchymal features of IPF. Nevertheless, it is still unclear to what extent this EMT-like process contributes to the development of fibrosis since it is not well documented if epithelial cells can acquire enough mesenchymal features to be classified as fibroblasts<sup>51</sup>. EMT and its reverse process, mesenchymal-epithelial transition (MET), are not black and white cell states but rather include many intermediate phases with partial EMT phenotypes<sup>52</sup>.

A key characteristic of EMT is the disruption of all tight junctions between adjacent epithelial cells. This is an important feature since it enables programmed migration and generates the epithelial secretome profile found in IPF. The main path to remodeling the lungs disrupted architecture in IPF is through the regeneration of an intact alveolar epithelium. Unfortunately our current knowledge of lung epithelial regeneration is incomplete since most such studies have been conducted on wound healing experiments in skin injury models. It is clear that there is a need to enrich our understanding of the molecular mechanisms of lung re-epithelization using various, lung specific injury models.

---

## 2.5 The Bleomycin mouse model of pulmonary fibrosis

The most commonly used mouse model for studying IPF and other interstitial lung diseases (ILDs) is the bleomycin mouse model in C57Bl6/J mice<sup>53</sup>. Bleomycin administration to murine lungs induces DNA strand breaks, leading to alveolar epithelial cell death followed by tissue injury along with an increasing inflammatory response and a gradual establishment of fibrogenesis by fibroblast activation and collagen deposition. These events are subsequently followed by tissue repair and almost complete regeneration of the destroyed tissue, enabling the study of both the injury and repair of the lungs in the context of fibrosis and regeneration<sup>54</sup>. Single dose administration of bleomycin leads to transient pulmonary fibrosis and reproduces several features of human ILD, defining this model to be very suitable for the study of both the development of fibrosis as well as its resolution over time<sup>55</sup>.

## 2.6 Aims of the study

**Publication I:** In order to study age related alterations at cell type level we employed microfluidic based single-cell RNA transcriptomics on whole lung homogenates of young (3 months old) and aged (24 months old) mice. These data were coupled with bulk proteomic and transcriptomic data in an attempt to create an atlas of the aged murine lung that represents a reference tool available for researchers of lung biology and aging. We examined gene expression patterns for over 30 cell types and identified changes with regard to aging in all cell types. Furthermore, we showed that aging is accompanied by an increase in transcriptional noise, a phenomena that had very recently been identified in human<sup>56,57</sup>, hinting that this feature could in fact be a global hallmark of aging that possibly affects most cell types in both humans and mice. **Publication I** in addition to all the findings that refer to the known hallmarks of aging, highlights a rather novel alteration of the stem cells of the alveolar epithelial barrier that had not been previously extensively studied. Our observations of increased cholesterol biosynthesis and high neutral lipid content in AEC2s of aged mice may refer to previous studies showing the lipid composition of the pulmonary surfactant being changed with age<sup>58</sup>. In our study we report a strong correlation between the aged AEC2 phenotype and that of Insig1/2 knockout mice that accumulate neutral lipids, resulting in lipotoxicity-driven lung inflammation<sup>59</sup>. This indicated that, to some extent, the observed chronic inflammation in the aged lung could be

---

influenced by deregulated lipid homeostasis. In addition, other hallmarks of aging such as epithelial senescence may be contributing to the inflammatory phenotype observed, since AEC2 specific deletion of telomerase in mice has been shown to promote the development of a pro-inflammatory tissue microenvironment, resulting to a decreased ability of resolving any acute lung injury<sup>60</sup>. Finally, up to date most studies of lung aging have focused primarily on accumulating fibroblasts and impaired alveolar epithelium. Through our unbiased approach we have revealed compositional changes in other respiratory compartments discovering a clear alteration of the airway epithelial cells during aging which results in an overall increase in the number of ciliated cells aligning the airways.

**Publication II:** In this study we attempted to investigate the kinetics of murine lung regeneration at the single cell level using the bleomycin murine injury model. We resolved the gene expression changes within 28 cell types focusing primarily on the alveolar and airway epithelium. Our work describes how airway and alveolar stem cells initially converge onto a novel Krt8+ transitional stem cell state, termed Krt8+ alveolar differentiation intermediate (ADI), prior to initiating proper regeneration of AEC1s. We identified these Krt8+ ADI cells in multiple mouse lung injury models as well as in human lung fibrosis. The Krt8+ ADI cells feature hallmarks of EMT, senescence and p53 activation and comprise a transient cell state governed by a squamous cell phenotype. In our attempt to describe the evolution of these cells we followed up our experiments with a daily sampling of lung tissue after bleomycin induced injury. Our research describes the gene expression profile of Krt8+ ADI cells throughout the development and resolution of fibrosis and identifies a transcriptional convergence of AEC2s and MHCII+ club cells towards the newly described Krt8+ ADI state. Finally we show that lung repair through terminal AEC1 differentiation of Krt8+ ADI cells occurs after injury and is paramount to sustaining epithelial regeneration. Our data indicate that the transcriptional signature of KRT5<sup>-</sup>/KRT17<sup>+</sup> basaloid cells<sup>50</sup> in IPF tissues resembles that of Krt8+ ADI described in **Publication II**. Driven by our findings we propose that chronic lung disease and in particular in IPF may be a result of a dysfunctional molecular differentiation checkpoint that results in a persistent intermediate regenerative cell state that should have otherwise been transient.

---

**Publications I and II** are the result of a greater attempt to describe the cellular architecture of the respiratory tract in normal health and in disease. These two studies are complementary to each other as they both similarly add to the greater pool of knowledge in the context of aging, injury repair and regeneration. Our research addressed many questions that existed in the field of lung regeneration and injury and highlights novel aspects of aging and IPF. Future research aiming to follow up the context of this thesis includes the study of injury repair and regeneration in the aged lung. Ongoing experiments that have been initiated in our lab will tackle these questions in detail. Administering bleomycin to aged mice will generate a directly comparable dataset that will clarify, to a previously unprecedented level, if and how aging leads to impaired regeneration of the alveolar epithelium, and which cellular processes are involved.

---

### 3. Paper I






Angelidis, I., Simon, L.M., Fernandez, I.E., Strunz, M., Mayr, C.H., Greiffo, F.R., Tsitsiridis, G., Ansari, M., Graf, E., Strom, T.M. and Nagendran, M., 2019. An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nature communications*, 10(1), pp.1-17.

ARTICLE

<https://doi.org/10.1038/s41467-019-08831-9>

OPEN

# An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics

Ilias Angelidis<sup>1</sup>, Lukas M. Simon <sup>2</sup>, Isis E. Fernandez<sup>1</sup>, Maximilian Strunz<sup>1</sup>, Christoph H. Mayr<sup>1</sup>, Flavia R. Greiffo<sup>1</sup>, George Tsitsiridis<sup>2</sup>, Meshal Ansari<sup>1,2</sup>, Elisabeth Graf<sup>3</sup>, Tim-Matthias Strom<sup>3</sup>, Monica Nagendran<sup>4</sup>, Tushar Desai <sup>4</sup>, Oliver Eickelberg <sup>5</sup>, Matthias Mann <sup>6</sup>, Fabian J. Theis <sup>2,7</sup> & Herbert B. Schiller<sup>1</sup>

Aging promotes lung function decline and susceptibility to chronic lung diseases, which are the third leading cause of death worldwide. Here, we use single cell transcriptomics and mass spectrometry-based proteomics to quantify changes in cellular activity states across 30 cell types and chart the lung proteome of young and old mice. We show that aging leads to increased transcriptional noise, indicating deregulated epigenetic control. We observe cell type-specific effects of aging, uncovering increased cholesterol biosynthesis in type-2 pneumocytes and lipofibroblasts and altered relative frequency of airway epithelial cells as hallmarks of lung aging. Proteomic profiling reveals extracellular matrix remodeling in old mice, including increased collagen IV and XVI and decreased Fraser syndrome complex proteins and collagen XIV. Computational integration of the aging proteome with the single cell transcriptomes predicts the cellular source of regulated proteins and creates an unbiased reference map of the aging lung.

<sup>1</sup>Helmholtz Zentrum München, Institute of Lung Biology and Disease, Member of the German Center for Lung Research (DZL), Munich 85764, Germany. <sup>2</sup>Helmholtz Zentrum München, Institute of Computational Biology, Munich 85764, Germany. <sup>3</sup>Helmholtz Zentrum München, Institute of Human Genetics, Munich 85764, Germany. <sup>4</sup>Department of Internal Medicine, Division of Pulmonary and Critical Care, Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Stanford 94305 CA, USA. <sup>5</sup>Department of Medicine, Division of Respiratory Sciences and Critical Care Medicine, University of Colorado, Aurora 80045 CO, USA. <sup>6</sup>Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Munich 82152, Germany. <sup>7</sup>Department of Mathematics, Technische Universität München, Munich 85748, Germany. These authors contributed equally: Ilias Angelidis, Lukas M. Simon. Correspondence and requests for materials should be addressed to F.J.T. (email: [fabian.theis@helmholtz-muenchen.de](mailto:fabian.theis@helmholtz-muenchen.de)) or to H.B.S. (email: [herbert.schiller@helmholtz-muenchen.de](mailto:herbert.schiller@helmholtz-muenchen.de))



The intricate structure of the lung enables gas exchange between inhaled air and circulating blood. As the organ with the largest surface area (~70 m<sup>2</sup> in humans), the lung is constantly exposed to a plethora of environmental insults. A range of protection mechanisms are in place, including a highly specialized set of lung-resident innate and adaptive immune cells that fight off infection, as well as several stem and progenitor cell populations that provide the lung with a remarkable regenerative capacity upon injury<sup>1</sup>. These protection mechanisms seem to deteriorate with advanced age, since aging is the main risk factor for developing chronic lung diseases, including chronic obstructive pulmonary disease (COPD), lung cancer, and interstitial lung disease<sup>2,3</sup>. Advanced age causes a progressive impairment of lung function even in otherwise healthy individuals, featuring structural and immunological alterations that affect gas exchange and susceptibility to disease<sup>4</sup>. Aging decreases ciliary beat frequency in mice, thereby decreasing mucociliary clearance and partially explaining the predisposition of the elderly to pneumonia<sup>5</sup>. Senescence of the immune system in the elderly has been linked to a phenomenon called ‘inflammaging’, which refers to elevated levels of tissue and circulating pro-inflammatory cytokines in the absence of an immunological threat<sup>6</sup>. Several previous studies analyzing the effect of aging on pulmonary immunity point to age-dependent changes of the immune repertoire as well as activity and recruitment of immune cells upon infection and injury<sup>4</sup>. Vulnerability to oxidative stress, pathological nitric oxide signaling, and deficient recruitment of endothelial stem cell precursors have been described for the aged pulmonary vasculature<sup>7</sup>. The extracellular matrix (ECM) of old lungs features changes in tensile strength and elasticity, which were discussed to be a possible consequence of fibroblast senescence<sup>8</sup>. Using atomic force microscopy, age-related increases in stiffness of parenchymal and vessel compartments were demonstrated recently<sup>9</sup>; however, the causal molecular changes underlying these effects are unknown.

Aging is a multifactorial process that leads to these molecular and cellular changes in a complicated series of events. The hallmarks of aging encompass cell-intrinsic effects, such as genomic instability, telomere attrition, epigenetic alterations, loss of proteostasis, deregulated nutrient sensing, mitochondrial dysfunction, and senescence, as well as cell-extrinsic effects, such as altered intercellular communication and extracellular matrix remodeling<sup>2,3</sup>. The lung contains potentially at least 40 distinct cell types<sup>10</sup>, and specific effects of age on cell-type level have never been systematically analyzed.

In this study, we build on rapid progress in single-cell transcriptomics<sup>11,12</sup> which recently enabled the generation of a first cell-type resolved census of murine lungs<sup>13</sup>, serving as a starting point for investigating the lung in distinct biological conditions as shown for lung aging in the present work. We computationally integrate single-cell signatures of aging with state-of-the-art whole lung RNA-sequencing (RNA-seq) and mass spectrometry-driven proteomics<sup>14</sup> to generate a multi-omics whole organ resource of aging-associated molecular and cellular alterations in the lung.

## Results

**Lung aging atlas reveals deregulated transcriptional control.** To generate a cell-type resolved map of lung aging we performed highly parallel genome-wide expression profiling of individual cells using the Dropseq workflow<sup>15</sup> which uses both molecule and cell-specific barcoding, enabling great cost efficiency and accurate quantification of transcripts without amplification bias<sup>16</sup>. Single-cell suspensions of whole lungs were generated from 3-month-old mice ( $n = 8$ ) and 24-month-old mice ( $n = 7$ ). After quality

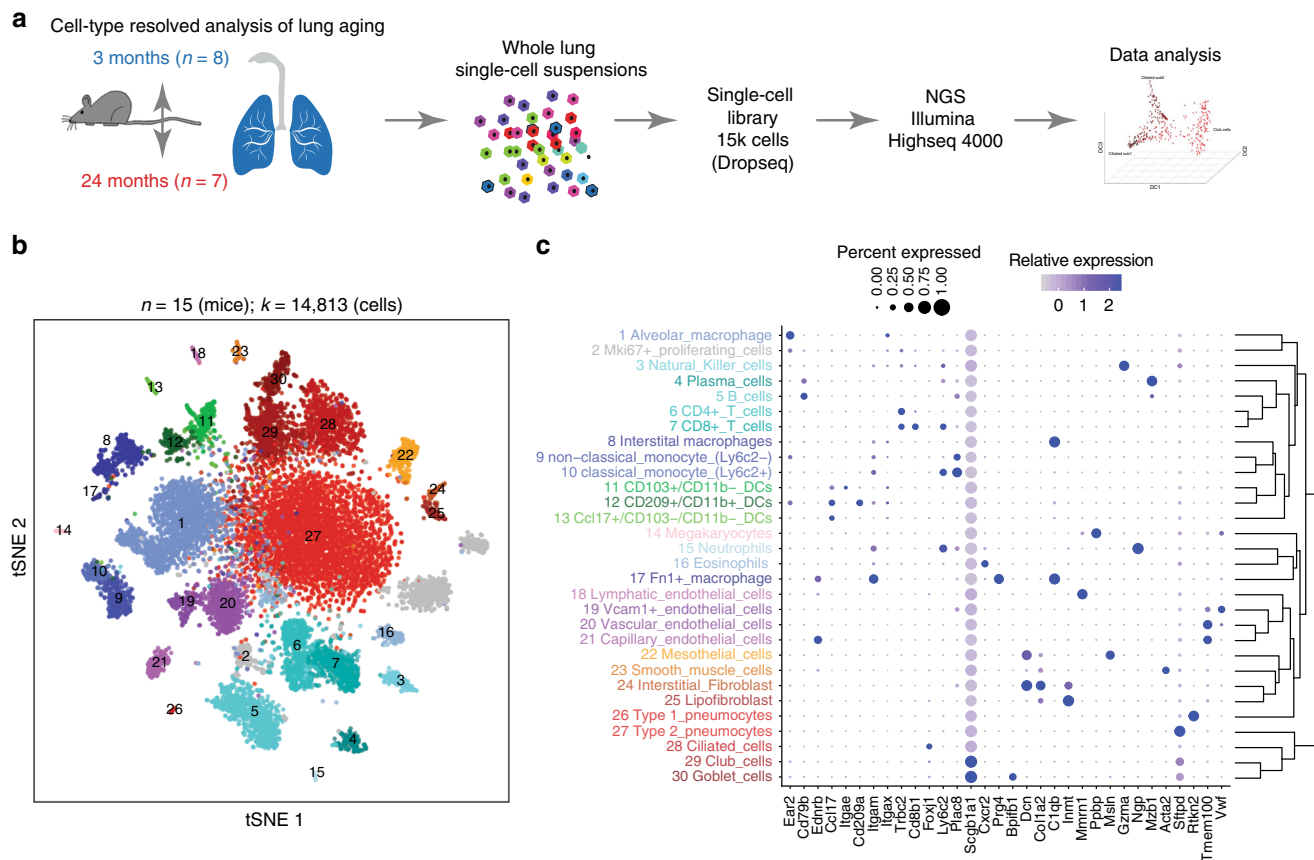
control, a total of 14,813 cells (7672 young, 7141 old) were used for downstream analysis (Fig. 1a). Quality metrics including number of unique molecular identifiers (UMI), genes detected per cell, and reads aligned to the mouse genome were comparable across mice (Supplementary Fig. 1a–c). To ensure that cell-type discovery is not confounded by aging effects, we only used highly variable genes between cell types (see Methods for details). Unsupervised clustering analysis revealed 36 distinct clusters corresponding to 30 cell types, including all major known epithelial, mesenchymal, and leukocyte lineages (Fig. 1b, c). We observed very good overlap across mouse samples (Silhouette coefficient:  $-0.074$ ) and most clusters were derived from >70% of the mice of both age groups (Supplementary Fig. 1d and e). The definition of cell types (clusters in  $t$ -distributed stochastic neighbor embedding (tSNE) map) was very comparable between old and young mice, indicating that the cell-type identity was not strongly confounded by the aging effects (Supplementary Fig. 1f). Two clusters exclusively contained cells from a single mouse and were removed from downstream analysis. Interestingly, we identified even rare (<1%, 43 cells) cell types such as megakaryocytes, which were recently identified as an unexpected tissue-resident cell type in mouse lung<sup>17</sup>. Of note, some samples contributed as little as a single cell to this megakaryocyte cluster, emphasizing the power and accuracy of the computational workflow used here for data integration from multiple mice.

We used differential gene expression analysis to determine cell type-specific marker genes with highly different levels between clusters (Fig. 1c, Supplementary Data 1). The clusters were annotated with assumed cell-type identities based on (1) known marker genes derived from expert annotation in literature and (2) enrichment analysis using Fisher’s exact test of gene expression signatures of isolated cell types from databases including ImmGen<sup>18</sup> and xCell<sup>19</sup>. Correlation analysis of marker gene signatures revealed that similar cell types clustered together, implying correct cell-type annotation (Fig. 1c).

We used the matchScore tool<sup>20</sup> to compare the cluster identities of our dataset with the lung data in the recently published Mouse Cell Atlas (MCA)<sup>13</sup>, and found very good agreement in cluster identities and annotations (Supplementary Fig. 2a). Moreover, when comparing our cluster identities to the MCA peripheral blood data, only weak correspondence was observed (Supplementary Fig. 2b), which was similar in the MCA peripheral blood versus MCA lung comparison (Supplementary Fig. 2c). One notable exception in this comparison is the cluster of red blood cells in our dataset which achieved high correspondence with the MCA peripheral blood cluster annotated as Erythroblast\_Hbb-a2\_high. The red blood cells serve as a control and illustrate matchSC values for a correct overlap (Supplementary Fig. 2d). Taken together, these findings indicate that very little blood-derived contamination was present.

Additionally, we noticed one cluster of mainly proliferating cells showing high expression levels for S and G2M cell-cycle marker genes (Supplementary Fig. 3a and b). Young mice showed a higher fraction of cells in this cluster compared to old mice (Supplementary Fig. 3c; Generalized linear binomial model,  $p < 0.001$ ). Next, we isolated this cluster and corrected the gene expression levels for cell-cycle phase (Supplementary Fig. 3 d and e). Subsequent unsupervised clustering analysis revealed that these proliferating cells belong to T cells, type-2 pneumocytes, and alveolar macrophages (Supplementary Fig. 3f–i).

It was suggested that aging is a consequence of increased transcriptional instability rather than the result of a coordinated transcriptional program, and that an aging-associated increase in transcriptional noise can lead to fate drifts and ambiguous cell-type identities<sup>21,22</sup>. Therefore, we quantified transcriptional noise following previous work<sup>22</sup> and accounted for differences in total



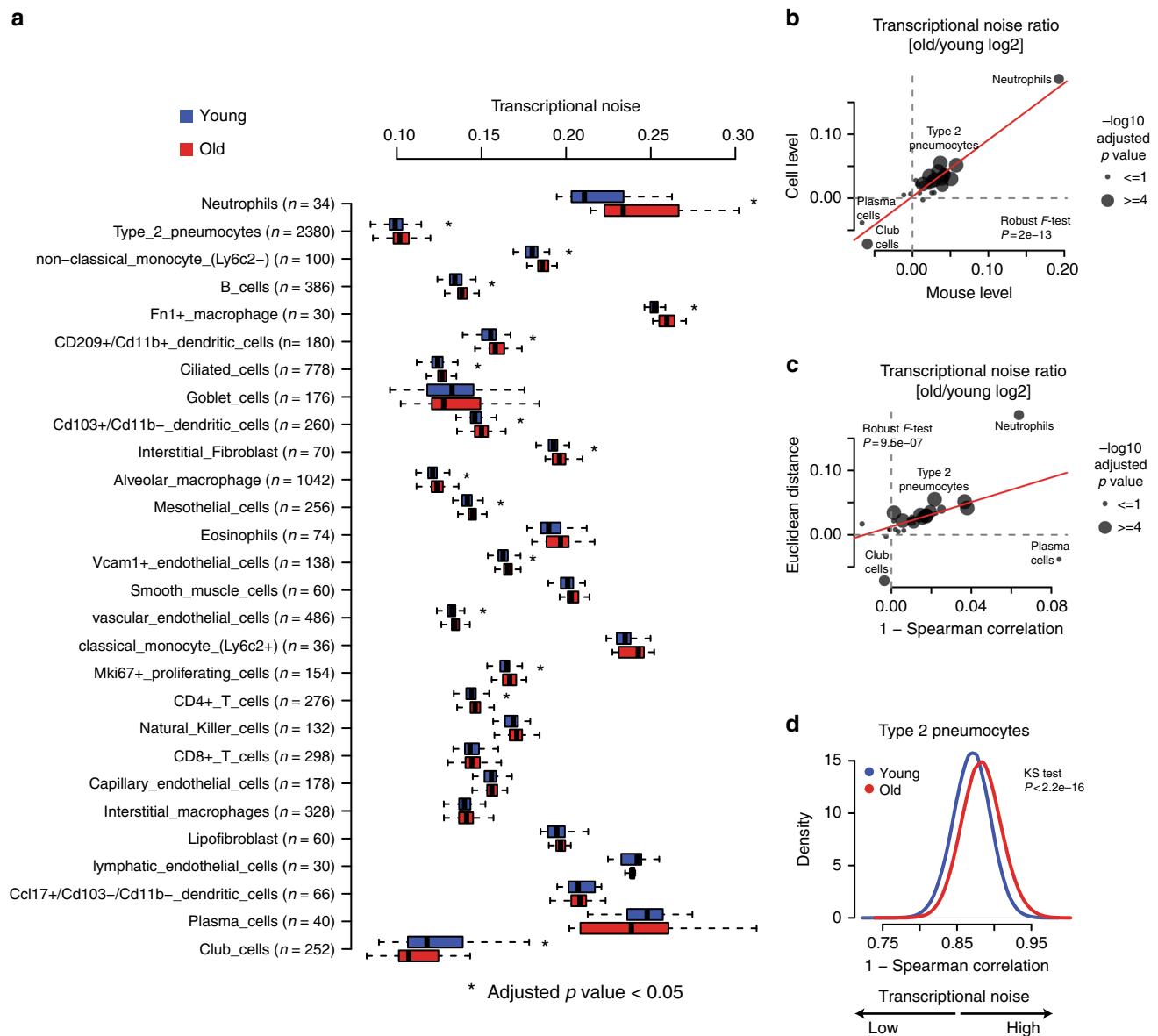
**Fig. 1** A single-cell atlas of mouse lung reveals major cell-type identities. **a** Experimental design—whole lung single-cell suspensions of young and old mice were analyzed using the Dropseq workflow. **b** The *t*-distributed stochastic neighbor embedding (tSNE) visualization shows unsupervised transcriptome clustering, revealing 30 distinct cellular identities. **c** The dotplot shows (1) the percentage of cells expressing the respective selected marker gene using dot size and (2) the average expression level of that gene based on unique molecular identifier (UMI) counts. Rows represent hierarchically clustered cell types, demonstrating similarities of transcriptional profiles

UMI counts and cell-type frequencies (see Methods for details). We observed an increase in transcriptional noise with aging in most cell types (Fig. 2a). To further exclude technical confounding we additionally averaged the transcriptional noise scores per mouse and obtained highly concordant results (Fig. 2b). To further substantiate this finding we quantified transcriptional noise in an alternative manner using Spearman's correlations between cells. This analysis confirmed our finding that transcriptional noise is increased with aging (Fig. 2c, d) and is in line with previous reports in the human pancreas<sup>22</sup> or mouse CD4+ T cells<sup>21</sup>.

**Multi-omics data integration of mRNA and protein.** To validate the completeness of our single-cell RNA-sequencing (scRNA-seq) data and capture age-dependent alterations in both mRNA and protein content for the whole lung, we generated two additional cohorts of young and old mice (Fig. 3a, Supplementary Figure 4 and Supplementary Data 2): (1) bulk RNA-seq data of three replicates of young (3 months) and old mice (22 months) and (2) state-of-the-art shotgun proteomics data of four replicates of young (3 months) and old mice (24 months). To compare the whole lung bulk transcriptome with single-cell data we generated in silico bulk samples from the scRNA-seq data by summing expression counts from all cells for each mouse individually (Supplementary Data 2). Differential gene expression analysis from in silico bulks and real whole lung bulk sequencing revealed a total of 2362 and 9245 differentially expressed genes (negative

binomial generalized linear model, false discovery rate (FDR) <10%) between the two age groups, respectively (Supplementary Fig. 4a, b, Supplementary Data 2). From whole lung tissue we quantified 5212 proteins across conditions and found 213 proteins to be significantly regulated with age (two-sided *t*-test, FDR < 10%, Supplementary Fig. 4c, Supplementary Data 2). We observed very good agreement between the real and in silico bulk data, thus excluding strong biases by the single-cell isolation procedures (Fig. 3b). Furthermore, we also observed strong correspondence between the age-dependent alterations in all three data sets (Fig. 3c), indicating that we were able to pick up robust age-dependent changes with three independent experimental settings. Significant correlation was observed between the gene-level fold changes derived from RNA-seq, scRNA-seq, and protein expression data (Supplementary Fig. 4d–f).

Prediction of the upstream regulators<sup>23</sup> of the observed expression changes in either the transcriptome or proteome data gave very similar results (Fig. 2d). In both datasets from independent mouse cohorts, we discovered a pro-inflammatory signature, which included upregulation of *Il6*, *Il1b*, *Tnf*, and *Ifng*, as well as the downregulation of *Pparg* and *Il10* (Fig. 2d). Furthermore, to reveal common or distinct regulation of gene annotation categories in the transcriptome or proteome, we performed a two-dimensional annotation enrichment analysis<sup>24</sup> (Supplementary Data 3). Again, most gene categories regulated by age were showing the same direction in transcriptome and proteome so that the positive Pearson's correlation of the annotation enrichment scores was highly significant (Fig. 2e).

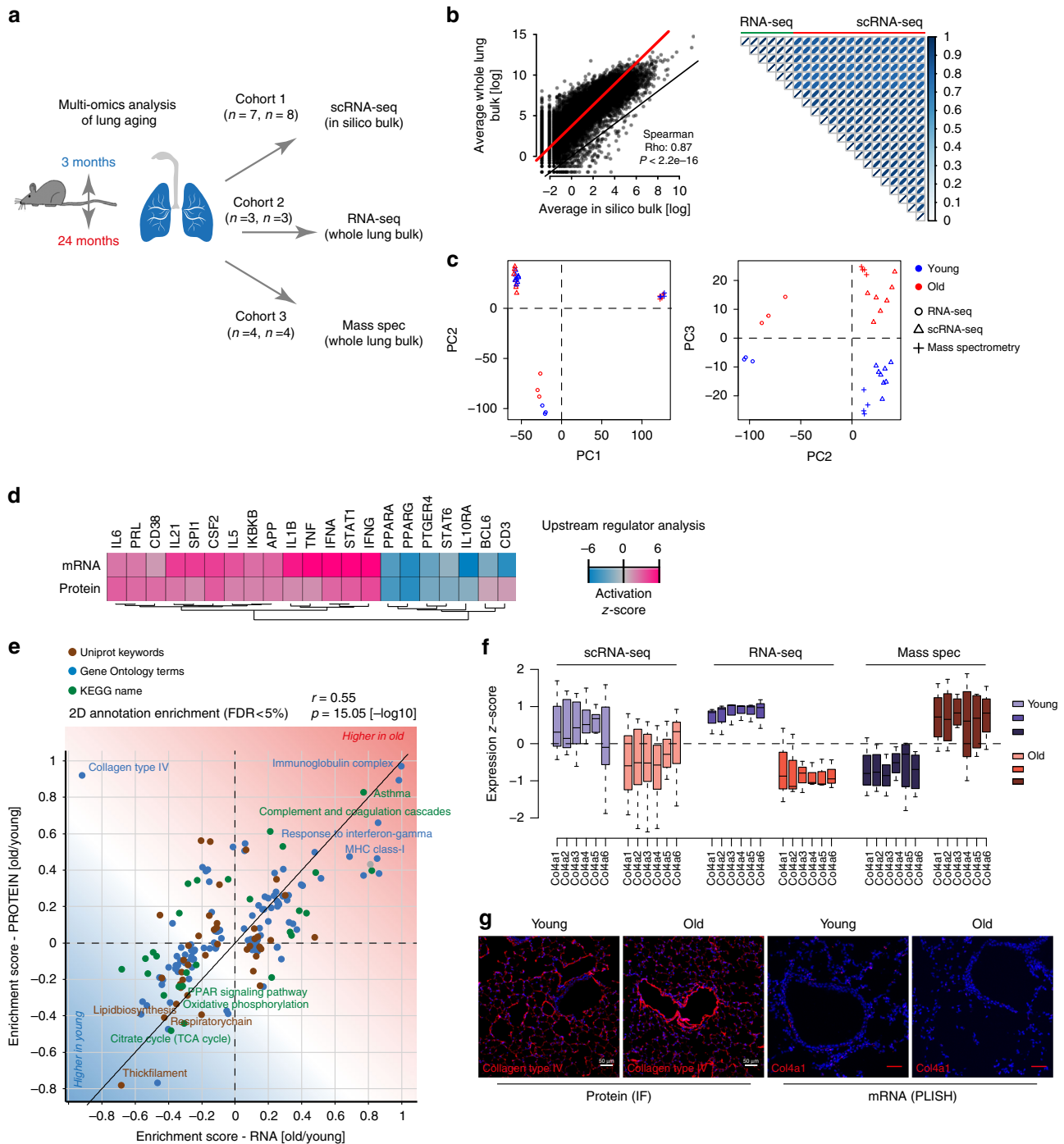


**Fig. 2** Most cell types show increased transcriptional noise with aging. **a** Boxplot illustrates transcriptional noise by age and cell type for the indicated number of cells. For all boxplots, the box represents the interquartile range, the horizontal line in the box is the median, and the whiskers represent 1.5 times the interquartile range. Blue and red colors indicate young and old cells, respectively. Asterisk indicates significant changes (Wilcoxon's rank sum test, adjusted  $p$  value < 0.05). Cell types are ordered by decreasing transcriptional noise ratio between old and young cells. **b** Scatterplot shows the log<sub>2</sub> ratio of transcriptional noise between old and young samples as calculated using mouse averages ( $n = 15$ ) and single cells on the X and Y axes, respectively. **c** Scatterplot depicts the log<sub>2</sub> ratio of transcriptional noise between old and young samples as calculated using 1-Spearman correlation and the Euclidean distance between cells on the X and Y axes, respectively. For both panels, the size of the dots corresponds to the negative log<sub>10</sub> adjusted  $p$  value of the cell type-resolved differential transcriptional noise test and the red lines correspond to the robust linear model regression fit. **d** As an example, the distribution of 1-Spearman correlation coefficients between all pairs of young and old cells is shown for type-2 pneumocytes. Larger values represent increased transcriptional noise. Blue and red colors indicate young and old samples

We observed several hallmarks of aging, including a decline in mitochondrial function and upregulation of pro-inflammatory pathways ('inflammaging'). Interestingly, we detected a strong increase in immunoglobulins in both datasets, as well as higher levels of major histocompatibility complex (MHC) class I, which is consistent with the observed increase in the interferon pathway (Fig. 2e). Many extracellular matrix genes, such as collagen III, were downregulated on both the mRNA and protein levels, while the levels of all basement membrane-associated collagen IV genes were increased on the protein level, but decreased at the mRNA level in both transcriptome datasets (Fig. 2f) and in proximity

ligation in situ hybridization of mRNA in tissue sections (Fig. 2g). The differential regulation of collagen IV transcripts and proteins highlights the importance of combined RNA and protein analysis. We validated the increased protein abundance of collagen IV using immunofluorescence and found that interestingly the main increase in collagen IV in old mice was found around airways and vessels (Fig. 2g).

**Altered frequency of airway epithelial cells upon aging.** Single-cell RNA-seq can disentangle relative frequency changes of cell types from real changes in gene expression within a given cell



type. We analyzed age-dependent alterations of relative frequencies of the 30 cell types represented in our dataset. Since the cell-type frequencies are proportions, the data are compositional. Therefore, it is impossible to statistically discern if a relative change in cell-type frequency is caused by the increase of a given cell type or the decrease of another. However, after performing dimension reduction using multidimensional scaling of the cell-type proportions, we observed a significant association between the first coordinate and age (Fig. 4a, b; Wilcoxon test,  $p < 0.005$ ), indicating that cell-type frequencies differed between young and old mice. Interestingly, the Dropseq data showed a relative increase in ciliated cells in old mice so that the ratio of club to ciliated cells was altered (Fig. 4c, d). Relative frequency differences in scRNA-seq data can be biased by tissue isolation

artifacts. We therefore validated the change in club to ciliated cell proportions by deconvolving the whole lung bulk expression data using our single-cell gene expression profiles (Fig. 4e). Indeed, we found that the ciliated cell marker genes signature was significantly upregulated in old compared to young mouse lungs (Fig. 4f). Interestingly, this analysis also revealed marked increase of various immune cell populations, including CD4+ and CD8+ T cells, eosinophils, and classical monocytes (Fig. 4e). We additionally validated this finding in situ by quantifying airway club and ciliated cells using immunostainings of Foxj1 (ciliated cell marker) and CC10 (club cell marker) (Fig. 4g). In addition, in this analysis the ciliated cells were increased in old mice (Fig. 4h), leading to a significantly altered ratio of club to ciliated cells in aged mouse airways (Fig. 4i).



**Fig. 3** Multi-omic data integration uncovers uncoupled regulation of RNA and protein. **a** Experimental design—three independent cohorts of young and old mice were analyzed by single-cell RNA-sequencing (scRNA-seq), bulk RNA-seq, and mass spectrometry-driven proteomics respectively. **b** On the left, gene expression profiles from whole lung bulk samples ( $n = 6$ ) and in silico bulk samples ( $n = 15$ ) were averaged and plotted on X and Y axes, respectively. Red and black lines indicate linear model fit and the diagonal. On the right, correlation heatmap shows Pearson's correlation between all bulk and in silico bulk samples. **c** Normalized bulk (RNA-seq) and in silico bulk (scRNA-seq) data were merged with proteome data (mass spectrometry) and quantile normalized. The first two principal components show clustering by data modality. The third principal component separates young from old samples across all three data modalities. Blue and red colors indicate young and old samples, respectively. **d** Gene expression and protein abundance fold changes were used to predict upstream regulators that are known to drive gene expression responses similar to the ones experimentally observed. Upstream regulators could be cytokines or transcription factors. The color-coded activation z-score illustrates the prediction of increased or decreased activity upon aging. **e** The scatter plot shows the result of a two-dimensional annotation enrichment analysis based on fold changes in the transcriptome (x-axis) and proteome (y-axis), which resulted in a significant positive correlation of both datasets. Types of databases used for gene annotation are color coded as depicted in the legend. **f** Expression of collagen IV genes in the in silico bulk (scRNA-seq), bulk (RNA-seq), and proteome (mass spec) experiments, respectively. The box represents the interquartile range, the horizontal line in the box is the median, and the whiskers represent 1.5 times the interquartile range. **g** Immunofluorescence image of collagen type IV using confocal microscopy at  $\times 25$  magnification and proximity ligation in situ hybridization (PLISH) staining of *Col4a1* mRNA. Note the increased fluorescence intensity of the protein around vessels in old mice along with the decreased mRNA expression (scale bar: 50  $\mu\text{m}$ )

### Altered composition of the pulmonary extracellular matrix.

The ECM can act as a solid phase-binding interface for hundreds of secreted proteins, creating an information-rich signaling template for cell function and differentiation<sup>25</sup>. Alterations in ECM composition and possibly architecture in the aging lung have been suggested<sup>26</sup>, but experimental evidence using unbiased mass spectrometry is scarce. From the 5138 proteins quantified in the tissue proteome (Fig. 5a), we identified 32 Matrisome proteins with significant change upon aging (two-sided *t*-test, FDR < 10%, Fig. 5b, Supplementary Data 2). Collagen XIV, a collagen of the FACIT (Fibril Associated Collagens with Interrupted Triple helices) family of collagens that is associated with the surface of collagen I fibrils and may function by integrating collagen bundles<sup>27</sup>, was downregulated in old mice (Fig. 5b). Collagen XIV is a major ECM binding site for the proteoglycan Decorin<sup>28</sup>, which is known to regulate TGF-beta activity<sup>29,30</sup>. Interestingly, our scRNA-seq data localized collagen XIV expression to interstitial fibroblasts that together with mesothelial cells also expressed Decorin and were distinct from the lipofibroblasts that showed very little expression of this particular collagen (Fig. 5c). Thus, the combination of tissue proteomics with single-cell transcriptomics enabled us to predict the cellular source of the regulated proteins, which can be explored in the online webtool (<https://theislab.github.io/LungAgingAtlas>). In the webtool the cell-type specificity of any gene query can be exported as dot plot in pdf format.

We previously developed the quantitative detergent solubility profiling (QDSP) method to add an additional dimension of protein solubility to tissue proteomes<sup>31–33</sup>. In QDSP, proteins are extracted from tissue homogenates with increasing stringency of detergents, which typically leaves ECM proteins enriched in the insoluble last fraction. This enables better coverage of ECM proteins and analysis of the strength of their associations with higher-order ECM structures such as microfibrils or collagen networks. We applied this method to young and old mice and compared protein solubility profiles between the two groups (Fig. 6a). Differential comparison of the solubility profiles between young and old mice revealed 74 proteins, including 8 ECM proteins, with altered solubility profiles (two-way analysis of variance (ANOVA), FDR < 20%) (Supplementary Data 4).

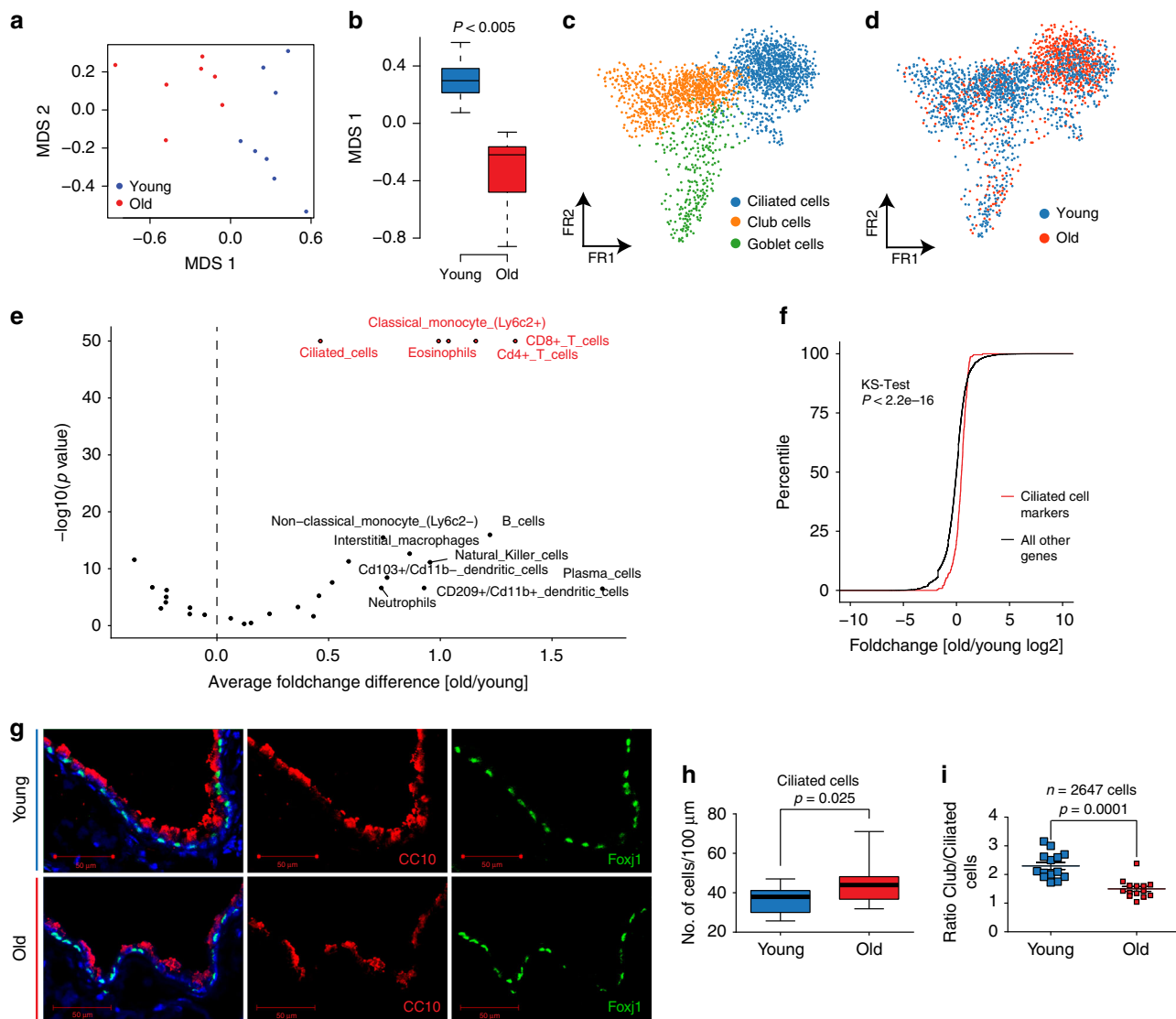
Using principal component analysis of 432 secreted extracellular proteins we found that the protein solubility fractions separated in component 1, while the age groups separated in component 4 of the data (Fig. 6b). Thus, principal component analysis enabled the stratification of secreted proteins by their biochemical solubility and their differential behavior upon aging (Fig. 6c). This analysis also showed that neither the abundance nor the solubility of many ECM proteins, including collagen I and

basement membrane laminins, was altered (Fig. 6c). While the most abundant basement membrane laminin chain (Lamc1) was unaltered in both abundance (Fig. 6d) and solubility (Fig. 6g), serving as a control for overall integrity of the basement membrane and the quality of our data, the basement membrane-associated trimeric Fraser Syndrome complex (consisting of Fras1, Frem1, and Frem2) was downregulated (Fig. 6e) and more soluble (Fig. 6h) in old age. Incorporation of the Fraser syndrome complex within the basement membrane (rendering it more insoluble) has been shown to depend on extracellular assembly of all three proteins<sup>34</sup>, indicating that this assembly and/or the expression of either one or all subunits of the complex is perturbed in old mice. Fraser syndrome is a skin-blistering disease which points to an important function of the Fraser syndrome complex proteins in linking the epithelial basement membrane to the underlying mesenchyme<sup>34</sup>. In the lungs of adult mice, expression is restricted to the mesothelium; Fras1<sup>−/−</sup> mice develop lung lobulation defects<sup>35</sup>. Interestingly, the solubility of the downregulated collagen XIV (Fig. 6f) was also significantly changed (Fig. 6i).

**Cell type-specific effects of aging.** Cell type-resolved differential gene expression testing between age groups in the single-cell data sets identified 391 significantly regulated genes (Wilcoxon rank sum test, FDR < 10%) (Fig. 7a; Supplementary Data 5). Alveolar macrophages and type-2 pneumocytes, the two cell types with highest number of cells in the dataset, are discussed as an example for the type of insight that can be gained from our cell type-resolved resource. Both cell types showed a clearly altered phenotype in aged mice.

In alveolar macrophages, we found 125 significantly regulated mRNAs (FDR < 10%, Fig. 7b), including the downregulation of the genes for Eosinophil cationic protein 1 & 2 (*Ear1* and *Ear2*), which have ribonuclease activity and are thought to have potent innate immune functions as antiviral factors<sup>36</sup>. We observed higher levels of the C/EBP beta (*Cebpb*), which is an important transcription factor regulating the expression of genes involved in immune and inflammatory responses<sup>37,38</sup>. Several genes that have been shown to be upregulated in lung injury, repair, and fibrosis<sup>33</sup>, such as *Spp1*, *Gpnmb*, and *Mfge8*, were also induced in alveolar macrophages of old mice, which may be a consequence of the ongoing 'inflammaging'.

