



Dissertation

Zum Erwerb des Doctor of Philosophy (Ph.D.)

an der Medizinischen Fakultät der

Ludwig-Maximilians-Universität zu München

A single cell transcriptomics and proteomics
perspective on
stromal cell diversity and fibrosis in the lung

vorgelegt von

Christoph Mayr

aus Dachau

am

23.03.2020

Supervisors: Prof. Dr. med. Jürgen Behr

Second evaluator: Dr. rer. nat. Herbert Schiller

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

Date of oral defense: 05.11.2020

Eidesstattliche Erklärung

Ich, Christoph Mayr, versichere hiermit an Eides statt, dass die vorgelegte Dissertation mit dem Thema

„A single cell transcriptomics and proteomics perspective on stromal cell diversity and fibrosis in the lung“

von mir selbständig angefertigt wurde, ich mich außer der Angegebenen keiner weiteren Hilfsmittel bedient habe und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

Außerdem erkläre ich hiermit, dass die hier vorgelegte Dissertation nicht in gleicher oder in ähnlicher Form bei einer anderen Stelle zur Erlangung eines akademischen Grades eingereicht wurde und dass ich mich nicht anderweitig einer Doktorprüfung ohne Erfolge unterzogen habe.

München, den 23. März 2020

Christoph Mayr

Confirmation of congruency

I hereby declare that the electronic version of the submitted thesis, entitled

“A single cell transcriptomics and proteomics perspective on stromal cell diversity and fibrosis in the lung”

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

Munich, March 23th, 2020

Christoph Mayr

Summary

Single cell genomics is revolutionizing biology and medicine by enabling a novel cell type specific viewpoint on organs and the specific cellular changes in disease. The structure of the lung, fabricated by at least 40 different unique cell types with specific functions, is uniquely designed to accomplish delivery of oxygen and removal of carbon dioxide from the circulation. Idiopathic pulmonary fibrosis destroys the structure of the lungs by progressive accumulation of scar tissue, produced by stromal cells, and ultimately leads to respiratory failure..

The aim of this thesis was to dissect molecular and cellular heterogeneities across different forms of pulmonary fibrosis, on the level of single cells in human patient lungs, as well as body fluids, such as bronchoalveolar lavage and plasma. I also investigated the poorly characterized diversity of stromal cell identities in mouse and human lungs and characterized their activation states in the bleomycin lung injury model and human lung fibrosis.

Single cell transcriptomics of large fibrosis patient cohorts allowed us to assess cell state changes in human fibrosis at cellular resolution. Mass spectrometry driven proteomics enabled us to study bronchoalveolar lavage and plasma protein compositions in patients from several independent ILD cohorts. Ultimately, a multi-modal analysis strategy was used to integrate the cell state descriptions on single cell level with the lavage and plasma proteomes and associated clinical meta-data. This revealed that fluid proteome signatures were predictive of specific cell state changes in the lung, that were associated with diagnosis, lung function, smoking and injury status. Many of the predictive proteins in body fluids were derived from activated fibroblasts of the fibrotic lung.

The second part of the thesis highlights the previously uncharacterized diversity of fibroblasts and other stromal cells in the lung. Using single cell transcriptomics I identified the unique transcriptional signatures of six distinct stromal cell identities. Using multiplexed single molecule fluorescence in situ hybridization, I located these six cell types by expression of sets of specific marker genes to their stereotypic locations in the tissue context. We distinguish adventitial, peribronchiolar, alveolar lipofibroblasts, a novel Hhip+ fibroblast type localized around bronchi and sparsely within alveoli, from pericytes and smooth muscle cells.

High resolution time-series single cell analysis of the bleomycin mouse model of lung fibrosis revealed differentiation trajectories and distinct activation states for all those stromal cell types. In contrast to a model in which various stromal cells such as alveolar fibroblasts and pericytes may converge on the same myofibroblast state, my work shows that all six stromal cell identities adopt very specific and unique activated states during injury repair. These results provide a molecular framework to physically isolate distinct stromal cell types in both their ground state and activated states and study their individual effect on alveolar stem cells in future work.

Zusammenfassung

Die Analyse der Genexpression von einzelnen Zellen revolutioniert die Biologie und Medizin, indem sie eine neue zelltypspezifische Sichtweise auf Organe und die spezifischen zellulären Veränderungen bei Krankheiten ermöglicht. Die Struktur der Lunge, die aus mindestens 40 verschiedenen, einzigartigen Zelltypen mit spezifischen Funktionen besteht, ist so konzipiert, dass sie die Zufuhr von Sauerstoff und den Abfluss von Kohlendioxid aus dem Blutkreislauf ermöglicht. In der idiopathischen Lungenfibrose wird die Struktur der Lunge durch die von Stromazellen erzeugte, fortschreitende Anhäufung von Narbengewebe zerstört, was schließlich zum Atemversagen führt.

Das Ziel dieser Arbeit war es, molekulare und zelluläre Unterschiede über verschiedene Formen der Lungenfibrose hinweg zu analysieren. Dies geschah auf Ebene einzelner Zellen in der Lunge von Patienten sowie in deren Körperflüssigkeiten wie etwa der bronchoalveolären Lavage und des Blutplasmas. Ich untersuchte darüber hinaus die bislang schlecht charakterisierte Vielfalt der Stromazellen in der Lunge von Mäusen und Menschen und charakterisierte dabei ihre Aktivierungszustände im Bleomycin Mausmodell sowie in der menschlichen Lungenfibrose.

Die Analyse der Einzelzell-Transkriptome einer großen Kohorte von Patienten ermöglichte es, die krankheitsbedingten Zustandsänderungen in der Fibrose mit einer zellulären Auflösung zu beurteilen. Durch den Einsatz von Massenspektrometrie konnten außerdem Proteome der bronchoalveolären Lavage und die Zusammensetzung von Plasmaproteinen bei Patienten aus mehreren unabhängigen ILD-Kohorten untersucht werden. Mithilfe einer bioinformatischen Analysestrategie konnte die Beschreibungen des Zellzustandes auf Einzelzellebene mit den Lavage- und Plasmaproteomen, sowie den zugehörigen klinischen Daten der Patienten, korreliert werden. Dadurch wurde gezeigt, dass die Proteomsignaturen prädiktiv für spezifische Zellzustandsänderungen in der Lunge sind, die mit der Diagnose, der Lungenfunktion und dem Raucher- und Krankheitsstatus der Patienten in Verbindung stehen. Viele der prädiktiven Proteine in den Körperflüssigkeiten stammten dabei von aktivierten Fibroblasten aus fibrotischem Lungengewebe.

Der zweite Teil der Arbeit beleuchtet die bisher nicht charakterisierte Diversität von Fibroblasten und anderen Stromazellen in der Lunge. Mit Hilfe von Einzelzell-Transkriptomik identifizierte ich die individuellen transkriptionellen Signaturen von sechs verschiedenen Stromazelltypen. Mittels Fluoreszenz-in-situ-Hybridisierung lokalisierte ich diese sechs Zelltypen durch ihre spezifischen Markergene in ihren jeweiligen Nischen im histologischen Gewebekontext. Wir identifizierten neben Perizyten und glatten Muskelzellen auch adventitielle, peribronchioläre und alveoläre Lipofibroblasten, sowie einen neuartigen Hhip⁺-Fibroblastentyp, der um die Bronchien herum und spärlich in den Alveolen lokalisiert ist.

Eine zeitlich hochauflösende Untersuchung des Bleomycin-Mausmodells der Lungenfibrose auf Einzelzellebene ergab für all diese Stromazelltypen unterschiedliche Differenzierungsverläufe und Aktivierungszustände. Im Gegensatz zu einem Modell, in dem verschiedene Stromazellen, wie Alveolarfibroblasten und Perizyten auf denselben Myofibroblastenzustand konvergieren können, zeigt meine Arbeit, dass alle sechs Stromazelltypen während des Regenerationsprozesses sehr spezifische und einzigartige Aktivierungszustände annehmen. Diese Ergebnisse bieten eine Grundlage, um verschiedene Stromazelltypen sowohl in ihrem Grundzustand als auch in ihren aktivierten Zuständen physikalisch zu isolieren, um ihre individuelle Wirkung auf alveoläre Stammzellen in zukünftigen Arbeiten zu untersuchen.

Contents

Eidesstattliche Erklärung.....	3
Confirmation of congruency	4
Summary	5
Zusammenfassung	7
1 Introduction	14
1.1 Lung structure and composition	14
1.1.1 Airway cells of the lung	14
1.1.1 Alveolar compartment.....	15
1.2 Pulmonary fibrosis	19
1.2.1 Wound repair	19
1.2.2 Interstitial lung disease (ILD).....	20
1.2.3 Idiopathic pulmonary fibrosis (IPF).....	22
1.2.4 Pathogenesis of idiopathic pulmonary fibrosis (IPF).....	23
.....	24
1.2.5 The role of fibroblasts in fibrosis	25
1.2.6 Bleomycin injury: an <i>in vivo</i> fibrosis model in mice	26
1.3 Biomarkers.....	27
1.3.1 General concept of biomarkers	28
1.3.2 Clinical proteomics based on mass spectrometry	28
1.3.1 Single cell RNA sequencing	30
2 Objectives.....	31
2.1 First chapter.....	31
2.2 Second chapter.....	32
3 Material and Methods	34
3.1 Methods human lung project.....	34
3.1.1 Lung samples	34
3.1.2 Lung tissue processing	35

3.1.3	Single cell sequencing of human tissue using Dropseq	35
3.1.4	Processing of single cell data from Munich	35
3.1.5	Analysis of single cell data from Munich	36
3.1.6	Integration of Chicago, Nashville and Munich single cell data.....	36
3.1.7	Differential expression analysis.....	37
3.1.1	Cell type signature enrichment analysis.....	38
3.1.2	Random forest prediction.....	38
3.1.3	Clinical parameters.....	38
3.1.4	BAL procedure.....	39
3.1.5	Mass spectrometry	39
3.1.6	Mass spectrometry bioinformatics and statistical analyses	40
3.2	Methods for the mouse lung mesenchyme project.....	40
3.2.1	Animal handling and bleomycin administration	40
3.2.2	Generation of single cell suspension from whole lung.....	41
3.2.3	Magnetic-activated cell sorting (MACS)	42
3.2.4	Processing of WT and bleomycin mouse single cell data.....	42
3.2.5	Analysis of mouse the WT and bleomycin single cell data	43
3.2.6	SCRINSHOT staining	44
3.2.7	SCRINSHOT data analysis.....	44
4	Results	47
4.1	Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids	47
4.1.1	Introduction.....	47
4.1.1	An integrated single cell atlas of human lung fibrosis.....	49
4.1.2	Human lung bronchoalveolar lavage fluid (BALF) proteomes reflect changes in disease activity	57
4.1.3	Correspondence of fluid proteins with transcriptional changes in specific cell types	67
4.1.4	Smoking induces highly persistent cell state changes in the lung	70

4.1.5	Biomarkers of lung health in plasma proteomes.....	74
4.1.6	Discussion	78
4.2	Mesenchymal heterogeneity in the lung	82
4.2.1	Introduction.....	82
4.2.1	Fibroblast heterogeneity in the adult wild-type mouse lung	84
4.2.2	Spatial mapping of mouse lung fibroblast subtypes	89
4.2.3	Longitudinal changes in fibroblast subtypes in bleomycin time course	98
4.2.4	Comparison of mesenchymal cell states in mouse and human lung injury on a single cell level.....	104
4.2.5	Discussion	110
5	Conclusions and future directions.....	113
6	Bibliography	118
	Publications	131
	Ackowlegmenets.....	142
	Curriculum Vitae	Fehler! Textmarke nicht definiert.