In alveolar type-2 pneumocytes, 121 mRNAs were significantly regulated (Wilcoxon rank sum test, FDR < 10%, Fig. 7c). We observed a strong increase of the MHC class I genes *H2-K1*, *H2-Q7*, *H2-D1*, and *B2m* (Fig. 7c), which we validated using an

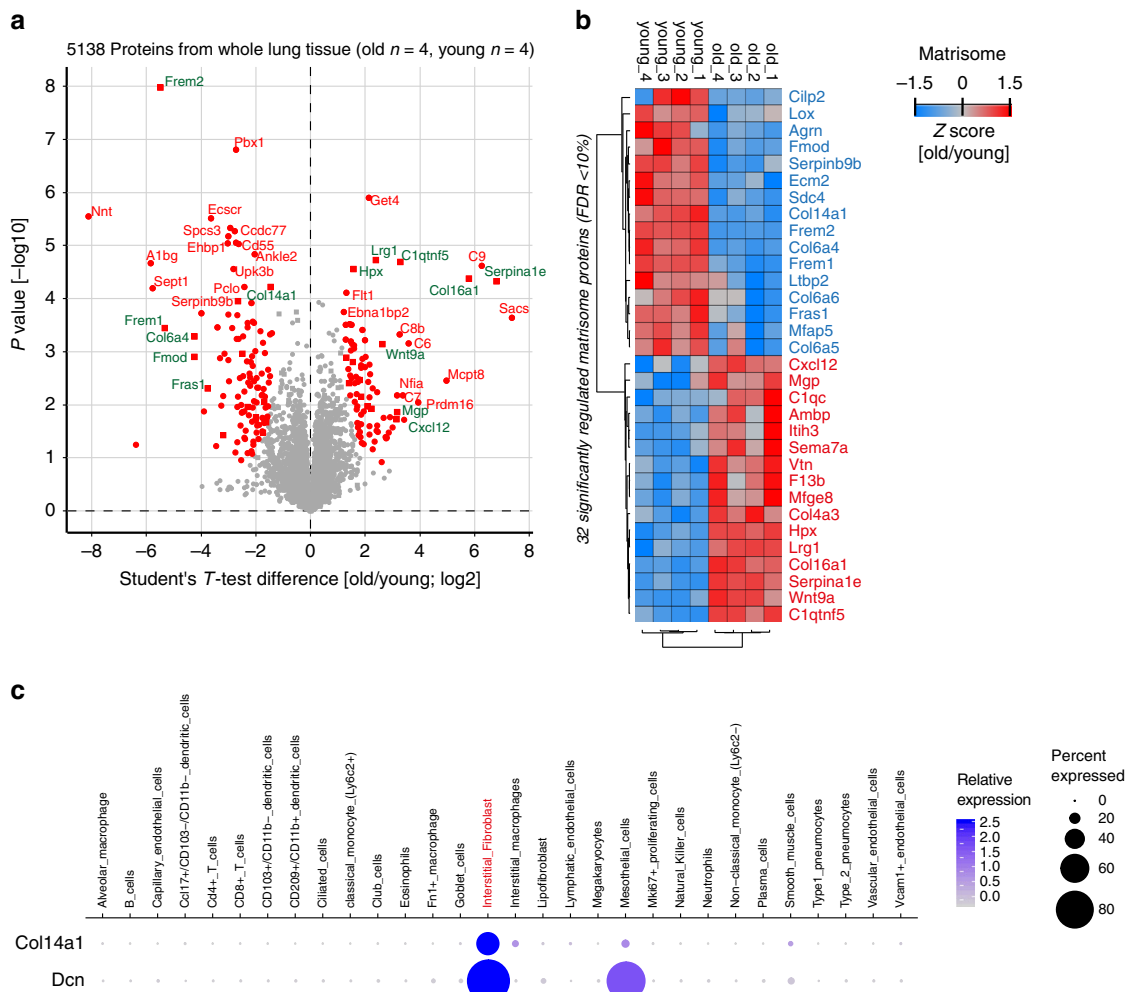


**Fig. 4** Cell-type deconvolution reveals increase of ciliated cells in airways of old mice. **a** The multidimensional scaling (MDS) plot shows the mouse-wise euclidean distances of cell-type proportions for the two age groups. **b** The box plot shows the significant difference in the multidimensional scaling component 1 of cell-type proportions between young ( $n = 8$ ) and old ( $n = 7$ ). The box represents the interquartile range, the horizontal line in the box is the median, and the whiskers represent 1.5 times the interquartile range. **c** The Fruchterman-Reingold (FR) embedding of the airway epithelial cells in the dataset reveals distinct clusters of airway cell identity. **d** The indicated color code shows the distribution of young and old cells to the three clusters presented in (**c**). Note the increased density of old cells in the ciliated cell cluster. **e** The volcano plot shows negative  $\log_{10}$  enrichment  $p$  values of cell-type marker signatures in the differential expression results of the bulk RNA-seq data from young and old mice. **f** The empirical density plot shows significant enrichment for ciliated cell-type marker genes (red line) compared to all other genes (black line) in the distribution of fold changes derived from the bulk differential expression analysis. **g** Club and ciliated cells were stained using a CC10 and Foxj1 antibody respectively (scale bar: 50  $\mu\text{m}$ ). **h** The boxplot depicts the quantification of ciliated cells from counting a total of 2647 club and ciliated cells in 14 individual airways of  $n = 2$  mice of each indicated age group. **i** Ratio of ciliated to club cells in 14 individual airways of two mice for each indicated age group. The  $p$  values are derived from an unpaired, two-tailed  $t$ -test using Welch's correction

independent flow cytometry experiment on epithelial cells marked by Epcam expression (Fig. 7k, l). Elevated MHC class I levels likely result in increased presentation of self-antigens to the immune system and are consistent with our observation of a prominent interferon-gamma signature in old mice (Fig. 3d), which is known to activate MHC class I expression<sup>39</sup>. Type-2 pneumocytes of old mice featured a highly significant upregulation of the enzyme Acyl-CoA desaturase 1 (*Scd1*), which is the fatty acyl  $\Delta 9$ -desaturating enzyme that converts saturated fatty acids into monounsaturated fatty acids (Fig. 7c). The age-dependent upregulation of *Scd1* in type-2 pneumocytes may have important implications since *Scd1* is thought to induce adaptive

stress signaling that maintains cellular persistence and fosters survival and cellular functionality under distinct pathological conditions<sup>40</sup>.

To perform global validation of the cell type-resolved differential gene expression analysis for a large number of genes we flow-sorted epithelial cells and macrophages from an additional cohort of young and old mice (see Supplementary Fig. 5 for gating strategy) and performed bulk RNA-seq on these isolated cell types from young ( $n = 4$ ) and old ( $n = 4$ ) mice. Principal component analysis (PCA) was performed using the scRNA-seq-derived signatures of alveolar macrophages and type-2 pneumocytes. Gene expression profiles of flow-sorted epithelial



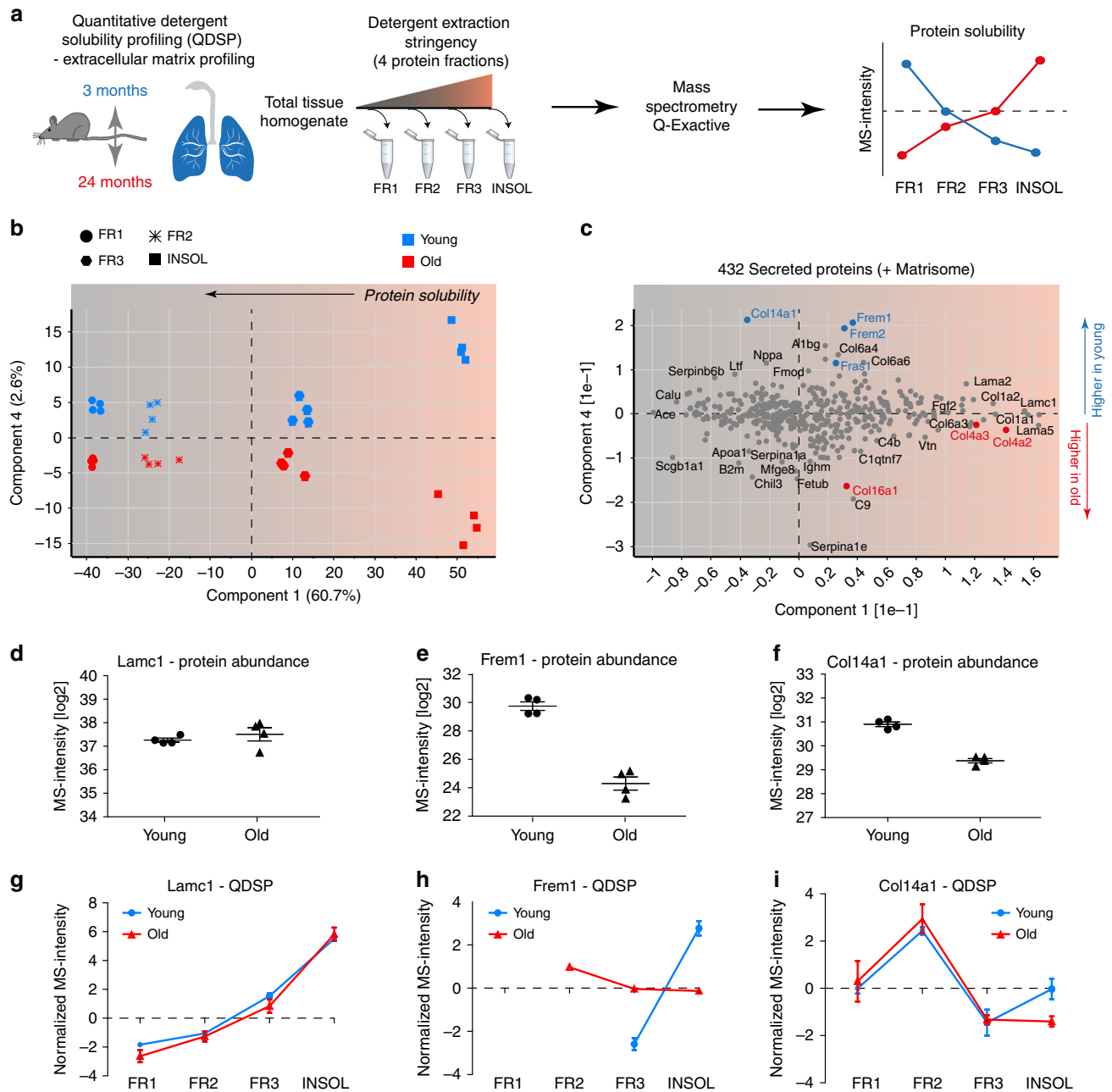
**Fig. 5** Single-cell RNA-sequencing (scRNA-seq) predicts cellular origin of age-dependent protein alterations. **a** Proteins regulated with a false discovery rate < 10% are highlighted in red in the volcano plot showing the indicated fold changes and *p* values derived from *t*-test statistic. Matrisome proteins are labeled with green gene names. **b** The *z*-score values of 32 significantly regulated extracellular matrix proteins were grouped by unsupervised hierarchical clustering (Pearson’s correlation). **c** The dot plot shows mRNA expression specificity of *Col14a1* and its binding partner Decorin (*Dcn*) in the scRNA-seq data

cells and macrophages were projected into this PCA space (see Methods for details) showing good overlap of cell-type identity, thereby confirming the scRNA-seq-based cell-type annotation (Fig. 7d, e). Next, age-dependent alterations in the flow-sorted bulk RNA-seq data were identified (Supplementary Data 2). Significant agreement with the scRNA-seq-derived results was observed (Fisher’s exact test,  $p < 2.2e-16$ , Fig. 7f–j), thus validating the power of scRNA-seq to derive age-dependent changes in gene expression.

To obtain a meta-analysis of changes in previously characterized gene expression modules and pathways, we used cell type-resolved mRNA fold changes for gene annotation enrichment analysis (Supplementary Fig. 6a and b, Supplementary Data 6) and upstream regulator analysis (Supplementary Fig. 6c–e). The analysis revealed cell type-specific alterations in gene expression programs upon aging. For instance, comparing club cells to type-2 pneumocytes showed that *Nrf2* (*Nfe2l2*)-mediated oxidative stress responses were higher in type-2 pneumocytes of old mice and lower in club cells (Supplementary Fig. 6c). Aging is known to affect growth signaling via the evolutionary conserved Igf-1/Akt/mTOR axis. Interestingly, we found evidence for increased mammalian target of rapamycin (mTOR) signaling in type-2 and club cells, but not in ciliated and goblet cells (Supplementary

Fig. 6c). Mesenchymal cells showed remarkable differences in their aging response (Supplementary Fig. 6d). For instance, we observed the pro-inflammatory *Il1b* signature in capillary endothelial cells, as well as in mesothelial and smooth muscle cells, but not in the other mesenchymal cell types. In myeloid cell types we found both differences and similarities in the aging response (Supplementary Fig. 6e). While an increased interferon-gamma and reduced *Il10* signature in old mice was consistently observed, other effects were more specific, such as the increase in *Stat1* target genes in classical monocytes (*Ly6c2+*), which was not observed in non-classical monocytes (*Ly6c2-*).

**Increased cholesterol biosynthesis in aged cell types.** Pulmonary surfactant homeostasis is a tightly regulated process that involves synthesis of lipids by type-2 pneumocytes and lipofibroblasts<sup>41</sup>. Lipid metabolism in alveolar type-2 cells is regulated by sterol-response element-binding proteins (SREBPs) such as *Srebf2* and their negative regulators *Insig1* and *Insig2*. Deletion of *Insig1/2* in mouse type-2 pneumocytes activated SREBPs and led to the accumulation of neutral lipids (cholesterol esters and triglycerids) in type-2 pneumocytes and alveolar macrophages, accompanied by lipotoxicity-related lung inflammation and tissue



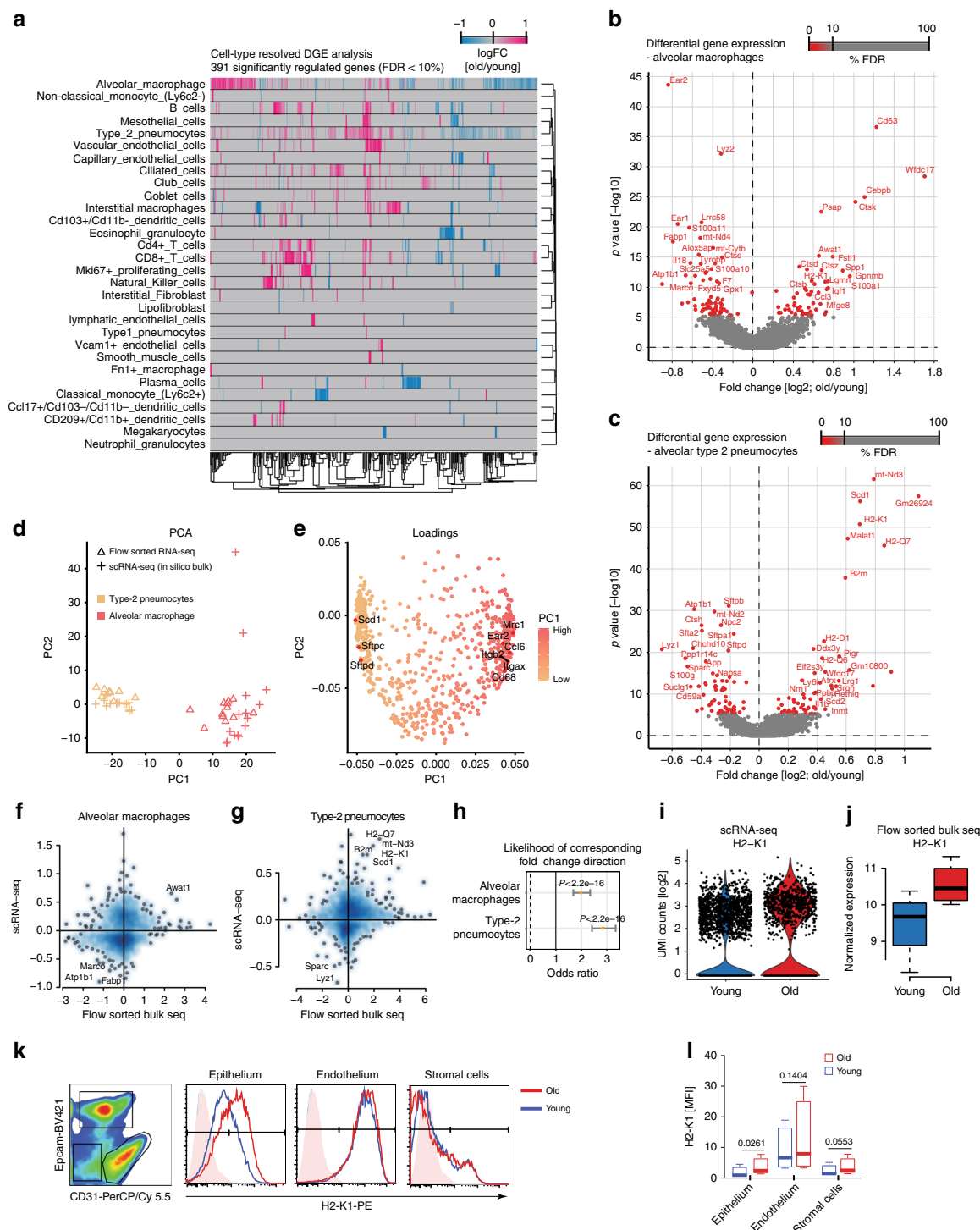
**Fig. 6** Proteome-wide detergent solubility profiling reveals changes in the extracellular matrix (ECM) architecture. **a** Experimental design—extraction of proteins from whole lung homogenates with increasing detergent stringency results in four distinct protein fractions, which are analyzed by mass spectrometry (MS). **b** The projections of a principal component analysis (PCA) of 432 proteins with the annotation ‘secreted’ in the Uniprot and/or Matrisome database separate the four protein fractions, indicated by symbol shape, in component 1 and the age groups, as indicated by color, in component 4. **c** The loadings of the PCA are shown. **d–f** Relative differences in MS intensity (abundance) of the indicated proteins. **g–i** The normalized MS intensity across the four protein fractions from differential detergent extraction highlights changes in protein solubility between young and old mice for the indicated proteins. Error bars represent the standard error of the means ( $n = 4$ )

remodeling<sup>42</sup>. Interestingly, we observed very similar gene expression changes in type-2 pneumocytes of old mice as reported for the *Insig1/2* deletion. Consistently, the upstream regulator analysis predicted increased activity of *Srebf2* and reduced activity of *Insig1* specifically in type-2 pneumocytes of old mice (Supplementary Fig. 6c). The upstream regulator analysis was based on 25 known targets of SREBP/Insig1, all of which were increased in aged type-2 pneumocytes (Fig. 8a). Using gene annotation enrichment analysis on the universal protein resource (Uniprot) Keywords, Gene Ontology (GO) terms, and Kyoto

encyclopedia of genes and genomes (KEGG) pathways (Supplementary Data 6), we found increased cholesterol biosynthesis as the top hit in type-2 pneumocytes and lipofibroblasts and no other cell type (Fig. 8b). Indeed, most of the *Insig1/2* target genes are directly involved in cholesterol biosynthesis (Fig. 8c).

To confirm the increased cholesterol biosynthesis and analyze the actual lipid content of the cells, we performed immunofluorescence of the type-2 pneumocyte marker prosurfactant protein C (proSP-C) together with the LipidTox compound that stains neutral lipids. Increased LipidTox staining in aged lungs





was specific to alveolar type-2 cells (Fig. 8d). In addition, we used the Nile red dye to stain neutral lipids in cells of a whole lung suspension after depletion of leukocytes. Using flow cytometry we quantified the Nile red lipid staining and found a significant increase in mean fluorescence intensity (Fig. 8e–g) in the CD45-negative cells of old mice. CD45+ cells were not significantly altered, indicating that the increase in neutral lipid content is specific to epithelial cells and fibroblasts. Thus, we have shown that increased cholesterol biosynthesis and neutral lipid content in type-2 pneumocytes and lipofibroblasts is a hallmark of lung aging.

## Discussion

Enabling healthy aging is one of the prime goals of the modern society. In order to better understand age-related chronic lung diseases such as COPD, lung cancer, or fibrosis, intense efforts in integrated multi-omics systems biology tools for the analysis of lung aging are needed<sup>26</sup>. In this work, we present a single-cell survey of mouse lung aging and computationally integrate single-cell transcriptomics data with bulk proteomics and transcriptomics of whole lung to build a draft of an atlas of the aging lung. Atlasing efforts are generally organized in stages so that more detailed maps of cellular phenotypes will be integrated at

**Fig. 7** Single-cell RNA-sequencing (scRNA-seq) enables cell type-resolved differential expression analysis. **a** Heatmap displays fold changes derived from the cell type-resolved differential expression analysis. Rows and columns correspond to cell types and genes, respectively. Negative fold change values (blue) represent higher expression in young compared to old. Positive fold change values are colored in pink. **b, c** Volcano plots visualize the differential gene expression results in **b** alveolar macrophages and **c** type-2 pneumocytes. X and Y axes show average log<sub>2</sub> fold change and  $-\log_{10}$  p value, respectively. **d** Scatterplot illustrates principal component analysis (PCA) of in silico bulk samples of alveolar macrophages and type-2 pneumocytes and the projected flow-sorted bulk samples. Color and shape indicate cell-type identity and data modality. PCA loadings show that well-known marker genes define the first principal component corresponding to cell-type identity (**e**). Fold changes derived from the flow-sorted bulk samples and the cell type-resolved differential expression analysis are depicted on the X and Y axes respectively for alveolar macrophages (**f**) and type-2 pneumocytes (**g**). The likelihood of corresponding fold change direction was highly enriched between the scRNA-seq and flow-sorted bulk data for both cell types (**h**). X-axis shows the odds ratio including 95% confidence interval. Black vertical line illustrates an odd ratio of one representing equal likelihood. Increased expression of H2-K1 in old compared to young mice was observed for type-2 pneumocytes in the scRNA-seq (**i**) and flow-sorted bulk (**j**) data ( $n = 4$  young and  $n = 4$  old mice). For (**j**), the box represents the interquartile range, the horizontal line in the box is the median, and the whiskers represent 1.5 times the interquartile range. **k** The indicated cell lineages were gated by flow cytometry as shown in the left panel in a CD31 and Epcam co-staining and evaluated for H2-K1 expression on protein level. The histograms show fluorescence intensity distribution of the H2-K1 cell surface staining for the indicated lineages and age groups. **l** Boxplot shows mean fluorescence intensity for H2-K1 in the indicated cell types taken from 4 young and 4 old mice. The p values are from a two-sided t-test. The box represents the interquartile range, the horizontal line in the box is the median, and the whiskers represent 1.5 times the interquartile range

later stages to initial drafts of the atlas. The intention in this study was to perform an integrated analysis of aging effects at a depth of current state of the art of proteomics and transcriptomics. The lung aging atlas and associated raw data can be accessed at <https://theislab.github.io/LungAgingAtlas> (Supplementary Fig. 7). It features five dimensions that can be navigated through gene and cell type-specific queries: (1) cell type-specific expression of genes and marker signatures for 30 cell types, (2) regulation of gene expression by age on cell-type level, (3) cell type-resolved pathway and gene category enrichment analysis, (4) regulation of protein abundance by age on tissue level, and (5) regulation of protein solubility by age.

The highly multiplexed nature of droplet-based single-cell RNA sequencing used in this study allows the direct analysis of thousands of individual cells freshly isolated from whole mouse lungs, providing unbiased classification of cell types and cellular states. Two previous studies have analyzed aging effects using single-cell transcriptomics and found increased transcriptional variability between cells in human pancreas and T cells<sup>21,22</sup>. In this study, we identify aging-associated increased transcriptional noise, which may result from deregulated epigenetic control, in most cell types of the lung, indicating that this phenomenon is a general hallmark of aging that likely affects most cell types in both mice and humans. This concept is supported by our study and it will be interesting when and how future investigations will shed light on the molecular mechanisms driving this phenomenon.

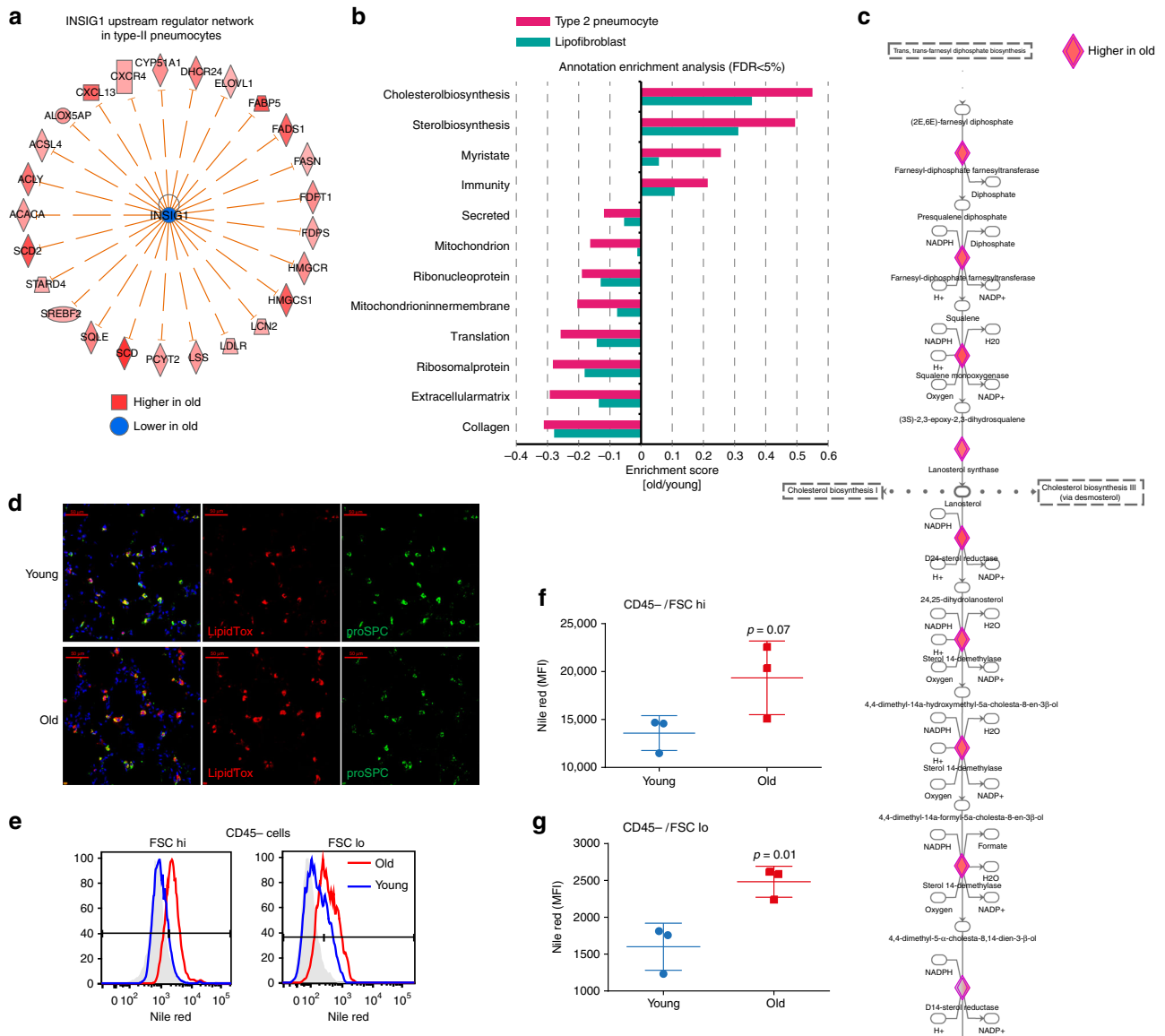
We have used three independent cohorts of young and old mice and uncovered remarkably well-conserved aging signatures in both mRNA and protein. Thus, the three datasets validate each other and show that (1) single-cell analysis can be highly representative of biological changes in total tissue, and (2) the analysis of protein and mRNA content can lead to overall similar results with important differences. Hallmarks of aging, such as the downregulation of mitochondrial oxidative phosphorylation and the upregulation of pro-inflammatory signaling pathways, were consistently observed in all datasets. On the level of individual genes/proteins, however, we often observed interesting differences, which indicates that for functional analysis of a particular gene/protein, it remains essential to also analyze the protein, which ultimately executes biological functions.

The example of basement membrane collagen IV genes that were all downregulated on the mRNA level but upregulated on the protein level illustrates that protein post-transcriptional regulation is indeed important. In particular, the abundance of ECM proteins, which often have long half-lives and are thus likely more often regulated on the posttranscriptional level, could frequently

show decoupling of protein and mRNA. Next to mass spectrometry-based methods, single-cell methods combining mRNA and protein analysis, such as cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq)<sup>43</sup>, will become ever more important in the near future. We show that the combination of single cell-resolved mRNA analysis and bulk proteomics is highly complementary using the single-cell expression data to understand the most likely cellular origin of proteins that showed altered abundance with age. Spatial transcriptomics methods for high-throughput detection of transcripts in single cells in situ are currently quickly evolving<sup>44,45</sup>. Traditional antibody-based methods for single-cell protein analysis in situ are however not well multiplexed and do not easily scale for high throughput. Thus, to fully develop the enormous potential of single-cell multi-omics data integration, the field depends on current and future developments in multiple omics layers on single-cell level in situ<sup>46,47</sup>.

We analyzed the foundations of lung tissue architecture by quantifying compositional and structural changes in the aged extracellular matrix using state-of-the-art mass spectrometry workflows. The ECM is not only key as a scaffold for the lungs overall architecture, but also an important instructive niche for cell fate and phenotype<sup>25,48</sup>. Recent proteomic studies identified at least 150 different ECM proteins, glycosaminoglycans, and modifying enzymes in the lung, and these assemble into intricate composite biomaterials that are characterized by specific biophysical and biochemical properties<sup>32,33,49</sup>. Due to this complexity of the ECM, both in terms of composition and posttranslational modification and the assembly of ECM proteins into supramolecular structures, it is presently unclear on which level and how exactly the aging process affects the lung ECM scaffold. We used detergent solubility profiling to screen for differences in protein crosslinking and complex formation within the ECM. Surprisingly, most solubility profiles were not significantly altered with age, indicating that aging-related ECM remodeling does not involve large differences in covalent protein crosslinks. However, we observed a few very strong changes in the ECM which have not yet been reported in the context of aging and are open for future investigation into their functional implications.

In order to stabilize the alveolar structure during breathing-induced expansion and contraction, type-2 pneumocytes produce and secrete pulmonary surfactant, which is a thin film of phospholipids and surfactant proteins<sup>41</sup>. The lipid composition of pulmonary surfactant has been shown to change with age<sup>50</sup>, and electron microscopy of surfactant and the lipid-loaded lamellar



**Fig. 8** Aging increases cholesterol biosynthesis in type-2 pneumocytes and lipofibroblasts. **a** The graph shows genes known to be negatively regulated by *Insig1* that were found to be upregulated in type-2 pneumocytes of old mice. **b** Selected gene categories found to be significantly (false discovery rate (FDR) < 5%) upregulated (positive enrichment scores) or downregulated (negative enrichment scores) in the indicated cell types. **c** Segment of the cholesterol biosynthesis pathway. Diamond-shaped nodes represent enzymes that were found to be upregulated in type-2 pneumocytes of old mice. The biochemical intermediates are named in between the enzyme nodes. **d** Immunofluorescence staining of lung sections of young and old mice shows type-2 pneumocytes expressing pro-SPC and neutral lipids marked by LipidTox staining (scale bar: 50 μm). **e** Quantification of Nile red stainings using flow cytometry. Histograms show flow cytometry analysis of Nile red in aged (red) and young (blue) mice; unstained control is represented in gray. Cells were stratified by size in bins of large (FSC hi) and small (FSC lo) cells using the forward scatter. **f**, **g** Nile red mean fluorescence intensity (MFI) quantification across three individual mice for **f** CD45-negative and forward scatter (FSC) high, and **g** FSC low cells. The *p* values are from an unpaired, two-sided *t*-test

bodies in type-2 pneumocytes revealed ultrastructural disorganization with age<sup>51</sup>. This may be related to our finding that cholesterol biosynthesis and neutral lipid content is upregulated in type-2 cells of old mice. It is currently unclear at which level the homeostasis of lipid metabolism is altered in the aged lungs. We found strong similarity of the aged type-2 phenotype with the phenotype in *Insig1/2* knockout mice that accumulated neutral lipids, accompanied by lipotoxicity-related lung inflammation and tissue remodeling<sup>42</sup>. Thus, it is possible that part of the chronic inflammation we observed in the aged lung is influenced by deregulation of lipid homeostasis. The inflammatory phenotype may also be related to epithelial senescence, as mice with a type-2 pneumocyte-specific deletion of telomerase, and thus

premature aging with increased senescence in these cells, developed a pro-inflammatory tissue microenvironment and were less efficient in resolving acute lung injury<sup>52</sup>.

In summary, we have demonstrated that the lung aging atlas presented here contains a plethora of information on molecular and cellular scale and serves as a reference for the large community of scientists studying chronic lung diseases and the aging process.

**Methods**

**Ethics statement.** Pathogen-free C57BL/6 mice were obtained from Charles River and housed in rooms maintained at constant temperature and humidity with a 12 h light cycle. Animals were allowed food and water ad libitum. For this study, organs were obtained from mice that had to be killed because of excessive breeding.

Animal handling was performed according to strict governmental and international guidelines and ethical oversight by the local government for the administrative region of Upper Bavaria, Germany.

**Generation of single-cell suspensions from whole mouse lung.** After killing mice, lung tissue was perfused with sterile saline from the right to the left ventricle of the heart and subsequently inflated via a catheter in the trachea by an enzyme mix containing dispase (50 caseinolytic units/ml), collagenase (2 mg/ml), elastase (1 mg/ml), and DNase (30 µg/ml). After tying off the trachea, the lung was removed and immediately minced to small pieces (approximately 1 mm<sup>2</sup>). The tissue was transferred into 4 ml enzyme mix for enzymatic digestion for 30 min at 37 °C. Enzyme activity was inhibited by adding 5 ml of phosphate-buffered saline (PBS) supplemented with 10% fetal calf serum (FCS). Dissociated cells in suspension were passed through a 70 µm strainer and centrifuged at 500 × g for 5 min at 4 °C. Red blood cell lysis (Thermo Fisher 00-4333-57) was done for 2 min and stopped with 10% FCS in PBS. After another centrifugation for 5 min at 500 × g (4 °C) the cells were counted using a Neubauer chamber and critically assessed for single-cell separation and viability. A total of 250,000 cells were aliquoted in 2.5 ml of PBS supplemented with 0.04% of bovine serum albumin and loaded for DropSeq at a final concentration of 100 cells/µl.

**Single-cell RNA sequencing.** Dropseq experiments were performed according to the original Dropseq protocol<sup>15,16</sup>. Using a microfluidic polydimethylsiloxane device (Nanoshift), single cells (100/µl) from the lung cell suspension were co-encapsulated in droplets with barcoded beads (120/µl, purchased from ChemGenes Corporation, Wilmington, MA) at rates of 4000 µl/h. Droplet emulsions were collected for 15 min/each prior to droplet breakage by perfluorooctanol (Sigma-Aldrich). After breakage, beads were harvested and the hybridized mRNA transcripts reverse transcribed (Maxima RT, Thermo Fisher). Unused primers were removed by the addition of exonuclease I (New England Biolabs), following which beads were washed, counted, and aliquoted for pre-amplification (2000 beads/reaction, equals ~100 cells/reaction) with 12 PCR cycles (Smart PCR primer: AAGCAGTGGTATCAACGCAGAGT (100 µM), 2× KAPA HiFi Hotstart Ready-mix (KAPA Biosystems), cycle conditions: 3 min 95 °C, 4 cycles of 20 s 98 °C, 45 s 65 °C, 3 min 72 °C, followed by 8 cycles of 20 s 98 °C, 20 s 67 °C, 3 min 72 °C, then 5 min at 72 °C)<sup>15</sup>. PCR products of each sample were pooled and purified twice by 0.6× clean-up beads (CleanNA), following the manufacturer's instructions. Prior to tagmentation, complementary DNA (cDNA) samples were loaded on a DNA High Sensitivity Chip on the 2100 Bioanalyzer (Agilent) to ensure transcript integrity, purity, and amount. For each sample, 1 ng of pre-amplified cDNA from an estimated 1000 cells was tagmented by Nextera XT (Illumina) with a custom P5 primer (Integrated DNA Technologies). Single-cell libraries were sequenced in a 100 bp paired-end run on the Illumina HiSeq4000 using 0.2 nM denatured sample and 5% PhiX spike-in. For priming of read 1, 0.5 µM Read1CustSeqB (primer sequence: GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTAC) was used.

**Bioinformatic processing of scRNA-seq reads.** The Dropseq core computational pipeline<sup>15</sup> was used for processing next-generation sequencing reads of the scRNA-seq data. STAR (version 2.5.2a) was used for mapping<sup>53</sup>. Reads were aligned to the mm10 genome reference (provided by the Dropseq group via the Gene Expression Omnibus (GEO) accession code GSE63269). For cell filtering, we considered all barcodes with more than 200 genes detected within the top 1200 barcodes by total UMI counts. Samples muc3838, muc3839, muc3840, and muc3841 were sequenced at lower depth in which case we considered the top 800, 500, 500, and 500 barcodes by total UMI counts corresponding to the expected number of cells, respectively.

**Single-cell data analysis.** After constructing the single-cell gene expression count matrix, we used the R package Seurat<sup>54</sup> and custom scripts for analysis.

For unsupervised clustering and visualization, we first defined highly variable genes within each mouse sample separately following the Seurat standard approach. Next, genes appearing in >4 mouse samples in the set of highly variable genes were defined as a set of consensus highly variable genes. To minimize the effect of cell cycle on clustering we removed cell-cycle genes<sup>55</sup> from the set of consensus highly variable genes. All 14,813 cells passing quality control were merged into one count matrix and normalized and scaled using Seurat's `NormalizeData()` and `ScaleData()` functions, in which we regressed out the total UMI count. The reduced set of consensus highly variable genes was used as the feature set for independent component analysis using Seurat's `RunICA()` function. The first 30 independent components were used for tSNE visualization and Louvain clustering using the Seurat functions `RunTSNE()` and `FindClusters()`, respectively.

To quantitatively assess the clustering overlap across mouse samples, the Silhouette coefficient was calculated. The Silhouette coefficient was calculated between the Euclidean distance of the 50 independent components and the mouse sample indicator. The Silhouette coefficient ranges from -1 to 1 and values close to zero indicate random clustering with regard to the specified indicator.

The Seurat `FindAllMarkers()` function was used to identify cluster-specific marker genes. Based on manual annotation and with guidance of the enrichment analysis (see below), the 36 clusters were assigned to 30 cell-type identities. Using the annotation of cell-type identities, the `FindAllMarkers()` function was called to identify the final set of cell-type markers used throughout this analysis.

An important technical detail needed our attention and is briefly described here. As infrequently discussed in the community but not yet addressed, we also observed 'ambient mRNA' effects, which we believe are the consequence of free mRNA released from dying cells hybridizing with beads in droplets during the microfluidic capture of single cells in the Dropseq workflow. The ambient mRNAs are typically derived from highly abundant transcripts and this artifact is inherent to all droplet-based methods (including the commercially available 10× platform). Here, it can be exemplified by the *Scgbl1* gene in Fig. 1c that is known to be highly specific for club and goblet cells but was nevertheless detected in almost 100% of the cells in our data. However, the UMI count levels were much higher in club and goblet cells (representing the real source of expression), indicating that the mRNA counts observed in all other clusters were of ambient mRNA background. To independently confirm this we therefore determined all genes that showed ambient mRNA background by analyzing the identity of genes on beads at the tail-end of the total UMI count distribution (on average 10 UMIs per barcode), representing empty beads that were never in contact with a real cell but nevertheless contain information from free-floating ambient mRNA. We identified 153 genes (Supplementary Data 7) with an 'ambient mRNA' effect and accounted for this effect in the cell type-resolved differential expression analysis (see below for details).

To aid the assignment of cell type to clusters derived from unsupervised clustering, we performed cell-type enrichment analysis. Cell-type gene signatures obtained from bulk-level gene expression were downloaded from the ImmGen and xCell resources. Each gene signature obtained from our clustering was statistically evaluated for overlap with gene signatures contained in these two resources. Mouse gene symbols were capitalized to map to human gene symbols. Overlap between gene signatures was evaluated using Fisher's exact test.

Cell-type marker signatures in our data (Supplementary Data 1) were compared to cell-type marker signatures in the MCA<sup>13</sup>. `MatchScore`<sup>20</sup> was used to quantify overlap between cell-type marker signatures derived from our study and the MCA. Marker genes with adjusted *p* value < 0.1 and average log fold change > 1 were considered.

Transcriptional noise in the gene expression profiles was quantified following previous work<sup>22</sup>. For each cell type with at least 10 old and young cells, we quantified transcriptional noise in the following manner. To account for differences in total UMI counts, all cells were downsampled so that all cells had equal number of total UMI counts. To account for differences in cell-type frequency, cell numbers were down-sampled so that equal numbers of young and old cells were used. Next, genes were divided into 10 equally sized bins based on mean expression and the top and bottom bins excluded. Within each bin, the 10% of genes with the lowest coefficient of variation were selected. Subsampled raw count data were reduced to this set of genes and square-root transformed. Next, the euclidean distance between each cell and the corresponding cell-type mean within each age group was calculated. This euclidean distance was used as one measure of transcriptional noise for each cell. Additionally, we average the euclidean distances for each mouse and calculated the transcriptional noise ratio between young and old mice. Alternatively, we calculated Spearman's correlation coefficients on the down-sampled expression matrices across all genes between all pairwise cell comparisons within each cell type and age group. To be consistent with the sign of the metric we used 1-Spearman correlation coefficient as the second measure of transcriptional noise. To statistically assess the association between transcriptional noise and age within each cell type, Wilcoxon's rank sum test was used. The *p* values were subsequently corrected for multiple testing using the Bonferroni-Hochberg method as implemented in the R function `p.adjust()`.

Cell-type frequencies were calculated based on the counts of cells annotated to each cell type for each mouse. Counts were transformed to proportions using the `DR_data()` function of the `DirichletReg` R package which causes the values to shrink away from extreme values of 0 and 1. Next, the mouse-wise euclidean distances were calculated based on these proportions using the `dist()` R function followed by multidimensional scaling using the `isoMDS()` R function. To statistically assess the association between age and the first coordinate derived from the multidimensional scaling, Wilcoxon test was applied. Relative changes in cell-type frequencies were calculated by subtracting the median cell-type proportion of the young mice from the cell-type proportions of the old mice.

Cell type-resolved differential expression analysis was performed using the Seurat differential gene expression testing framework. Within each cell type, cells were grouped by age and differential testing performed using the Seurat `FindMarkers()` function. By inspecting barcodes with a very low number of UMI counts, we identified 153 potential ambient mRNAs. However, these mRNAs could represent true housekeeper genes which are constitutively expressed in all cells. Therefore, we removed 41 mRNAs which showed no cell type-specific expression effect (log<sub>2</sub> fold change < 1) in any of the cell types in the cell-type marker discovery analysis from this list. Next, to avoid differential testing of a gene in a cell type where expression levels are driven by the ambient effect, cell type-resolved differential expression testing of the remaining 112 ambient mRNAs was limited to



cell types in which the ambient mRNA showed moderate cell type-specific expression (adjusted  $p$  value  $< 0.25$ ).

The one-dimensional annotation enrichment analysis<sup>24</sup> was used for cell type-resolved pathway analysis. We used the freely available software package Perseus<sup>56</sup>, as previously described<sup>33</sup>. To predict the activity of upstream transcriptional regulators and growth factors based on the observed gene expression changes, we used the Ingenuity® Pathway Analysis platform (IPA®, QIAGEN Redwood City, [www.qiagen.com/ingenuity](http://www.qiagen.com/ingenuity)). The analysis uses a suite of algorithms and tools embedded in IPA for inferring and scoring regulator networks upstream of gene expression data based on a large-scale causal network derived from the Ingenuity Knowledge Base. The analytics tool Upstream Regulator Analysis<sup>23</sup> was used to compare the known effect (transcriptional activation or repression) of a transcriptional regulator on its target genes to the observed changes to assign an activation Z-score. Since it is a priori unknown which causal edges in the master network are applicable to the experimental context, the Upstream Regulator Analysis tool uses a statistical approach to determine and score those regulators whose network connections to dataset genes as well as associated regulation directions are unlikely to occur in a random model<sup>23</sup>. In particular, the tool defines an overlap  $p$  value measuring enrichment of network-regulated genes in the dataset, as well as an activation Z-score which can be used to find likely regulating molecules based on a statistically significant pattern match of up- and down-regulation, and also to predict the activation state (either activated or inhibited) of a putative regulator. In our analysis we considered genes with an overlap  $p$  value of  $> 7$  ( $\log_{10}$ ) that had an activation Z-score  $> 2$  as activated and those with an activation Z-score  $< -2$  as inhibited.

**Proteomics and multi-omics data integration.** For proteome analysis ~100 mg of fresh frozen total tissue (wet weight) of mouse lungs was homogenized in 500  $\mu$ l PBS (with protease inhibitor cocktail) using an Ultra-turrax homogenizer. After centrifugation the soluble proteins were collected and proteins were extracted from the insoluble pellet in three steps using buffers with increasing stringency using the QDSP protocol<sup>33</sup>. Lungs were perfused with PBS through the heart to remove blood. Then, ~100 mg of total lung tissue (wet weight) was homogenized in 500  $\mu$ l PBS (with protease inhibitor cocktail and EDTA) using an Ultra-turrax homogenizer. After centrifugation the soluble proteins were collected and proteins were extracted from the insoluble pellet in three steps using buffers with increasing stringency (*buffer 1*: 150 mM NaCl, 50 mM Tris-HCl (pH 7.5), 5% Glycerol, 1% IGPAL-CA-630 (Sigma, #I8896), 1 mM MgCl<sub>2</sub>, 1 $\times$  Protease inhibitors (+EDTA), 1% Benzonase (Merck, #70746-3), 1 $\times$  Phosphatase inhibitors (Roche, #04906837001); *buffer 2*: 50 mM Tris-HCl (pH 7.5), 5% Glycerol, 150 mM NaCl + fresh protease inhibitor tablet (+EDTA), 1.0% IGPAL® CA-630, 0.5% sodium deoxycholate, 0.1% SDS, 1% Benzonase (Merck, #70746-3); *buffer 3*: 50 mM Tris-HCl (pH 7.5), 5% Glycerol, 500 mM NaCl, protease inhibitor tablet (+EDTA), 1.0% IGPAL® CA-630, 2% sodium deoxycholate, 1% SDS, 1% Benzonase (Merck, #70746-3)). Insoluble pellets were resuspended in detergent containing buffers and incubated for 20 min on ice (except for buffer 3, which was used at room temperature), followed by separation of soluble and insoluble material using centrifugation for 20 min at 16,000  $\times$  g. The PBS from the tissue homogenate and the NP40 soluble fraction (buffer 1) was pooled which, together with the two fractions derived from ionic detergent extraction (buffer 2 and 3), resulted in a total of three soluble fractions and one insoluble pellet that were subjected to liquid chromatography-tandem mass spectrometry (LC-MS/MS) analysis. Soluble proteins were precipitated with 80% acetone and subjected to in solution digestion using a modified published protocol<sup>57</sup>. In brief, protein reduction (10 mM TCEP) and alkylation (50 mM CAA) were performed at once in 6 M guanidium hydrochloride (100 mM Tris-HCl pH 8) at 99 °C for 15 min. Subsequent protein digestion was done in two steps. The first digestion was done at 37 °C for 2 h with LysC (1:50 enzyme to protein ratio) in 10 mM Tris-HCl (pH 8.5) containing 2 M guanidium hydrochloride (Gdm), 2.7 M Urea, and 3% acetonitrile. The second digestion step was done using fresh LysC (1:50 enzyme to protein ratio) and trypsin (1:20 enzyme to protein ratio) in 600 mM Gdm, 800 mM Urea, and 3% acetonitrile at 37 °C overnight. For the insoluble protein pellet, which is strongly enriched for insoluble ECM proteins, we optimized the in-solution digestion protocol with additional steps involving extensive mechanical disintegration and ultra-sonication aided digestion. The insoluble material was cooked, reduced, and alkylated in 6 M Gdm for 15 min and then subjected to 200 strokes in a micro-dounce device, which reduced the particle size of the insoluble protein meshwork. We then proceeded with the two-step digestion protocol described above, which was additionally aided by 15 min ultrasonication (Bioruptor, Diagenode) in the presence of the enzymes in both digestion steps. Peptides were purified using stage-tips containing a polystyrene-divinylbenzene copolymer modified with sulfonic acid groups (SDB-RPS) material (3 M, St. Paul, MN 55144-1000, USA) as previously described<sup>57</sup>.

Mass spectrometry data were acquired on a Quadrupole/Orbitrap type Mass Spectrometer (Q-Exactive, Thermo Scientific) as previously described<sup>33</sup>. Approximately 2  $\mu$ g of peptides were separated in a 4 h gradient on a 50 cm long (75  $\mu$ m inner diameter) column packed in-house with ReproSil-Pur C18-AQ 1.9  $\mu$ m resin (Dr. Maisch GmbH). Reverse-phase chromatography was performed with an EASY-nLC 1000 ultra-high pressure system (Thermo Fisher Scientific), which was coupled to a Q-Exactive Mass Spectrometer (Thermo Scientific). Peptides were loaded with buffer A (0.1% (v/v) formic acid) and eluted with a nonlinear 240 min

gradient of 5–60% buffer B (0.1% (v/v) formic acid, 80% (v/v) acetonitrile) at a flow rate of 250 nl/min. After each gradient, the column was washed with 95% buffer B and re-equilibrated with buffer A. Column temperature was kept at 50 °C by an in-house designed oven with a Peltier element<sup>58</sup> and operational parameters were monitored in real time by the SprayQc software<sup>59</sup>. MS data were acquired with a shotgun proteomics method, where in each cycle a full scan, providing an overview of the full complement of isotope patterns visible at that particular time point, is followed by up to 10 data-dependent MS/MS scans on the most abundant not yet sequenced isotopes (top10 method)<sup>60</sup>. Target value for the full scan MS spectra was  $3 \times 10^6$  charges in the 300–1650  $m/z$  range with a maximum injection time of 20 ms and a resolution of 70,000 at  $m/z$  400. Isolation of precursors was performed with the quadrupole at window of 3 Th. Precursors were fragmented by higher-energy collisional dissociation with normalized collision energy of 25% (the appropriate energy is calculated using this percentage, and  $m/z$  and charge state of the precursor). MS/MS scans were acquired at a resolution of 17,500 at  $m/z$  400 with an ion target value of  $1 \times 10^5$ , a maximum injection time of 120 ms, and fixed first mass of 100 Th. Repeat sequencing of peptides was minimized by excluding the selected peptide candidates for 40 s.