1.1 Lung structure and composition

1 Introduction

In times of increasing environmental awareness, due to the effects of human-made pollution on nature becoming more and more obvious, the lung, as the biggest interface to our surroundings, needs special attention. The lung is constantly subject to harmful exposure and therefore the internal organ most vulnerable to inhaled substances, particles or infectious organisms in ambient air. Globally, over three billion people are exposed to toxic air or tobacco smoke, making respiratory diseases an immense economical and health burden and the third leading cause of death worldwide (Collaborators 2016).

1.1 Lung structure and composition

For the recognition and awareness of abnormal structures upon diseases, it is essential to have an understanding of normal lung anatomy. The lung is the only respiratory organ found in mammals and is uniquely designed to accomplish its function of facilitating the delivery of oxygen and the removal of carbon dioxide from the circulation. Residing in the chest cavity, the lungs are a paired but asymmetrical organ that can be divided into structural units. Anatomically, the left and the right lung split up in lobes that consist of pulmonary segments in which the airway tree branching, starting with the trachea, reaches from the main bronchus into several subdivisions of bronchioles. Distal to this tracheobronchial tree, the active gas exchange happens in the alveoli, which comprise the alveolar sacs that derive from the alveolar ducts, branching off the respiratory bronchioles. The bronchi together with the pulmonary arteries are surrounded by connective tissue in bronchovascular bundles. Smooth muscle, basement membrane and connective tissue provide the basic structure of the airways (Fig. 1) (Tomashefski J.F. 2008).

On a cellular level, both airway and alveolar regions contain their own cell populations with distinct cell frequencies and cellular functions. The bronchial tree is covered with a continuous layer of epithelial cells that play a crucial role in maintaining the air flow, and are central to the defense of the lung against inhaled exposures and in regeneration processes.

1.1.1 Airway cells of the lung

The epithelium in the airways consist of the six major cell types, that are ciliated cells, secretory, goblet and clara cells, club cells, basal cells, and few rare cell types such as the neuroendocrine cells (Fig. 1) (Crystal et al. 2008). Located below the surface epithelial cell layer, basal cells serve as progenitor cells from which goblet, ciliated, and secretory cells differentiate (Whitsett 2018).

1.1 Lung structure and composition

Club and goblet cells main functions are the secretion of among others peroxidases, lactoferrin, albumin, lysozyme, and mucus, with the latter being constituted from over 20 mucins genes (Stripp 2008, Corfield 2015).

On top of the epithelium lays the airway fluid which, in cooperation with the cilia of ciliated cells, catches foreign particles and organisms that want to enter the airways and pushes those intruders out of the airways (Ng et al. 2004). Underneath lays the basement membrane, which is an approximately 7 μm thin layer of ECM that provides support but also signaling for and to the epithelium. Its major molecular constituents are collagens, fibronectin, complexes of laminins and nidogens, and proteoglycans (Laitinen et al. 1994). The described epithelial cell populations vary in their relative frequencies, as the airways branch from the level of the large airway down the tree to the respiratory bronchioles. Thereby, the number of cartilage cells decreases and more clara secretory cells emerge, while at the same time the cartilage disappears from the airway walls.

1.1.1 Alveolar compartment

Where bronchiolar epithelial cells are no longer present, the airway tree ends in the distal part of lung and leads into the alveolar sacs; structures from which multiple alveoli branch off in all directions. There are about 300 million alveoli in a 70 kg man, with an alveolar surface for gas exchange of about 143 m^2 , but at the same time these compartments only have diameters of 150 to 500 μm (average 250 μm) (Crapo RO 1998). The cell populations in the alveolar region can be categorized in compartments such as the epithelium, mesenchyme, endothelium, and immune cells.

Alveolar epithelial cells are constituted by alveolar type I (AT-1) and type II (AT-2) cells. AT-1 cells are thinly shaped and span the alveolar wall. Due to their flat shape, they provide a huge surface area, covering over 90% of the alveolar surface and facilitate the gas exchange due to their direct contact with the endothelium (Sirianni et al. 2003). In contrast, the AT-2 cell is a cuboidal cell and possesses several lamellar inclusion bodies containing lipids, which are needed for the production of surfactant. Besides secreting surfactant, AT-2 cells also serve as stem cells, being able to repopulate AT-1 cells or alternatively producing more AT-2 cells after injury by clonal expansion (Evans et al. 1975, Barkauskas et al. 2013). Together with stem cells from the club cell lineage, AT-2 cells were recently shown to be able to undergo transcriptional convergence into a newly discovered Krt8 expressing intermediate cell state. This cell state has important functions in the injury state and serves as progenitor that later can differentiate into AT-1 cells (Strunz et al. 2019).

1.1 Lung structure and composition

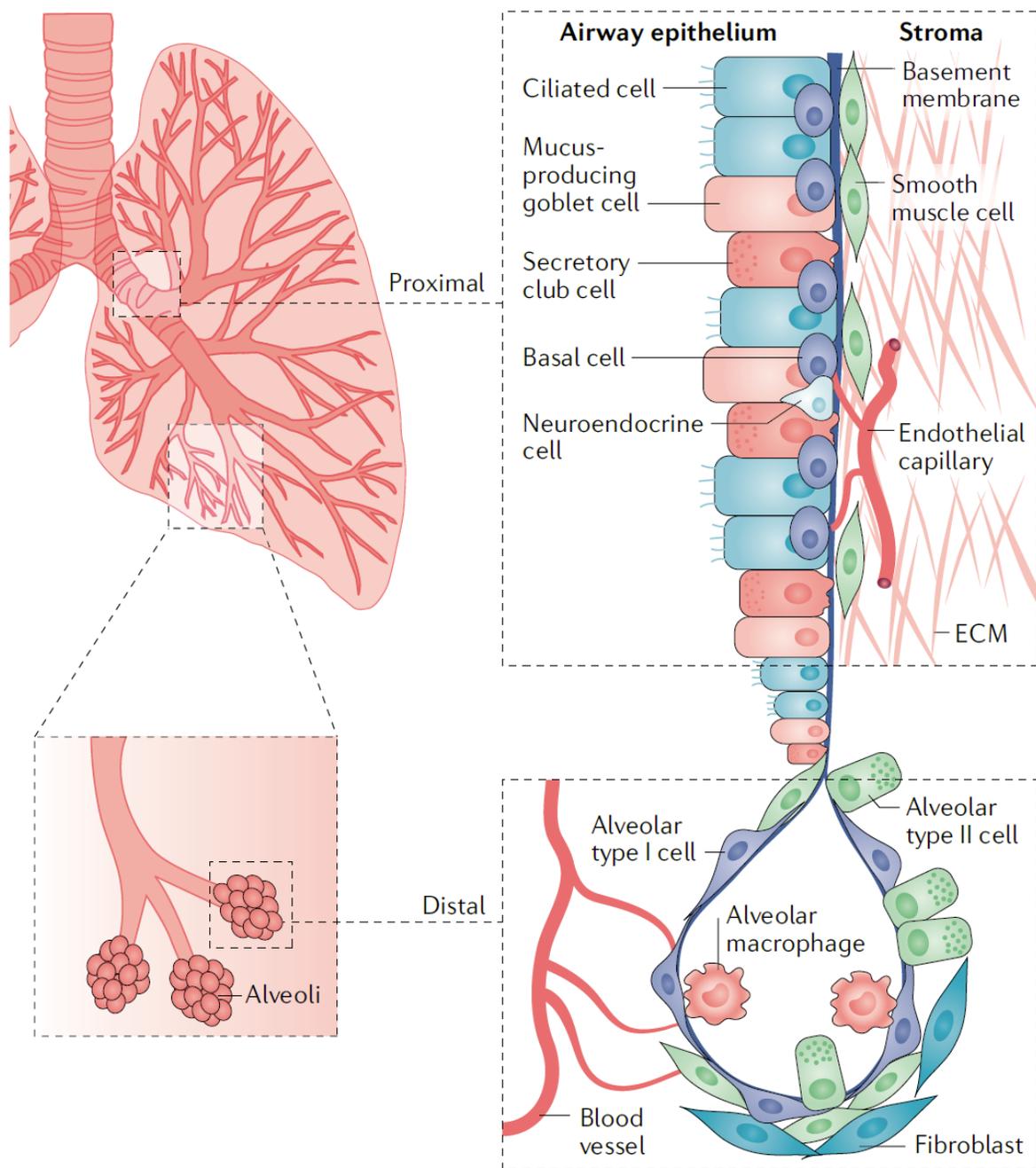


Figure 1. The heterogeneous microenvironment of the lung. The schematic shows anatomic regions of a normal healthy lung, divided in proximal and distal airways and the cell types they encompass. The proximal airways are hereby composed of a variety of epithelial cells, with special focus on mucus producing cells as well as ciliated cells for immune defense. The distal airways and the alveoli is where the gas exchange is happening, facilitated by the thin AT-1 cells and vascular capillaries. All of that structure is supported by extracellular matrix (ECM) produced by fibroblasts. Figure taken from (Altorki et al. 2019).

1.1 Lung structure and composition

Together with the AT-1 cell, the endothelial cell cytoplasm is spread thinly on both sides of the basal membrane, forming a barrier between air and blood with a thickness of around 0.6 μm . The pulmonary endothelium around the alveoli is the largest and most dense vascular bed in the human body. At first glance, it represents a tight barrier between the circulating blood and the vascular wall. However, due to its unique structure of endothelial cells that are connected with each other by loose junctions, the endothelium allows for passage of fluids and macromolecules into the interstitium (Renkin 1992). In larger veins or arteries, smooth muscle cells (SMCs) enclose the vessels, however, in the fine capillary structures of the alveolus, pericytes are the supporting cell type (He et al. 2018).

Although not the focus of this thesis, it should be noted that immune cells of both the lymphoid lineage such as natural killer (NK) cells or T- and B- cells, as well as from myeloid lineage such as basophils, monocytes or macrophages are present in the lung and fulfill important functions in the innate and adaptive immunological defense. Especially the macrophages should be mentioned, that can be classified into interstitial or alveolar macrophages. Depending on their location they can interact with lymphocytes and actively participate in antigen presentation in the interstitium or remove trapped particles or pathogens via the mucosal clearance (Benoit et al. 2008, Balhara et al. 2012).

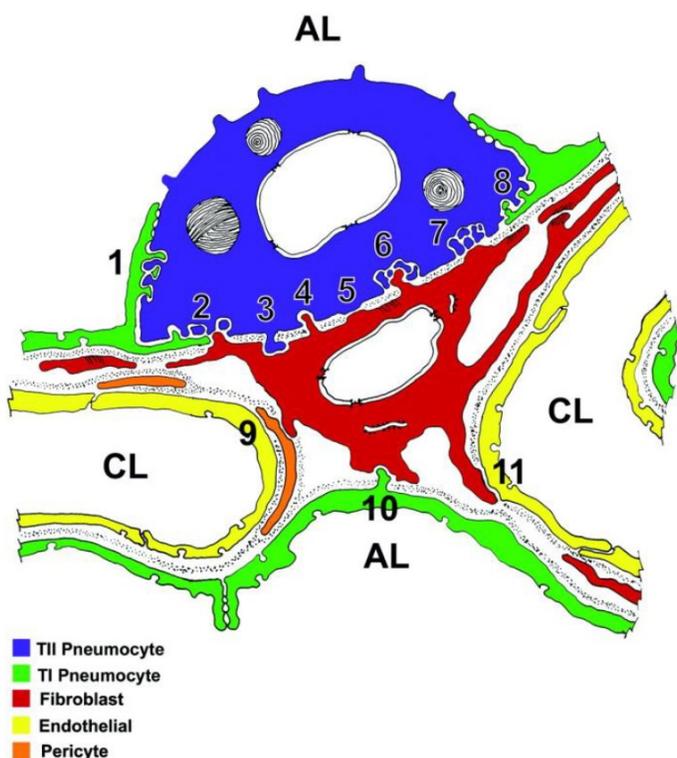


Figure 2. The alveolar niche in the lung. The figure shows a schematic of the alveolar niche in the lung, made out of real composite images of histologic electron microscopy serial sections. It illustrates nicely how the AT-1 cell (green) and the endothelial cells (yellow) are thin and span the alveolar lumen (AL) and the capillary lumen (CL), only interrupted by the stem cell like AT-2 cell. The fibroblast (red) sits in the middle of all of this and links the endothelium to the epithelium by contacts through basal lamina (stippling) holes of both endothelial cells and AT-2. In the same manner, AT-2 cells are in contact with the fibroblast. Figure taken from (Sirianni et al. 2003).

1.1 Lung structure and composition

In the adult lung, the alveolar niche, as most distal part in the airway tree, is where it all comes together (Fig. 2). Epithelial cell populations receive multiple signals from the surrounding mesenchymal cells, which modulate the homeostasis, the ability to differentiate and proliferate, as well as injury repair (Fig. 2) (as outlined in 1.2.1 Wound repair). But also endothelial cells interact with fibroblasts, that support capillary structures by remodeling the local ECM components and supplying various growth factors (Hughes 2008, Costa-Almeida et al. 2015). The mesenchymal stromal cells in the lung are comprised of smooth muscle cells, the already mentioned endothelial cells and several populations of fibroblasts that are poorly characterized (Zepp et al. 2017). Pulmonary fibroblasts are mesenchymal cells of the interstitium. Their plasticity and heterogeneity have complicated their study so far, nevertheless they are of great interest due to their various functions and implications in fibrotic diseases.

On the one hand, in homeostasis, fibroblasts are involved with the epithelium in several ways. Lipofibroblasts assist AT-2 cells in their production of surfactant, composed of approximately 90% lipids and 10% proteins, that is crucial for lowering surface tension and as innate immune defense (Griffin et al. 1993, Haagsman et al. 2001, Torday et al. 2016). Another important component is ECM that is produced by fibroblasts, which acts as a reservoir for growth factors and provides mechanical stimuli, which are then translated into biochemical signals through integrins via mechanotransduction (Hsu et al. 2011, Hsu et al. 2012). Signaling also happens via other critical pathways. For example, fibroblast growth factor (FGF), produced by fibroblasts, was shown to break the quiescence of club stem cells which induces proliferation and initiates epithelial repair (Volckaert et al. 2014). Fibroblasts, reacting to signaling through the Wnt-pathway were also found in the proximity of AT-2 cells, enhancing their expansion (Zepp et al. 2017). Sonic hedgehog (SHH) is another crucial signaling pathway, for actively maintaining the quiescence of the epithelium during homeostasis and gets activated during injury (Peng et al. 2015).

On the other hand, fibroblasts express receptors for pro-fibrotic molecules such as, platelet-derived growth factor (PDGF) and transforming growth factor-beta (TGF- β), which are known drivers of pathological fibroblast accumulation and persistence (Barron et al. 2016). Upon an activation stimulus, pulmonary fibroblasts can migrate and differentiate into active myofibroblasts. These cells are characterized as highly contractile fibroblasts that express large amounts of alpha smooth muscle actin (alpha-SMA) (Selman et al. 2016).

1.2 Pulmonary fibrosis

1.2.1 Wound repair

The ability to respond to injury and to repair tissue is a fundamental property of all multicellular organisms. The lungs are constantly exposed to various injuries but show an astonishing ability to resolve and repair these challenges with finely tuned biological machinery. In all organs, the normal, physiological response to injury occurs in overlapping but distinct stages: coagulation, inflammation, new tissue formation, and remodeling (Gurtner et al. 2008) (Fig. 3).

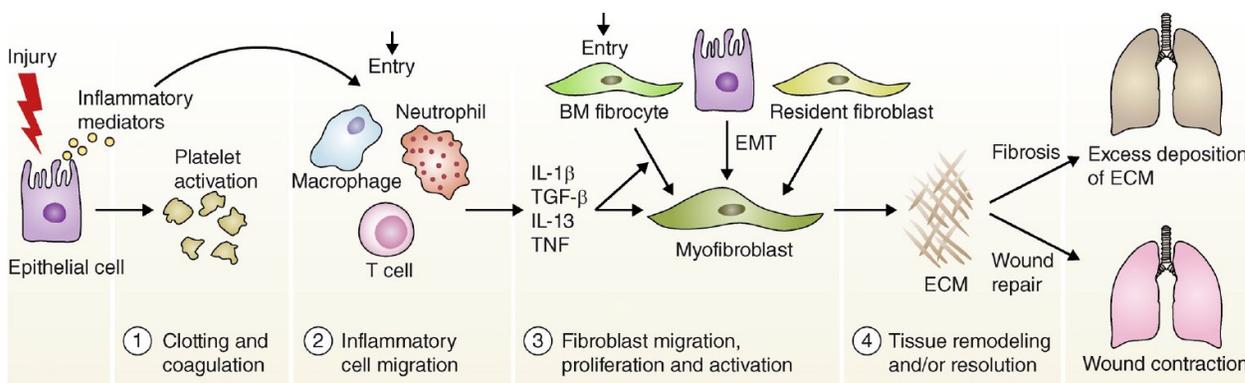


Figure 3. Wound repair and tissue remodeling in the lung. The scheme depicts the four stages of physiological lung wound healing. First, epithelial damage activates platelets and leads to secretion of inflammatory mediators. Inflammatory cells get recruited to the site of injury to eliminate invading pathogens and secrete pro-fibrotic factors such as TGF- β 1. Activated fibroblasts and myofibroblasts secrete ECM proteins to build an intermediate repair matrix, that facilitates contraction of the wound and migration to the site of injury. The origin of these myofibroblasts is however heavily disputed. If a dysregulation of this process happens, and excessive amounts of ECM are deposited which leads to destruction of functional tissue, the process is regarded as fibrotic tissue remodeling. Figure taken from (Wynn 2011).

At first, circulating platelets are activated by released inflammatory mediators to trigger the formation of a plug in the process of coagulation to prevent ongoing blood and fluid losses, infection, and to remove dead and dying tissue. Hemostasis is afterwards achieved by a provisional fibrin matrix, which becomes the scaffold for infiltrating cells. In the second step, platelet aggregation and secretion of chemokines, leads to recruitment and migration of neutrophils, lymphocytes and macrophages to the site of injury, to fight possible infections and remove cell debris. The inflammatory cells also secrete a variety of cytokines such as TGF- β 1, Platelet-PDGF, tumor necrosis factor alpha (TNF- α) and interleukin six and 13 (IL-6 and IL-13), that increase the inflammatory response and spark the activation, migration and proliferation of fibroblasts to myofibroblasts. The origin of myofibroblasts is not completely understood yet, and will be addressed in the second part of this thesis. However, besides the classical α -smooth muscle actin-expressing myofibroblasts, multiple mesenchymal cell populations and fibroblasts

1.2 Pulmonary fibrosis

subtypes seem to transform into activated-states to secrete distinct ECM components that form new tissue. In a process called fibrinogenesis, several activated fibroblast populations assemble a provisional repair matrix, which acts as scaffold and regulator for developmental programs. Morphogens enable the communication between mesenchymal and epithelial cells which is crucial to mediating re-epithelialization and reconstitution of the tissue. Finally, all of the processes activated after injury wind down and cease. Most of the recruited cells undergo apoptosis or exit from the wound, the provisional ECM is degraded and remodeled to rebuild the parenchymal tissue architecture (Gurtner et al. 2008, Kisseleva et al. 2008, Hecker et al. 2011, Hochreiter-Hufford et al. 2013, Wynn et al. 2016)

However, if this process becomes dysregulated, it can lead to the development of a permanent fibrotic “scar,” which is characterized by the excess accumulation of ECM components (e.g., hyaluronic acid, fibronectin, proteoglycans, and interstitial collagens) at the site of tissue injury. Consequently, fibrinogenesis is often defined as an out of control wound healing response.

1.2.2 Interstitial lung disease (ILD)

Interstitial lung diseases (ILD) or diffuse parenchymal lung disease (DPLD) are a heterogeneous group of over 200 clinically defined parenchymal pulmonary disorders. They are characterized by persisting inflammation and gradual scarring of tissue in the alveolar region that ultimately leads to pulmonary fibrosis, destruction of the lung parenchyma and respiratory failure that results in lung fibrosis. In consequence, patients exhibit impaired gas exchange capacity, phases of exacerbations, and sometimes heart failure owed to chronic lung disease (Thomeer et al. 2001, Neurohr et al. 2015). Only 35-40% of patients are diagnosed with a known aetiology (e.g. autoimmune ILDs, hypersensitivity pneumonitis, drug or radiation induced ILD), while most ILDs have an unknown aetiology (e.g. idiopathic interstitial pneumonias (IIP), sarcoidosis or connective tissue diseases related ILD (CTD-ILD)) (Thomeer et al. 2001). CTD-ILDs encompass various systemic immune disease, where autoantibodies lead to inflammation and organ damage (Wells et al. 2014). IIPs can be further grouped into the ‘rare IIPs’ (less than 1%) and ‘major IIPs’, including chronic fibrosing IIPs like idiopathic pulmonary fibrosis (IPF; approx. 50%) and idiopathic non-specific interstitial pneumonia (NSIP; approx. 25%), smoking related IIPs like respiratory bronchiolitis ILD (RB-ILD; approx. 10%) and acute/subacute IIPs like cryptogenic organizing pneumonia (COP; approx. 5%) (Travis et al. 2013, Neurohr et al. 2015). Furthermore, 10-14% of patients remain unclassifiable IIP, due to nonspecific clinical findings but also because the patients are unwilling or, due to their condition, unable to undergo further testing, such as lung biopsy (Ryerson et al. 2013, Hyldgaard et al. 2014, Jo et al. 2017, Guler et al. 2018). Accurate and early diagnosis of these ILD can be challenging, but is crucial for optimal treatment of patients.

1.2 Pulmonary fibrosis

An estimate of one hundred million deaths worldwide in the 20th century is supposed to be caused by tobacco (Laniado-Laborín 2009). Smoking is a major risk factor and involved in six of the eight leading causes of deaths, including respiratory, cardiovascular and several malignant as well as. Smoking might also be critically entangled in the genesis of ILD entities. Indeed, some ILD are directly smoking-related, such as RB-ILD and desquamative interstitial pneumonia (DIP). On a histologic spectrum both present with an accumulation of macrophage, and can only be distinguished either on the extent and distribution of macrophage influx, or in general by a lower amount of fibrosis (Travis et al. 2013).