MS raw files were analyzed by the MaxQuant<sup>61</sup> (version 1.4.3.20) and peak lists were searched against the human Uniprot FASTA database (version Nov 2016), and a common contaminants database (247 entries) by the Andromeda search engine<sup>62</sup> as previously described<sup>33</sup>. As fixed modification cysteine carbamidomethylation and as variable modifications, hydroxylation of proline and methionine oxidation was used. False discovery rate was set to 0.01 for proteins and peptides (minimum length of seven amino acids) and was determined by searching a reverse database. Enzyme specificity was set as C-terminal to arginine and lysine, and a maximum of two missed cleavages were allowed in the database search. Peptide identification was performed with an allowed precursor mass deviation up to 4.5 ppm after time-dependent mass calibration and an allowed fragment mass deviation of 20 ppm. For label-free quantification in MaxQuant the minimum ratio count was set to two. For matching between runs, the retention time alignment window was set to 30 min and the match time window was 1 min. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [1] partner repository with the dataset identifier PXD012307.

QDSP analysis intensities were first normalized such that the mean  $\log_2$  intensities of the young and the old samples are zero, respectively. Using the normalized intensities, a two-way ANOVA with the two-factor treatment (old/young) and solubility fraction (FR1, FR2, FR3, INSOL) and the corresponding interaction term was performed using the R function `aov()`. Proteins significant in the interaction term correspond to proteins for which the solubility profile changes between young and old mice. Therefore, the corresponding  $p$  value was used for filtering the significantly changed profiles after FDR correction.

**scRNA-seq, bulk RNA-seq, and proteome integration.** In silico bulk samples were generated by summing UMI counts across all cells within one mouse sample. Differential gene expression analysis of in silico bulk samples was performed using the R package DESeq2 (v1.20.0)<sup>63</sup>. To integrate scRNA-seq, bulk RNA-seq, and protein data, the following approach was used. Raw count data from the in silico bulk and whole lung tissue bulk were normalized using the `voom()` function of the limma R package<sup>64</sup>. Next, in silico bulk, whole lung tissue bulk, and protein data were merged on a set of genes present in all three data sets and quantile normalized. This merged and quantile normalized expression matrix was then subjected to PCA.

Some statistical and bioinformatics operations, such as normalization, pattern recognition, cross-omics comparisons, and multiple-hypothesis testing corrections, were performed with the Perseus software package<sup>56</sup>. The two-dimensional annotation enrichment test used to compare proteome and transcriptome is based on a two-dimensional generalization of the nonparametric two-sample test. The false discovery rate is stringently controlled by correcting for multiple hypothesis testing<sup>24</sup>.

**Flow cytometry.** Isolated total lung cell suspensions were used to detect and quantify cell populations and activation by flow cytometry. We depleted red blood cells by positive selection of Ter199 cells, followed by CD45 bead separation (Miltenyi Biotec; Bergish Gladbach, 130-052-301). Next, we analyzed cells by fluorescence-activated cell sorting (FACS) cell suspensions before and after CD45 separation and stained cell suspensions with anti-mouse CD31 (Biolegend, 102419), EpCAM (Biolegend, 118225), and H2-K1 (Thermo Fisher Scientific, Waltham, 12-5998-81). Cells were stained in the dark at 4 °C for 20 min. CD45 lineage-negative cells were stained with Nile red (Santa Cruz Biotechnology, sc-203747) in a 1:1000 dilution for 10 min at 4 °C, as previously reported<sup>62</sup>. Cells were sorted using the CD45-negative fraction of the cell isolate stained for anti-mouse CD31, and EpCAM antibodies. Epithelial cells were sorted as CD31<sup>-</sup> cells and EpCAM<sup>+</sup> cells. For sorting macrophages we used the CD45-positive fraction and stained with anti-mouse CD11c (Biolegend, 117310), CD11b (Biolegend, 101216), MHC II (Biolegend, 107615), Siglec-F (552126, BD Pharmingen), and Ly6G (Biolegend, 127627) antibodies. For flow cytometry sorting, neutrophils were excluded by selection of Ly6G-negative cells. Macrophages were sorted as MHCII<sup>+</sup>, CD11c<sup>+</sup>, CD11b<sup>+</sup> as previously described<sup>65</sup>. Data acquisition was performed in

a BD Fortessa flow cytometer (Becton Dickinson, Heidelberg, Germany). All stainings were performed per 300,000 cells in the following dilutions: CD31 (1:300), EpCAM (1:50), H2-K1 (1:50), CD11c (1:100), Siglec-F(1:20), CD11b (1:25), MHCII (1:50), and Ly6G (1:10).

Data were analyzed using the FlowJo software (TreeStart Inc., Ashland, OR, USA). Data were reported as absolute numbers (cells/ $\mu$ l), normalized by bead counts (BD Truecount TM Beads tubes; BD Biosciences, Heidelberg, Germany) (Supplementary Fig 5). For H2-K1 and Nile red, data were analyzed by mean fluorescence intensity (MFI). Negative thresholds for gating were set according to isotype-labeled and unstained controls.

**Bulk RNA-sequencing and analysis.** RNA was isolated from whole lung tissue using the Qiagen RNeasy® Mini Kit (#74104) according to the manufacturer's recommendations. The RNA isolate was thereafter enriched for poly-A templates and submitted for whole mRNA sequencing on the Illumina HiSeq4000.

Whole lung tissue bulk RNA next-generation sequencing reads were aligned to the mouse reference genome mm10 using STAR<sup>53</sup> (version 2.2.1). Read summarization was performed using the featureCounts<sup>63</sup> (version 1.5.0) tool. To statistically evaluate the agreement between the in silico bulk and true bulk RNA-seq data, Spearman's correlation coefficients were calculated on the gene expression profiles between all sample pairs and the averages of both modalities. Differential gene expression analysis of whole lung tissue bulk samples was performed using the R package DESeq2<sup>66</sup> (v1.20.0).

To identify potential age-dependent alterations in tissue composition, the whole lung tissue bulk RNA-seq were integrated with the scRNA-seq-derived cell-type signatures. Kolmogorov-Smirnov test was used to statistically evaluate the enrichment of cell-type marker genes in the fold changes derived from the differential expression analysis of the whole lung tissue bulk RNA-seq. The *p* values were limited to the range from 1 to 1e−50.

Flow-sorted macrophages and epithelial cells were immediately lysed after sorting and cDNA synthesis was performed using the Smart-Seq® v4 Ultra® Low Input RNA Kit for Sequencing (TaKaRa, 634896). For each sample, 200 pg of pre-amplified cDNA from an estimated 2000 cells was tagged by Nextera XT (Illumina) according to the manufacturer's protocol and submitted for sequencing on the Illumina HiSeq4000.

Flow-sorted bulk RNA next-generation sequencing reads were aligned to the mouse reference genome mm10 using STAR<sup>53</sup> (version 2.2.1). Read summarization was performed using the featureCounts<sup>63</sup> (version 1.5.0) tool. To increase comparability between bulk and single-cell RNA-seq data, a total of 30 in silico bulk samples were generated by summing the counts from all cells belonging to the alveolar macrophages and type-2 pneumocytes clusters for each mouse. Next, PCA was calculated for these in silico bulk samples using the alveolar macrophages and type-2 pneumocytes marker genes with adjusted *p* value < 0.1 and fold change > 0 (Supplementary Data 1). Subsequently, the flow-sorted bulk RNA-seq samples were projected into this PCA space to show correspondence between the scRNA-seq-derived in silico bulk samples and the flow-sorted RNA-seq samples.

Differential expression analysis of flow-sorted bulk RNA-seq samples was conducted using the R package limma<sup>64</sup>. To statistically evaluate the agreement between the age-dependent alterations measured in the scRNA-seq and flow-sorted bulk RNA-seq data, Fisher's exact test was used. Fisher's exact test assesses the likelihood of genes having the same fold change direction (up- or down-regulation in old compared to young).

**Proximity ligation in situ hybridization (PLISH).** Samples were prepared and processed for PLISH and immunostaining as described in Nagendran et al.<sup>67</sup> with some modifications. 14  $\mu$ m mouse lung cryosections were collected on superfrost slides and allowed to air dry for 10 min. The slides were immersed in prewarmed 10 mM citrate buffer containing 0.05% lithium dodecyl sulfate at 100 °C in a water bath for 5 min. The slides were quickly removed, rinsed with diethyl pyrocarbonate (DEPC)-treated water and air dried. Seal chambers (GBL621505 Sigma-Aldrich) were applied and the sections were rehydrated with DEPC-treated water for 1 min. The samples were incubated with 0.025 mg/ml Pepsin (10108057001 Roche; from Sigma-Aldrich) in 0.1 M HCL for 5 min at 37 °C followed by a quick rinse with 1× PBS and the addition of H probes for *Col4a1*.

H probe sequences were: Col4a1 NM\_009931.2:mmHLC2-VB01-Col4a1-5315 AGGTCAGGAATACTTACGTCGGTTATGGTAGGGTTCATTGCTGTTCACA, mmHRC2-VB01-Col4a1-5315 AGGTACACAGGATATAATCTTATAGGTCGAGTAGTATAGCCAGGTT, mmHLC2-VB01-Col4a1-5385 AGGTACAGGAATACTTACGTCGGTTATGGAGTTACGCGAATCCCTATAA, mmHRC2-VB01-Col4a1-5385 CCAACGAAGCGGGGTGTGTTTATAGGTCGAGTAGTATAGCCAGGTT, mmHLC2-VB01-Col4a1-5910 AGGTACAGGAATACTTACGTCGGTTATGGTTGACCTGCCAATTGCTGA, mmHRC2-VB01-Col4a1-5910 AACAGGCTCTACGCTAGAACTTATAGGTCGAGTAGTATAGCCAGGTT, mmHLC2-VB01-Col4a1-5848 AGGTACAGGAATACTTACGTCGGTTATGGATTATTTATTTTCCATCTA, mmHRC2-VB01-Col4a1-5848 ATATATATATTTTACTTTTATAGGTCGAGTAGTATAGCCAGGTT,

mmHLC2-VB01-Col4a1-5753

AGGTCAGGAATACTTACGTCGGTTATGGAGGTTTGTGTTTGGGGCTGA,

mmHRC2-VB01-Col4a1-5753

CATAGTACCACACAGGGCATTATAGGTCGAGTAGTATAGCCAGGTT.

Connector circle CCC2.1: 5'

ATTCTGACCTAACAAACATGCGTCTATAGTGGAGCCACATAAT-TAAACCTGGCTAT 3'.

Variable bridge VB01-P1:

ACTACTCGACCTATAACCATAACGACGTAAGT.

Label probe: LP1m-Cy5: 5'Cy5/ CTACTACTCGACCTATA.

**Immunofluorescence and histology.** For immunofluorescence microscopy, mouse lungs were perfused with PBS, fixed in 4% paraformaldehyde (pH 7.0), and embedded in paraffin for formalin-fixed, paraffin-embedded sections. The paraffin sections (3.5  $\mu$ m) were deparaffinized and rehydrated, and the antigen retrieval was accomplished by pressure-cooking (30 s at 125 °C and 10 s at 90 °C) in citrate buffer (10 mM, pH 6.0). After blocking for 1 h at room temperature with 5% bovine serum albumin, the lung sections were incubated with the primary antibodies overnight at 4 °C, incubated with the secondary antibodies (1:250) for 2 h, followed by 4',6-diamidino-2-phenylindole (DAPI; Sigma-Aldrich, 1:2000) for 20 min at room temperature. Images were acquired with an LSM 710 microscope (Zeiss). The following primary (1) and secondary (2) antibodies were used: (1) CC10 rabbit (Santa Cruz, sc-25554, 1:100), Foxj1 mouse (Santa Cruz, sc-53139, 1:50), collagen IV rabbit (Abcam, ab6586, 1:100), (2) donkey anti-mouse Alexa Fluor (AF) 647 (Invitrogen, A21447), donkey anti-rabbit AF 568 (Invitrogen, A10042), and donkey anti-goat AF 488 (Invitrogen, A21202). Counterstain with LipidTox was performed using HCS LipidTOX deep red neutral lipid stain (Invitrogen, H34477, 1:200).

The frequency of ciliated (nuclear Foxj1+) and club cells (CC10+) were quantified by counting 2647 cells, covering a total length of 22 mm airway in 28 individual airways (young, *n* = 14; old *n* = 14) of 2 mice of each age group. We normalized cell numbers to the total length of their respective airway using the ZEN 2.3 SP1 software for image processing.

**Reporting summary.** Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

**Code availability.** The code to reproduce the analyses and figures described in this study can be found at: [github.com/theislab/2018\\_Angelidis](https://github.com/theislab/2018_Angelidis).

## Data availability

Proteome raw data can be downloaded from the PRIDE repository under the accession number [PSX012307](https://www.ebi.ac.uk/pride/archive/study/PSX012307). scRNA-seq, whole lung tissue bulk and flow-sorted cell populations bulk raw data, can be downloaded from the Gene Expression Omnibus under the accession number [GSE124872](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE124872). The whole lung aging atlas can be accessed via an interactive user-friendly webtool at: <https://theislab.github.io/LungAgingAtlas>. All other data supporting the findings of this study are available from the corresponding authors upon reasonable request.

Received: 23 June 2018 Accepted: 1 February 2019

Published online: 27 February 2019

## References

- Hogan, B. L. et al. Repair and regeneration of the respiratory system: complexity, plasticity, and mechanisms of lung stem cell function. *Cell Stem Cell* **15**, 123–138 (2014).
- Lopez-Otin, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, 1194–1217 (2013).
- Meiners, S., Eickelberg, O. & Konigshoff, M. Hallmarks of the ageing lung. *Eur. Respir. J.* **45**, 807–827 (2015).
- Lowery, E. M., Brubaker, A. L., Kuhlmann, E. & Kovacs, E. J. The ageing lung. *Clin. Interv. Aging* **8**, 1489–1496 (2013).
- Bailey, K. L. et al. Aging causes a slowing in ciliary beat frequency, mediated by PKCepsilon. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **306**, L584–L589 (2014).
- Panda, A. et al. Human innate immunosenescence: causes and consequences for immunity in old age. *Trends Immunol.* **30**, 325–333 (2009).
- Jane-Wit, D. & Chun, H. J. Mechanisms of dysfunction in senescent pulmonary endothelium. *J. Gerontol. A Biol. Sci. Med. Sci.* **67**, 236–241 (2012).
- Brandenberger, C. & Muhlfield, C. Mechanisms of lung aging. *Cell Tissue Res.* **367**, 469–480 (2017).
- Sicard, D. et al. Aging and anatomical variations in lung tissue stiffness. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **314**, L946–L955 (2018).
- Franks, T. J. et al. Resident cellular components of the human lung: current knowledge and goals for research on cell phenotyping and function. *Proc. Am. Thorac. Soc.* **5**, 763–766 (2008).
- Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338 (2017).

12. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).
13. Han, X. et al. Mapping the mouse cell atlas by Microwell-Seq. *Cell* **173**, 1307 (2018).
14. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
15. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
16. Ziegenhain, C. et al. Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* **65**, 631–643 e634 (2017).
17. Lefrancais, E. et al. The lung is a site of platelet biogenesis and a reservoir for haematopoietic progenitor. *Nature* **544**, 105–109 (2017).
18. Heng, T. S. & Painter, M. W; Immunological Genome Project Consortium. The Immunological Genome Project: networks of gene expression in immune cells. *Nat. Immunol.* **9**, 1091–1094 (2008).
19. Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **18**, 220 (2017).
20. Mereu, E. et al. matchScore: matching single-cell phenotypes across tools and experiments. Preprint at <https://doi.org/10.1101/314831> (2018).
21. Martinez-Jimenez, C. P. et al. Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science* **355**, 1433–1436 (2017).
22. Enge, M. et al. Single-cell analysis of human pancreas reveals transcriptional signatures of aging and somatic mutation patterns. *Cell* **171**, 321–330 e314 (2017).
23. Kramer, A., Green, J., Pollard, J. Jr. & Tugendreich, S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **30**, 523–530 (2014).
24. Cox, J. & Mann, M. 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinforma.* **13**(Suppl. 16), S12 (2012).
25. Burgstaller, G. et al. The instructive extracellular matrix of the lung: basic composition and alterations in chronic lung disease. *Eur. Respir. J.* **50**, pii: 1601805 (2017).
26. Budinger, G. R. S. et al. The intersection of aging biology and the pathobiology of lung diseases: a joint NHLBI/NIA Workshop. *J. Gerontol. A Biol. Sci. Med. Sci.* **72**, 1492–1500 (2017).
27. Ricard-Blum, S. The collagen family. *Cold Spring Harb. Perspect. Biol.* **3**, a004978 (2011).
28. Ehnis, T., Dieterich, W., Bauer, M., Kresse, H. & Schuppan, D. Localization of a binding site for the proteoglycan decorin on collagen XIV (undulin). *J. Biol. Chem.* **272**, 20414–20419 (1997).
29. Kolb, M., Margetts, P. J., Sime, P. J. & Gauldie, J. Proteoglycans decorin and biglycan differentially modulate TGF-beta-mediated fibrotic responses in the lung. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **280**, L1327–L1334 (2001).
30. Yamaguchi, Y., Mann, D. M. & Ruoslahti, E. Negative regulation of transforming growth factor-beta by the proteoglycan decorin. *Nature* **346**, 281–284 (1990).
31. Wierer, M. et al. Compartment-resolved proteomic analysis of mouse aorta during atherosclerotic plaque formation reveals osteoclast-specific protein expression. *Mol. Cell. Proteom.* **17**, 321–334 (2018).
32. Schiller, H. B. et al. Deep proteome profiling reveals common prevalence of MZB1-positive plasma B cells in human lung and skin fibrosis. *Am. J. Respir. Crit. Care Med.* **196**, 1298–1310 (2017).
33. Schiller, H. B. et al. Time- and compartment-resolved proteome profiling of the extracellular niche in lung injury and repair. *Mol. Syst. Biol.* **11**, 819 (2015).
34. Kiyozumi, D., Sugimoto, N. & Sekiguchi, K. Breakdown of the reciprocal stabilization of QBRICK/Frem1, Fras1, and Frem2 at the basement membrane provokes Fraser syndrome-like defects. *Proc. Natl. Acad. Sci. USA* **103**, 11981–11986 (2006).
35. Petrou, P., Pavlakis, E., Dalezios, Y., Galanopoulos, V. K. & Chalepakis, G. Basement membrane distortions impair lung lobation and capillary organization in the mouse model for fraser syndrome. *J. Biol. Chem.* **280**, 10350–10356 (2005).
36. Cormier, S. A. et al. T(H)2-mediated pulmonary inflammation leads to the differential expression of ribonuclease genes by alveolar macrophages. *Am. J. Respir. Cell Mol. Biol.* **27**, 678–687 (2002).
37. Ruffell, D. et al. A CREB-C/EBPbeta cascade induces M2 macrophage-specific gene expression and promotes muscle injury repair. *Proc. Natl. Acad. Sci. USA* **106**, 17475–17480 (2009).
38. Gorgoni, B., Maritano, D., Marthyn, P., Righi, M. & Poli, V. C/EBP beta gene inactivation causes both impaired and enhanced gene expression and inverse regulation of IL-12 p40 and p35 mRNAs in macrophages. *J. Immunol.* **168**, 4055–4062 (2002).
39. Zhou, F. Molecular mechanisms of IFN-gamma to up-regulate MHC class I antigen processing and presentation. *Int. Rev. Immunol.* **28**, 239–260 (2009).
40. Koerberle, A., Loser, K. & Thurmer, M. Stearoyl-CoA desaturase-1 and adaptive stress signaling. *Biochim. Biophys. Acta* **1861**, 1719–1726 (2016).
41. Whitsett, J. A., Wert, S. E. & Weaver, T. E. Diseases of pulmonary surfactant homeostasis. *Annu. Rev. Pathol.* **10**, 371–393 (2015).
42. Plantier, L. et al. Activation of sterol-response element-binding proteins (SREBP) in alveolar type II cells enhances lipogenesis causing pulmonary lipotoxicity. *J. Biol. Chem.* **287**, 10099–10114 (2012).
43. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
44. Lein, E., Borm, L. E. & Linnarsson, S. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science* **358**, 64–69 (2017).
45. Moor, A. E. & Itzkovitz, S. Spatial transcriptomics: paving the way for tissue-level systems biology. *Curr. Opin. Biotechnol.* **46**, 126–133 (2017).
46. Schulz, D. et al. Simultaneous multiplexed imaging of mRNA and proteins with subcellular resolution in breast cancer tissue samples by mass cytometry. *Cell Syst.* **6**, 531 (2018).
47. Mondal, M., Liao, R., Xiao, L., Eno, T. & Guo, J. Highly multiplexed single-cell in situ protein analysis with cleavable fluorescent antibodies. *Angew. Chem. Int. Ed. Engl.* **56**, 2636–2639 (2017).
48. Hynes, R. O. Stretching the boundaries of extracellular matrix research. *Nat. Rev. Mol. Cell Biol.* **15**, 761–763 (2014).
49. Decaris, M. L. et al. Proteomic analysis of altered extracellular matrix turnover in bleomycin-induced pulmonary fibrosis. *Mol. Cell. Proteom.* **13**, 1741–1752 (2014).
50. Moliva, J. I. et al. Molecular composition of the alveolar lining fluid in the aging lung. *Age (Dordr.)* **36**, 9633 (2014).
51. Walski, M. et al. Pulmonary surfactant: ultrastructural features and putative mechanisms of aging. *J. Physiol. Pharmacol.* **60**(Suppl. 5), 121–125 (2009).
52. Alder, J. K. et al. Telomere dysfunction causes alveolar stem cell failure. *Proc. Natl. Acad. Sci. USA* **112**, 5099–5104 (2015).
53. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, (15–21) (2013).
54. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
55. Kowalczyk, M. S. et al. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res.* **25**, 1860–1872 (2015).
56. Tyanova, S. et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).
57. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11**, 319–324 (2014).
58. Thakur, S. S. et al. Deep and highly sensitive proteome coverage by LC-MS/MS without pre-fractionation. *Mol. Cell. Proteomics* **10**, M110 003699 (2011).
59. Scheltema, R. A. & Mann, M. SprayQC: a real-time LC-MS/MS quality monitoring system to maximize uptime using off the shelf components. *J. Proteome Res.* **11**, 3458–3466 (2012).
60. Michalski, A. et al. Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol. Cell. Proteomics* **10**, M111 011015 (2011).
61. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
62. Cox, J. et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
63. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
64. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
65. Misharin, A. V., Morales-Nebreda, L., Mutlu, G. M., Budinger, G. R. & Perlman, H. Flow cytometric analysis of macrophages and dendritic cell subsets in the mouse lung. *Am. J. Respir. Cell Mol. Biol.* **49**, 503–510 (2013).
66. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
67. Nagendran, M., Riordan, D. P., Harbury, P. B. & Desai, T. J. Automated cell-type classification in intact tissues by single-cell molecular profiling. *Elife* **7**, pii: e30510 (2018).

## Acknowledgements

We thank Silvia Weidner and Daniela Dietel for excellent technical assistance. We also thank Gabi Sowa, Igor Paron and Korbinian Mayr for expert support of the proteomics pipeline. We thank Sandy Lösecke for technical assistance in next generation sequencing and Thomas Schwarzmayr for support with High-seq 4000 sequencing raw data. L.M.S. acknowledges funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 753039. This work was supported by the German Center for Lung Research (DZL), the Helmholtz Association, and the German Federal Ministry of Education and Research (BMBF), project Single Cell Genomics Network Germany.



## Author contributions

H.B.S. conceptualized and supervised the entire project and wrote the paper. F.J.T. supervised single-cell analysis and multi-omics data integration. I.A. and M.S. performed single-cell transcriptomics experiments. L.M.S., I.A., M.A., and H.B.S. analyzed single-cell transcriptomics data. H.B.S. performed proteomics experiments. H.B.S. and L.M.S. analyzed the proteomics data and performed transcriptomics and proteomics data integration. I.A. and C.H.M. performed histology and immunofluorescence microscopy. I.E.F. and F.R.G. performed flow cytometry experiments. M.N. and T.D. performed PLISH experiments. E.G. and T.-M.S. performed next-generation sequencing of single-cell libraries. L.M.S. and G.T. set up the interactive webtool. O.E. assisted in data interpretation and M.M. provided important support with mass spectrometry equipment. All authors read, edited, and approved the manuscript.

## Additional information

**Supplementary Information** accompanies this paper at <https://doi.org/10.1038/s41467-019-08831-9>.

**Competing interests:** The authors declare no competing interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

**Journal peer review information:** *Nature Communications* thanks G. R. Scott Budinger and the other anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019



# An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics

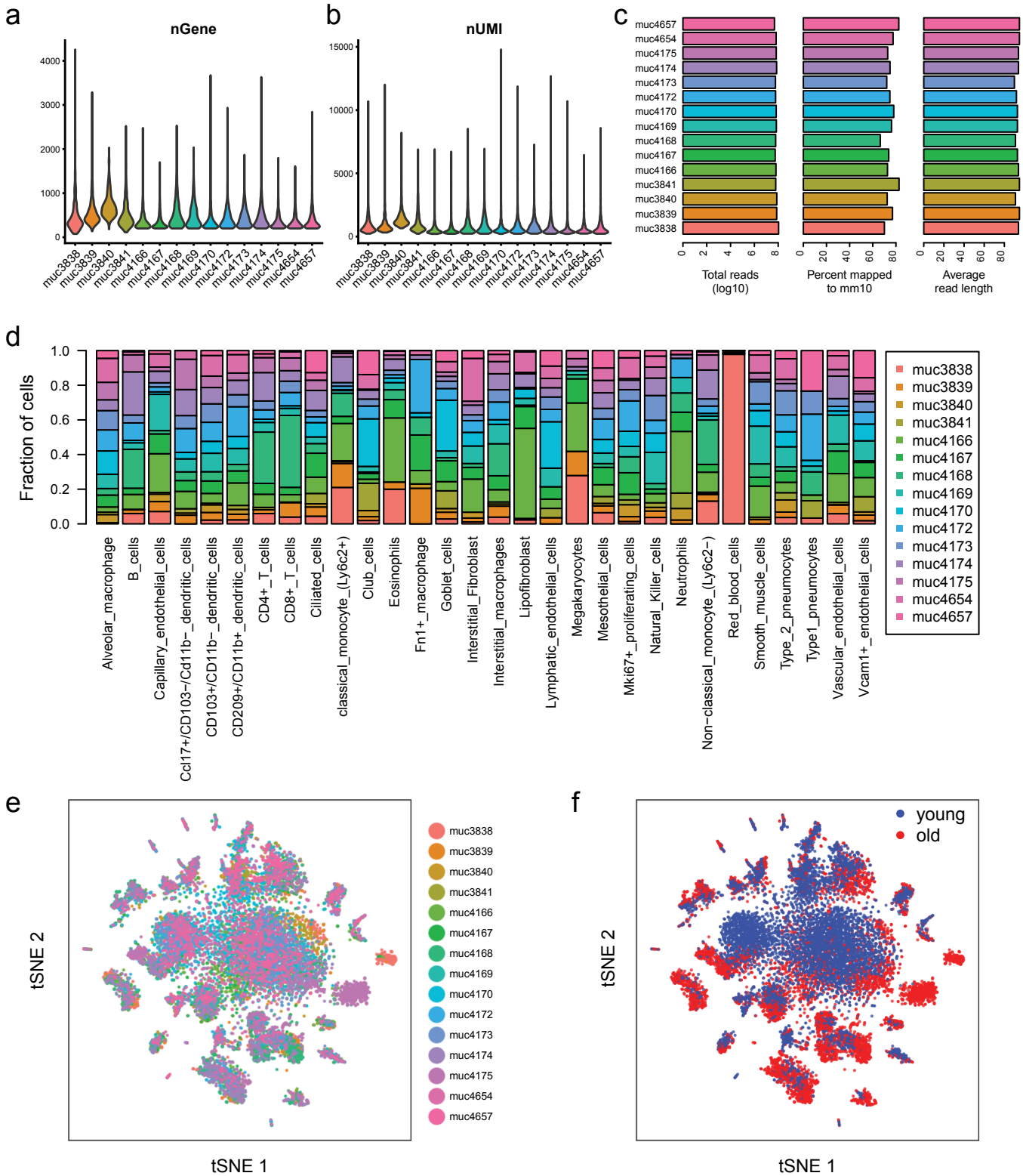
Ilias Angelidis<sup>1\*</sup>, Lukas M. Simon<sup>2\*</sup>, Isis E. Fernandez<sup>1</sup>, Maximilian Strunz<sup>1</sup>, Christoph H. Mayr<sup>1</sup>, Flavia R. Greiffo<sup>1</sup>, George Tsitsiridis<sup>2</sup>, Meshal Ansari<sup>1,2</sup>, Elisabeth Graf<sup>3</sup>, Tim-Matthias Strom<sup>3</sup>, Monica Nagendran<sup>4</sup>, Tushar Desai<sup>4</sup>, Oliver Eickelberg<sup>5</sup>, Matthias Mann<sup>6</sup>, Fabian J. Theis<sup>2,7,#</sup>, and Herbert B. Schiller<sup>1,#</sup>

1. Helmholtz Zentrum München, Institute of Lung Biology and Disease, Member of the German Center for Lung Research (DZL), Munich, Germany
2. Helmholtz Zentrum München, Institute of Computational Biology, Munich, Germany
3. Helmholtz Zentrum München, Institute of Human Genetics, Munich, Germany
4. Department of Internal Medicine, Division of Pulmonary and Critical Care, Stanford University School of Medicine, Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Stanford, United States of America
5. University of Colorado, Department of Medicine, Division of Respiratory Sciences and Critical Care Medicine, Denver, CO, USA
6. Max Planck Institute of Biochemistry, Department of Proteomics and Signal Transduction, Martinsried, Germany
7. Department of Mathematics, Technische Universität München, Munich, Germany

\*... these authors contributed equally to this work

# ... correspondence to Fabian J. Theis (fabian.theis@helmholtz-muenchen.de)  
and Herbert B. Schiller (herbert.schiller@helmholtz-muenchen.de)

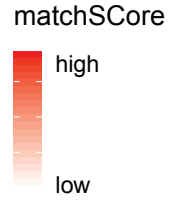
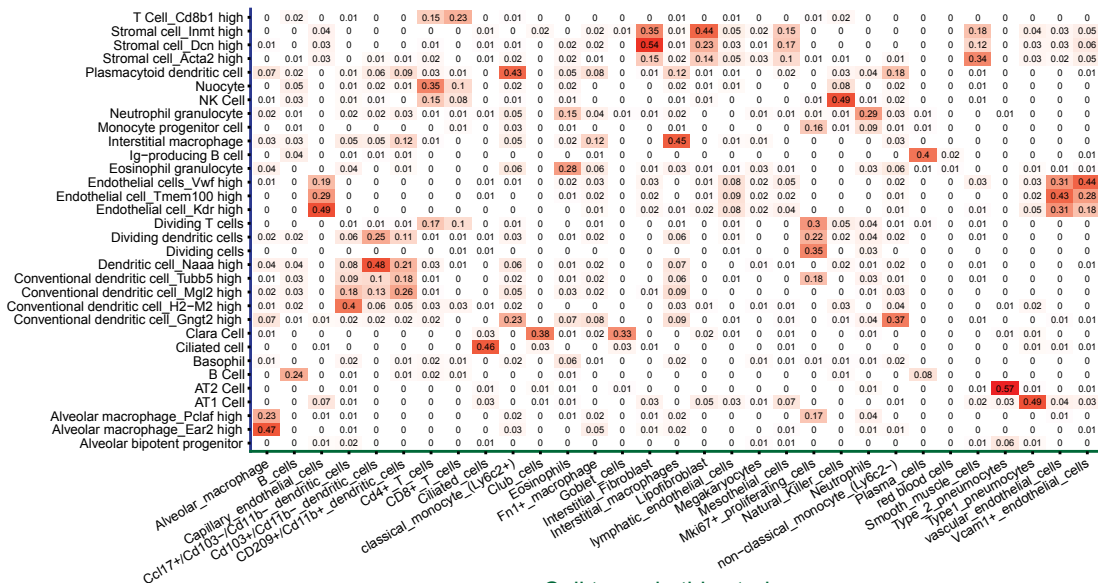
## Supplementary Figures



**Supplementary Figure 1. High technical reproducibility enables integration of the 15 mouse experiments.** (a, b) The violin plots show the distribution of the (a) number of genes detected per cell and (b) total UMI counts per cell across mice, respectively. (c) scRNA-seq alignment statistics show comparable values across mice. (d) Cell type identity and the fraction of cells per mouse are shown on the X and Y axes respectively. (e, f) tSNE visualization colored by (e) mouse sample and (f) age group.

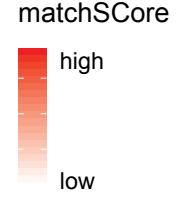
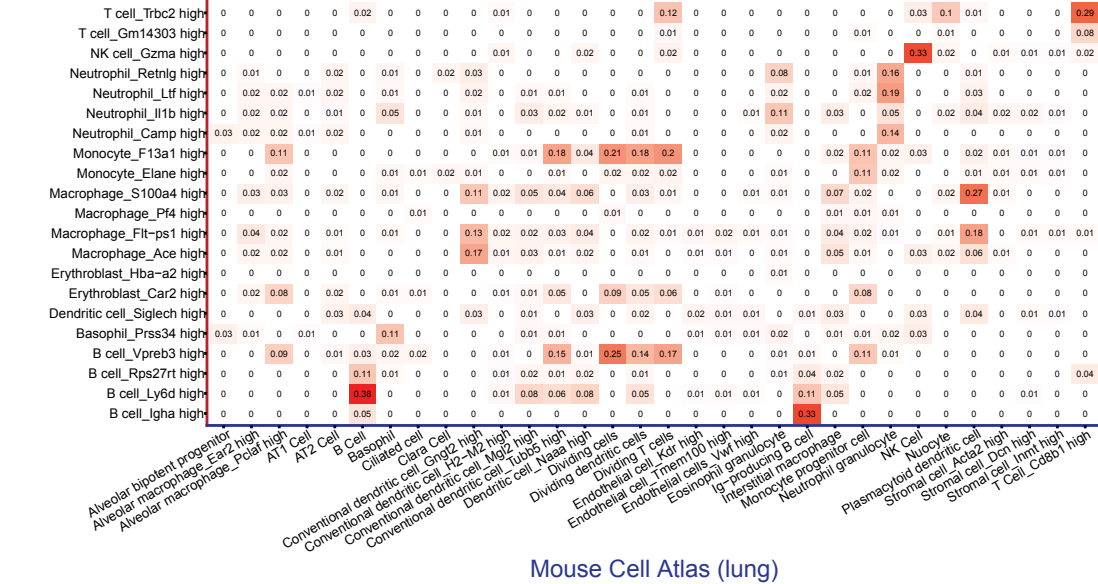
a

Mouse Cell Atlas (lung)



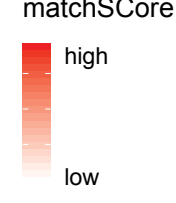
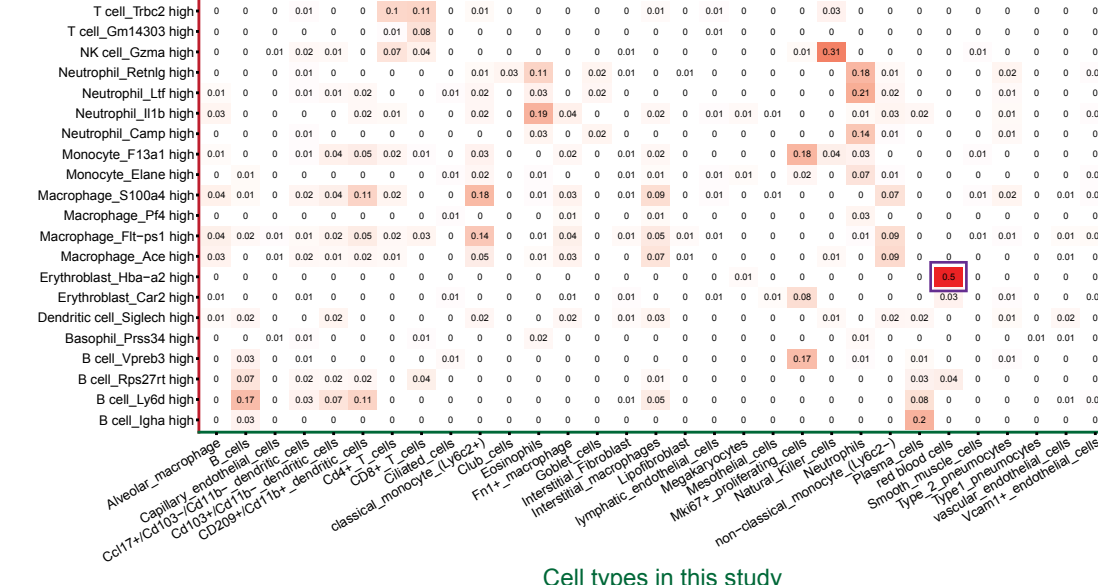
b

Mouse Cell Atlas (peripheral blood)



c

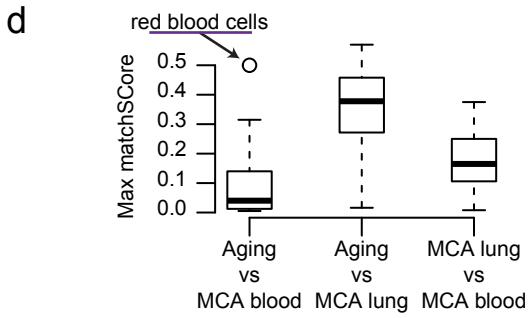
Mouse Cell Atlas (peripheral blood)



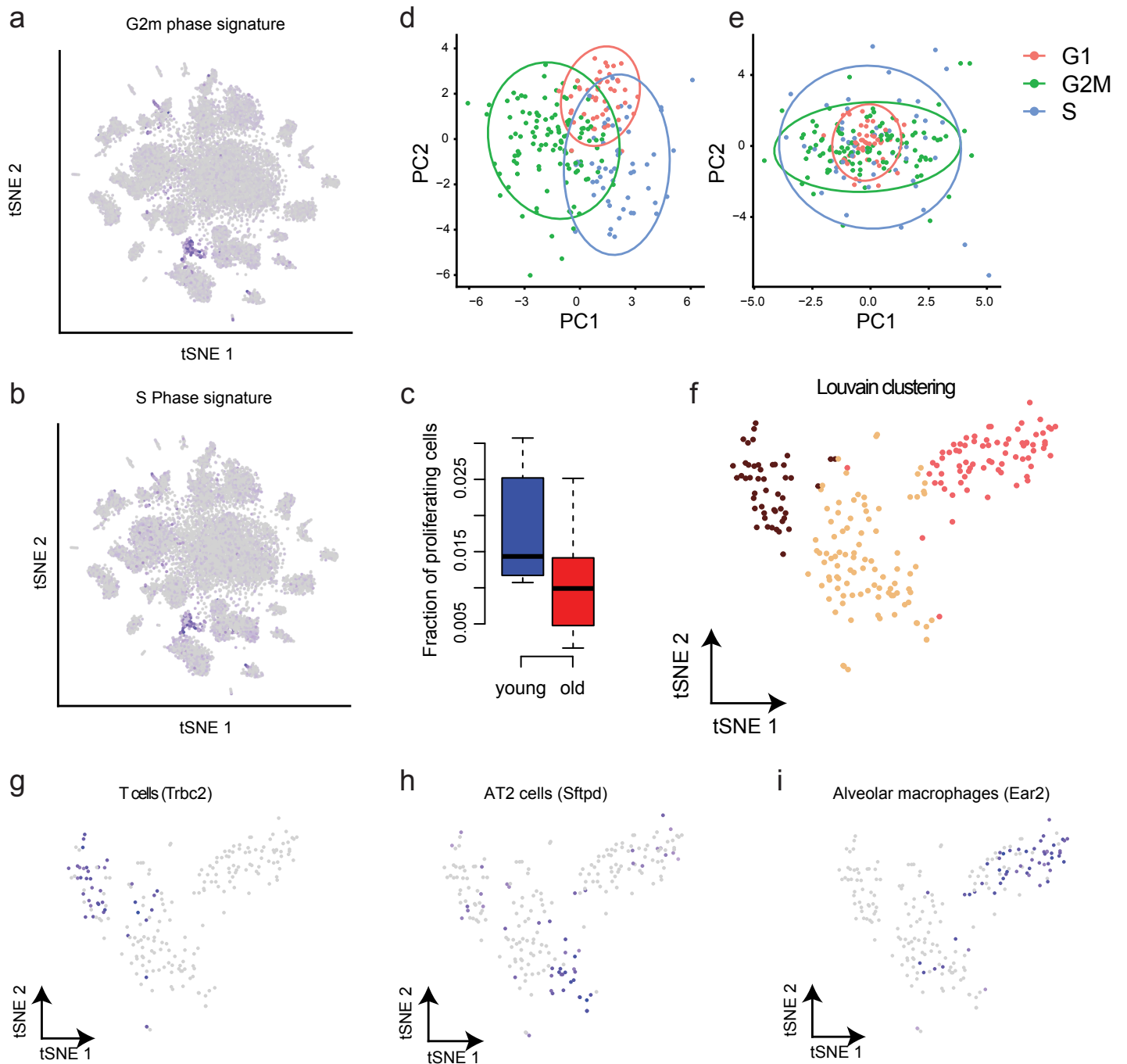
Cell types in this study

Mouse Cell Atlas (lung)

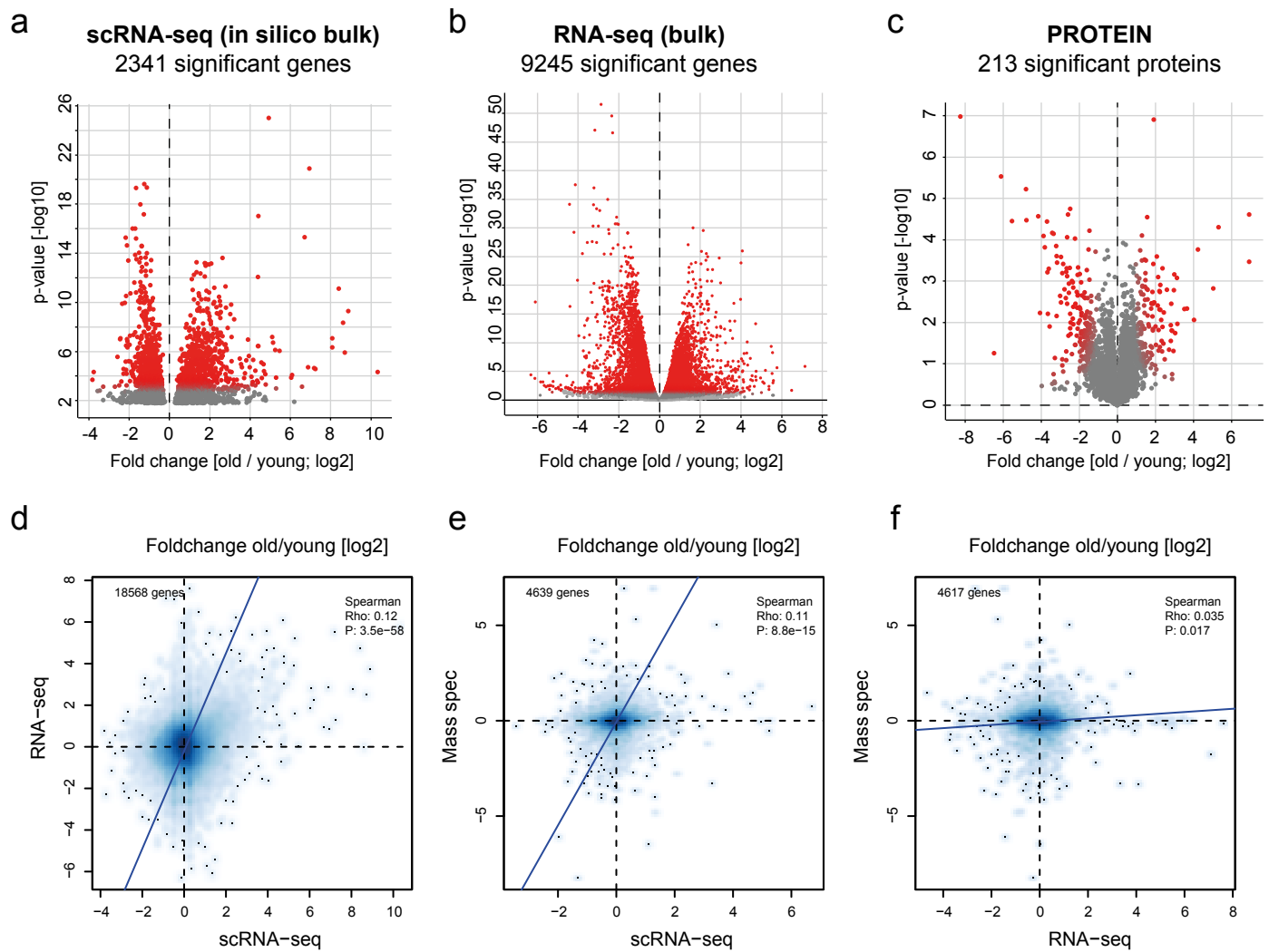
Cell types in this study



**Supplementary Figure 2. Comparison with the Mouse Cell Atlas validates lung cell identities.** (a-c) The matchSC score comparison between the clusters in this study, the MCA lung and peripheral blood signatures is shown. Red and white colors indicate high and low matchSC scores, respectively. The outlier in panel c represents red blood cells (purple rectangle). (d) The box plot shows the distribution of maximal matchSC scores for each cluster across the comparisons between these three data sets. The box represents the interquartile range, the horizontal line in the box is the median, and the whiskers represent 1.5 times the interquartile range. The outlier in the comparison between cell types in this study and the MCA blood data represents red blood cells (underlined in purple).

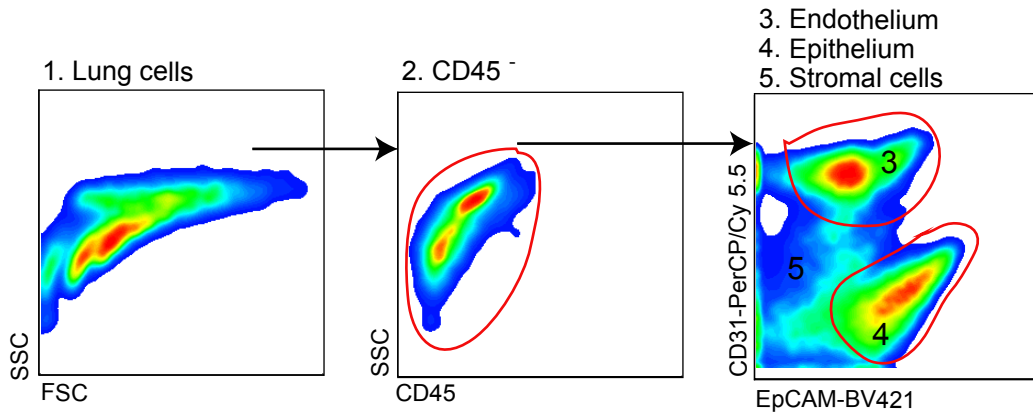


**Supplementary Figure 3. Cell-cycle analysis reveals reduced proliferative capacity of T cells, Alveolar macrophages and Type-2 pneumocytes in aged lungs.** (a, b) The `Mki67+ proliferating cell' cluster (Fig. 1) showed high expression of (a) G2M- and (b) S-phase cell cycle signatures. (c) A higher fraction of proliferating cells was observed in young (n = 8 animals) compared to old (n = 7 animals) mice. The box represents the interquartile range, the horizontal line in the box is the median, and the whiskers represent 1.5 times the interquartile range. (d) PCA based on cell cycle marker genes revealed clustering by cell cycle phase and (e) the removal of this effect after regressing out the cell cycle effect. Cells are colored by cell cycle phase as assigned by Seurat. (f) Unsupervised Louvain clustering revealed three distinct cell clusters. (g-i) tSNE visualization colored by the expression of cell type marker genes (g) Trbc2, (h) Sftpd and (i) Ear2 corresponding to T cells, Type 2 pneumocytes and alveolar macrophages, respectively.

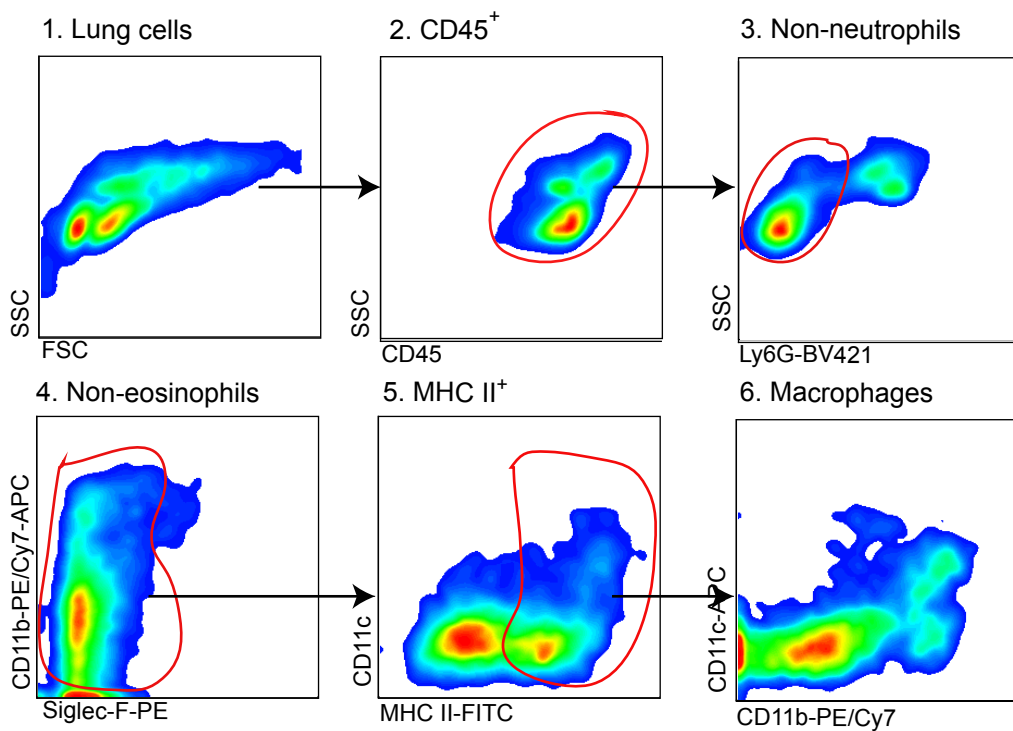


**Supplementary Figure 4. Multi-omics lung aging data displays significant correspondence.** Volcano plots show the significantly regulated genes from (a) in from scRNA-seq, (b) bulk RNA-seq and (c) mass spectrometry. (d-f) Differential expression results from multi-omics experiments show significant correspondence. X and Y axes illustrate the log2 fold changes calculated from the (d) RNA-seq and scRNA-seq (in silico bulk) experiments, (e) the mass spectrometry (protein) and scRNA-seq (in silico bulk) experiments, and (f) the mass spectrometry (protein) and RNA-seq experiments. Blue line indicates the Deming regression fit. Black dotted horizontal and vertical lines indicate 0 values (no differential expression) for the in silico bulk and protein data, respectively.

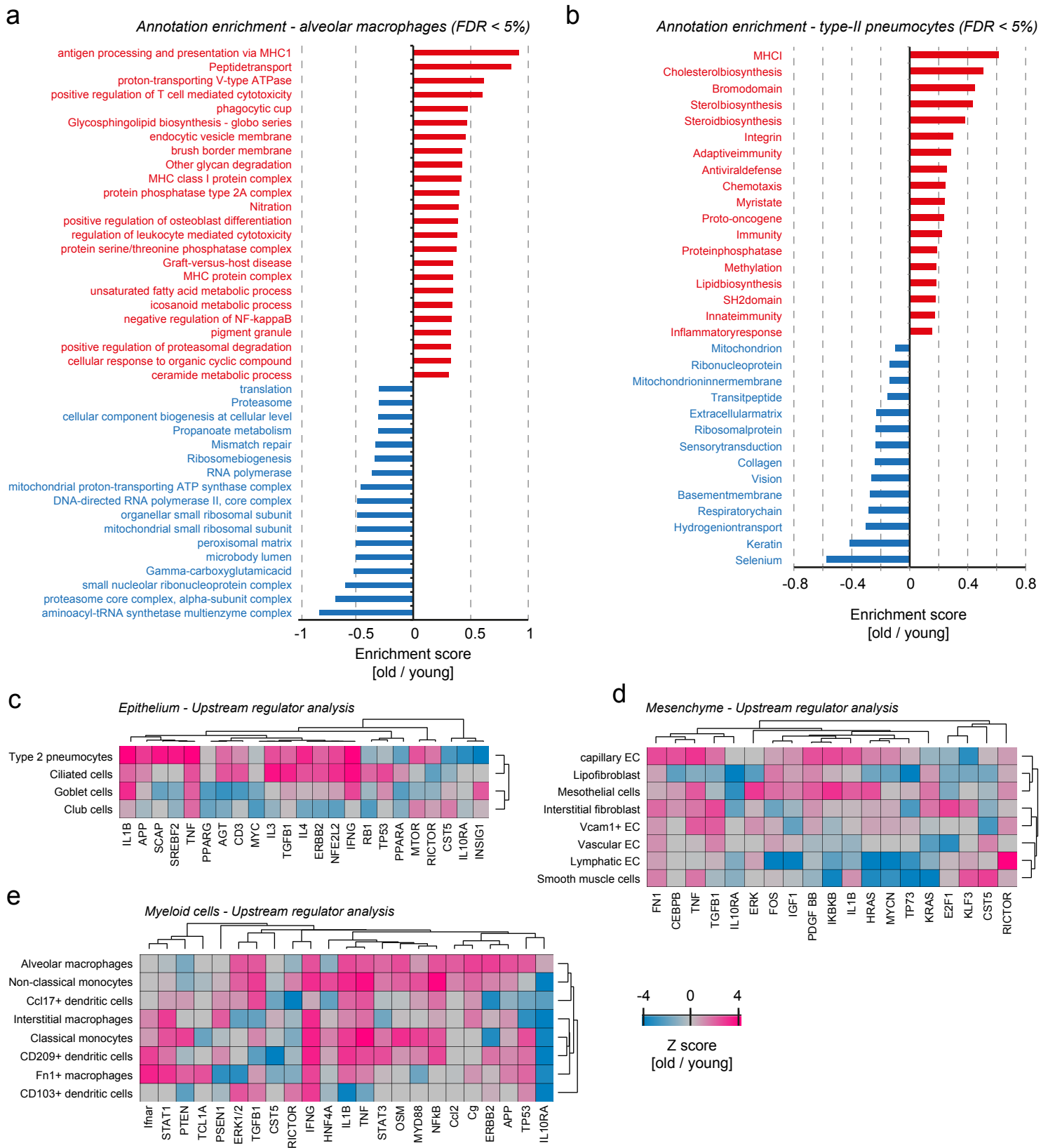
a Lung cells FACS analysis



b Lung macrophages sorting strategy



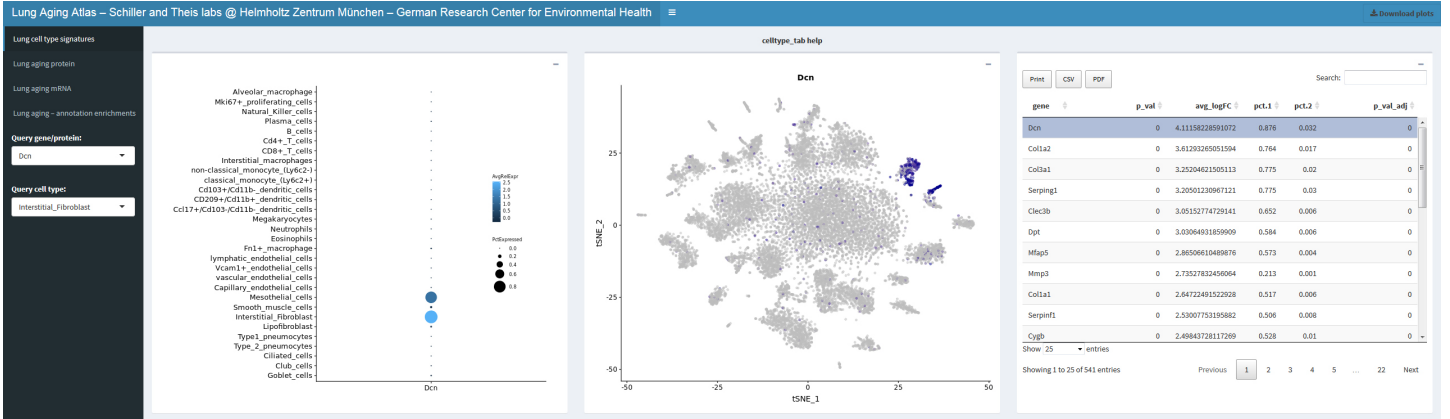
**Supplementary Figure 5. scRNA-seq data validation using bulk RNA-seq of flow sorted cell populations from young and old mice.** Cells were sorted by using the (a) CD45 negative fraction of the cell isolate stained for anti-mouse CD31, and EpCAM antibodies. Epithelial cells were sorted as CD31- cells and EpCAM+ cells (a-4). For sorting macrophages we used the (b) CD45 positive fraction and stained with anti-mouse CD11c, CD11b, MHC II, Siglec-F and Ly6G antibodies. For flow cytometry sorting, neutrophils were excluded by selection of Ly6G negative cells. Macrophages were sorted as MHCII+, CD11c+,CD11b+.



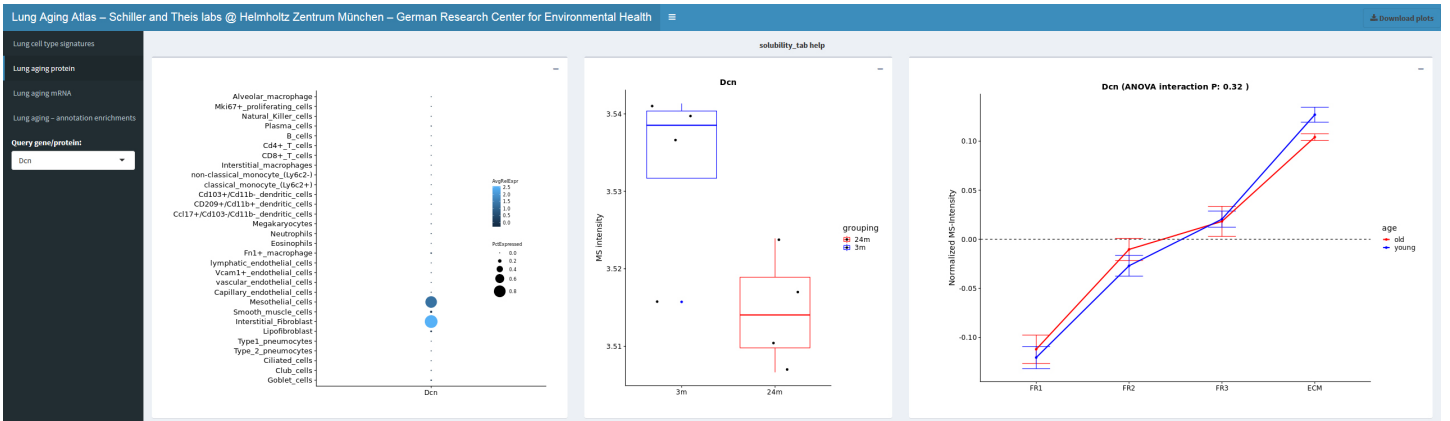
**Supplementary Figure 6. Pathway and upstream regulator analysis reveals cell type specific effects of aging.** (a, b) The bar graph shows the result of a gene annotation enrichment analysis for (a) alveolar macrophages, and (b) type-II pneumocytes, respectively. Gene categories with positive (upregulated in old) and negative scores (downregulated in old) are highlighted in red and blue respectively. (c-e) Upstream regulators are predicted based on the observed gene expression changes for (c) epithelial, (d) mesenchymal, and (e) myeloid cells. Cell types and regulators were grouped by unsupervised hierarchical clustering (Pearson correlation) and the indicated transcriptional regulators and cytokines, growth factors and ECM proteins are color coded based on the activation score as shown.



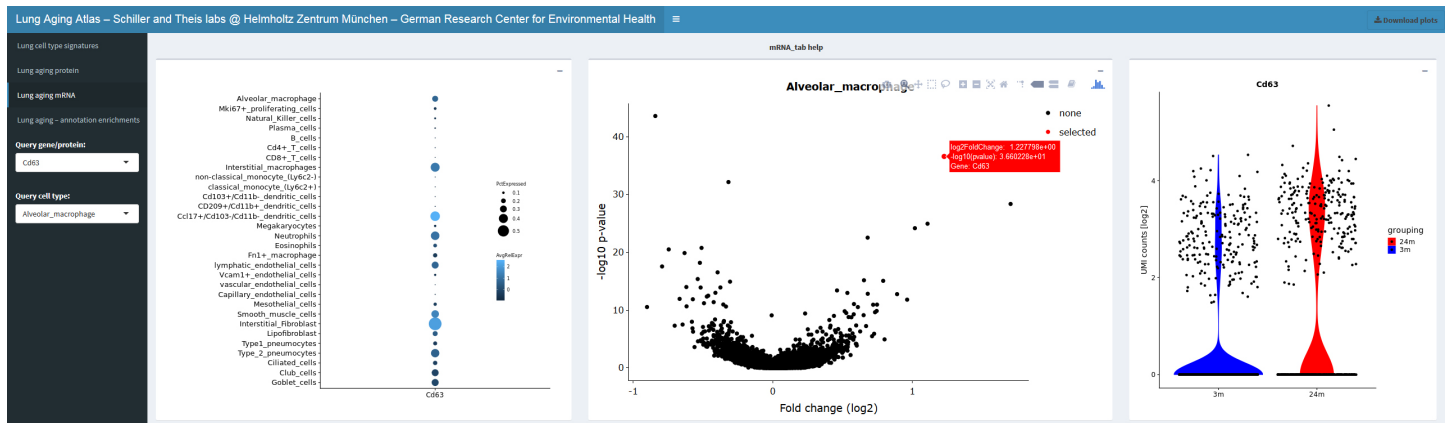
**a Lung cell type signatures - cell type and gene specific queries, list of most significant marker genes for 30 cell types**



**b Lung aging protein - cellular source, protein abundance, and protein solubility**



**c Lung aging mRNA - cellular source, mRNA volcano plot [old/young], and UMI counts**



**d Lung aging - annotation enrichments and pathways**

Cell type	Type	Name	Score	Benj. Hoch. FDR
Alveolar_macrophage	GOBP name	antigen processing and presentation of endogenous antigen	0.921843739	0.00772115
Alveolar_macrophage	GOBP name	antigen processing and presentation of endogenous peptide antigen	0.921843739	0.007487126
Alveolar_macrophage	GOBP name	antigen processing and presentation of endogenous peptide antigen via MHC class I	0.914893929	0.05427432
Alveolar_macrophage	Keywords	Peptide transport	0.84951033	0.061790985
Alveolar_macrophage	GOBP name	regulation of T cell mediated cytotoxicity	0.62760102	0.112333456
Alveolar_macrophage	GOCC name	proton-transporting V-type ATPase, V1 domain	0.610859501	0.071893691
Alveolar_macrophage	GOBP name	positive regulation of T cell mediated cytotoxicity	0.600715884	0.046929143
Alveolar_macrophage	GOCC name	phagocytic cup	0.472170022	0.093519296
Alveolar_macrophage	KEGG name	glycophingolipid biosynthesis - globbo series	0.469584796	0.08068099
Alveolar_macrophage	GOCC name	endocytic vesicle membrane	0.454083233	0.001432512
Alveolar_macrophage	GOCC name	phagocytic vesicle membrane	0.454083233	0.001337011
Alveolar_macrophage	KEGG name	Collecting duct acid secretion	0.444581877	0.13627242
Alveolar_macrophage	GOCC name	brush border membrane	0.424513369	0.031236554
Alveolar_macrophage	KEGG name	Other glycan degradation	0.423316648	0.050696909
Alveolar_macrophage	GOCC name	MHC class I protein complex	0.417174324	0.03723089

**Supplementary Figure 7. A user friendly and interactive webtool enables navigating the Lung Aging Atlas.**


(a) The first tab 'Lung cell type signatures' provides a cell type dotplot (left panel) and a color coded tSNE map (middle panel) for gene specific queries and illustrates cell type specific expression of any gene of interest. A cell type query produces a list of top marker genes for the cell type of interest (right panel). (b) The panel 'Lung aging protein' features a dot plot to illustrate the most likely cellular source of the protein of interest (left panel), a box plot to show alterations in total lung tissue protein abundance in old mice (middle panel), and a line plot to show protein solubilities. Protein solubility is measured by relative quantification of protein abundance across four fractions. Fraction 1 (FR1) contains proteins with highest and fraction 4 (ECM) with lowest solubility. Curves that peak on the right (ECM) thus represent insoluble proteins. (c) The tab 'Lung aging mRNA' again features the dotplot (left panel), a volcano plot that shows fold changes [old/young] on the x-axis and  $-\log_{10}$  p-values on the y-axis (middle panel), and a violin plot of the  $\log_2$  UMI counts illustrating mRNA abundance in young and old mice. The dot and violin plot are navigated with the gene specific query, while the volcano plot requires navigation via the cell type query. The volcano plot has a toggle over function that allows identification of genes and can thus be used to browse through differential gene expression between young and old cells of any cell type of interest. (d) In the tab 'Lung aging - annotation enrichments', the gene annotation enrichments between old and young can be browsed for all 3 cell types.

---

## 4. Paper II

Strunz, M., Simon, L.M., Ansari, M., Kathiriya, J.J., Angelidis, I., Mayr, C.H., Tsidiridis, G., Lange, M., Mattner, L.F., Yee, M. and Ogar, P., 2020. Alveolar regeneration through a Krt8+ transitional stem cell state that persists in human lung fibrosis. *Nature communications*, 11(1), pp.1-20.

# Alveolar regeneration through a Krt8+ transitional stem cell state that persists in human lung fibrosis

Maximilian Strunz<sup>1,14</sup>, Lukas M. Simon<sup>2,3,14</sup>, Meshal Ansari <sup>1,2</sup>, Jaymin J. Kathiriya <sup>4</sup>, Ilias Angelidis <sup>1</sup>, Christoph H. Mayr<sup>1</sup>, George Tsidiridis<sup>2</sup>, Marius Lange <sup>2,5</sup>, Laura F. Mattner<sup>1</sup>, Min Yee<sup>6</sup>, Paulina Ogar<sup>1</sup>, Arunima Sengupta<sup>1</sup>, Igor Kukhtevich<sup>7</sup>, Robert Schneider<sup>7</sup>, Zhongming Zhao <sup>3</sup>, Carola Voss <sup>1</sup>, Tobias Stoeger <sup>1</sup>, Jens H. L. Neumann<sup>8</sup>, Anne Hilgendorff<sup>1,9</sup>, Jürgen Behr<sup>1,10,11</sup>, Michael O'Reilly<sup>6</sup>, Mareike Lehmann <sup>12</sup>, Gerald Burgstaller <sup>1</sup>, Melanie Königshoff<sup>12,13</sup>, Harold A. Chapman <sup>4</sup>, Fabian J. Theis <sup>2,5</sup>✉ & Herbert B. Schiller <sup>1</sup>✉

The cell type specific sequences of transcriptional programs during lung regeneration have remained elusive. Using time-series single cell RNA-seq of the bleomycin lung injury model, we resolved transcriptional dynamics for 28 cell types. Trajectory modeling together with lineage tracing revealed that airway and alveolar stem cells converge on a unique Krt8+ transitional stem cell state during alveolar regeneration. These cells have squamous morphology, feature p53 and NFκB activation and display transcriptional features of cellular senescence. The Krt8+ state appears in several independent models of lung injury and persists in human lung fibrosis, creating a distinct cell-cell communication network with mesenchyme and macrophages during repair. We generated a model of gene regulatory programs leading to Krt8+ transitional cells and their terminal differentiation to alveolar type-1 cells. We propose that in lung fibrosis, perturbed molecular checkpoints on the way to terminal differentiation can cause aberrant persistence of regenerative intermediate stem cell states.

<sup>1</sup>Institute of Lung Biology and Disease and Comprehensive Pneumology Center with the CPC-M bioArchive, Helmholtz Zentrum Muenchen, Member of the German Center for Lung Research (DZL), Munich, Germany. <sup>2</sup>Institute of Computational Biology, Helmholtz Zentrum München, Munich, Germany. <sup>3</sup>Center for Precision Health, School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX, USA. <sup>4</sup>Biomedical Center, University of California San Francisco, San Francisco, CA, USA. <sup>5</sup>Department of Mathematics, Technische Universität München, Munich, Germany. <sup>6</sup>Department of Pediatrics, University of Rochester, Rochester, NY, USA. <sup>7</sup>Institute of Functional Epigenetics, Helmholtz Zentrum München, Munich, Germany. <sup>8</sup>Institute of Pathology, Ludwig Maximilians University Hospital Munich, Munich, Germany. <sup>9</sup>Member of the German Center for Lung Research (DZL), Center for Comprehensive Developmental Care (CDeCLMU), Department of Neonatology, Perinatal Center Grosshadern, Hospital of the Ludwig-Maximilians University (LMU), Munich, Germany. <sup>10</sup>Member of the German Center for Lung Research (DZL), Department of Internal Medicine V, Ludwig Maximilians University Hospital (LMU) Munich, Munich, Germany. <sup>11</sup>Asklepios Fachkliniken in Munich-Gauting, Munich, Germany. <sup>12</sup>Comprehensive Pneumology Center (CPC), Research Unit Lung Repair and Regeneration, Helmholtz Zentrum München, Member of the German Center for Lung Research (DZL), Munich, Germany. <sup>13</sup>University of Colorado, Department of Pulmonary Sciences and Critical Care Medicine, Denver, CO, USA. <sup>14</sup>These authors contributed equally: Maximilian Strunz, Lukas M. Simon. ✉email: [fabian.theis@helmholtz-muenchen.de](mailto:fabian.theis@helmholtz-muenchen.de); [herbert.schiller@helmholtz-muenchen.de](mailto:herbert.schiller@helmholtz-muenchen.de)

Lung disease is a major health burden accounting for one in six deaths globally<sup>1</sup>. The lung's large surface area is exposed to a great variety of environmental and microbial insults causing injuries to its epithelium that require a stem cell driven regenerative response. Lineage tracing studies revealed that depending on the location within the lung and the severity of injury, different stem cell populations can be engaged<sup>2–4</sup>. The cell-intrinsic properties and niche signals driving these processes are not well understood and likely involve tight spatiotemporal control of crosstalk with the various immune and mesenchymal cell types that are activated or recruited after injury<sup>5,6</sup>. Importantly, many of the functionally relevant cell states appear transiently after injury. For instance, the conversion of fibroblasts to myofibroblasts during fibrogenesis has been shown to be reversible<sup>7</sup>. Similarly, the recruitment of monocytes early after injury results in a continuum of macrophage states that evolve as their microenvironment changes over time<sup>8</sup>. This implies that functionally important cell states are limited in time and space by yet to be resolved regulatory mechanisms.

In the alveolar compartment, gas exchange is enabled by ultra-thin extensions of alveolar type-1 pneumocytes (AT1) forming the alveolar surface area. The surfactant-producing cuboidal alveolar type-2 pneumocytes (AT2) have been shown to act as alveolar stem cells by self-renewing and differentiating into squamous AT1 cells, during both homeostatic turnover and injury<sup>9</sup>. In very severe cases of injury with massive loss of AT2 cells, both AT1 and AT2 cells can be replenished by airway-derived stem cell populations<sup>10–13</sup>. The molecular details and spatiotemporal organization of such decisive signals, gene programs and pathways during recovery of the AT1 cell layer have not been resolved.

Using single-cell RNA sequencing (scRNAseq) methods it is now possible to predict the future state of individual cells based on RNA velocity<sup>14</sup> and model cell fate trajectories in pseudotime<sup>15,16</sup>. These methods are highly complementary with traditional lineage tracing and longitudinal single-cell analysis of a dynamic system<sup>17</sup>, combined with computational methods is unbiased and allows for discovery in high-throughput. Furthermore, the dynamics of cell–cell communication networks can be computationally approximated from scRNAseq datasets by the integration of receptor–ligand databases<sup>18,19</sup>. Here, we ask if we can leverage these ideas for the problem of gene regulation during epithelial regeneration.

We chart the cell type specific gene expression trajectories in whole-lung single-cell suspensions after bleomycin induced lung injury to provide a resource of the gene expression dynamics and routes of cell–cell communication during regeneration after bleomycin induced lung injury. In this analysis we discover an intermediate alveolar epithelial cell state forming a unique cellular niche that peaks in frequency during the fibrogenic phase of tissue repair together with the appearance of myofibroblasts and M2-macrophages. Using high resolution pseudotime modeling and lineage tracing we demonstrate transcriptional convergence of airway and alveolar stem cells into the transitional stem cell state and reveal candidate transcriptional regulators. Disease relevance of the regenerative intermediate stem cell state described in this work is emerging from our observation that this cell state accumulates and persists in lung fibrosis.

## Results

**A time-resolved single cell analysis of lung regeneration.** To comprehensively chart the cellular dynamics of all major cell lineages during regeneration after bleomycin-mediated acute lung injury, we collected whole-organ single cell suspensions from six time points after injury (day 3, 7, 10, 14, 21, and 28) and

uninjured control lungs (PBS) with on average four replicate mice per time point. Using the Dropseq workflow<sup>20</sup>, we generated single cell transcriptomes from ~1000 cells per individual mouse, resulting in a final data set with 29,297 cells after quality control filtering.

Single cell transcriptional profiles were visualized in two dimensions using the Uniform Manifold Approximation and Projection (UMAP) method<sup>21</sup> (Fig. 1a). We identified 26 cell type identities that were manually annotated using canonical marker genes and previously published scRNAseq datasets of the mouse lung<sup>22,23</sup> (Supplementary Fig. 1a, b; Supplementary Data 1). Most cell clusters contained cells from both conditions (Supplementary Fig. 1c) and we found good reproducibility of quality metrics across samples (Supplementary Fig. 1d). Linear discriminant analysis confirmed good agreement of cell type frequencies between conditions with 93% accuracy, demonstrating high replicability of the mouse replicates (Supplementary Fig. 1e–g, and Supplementary Fig. 2).

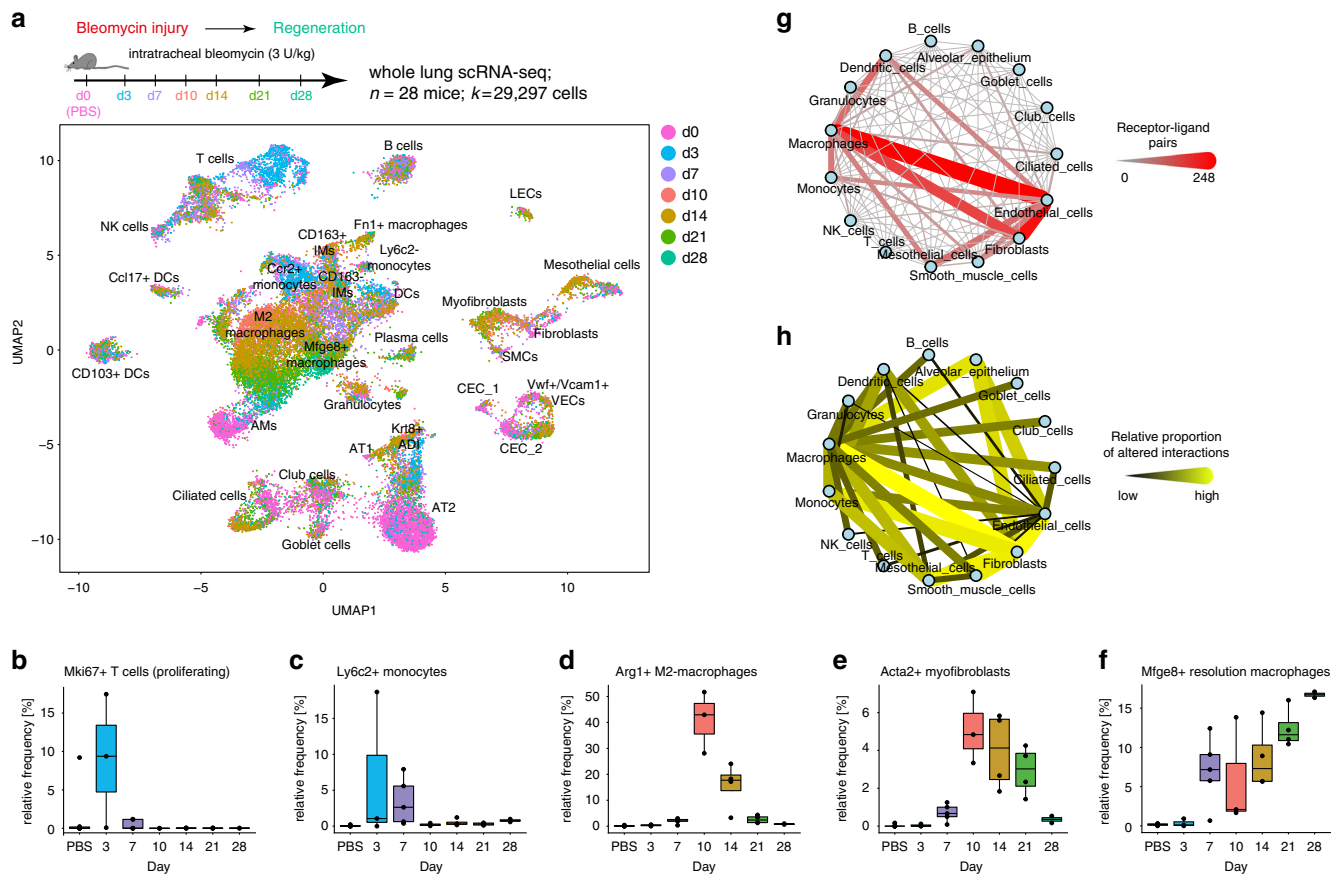
Cell frequency dynamics showed an expansion of T cells early after injury at day 3 (Fig. 1b), recruitment of Ly6c2+ monocytes from blood within the first weeks after injury (Fig. 1f), the appearance of Arg1+M2-macrophages peaking at day 10 (Fig. 1c), transient formation of Acta2+ myofibroblasts (Fig. 1d; Supplementary Fig. 3), and appearance of a Mfge8+ resolution macrophage state peaking at day 28 (Fig. 1e). Subclustering macrophages revealed several distinct phenotypes at different time points (Supplementary Fig. 4a–c). Previously published bulk RNAseq signatures from lineage tracing of monocyte-derived macrophages in the bleomycin model<sup>8</sup> were used to score individual cells, revealing that recruited monocytes give rise to several different macrophage identities (Supplementary Fig. 4d–g).

A total of 6660 genes showed significant changes after injury in at least one cell type (FDR < 0.1, Supplementary Data 2). The results of this analysis can be interactively explored with our webtool at [github.com/theislab/LungInjuryRegeneration](https://github.com/theislab/LungInjuryRegeneration), which provides a user-friendly resource of gene expression changes in the whole lung during injury repair.

We constructed a putative cell–cell communication network by mapping known receptor–ligand pairs across cell types (see “Methods” for details) (Fig. 1g), and integrated longitudinal expression dynamics of receptor–ligand pairs. This analysis revealed considerable alterations in possible communication routes between macrophages and fibroblasts, as well as striking differences in communication of these cell types with the alveolar epithelium (Fig. 1h).

To validate the scRNAseq data we performed extensive comparisons with our previously published bulk RNA-seq and proteomics data from day 14 after bleomycin treatment<sup>24</sup>. We generated in silico bulk samples by summing counts across all cells of each mouse replicate. Significant correlations were observed across all three modalities (Fig. 2a, b), and samples clustered by data modality but also injury status, cross-validating the global injury-induced expression changes (Fig. 2c). Interestingly, the shared bleomycin induced features across data modalities mostly showed cell type specific expression in the alveolar epithelium, fibroblasts and macrophages (Fig. 2d), with a peak at day 10 and resolving during the regeneration time course (Fig. 2e). To validate changes in cell type frequency observed at the cellular level, we performed bulk deconvolution analysis, testing for enrichment of cell type marker signatures in the bulk RNA-seq data. This analysis revealed cell types and states with significantly increased frequency after bleomycin injury (Fig. 2f, g).

**Unique squamous Krt8+ cells in alveolar regeneration.** One of the clusters with significantly enriched frequency after injury



**Fig. 1** Longitudinal single cell RNA-seq reveals cell state and cell communication dynamics. **a** Single cell suspensions from whole-mouse lungs were analyzed using scRNAseq at the indicated time points after bleomycin-mediated lung injury. The color code in the UMAP embedding shows shifts of the indicated cell types in gene expression space during the regeneration time course. **b–f** Relative frequency of the indicated cell types relative to all other cells was calculated for individual mice at the indicated time points after injury ( $n = 4$ ) and for PBS treated control mice ( $n = 7$ ). The boxes represent the interquartile range, the horizontal line in the box is the median, and the whiskers represent 1.5 times the interquartile range. **g** The network shows 15 meta-cell type identities (see Supplementary Fig. 1d) and their putative communication structure. Edge weight and color illustrate the number of receptor-ligand pairs between cell types. **h** The edges represent the relative proportion of receptor-ligand pairs between cell types with altered expression after injury.

represented a so far undescribed cell state in the alveolar epithelium, marked by high expression of Keratin-8 (Krt8) and a highly distinct set of genes. Subclustering of alveolar epithelial cells resulted in four distinct clusters (Fig. 3a), which largely represented different time points (Fig. 3b). Notably, AT1 and AT2 cells were connected by cells mainly derived from intermediate time points. We identified AT2 cells marked by *Sftpc* expression, and an activated AT2 state marked by injury-induced genes, such as *Lcn2* and *Il33* (Fig. 3d, e). The *Krt8+* cells showed some transcriptional similarity to AT1 cells, however, were clearly distinct and did not highly express the canonical marker genes for AT2 and AT1 (Fig. 3d, e). To analyze a possible transition of AT2 cells to these cells we used scVelo (see Methods for details) which uses the ratio of spliced to unspliced reads to infer RNA velocities<sup>14</sup> and computationally predicts the future state of individual cells. This RNA velocity analysis suggested that alveolar *Krt8* high cells were derived from activated AT2 cells and might give rise to AT1 cells (Fig. 3c, d). Thus, we named the cell state *Krt8+* alveolar differentiation intermediate (ADI).

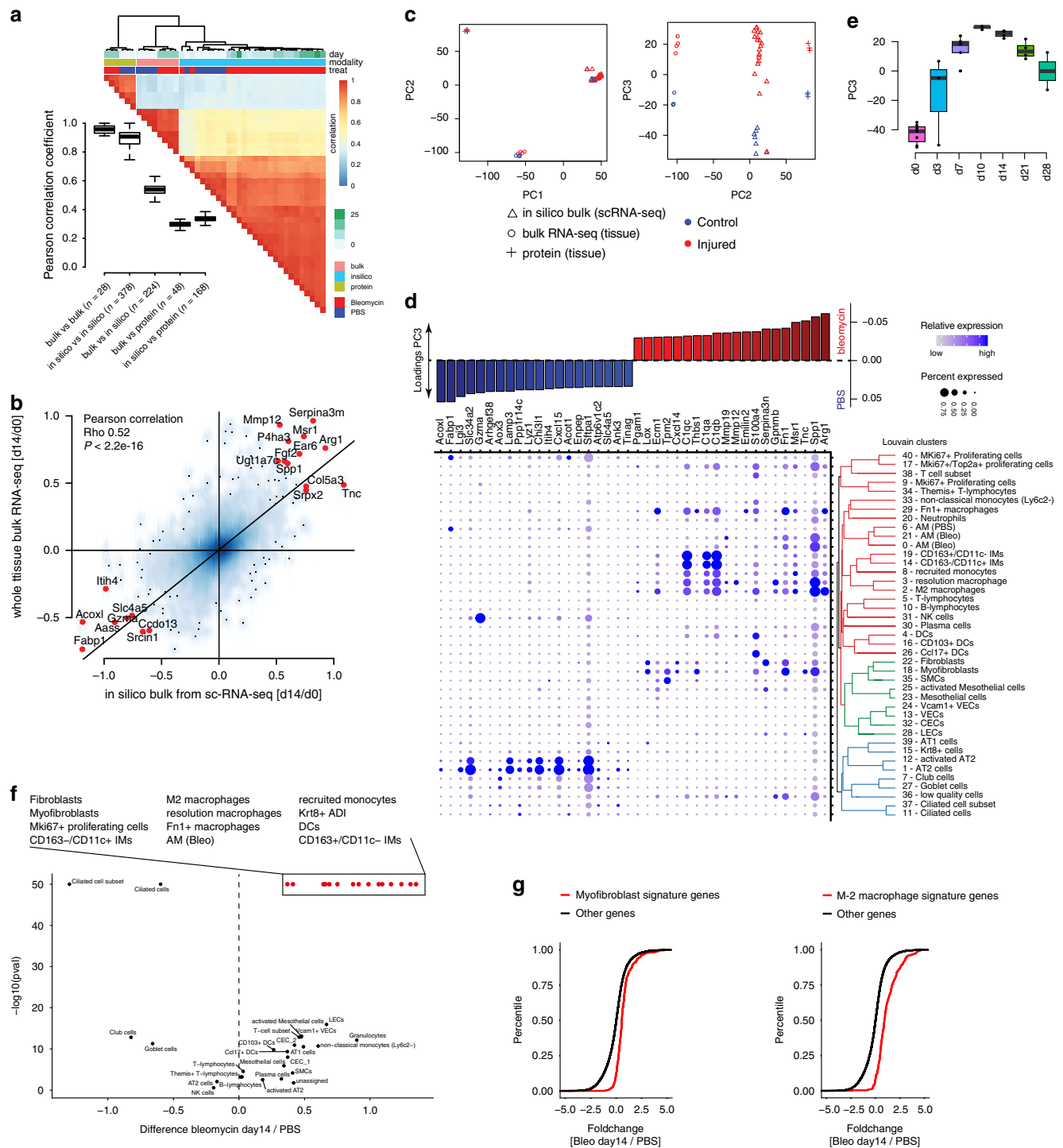
Immunostainings of *Krt8* in lung sections confirmed its transient de novo expression in lung parenchyma. We observed a peak of alveolar *Krt8* expression around day 10–14 after injury (Fig. 3f, g; Supplementary Fig. 5a). In contrast, the uninjured control lungs and fully regenerated lungs at eight weeks after injury showed *Krt8* expression only in airways (Fig. 3f). The transient burst of *Krt8* protein expression was additionally

validated on whole tissue level using mass spectrometry (Fig. 3h) and flow cytometry (Fig. 3i; Supplementary Fig. 5b, c).

Alveolar *Krt8+* cells featured high expression of pro-fibrogenic proteins, including the low-affinity epidermal growth factor receptor ligands *Areg* and *Hbegf*, as well as the integrin *Itgb6*, validated by flow cytometry (Supplementary Fig. 5b, c), and immunostainings (Supplementary Fig. 5d, e). A recent report has highlighted the important role of the *Yap/Taz* signaling pathway in alveolar regeneration<sup>25</sup>. We found high levels of nuclear *YAP* in *Krt8+* ADI cells and also some myofibroblasts, indicating active *Yap/Taz* signaling (Supplementary Fig. 5f, g). The expression of many *Yap/Taz* target genes can also be activated by *TGF-beta* signaling. We therefore assessed if *Krt8+* ADI cells also feature high levels of phospho-SMADs and found that pSMAD2 staining in fibrotic areas after bleomycin injury specifically marked *Acta2+* myofibroblasts and was surprisingly absent in *Krt8+* ADI (Supplementary Fig. 5h, i).

Morphometric analysis on 300 micron-thick precision cut lung slices revealed that *Krt8* is expressed only at very low levels in cuboidal AT2 cells in the uninjured lung, while in bleomycin injured lungs it is increased in still cuboidal AT2 cells expressing *Sftpc* and at highest levels in *Sftpc* negative cells with squamous morphology (Fig. 3k). In comparison to AT2 cells, the *Krt8+* cells showed a significantly reduced sphericity factor and also AT2 cells with upregulated *Krt8* after injury were found to assume a significantly flatter shape (Fig. 3j).





To determine if the appearance of alveolar Krt8+ ADI is specific to the bleomycin injury model, we turned to two other independent mouse models that are not based on DNA damage for the injury. Alveolar Krt8 expression was increased in a model of neonatal hypoxia and hyperoxia with Influenza type-A infection<sup>26</sup> (Supplementary Fig. 6a), as well as exposure of adult mice to hyperoxia, which has been shown to preferentially kill alveolar AT1 cells<sup>27</sup> (Supplementary Fig. 6b).

**A sky dive into epithelial cell transitions after injury.** To model the generation of Krt8+ ADI at higher temporal and cellular

resolution we sorted EpCam+ cells and sampled single cell transcriptomes daily up to day 13. We also included later time points up to day 54 after injury to analyze the recovery of the system back to baseline with fully regenerated AT1 cells. In total, we collected 18 time points after injury using two replicate mice each ( $n = 36$  mice;  $k = 34575$  cells) (Fig. 4a).

Cell type identities were consistent with the first whole-lung experiment and we identified rare neuroendocrine cells and basal cells in addition (Fig. 4b; Supplementary Fig. 7). We observed gene expression changes of AT2 cells with cell state densities moving towards the Krt8+ ADI state already at early time points starting at day 2 (Fig. 4d). A continued presence of Krt8+ ADI

**Fig. 2 Bulk deconvolution reveals cellular source of regulated proteins and cell state frequency changes.** **a** Pairwise Pearson correlation was calculated across whole lung bulk RNA-seq (bulk,  $n = 4$ ), in silico bulk scRNA-seq (in silico, Bleo  $n = 4$ , PBS  $n = 7$ ) and proteomics samples (protein,  $n = 4$ ). Bulk and proteomics data contain samples from day-14 after bleomycin-induced injury and controls<sup>24</sup>. Red and blue colors indicate high and low correlation values, respectively. Columns are ordered by unsupervised hierarchical clustering. Colored bars on top of heatmap indicate time point, data modality and injury status of each sample. Boxplot displays the distribution of Pearson correlation coefficients across comparisons between various data modalities; boxes represent the interquartile range, the horizontal line the median, and the whiskers 1.5 times the interquartile range. **b** Scatter plot depicts fold changes calculated between day 0 and 14 for the bulk (y-axis) and in silico bulk (x-axis) RNAseq samples. The black line represents the Deming regression line. Top 20 genes with the highest average fold change in both modalities are highlighted. Statistical significance was assessed using Pearson correlation ( $p < 2.2e-16$ ). **c** Data from all three modalities was integrated. The first two principal components show clustering by data modality. The third principal component separates bleomycin samples from controls across all three data modalities. Blue and red colors indicate control and bleomycin samples. **d** Barplot on top depicts genes with the highest loadings for principal component 3. **e** The box plot shows the time-resolved loading of PC3 peaking at day 10. The boxes represent the interquartile range, the horizontal line in the box is the median, and the whiskers represent 1.5 times the interquartile range (Bleo,  $n = 4$  per timepoint; PBS,  $n = 7$ ). **f** Volcano plot illustrates results from the bulk deconvolutions analysis. X axis indicates mean fold change of cell type markers between day 14 and PBS bulk samples. Y axis displays the  $-\log_{10}$   $p$ -value derived from a two-sided Kolmogorov-Smirnov test.  $P$ -values were limited to a minimum of  $1e-50$  for visualization purposes. **g** Empirical cumulative density plots show two exemplary cell types Myofibroblasts (right) and M2 macrophages (left). Red and black lines correspond to the distribution of cell type markers and all other genes, respectively.

cells was then seen until day 36 (Fig. 4d, Supplementary Fig. 7). Scoring single cells for enrichment of gene programs revealed that in comparison to the other epithelial cells, Krt8+ ADI displayed high scores for genes involved in epithelial–mesenchymal transition (EMT), cell senescence, and the p53, MYC, TNFA via NFkB, and oxidative phosphorylation pathways (Fig. 4e). All these pathways have been characterized by expression of a host of secreted factors that may promote fibrogenesis. Statistical analysis of pathway enrichment confirmed the strong and specific enrichment of genes previously associated with wound healing, angiogenesis and the p53 pathway in the Krt8+ ADI cells (Fig. 4f).

We hypothesize that the Krt8+ ADI cell state with its unique gene expression program serves important niche functions to coordinate other cell types during tissue regeneration. In the receptor-ligand database (Fig. 1) the Krt8+ ADI show their largest number of receptor-ligand pairs with fibroblasts, macrophages and (capillary) endothelial cells (Fig. 4f; Supplementary Data 6). Interestingly, in the endothelial cell (EC) connectome with Krt8+ ADI and AT1, the capillary ECs received signals via the endothelin-receptor (Ednrb) expressed on ECs via the ligand endothelin-1 (Edn1), which was specifically expressed on Krt8+ ADI and not on AT1 (Fig. 4g). Conversely, the AT1 cells displayed a large number of ligands, including Vegfa and Sema3e that bind receptors, such as Flt1 or Nrp1/2 on ECs, which were not expressed on Krt8+ ADI. Similar selective differences between Krt8+ ADI and AT1 were observed for receptors such as the urokinase plasminogen activator receptor (Plaur) specifically expressed on Krt8+ ADI but not AT1, binding to the EC-derived ligand PAI-1 (Serpine1) (Fig. 4h).