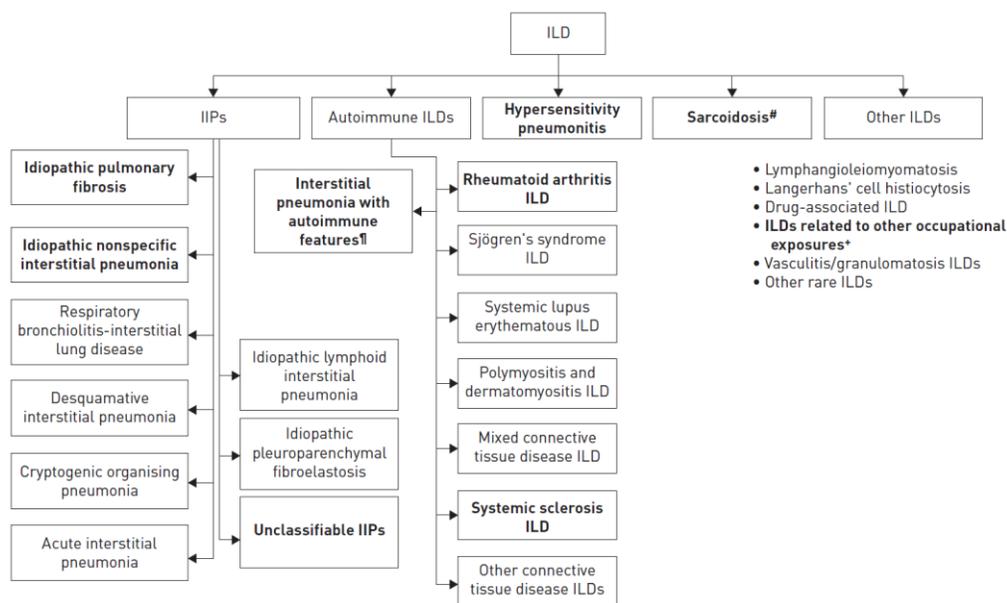


Figure 4. Interstitial lung diseases (ILD) are a heterogeneous group. Types of ILD that are likely to present with a progressive-fibrosing phenotype (indicated in bold). IIPs: idiopathic interstitial pneumonias. #: stage IV sarcoidosis only; ¶: not an established clinical diagnosis; +: e.g. asbestosis, silicosis. Figure taken from (Cottin et al. 2018)).

As highly diverse as the classification of disorders, are the clinical courses ILD patients experience, ranging from slow to rapid progressive (King et al., 2011). Progressive can be defined as a relative decline of $\geq 10\%$ in forced vital capacity (FVC), a relative decline of $\geq 15\%$ in diffusing capacity of the lung for carbon monoxide (DLCO), or worsening symptoms within a 24-month period (Cottin et al. 2018). Disease progression is often accelerated by acute exacerbations (AE), which occur with currently unpredictable probability and are associated with a high mortality (Collard et al., 2007; Collard et al., 2016).

1.2 Pulmonary fibrosis

1.2.3 Idiopathic pulmonary fibrosis (IPF)

Idiopathic pulmonary fibrosis (IPF) is the most studied and most prevalent common ILD. It is a specific form of chronic, progressive fibrosing interstitial pneumonia, which is associated with the poor prognosis of mortality 3 to 5 years after diagnosis being at 50% (Fig. 5) (Schwartz et al. 1994, King et al. 2001, Selman et al. 2002). As per multinational clinical definition IPF is of unknown cause, occurring primarily in older adults, limited to the lungs and requires the exclusion of other forms of ILD. Patients present with progressive worsening of dyspnea and lung function, histopathologically visible as lung scarring and a radiological pattern known as usual interstitial pneumonia (UIP). IPF is generally associated with older people, with most patients being between 50-70 years old at the time of diagnosis and it often affects men more than women, with the majority of patients having a smoking history (Annesi-Maesano et al. 2013, Patterson et al. 2017). Smoking is discussed as one of the risk factors for IPF but the questions of cause or consequence remains (Baumgartner et al. 1997). IPF is self-sustaining and causes progressive worsening in lung function, respiratory symptoms and thereby quality of life, making it a currently incurable disease with limited understanding of its molecular pathophysiology, and lung transplantation as only definitive therapy (Raghu et al. 2011).

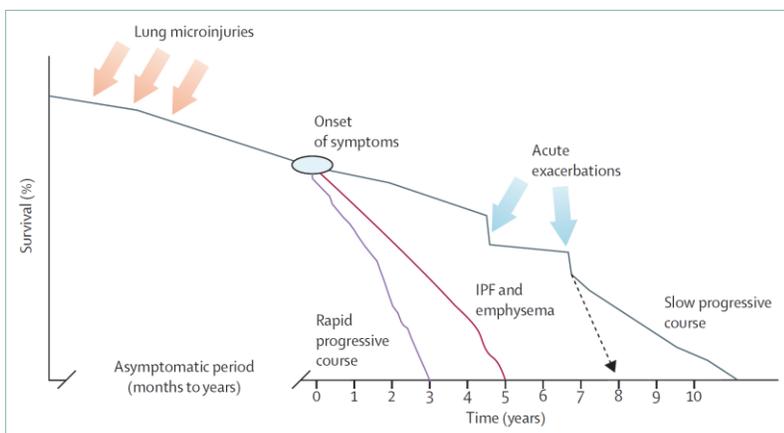


Figure 5. Clinical phenotypes of IPF. Most patients follow a relatively slow clinical and functional decline (slowly progressive) after diagnosis. Some patients however present with acute exacerbations that could initiate the terminal phase of their disease. A few patients have a rapidly progressive clinical course, partially due to comorbidities. Heavy smokers might develop emphysema on top, with shorter survival compared with patients with IPF alone. Figure taken from (Raghu et al. 2011)

Currently, the only approved treatments to slow disease progression in patients with IPF are *Nintedanib* and *Pirfenidon*, both antifibrotic therapies. One-year treatment of IPF patients with *Pirfenidon* slows down the decline in FVC and reduces the risk of death, disease progression and respiratory-related hospitalization in general (Noble et al. 2011, King et al. 2014, Noble et al. 2016, Ley et al. 2017, Nathan et al. 2017). While proven to act anti-inflammatory, antioxidant and antifibrotic, the exact molecular mode of action is not well known. *Nintedanib* on the other hand, is a triple receptor tyrosine kinase inhibitor, targeting the platelet-derived growth factor (PDGF),

1.2 Pulmonary fibrosis

vascular endothelial growth factor (VEGF), and basic fibroblast growth factor (FGF) receptor. It also has shown significant reduction on lung function decline and time to the first exacerbation. (Richeldi et al. 2011, Richeldi et al. 2014).

So far, treatment options are limited, which contributes to the low survival of patients. However, while IPF is the main reason for death, the cause of death in 30–40% of these patients is related to comorbidities, such as cardiovascular and thromboembolic disorders, gastro-oesophageal reflux disease (GORD), depression, sleep disorders, pulmonary hypertension, emphysema, diabetes and lung cancer. This association might be explained in parts by the elderly population, which is mainly affected by IPF, yet, a lot of the comorbidities have a negative impact on the prognosis of the disease (Collard et al. 2012, Kreuter et al. 2015, Kreuter et al. 2016, King et al. 2017).

However, as the term “idiopathic” implicates, the etiology of the disease has not been fully understood yet. A better understanding of IPF pathogenesis is needed to help to arise new opportunities for novel therapeutics and for the discovery of new biomarkers enabling early diagnosis of IPF patients. Importantly, over 16 novel therapies, exploring multiple biological pathways are currently in clinical phase II or phase III trials, validating both the complexity of the disease and the continuous effort to combat it (Somogyi et al. 2019).

1.2.4 Pathogenesis of idiopathic pulmonary fibrosis (IPF)

As described above, wound healing is a critical biological process, following injury, by replacing damaged or dead cells with connective tissue, resulting in the formation of a scar. Therefore scars are part of the normal physiology with useful functions by preventing blood loss at the site of injury, creating a barrier to prevent pathogen invasion, and providing a provisional matrix for subsequent re-epithelialization (Hecker et al. 2011). However, due to a dysregulated or aberrant repair mechanism, in pulmonary fibrosis, regeneration is insufficient or incomplete, leading to permanent scarring and remodeling of the lung interstitium, which destroys the original tissue architecture and function (Wynn 2011, Wuyts et al. 2013, Rockey et al. 2015). Fibrosis typically results from conditions persisting for several weeks, months or years in which tissue destruction and repair occur simultaneously leading to excessive connective tissue deposits. In the lung, the most recent understanding of the pathomechanisms of IPF is, that repetitive injury to epithelial cells in a genetically predisposed, elderly person could induce pro-fibrotic reprogramming on an epigenetic level, sustained epithelial cell senescence, dysregulated amplified production of fibrotic mediators and constant activation of fibroblasts. All these factors combined will result in uncontrolled wound repair that eventually will end in IPF (Martinez et al. 2017).

1.2 Pulmonary fibrosis

Despite having an unknown aetiology, several risk factors are known to play a role in the pathogenesis of IPF, including genetic variants, environmental exposures, chronic viral infections, smoking, or and certain comorbidities. Persistent inflammation is often involved in fibrotic diseases in different organs (Thannickal et al. 2004, Wynn 2011). In IPF, however, the role of inflammation is controversial, as fibrotic remodeling is happening without a remarkable inflammatory response. This is in agreement with clinical observations that anti-inflammatory therapies exhibit low efficiency and might even be harmful (Thannickal et al. 2004, Rafii et al. 2013). With age being the strongest risk factor for IPF, lung ageing or ageing of the cells in lung could be accelerated for IPF development. And indeed, AT-2 cells, that play an important role in regeneration of epithelial cells show almost all hallmarks of ageing cells in IFP lungs, including genomic instability, cellular senescence and altered intercellular communication (Selman et al. 2014, Lehmann et al. 2017). IPF remains a complex disorder resulting from interplay between multiple of these risk factors and with the relative contribution of each factor being depended on the individual patient (Selman et al. 2014). The complexity of IPF is also characterized by a wealth of interactions between epithelial and in particular mesenchymal cells, which are the main cell types involved in the continuous fibrotic remodeling, and the surrounding ECM (Wynn 2011).