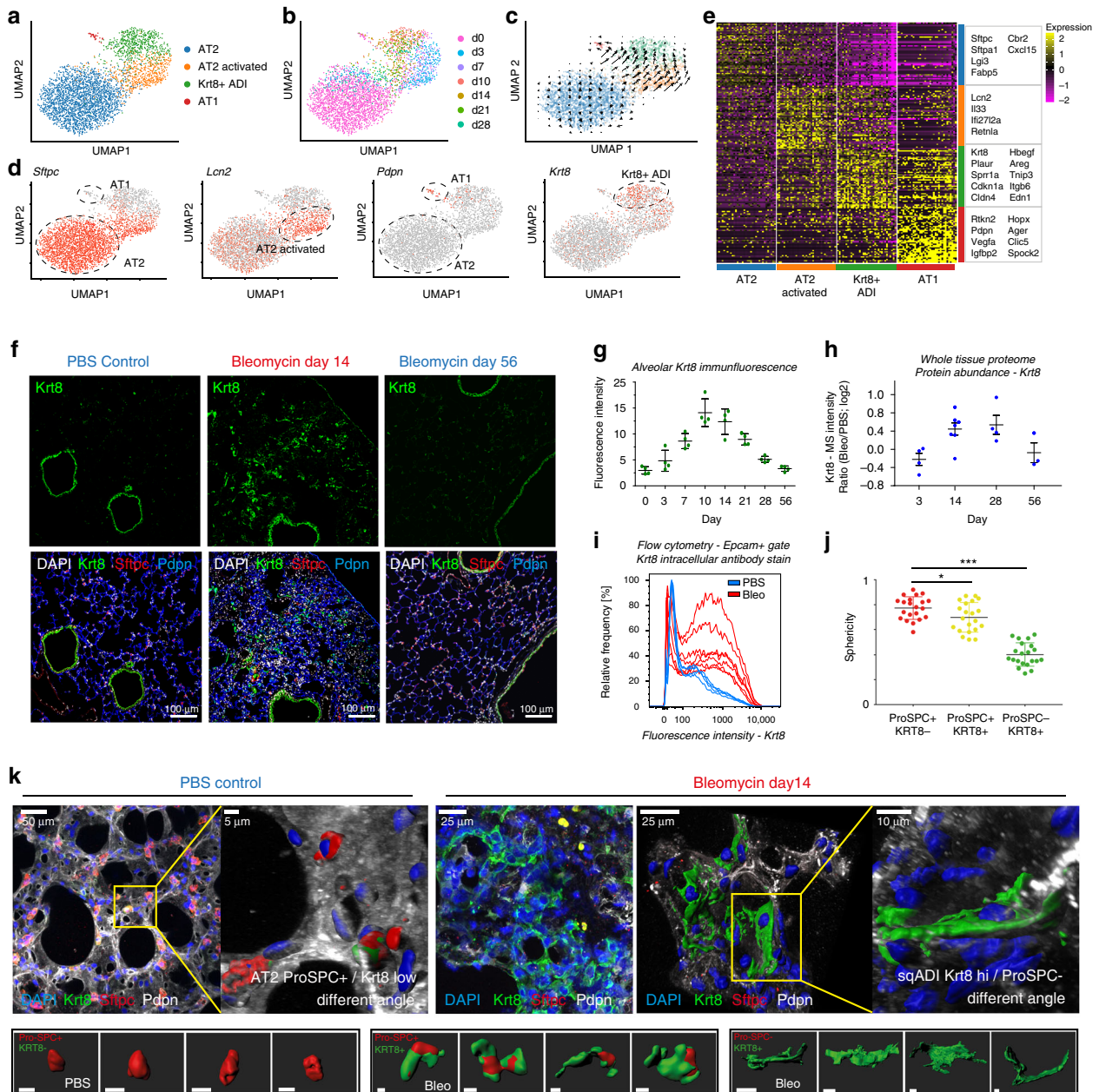
**Involvement of airway stem cells in alveolar regeneration.** To analyze global connectivity and potential trajectory topology in the epithelial cell state transitions we applied partition-based graph abstraction (PAGA) (Fig. 5a), which provides an interpretable graph-like map of the data manifold<sup>28</sup>. Interestingly, the PAGA map revealed several nodes with high connectivity between cell types that represented potential transdifferentiation bridges. In particular, we observed a subset of airway club cells (cluster 10) with connectivity to all alveolar cells including Krt8+ ADI, and an activated AT2 cell state (cluster 9) which also featured high connectivity to Krt8+ ADI. We simulated gradual differentiation intermediates by generating in silico doublets combining AT1 with cluster 10 and 9 (Fig. 5b). The simulated doublets mapped between these clusters and AT1 samples, while Krt8+ ADI cells mapped orthogonal to linear differentiation

trajectories towards AT1. This demonstrates that the Krt8+ ADI state is highly distinct and does not resemble a linear gene expression intermediate from stem cells towards AT1 (Fig. 5b).

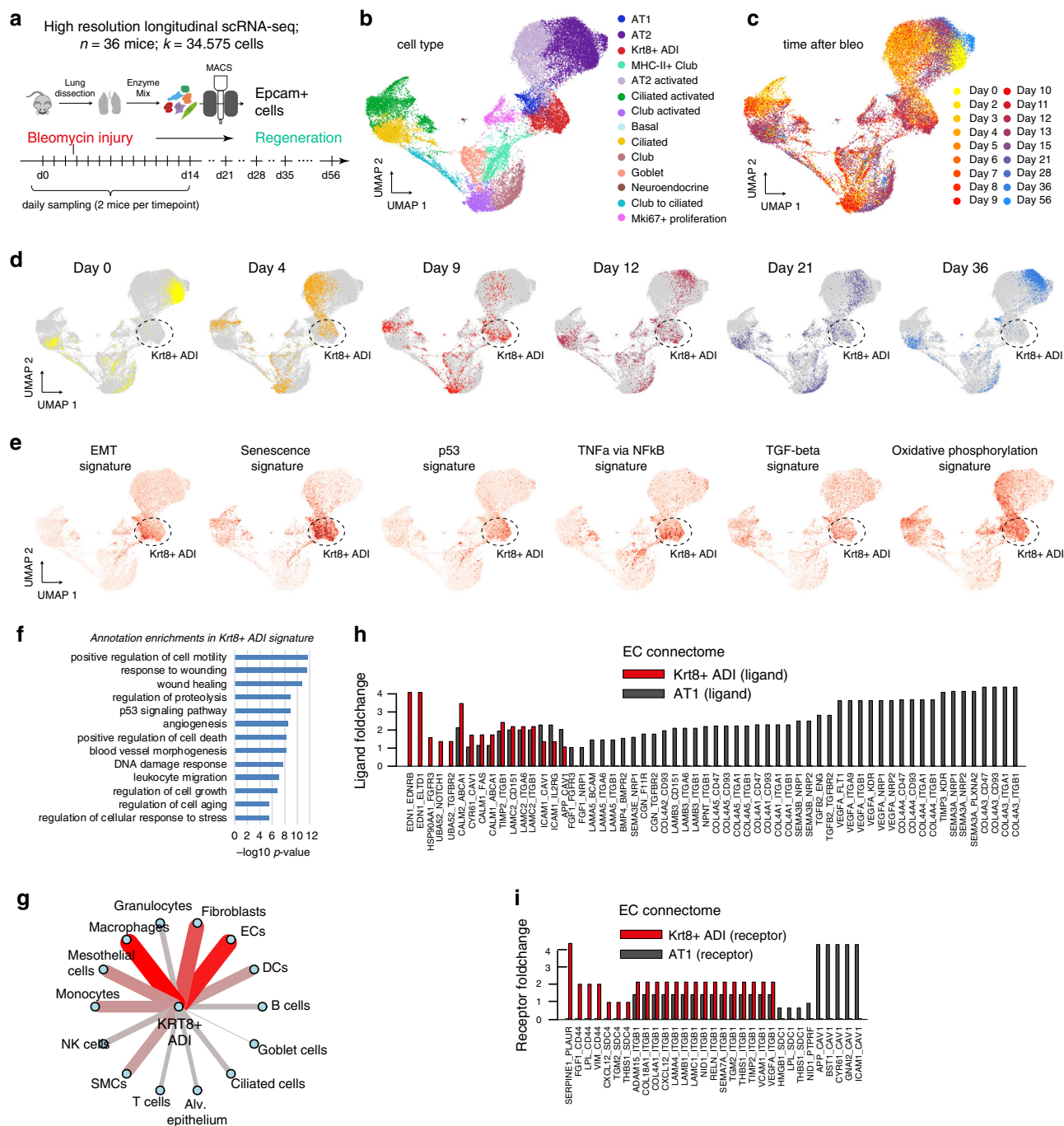
Two clusters (7 and 8) mainly represented club cells at different times after injury, which we termed club and club activated, respectively (Fig. 5c, d, h). Cluster 10, however, was highly distinct and surprisingly marked by high expression of MHC-II complex genes (e.g. H2-Ab1) and the cysteine proteinase inhibitor Cystatin-C (Cst3), which is typically co-expressed with MHC-II in dendritic cells<sup>29</sup> (Fig. 5e, f). Of note, MHC-II positive club cells were not doublet artefacts as evidenced by comparison to artificially generated club and dendritic cell doublets (Fig. 5g). We additionally validated the MHC-II + club cell state using immunofluorescence of Cst3 that stained a rare subset of Scgb1a1+ airway club cells (Fig. 5i). Taken together, our data suggests the existence of a distinct cell state within the club cell lineage, marked by high expression of MHC-II genes, that features high connectivity to alveolar epithelial cell identities. Importantly, a recent report described a very similar gene signature in club-like epithelial progenitors that regenerated both AT2 and AT1 cells in the bleomycin model<sup>30</sup>, suggesting that we have identified the same stem cell in our data.

We occasionally found rare cells with high levels of *Krt8* expression in the alveolar space of uninjured control lungs (Supplementary Fig. 7, 10a), suggesting that the same cell state observed after injury may be a natural intermediate of homeostatic cell turnover. These pre-existing alveolar Krt8+ cells did not undergo proliferative expansion. The relative frequency of Ki67+ proliferating cells in the single cell data manifold (cluster 14) peaked at day 15 (Supplementary Fig. 9a). Counting Ki67+ cells in immunostainings confirmed the peak of cell proliferation around day 14 with a sudden drop in proliferation rates around day 28 (Supplementary Fig. 9d, e). Cell cycle regression within the proliferative cells enabled us to deconvolve cell type identity (Supplementary Fig. 9b), revealing that Krt8+ ADI cells, AT2, club, and the MHC-II + club cells all proliferated after injury (Supplementary Fig. 9c). We validated proliferating Krt8+ cells in co-immunostainings Ki67+ at day 10 after injury (Supplementary Fig. 9f). Importantly, the massive expansion of Krt8+ ADI over time happened without spiking numbers of Krt8+/Ki67+ cells preceding this (Supplementary Fig. 10b). Using tamoxifen labeling in SPC-CreERT2 and Sox2-CreERT mice we found that the rare pre-existing Krt8+ ADI cells were 80% labeled in the SPC-CreERT2 mice (Supplementary Fig. 10c–e), suggesting that these cells are derived from AT2, possibly during normal homeostatic turnover.





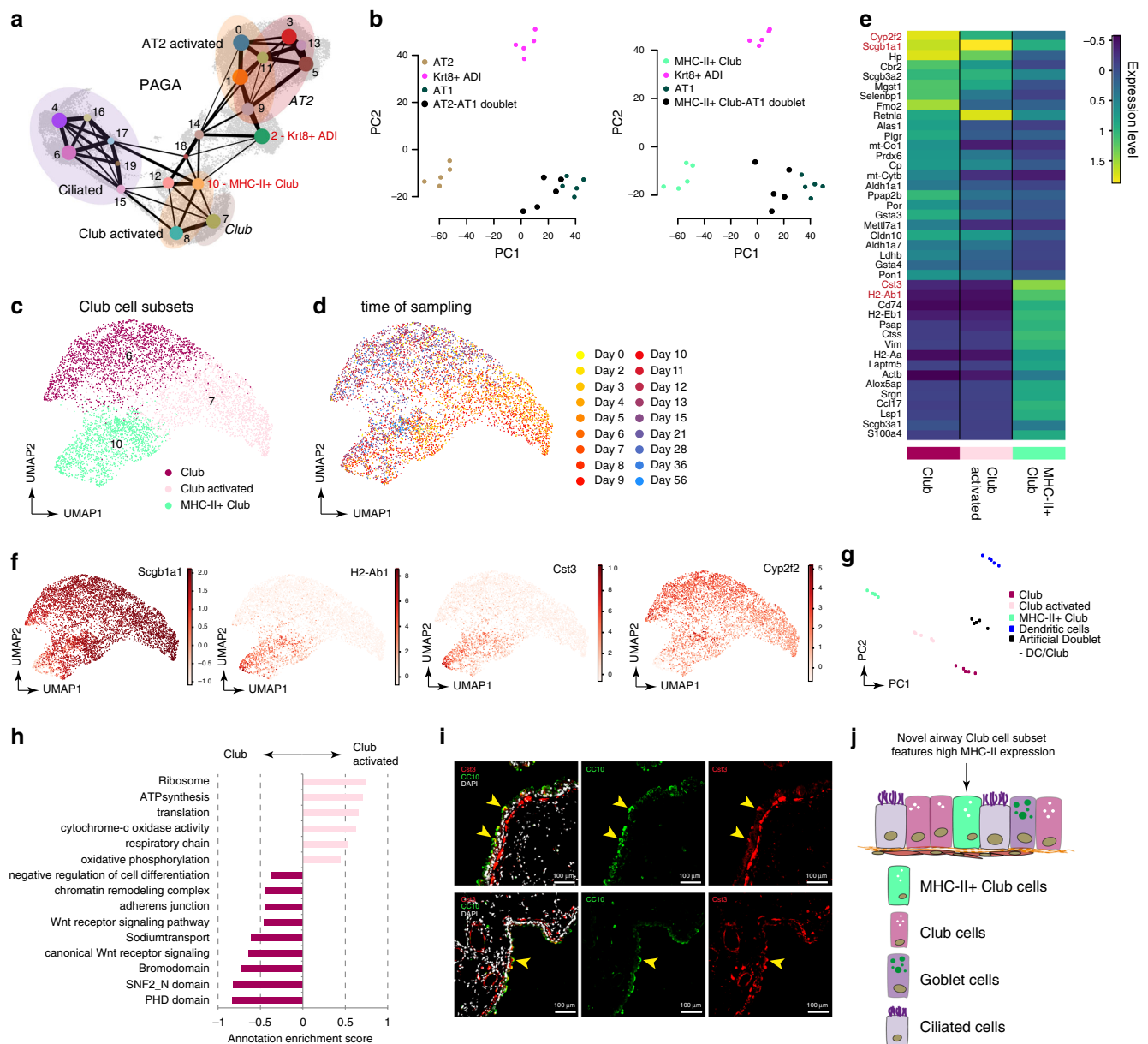
**Fig. 3** Alveolar regeneration features a transient squamous cell state marked by *Krt8* expression. **a–d** UMAP embedding of alveolar epithelial cells shows **(a)** four distinct cell states, and **(b)** the time points of sampling, and **(c)** the RNA velocity vectors, indicating AT2 cell differentiation towards the alveolar *Krt8*+ cell state after bleomycin-mediated injury, and **(d)** gene expression of the indicated marker genes. **e** Heatmap of top 50 differentially expressed genes across alveolar cell states, with selected marker genes in boxes. **f** Fluorescent immunostainings from the indicated conditions show nuclei (DAPI) in white, *Krt8* in green, *Sftpc* (AT2 cells) in red, and *Pdpn* (AT1 cells) in blue (scale bar 100 microns). **g** Quantification of *Krt8* mean fluorescence intensity in alveolar space (excluding airways;  $n = 4$  per time point, mean with SD). **h** Protein abundance of *Krt8* in total lung homogenates was assessed by mass spectrometry<sup>24</sup>. Individual data points show log<sub>2</sub> ratio of *Krt8* MS-intensity after bleomycin injury [ $n(d3) = 4$ ,  $n(d14) = 7$ ,  $n(d28) = 4$ ,  $n(d56) = 3$ ] versus PBS control mice ( $n = 4$ ). The mean and standard error of the mean is shown. **i** *Krt8* fluorescence intensity quantified by flow cytometry in epithelial cells. PBS control ( $n = 5$ , blue color) and day 10 after bleomycin ( $n = 7$ , red color) is shown. **j** Alveolar cell sphericity analysis of 21 cells per condition revealed elongated cell shapes for alveolar *Krt8*+ cells in IF-stained precision cut lung slices (in **k**). Sphericity of 1 indicates round, cuboidal cells, 0 indicates flat cells. PBS,  $n = 2$ ; Bleo,  $n = 2$ . One-way ANOVA with Dunnett's post testing:  $*p = 0.0376$ ,  $***p < 0.0001$ . **k** Maximum projections of confocal z-stacks taken from immunostained 300 micron-thick precision cut lung slices (PCLS) are shown for a representative PBS control mouse and a mouse at day 14 after bleomycin injury. Nuclei (DAPI) are colored blue, *Krt8* appears in green, *Sftpc* (AT2 cells) in red, and *Pdpn* (AT1 cells) in white. Image data representation stems from  $n = 5$  samples. Small images below show examples taken for cell morphometric analysis (in **j**). All scale bars in small single-cell images represent 15  $\mu\text{m}$ .



**Fig. 4 Krt8+ADI cells feature unique pathway and cell-cell communication activities.** **a** A high-resolution longitudinal data set was generated by subjecting sorted cells from the epithelial compartment to scRNAseq at the 18 indicated time points. UMAP embedding displays cells colored by **(b)** cell type identity and **(c)** time point. **d** The colored dots on the UMAP illustrate densities and distribution of cells at individual time points after bleomycin injury. Note the time dependent movement of cells within the data manifold. **e** UMAP embedded visualizations of single cells colored by gene expression signature scores for the indicated pathways (MSigDB Hallmark gene sets). **f** The indicated terms were significantly enriched in the Krt8+ ADI signature compared to all other epithelial cell states. **g** The cell-cell communication network displays the number of receptor-ligand pairs between the molecular markers of the Krt8+ ADI state and all other meta cell type identities (Fig. 1). **h, i** The bar graphs show the average log<sub>2</sub> fold change of either **(h)** receptors or **(i)** ligands within the endothelial cell (EC) connectome for Krt8+ ADI and AT1 cells.

**Transcriptional convergence of alveolar and airway stem cells.** RNA velocity vectors overlaid onto the UMAP embedding predicted transdifferentiation of club cells towards ciliated and goblet cells, which is in agreement with previous literature<sup>2</sup> (Fig. 6a). Interestingly, RNA velocities also strongly suggested a dual origin of alveolar Krt8+ ADI cells from AT2 and airway cells, in particular from Scgb1a1+ club cells (Fig. 6a, b). Club cells

and MHC-II+club cells show differentiation bridges towards AT2 cells and Krt8+ ADI (Fig. 6b). As MHC-II+club cells showed very high connectivity to Krt8+ ADI and were closest in the UMAP embedding, we restricted the analysis to the activated AT2, MHC-II+club and Krt8+ ADI states, and calculated terminal state likelihoods based on RNA velocities, which showed differentiation of both activated AT2 and MHC-II+



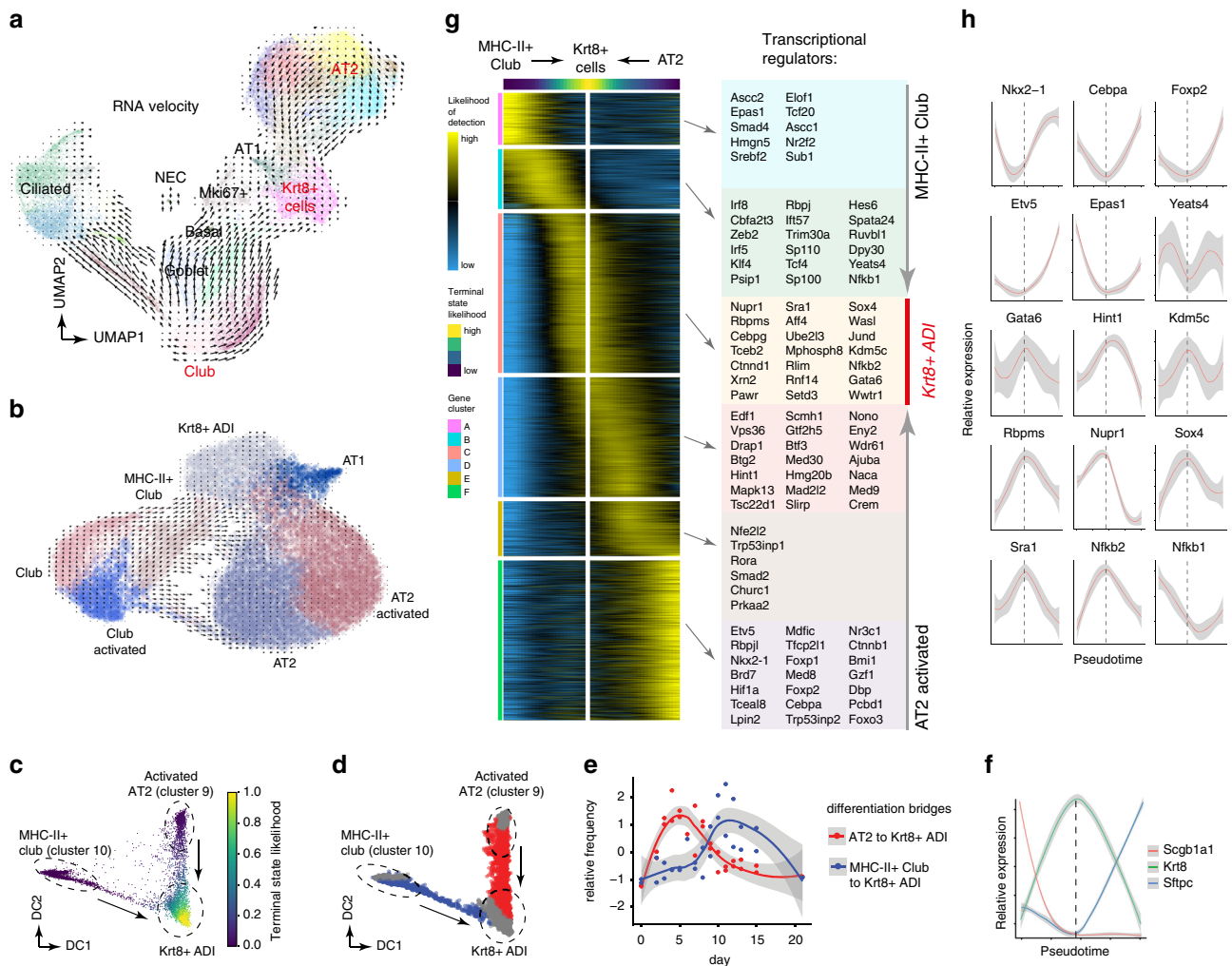
**Fig. 5** A distinct club cell state shows high connectivity to alveolar cell identities after injury. **a** The PAGA graph visualizes potential cell-type transitions and the topology of the data manifold. Nodes represent Louvain clusters and thicker edges indicate stronger connectedness between clusters. **b** Principal component analysis of artificially generated cluster-specific and doublet in silico bulk samples shows that Krt8+ ADI cells map orthogonal with respect to linear activated AT2 to AT1 (left) and MHC-II+ club to AT1 (right) differentiation profiles. In silico bulk samples are colored by cluster as derived from the PAGA map in (**a**). Artificially generated doublets are colored in black. **c**, **d** The plots visualize the UMAP embedding of Club cells colored by Louvain clustering (**c**) and by time point (**d**). **e** The heatmap shows the average expression levels of marker genes across the three club cell clusters. **f** UMAP embeddings show distinct expression patterns for selected marker genes. **g** Principal component analysis of artificially generated cluster-specific and doublet in silico bulk samples shows that MHCII + club cells map orthogonal with respect to a linear connection between dendritic and club cells. Dendritic cell samples and artificially generated doublets are colored in blue and black, respectively. **h** The bar graph shows the annotation enrichment score<sup>78</sup> for selected examples of gene categories with significant enrichment (FDR < 5%) in either activated Club (positive scores) or Club cells (negative scores). **i** Immunofluorescence staining of mouse airways shows CC10+ club cells (green) and Cst3+ cells (red), DAPI (white). Note the partial overlap of Cst3+/CC10+ airway cells (highlighted by yellow arrowheads). Scale bar = 100 microns; representative images from n = 3 bleo-treated mice. **j** Revised model of club cell heterogeneity in mouse airways.

airway club cells towards Krt8+ ADI (Fig. 6c). Even though MHC-II+ club cells (cluster 10) showed high connectivity with alveolar cells (Fig. 5b), the data indicates that also other Scgb1a1+ club cells can give rise to alveolar cells during injury repair.

Restricting the analysis to cells “bridging” from the AT2 and MHC-II+ club cells to Krt8+ ADI (Fig. 6d), we found that the

AT2 conversion preceded the MHC-II+ club to Krt8+ ADI differentiation by about one week (Fig. 6e). This may indicate that alveoli with surviving AT2 cells regenerate faster than alveoli with total loss of AT2 that require recruitment of distal airway stem cells. Thus, the data shows convergence of transcriptional states from distinct lineages (airway stem cells versus alveolar AT2) even at different times after injury.



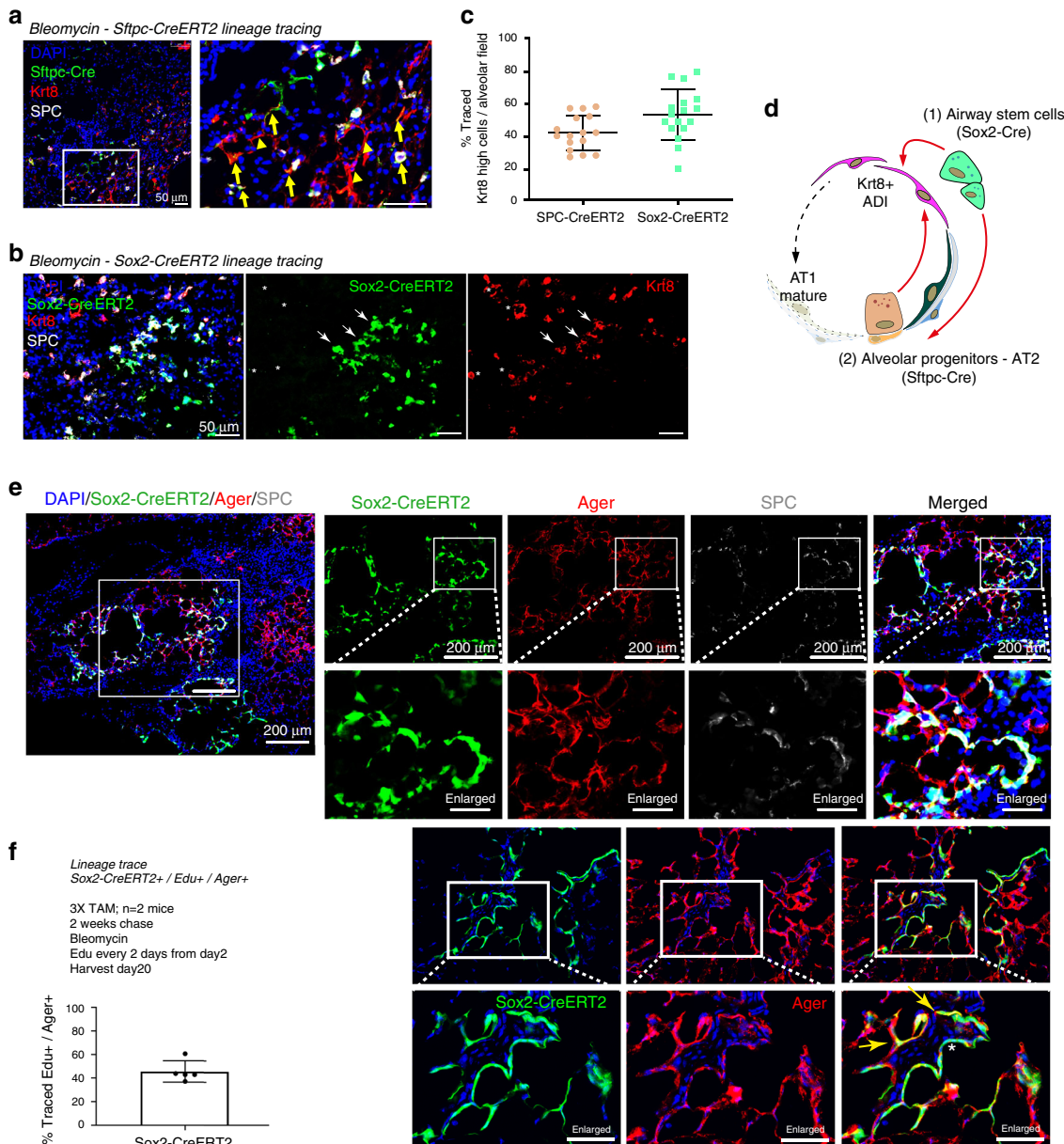


**Fig. 6 Transcriptional convergence of MHC-II+ club and AT2 cells onto the alveolar Krt8+ ADI cell state.** **a** Velocity plot displays the UMAP embedding colored by Louvain clusters with velocity information overlaid (arrows). **b** Velocity plot of a subset of the data only showing alveolar identities and club cell subsets. RNA velocity shows contribution of Scgb1a1+ club cells to both Krt8+ ADI and AT2 identities. **c** Diffusion map of Louvain clusters 2, 10, and 9 colored by inferred terminal state likelihood reveals two distinct transdifferentiation trajectories from activated AT2 and MHC-II + club cells towards a Krt8+ cell state. **d** Diffusion map colored by groupings derived from Gaussian Mixed Model Clustering. Red and blue colors represent AT2 and MHC-II + club cell differentiation bridges towards the Krt8+ ADIs. Grey colors represent cells at endpoints. **e** The lines indicate smoothed relative frequencies across time points of cells within the AT2 (red) and MHC-II + club cell (blue) differentiation bridges. **f** The lines illustrate smoothed expression levels of Scgb1a1, Krt8, and Sftpc across the trajectory, marking cell identities. The dashed vertical line indicates the peak of Krt8 expression. **g** The heatmap shows the gene expression patterns along the differentiation trajectory based on the inferred likelihood of detection for 3036 altered genes. **h** Line plots show the smoothed relative expression levels of selected transcriptional regulators across the converging trajectories. The dashed vertical line indicates the peak of Krt8 expression. For **(e)**, **(f)**, and **(h)**, gray colors represent the 95% confidence interval derived from the smoothing fit.

We identified 3036 genes showing distinct expression patterns along these differentiation trajectories (Fig. 6f, g; Supplementary Data 4). We observed a gradual decline in expression of the Homeobox protein Nkx-2.1, critical for lung development and lung epithelial identity<sup>31</sup>, as well as Foxp2, which is one of the key transcriptional repressors involved in the specification and differentiation of the lung epithelium<sup>32,33</sup>, in both MHC-II + club and AT2 cells during conversion to Krt8+ ADI (Fig. 6h). Also, expression of the transcription factor Cebpa with important functions in lung development and maintenance of both club and AT2 cell identity<sup>34–36</sup> reached a minimum at the Krt8+ ADI state. AT2 cell conversion into Krt8+ ADI was marked by a drastic reduction of the transcription factor Etv5, which has been shown to be essential for the maintenance of AT2 cells<sup>37</sup> (Fig. 6h). Conversely, the differentiation towards the Krt8+ ADI signature expression was characterized by a gradual increase in one of the master regulators of AT1 cell differentiation Gata6<sup>38,39</sup> in both

MHC-II + club and AT2 cells. Both trajectories converged on a large number of alveolar Krt8+ ADI specific genes representing distinct pathways (Fig. 4) and their transcriptional regulators, including the stress-induced p53 interactor Nupr1, a master regulator of epithelial to mesenchymal transition Sox4, and many other genes, including chromatin remodeling factors such as the histone demethylase Kdm5c (Fig. 6h). To validate our findings we re-analysed the scRNAseq data set from whole-lung suspensions (Fig. 1), which confirmed differentiation of AT2 cells onto the Krt8+ ADI state and contribution of airway cells to alveolar fates (Supplementary Fig. 10).

To experimentally validate this computational analysis, we used Sftpc-CreERT2 (AT2 cells) and Sox2-CreERT2 (airway cells) reporter mice to trace the origin of Krt8+ ADI cells back to these lineages. In the quantification of these two independent lineage tracing experiments (Fig. 7a, b), we observed that approximately half of the alveolar Krt8+ ADI cells were derived from either

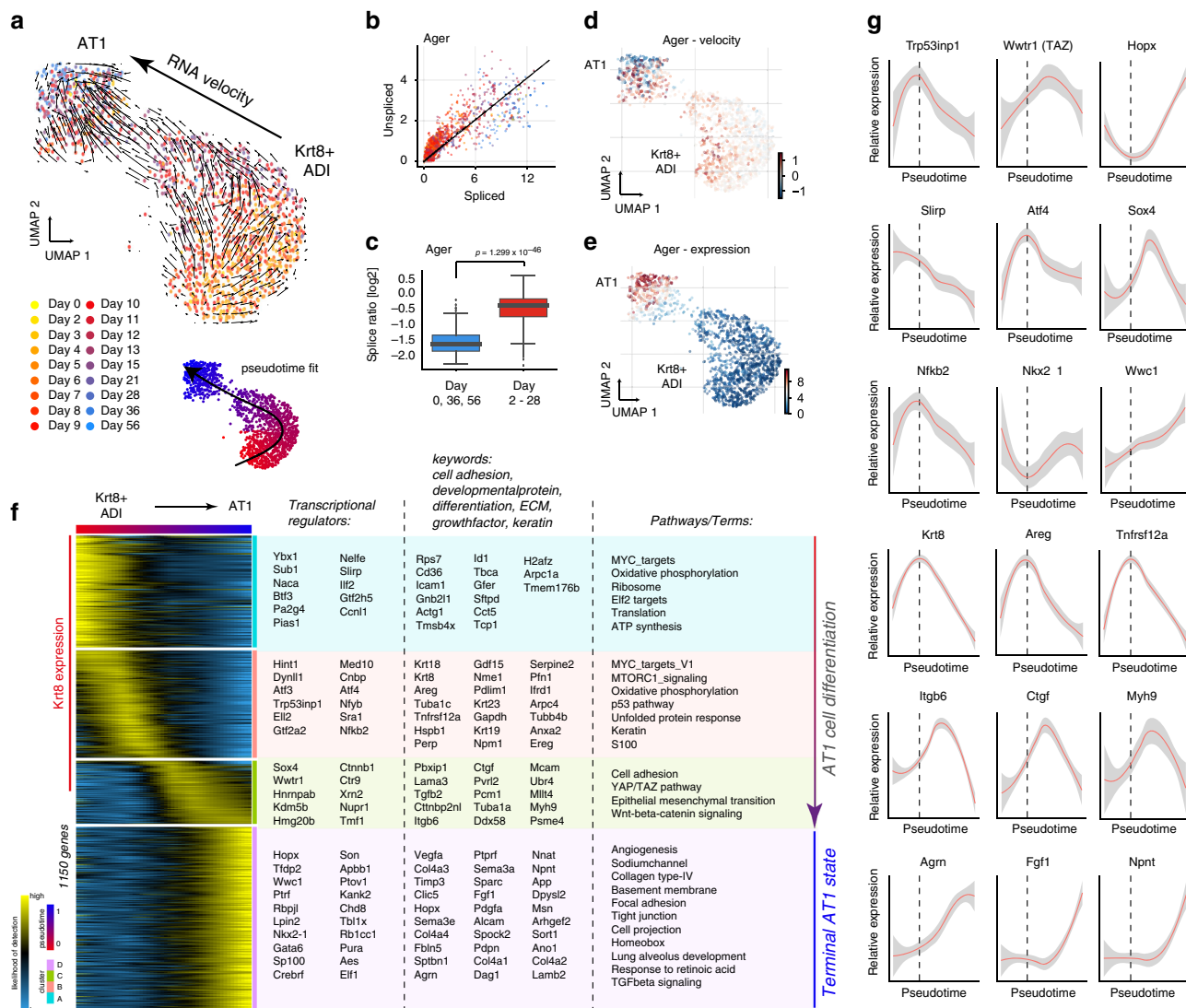


**Fig. 7** Lineage tracing validates dual origin of Krt8+ ADI. **a, b** Immunostainings of Krt8 and Sftpc (SPC) in **(a)** Sftpc-CreERT2-labeled mice ( $n = 2$ ) and **b** Sox2-CreERT2-labeled mice ( $n = 2$ ; lobes analyzed per mouse:  $n = 3$ ). Arrows indicate lineage positive and stars lineage negative Krt8+ ADI. **c** Quantification of lineage labeled alveolar cells with high Krt8 expression. Each point in the graph represents a large region (at least 1.2 mm<sup>2</sup> area) and cells from at least three lobes/mouse at two different levels (>100  $\mu$ m apart) were analyzed in 16 fields of view (Sftpc-CreERT2:  $n$  cells = 1382; Sox2-CreERT2:  $n$  cells = 1833). **d** Lineage tracing experiments validate the scRNAseq experiments and show convergence of distinct alveolar progenitors into Krt8+ ADI. **e** Immunostainings of the AT1 marker Ager and the AT2 marker SPC with the Sox2-CreERT2 lineage label at day 14 after bleomycin. Heavily injured regions show only little endogenous SPC+ cells but also Sox2-traced Ager+ and SPC+ cells. The experiment was performed on  $n = 2$  mice and  $n = 3$  lobes/mouse were analyzed. Flat lineage labeled cells can be observed Ager+ (yellow arrow) and Ager- (asterisk). **f** Sox2-CreERT2+/Ager+ AT1 cells that previously proliferated upon bleomycin injury were quantified using the indicated EdU pulse chase labeling strategy. The percentage of EdU chased and lineage labeled AT1 cells is shown. Each dot represents cell counts from at least 2 large regions from two mice,  $n(\text{mice}) = 2$ ; data represented with mean and SD.

Sftpc-CreERT2 or Sox2-CreERT2+ airway cells in the bleomycin model (Fig. 7c, d). Of note, Sox2-CreERT2+ airway cells gave rise to both Sftpc+ and Sftpc- cells (Fig. 7b), Ager+ (AT1 marker) and Ager- squamous cells covering alveolar surfaces (Fig. 7e). We found increased contribution of Sox2-lineage labeled cells in severely injured areas. Using an EdU pulse chase labeling strategy we chased proliferating cells every other day for 20 days and counted Sox2-CreERT2+/Ager+ cells, revealing that around 40% of newly formed AT1 cells were airway derived (Fig. 7f;

Supplementary Fig. 10f). In conclusion, the lineage tracing data confirms a dual origin of Krt8+ ADI cells and substantiates our prediction of a substantial contribution of airway derived stem cells in alveolar regeneration after bleomycin injury.

**Cell state trajectory model of AT1 cell regeneration.** Analysis of RNA velocity information within the subset of Krt8+ ADI and AT1 cells indicated differentiation of Krt8+ ADI toward



**Fig. 8 Terminal differentiation trajectory modeling of Krt8+ ADI to AT1.** **a** Velocity plot displays the UMAP embedding colored by time point with velocity information overlaid (arrows), indicating terminal differentiation of Krt8+ ADI into AT1 cells. **b** The velocity phase plot shows the number of spliced and unspliced reads of the AT1 marker Ager for each cell (points) on the X and Y axes, respectively. Cells are colored by time point and the black line represents the linear steady-state fit. Cells above and below the diagonal are predicted to be in inductive or repressive states, respectively. **c** The Boxplot shows the log2 ratio of unspliced over spliced Ager reads for days 0, 36 and 56 (blue,  $n = 100$  cells) and all other time points (red,  $n = 1193$  cells). To avoid division by zero, one was added to both counts. Statistical significance was assessed by using Wilcoxon rank-sum test (two-sided). The boxes represent the interquartile range, the horizontal line in the box is the median, and the whiskers represent 1.5 times the interquartile range. UMAP embedding colored by Ager velocity (**d**) and expression (**e**) displays a gradual increase along the inferred trajectory. **f** The heatmap shows the gene expression patterns across the differentiation trajectory for 1150 altered genes. **g** The line plots illustrate smoothed expression across the differentiation trajectory for a number of exemplary genes. Gray colors represent the confidence interval derived from the smoothing fit. The dotted line indicates the peak of Krt8 expression.

AT1 cells (Fig. 8a). The ratio of spliced and unspliced reads revealed gradual induction of transcription of the AT1 cell marker Ager in Krt8+ ADI cells around day 14 (Fig. 8b). Days 0, 36 and 56 representing a baseline with mature AT1 cells contained a significantly lower ratio of unspliced over spliced Ager reads compared to all other time points (Fig. 8c; Wilcoxon Rank Sum test,  $P < 1e-46$ ). A gradual decrease in Ager mRNA velocity (Fig. 8d), was reflected with a gradual increase of Ager expression (Fig. 8e). Using this information, we modeled a pseudotime trajectory and determined gene expression dynamics for 1150 significantly regulated genes along the putative Krt8+ ADI to AT1 transition (Fig. 8f; Supplementary Data 5).

The differentiation trajectory was split in four phases that were marked by distinct sets of transcriptional regulators,

developmental genes and signaling pathways (Fig. 8f, g). The initial phase was marked by genes and pathways consistent with cell growth after exit from the cell cycle (e.g. MYC targets). This was followed by the induction of stress-related signaling pathways, such as the p53 pathway and the unfolded protein response pathway, featuring increased expression of the corresponding transcriptional regulators such as Trp53inp1 and Atf4, and the peak of Krt8 expression (Fig. 8f, g). Next, a critical pre-AT1 stage was marked by the downregulation of the Krt8 signature and the induction of a gene expression program with similarities to the epithelial to mesenchymal transition (EMT), together with one of its master regulators Sox4<sup>40</sup> (Fig. 8g). We further observed pre-AT1 specific expression of important transcriptional regulators such as TAZ (Wwtr1) and beta-catenin



(Ctnnb1), and pro-fibrogenic proteins such as integrin beta-6 (Itgb6), and connective tissue growth factor (Ctgf). The non-muscle myosin heavy chain IIa (Myh9) also peaked in pre-AT1 cells, suggesting important additional cytoskeletal rearrangements and increased cell contractility in the already squamous Krt8+ ADI cells in the final steps of maturation towards AT1 cells (Fig. 8g).

Terminally differentiated AT1 cells were characterized by high expression of the transcription factors Hopx, Gata6 and Wwc1, as well as a large number of developmentally important factors, including extracellular matrix proteins and growth factors, such as Fgf1, Npnt and Agrn (Fig. 8g). It has long been noted that isolated AT2 cells spontaneously drift toward AT1 fate in vitro, suggesting that plasticity may be a cell-intrinsic property and that AT2 cell identity in vivo is actively maintained by niche signals. Interestingly, during a five-day AT2 to AT1 in vitro differentiation, Krt8 protein levels were shown to be highest at day 3, followed by the AT1 marker Pdpn peaking later at day 5<sup>41</sup>. We repeated this experiment and subjected isolated AT2 cells to inhibition of Wnt/ $\beta$ -catenin/TCF-mediated transcription, which significantly reduced the induction of Krt8 expression and levels of the AT1 cell marker Pdpn in comparison to controls (Supplementary Fig. 12, Supplementary Fig. 13).

**Aberrant persistence of Krt8+ ADI is linked to fibrosis.** We here identified and characterized the transient appearance of Krt8+ ADI cells during lung regeneration. Single cell analysis of human lung fibrosis recently identified a disease specific cell state that was termed aberrant basaloid cell (KRT17+/KRT5-) based on some similarities to airway basal cells<sup>42,43</sup>. It is currently unclear if these cells are indeed airway derived or could represent stem cells undergoing alveolar repair. We re-analysed available human single cell data to extract a full gene expression signature characterizing KRT5-/KRT17+ human basaloid cells (Fig. 9a, b). Scoring the human basaloid cell signature on single cells in the mouse data manifold revealed that Krt8+ ADI cells described in this work are very similar to KRT5-/KRT17+ cells in IPF (Fig. 9c). A systematic cross-species comparison of epithelial cell state identities confirmed that human KRT5-/KRT17+ basaloid cells are most closely related to mouse Krt8+ ADI (Fig. 9e).

A recent landmark study showed that blocking alveolar stem cell differentiation, through deletion of the RhoGTPase Cdc42 in a model of pneumonectomy induced regeneration, leads to the accumulation and persistence of a unique AT2 derived cell state. These mice have progressive lung fibrosis with the typical periphery-to-center pattern of disease progression as seen in IPF patients<sup>44</sup>. Using quantitative comparisons of the gene expression signatures measured in this study we found that this AT2 derived cell state is also very similar if not identical to the Krt8+ ADI cells discovered by us (Fig. 9d), suggesting that persistence of Krt8+ ADI may directly mediate progressive lung fibrosis.

To further validate that KRT8+ alveolar cells can also be observed in human acute lung injury and chronic lung disease associated with alveolar injury, we stained human tissue sections and did not detect any expression of KRT8 in the alveolar space of non-injured control lungs ( $n = 7$ ; Fig. 9f). In sharp contrast, we observed very strong alveolar KRT8 expression in human acute respiratory distress syndrome (ARDS,  $n = 2$ ) caused by Influenza-A and pneumococcal infection and interstitial lung disease patients with various diagnoses ( $n = 5$ ; Fig. 9g, h). Finally, we also co-stained KRT8 with KRT17 and observed co-expression in both flat epithelial cells and bronchiolized epithelia in fibrotic areas but not in controls (Fig. 9i).

## Discussion

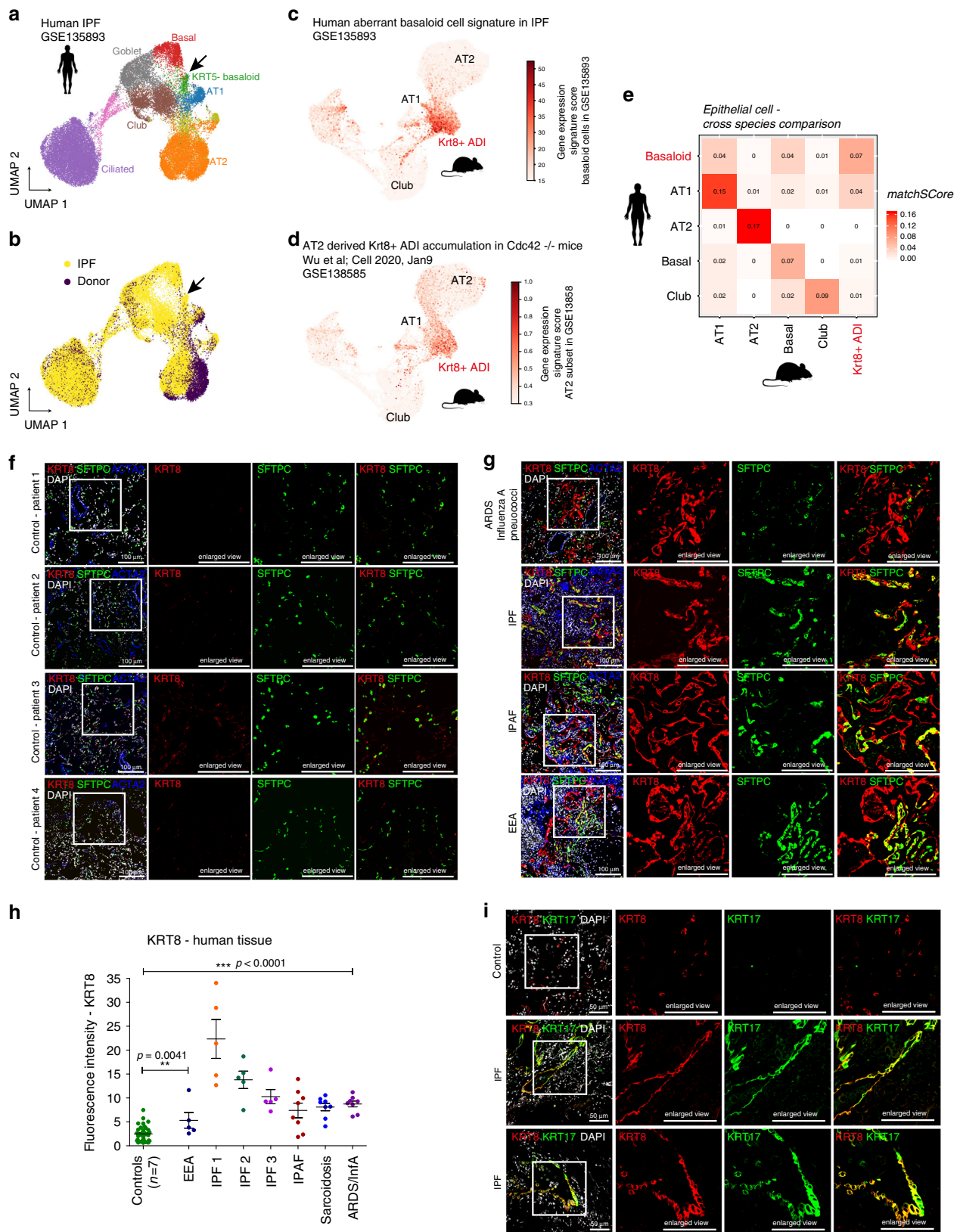
In this work, we describe the dynamics of mouse lung regeneration at single cell resolution and discover the transcriptional convergence of airway and alveolar stem cells to a Krt8+ transitional stem cell state that precedes the regeneration of AT1 cells (Fig. 10). The discovery of Krt8+ ADI cells in several independent mouse lung injury models and human lung fibrosis sheds a new light on reports of EMT<sup>45</sup>, senescence and p53 activation<sup>46–49</sup> in lung injury, repair and fibrosis. We conceptualize these observations with the appearance of this transient stem cell derived Krt8+ ADI state with its unique transcriptomic signature. Using the power of pseudotemporal modeling<sup>15,16,50</sup> we analyze gene regulation during stem cell differentiation, laying out the sequence of gene programs and transcriptional regulators. Our cell state trajectory model was validated by correspondence with the real time points of sampling, RNA velocities of individual cells and lineage tracing experiments. The receptor-ligand analysis revealed potential routes of cell–cell communication and their dynamics over time. All data and code is freely available at our interactive webtool and github repository ([github.com/theislabs/LungInjuryRegeneration](https://github.com/theislabs/LungInjuryRegeneration)).

Even though the Krt8+ ADI gene programs resemble features found in EMT, we do not see conversion of epithelial cells to anything similar to fibroblasts. It seems that we rather see an overlap in gene expression patterns between cells undergoing genuine epithelial–mesenchymal transition (e.g. neural crest cells) and airway and alveolar stem cells changing their morphology. Morphologically, the terminal differentiation of AT1 cells in development has been shown to occur via a non-proliferative two-step process of cell flattening and cell folding<sup>51</sup>. We have shown that Krt8+ ADI cells in adult regeneration feature mostly squamous morphology and may thus correspond with this first phase of cell flattening. In the developmental cell folding phase, AT1 cells increase their size ten-fold to span multiple alveoli and establish the honeycomb alveolar structure in coordination with myofibroblasts and capillary vessels<sup>51</sup>. In this process, AT1 cells express a large number of morphogens, such as *Vegfa* and semaphorins that stimulate angiogenesis and thus likely play an active signaling role in the coordination of alveolar morphogenesis. We confirm the specific expression of these morphogens only in mature AT1 in our study and show in contrast that Krt8+ ADI express a distinct set of morphogens, including Endothelin-1 (Edn1) that likely serves the paracrine stimulation of capillary endothelial cells after injury.

In lung development, the generation of the distal epithelium has been proposed to be driven by a bipotent progenitor co-expressing both AT1 and AT2 markers<sup>52</sup>. Additionally, a recent scRNAseq analysis of the mouse lung epithelium at birth identified a similar AT1/AT2 cluster that may be interpreted as a bipotent progenitor state<sup>53</sup>. In our preliminary analysis, both published developmental signatures do not correspond well to the injury induced Krt8+ ADI signature. Additional experiments will be needed to better understand the differences of epithelial lineage trajectories in lung development versus adult homeostasis and regeneration. Interestingly, we do find rare Krt8 high alveolar cells in the parenchyma of the normal adult mouse lung. These cells are largely lineage labeled in the SPC-CreERT2 analysis but not in Sox2-CreERT2 labeling, suggesting that rare Krt8 high cells in normal homeostasis are derived from AT2 and are possibly a naturally occurring intermediate en route to AT1. We show that upon injury the bulk of AT2 cells and also airway cells differentiate into Krt8+ ADI, producing high frequencies of these cells without massive proliferation of rare stem cells at early time points.

The Krt8+ ADI cells display a highly distinct receptor-ligand connectome with mesenchyme and macrophages, and are a

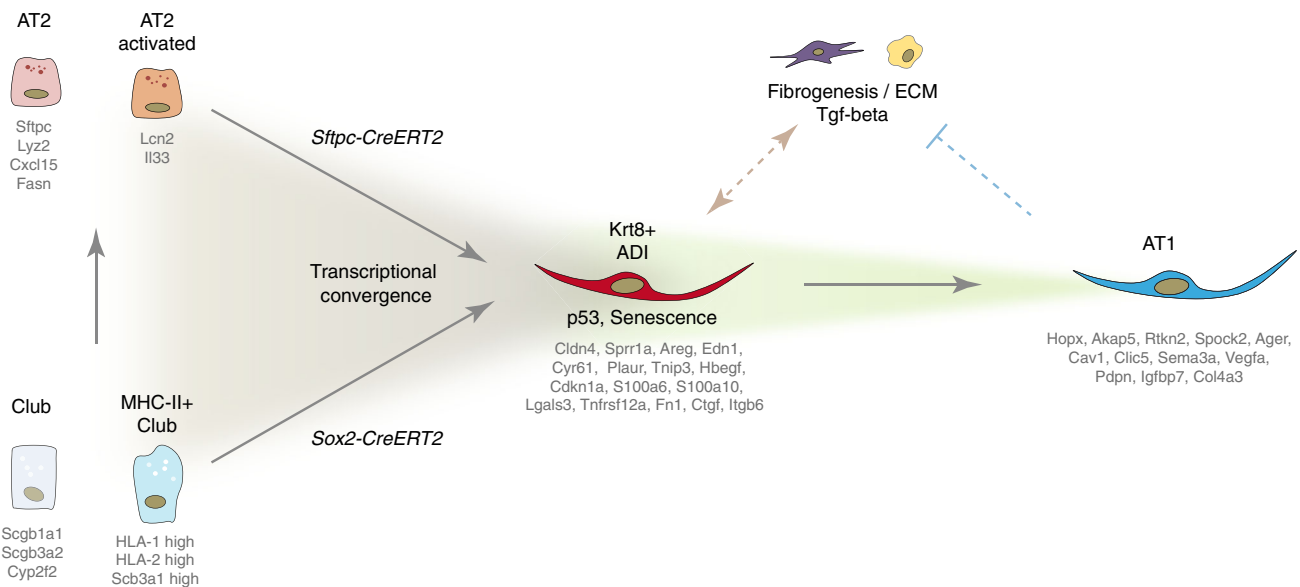




specific source of pro-fibrogenic factors such as *Ctgf*, *Itgb6*, *Areg*, *Hbegf*, *Edn1*, and *Lgals3*, all of which are antifibrotic targets that have been tested in pre-clinical and clinical studies. Thus, the availability of these factors (we have validated *Areg*, *Hbegf*, and *Itgb6* on protein level) during the fibrogenic phase around

10 days after injury is likely dependent on the Krt8+ ADI cell state, while its transient nature elegantly enables the system to temporally limit their expression. Many pathways that peaked in the Krt8+ ADI state represent environmental stress- and inflammation-induced gene programs, represented by their

**Fig. 9 Cells similar to Krt8+ADI persist in a mouse model of progressive lung fibrosis and human disease.** **a, b** Re-analysis of human lung fibrosis single cell data from GSE135893 for epithelial cells only. The indicated cell type identities (**a**) and disease status (**b**) show a relative increase of airway epithelial cell types in lung fibrosis (IPF) and appearance of a disease specific cell state termed aberrant KRT5<sup>-</sup>/KRT17<sup>+</sup> basaloid cell (arrow)<sup>42,45,79</sup>. **c, d** The indicated human (**c**) and mouse (**d**) gene signatures downloaded from the Gene Expression Omnibus were scored on single cells in our mouse epithelial data manifold. Higher scores indicate higher similarity in gene expression to the indicated signatures. **e** The matchScore matrix shows the degree of similarity of the indicated cell state signatures across species. **f** FFPE sections from non-fibrotic controls were stained against KRT8 (red), SFTPC (green), and ACTA2 (blue). Scale bar = 100 microns. **g** Human lung tissue sections were stained as in **f**, revealing pronounced KRT8 expression at the site of acutely injured lesions (ARDS diagnosis) and fibrotic regions of ILD patient lungs (IPAF, IPF and EAA diagnosis). Scale bar = 100 microns. **h** Fluorescence intensity of KRT8 stainings was quantified from representative areas of control tissue [ $n(\text{patients}) = 7, n(\text{areas}) = 36$ ], EEA tissue [ $n(\text{patients}) = 1, n(\text{areas}) = 5$ ], IPF tissue [ $n(\text{patients}) = 3, n(\text{areas per single patient}) = 5$ ], IPAF tissue [ $n(\text{patients}) = 1, n(\text{areas}) = 8$ ], Sarcoidosis tissue [ $n(\text{patients}) = 1, n(\text{areas}) = 8$ ], and ARDS tissue [ $n(\text{patients}) = 1, n(\text{areas}) = 8$ ]. One-way ANOVA statistical analysis:  $***p < 0.0001, **p = 0.0041$ . **i** FFPE sections from non-fibrotic controls or IPF patients were stained against KRT8 (red) and KRT17 (green). Scale bar = 50 microns; representative images from 2x IPF patients and 2x controls.



**Fig. 10 A revised model of alveolar regeneration.** We identify convergence of alveolar and airway stem cells on an injury-induced transitional cell state characterized by a unique transcriptional signature, including high levels of Krt8 expression, that precedes the regeneration of AT1 cells. In this process, stem cells lose cell identity genes, gain specific gene programs including p53 and NF $\kappa$ B target genes, and undergo a drastic change in shape towards a squamous morphology. Krt8<sup>+</sup> ADI cells feature a highly distinct connectome of receptor-ligand pairs with endothelial cells, fibroblasts, and macrophages. The Krt8<sup>+</sup> ADI cell state persists in models of progressive lung fibrosis and human IPF patients, suggesting that the cell state transitions described in this work are coordinated in space and time by cell intrinsic and tissue niche checkpoints that may be derailed in disease.

transcriptional master regulators, including Trp53, Atf3/4, Nupr1, Hif1a, NF $\kappa$ B and TGF- $\beta$ . The proliferation of AT2 cells after lung injury involves IL1- $\beta$  and Tnf- $\alpha$  driven NF $\kappa$ B activation and accordingly was lost in AT2-specific IL1-receptor knock-out mice<sup>54</sup>, which provides a molecular link between inflammation and epithelial regeneration that is consistent with our results. We propose that inflammatory stimuli can promote cell plasticity by inducing epithelial cell states with a higher susceptibility for alternative fate programs.

Our trajectory model predicts distinct transcriptional regulators as candidate switch points in terminal AT1 differentiation, including TAZ (Wwtr1), Sox4 and beta-catenin (Ctnnb1). The mechanistic importance of Wwtr1 (YAP/TAZ - Hippo pathway) in AT2 to AT1 transdifferentiation has recently been demonstrated by using small molecule inhibition and conditional knock-out in mouse lung organoids and in vivo injury experiments<sup>25,55</sup>. An important role of beta-catenin and the canonical Wnt signaling pathways has been suggested based on in vitro differentiation of isolated AT2 cells<sup>41</sup>. Moreover, the TGF- $\beta$  pathway has been proposed to mediate cell cycle arrest in AT2 cells followed by transdifferentiation into AT1 cells<sup>56</sup>. Here, we

found high Yap/TAZ activity in Krt8<sup>+</sup> ADI cells using immunostainings and reduced formation of Krt8<sup>+</sup> ADI and AT1 cell states upon Wnt/beta-catenin inhibition in vitro. A functional role of Sox4 in switching towards AT1 fate as suggested by our model awaits experimental validation.

Various potential stem cell sources for AT1 cells after injury have been described, including AT2 cells, bronchioalveolar stem cells (BASC) and p63(+)/Krt5(+) distal airway stem cells (DASC). We did not see an important contribution of Krt5<sup>+</sup> cells in our data, however, the Sox2-CreERT2 lineage tracing confirmed that after bleomycin injury a substantial fraction of Krt8<sup>+</sup> ADI (and AT1/AT2) was derived from airway cells. We found that Scgb1a1<sup>+</sup> club cells give rise to both AT2 and Krt8<sup>+</sup> ADI, with a MHC-II<sup>+</sup> subset of club cells showing a particularly strong connectivity with alveolar cell identity after injury. Comparing the signature of these MHC-II club cells with the recently described H2-K1-high epithelial progenitors<sup>30</sup> suggests that these cells are identical. Strong connectivity in PAGA analysis and the observed continuous trajectories in UMAP space indicate direct conversion of airway stem cells into Krt8<sup>+</sup> ADI, as presented in our model. However, we cannot formally exclude the possibility that airway stem cells initially give rise to AT2 and

subsequently differentiate towards Krt8+ ADI. Interestingly, we find that direct AT2 cell differentiation to Krt8+ ADI precedes the differentiation of airway stem cells, suggesting that these two processes happen at different locations, possibly reflecting heterogeneity in the local severity of injury and AT2 cell survival.

In idiopathic pulmonary fibrosis patients, the aberrant activity of p53, TGF- $\beta$ , Hippo and Wnt pathway genes has been reported<sup>57</sup>, and a p53/p21 mediated cellular senescence program in AT2 cells, which is also reflected in the Krt8+ ADI signature, was recently proposed to drive progressive lung fibrosis in mice<sup>46</sup>. Furthermore, local hypoxia signaling has been implicated in dysplastic abnormal epithelial barriers<sup>58</sup>, which we suggest may represent an accumulation of transitional or aberrant cell states blocked in their commitment towards AT2/AT1 cell fate. Our analysis shows that the transcriptional signature of KRT5–/KRT17+ basaloid cells<sup>45</sup> in IPF tissues is highly similar to the Krt8+ ADI described here. Based on these findings we propose that IPF in particular and chronic lung diseases in general may be rooted in defective molecular cell differentiation checkpoints that lead to aberrant persistence of (normally transient) regenerative intermediate cell states. Indeed, our quantitative analysis demonstrates that the Krt8+ ADI state is identical to a cell state that accumulates and persists in mice with AT2 cell specific deletion of the Rho GTPase Cdc42, which leads to progressive fibrosis similar to IPF after pneumectomy<sup>44</sup>. Thus, the defective terminal differentiation of stem cells into AT1 may be a key event in pathogenesis of progressive fibrosis in IPF patients. Future therapeutic approaches may specifically aim at (re)programming Krt8+ ADI into AT1 to avoid self-amplifying paracrine feedback loops in tissue regions that are still in the early stage of disease progression.

## Methods

**Mouse experiments-bleomycin treatment.** Pathogen-free female C57BL/6J mice were purchased from Charles River Germany and maintained at the appropriate biosafety level at constant temperature and humidity with a 12 h light cycle. Animals were allowed food and water ad libitum. Animal handling, bleomycin/PBS administration, and organ withdrawal were performed in accordance with the governmental and international guidelines and ethical oversight by the local government for the administrative region of Upper Bavaria (Germany), registered under 55.2-1-54-2532-130-2014 and ROB-55.2-2532.Vet\_02-16-208.

**Human tissues.** Resected lung tissue and lung explant material was obtained from the CPC-M bioArchive at the Comprehensive Pneumology Center (CPC), Munich. ILD diagnosed lung tissue ( $n = 6$ ) is derived from lung explant material obtained during lung transplantation, reflecting non-resolving end-stage fibrotic disease. Healthy control tissue ( $n = 7$ ) was derived from tumor resection in non-chronic lung disease (CLD) patients. The tissue section from a patient with ARDS ( $n = 2$ ) has been provided by the Institute of Pathology, Ludwigs Maximilians University, Munich.

All participants gave written informed consent; the study was approved by the local ethics committee of the Ludwig Maximilians University, Munich, Germany (#333-10).

**Experimental design and animal treatment.** Mice were divided randomly into two groups: (A) saline-only (PBS), or (B) bleomycin (Bleo). Lung injury and pulmonary fibrosis were induced by single-dose administration of bleomycin hydrochloride (Sigma Aldrich, Germany), which was dissolved in sterile PBS and given at 2 U/kg (oropharyngeal instillation) and 3U/kg (intratracheal instillation) bodyweight. The control group was treated with sterile PBS only. Mice were sacrificed at designated time points (days 1–14, 21, 28, 35, 56) after instillation. Treated animals were continuously under strict observation with respect to phenotypic changes, abnormal behavior and signs of body weight loss.

**Generation of single cell suspensions for whole-lung tissue.** Lung single cell suspensions were generated as previously described<sup>24</sup>. Briefly, after euthanasia, lung tissue was perfused with sterile saline through the heart and the right lung was tied off at the main bronchus. The left lung lobe was subsequently filled with 4% paraformaldehyde for later histologic analysis. Right lung lobes were removed, minced (tissue pieces at  $\sim 1$  mm<sup>2</sup>), and transferred for mild enzymatic digestion for 20–30 min at 37 °C in an enzymatic mix containing dispase (50 caseinolytic U/ml), collagenase (2 mg/ml), elastase (1 mg/ml), and DNase (30  $\mu$ g/ml). Single cells were

harvested by straining the digested tissue suspension through a 40 micron mesh. After centrifugation at 300 g for 5 min, single cells were taken up in 1 ml of PBS (supplemented with 10% fetal calf serum), counted and critically assessed for single cell separation and overall cell viability. For Dropseq, cells were aliquoted in PBS supplemented with 0.04% of bovine serum albumin at a final concentration of 100 cells/ $\mu$ l.

**Production of microfluidic devices for Dropseq.** Microfluidic devices needed for scRNAseq using the Dropseq platform were fabricated by means of standard soft lithography. In brief, by using photolithography, a polydimethylsiloxane (PDMS) master mold for the Dropseq device design (CAD file available as a download from: <http://mccarrollab.org/dropseq/>) was fabricated from a SU-8 photoresist (MicroChem, USA), and spin-coated on a 3" silicon wafer to generate 125  $\mu$ m-thick uniform layers. Afterwards, the master mold was filled with a 10:1 mixture of base to curing agent of the PDMS kit Sylgard 184 (Dow Corning, USA) and left at 60 °C in an oven for 4 h to crosslink the PDMS. After crosslinking, the PDMS replica was cut and peeled off from the master mold, as well as all necessary inlets/outlets for tubing connection were made in it using a 1 mm puncher. Next, the replica was sealed with a 2"  $\times$  3" microscopic slide, after the treatment of both in O<sub>2</sub> plasma. The assembled microfluidic device was treated with Aquapel (Pittsburgh Glass Works, USA) to make all inner surfaces evenly hydrophobic.

**Single cell RNA-sequencing using Dropseq.** Dropseq experiments were performed according to the original protocols<sup>24,27</sup>. Using the microfluidic device, single cells (100/ $\mu$ l) were co-encapsulated in droplets with barcoded beads (120/ $\mu$ l, purchased from ChemGenes Corporation, Wilmington, MA) at rates of 4000  $\mu$ l/h. Droplet emulsions were collected for 10–20 min/each prior to droplet breakage by perfluorooctanol (Sigma-Aldrich). After breakage, beads were harvested and the hybridized mRNA transcripts reverse transcribed (Maxima RT, Thermo Fisher; Template-switch oligonucleotide primer: AAGCAGTGGTATCAACGCAGAGTG AATrGrGrG (50  $\mu$ M)). Unused primers were removed by the addition of exonuclease I (New England Biolabs), following which, beads were washed, counted, and aliquoted for pre-amplification (2000 beads/reaction, equals ca. 100 cells/reaction) with 12 PCR cycles (Smart PCR primer: AAGCAGTGGTATCAACGCAGAGT (100  $\mu$ M), 2x KAPA HiFi Hotstart Ready-mix (KAPA Biosystems), cycle conditions: 3 min 95 °C, 4 cycles of 20 s 98 °C, 45 s 65 °C, 3 min 72 °C, followed by 8 cycles of 20 s 98 °C, 20 s 67 °C, 3 min 72 °C, then 5 min at 72 °C)<sup>27</sup>. PCR products of each sample were pooled and purified twice by 0.6x clean-up beads (CleanNA), following the manufacturer's instructions. Prior to tagmentation, complementary DNA (cDNA) samples were loaded on a DNA High Sensitivity Chip on the 2100 Bioanalyzer (Agilent) to ensure transcript integrity, purity, and amount. For each sample, 1 ng of pre-amplified cDNA from an estimated 1000 cells was tagmented by Nextera XT (Illumina) with a custom P5-primer (Integrated DNA Technologies; primer sequence: AATGATACGGCGACCACCGAGATCTACACGCTGTCCG CGGAAGCAGTGGTATCAACGCAGAGT\*<sup>\*</sup>A\*<sup>\*</sup>C (10  $\mu$ M)). Single-cell libraries were sequenced in a 100 bp paired-end run on the Illumina HiSeq4000 using 0.2 nM denatured sample and 5% PhiX spike-in. For priming of read 1, 0.5  $\mu$ M Read1CustSeqB (primer sequence: GCCTGTCCGCGGAAGCAGTGGTATCAAC GCAGAGTAC) was used.

Quality metrics, including the number of unique molecular identifiers (UMI), genes detected per cell and reads aligned to the mouse genome were comparable across all mice (Supplementary Fig. 1). Every time point was analyzed together with control mice that were instilled with phosphate-buffered saline (PBS). UMI-based counting of mRNA copies was used to determine differential gene expression between single cells. We used the six batches of PBS control mice to exclude dominant batch effects observing very good overlap across mouse samples (Silhouette coefficient:  $-0.08$ ) (Supplementary Fig. 1).

**Processing of the whole-lung data set.** For the whole-lung data set, the Dropseq computational pipeline was used (version 2.0) as described by Macosko et al.<sup>20</sup>. Briefly, STAR (version 2.5.2a) was used for mapping<sup>59</sup>. Reads were aligned to the mm10 reference genome (provided by the Dropseq group, GSE63269). For barcode filtering, we excluded barcodes with <200 detected genes. As 1000 cells were expected per sample, the first 1200 cells were used before further filtering. A high proportion (>10%) of transcript counts derived from mitochondria-encoded genes may indicate low cell quality, and we removed these unqualified cells from downstream analysis. Cells with a high number of UMI counts may represent doublets, thus only cells with less than 5000 UMIs were used in downstream analysis.

**Analysis of the whole-lung data set.** The computational analysis of the whole-lung data set was largely performed using the R package Seurat<sup>60</sup>. Count matrices were merged using Seurat version 2.3. The merged expression matrix was normalized using the Seurat NormalizeData() function. To mitigate the effects of unwanted sources of cell-to-cell variation, we regressed out the number of UMI counts using the Seurat function ScaleData(). Highly variable genes were calculated per sample, selecting the top 7000 genes with a mean expression between 0.01 and 8. After excluding homologs of known cell-cycle marker genes<sup>61</sup>, a total of 18893 genes were subjected to independent component analysis. The first 50 independent



components were used as input to the FindClusters() function with the ‘resolution’ parameter set to two and the RunUMAP() function with the “n\_neighbors” parameter set to ten.

**Multi-omic data integration:** to confirm global expression changes observed at the single-cell level, we integrated previously published bulk RNAseq and proteomics data obtained from whole-mouse lungs 14 days after bleomycin-induced injury and controls<sup>24</sup>. Multi-omic data integration was performed following previous work<sup>23,24</sup>. Briefly, in silico bulk samples were generated by summing all counts within a mouse sample. Both the in silico bulk and whole-lung tissue bulk data were normalized using the voom() function of the limma R package<sup>62</sup>. Next, in silico bulk, whole-lung tissue bulk, and proteomics data were merged on a set of genes present in all three data sets and quantile normalized. This merged and quantile normalized expression matrix was then subjected to principal component analysis (PCA).

**Bulk deconvolution analysis:** to interpret the expression changes observed in the bulk RNAseq data at the cellular level and to validate the cell type frequency changes observed at the single cell level, we performed deconvolution analysis. Fold changes between the bleomycin model at day 14 and controls were obtained from Table EV4 from Schiller et al.<sup>24</sup>. For each cell type, marker genes with average fold change greater than zero and adjusted *p*-value < 0.25 were tested for enrichment in the fold changes by comparison to all other genes using the Kolmogorov-Smirnov test. For visualization purposes the minimum *p*-value was set to 1e-50.

**Discovery of cell type identity marker genes:** to identify cluster-specific marker genes, the Seurat FindAllMarkers() function was applied, restricted to genes detected in more than 10% of cells and with an average fold change difference of 0.25 or more. Based on these derived marker genes and manual curation we assigned all clusters to cell type and meta-cell type identities (Supplementary Fig. 1d). Cell type frequencies were calculated by dividing the number of cells annotated to a specific cell type identity, by the total number of cells for each mouse sample. In droplet-based scRNAseq data, background mRNA contamination by the so-called “ambient RNAs” is frequently observed. These mRNAs are believed to stem from dying cells which release their content upon cell lysis. This contamination is distributed to many droplets and leads to a blurred expression signal that does not stem solely from the single cell in the droplet but also from the solution that contains it. We used the function inferNonExpressedGenes() from SoupX<sup>63</sup> to identify a set of 80 ambient RNAs and accounted for these in the downstream analysis.

**Time course differential expression analysis:** to identify genes that show differential expression patterns across time within a given cell type we performed the following analysis. We used the R packages splines and lmer for our modeling approach. First, we manually combined the Louvain clusters into 26 cell types to generate a more coarse grained cell type annotation for the time course differential expression analysis (Supplementary Fig. 1d). Within each of these groups we modeled gene expression as a binomial response where the likelihood of detection of each gene within each mouse sample was the dependent variable. Therefore, the sample size of the model was the number of mouse samples ( $n = 28$ ) and not the number of cells. To assess significance we performed a likelihood-ratio test between the following two models. For the first model, the independent variables contained an offset for the log-transformed average total UMI count and a natural splines fit of the time course variable with two degrees of freedom. The independent variables of the second model just contained the offset for the log-transformed average total UMI count. The dependent variable of both models was the number of cells with UMI count greater than zero out of all cells for a given cell type and mouse sample. To account for potential false positive signal derived from ambient RNA levels, we calculated cell type marker genes for the 26 cell type annotation using the Seurat FindAllMarkers() function. For all 80 candidate ambient RNAs, we consequently set all regression *p*-values to one in cell types where the gene was not simultaneously a marker gene with an adjusted *p*-value of < 0.1 and a positive average log fold change.

**Cell-cell communication analysis:** to identify cell-cell communication networks, we downloaded a list of annotated receptor-ligand pairs<sup>64</sup>. Next, we integrated this information with the cell type marker genes from Supplementary Data 1. Cell-cell communication networks were generated in the following manner. An edge was created between two cell types if these two cell types shared a receptor-ligand pair between them as marker genes.

**Macrophage analysis:** it is not entirely understood whether monocyte-derived macrophages contribute to the development of lung fibrosis. To see if our data reflects published models of monocyte recruitment, we integrated bulk RNAseq data from FACS sorted macrophage populations after bleomycin-induced lung fibrosis<sup>8,65</sup>. This data set contained bulk RNAseq gene expression of tissue-resident alveolar macrophages (TR-AMs), monocyte-derived alveolar macrophages (Mo-AMs), interstitial macrophages (IM), and monocytes (Mono) for both day 14 and day 19 after bleomycin injury, including additional measurements for TR-AMs at day 0. To derive a gene expression signature from the bulk RNAseq data, we used the R package limma<sup>66</sup>. We followed the standard limma workflow<sup>65</sup> to find genes which are differentially expressed between these four populations. Next, we subset our scRNAseq data set to only clusters expressing known macrophage markers and selected a new set of variable genes. Following this the PCA and UMAPs were recreated for this subset, using 20 PCs and 20 n\_neighbors in Seurat’s functions. The macrophages from our data were scored according to their similarity to these bulk-derived signatures using Pearson correlation. For each of the four bulk-

derived groups, the log fold changes of the 500 most differentially expressed genes were correlated with the scaled expression values of each macrophage cell in our scRNAseq data. To separate potential monocyte-derived macrophages from interstitial macrophages, we assigned each cell to the category with the higher correlation coefficient as long as the difference was > 0.05. Otherwise, the cell was labeled unassigned.

**Processing of the high-resolution epithelial data set.** The high-resolution gene expression matrix was generated as specified for the whole-lung data set with the following changes. To lessen the technical bias introduced by ambient RNA, we applied SoupX the pCut parameter set to 0.3 within each sample before merging the count matrices together. The merged expression table was then pre-processed as described in the section “Processing of the whole-lung data set”, with minor alterations. To account for the fact that a certain fraction of the counts was removed, the upper threshold for the number of total UMI counts per cell was set to 3000.

**Analysis of the high-resolution epithelial data set.** The computational analysis of the whole-lung data set was performed using a combination of the Seurat<sup>60</sup> and Scanpy<sup>67</sup> code. Cell-cycle effects, the percentage of mitochondrial reads, and the total number of UMI counts are often viewed as unwanted sources of variation and were therefore regressed out using the Seurat functions CellCycleScoring() and ScaleData(). Genes which had a variable expression in at least two samples (17038 genes) were used for the principal component analysis. The majority of the cells were airway and alveolar epithelial cells, although non-epithelial cells were also captured. To filter the data further, the cells were clustered and clusters expressing non-epithelial markers were excluded from the data set. The cleaned object was then converted to a h5ad file for downstream analysis using the python package Scanpy. The aligned bam files were used as input for Velocyto<sup>14</sup> to derive the counts of unspliced and spliced reads in loom format. Next, the sample-wise loom files were combined, normalized and log transformed using scvelos (<https://github.com/theislab/scvelo>)<sup>68</sup> functions normalize\_per\_cell() and log1p(). After merging the loom information to the exported h5ad file using scvelos merge() function the object was scaled and the neighbourhood graph constructed with Batch balanced KNN (BBKNN)<sup>69</sup> to account for the different PCR cycles used in the experiment with neighbors\_within\_batch set to 15 and n\_pcs to 40. Two dimensional visualization and clustering was carried out with the Scanpy functions t.louvain() at resolution two and t.umap(). The neuroendocrine cells (NEC) formed a distinct cluster in the UMAP, however, they were only assigned to a single cluster at higher resolutions. To separate them from basal cells we captured the NEC with dbSCAN using the UMAP coordinates and assigned them as cluster 21. After manual curation of the markers the remaining 20 clusters were combined, leading to thirteen final meta cell types. Relative frequencies were calculated as described for the whole-lung data set. To better visualize the dynamic changes of each cell type over time, values were scaled to a minimum of 0 and a maximum of 1 using numpy’s interp() function for each cell type annotation separately. Smoothed line plots of the scaled frequencies were generated by employing the lmplof() function of the python module seaborn with default parameters.

**Cell-cycle analysis:** the proliferating cells (Louvain cluster 14, Fig. 4d) of the high-resolution data set were subjected to cell type deconvolution analysis. Cell cycle phases (S.Score, G2M.Score) were regressed out using the Seurat ScaleData() function. Next, PCA was calculated using all unique marker genes from Supplementary Data 3 and the Seurat RunPCA() function. UMAP embedding and Louvain clusters were calculated using the first 20 principal components with the Seurat RunUMAP() and FindClusters() functions, respectively. Upon manual curation of the marker genes for the generated embedding, we identified four distinct clusters. Next, the frequency of proliferating cells was calculated by dividing the number of cells in cluster 14, by the number of total cells for each mouse sample.

**PAGA analysis:** to assess the global connectivity topology between the Louvain clusters we applied Partition-based graph abstraction (PAGA)<sup>28</sup>. We applied the t.paga() function integrated in the Scanpy package to calculate connectivities and used the Louvain clusters as partitions. The weighted edges represent a statistical measure of connectivity between the partitions. Connections with a weight < 0.3 were removed.

**Velocity analyses:** to infer future states of individual cells we made use of the spliced and unspliced information. We employed scvelo<sup>68</sup> (<https://github.com/theislab/scvelo>). The previously normalized and log transformed data was the starting point to calculate first and second order moments for each cell across its nearest neighbors (scvelo.pp.moments(n\_pcs = 40, n\_neighbors = 15)). Next, the velocities were estimated and the velocity graph constructed using the scvelo.tl.velocity() with the mode set to ‘stochastic’ and scvelo.tl.velocity\_graph() functions. Velocities were visualized on top of the previously calculated UMAP coordinates with the scvelo.tl.velocity\_embedding() function. To compute the terminal state likelihood of a subset of cells, the function scvelo.tl.terminal\_states() with default parameters was used.

**Trajectory differential expression analysis:** to identify genes showing significantly altered expression along the differentiation trajectories toward the Krt8+ cell state, the following approach was used. The high-resolution data set was restricted to cells from Louvain clusters 2, 10, 11, and 12 for the convergence and AT1

trajectories. The convergence (Louvain clusters 2, 10, 11) and AT1 (Louvain clusters 2, 12) trajectories were analyzed independently. For the convergence and AT1 trajectories we used diffusion map and UMAP as the cellular embeddings, respectively. The `dbSCAN()` function from the DBSCAN R package was used to identify outlier cells which were subsequently removed from further analysis. The R package `slingshot`<sup>70</sup> was used to infer the pseudotemporal ordering along the trajectory of the cellular embeddings of all remaining cells. Next, the analysis was restricted to genes detected in at least 5% of cells. The R package `tradeSeq`<sup>71</sup> was used to identify genes differentially expressed along the trajectories. Despite the fact that *p*-values derived from pseudotemporal analyses are inflated they can be used to prioritize candidate genes. Heatmaps were restricted to genes with Benjamini-Hochberg adjusted *p*-values < 0.05. Gene expression patterns along pseudotemporal trajectories were visualized using local polynomial regression fitting as implemented in the R `loess()` function with default parameters.

**In silico doublet simulation:** to exclude the potential artefacts derived from cell doublets we performed the following analyses. In silico bulk samples were generated by summing the counts across cells randomly sampled from specific cell clusters or mixtures thereof. More precisely, we randomly selected 600 cells from AT1, AT2, club and Krt8+ ADI cell clusters in silico bulk samples per cell identity. Doublets were generated by randomly selecting 300 cells from the AT2 and AT1 clusters as well as Club and AT1 clusters and subsequently aggregated into in silico samples. This procedure was repeated five times to generate five in silico samples per condition. Counts were normalized using the `voom()` function of the `limma` R package and subjected to principal component analysis. Analogous procedure was performed for the club cell analysis.

**Integration of whole-lung and high resolution data sets:** epithelial cells of the whole-lung data set were re-analysed to validate findings derived from the high resolution data set. The principal components were re-calculated on this subset using a new set of variable genes, in order to emphasize changes in the epithelium specifically. Following the procedure described above, UMAP visualization and RNA velocities were generated using `Scanpy` and `scvelo`.

**Pathway analysis.** To predict the activity of pathways and cellular functions based on the observed gene expression changes, we used the Ingenuity Pathway Analysis platform (IPA, QIAGEN Redwood City, [www.qiagen.com/ingenuity](http://www.qiagen.com/ingenuity)) as previously described<sup>24</sup>. The analysis uses a suite of algorithms and tools embedded in IPA for inferring and scoring regulator networks upstream of gene-expression data based on a large-scale causal network derived from the Ingenuity Knowledge Base. We used the upstream regulator tool in IPA to derive pathway *z*-scores across cell type identities by loading the marker gene list fold changes of our single cell louvain clusters (logFC relative to all other cells) for comparison of the indicated cell type identities. The missing values represent cell type signatures that did not have significant overlap with the respective pathways in IPA. The upstream regulator tool in IPA defines an overlap *p* value measuring enrichment of network-regulated genes in the data set, as well as an activation *Z*-score which can be used to find likely regulating molecules based on a statistically significant pattern match of up- and down-regulation, and also to predict the activation state (either activated or inhibited) of a putative regulator. In our analysis we considered pathways/genes with an overlap *p* value of > 7 (log10) that had an activation *Z*-score > 2 as activated and those with an activation *Z*-score < -2 as inhibited.

**Magnetic-activated cell sorting.** Cells from whole-lung single cell suspensions were strained using a 40 µm mesh size and red blood cells were eliminated by lysis (RBC lysis buffer, ThermoFisher). For positive epithelial cell selection, cells were stained with CD326-AlexaFluor647 antibody (Biolegend, 118212) for 30 min at 4 °C in the dark, and after washing, incubated with microbeads specific against AlexaFluor647 (Miltenyi Biotec, 130-091-395) for 15 min at 4 °C. MACS LS columns (Miltenyi Biotec, 130-042-401) were prepared according to the manufacturer's instructions. Cells were applied to the columns and positively-labeled epithelial cells were retained in the column. The flow-through was collected separately for later mesenchymal cell enrichment (negative magnetic-activated cell sorting (MACS) selection) and kept on ice. Epithelial cells were eluted from the LS columns and used for either Dropseq runs. Mesenchymal cells from the flow-through were further enriched by negative depletion of CD31+ (Invitrogen, 17-0311-82), CD45+ (Biolegend, 103112), Lyve1+ (Invitrogen, 50-0443-82), Ter119+ (Biolegend, 116218), and CD326+ cells (Biolegend, 118212). After antibody staining, 100 µl per 10 million cells of MACS dead cell removal beads (Miltenyi Biotec, 130-090-101) were added and incubated according to the product's accompanying protocols. Depletion of undesired cell types was achieved by the use of microbeads specific for APC (Miltenyi Biotec, 130-090-855), which ensured magnetic retention of these cells. Likewise to epithelial cells, negatively-selected mesenchymal cells were applied to the Dropseq workflow.

**Flow cytometry.** Isolated total lung cell suspensions were used to detect and quantify cell populations by flow cytometry. After depletion of red blood cells by red blood cell lysis buffer (Invitrogen, ThermoFisher), cell suspensions were stained with anti-mouse CD45-PE-Vio770 (Miltenyi Biotec, 130-110-661), CD326-BV421 (Biolegend, 118225), Krt8/TROMA-I (DSHB-Developmental Studies Hybridoma Bank at the University of Iowa), and αvβ6-specific monoclonal antibody 6.3G9

(Itgb6-3G9; kindly provided by Prof. Dr. Dean Sheppard, available through Biogen Idec, USA). Cells were stained for surface markers in the dark at 4 °C for 20 min, followed by cell fixation and permeabilization (Fix & Perm, Life Technologies, G45004) for intracellular staining of Krt8. Epithelial cells were selected using the CD45-negative fraction of the cell isolate that stained positively for CD326. Within the epithelial cell gate, Krt8+, Itgb6+, or Krt8+/Itgb6+ cells were identified and quantified by their geometric mean fluorescence signal intensity. For exclusion of non-specific antibody binding and autofluorescence signal, fluorescence minus one (FMO) controls were included in the measurement. All stainings were performed per 1,000,000 cells in the following dilutions: CD326 (1:500), CD45 (1:20), Krt8 (1:35), Itgb6 (1:1000). Data was acquired in a BD LSRII flow cytometer (Becton Dickinson, Heidelberg, Germany) and analyzed by mean fluorescence intensity (MFI) using the FlowJo software (TreeStar Inc., Ashland, OR, USA). Negative thresholds for gating were set according to isotype-labeled and unstained controls.

**Precision cut lung slices (PCLS).** Precision cut lung slices were generated as previously described<sup>72</sup>. Briefly, using a syringe pump, the mouse lungs were filled via a tracheal cannula with 2% (w/v) warm, low gelling temperature melting point agarose (Sigma Aldrich, A9414) in sterile DMEM/Ham's F12 cultivation medium (Gibco, 12634010), supplemented with 100 U/ml penicillin, 100 µg/ml streptomycin, and 2.5 µg/ml amphotericin B (Sigma Aldrich, A2942). Afterwards, the lungs were removed and transferred on ice in cultivation medium for 10 min to allow for gelling of the agarose. Each lung lobe was separated and cut with a vibratome (Hydrax V55; Zeiss, Jena, Germany) in 300 µm thick sections. The PCLS were immediately fixed in -20 °C-cold methanol for 20 min and subsequently stained for immunofluorescence microscopy.

**Immunofluorescence microscopy of PCLS and analysis.** Methanol-fixed PCLS were stained and imaged as previously described<sup>73</sup>. Shortly, primary antibodies were diluted in 1% bovine serum albumin (BSA, Sigma Aldrich, 84503) in PBS (1:100), incubated for 16 h at 4 °C and subsequently washed three times with PBS for 5 min each. Secondary antibodies were diluted in 1% bovine serum albumin in PBS (1:200), incubated for 4 h at room temperature and subsequently washed three times with PBS for 5 min each. Primary antibodies were: rat anti-Krt8/TROMA-I (1:200; DSHB-Developmental Studies Hybridoma Bank at the University of Iowa), rabbit anti-pro-SPC (1:200; Millipore, AB3786), goat anti-Pdpn (1:200; R&D Systems, AF3244). Cell nuclei were stained with DAPI (40,6-diamidino-2-phenylindole, Sigma-Aldrich, 1:2,000). Confocal high-resolution 3D imaging of the PCLS was accomplished by placing the PCLS into a glass-bottomed 35 mm CellView cell culture dish (Greiner BioOne, 627870) as a wet chamber. Images of PCLS were acquired as *z*-stacks using an inverted microscope stand with an LSM 710 (Zeiss) confocal module operated in multitrack mode using the following objectives: Plan-Apochromat W 40x/1.0 M27 and Plan-Apochromat W 63x/1.3 M27. The automated microscopy system was driven by ZEN2009 (Zeiss) software, version 5.5. The acquired confocal fluorescent *z*-stacks were surface rendered in Imaris 9.3 software (Bitplane) and its statistical analysis tool (MeasurementPro) was used for 3D cell shape analysis using morphometric parameter sphericity as a readout (a value of 1 corresponds to a perfect sphere).

**Immunofluorescence microscopy.** After euthanasia, mouse lungs were immediately inflated with 4% paraformaldehyde. For frozen OCT embedding, tissue was fixed for 1 h at room temperature. Thin lung sections (7 µm) were cut on a cryostat. Sections were incubated with 0.1% sodium borohydride (PBS) to reduce background fluorescence, followed by blocking in PBS plus 1% bovine serum albumin, 5% non-immune horse serum (UCSF Cell Culture Facility), 0.1% Triton X-100 (Carl Roth, 3051.3) and 0.02% sodium azide (Sigma Aldrich, S2002). Slides were then incubated in primary antibodies overnight at 4 °C followed by secondary antibody incubation at 1:1,000 dilutions at room temperature for > 1 h. Slides were counterstained with 1 µM DAPI for 5 min at room temperature and mounted using Prolong Gold (Life Technologies, P36930). The following antibodies were used: rabbit anti-pro-SPC (1:2,500; Millipore, AB3786), and rat anti-Krt8 (0.9 µg/ml; TROMA-I (Krt8); University of Iowa Hybridoma Bank). Slides were imaged using a Leica Microscope (DM6B-Z; Leica Biosystems) or Axiovision Imager M1 (Carl Zeiss AG).

For formalin-fixed, paraffin-embedded (FFPE) lung tissue, sections were cut at 3.5 µm, followed by deparaffinization, rehydration, and antigen retrieval by pressure-cooking (30 s at 125 °C and 10 s at 90 °C) in citrate buffer (10 mM, pH 6.0). After blocking for 1 h at room temperature with 5% bovine serum albumin, lung sections were incubated in primary antibodies overnight at 4 °C, followed by secondary antibody (1:250) incubation for 2 h at room temperature. The following primary (1) and secondary (2) antibodies were used: (1) rat anti-Krt8 (170 µg/ml; University of Iowa Hybridoma Bank, 1:200), rabbit anti-pro-SPC (1:200; Millipore, AB3786), goat anti-Pdpn (1:200; R&D Systems, AF3244), rabbit anti-SPC (1:150; Sigma-Aldrich, HPA010928), mouse anti-alphaSMA (1:1,000, Sigma-Aldrich, A5228), rabbit anti-Areg (1:50; LSBIO, LS-B13911), rabbit anti-Hbegf (1:200; Bioss Antibodies, bs-3576R), rabbit anti-Ki67 (1:200; Abcam, ab16667), mouse anti-CC10 (1:200; Santa Cruz, sc-365992), rabbit anti-Cst3 (1:100; Abcam, ab109508), rabbit anti-Yap (1:500; Abcam, ab205270), rabbit anti-pSmad2 (Ser465/467) (1:1000; Cell Signaling, 3101), rabbit anti-Krt17 (1:200; Sigma, HPA000452); (2) donkey anti-rabbit AlexaFluor568 (Invitrogen, A10042), donkey anti-rat

AlexaFluor488 (Invitrogen, A21208), donkey anti-goat AlexaFluor647 (Invitrogen, A21447), goat anti-mouse AlexaFluor647 (Invitrogen, A21236). Images were acquired with an LSM 710 microscope (Zeiss).

**Microscopic image analysis-quantification.** The fluorescence intensity of Krt8 expression in selected regions of immunofluorescence microscopy images was measured excluding airways using FIJI (ImageJ, v.1.8.0)<sup>74</sup>. For quantification of Krt8 expression in the human FFPE sections, and likewise, for the Hbepf and Areg quantification in the mouse sections, the mean overall fluorescence intensities were measured. For quantification of cell proliferation, cells were stained with Ki67 and Krt8 and counted manually for Ki67 positive cells.

**Lineage tracing experiments.** SPC-CreERT2 (Sftpc<sup>tm1(cre/ERT2,rtTA)Hap</sup>) mice were crossed with Gt(ROSA)26Sor<sup>tm4(ACTB-tdTomato,-EGFP)Luo</sup> mice. Sox2-CreERT2 mice were crossed with Ai14-tdTomato (Gt(ROSA)26Sor<sup>tm14(CAG-tdTomato)Hze</sup>). Four doses (SPC-CreERT2) or three doses (Sox2-CreERT2) of 0.25 mg/kg body weight tamoxifen in 50 µl corn oil. A chase period of >21 days was used to ensure the absence of residual tamoxifen before injury. Bleomycin (1.5 U/kg) was delivered to mouse lungs via oral aspiration in 40 µL sterile PBS. Lungs were harvested at 10 days or 14 days following injury.