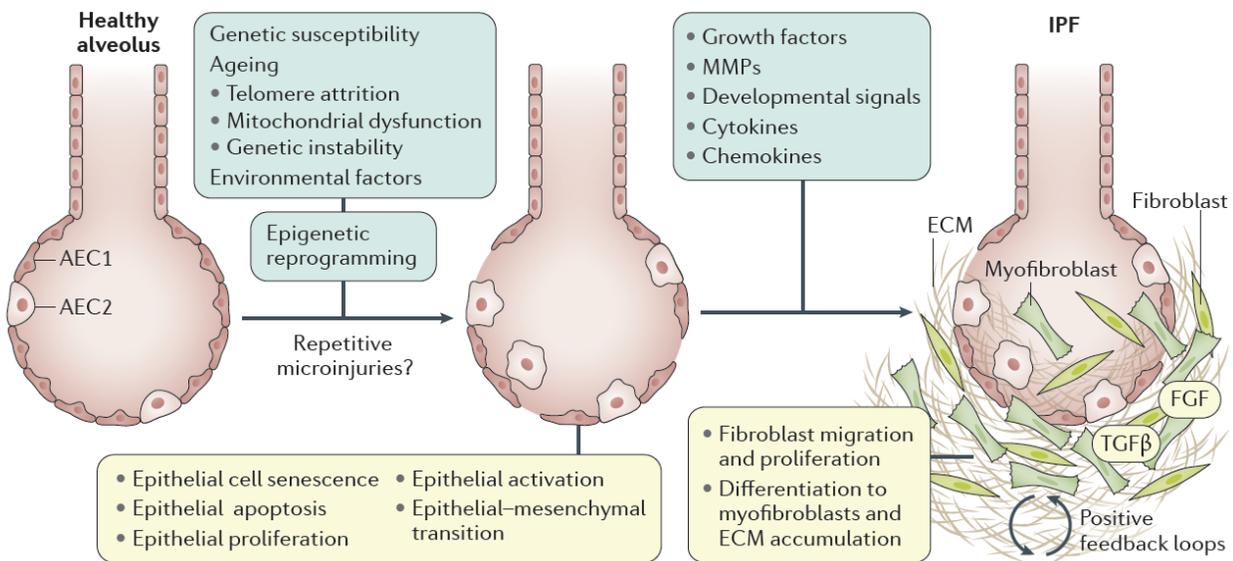


Figure 6. A pathogenic model of idiopathic pulmonary fibrosis. The interplay of the three factors, genetic susceptibility, especially changes to health of epithelial cells, environmental factors, such as smoking, and ageing associated cell changes lead to an epigenetic reprogramming that promotes dysregulated epithelial activation and ultimately may result in IPF. Repetitive micro injuries might induce or even accelerated the process. Dysfunctional wound healing will then trigger fibrotic tissue remodeling of the lung as described above. Figure taken from (Martinez et al. 2017)

1.2 Pulmonary fibrosis

1.2.5 The role of fibroblasts in fibrosis

It is widely accepted that mesenchymal cells especially fibroblasts have an essential role in wound healing by establishing tissue tension and building up a provisional repair matrix. In pulmonary fibrotic disease, however, these activated fibroblasts or myofibroblasts migrate to the site of injury, proliferate in numbers and accumulate in fibrogenic patches where they secrete excessive amounts of ECM proteins, such as elastin, fibronectin and collagens (Thannickal et al. 2004).

Myofibroblasts are elongated spindle-like shaped highly contractile cells, thereby sharing features with fibroblasts and SMCs, and can be identified by large stress fibers in the fibrotic foci (Jordana et al. 1994). These foci are regions of highly proliferative myofibroblast and ECM proteins, especially α SMA, with its gene Acta2, that can be seen in CT analysis of the patients in the areas of honeycombing (Fernandez et al. 2012).

Fibroblasts have to get activated and proliferate upon injury, however, the origin and source of myofibroblasts remains elusive. Different studies have proposed multiple possible origins, partially contradicting each other. One hypothesis proposes that myofibroblasts stem from resident fibroblasts with expression of high affinity type-2 TGF- β receptor (Hoyles et al. 2011). Another theory is that they might stem from bone-marrow derived circulating fibrocytes that migrate to the site of injury and differentiate into myofibroblasts (Phillips et al. 2004). Also often proposed are pericytes as the origin for myofibroblasts due to similar gene expression between these cell types (Hung et al. 2013). Some studies hypothesized that myofibroblast might evolve from epithelial cells, undergoing epithelial-to-mesenchymal transition (EMT) (Kim et al. 2006). However, this has been very clearly countered by a lineage tracing experiment of AT-2 cells and airway basal cells that after differentiation did not end up as mesenchymal cells, and was never proven again (Rock et al. 2011).

Myofibroblasts isolated from IPF tissue were shown to exhibit a pathological phenotype of uncontrolled proliferation, migration and they also seem get more resistant to programmed cell death (Maher et al. 2010). The changes in the ECM composition that myofibroblasts induce in IPF, as well as altered external stimuli such as growth factors induce a feedback loop for myofibroblasts (Hinz et al. 2001). With the production of an aberrant ECM with excessive amounts of α SMA positive stress-fibers, normal lung tissue is replaced, hindering the gas exchange. In addition, myofibroblasts were suggested to provoke epithelial apoptosis and obstruct cell regeneration by an imbalance in the epithelial-mesenchymal cross-talk, mediated by TGF- β , IL-6 or TNF- α signaling (Moodley et al. 2003, Frankel et al. 2006, Zhang et al. 2017). Precise mechanism and pathways are however still unknown and require further investigation into the receptor-ligand interactions between epithelium and mesenchyme.

1.2 Pulmonary fibrosis

1.2.6 Bleomycin injury: an *in vivo* fibrosis model in mice

The airway tree is complex and cell populations differ both by airway branching generation but also by species. Despite the known differences between animal models and humans, most of the current knowledge in basic lung research is driven mostly by studies of the mouse airways and lung (Crystal et al. 2008). Several *in vivo* animal models are used to conduct research on the pathomechanism of ILD, as access to human material is usually restricted or ethically reprehensible. The most common model to mimic human fibrosis features but also tissue repair and remodeling, is the bleomycin-induced lung injury.

Bleomycin is an antibiotic glycopeptide originally isolated and obtained from *Streptomyces verticillus* in 1966 (Meadors et al. 2006). Due to its anti-neoplastic properties, bleomycin has originally been used as a drug against skin tumors and squamous cell carcinomas, with higher doses, however, leading to pulmonary toxicity and development of ILD (Pedersen-Bjergaard et al. 1991, Saxman et al. 1997, Sleijfer 2001, O'Sullivan et al. 2003).

Bleomycin exerts its anti-neoplastic effects by inducing single- or double-strand DNA breaks via a complex of bleomycin, oxygen, and iron, and by working as endonuclease, though inhibition of DNA replication and DNA polymerase activity (Sikic 1986, Chandler 1990). The high toxicity which is used to combat cancer cells in the skin, but causes lesions in the lung is due to a lack of the bleomycin hydrolase in human and certain rodents that enzymatically could deactivate bleomycin. Histologic hallmarks of the bleomycin lung include alveolitis, alveolar and interstitial fibrosis, and dysplasia of the alveolar cells, with the distribution of fibrotic lesions to be found in both subpleural and periseptal areas (Comrie et al. 2015). Together with other observations such as clusters of mesenchymal cells, collagen deposition and damaged basement membrane, these hallmarks illustrate that bleomycin lung injury is able to reproduce many typical features of the human IPF disease (Moeller et al. 2008, Mouratis et al. 2011).

However, it should be noted that the model has limitations, with some pathological features not being mimicked, missing or are attributed to other ILD. The progressive and irreversible nature of fibrosis, rendering it incurable in human, is not present in the bleomycin model that shows resolution of fibrosis after day 56. Additionally, the inflammatory response which occurs early after bleomycin administration is more similar to acute lung injury than fibrosis. Despite these disadvantages, the bleomycin model remains a vital tool for investigation of fibrosis, if the correct time points are chosen; and even the limitation of occurring resolution can be turned around, by embracing the possibility to study rare regenerative processes.

Administration of bleomycin to mice can be achieved by different routes of intratracheal instillation, intratracheal micro sprayer aerosolization, intranasal application by face mask aerosolization, or oropharyngeal aspiration (Bivas-Benita et al. 2005, Egger et al. 2013, Bale et al. 2016, Barbayianni

1.3 Biomarkers

et al. 2018). A recent study comparing delivery routes, reported that oropharyngeal aspiration outperformed intratracheal delivery methods, with a greater ease of handling reduced animal burden and the reduction of experimental animal numbers (Barbayianni et al. 2018).

Bleomycin can be administered in a single dose or multiple doses, respectively. While repetitive administration mimics more the human disease by developing permanent fibrosis, a single experiment can take up to four month and needs substantial mouse numbers. Single dose administration of bleomycin leads to transient pulmonary fibrosis, and thus represents a more acute model that allows also the study of successful tissue restoration (Degryse et al. 2011, Schiller et al. 2015, Cao et al. 2016). In the bleomycin model, certain time points after administration have been established to capture the main aspects of injury and repair. At day three, first signs of inflammation and acute damage in the lung epithelium can be observed. Day seven is characterized by severe inflammatory reactions and early sings of fibrosis. Day 14 is the classical time point to observe active fibrogenesis with the deposition of ECM proteins. The climax of fibrosis is studied at day 21 when the provisional matrix is fully assembled for subsequent repair and re-epithelialization. Afterwards, from day 28 to day 56, tissue remodeling takes place and the lung parenchyma is almost completely regenerated. While these time points serve to present the critical phases of tissue repair, a higher temporal resolution study of the processes involved, with daily sampling as presented in the second part of this thesis, can be crucial to discover novel aspects and better interpret known observations (Strunz et al. 2019).

1.3 Biomarkers

ILDs in general with the multitude of unknown aetiologies and IPF in particular, are complex disorders with multiple pathogenetic pathways, different disease profiles and diverse responses to treatment. Most of the clinical disease classifications are not necessarily reflecting cellular and molecular pathophysiology. This can lead to the effect that patients with the same diagnosis might be highly heterogeneous, while patients with different diagnosis but similar histopathology might share important diagnostic and therapeutic aspects. Personalized medicine is aimed at distinguishing such disease subsets taking into account individual lifestyle, environmental exposures, genetic profiles and molecular pathways. The most important step in this concept is the identification of biomarkers that can be used to perform clinical classification diagnosis, prognosis, risk and treatment stratification based on biomarker signatures (Kokosi et al. 2018). Such signatures have the clear advantage, to be more robust and reliable compared to single protein or gene markers, but require next generation methodology such as MS-based proteomics or single cell sequencing to be identified.

1.3 Biomarkers

1.3.1 General concept of biomarkers

As announced by the US National Institutes of Health (NIH), a biomarker is a defined characteristic that can be quantified as an indicator of a normal biological process, pathogenic process, or a response to an exposure or intervention (Biomarker Definition Group, 2001).

Biomarkers play an important role for the decision making in the clinics. On the report of a survey at the Institute of Laboratory Medicine at the Klinikum Großhadern, 77% of all clinical decisions are based on laboratory tests, with most of these (42%) measuring the enzymatic activities or concentrations of proteins (Fig. 7a) (Geyer et al. 2017).

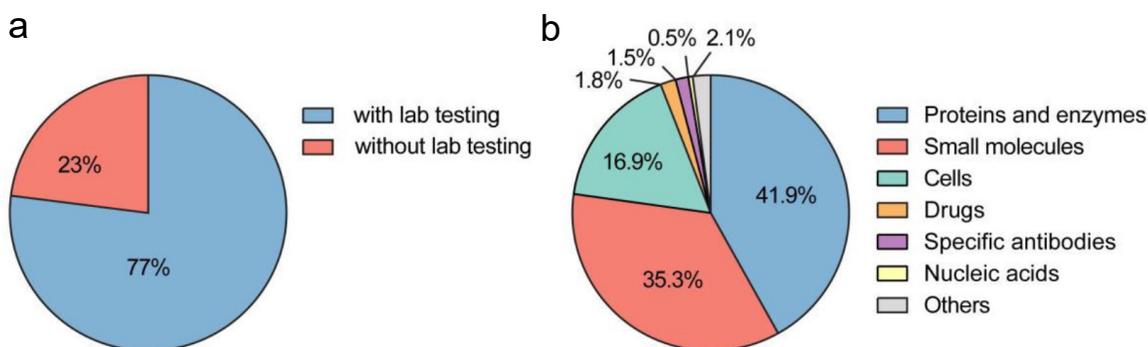


Figure 7. Statistics of clinical tests. (a) Proportion of clinical decisions that are made based on laboratory testing. (b) Proportion of clinical tests that are based on different molecule types. Adapted from (Geyer et al. 2017)

A vast majority of those assays, based on immune- or enzymatic reactions, target single proteins, due to limitations in multiplexing, antigen-antibody recognition or even detection of clinically important protein variants. Therefore, mass-spectrometry (MS) based methods offer an attractive alternative to the classical assays, by being able to discover multi-protein panels with high sensitivity. The last years have seen dramatic improvements in all areas of MS-based proteomic workflows, enabling broad application of quantitative proteomics in biological but also clinical research.

1.3.2 Clinical proteomics based on mass spectrometry

The proteome is the best direct molecular representation of a phenotype, be it a cell-line, a mouse or individual patients with a specific disease. Individual proteins are already applied as indicators of diseases in clinical practice and most drugs influence proteins. Clinical proteomics could be used in multiple ways to contribute on biomarkers: protein signatures for diagnosis, correlating protein patterns for disease sub-classification, predicting disease progression and finding new targets for treatment strategies (Geyer et al. 2017). Tissues are generally only available from

1.3 Biomarkers

patients where surgery is a necessary treatment step. In most cases, they have to be obtained after the patients die, which is giving a bias to the group. In contrast, body fluids such as blood plasma or BALF are accessible by minimally invasive procedures (Geyer et al. 2019). In addition, this gives the advantage that the disease progression can be monitored longitudinal over time, allowing for personalized projection (Geyer et al. 2016). A classical MS workflow starts with the sample preparation where first the proteins have to be extracted from cells or enriched by precipitation from a liquid like BALF, before they are digested into peptides by trypsin and other enzymes. Next, the peptides are separated, based on their hydrophobicity on a high performance liquid chromatography (HPLC) to allow for a sequential acquisition of the ionized peptides. The fact that an ion behavior in the vacuum is strictly dependent on its mass over charge ratio (m/z) can be used for identification and quantification. As peptides are combinations of amino acids with distinct masses, it is possible to determine their sequences and the proteins they belonged to, reaching a depth of up to 10 000 proteins (Kulak et al. 2017).

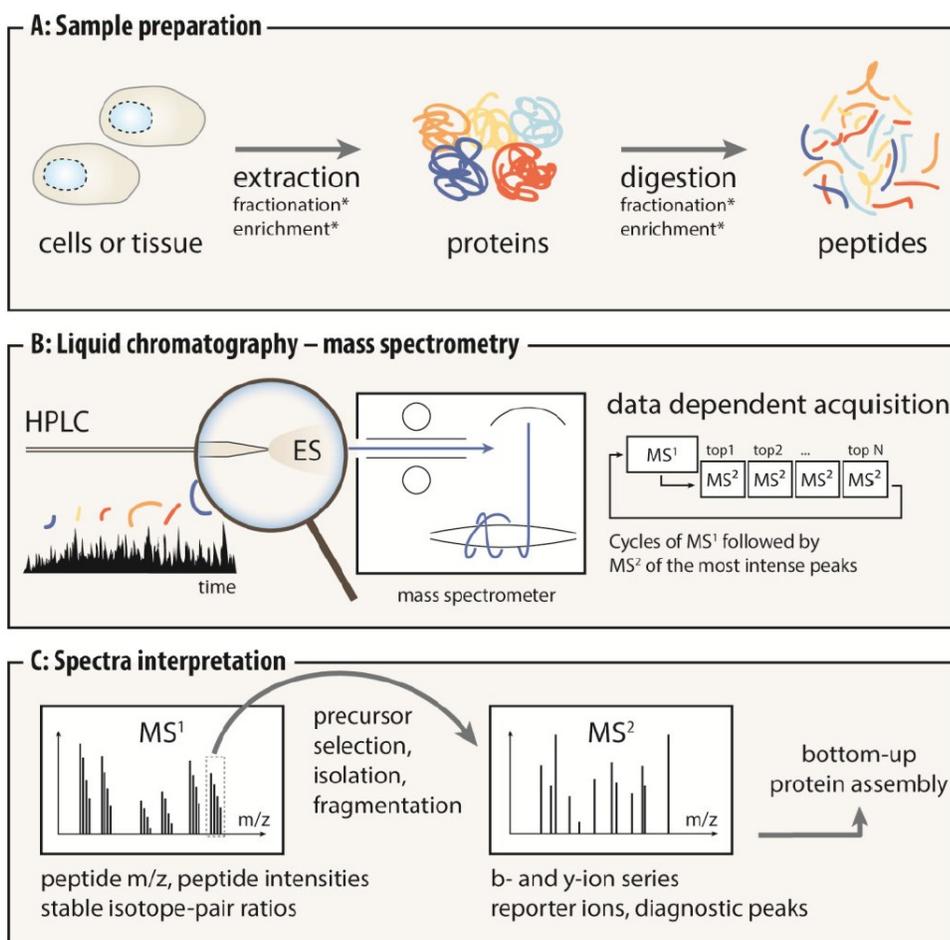


Figure 8. Shotgun proteomics workflow. (A) In the sample preparation, proteins are extracted and enzymatically digested to peptides. (B) Peptides are separated by a high-performance liquid chromatography (HPLC) system and ionized by electrospray for mass spectrometry analysis. (C) Bioinformatics spectra interpretation uses the information from the full and MS2 spectra for data searching. Figure taken from (Hein 2013)

1.3 Biomarkers

1.3.1 Single cell RNA sequencing

While MS-based proteomic analysis is powerful and sensitive on a protein level, it is still a “bulk” method that captures the composition of the whole piece of tissue that is lysed. Although some cell types within an organ can fulfill equal functions, the transcriptomic state of each individual cell may mirror solely the activity of only few active genes in that cell (Huang 2009). Therefore it is not possible to assign changes in protein expression to certain cell types or cell states, a disadvantage that may be hindering more specific treatment or interference options and cell type specific biomarker signatures in disease.

However, after the Human Protein Atlas, the science community is now aiming to generate the Human Cell Atlas as a cell-based reference of the human body, to solve these issues (Uhlen et al. 2015, Regev et al. 2017). This is possible with the rise of single cell genomics which is revolutionizing biology and medicine. The profiling of single cells with a high resolution allows to capture subtle gene expression changes that guide biological processes, define cell fate decisions, cell-to-cell variability and the distinct transcriptomic signals of individual cells (Vaquerizas et al. 2009). Over 40 different unique cell types with specific functions have been described in the lung (Franks et al. 2008)

One of the first successful single cell techniques to be developed was the Drop-seq method, which uses microfluidics to encapsulate single cells together with unique primers as barcodes on beads into aqueous droplets. The smart barcoding approach in Drop-seq allows the massively parallel, analysis of mRNA transcripts from thousands of individual cells simultaneously, while the cell of origin for each transcript can always be traced back (Macosko et al. 2015).

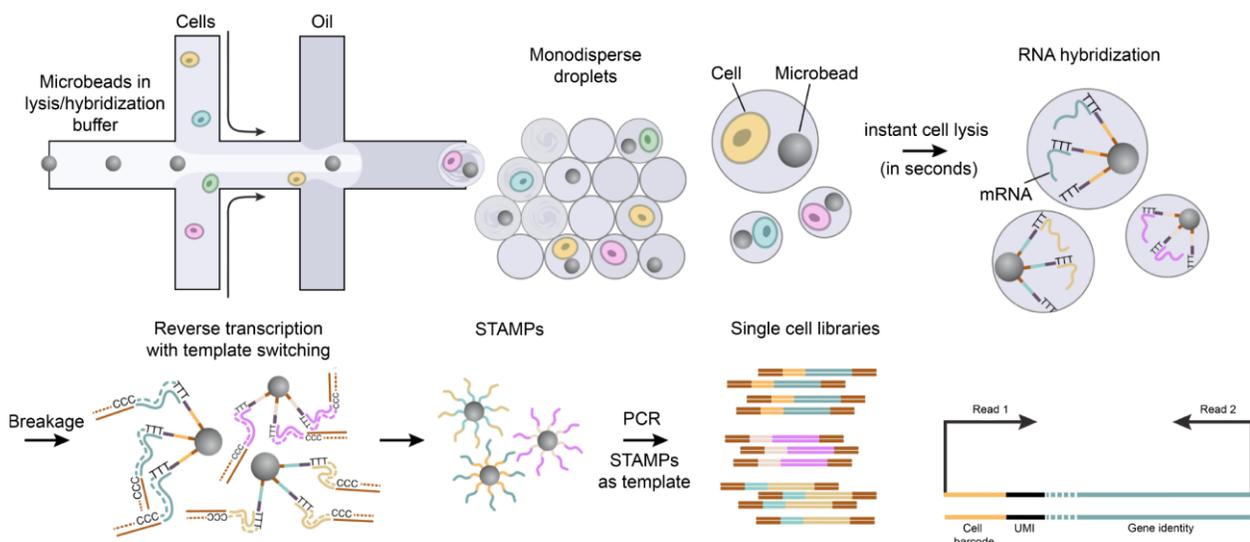


Figure 9. The principle of Drop-Seq. Cells are encapsulate together with microbeads that capture mRNA transcripts after cell lysis. Captured transcripts are transferred into cDNA, removed from the beads and amplified for sequencing. Libraries are sequences in a paired-end run to capture both the cell and gene identity. Single-cell Transcriptomes Attached to MicroParticles (STAMPS). Fiure taken from (Macosko et al. 2015).

2 Objectives

2.1 First chapter

Single cell genomics is conquering the world of biology and medicine by enabling a novel cell type specific viewpoint on organs and pathological changes. Technological advances now allow the profiling of genomes, transcriptomes and epigenomes at high resolution down to single cells. In the lung over 40 different unique cell types with specific functions have been described and even new ones are found (Franks et al. 2008). Also data from lung disease such as IPF is being published, enabling a novel cell type specific viewpoint on pathological changes in disease. While powerful, single cell mRNA sequencing relies on a piece of lung tissue, which involves invasive procedures such as lung biopsy, resulting in most of the samples stemming from deceased or transplanted patients. Additionally, mRNA expression does not necessarily reflect expression of proteins that are the final biomolecules performing critical processes within cells in the organism. New approaches are needed to dissect these heterogeneities on a molecular level to provide more personalized patient monitoring and care. MS-based proteomics on the other hand, enables analysis of body fluids, such as (BALF) or blood plasma, that be monitored longitudinal over time and allowing for personalized projection. The present study aimed to extend current knowledge of ILD on a transcriptional and protein level. The specific aims of this first chapter were as follows:

1. *Integrating an atlas of human fibrotic lung diseases on a single cell level*

Human single cell data had already been collected but was lacking analysis. With the publication of other human lung scRNAseq data (Travaglini et al. 2019) providing convincing marker genes, the first aim of this thesis was to annotate the Munich data into cell type specific clusters. Taking into account diseases specific cell states, an atlas of the human fibrotic lung should be generated. With scRNAseq data on human fibrosis already published by other groups, but often lacking relevance due to low patient numbers, the ultimate aim was to integrate the Munich atlas with two other published datasets (Reyfman et al. 2018, Habermann et al. 2019), to generate the largest human lung fibrosis atlas on a single cell level, providing the statistical power for subsequent analysis.