**Edu labeling and Sox2-CreERT2.** Sox2-CreERT2/tidTomato mice were labeled with tamoxifen dissolved in corn oil (3 doses, 250 mg/kg) followed by two weeks of chase before injuring with bleomycin dissolved in 1X PBS (2.1U/kg). Proliferating cells were labeled with 5-Ethynyl-2'-deoxyuridine (Edu) every other day starting two days after bleomycin injury (50 mg/kg dissolved in 1X PBS, intraperitoneal injection). Lungs were harvested twenty days post bleomycin injury, embedded in optimal cutting temperature compound (OCT) and stained for Edu using Click-iT Edu Imaging kit (ThermoFisher, c10086). Images were quantified by counting the total number of proliferating AEC1s (Edu+/RAGE1+) in two-three lobes/mouse ( $n = 2$  mice). Each dot represents one large region (>600 DAPI+ cells each) from one lobe.

**Uninjured labeling.** SPC-CreERT2 or Sox2-CreERT2 mice were labeled with tamoxifen (4 or 3 doses at 250 mg/kg, respectively). Lungs were harvested at least one week post last dose of tamoxifen and OCT embedded sections were stained for KRT8. At least one lobe/mouse was quantified for SPC-CreERT2 mice ( $n = 3$ ). Each dot represents quantification of a large region from one lobe of a mouse. We found no labeling of alveolar cells in Sox2-CreERT2 mice.

**Hypoxia/Hyperoxia+InfA infection model.** Wild-type or bi-transgenic Sftpc<sup>CreERT2</sup>; Rosa26R<sup>mTmG</sup> mice were exposed to 12% (hypoxia), 21% (room air) or 100% (hyperoxia) oxygen between postnatal days 0–4<sup>75</sup>. All mice were then exposed to room air until they were 8 weeks old. Sftpc<sup>CreERT2</sup>; Rosa26R<sup>mTmG</sup> mice were administered tamoxifen (Sigma Aldrich, T5648) (0.25 g/kg) or corn oil vehicle by single daily injections for four consecutive days<sup>76</sup>. On the seventh day, the mice were infected with influenza A virus (HKx31, H3N2) and lungs were harvested on post-infection day 14. Lungs were inflation fixed overnight in 10% neutral buffered formalin, embedded in paraffin, sectioned and stained with antibodies against pro-SPC (Seven Hills Bioreagents, Cincinnati, OH);

Tialpha (1:100; Syrian Hamster, clone 8.1.1, DSHB-Developmental Studies Hybridoma Bank at the University of Iowa) and Krt8/TROMA-I (DSHB-Developmental Studies Hybridoma Bank at the University of Iowa); Sections were incubated with fluorescently labeled secondary antibody and stained with 4', 6-diamidino-2-phenylindole (DAPI). Slides were visualized with a Nikon E-800 fluorescence microscope (Nikon Instruments, Microvideo Instruments, Avon, MA). Images were captured with a SPOT-RT digital camera (Diagnostic Instruments, Sterling Heights, MI).

**pmATII isolation and culture.** Primary mouse ATII cells (pmATII) were isolated from 8 to 10 week-old, pathogen-free, female C57BL6/N mice (Charles River Laboratories, SUzfeld, Germany) as previously described<sup>59,77</sup>. Briefly, lungs were filled with dispase (Corning, New York, NY, USA) and low-gelling temperature agarose (Sigma Aldrich, Saint Louis, MO, USA) before mincing and filtering through 100-, 20-, and 10-µm nylon meshes (Sefar, Heiden, Switzerland). Fibroblasts were depleted by adherence on non-coated plastic plates. Macrophages and white blood cells were depleted using CD45-specific magnetic beads (Miltenyi Biotec, Bergisch Gladbach, Germany), and endothelial cells with CD31-specific magnetic beads, respectively. Cell depletion was performed according to the manufacturer's instructions. pmATII cells were resuspended in DMEM containing 10% FCS (PAA Laboratories, Pasching, Austria), 2 mM glutamine, 1% penicillin/streptomycin (both Life Technologies, Carlsbad, CA), 3.6 mg/ml glucose (Applchem GmbH, Darmstadt, Germany) and 10 mM HEPES (PAA Laboratories), and cultured for 24 h to allow for cell attachment. The medium was changed to medium containing 7.5 µM ICG-001 (Biomol) or the respective DMSO control, refreshed at day 3. Cells were cultured up to 5 days.

**Western blotting.** Cells were washed with PBS (PAA Laboratories), lysed in T-PER lysis buffer (Thermo Fisher Scientific, Waltham, MA), supplemented

with proteinase inhibitor cocktail tablets (Roche, Germany). Protein concentration was quantified using the Pierce BCA Protein Assay Kit (Pierce, Thermo Fisher Scientific) according to the manufacturer's instructions. In all, 10 µg of protein lysates were separated on SDS-polyacrylamide gel and transferred to nitrocellulose membranes. Membranes were blocked with 5% non-fat dried milk solution in TRIS-buffered saline containing 0.01% (v/v) Tween (TBS-T) (Applchem, Darmstadt, Germany) for 1 h and incubated with primary T1α (R&D Systems) or Krt8/TROMA-I (DSHB-Developmental Studies Hybridoma Bank at the University of Iowa) antibody at 4 °C overnight. Next, blots were incubated for 1 h at RT with secondary, HRP-conjugated, antibodies (GE-Healthcare), or HRP-conjugated anti-β-actin antibody (Sigma-Aldrich) prior to visualization of the bands using chemiluminescence reagents (Pierce ECL, Thermo Scientific, Ulm, Germany). Blots were recorded with the ChemiDocTMXR+ system and analyzed using the Image Lab 6.0.1 software (Biorad, Munich, Germany).

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

## Data availability

Bulk and scRNA-seq data are available via the Gene Expression Omnibus with the accession code GSE141259. Additionally, results can be explored using our interactive webtool at <https://theislab.github.io/LungInjuryRegeneration>.

## Code availability

Code to reproduce the analyses described in this manuscript can be accessed via: [https://github.com/theislab/2019\\_Strunz](https://github.com/theislab/2019_Strunz).

Received: 17 March 2020; Accepted: 24 June 2020;

Published online: 16 July 2020

## References

- 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.* <https://doi.org/10.1165/rcmb.2018-0416TR> (2019)
- Hogan, B. L. M. et al. Repair and regeneration of the respiratory system: complexity, plasticity, and mechanisms of lung stem cell function. *Cell Stem Cell* **15**, 123–138 (2014).
- Logan, C. Y. & Desai, T. J. Keeping it together: pulmonary alveoli are maintained by a hierarchy of cellular programs. *Bioessays* **37**, 1028–1037 (2015).
- Herriges, M. & Morrisey, E. E. Lung development: orchestrating the generation and regeneration of a complex organ. *Development* **141**, 502–513 (2014).
- Lodyga, M. et al. Cadherin-11-mediated adhesion of macrophages to myofibroblasts establishes a profibrotic niche of active TGF-β. *Sci. Signal.* **12**, ea03469 (2019).
- Gieseck, R. L. 3rd, Wilson, M. S. & Wynn, T. A. Type 2 immunity in tissue repair and fibrosis. *Nat. Rev. Immunol.* **18**, 62–76 (2018).
- El Agha, E. et al. Two-way conversion between lipogenic and myogenic fibroblastic phenotypes marks the progression and resolution of lung fibrosis. *Cell Stem Cell* **20**, 571 (2017).
- Misharin, A. V. et al. Monocyte-derived alveolar macrophages drive lung fibrosis and persist in the lung over the life span. *J. Exp. Med.* **214**, 2387–2404 (2017).
- Barkauskas, C. E. et al. Type 2 alveolar cells are stem cells in adult lung. *J. Clin. Invest.* **123**, 3025–3036 (2013).
- Vaughan, A. E. et al. Lineage-negative progenitors mobilize to regenerate lung epithelium after major injury. *Nature* **517**, 621–625 (2015).
- Zuo, W. et al. p63(+)/Krt5(+) distal airway stem cells are essential for lung regeneration. *Nature* **517**, 616–620 (2015).
- Zacharias, W. J. et al. Regeneration of the lung alveolus by an evolutionarily conserved epithelial progenitor. *Nature* **555**, 251–255 (2018).
- Liu, Q. et al. Lung regeneration by multipotent stem cells residing at the bronchioalveolar-duct junction. *Nat. Genet.* **51**, 728–738 (2019).
- La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
- Saelens, W., Cannoodt, R., Todorov, H. & Saeyns, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
- Haghverdi, L., Büttner, M., Wolf, F. A., Büttner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
- Fischer, D. S. et al. Inferring population dynamics from single-cell RNA-sequencing time series data. *Nat. Biotechnol.* **37**, 461–468 (2019).



18. Vento-Tormo, R. et al. Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* **563**, 347–353 (2018).
19. Schiebinger, G. et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* **176**, 1517 (2019).
20. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
21. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* <https://doi.org/10.1038/nbt.4314> (2018)
22. Han, X. et al. Mapping the mouse cell atlas by Microwell-Seq. *Cell* **173**, 1307 (2018).
23. Angelidis, I. et al. An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. <https://doi.org/10.1101/351353> (2018)
24. Schiller, H. B. et al. Time- and compartment-resolved proteome profiling of the extracellular niche in lung injury and repair. *Mol. Syst. Biol.* **11**, 819 (2015).
25. LaCanna, R. et al. Yap/Taz regulate alveolar regeneration and resolution of lung inflammation. *J. Clin. Invest.* **130**, 2107–2122 (2019).
26. O'Reilly, M. A., Marr, S. H., Yee, M., McGrath-Morrow, S. A. & Lawrence, B. P. Neonatal hyperoxia enhances the inflammatory response in adult mice infected with influenza A virus. *Am. J. Respir. Crit. Care Med.* **177**, 1103–1110 (2008).
27. Nabhan, A. N., Brownfield, D. G., Harbury, P. B., Krasnow, M. A. & Desai, T. J. Single-cell Wnt signaling niches maintain stemness of alveolar type 2 cells. *Science* **359**, 1118–1123 (2018).
28. Wolf, F. A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).
29. El-Sukkari, D. et al. The protease inhibitor Cystatin C is differentially expressed among dendritic cell populations, but does not control antigen presentation. *J. Immunol.* **171**, 5003–5011 (2003).
30. Kathiriyai, J. J., Brumwell, A. N., Jackson, J. R., Tang, X. & Chapman, H. A. Distinct airway epithelial stem cells hide among club cells but mobilize to promote alveolar regeneration. *Cell Stem Cell* **26**, 346–358.e4 (2020).
31. Yuan, B. et al. Inhibition of distal lung morphogenesis in *Nkx2.1*( $-/-$ ) embryos. *Dev. Dyn.* **217**, 180–190 (2000).
32. Li, S. et al. Foxp transcription factors suppress a non-pulmonary gene expression program to permit proper lung development. *Dev. Biol.* **416**, 338–346 (2016).
33. Shu, W. et al. Foxp2 and Foxp1 cooperatively regulate lung and esophagus development. *Development* **134**, 1991–2000 (2007).
34. Roos, A. B., Berg, T., Barton, J. L., Didon, L. & Nord, M. Airway epithelial cell differentiation during lung organogenesis requires C/EBP $\alpha$  and C/EBP $\beta$ . *Developmental Dyn.* **241**, 911–923 (2012).
35. Bassères, D. S. et al. Respiratory failure due to differentiation arrest and expansion of alveolar cells following lung-specific loss of the transcription factor C/EBP $\alpha$  in mice. *Mol. Cell. Biol.* **26**, 1109–1123 (2006).
36. Martis, P. C. et al. C/EBP $\alpha$  is required for lung maturation at birth. *Development* **133**, 1155–1164 (2006).
37. Zhang, Z. et al. Transcription factor Etv5 is essential for the maintenance of alveolar type II cells. *Proc. Natl Acad. Sci. USA* **114**, 3903–3908 (2017).
38. Yang, H., Lu, M. M., Zhang, L., Whitsett, J. A. & Morrisey, E. E. GATA6 regulates differentiation of distal lung epithelium. *Development* **129**, 2233–2246 (2002).
39. Cheung, W. K. C. et al. Control of alveolar differentiation by the lineage transcription factors GATA6 and HOPX inhibits lung adenocarcinoma metastasis. *Cancer Cell* **23**, 725–738 (2013).
40. Tiwari, N. et al. Sox4 is a master regulator of epithelial-mesenchymal transition by controlling Ezh2 expression and epigenetic reprogramming. *Cancer Cell* **23**, 768–783 (2013).
41. Mutze, K., Vierkotten, S., Milosevic, J., Eickelberg, O. & Königshoff, M. Enolase 1 (ENO1) and protein disulfide-isomerase associated 3 (PDIA3) regulate Wnt/ $\beta$ -catenin-driven trans-differentiation of murine alveolar epithelial cells. *Dis. Model. Mech.* **8**, 877–890 (2015).
42. Adams, T. S. et al. Single Cell RNA-seq reveals ectopic and aberrant lung resident cell populations in idiopathic pulmonary fibrosis. <https://doi.org/10.1101/759902>.
43. Mayr, C. H. et al. Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids. <https://doi.org/10.1101/2020.01.21.20018358>.
44. Wu, H. et al. Progressive pulmonary fibrosis is caused by elevated mechanical tension on alveolar stem cells. *Cell.* <https://doi.org/10.1016/j.cell.2019.11.027> (2019)
45. Kim, K. K. et al. Alveolar epithelial cell mesenchymal transition develops in vivo during pulmonary fibrosis and is regulated by the extracellular matrix. *Proc. Natl Acad. Sci. USA* **103**, 13180–13185 (2006).
46. Yao, C. et al. Senescence of alveolar stem cells drives progressive pulmonary fibrosis. *SSRN Electron. J.* <https://doi.org/10.2139/ssrn.3438364>.
47. Kobayashi, Y. et al. Persistence of a novel regeneration-associated transitional cell state in pulmonary fibrosis. <https://doi.org/10.1101/855155>.
48. Aoshiba, K., Tsuji, T. & Nagai, A. Bleomycin induces cellular senescence in alveolar epithelial cells. *Eur. Respir. J.* **22**, 436–443 (2003).
49. Lehmann, M. et al. Chronic WNT/ $\beta$ -catenin signaling induces cellular senescence in lung epithelial cells. *Cell. Signal.* **70**, 109588 (2020).
50. Tritschler, S. et al. Concepts and limitations for learning developmental trajectories from single cell genomics. *Development* **146**, dev170506 (2019).
51. Yang, J. et al. The development and plasticity of alveolar type 1 cells. *Development* **143**, 54–65 (2016).
52. Treutlein, B. et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014).
53. Guo, M. et al. Single cell RNA analysis identifies cellular heterogeneity and adaptive responses of the lung at birth. *Nat. Commun.* **10**, 37 (2019).
54. Katsura, H., Kobayashi, Y., Tata, P. R. & Hogan, B. L. M. IL-1 and TNF $\alpha$  contribute to the inflammatory niche to enhance alveolar regeneration. *Stem Cell Rep.* **12**, 657–666 (2019).
55. Sun, T. et al. TAZ is required for lung alveolar epithelial cell differentiation after injury. *JCI Insight*. **5**, e128674 (2019).
56. Riemondy, K. A. et al. Single cell RNA sequencing identifies TGF $\beta$  as a key regenerative cue following LPS-induced lung injury. *JCI Insight* **5**, e123637 (2019).
57. Xu, Y. et al. Single-cell RNA sequencing identifies diverse roles of epithelial cells in idiopathic pulmonary fibrosis. *JCI Insight* **1**, e90558 (2016).
58. Xi, Y. et al. Local lung hypoxia determines epithelial fate decisions during alveolar regeneration. *Nat. Cell Biol.* **19**, 904–914 (2017).
59. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
60. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
61. Kowalczyk, M. S. et al. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res.* **25**, 1860–1872 (2015).
62. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
63. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet based single cell RNA sequencing data. <https://doi.org/10.1101/303727>.
64. Ramiłowski, J. A. et al. A draft network of ligand–receptor-mediated multicellular signalling in human. *Nature Commun.* **6**, 7866 (2015).
65. Law, C. W. et al. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000Res.* **5**, 1408 (2016).
66. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
67. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
68. Bergen, V., Lange, M., Peidli, S., Alexander Wolf, F. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. <https://doi.org/10.1101/820936>.
69. Polański, K. et al. BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics.* (2019) <https://doi.org/10.1093/bioinformatics/btz625> (2019)
70. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
71. Van den Berge, K. et al. Trajectory-based differential expression analysis for single-cell sequencing data. *Nature Commun.* **11**, 1201 (2020).
72. Uhl, F. E. et al. Preclinical validation and imaging of Wnt-induced repair in human 3D lung tissue cultures. *Eur. Respir. J.* **46**, 1150–1166 (2015).
73. Burgstaller, G. et al. Distinct niches within the extracellular matrix dictate fibroblast function in (cell free) 3D lung tissue cultures. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **314**, L708–L723 (2018).
74. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
75. Yee, M., Gelein, R., Mariani, T. J., Lawrence, B. P. & O'Reilly, M. A. The Oxygen environment at birth specifies the population of alveolar epithelial stem cells in the adult lung. *Stem Cells* **34**, 1396–1406 (2016).
76. Yee, M. et al. Alternative progenitor lineages regenerate the adult lung depleted of alveolar epithelial type 2 cells. *Am. J. Respir. Cell Mol. Biol.* **56**, 453–464 (2017).
77. Lehmann, M. et al. Senolytic drugs target alveolar epithelial cell function and attenuate experimental lung fibrosis *in vivo*. *Eur. Respiratory J.* **50**, 1602367 (2017).
78. Cox, J. & Mann, M. 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinform.* **13**(Suppl. 16) S12 (2012).
79. Habermann, A. C. et al. Single-cell RNA-sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. <https://doi.org/10.1101/753806>.



## Acknowledgements

This study was supported with funding by the German Center for Lung Research (DZL) and the Helmholtz Association. Further, this work has been funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036AB. M. Lange acknowledges financial support by the DFG through the Graduate School of QBM (GSC 1006) and by the Joachim Herz Stiftung. The Krt8/TROMA-I monoclonal antibody, developed by Brulet, P./Kemler, R. was obtained from the Developmental Studies Hybridoma Bank, created by the NICHD of the NIH and maintained at the University of Iowa, Department of Biology, Iowa City, IA, 52242. The beta6 integrin (Itgb6, clone 3G9) antibody was kindly provided by Prof. Dr. Dean Sheppard at the University of California San Francisco. Human samples were kindly provided by the CPC-M BioArchive. We thank Dr. Kathrin Mutze for the pmATII cell treatments, and Dr. Elisabeth Graf and Sandy Lösecke for sequencing and technical support.

## Author contributions

H.B.S. designed and supervised the entire study and wrote the paper. H.B.S. and F.J.T. supervised single cell data analysis. M.S., L.M.S., M.A., M.La., and H.B.S. performed single cell data analysis. M.S., I.A., and C.H.M. collected single cell data. I.K. and R.S. produced microfluidic devices for single cell RNA-seq. L.M.S. performed multi-modal data integration. L.M.S., M.A., and G.T. generated the webtool and wrote custom code. M.S., J.J.K., and M.Y. performed bleomycin experiments and lineage tracing. M.S., J.J.K., P.O., and G.B. performed microscopy and quantitative image analysis. M.S., L.F.M., A.S., C.V., and M.Le. validated single cell results using flow cytometry, immunostainings, and in vitro cell assays. Z.Z., T.S., J.H.L.N., A.H., J.B., M.O.R., R.S., M.K., H.A.C., F.J.T., and H.B.S. provided resources. M.S., L.M.S., H.A.C., F.J.T., and H.B.S. interpreted results. All authors read and approved the paper. Correspondence to F.J.T. and H.B.S.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41467-020-17358-3>.

**Correspondence** and requests for materials should be addressed to F.J.T. or H.B.S.

**Peer review information** *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

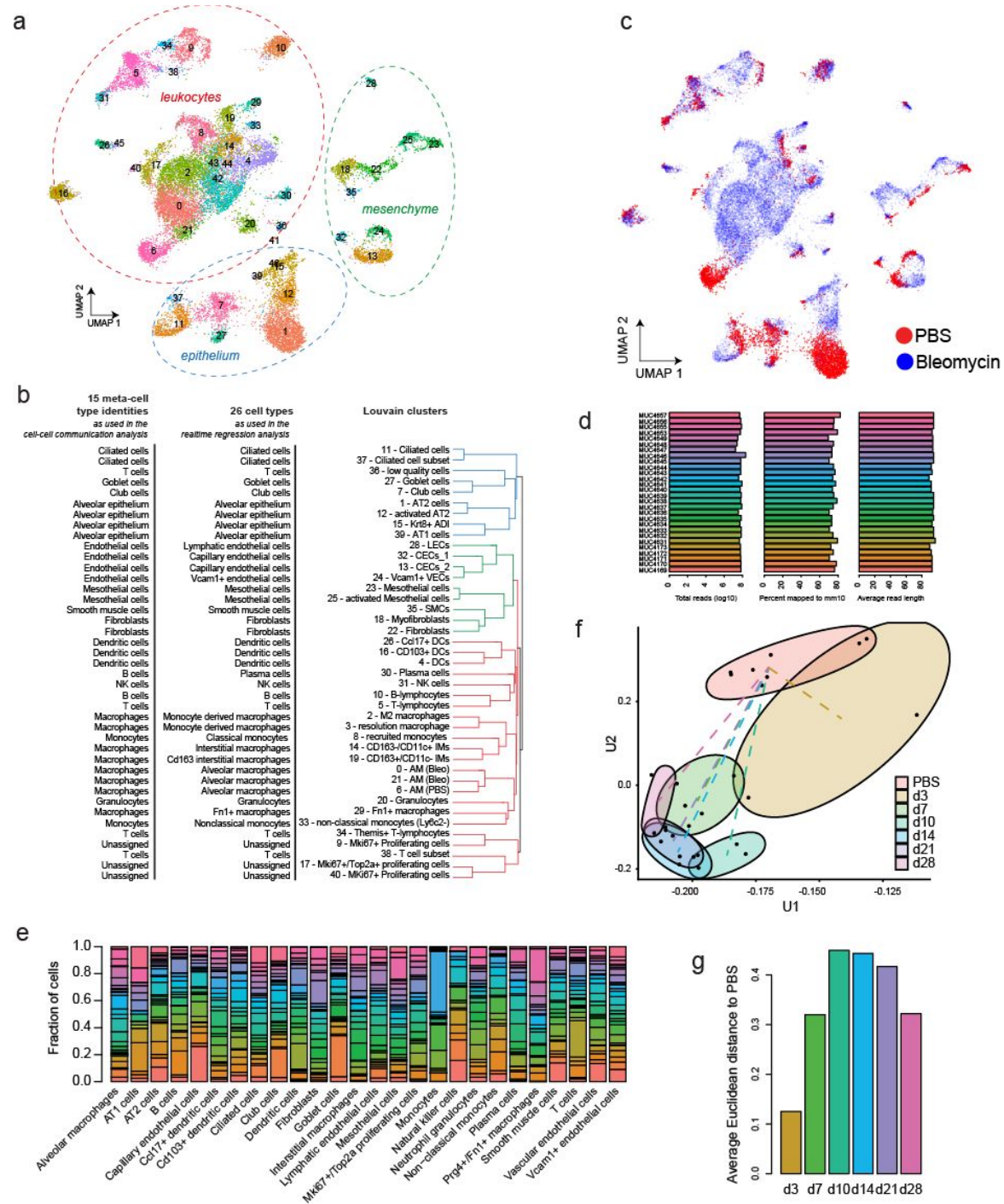
**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

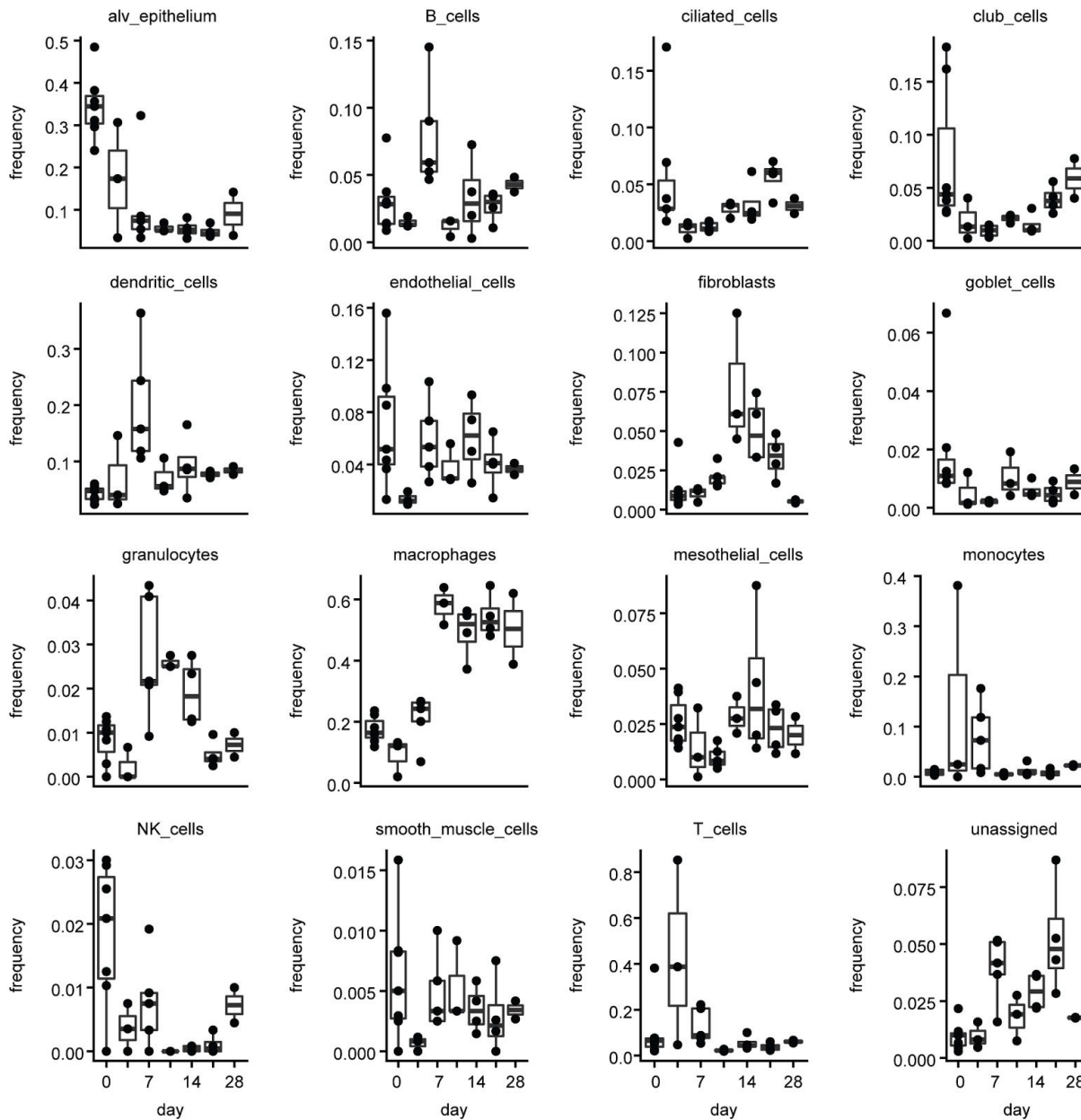
*Alveolar regeneration through a Krt8+ transitional stem cell state that persists in human lung fibrosis*



**Supplementary Figure 1. Good technical agreement of whole lung single cell transcriptomes of 28 individual mice.** (a) UMAP embedding colored by Louvain clusters demonstrates separation of cells into major lineages. (b) Unsupervised hierarchical clustering of the Louvain clusters recapitulates known hierarchical cell type topology. UMAP embeddings show good overlap between treatment conditions (c). (d) Alignment summary statistics are comparable across mouse samples. (e) Bar plot shows high overlap of mouse samples across cell types. (f) Scatter

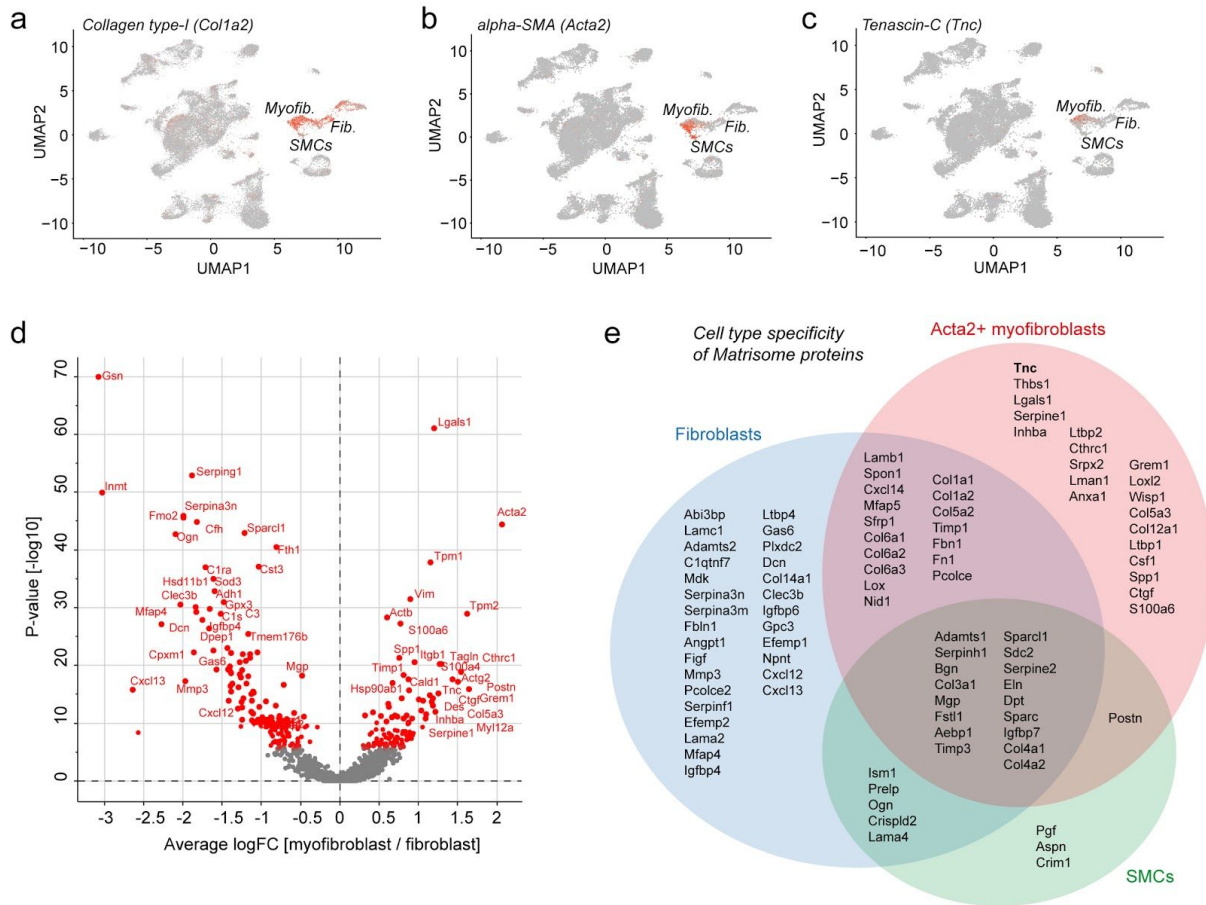
*Alveolar regeneration through a Krt8+ transitional stem cell state that persists in human lung fibrosis*

plot depicts coordinates from singular value decomposition of the arcsine square root transformation of the relative cell type frequencies. Dashed lines connect the mean coordinates between PBS and all other time points. Ellipses are colored by time point and encapsulate samples from the same time point. (g) Barplot displays mean Euclidean distance derived from the embedding in (f) between PBS and all other time points.



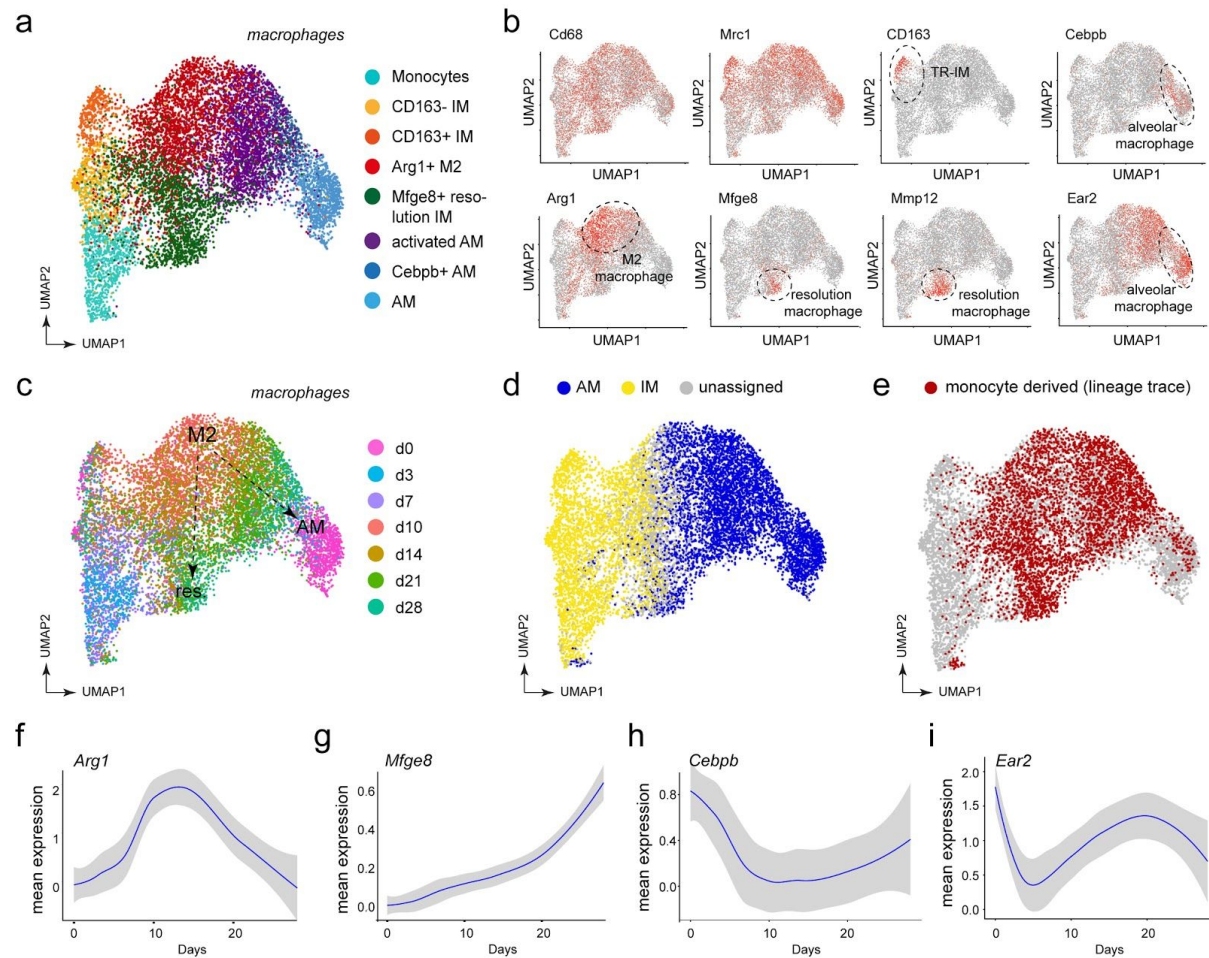
**Supplementary Figure 2.** Barplots display the relative frequencies (y-axis) of 16 meta-cell types across time points (x-axis) across 28 mouse replicates. Each dot represents one mouse sample. The boxes represent the interquartile range, the horizontal line in the box is the median, and the whiskers represent 1.5 times the interquartile range.

**Alveolar regeneration through a *Krt8+* transitional stem cell state that persists in human lung fibrosis**

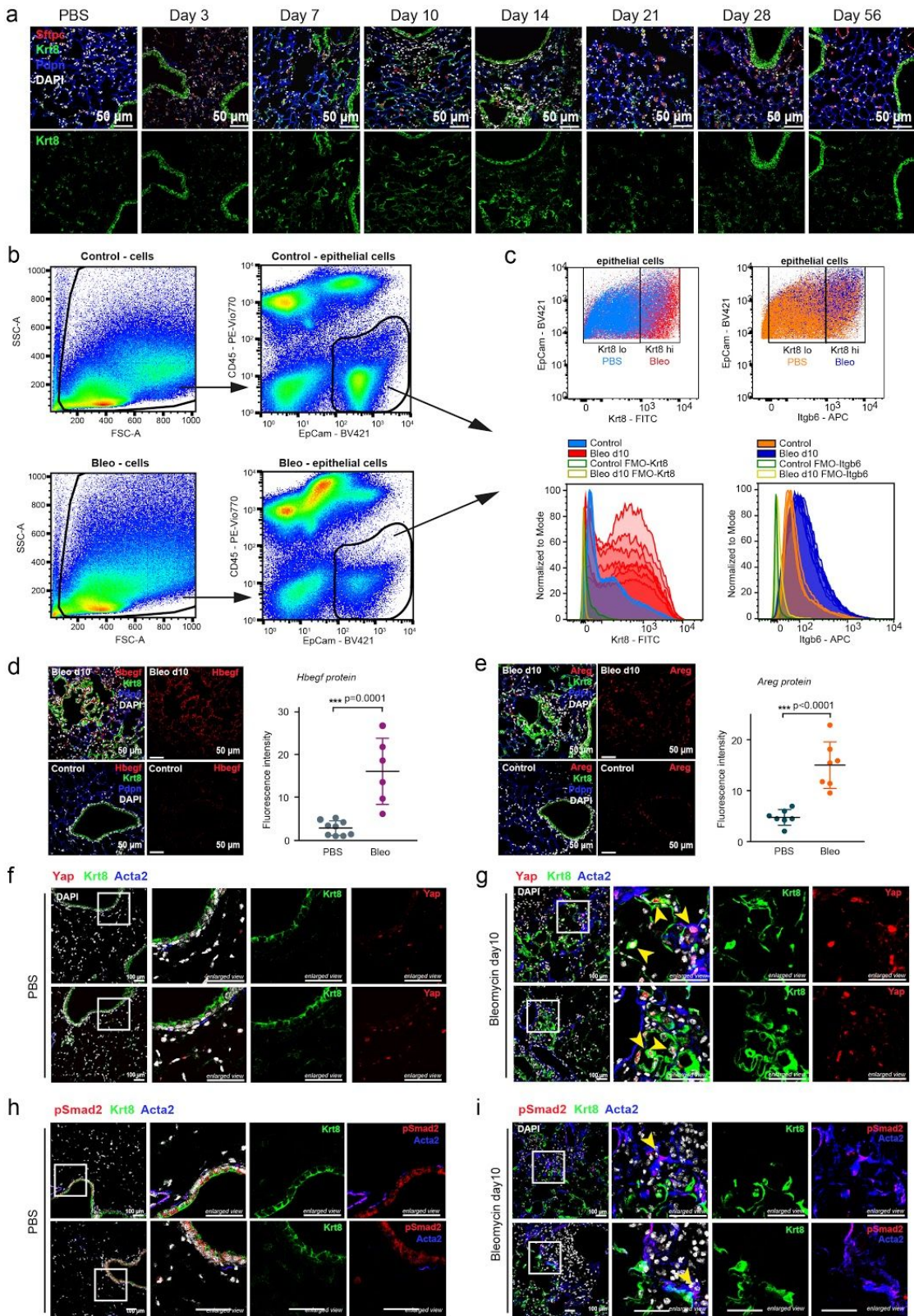


**Supplementary Figure 3. Transient appearance of the myofibroblast cell state upon lung injury.** (a-c) Relative expression levels of *Col1a2* (a), *Acta2* (b), and *Tnc* (c) are shown on the UMAP embedding. (d) The volcano plot shows differential gene expression between myofibroblasts (right side) and fibroblasts (left side). (e) Single cell analysis was used to derive the myofibroblast specific ECM components in comparison to other fibroblasts and smooth muscle cells.



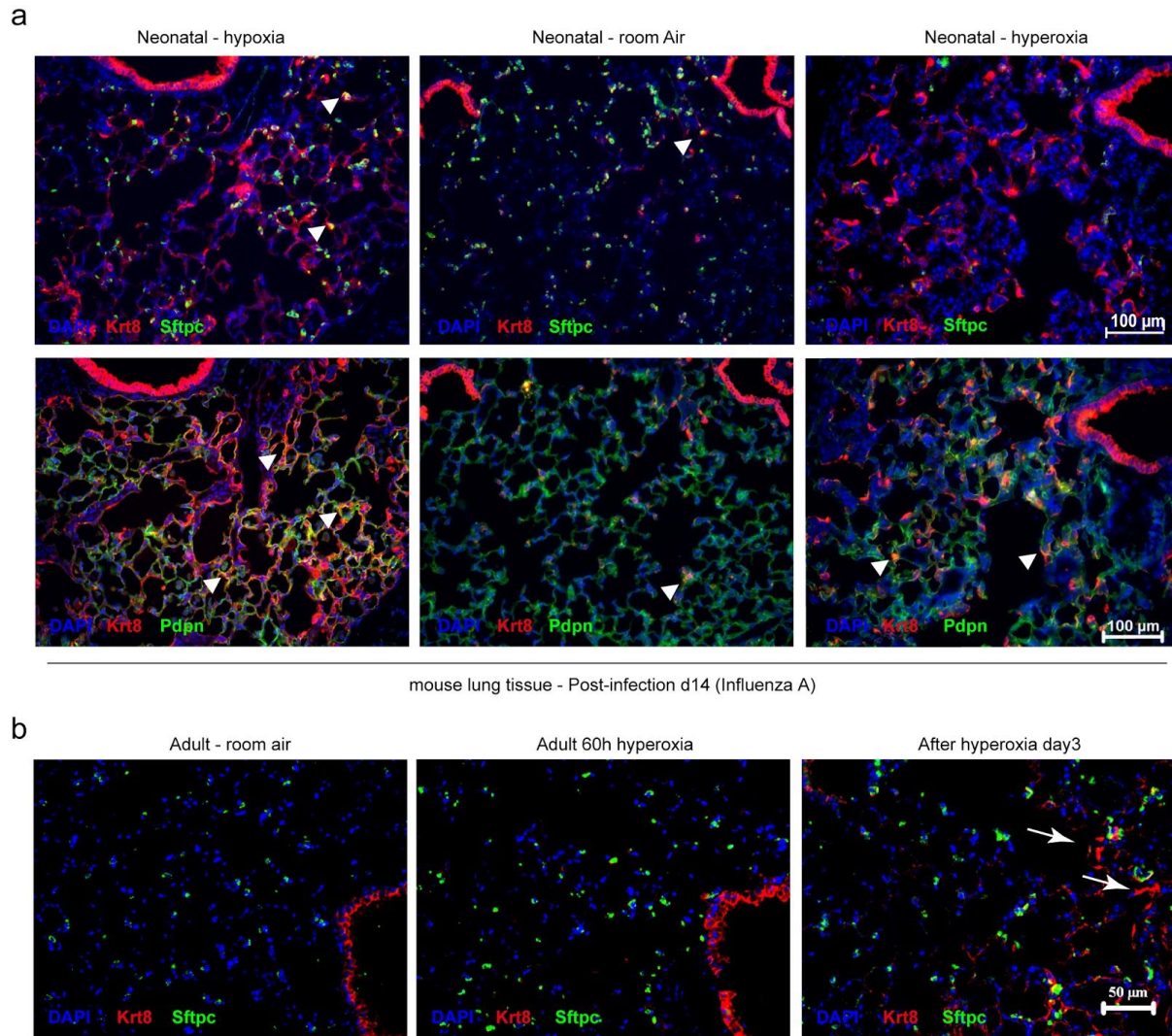


**Supplementary Figure 4. Dynamics of macrophage states in lung tissue regeneration.** (a, c) UMAP embedding of 10379 cells that express known macrophage markers is colored by (a) cluster identity and (c) time points. Following cells along the time course after reaching the peak of inflammation at day 10 and 14, two potential trajectories can be discerned. (b) Several macrophage populations can be identified. These clusters uniformly express the macrophage marker *Cd68* and *Mrc1* while also showing distinct expression of certain genes. (d) Previously published gene signatures from bulk RNA experiments were used to reveal potential origins of macrophage cells. In this data set, FACS-sorting allowed to differentiate between tissue-resident alveolar (AM), interstitial (IM) and monocyte-derived macrophage populations<sup>9</sup>. Similarity score of each cell is calculated as correlation to differentially expressed genes and corresponding log fold changes in the three sorted populations. Cells are assigned to either AM or IM category, if the difference in scores for either category is higher than 0.05. Alveolar macrophages in our data set indeed show the highest score on the tissue-resident AM. (e) Potentially monocyte-derived cells based on scoring (at threshold of 0.1). There is a separation in the potentially monocyte-derived cells, which concurs with the real-time trajectories in (c). (f-i) The line plots show the smoothed expression mean over time for the indicated genes within all macrophage subsets with a confidence interval of 0.95 (grey shades) across the 28 mouse replicates.