2.2 Second chapter

2. *Identifying protein biomarker signatures for disease diagnosis, progression, lung function and health status*

MS-based proteomics is becoming more and more sensitive, enabling reliable peptide identification from low amounts of input material (Aebersold et al. 2016). With body fluids of patients available in the form of BALF and blood plasma, the second aim was to use specialized MS-based workflows to identify protein signatures in three different cohorts from two centers. These biomarker panels should represent disease progression, reflect the patient on a lung function trajectory or in general be of help for diagnosis or accurate determination of the health status.

3. *Deconvolution of predictive biomarker signatures into specific cell types/states*

The last part of the thesis aimed at combining the results of the two before mentioned objectives. The highly significant cell type and state annotations as well as differential gene expression (DGE) between healthy and diseased, from the scRNAseq data, should be used to deconvolute the bulk protein data. In a first step, identified protein biomarker signatures should be assigned to changes of specific cell types in the lung. Second, the clinical data available to both the scRNAseq and MS cohorts could be used to predict clinical parameters such as lung function based on protein and gene expression profiles.

2.2 Second chapter

The lung is an architecturally complex organ comprising a heterogeneity of various epithelial and mesenchymal cell lineages. On the one hand, new discoveries have expanded and reinforced the conclusion that mesenchymal fibroblasts are crucial for forming and maintaining vascular networks, sensing damage, recruiting inflammatory cells and remodel the ECM. In fibrosis, the pathologic response, where the normal tissue is replaced with ECM, activated and α -smooth muscle expressing (*Acta2*/ α SMA) fibroblasts, the so called myofibroblasts are main source of ECM, which renders the lung into a hard block, hindering gas exchange. On the other hand, no more is known about the critical myofibroblast, where they come from and also the remaining mesenchymal cell populations are poorly defined, with markers discriminating them being urgently needed and cellular locations that could infer function and identify, hard to define. The bleomycin mouse model of fibrosis and regeneration is not a perfect comparison to the human situation but recapitulates the myofibroblast induction and ECM production very well, while on top giving the chance to investigate regeneration after injury as well; an aspect not occurring in the human disease.

2.2 Second chapter

Additionally, the bleomycin mouse model gives the unique opportunity to monitor development of fibrosis longitudinally over time, allowing for cell state projections. While single cell sequencing, where the RNA expression of each cell is captured, is the perfect tool for such analysis, the tissue is dissociated, which undesirably loses the spatial context of where each cell came from. Therefore combining scRNA-seq methods with spatial coordinate mapping methods, can be used to circumvent the problem of lost spatial cellular organization and used to define mesenchymal heterogeneity in the lung. The specific aims of this chapter were as follows:

1. *Fibroblast heterogeneity in the healthy adult wild-type mouse lung on single cell level*

The mesenchymal niche in the lung is surrounded by epithelial progenitors for which it provides signaling to modulate proliferations and differentiation, providing homeostasis. While the specialized epithelial cells are intensively studied, the heterogenous mesenchyme is poorly defined. Using scRNAseq, the first aim was to define and characterize the mesenchymal subtypes in the adult mouse lung on single cell level.

2. *In situ spatial mRNA mapping of fibroblast subtypes in the mouse lung*

Spatial location of cells in the tissue can infer function but cannot be studied if markers to discriminate cellular subsets are missing. Having annotated and defined the fibroblast subtypes by scRNAseq and distinct gene expression profiles, the second aim was to use these marker genes to perform “single cell resolution in situ hybridization on tissues” (SCRINSHOT) to spatially localize the subtypes in mouse lung and learn about their potential roles.

3. *Daily sampling of bleomycin injury to study activated fibroblast states*

The immense interest in fibroblast in the lung stems from the unsolved mystery behind the origin of the harmful myofibroblasts. With the healthy state of the fibroblast subtypes in the mouse extensively characterized, the third aim was to identify longitudinal changes of cellular states in the bleomycin model of lung fibrosis. Additionally, all subtypes in the injury state, especially myofibroblasts, should be characterized with a marker signature that goes beyond the so far known expression of *Acta2* to allow for future experiments on role and function.

3 Material and Methods

3.1 Methods human lung project

3.1.1 Lung samples

Human samples of the Munich cohorts (tissue, BAL fluid and plasma) were obtained from the BioArchive of the Comprehensive Pneumology Center Munich (CPC-M). All patients gave written informed consent and the study was approved by the local ethics committee of the Ludwig-Maximilians University of Munich, Germany (EK 333-10 and 382-10). Plasma samples of the Hannover cohort were obtained as part of a cooperation of the German Centers for Lung Research (DZL). Patients gave written informed consent to the DZL broad-consent form and the study was approved by the local ethics committee of the Medizinische Hochschule Hannover (2923-2015). Lung tissues used as healthy controls for single cell analysis were tumor free, uninvolved lung tissue freshly obtained during tumor resections performed at the lung specialist clinic “Asklepios Fachkliniken Munich-Gauting”. ILD lung tissue for single cell analysis was freshly obtained after lung transplantation at the University Hospital Munich. BAL fluid samples of the BAL fluid cohort and matched plasma samples were collected at the lung specialist clinic and included mainly first ILD evaluations. Plasma samples from an independent ILD cohort were obtained from the university hospital of the Ludwigs Maximilian University Munich from patients seen in the ILD outpatient clinic during routine visits or in the inpatient unit during evaluation for lung transplantation (Mayr et al. 2020).

The diagnosis of IPF was made in accordance with the current guidelines (Raghu et al. 2015). All ILD diagnosis was made according to international guidelines and established criteria. Non-ILD patients of the BAL fluid cohort included patients who underwent BAL due to evaluation of asthma, COPD, lung cancer, hemoptysis or chronic cough (Mayr et al. 2020).

For transport from the surgeon to the laboratory, lung tissue samples for single cell analysis were stored in ice-cold DMEM-F12 media and packed in thermo stable boxes. Tissue was processed with a maximum delay of 2 h after surgery. On delivery to the lab, tissue samples were assessed visually for qualification for the study (Mayr et al. 2020).

3.1 Methods human lung project

3.1.2 Lung tissue processing

Lung tissue was processed in a routinely way that has been established in the lab, was already used multiple times and is already described in various publications (Angelidis et al. 2019, Strunz et al. 2019, Vieira Braga et al. 2019, Mayr et al. 2020). Briefly, the tissue was manually homogenized into smaller pieces and washed. The tissue pieces were then transferred into enzyme mix consisting of dispase, collagenase, elastase, and DNase for mild enzymatic digestion. Released cells were washed, red blood cell lysis performed, and the remaining cells counted and assess for single-cell separation and viability. Around 300,000 cells were then loaded into the dropseq chip.

3.1.3 Single cell sequencing of human tissue using Dropseq

Drop-seq experiments were also routinely performed in the lab, as it has been described previously in multiple publications (Angelidis et al. 2019, Strunz et al. 2019, Vieira Braga et al. 2019, Mayr et al. 2020). Briefly, using a microfluidic device, single cells from the lung cell suspension were co-encapsulated in droplets with barcoded beads. Droplet emulsions were collected, the hybridized mRNA transcripts reverse transcribed and amplified by PCR. PCR products were pooled and purified twice and an estimated 1,000 cells was tagged to the same fragment size. Single-cell libraries were sequenced in a paired-end run on the Illumina HiSeq4000 (Mayr et al. 2020).

3.1.4 Processing of single cell data from Munich

For the single cell data of human patients from the Munich cohort, the Dropseq computational pipeline was used (version 2.0) as previously described, but with some specific alterations (Macosko et al. 2015, Angelidis et al. 2019, Strunz et al. 2019, Mayr et al. 2020). Briefly, STAR (version 2.5.2a) was used for mapping (Dobin et al. 2013). Reads were aligned to the hg19 reference genome, to be consistent with all samples that have been collected over several years (GSE63269). For filtering, barcodes with less than 200 detected genes were excluded. For further filtering, the top barcodes based on UMI count per cell were selected, guided by the number of estimated cells per sample. As a certain degree of ambient RNA bias was observed, SoupX was applied to lessen this effect (Young et al. 2020). The *pCut* parameter was set to 0.3 within each sample and count matrices were merged together. This merged expression table was then pre-processed further. A high proportion (> 10%) of transcript counts derived from mitochondria-

3.1 Methods human lung project

encoded genes may indicate low cell quality of dying or physically damaged cells. Therefore, cells that were above this threshold were removed from the analysis. Accordingly, cells with a high number of UMI counts may represent doublets, and therefore only cells with less than 4000 UMIs were kept in the experiment. Genes were only considered if they were expressed in at least 3 cells in the data set (Dobin et al. 2013) (Mayr et al. 2020).

3.1.5 Analysis of single cell data from Munich

The downstream analysis of the Munich single cell data was performed, in the same manner as previously reported from the lab, but with specific alterations to this project, using the Scanpy Package (Wolf et al. 2018), a python package for the exploration of single-cell RNA-seq data (Strunz et al. 2019, Mayr et al. 2020). Following the common procedure, the expression matrices were normalized using *scran*'s (Lun et al. 2016) normalization based on size factors which are calculated and used to scale the counts in each cell. Next log transformation was used via *scanpy*'s function *pp.log1p()*. Highly variable genes were selected as follows. First the function *pp.highly_variable_genes()* was executed for each sample separately, returning the top 4,000 variable genes per sample. Next, a gene was only considered as variable if it was labelled as such in at least 2 samples, resulting in a total of 15,096 genes which were further used for the principal component analysis. In an additional step to mitigate the effects of unwanted sources of cell-to-cell variation, the number of UMI counts, percentage of mitochondrial DNA and the calculated cell cycle score was regressed out, using the function *pp.regress_out()* (Mayr et al. 2020).

For visualizing the whole Munich data set, the UMAP was generated using 50 components as input for *scanpy*'s *tl.umap()* with number of *neighbors* set to 10 and *min_dist* parameter to 0.4. To better align the data of the different patients and to account for possible batch effects, the python package *bbknn()* (batch balanced k nearest neighbours) (Polanski et al. 2020) was used with the same number of components and neighbors. Louvain clustering was calculated with a resolution of 6. The whole lung parenchymal dataset was split into subsets for COL1A2+ mesenchymal cells, EPCAM+ epithelial cells, CLDN5+ endothelial cells and PTPRC+ leukocytes. New UMAP embeddings of these subsets were calculated until clear separation of cluster identities was achieved that allowed for identification of cell states by exploring the highest expressed markers per cluster explored via *tl.rank_genes_groups()* and manual assessment of known marker gene expression (Mayr et al. 2020).

3.1.6 Integration of Chicago, Nashville and Munich single cell data

3.1 Methods human lung project

To improve statistical power, to ensure generalization across cohorts and to achieve a more balanced ratio of diseased and healthy patients, our Munich single-cell RNA-seq data set was combined with the filtered count matrices from the Chicago cohort (Reyfman et al. 2018) and the Nashville cohort (Habermann et al. 2019) (Mayr et al. 2020).

Before combining these, the count matrices from Chicago and Nashville were processed separately. The normalization using *scrn* and the log transformation of the two external datasets was performed as described for the Munich cohort. The effect of cell cycle, the percentage of mitochondrial reads and the number of UMI counts was regressed out cohort-wise as well (Mayr et al. 2020).