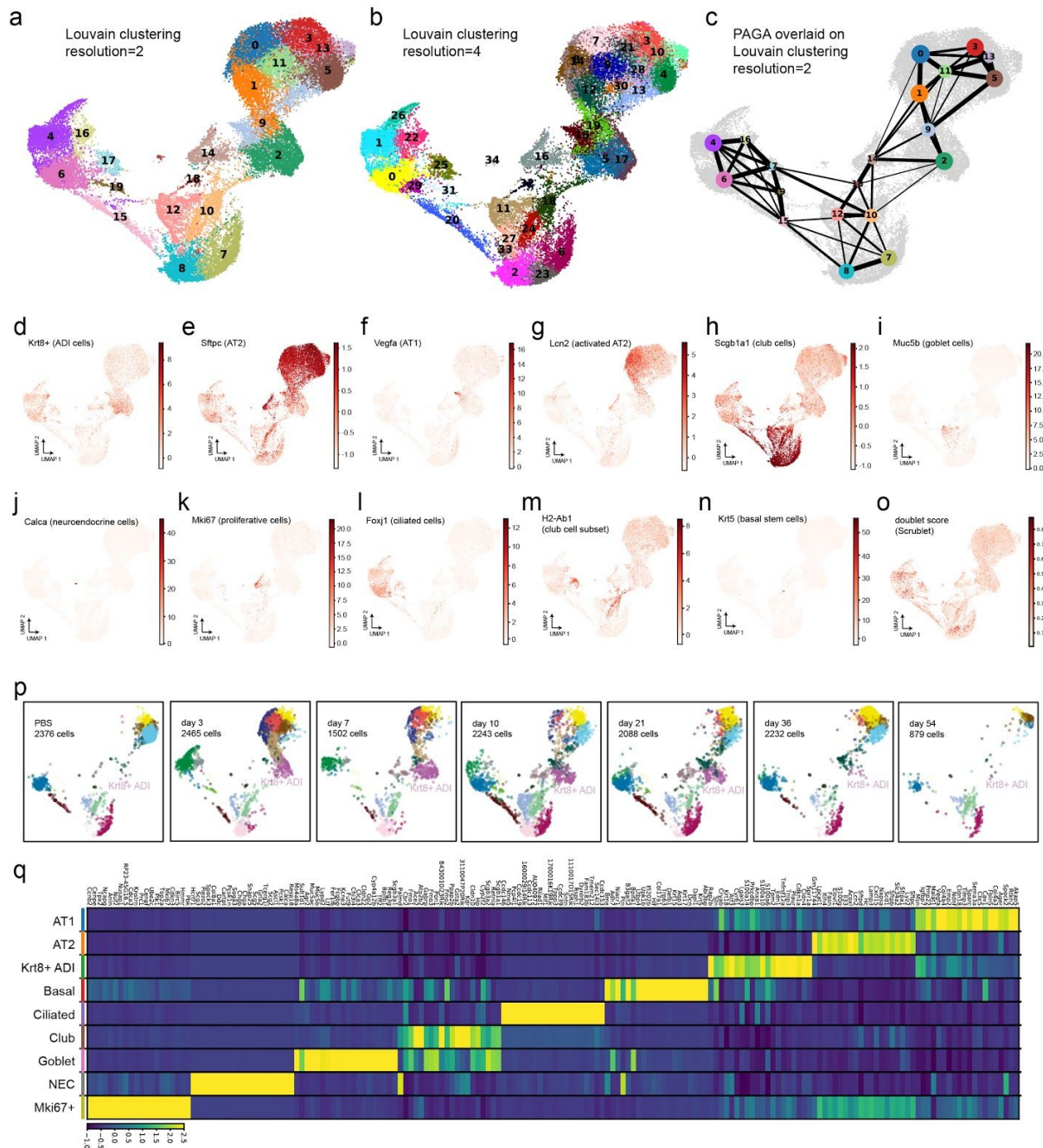


**Supplementary Figure 5. Protein validation of the alveolar Krt8+ ADI signature.** (a) Immunostaining of Krt8 (green) at the indicated time points after bleomycin injury. FFPE tissue sections were co-stained with the AT2 marker Sftpc (red), and the AT1 marker Pdpn (blue). Nuclei were labeled using DAPI (white). Scale bar = 50 microns; representative images from n=4 lungs/timepoint. (b) Gating strategy for the analysis of CD45-/Epcam+ epithelial cells. (c) The scatter plots and histograms show increased expression of Krt8 and Itgb6 at day 10 after bleomycin in Epcam+ epithelial cells. Highest expression of Itgb6 was observed on Krt8 high cells. Fluorescence-minus-one (FMO) controls were used for both the Krt8 and Itgb6 quantification. (d) Increased Hbegf (red) expression in bleomycin treated lung tissue, showing partial overlap with Krt8 (green) signal. Quantification of the mean fluorescence signal intensities confirmed increased Hbegf expression (unpaired t-test, one-sided, \*\*\* p = 0.0001, mean measure with SD). Sections were co-stained with Pdpn (blue); scale bar = 50 microns. Sections assessed in PBS n=9, in Bleo n=6. (e) Immunostainings of Areg (red) and Krt8 (green) expression in the lung, co-stained with Pdpn (blue) and quantified by mean fluorescence intensity. Unpaired t-test, one-sided, \*\*\* p < 0.0001, mean measure with SD. Scale bar = 50 microns. Sections assessed in PBS n=7, in Bleo n=7. (f-i) Tissue sections from either controls (f, h) or bleomycin day 10 (g, i) were stained with the indicated antibodies to assess activity of the Yap/Taz pathway (f, g) and the TGF-beta signaling by pSMAD nuclear translocation (h, i); representative images from n=2 lungs/condition.





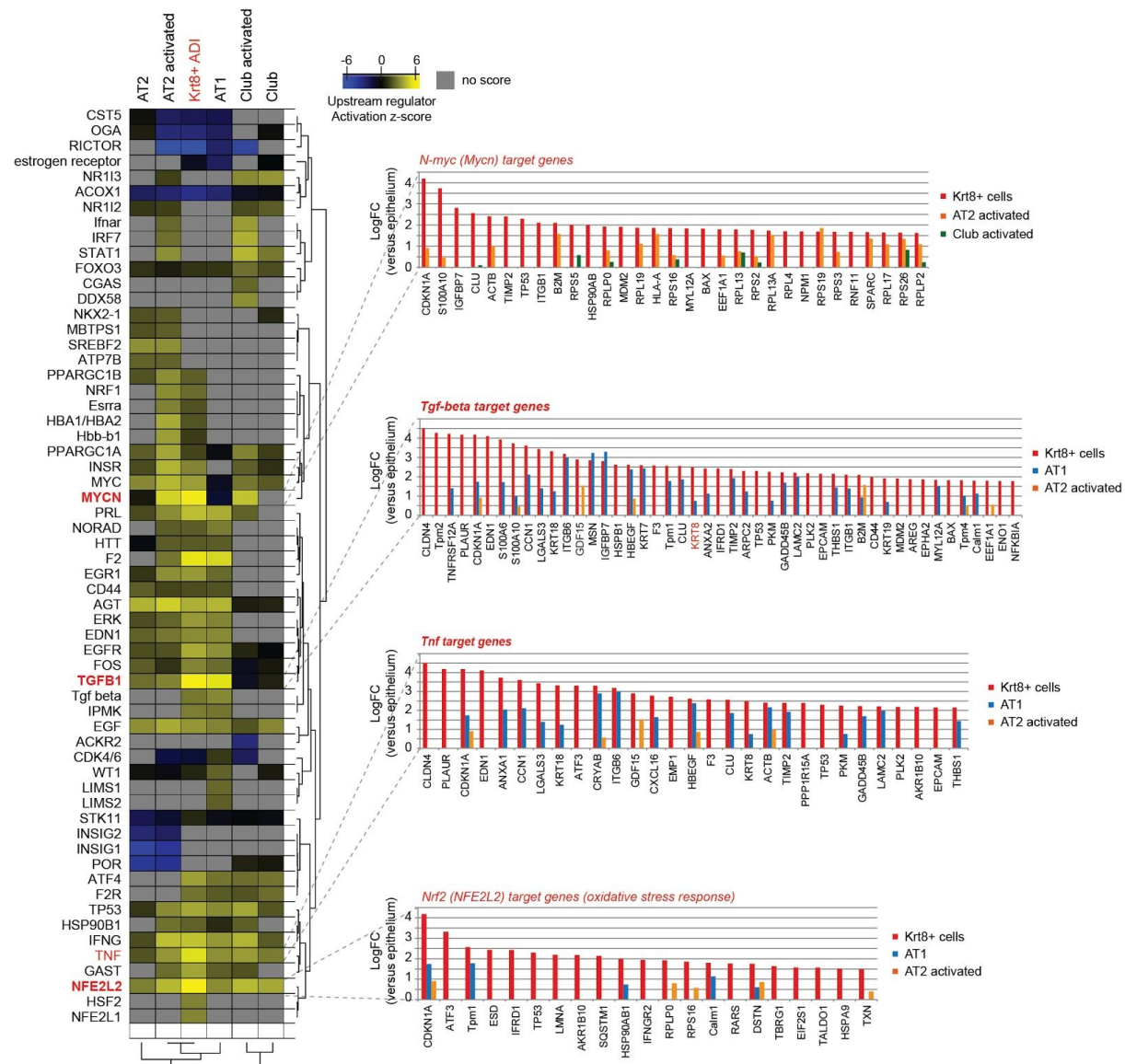
**Supplementary Figure 6. Appearance of Krt8+ ADI cells in two alternative mouse injury models.** (a) An aberrant oxygen environment at birth alters alveolar injury and repair following influenza A virus infection. Lungs of infected mice that were previously exposed to the indicated neonatal conditions were stained for Krt8 (red) and Sftpc (green). Scale bar = 100 microns. (b) A sixty-hour exposure of adult mice to hyperoxia leads to the emergence of Krt8+ cells in the alveolar space. Mice were sacrificed three days after the exposure period terminated. Lung tissue was stained for Krt8 (red) and Sftpc (green). Scale bar = 50 microns.



**Supplementary Figure 7. Feature plots of selected marker genes for epithelial cell types and states.** (a,b) UMAP embeddings showing the Louvain clustering as calculated with the resolution parameter=2 (a) and resolution parameter=4 (b). (c) PAGA graph representation of the data overlaid onto the UMAP representation with respective Louvain clustering, calculated with resolution parameter=2. (d-n) UMAP embeddings display distinct expression patterns for selected epithelial cell type marker genes: (d) Krt8 (ADI), (e) Sftpc (AT2 cells), (f) Vegfa (AT1 cells), (g) Lcn2 (activated AT2 cells), (h) Scgb1a1 (club cells), (i) Muc5b (goblet cells), (j) Calca (neuroendocrine cells), (k) Mki67 (proliferative cells), (l) Foxj1 (ciliated cells), (m) H2-Ab1 (club cell subset), (n) Krt5 (basal stem cells). Red colors indicate higher expression levels, (o) doublet score calculated with the doublet detection algorithm Scrublet. (p) UMAP shows epithelial cells at the indicated time points after injury color coded by their

**Alveolar regeneration through a *Krt8+* transitional stem cell state that persists in human lung fibrosis**

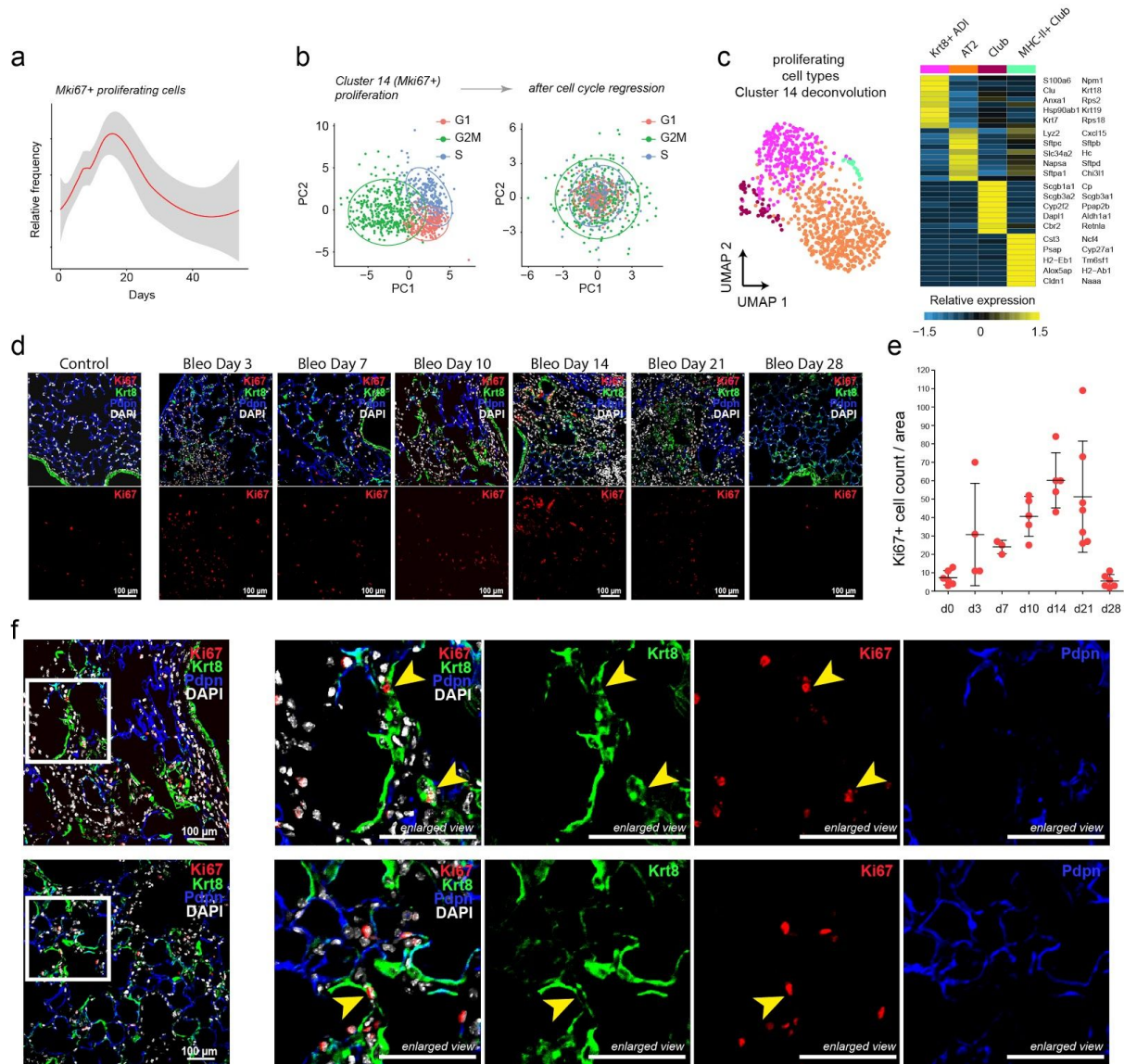
respective louvain cluster identity (Fig. 5). Note the massive increase of *Krt8+* ADI relative frequency during inflammation and fibrogenesis. (q) Heatmap shows the average expression levels for the top 20 genes with lowest adjusted p value of each cell type.



**Supplementary Figure 8. Gene programs with increased activity in *Krt8+* ADI.** Ingenuity upstream regulator analysis was used to score the activity of upstream regulators within the signatures of the indicated cell states. The activation z-scores were grouped by hierarchical clustering using their Pearson correlation. Bar graphs show target genes sorted by highest expression in *Krt8+* cells relative to all other cells.

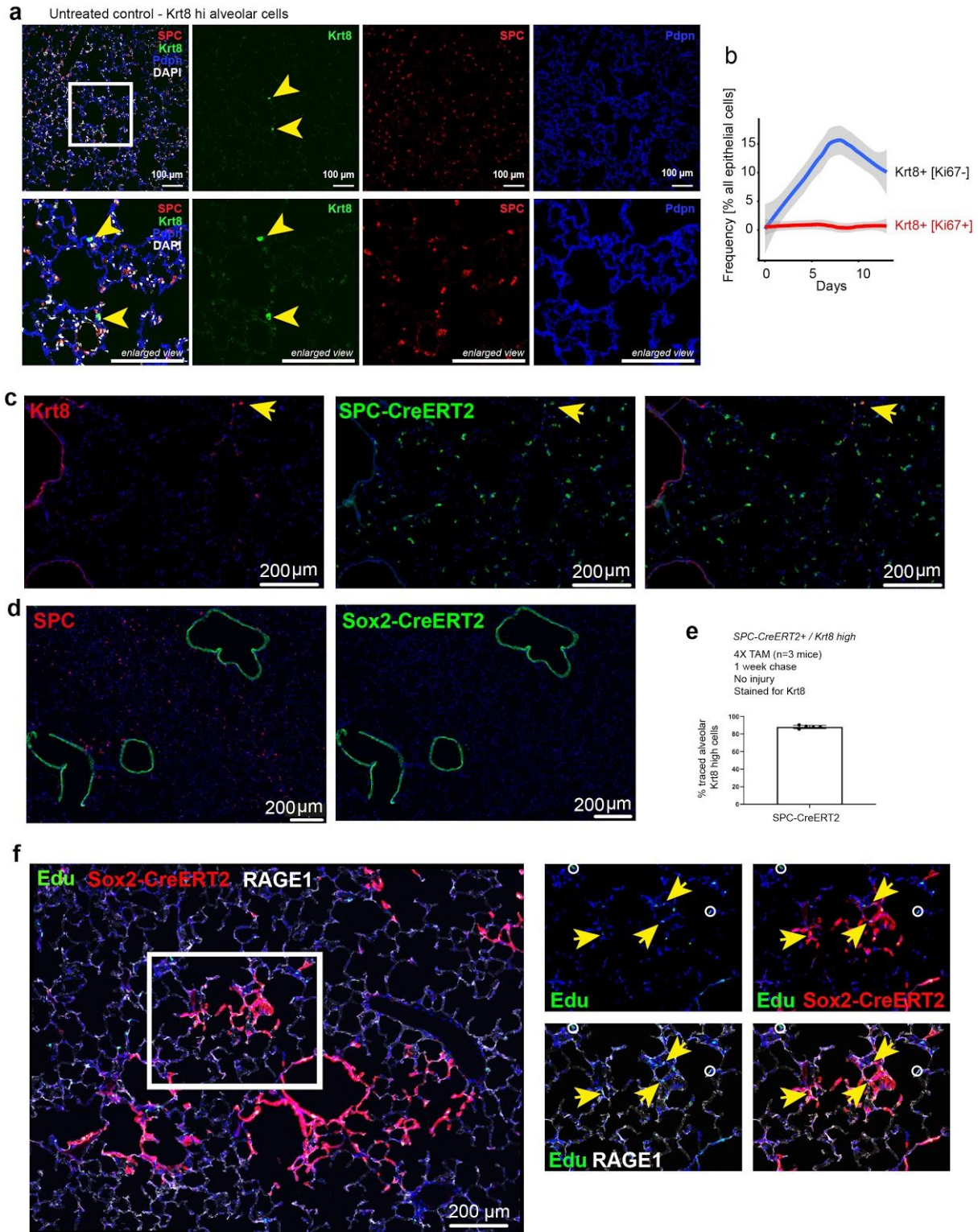


*Alveolar regeneration through a Krt8+ transitional stem cell state that persists in human lung fibrosis*



**Supplementary Figure 9. Cell cycle analysis shows no massive expansion of rare pre-existing Krt8+ ADI cells.** (a) The line plot shows the smoothed relative frequency of Mki67+ proliferating cells over time. Grey colors indicate 95% confidence interval of fit across the 36 mouse replicates. (b) The scatter plots show cells from proliferating cell cluster 14 before and after cell cycle regression, colored by inferred cell cycle phase. Regression removes cell cycle effects from principal component data manifold. Re-analysis of cell cycle corrected expression deconvolves cell type identities of proliferating cells. (c) UMAP of cell cycle corrected cluster 14 cells visualizes four distinct clusters, which contain Krt8+ ADI, AT2, club, and MHC-II+ club cells. Heatmap shows the average expression levels of selected marker genes. (d) Immunofluorescence stainings of control versus bleomycin treated lung sections (day 3, 7, 10, 14, 21, 28). Sections were stained for Krt8 (green), Ki67 (red), Pdpn (blue), and DAPI (white). Scale bar indicates 100 microns. (e) Ki67+ cells were quantified from the micrographs by counting Ki67+ cells in each ROI/field of view [mean with SD, n(d0)=6, n(d3)=4, n(d7)=3, n(d10)=5, n(d14)=5, n(d21)=7, n(d28)=6]. (f) Immunostaining as in (e) on day 10 post bleomycin injured lungs with enlarged views on proliferative Krt8+ ADI cells (Ki67+/Krt8+), highlighted with yellow arrowheads. Scale bar indicates 100 microns, n(mice)=2.

*Alveolar regeneration through a Krt8+ transitional stem cell state that persists in human lung fibrosis*

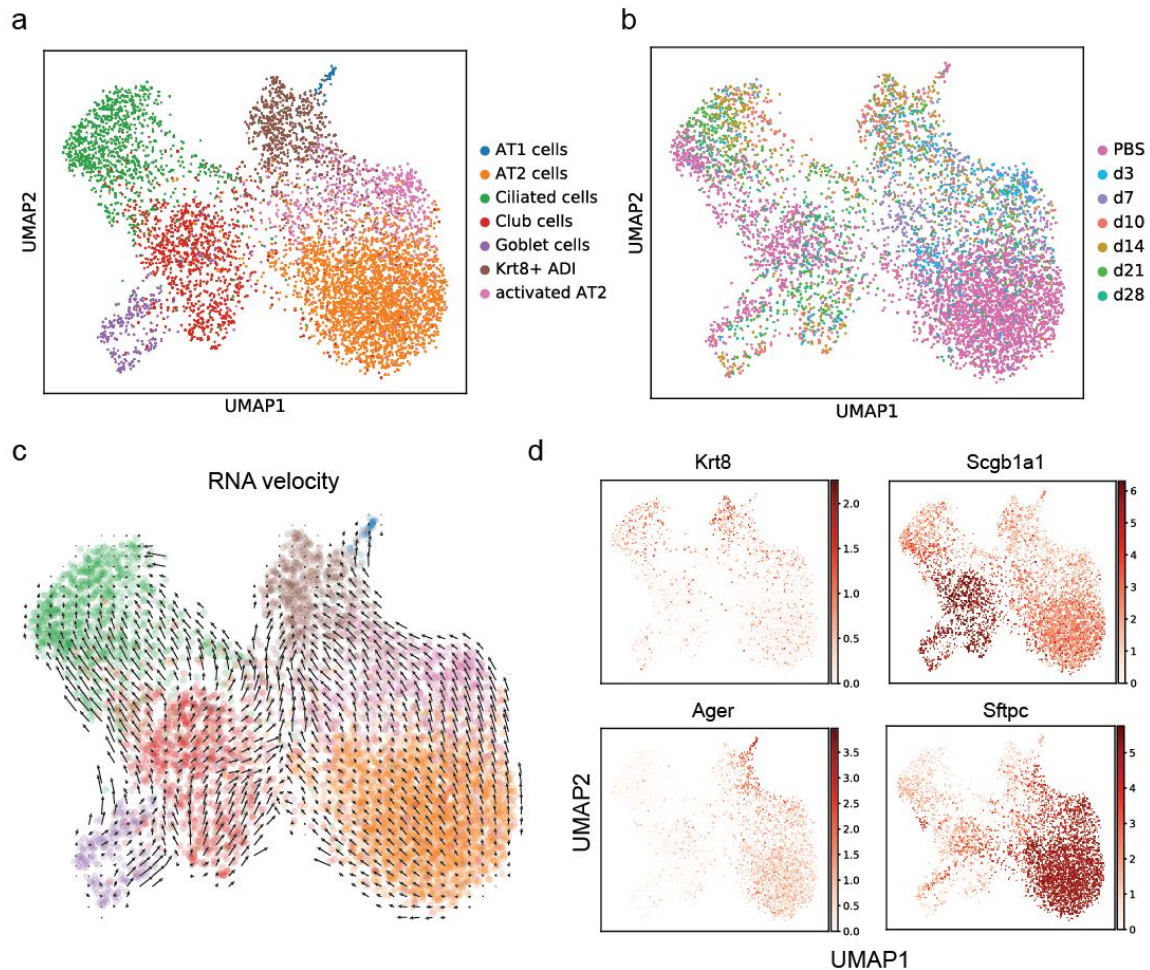


**Supplementary Figure 10. Rare Krt8 high alveolar cells in healthy lung parenchyma.** (a) Fluorescent immunostainings and confocal imaging of lung sections from untreated control lungs. Nuclei (DAPI) are colored in white, Krt8 appears in green, Sftpc (AT2 cells) in red, and Pdpn (AT1 cells) in blue. The scale bar indicates 100 microns. (b) Line plots show smoothed relative frequency of cells with a Krt8+ ADI signature stratified in Ki67+ (red



**Alveolar regeneration through a Krt8+ transitional stem cell state that persists in human lung fibrosis**

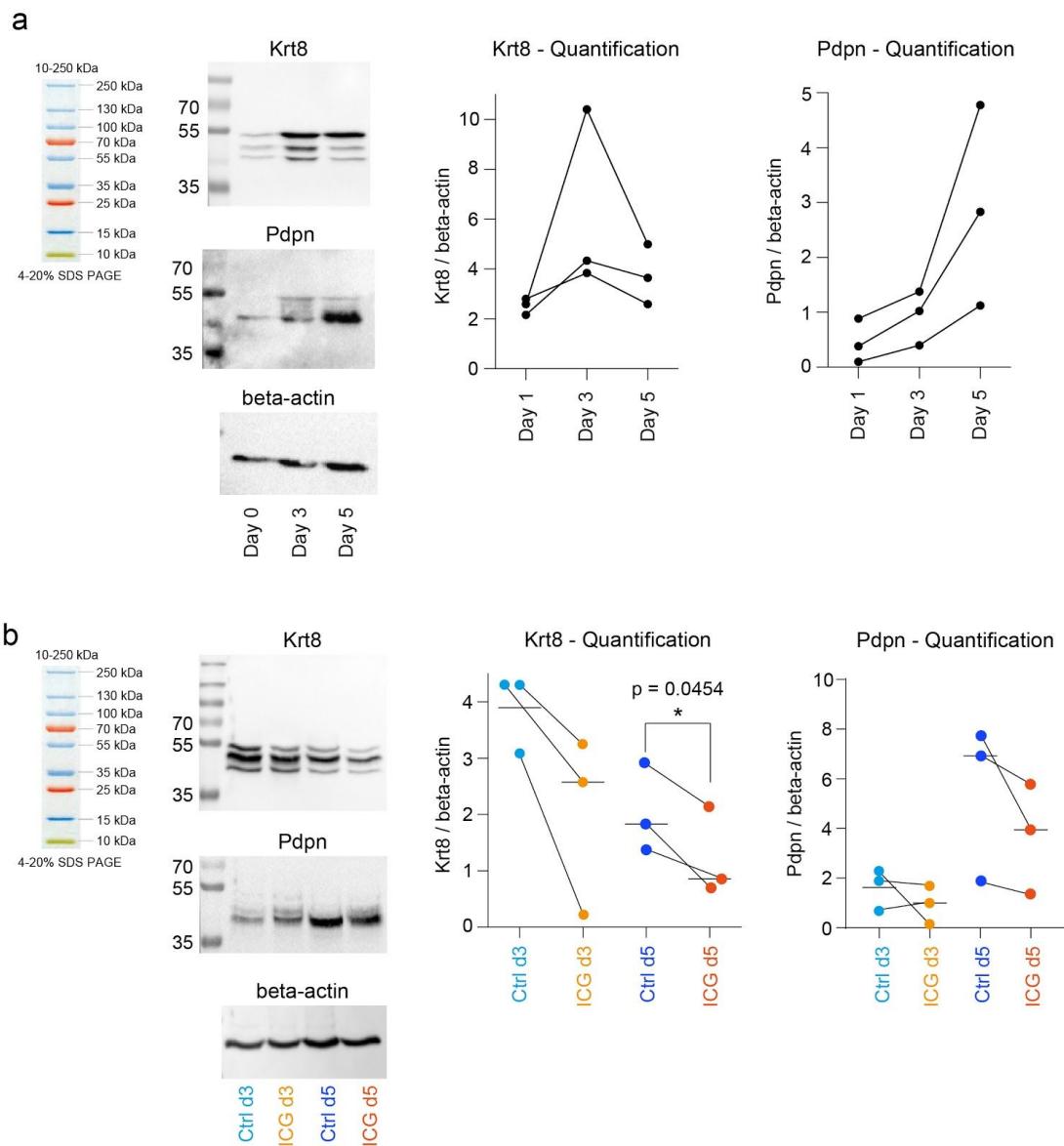
line) and Ki67- bins (blue line). Note the massive expansion of Krt8+ ADI over time without spiking numbers of Ki67+ cells preceding this. This indicates that most Krt8+ ADI are generated by differentiation of stem cells rather than expansion of pre-existing Krt8 high cells from the baseline. Grey colors indicate 95% confidence interval of a fit across the mouse replicates. (c, d) SPC-CreERT2 mice (c, n=3) or Sox2-CreERT2 mice (d) were labeled for 1 week using 4x tamoxifen (TAM) injections. Cryosections were stained for (c) Krt8 or SPC (d) and the lineage label as shown. (e) Quantification of SPC-CreERT2 traced cells with high expression of Krt8 in alveolar areas shows that most pre-existing rare Krt8 high alveolar cells are AT2 cell derived; n(mice)=3 Each data point represents quantification from one large region (n=6; each at least 2.5sqmm area). Error bars show standard deviation. (f) Cryosection of an Edu-labeled lung after bleomycin-injury in the Sox2-CreERT2 mice; the four enlargements as indicated by the white box show that proliferating Edu+/RAGE1+ cells have overlapping signal with the Sox2 lineage label (yellow arrows), indicating that Sox2 lineage-derived cells can give rise to RAGE1+ AT1 cells; untraced double positive Edu+/RAGE1+ cells are highlighted by a white circle. Experiment includes n=2 mice; at least 3 lobes/mouse were analyzed with similar results.



**Supplementary Figure 11. Targeted re-analysis of epithelial cells from whole lung data set.** (a, b) Data from epithelial cells in the whole lung data set (Fig. 1) was subjected to dimension reduction. UMAP visualizations colored by cell type (a) and time point (b) illustrate injury-specific cells connecting major cell types. (c) RNA

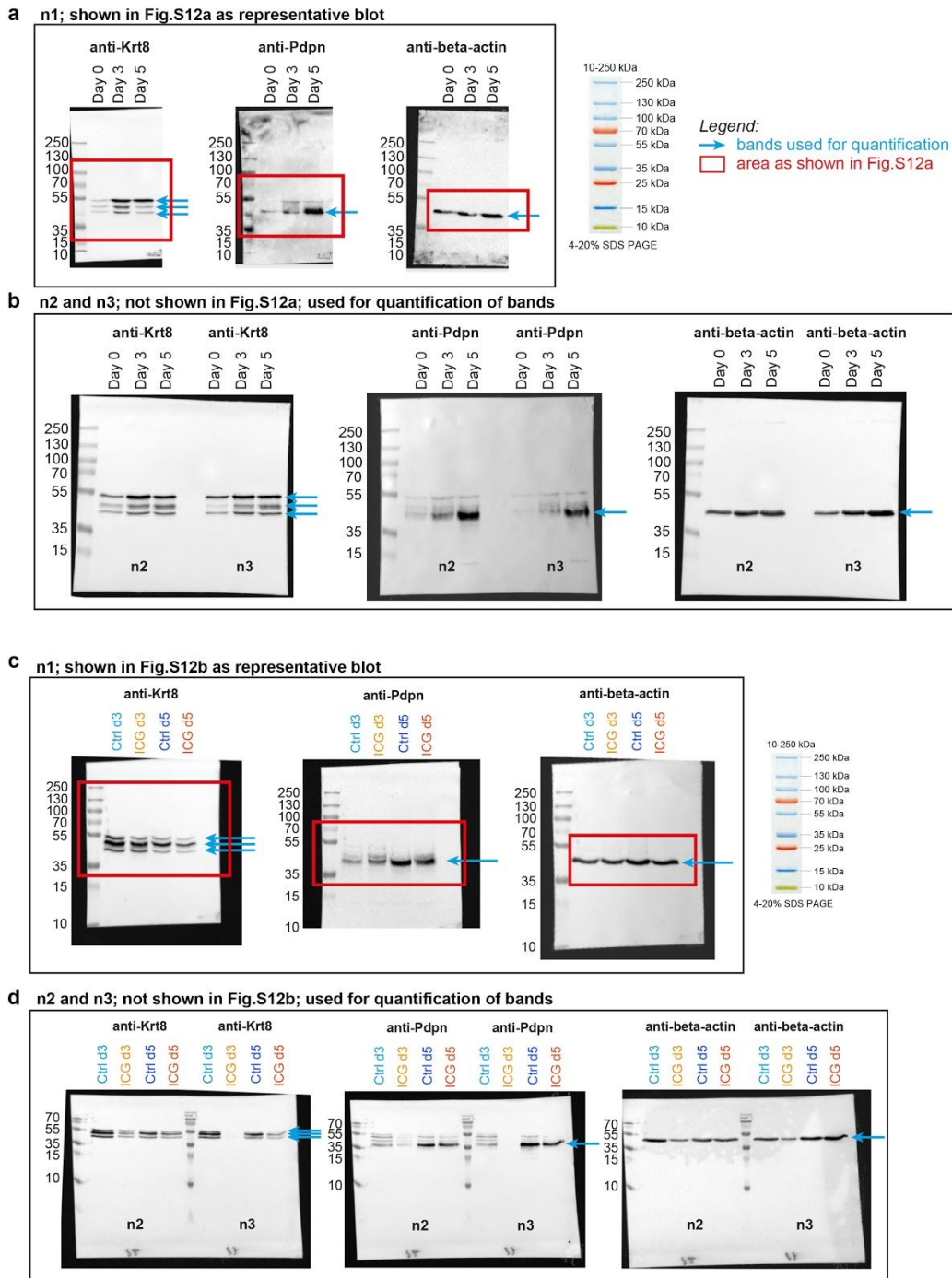
**Alveolar regeneration through a Krt8+ transitional stem cell state that persists in human lung fibrosis**

velocities predict both AT2 and airway cell derived Krt8+ ADI. (d) Expression of select marker genes is colored on top of the UMAP embedding. Grey and red colors correspond to low and high values, respectively.



**Supplementary Figure 12: AT1 cell differentiation involves Wnt/ $\beta$ -catenin/TCF-mediated transcription.** (a) Representative Western blot for the detection of Krt8 and Pdpn in MACS-negatively selected pmAT2 cells plated on plastic dishes; proteins were analyzed and quantified at day 1, day 3, and day 5, demonstrating that during in vitro differentiation the expression of Krt8 (peak expression at day 3) and Pdpn are increased over time (n of blots = 3). All gels/blots were processed in parallel. (b) Representative Western blot for likewise selected pmAT2 cells, treated with WNT inhibitor ICG-001 (start of inhibition at day 1). Inhibition was stopped at day 3 and day 5, respectively, and lysates loaded for Western blot analysis. WNT inhibition induced a reduction in both Krt8 and Pdpn levels compared to untreated control samples (n of blots = 3, data shown with mean). Quantification of the Krt8 protein at day 5 revealed a significant decrease of protein levels upon WNT inhibition (paired t-test, two-sided,  $p = 0.0454$ ). All gels/blots were processed in parallel.





**Supplementary Figure 13. Full-scan sizes of all western blots from supplementary figure 12.** A total of n=3 samples were used for all blots. (a) The first sample used as a representative blot. (b) The second and third sample blots which were included in the blot quantification. (c) The first sample used as a representative blot for the inhibition experiment. (d) The second and third sample blots which were included in the blot quantification of the inhibition experiment.

---

## References

1. Selman, M., López-Otín, C. & Pardo, A. Age-driven developmental drift in the pathogenesis of idiopathic pulmonary fibrosis. *European Respiratory Journal* vol. 48 538–552 (2016).
2. Nations, U. & United Nations. World Population Ageing 2019 Highlights. (2019) doi:10.18356/9df3caed-en.
3. Meiners, S., Eickelberg, O. & Königshoff, M. Hallmarks of the ageing lung. *Eur. Respir. J.* **45**, 807–827 (2015).
4. Miller, M. Structural and Physiological Age-Associated Changes in Aging Lungs. *Seminars in Respiratory and Critical Care Medicine* vol. 31 521–527 (2010).
5. Sorino, C. *et al.* Diagnosis of airway obstruction in the elderly: contribution of the SARA study. *International Journal of Chronic Obstructive Pulmonary Disease* 389 (2012) doi:10.2147/copd.s31630.
6. Gibson, P. G., McDonald, V. M. & Marks, G. B. Asthma in older adults. *The Lancet* vol. 376 803–813 (2010).
7. Aw, D., Silva, A. B. & Palmer, D. B. Immunosenescence: emerging challenges for an ageing population. *Immunology* vol. 120 435–446 (2007).
8. Murtha, L. A. *et al.* The Role of Pathological Aging in Cardiac and Pulmonary Fibrosis. *Aging and disease* vol. 10 419 (2019).
9. Plantier, L. *et al.* Physiology of the lung in idiopathic pulmonary fibrosis. *Eur. Respir. Rev.* **27**, (2018).
10. Schmidt, R. *et al.* Altered fatty acid composition of lung surfactant phospholipids in interstitial lung disease. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **283**, L1079–85 (2002).
11. Lorenz, R. J. Weibel, E. R.: Morphometry of the Human Lung. Springer Verlag, Berlin-Göttingen-Heidelberg 1963; 151 S., 109 Abb., DM 36,-. *Biometrische Zeitschrift* vol. 8 143–144 (1966).
12. Hiemstra, P. S., McCray, P. B. & Bals, R. The innate immune function of airway epithelial cells in inflammatory lung disease. *European Respiratory Journal* vol. 45 1150–1162 (2015).
13. Leiva-Juárez, M. M., Kolls, J. K. & Evans, S. E. Lung epithelial cells: therapeutically inducible effectors of antimicrobial defense. *Mucosal Immunology* vol. 11 21–34 (2018).
14. Boers, J. E., Ambergen, A. W. & Frederik B J. Number and Proliferation of Clara Cells in Normal Human Airway Epithelium. *American Journal of Respiratory and Critical Care Medicine* vol. 159 1585–1591 (1999).
15. Rock, J. R., Randell, S. H. & Hogan, B. L. M. Airway basal stem cells: a perspective on their roles in epithelial homeostasis and remodeling. *Disease Models & Mechanisms* vol. 3 545–556 (2010).
16. Sacco, O. *et al.* Epithelial cells and fibroblasts: structural repair and remodelling in the airways. *Paediatric Respiratory Reviews* vol. 5 S35–S40 (2004).

- 
17. Zepp, J. A. & Morrissey, E. E. Cellular crosstalk in the development and regeneration of the respiratory system. *Nature Reviews Molecular Cell Biology* vol. 20 551–566 (2019).
  18. Hogan, B. L. M. The Alveolar Stem Cell Niche of the Mammalian Lung. in *Molecular Mechanism of Congenital Heart Disease and Pulmonary Hypertension* 7–12 (Springer, Singapore, 2020).
  19. Niessen, C. M. Tight Junctions/Adherens Junctions: Basic Structure and Function. *Journal of Investigative Dermatology* vol. 127 2525–2532 (2007).
  20. Shen, L., Weber, C. R., Raleigh, D. R., Yu, D. & Turner, J. R. Tight Junction Pore and Leak Pathways: A Dynamic Duo. *Annual Review of Physiology* vol. 73 283–309 (2011).
  21. Loxham, M., Davies, D. E. & Blume, C. Epithelial function and dysfunction in asthma. *Clinical & Experimental Allergy* vol. 44 1299–1313 (2014).
  22. Steed, E., Balda, M. S. & Matter, K. Dynamics and functions of tight junctions. *Trends in Cell Biology* vol. 20 142–149 (2010).
  23. Rose, V. D. *et al.* Airway Epithelium Dysfunction in Cystic Fibrosis and COPD. *Mediators of Inflammation* vol. 2018 1–20 (2018).
  24. Aghapour, M., Raee, P., Moghaddam, S. J., Hiemstra, P. S. & Heijink, I. H. Airway Epithelial Barrier Dysfunction in Chronic Obstructive Pulmonary Disease: Role of Cigarette Smoke Exposure. *American Journal of Respiratory Cell and Molecular Biology* vol. 58 157–169 (2018).
  25. Kottmann, R. M. *et al.* Lactic Acid Is Elevated in Idiopathic Pulmonary Fibrosis and Induces Myofibroblast Differentiation via pH-Dependent Activation of Transforming Growth Factor- $\beta$ . *American Journal of Respiratory and Critical Care Medicine* vol. 186 740–751 (2012).
  26. Bagnato, G. & Harari, S. Cellular interactions in the pathogenesis of interstitial lung diseases. *Eur. Respir. Rev.* **24**, 102–114 (2015).
  27. Zhou, Y. *et al.* Inhibition of mechanosensitive signaling in myofibroblasts ameliorates experimental pulmonary fibrosis. *J. Clin. Invest.* **123**, 1096–1108 (2013).
  28. Spagnolo, P. & Cottin, V. Genetics of idiopathic pulmonary fibrosis: from mechanistic pathways to personalised medicine. *J. Med. Genet.* **54**, 93–99 (2017).
  29. Borie, R., Kannengiesser, C. & Crestani, B. Familial forms of nonspecific interstitial pneumonia/idiopathic pulmonary fibrosis. *Current Opinion in Pulmonary Medicine* vol. 18 455–461 (2012).
  30. Campo, I. *et al.* A large kindred of pulmonary fibrosis associated with a novel ABCA3 gene variant. *Respir. Res.* **15**, 43 (2014).
  31. Selman, M., King, T. E. & Pardo, A. Idiopathic Pulmonary Fibrosis: Prevailing and Evolving Hypotheses about Its Pathogenesis and Implications for Therapy. *Annals of Internal Medicine* vol. 134 136 (2001).
  32. King, T. E., Jr, Pardo, A. & Selman, M. Idiopathic pulmonary fibrosis. *Lancet* **378**, 1949–1961 (2011).
  33. Xu, Y. *et al.* Single-cell RNA sequencing identifies diverse roles of epithelial cells in idiopathic pulmonary fibrosis. *JCI Insight* **1**, e90558 (2016).
  34. Antoniades, H. N. *et al.* Platelet-derived growth factor in idiopathic pulmonary fibrosis. *Journal of Clinical Investigation* vol. 86 1055–1064 (1990).

- 
35. Khalil, N., O'Connor, R. N., Flanders, K. C. & Unruh, H. TGF-beta 1, but not TGF-beta 2 or TGF-beta 3, is differentially present in epithelial cells of advanced pulmonary fibrosis: an immunohistochemical study. *American Journal of Respiratory Cell and Molecular Biology* vol. 14 131–138 (1996).
  36. Miyazaki, Y. *et al.* Expression of a tumor necrosis factor-alpha transgene in murine lung causes lymphocytic and fibrosing alveolitis. A mouse model of progressive pulmonary fibrosis. *Journal of Clinical Investigation* vol. 96 250–259 (1995).
  37. Saleh, D. *et al.* Elevated expression of endothelin-1 and endothelin-converting enzyme-1 in idiopathic pulmonary fibrosis: possible involvement of proinflammatory cytokines. *American Journal of Respiratory Cell and Molecular Biology* vol. 16 187–193 (1997).
  38. Pan, L.-H. *et al.* Type II alveolar epithelial cells and interstitial fibroblasts express connective tissue growth factor in IPF. *European Respiratory Journal* vol. 17 1220–1227 (2001).
  39. Andersson-Sjöland, A. *et al.* Fibrocytes are a potential source of lung fibroblasts in idiopathic pulmonary fibrosis. *The International Journal of Biochemistry & Cell Biology* vol. 40 2129–2140 (2008).
  40. Pardo, A. *et al.* Up-regulation and profibrotic role of osteopontin in human idiopathic pulmonary fibrosis. *PLoS Med.* **2**, e251 (2005).
  41. Pardo, A. & Selman, M. Lung Fibroblasts, Aging, and Idiopathic Pulmonary Fibrosis. *Ann. Am. Thorac. Soc.* **13** Suppl 5, S417–S421 (2016).
  42. Xu, Y. D. *et al.* Release of biologically active TGF- $\beta$ 1 by alveolar epithelial cells results in pulmonary fibrosis. *American Journal of Physiology-Lung Cellular and Molecular Physiology* vol. 285 L527–L539 (2003).
  43. Horan, G. S. *et al.* Partial Inhibition of Integrin  $\alpha\beta$ 6 Prevents Pulmonary Fibrosis without Exacerbating Inflammation. *American Journal of Respiratory and Critical Care Medicine* vol. 177 56–65 (2008).
  44. Munger, J. S. *et al.* A Mechanism for Regulating Pulmonary Inflammation and Fibrosis: The Integrin  $\alpha\beta$ 6 Binds and Activates Latent TGF  $\beta$ 1. *Cell* vol. 96 319–328 (1999).
  45. Cosgrove, G. P. *et al.* Pigment Epithelium-derived Factor in Idiopathic Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine* vol. 170 242–251 (2004).
  46. Kotani, I. *et al.* Increased procoagulant and antifibrinolytic activities in the lungs with idiopathic pulmonary fibrosis. *Thrombosis Research* vol. 77 493–504 (1995).
  47. Mora, A. L., Rojas, M., Pardo, A. & Selman, M. Emerging therapies for idiopathic pulmonary fibrosis, a progressive age-related disease. *Nat. Rev. Drug Discov.* **16**, 810 (2017).
  48. Scotton, C. J. *et al.* Increased local expression of coagulation factor X contributes to the fibrotic response in human and murine lung injury. *Journal of Clinical Investigation* (2009) doi:10.1172/jci33288.
  49. Vaughan, A. E. *et al.* Lineage-negative progenitors mobilize to regenerate lung epithelium after major injury. *Nature* vol. 517 621–625 (2015).

- 
50. Kim, K. K. *et al.* Alveolar epithelial cell mesenchymal transition develops in vivo during pulmonary fibrosis and is regulated by the extracellular matrix. *Proceedings of the National Academy of Sciences* vol. 103 13180–13185 (2006).
  51. Rock, J. R. *et al.* Multiple stromal populations contribute to pulmonary fibrosis without evidence for epithelial to mesenchymal transition. *Proceedings of the National Academy of Sciences* vol. 108 E1475–E1483 (2011).
  52. Nieto, M. A. Epithelial plasticity: a common theme in embryonic and cancer cells. *Science* **342**, 1234850 (2013).
  53. Tzouvelekis, A., Eickelberg, O., Kaminski, N., Bouros, D. & Aidinis, V. *Pulmonary Fibrosis*. (Frontiers Media SA, 2019).
  54. Mouratis, M. A. & Aidinis, V. Modeling pulmonary fibrosis with bleomycin. *Current Opinion in Pulmonary Medicine* vol. 17 355–361 (2011).
  55. Schiller, H. B. *et al.* Time- and compartment-resolved proteome profiling of the extracellular niche in lung injury and repair. *Molecular Systems Biology* vol. 11 819 (2015).
  56. Martin, G. M. & Ladd, P. D. Faculty Opinions recommendation of Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature* (2017) doi:10.3410/f.727464464.793531568.
  57. Enge, M. *et al.* Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. *Cell* vol. 171 321–330.e14 (2017).
  58. Moliva, J. I. *et al.* Molecular composition of the alveolar lining fluid in the aging lung. *Age* **36**, 9633 (2014).
  59. Plantier, L. *et al.* Activation of Sterol-response Element-binding Proteins (SREBP) in Alveolar Type II Cells Enhances Lipogenesis Causing Pulmonary Lipotoxicity. *Journal of Biological Chemistry* vol. 287 10099–10114 (2012).
  60. Alder, J. K. *et al.* Telomere dysfunction causes alveolar stem cell failure. *Proceedings of the National Academy of Sciences* vol. 112 5099–5104 (2015).



---

## Acknowledgements

I feel extremely fortunate to have had the opportunity to start and finish my PhD in Munich in the laboratory of Herbert Schiller. I could not express enough how happy I am to have met and worked with you for the past four years and leaving your lab brings mixed feelings to me. I am sad to be disembarking the train that started its journey roughly five years ago with the establishment of your first lab as a young P.I, but in the same time I am glad to be leaving a better scientist and a more complete person than I had arrived. You have influenced my life and character in many ways and I really hope we will be seeing each other often in the future. I have said this many times but it still stands true. Had I had the opportunity to go back and do everything from scratch I would have taken the same path, same lab, same mentor, same jokes, same people, same successes same failures. Thank you, Herbert!

I really want to take this opportunity to also thank all the members of our awesome lab with special thanks to those who have been by my side almost from day one. Thank you, Max, Christoph, Laura, Gabi, Silvia and thank you Meshal! You may have been the late arrival but you are definitely a keeper! Many thanks to our Master students who have taught me more than I could have ever imagined. I discovered through you guys that teaching is like learning just twice as fast! Thank you, Mert, Paulina and Pawandeep for that!

Many thanks to the CPC and all the people involved in making it feel like a small family where PhD students belong in and can count on. Thank you Silke, Claudia, Mareike, Doreen and Darcy for making the Research school rock!

I extend my gratitude to my thesis committee, Prof. Dr. Jürgen Behr, Prof. Dr. Heiko Adler, Prof. Dr. Silke Meiners and Dr. Herbert Schiller for their immense support and guidance throughout these four years.

Many thanks to Bavaria and Munich for greeting me in the best way possible. I might be originally Greek however four years around the Isar makes me feel part Bavarian in heart. I know you don't mind that. Thank you, Munich!

Finally, I would like to thank and also dedicate my work to three very important people that have defined these past years and will be defining the rest of my life:

To **the friend I have had** since I arrived in Munich, we started this journey together and I wouldn't have it any other way. I love you from the bottom of my heart and will do so forever

*...to Angela*

To **the friend I lost** during the journey. Your loss will always be felt in my heart and I will always remember you

*...to Magda*

To **the friend I found** along the way. My little boy to whom I wish the best of luck in his life. Always be happy

*...to Stavros*