For a first lighter batch correction we defined the list of variable genes in a way to decrease cohort specific effect as follows. For both the Nashville and the Chicago data we considered a gene as highly variable if it is labelled highly variable in at least three patients of the respective data set. Next, the preprocessed count matrices from the three data sets were merged and genes retained their highly variable status if they were highly variable in at least two of the three cohorts, resulting in 3,854 variable genes (Mayr et al. 2020).

The concatenated object was scaled with *scanpy's pp.scale()* function and the principal components were calculated using the defined variable genes. As a second batch correction the neighborhood graph was calculated using the *bbknn* package, defining the individual patients as batch key, number of neighbors within batch as 5 and number of components as 40. As described for the Munich cohort, the whole combined object was subsetted and new embeddings were calculated in order to identify cell states (Mayr et al. 2020).

3.1.7 Differential expression analysis

To identify genes associated with ILD status in a cell-type specific manner we applied the following procedure. The R statistical software was used for the analysis. Since the outcome of interest (ILD status) varies at the donor ($n = 61$) as opposed to the cell level ($n = 233,638$) the analysis was framed as a likelihood of detection problem across all donors. For each donor and cell-type combination the likelihood of detection for each gene was calculated as the average number of cells with more than one count. As the likelihood of detection represents a probability and is bounded between 0 and 1, values were square-root transformed. Next, multiple linear regression was used to model the probabilities of detection. The square-root transformed detection probability was used as the dependent variable and the ILD status, age or smoking status, as the explanatory variable accounting for the total number of UMI counts, total number of cells and study indicator as covariates. The resulting t and p-values for the coefficient describing the ILD status were used in downstream analysis (Mayr et al. 2020).

3.1 Methods human lung project

3.1.1 Cell type signature enrichment analysis

To infer cell type frequency changes from bulk transcriptomics or proteomics data, signature enrichment analysis was applied. Cell type signatures were defined as sets of genes with significant cell type specific expression. Next, the enrichment of each signature was statistically evaluated in a ranked list of fold changes or correlation coefficients using the Kolmogorov-Smirnov test. The signed p-value score represents the $-\log_{10}$ p-value of the Kolmogorov-Smirnov test signed by the effect size. Negative and positive values represent depletion and enrichment of the given signature in the ranked list, respectively (Mayr et al. 2020).

3.1.2 Random forest prediction

To integrate scRNAseq with BALF data we used a random forest as implemented in the R package *randomForest*. First, BALF expression data was quantile normalized and scaled. Next, only features with an absolute correlation coefficient greater than 0.2 with lung function and present in the scRNAseq data were used to train a random forest to predict lung function. Then, *in-silico* bulk scRNAseq was calculated by taking the mean expression count of each gene across all cells for all samples. Finally, the in silico bulk data was quantile normalized and scaled before feeding it into the trained model to predict lung function (Mayr et al. 2020).

3.1.3 Clinical parameters

For all patients included in the final analysis clinical information were collected at the time of BAL fluid procedure or when plasma was taken, respectively. Clinical parameters included demographics (age, gender, smoking status, pack years, smoking abstinence, lung function [forced vital capacity (FVC) (% pred.), FVC (l), FVC (post broncholysis), expiratory reserve volume (ERV), forced expiratory volume in 1 second (FEV1) (l), FEV1/FVC (%), inspiratory capacity (IC) (l), total lung capacity (TLC) (l), TLC from baseline, residual volume (RV), RV (%TLC), diffusing capacity of the lung for carbon monoxide (DLCO) (VA) (mmol/min/kPa/l), DLCO (SB) (% pred.), DLCO (SB) (Hb corrected, % pred.), DLCO (SB) (mmol/min/kPa), mean expiratory flow (MEF) 25, MEF50, intrathoracic gas volumen (ITGV)], laboratory values [cholinesterase, alkaline phosphatase, C-reactive protein, alanine-aminotransferase (ALT), aspartate-aminotransferase (AST), gamma-glutamyltransferase (GGT), LDH] (Mayr et al. 2020).

3.1 Methods human lung project

3.1.4 BAL procedure

BALF was collected from 141 patients undergoing bronchoscopy from January 2013 until March 2016 at the Lungenfachklinik Gauting in Munich, Germany. Most of the patients underwent bronchoscopy due to ILD evaluation. BAL was performed with standard technique. In brief, 100 to 200 ml of sterile saline (0.9% NaCl) was instilled into the right middle lobe or the lingula in 20-ml injections which were each immediately aspirated. Cells of the BAL were analyzed by cytopsin analysis. The remaining cell-free BAL fluid was immediately stored at -80°C and transferred to the BioArchive of the CPC-M. For mass spectrometry, only the cell-free BAL fluids were analyzed. Of the 141 patients only 124 passed quality control and were included in the analysis (95 ILD and 29 non-ILD) (Mayr et al. 2020).

3.1.5 Mass spectrometry

3.1.5.1 BAL fluid proteomes

The BAL fluid depleted from cells was subjected to mass spectrometry analysis, which was performed in collaboration with the group Mathias Mann for the use of the MS machines. The protocol is established and has been routinely performed by the PhD student before, as reported in several publications (Schiller et al. 2017, Kammerl et al. 2019, Mayr et al. 2020, Petzold et al. 2020). Proteins in BALF were precipitated from 300µl BAL fluid using 80% ice cold acetone, followed by reduction and alkylation of proteins and overnight digestion into peptides using Trypsin and LysC proteases (1:100) as previously described (Schiller et al. 2015). Briefly, resulting peptides were purified using stage-tips. Approximately 2 µg of peptides were separated in four hour gradients on columns using, reverse-phase chromatography coupled to a Q-Exactive Mass Spectrometer. MS data were acquired with a shotgun proteomics method (top10 method). The resulting mass spectra and raw files were processed using the MaxQuant software (Cox et al. 2008), which enabled label free protein quantification across different samples (Tyanova et al. 2016) (Mayr et al. 2020).

3.1.5.2 Plasma proteomes

Plasma samples were prepared with the Plasma Proteome Profiling Pipeline, automated on an Agilent Bravo liquid handling platform and performed in collaboration with the group of Mathias

3.2 Methods for the mouse lung mesenchyme project

Mann (Geyer et al. 2016, Geyer et al. 2017, Geyer et al. 2019, Mayr et al. 2020). Briefly, plasma samples were diluted and mixed with PreOmics lysis buffer for reduction, alkylation and protein denaturation, followed by digestion with Trypsin and LysC. Resulting peptides were loaded on StageTip plugs, followed by consecutive purification steps. Plasma peptides were measured using LC-MS instrumentation consisting of an EASY-nLC 1200 ultra-high pressure system, online coupled to a Q Exactive HF Orbitrap. For each LC-MS/MS analysis about 0.5 µg peptides were injected. MS data was acquired with data independent acquisition. MS raw files were analyzed by Spectronaut software (version 12.0.20491.10.21239 (Bruderer et al. 2015, Geyer et al. 2016)) and were searched against the human Uniprot FASTA database (Mayr et al. 2020).

To compare BAL fluid and plasma proteomes, patients were grouped into DLCO low (DLCO < 50% pred.) and DLCO high (DLCO ≥ 50% pred.), respectively (Mayr et al. 2020).

3.1.6 Mass spectrometry bioinformatics and statistical analyses

Mass spectrometry raw files were processed using the MaxQuant software (Cox et al. 2008, Cox et al. 2011) (*version 1.5.3.34*). As previously described (Schiller et al. 2015, Schiller et al. 2017, Mayr et al. 2020), peak lists were searched against the human Uniprot FASTA database (*November 2016*), and a common contaminants database (247 entries) by the Andromeda search engine (Cox et al. 2011). Protein expression was corrected for age in the following manner. Age was regressed out from the protein expression data using the R function *aov()*. The residuals from this model were used in subsequent analysis. All other statistical and bioinformatics operations (such as normalization, data integration, annotation enrichment analysis, correlation analysis, hierarchical clustering, principal component analysis, Pearson correlation analysis, t-test statistics, ANOVA tests, or Fisher's exact test and multiple-hypothesis testing corrections), were run with the Perseus software package (version 1.5.3.0 and 1.6.1.1.) (Tyanova et al. 2016) (Mayr et al. 2020).

3.2 Methods for the mouse lung mesenchyme project

3.2.1 Animal handling and bleomycin administration

3.2 Methods for the mouse lung mesenchyme project

The mice used for experiments in this study were old pathogen-free female C57BL/6J purchased from Charles River, Germany. They were maintained at the appropriate biosafety level at constant temperature and humidity, with a 12 hour light cycle and allowed food and water *ad libitum* at the animal husbandry of the Helmholtz Zentrum München, Germany. All experimental procedures were conducted in accordance with the regulations of the Bavarian government and the guidelines of the European Union and are registered and approved under the registry ROB-55.2-2532.Vet_02-16-208.

For instillation, mice with 10 to 14 weeks of age were first sedated by intraperitoneal injection of medetomidin 500 µg/kg body weight, midazolam 5 mg/kg body weight, and fentanyl 50 µg/kg body weight in sterile sodium chloride solution at a maximal volume of 100 µl via a 27 gauge cannula. Oropharyngeal administration was achieved by pulling out the tongue and closing the nostrils to induce deep breathing through the oral cavity before application to the oropharynx. Mice were antagonized, directly after treatment by subcutaneous injection of atipamezol 2.5 mg/kg body weight, flumazenil 500 µg/kg bodyweight, and naloxon at 1.2 mg/kg body weight, dissolved in sterile sodium chloride solution at a maximal volume of 100 µl via a 27 gauge cannula (Strunz et al. 2019).

Control mice were administered with 80 µl of phosphate-buffered saline (PBS), while lung injury was induced by single-dose administration of bleomycin hydrochloride (Sigma Aldrich, Germany), dissolved in sterile PBS and given at 2U/kg bodyweight. Mice were randomly sacrificed at designated time points (21, 28, 35, 56) or even daily (days 1-14) after instillation, for a high resolution time-resolved investigation of bleomycin induced changes after lung injury. Treated animals were continuously under strict observation and mice were sacrificed in case of signs of phenotypic changes, such as abnormal behavior, weight loss or fur appearance caused by the administration (Strunz et al. 2019).

3.2.2 Generation of single cell suspension from whole lung

Mice were sacrificed by cervical dislocation of the spinal cord from the brain and dissected as fast as possible to prevent clogging of blood. The thorax was opened and heart and lung were perfused with up to 25 ml of PBS through the right ventricle of the heart. Mice were intubated through a cut in the trachea to fit a needle which allowed for filling of whole lung with either, 4% paraformaldehyde (PFA) for formalin-fixed paraffin embedded (FFPE) tissue blocks or enzyme mix for experiments, requiring a single cell suspension. Lungs, filled with enzyme mix, to allow for distribution of the enzyme down to the alveoli, were taken out with the heart *en block* and the lung tissue was minced to pieces around 1 mm³. For mild enzymatic digestion, the tissue was transferred to 50 ml falcons in up to 4 ml of enzymatic mix, containing dispase (50 caseinolytic U/ml), collagenase (2 mg/ml), elastase (1 mg/ml), and DNase (30 µg/ml) (Angelidis et al. 2019,

3.2 Methods for the mouse lung mesenchyme project

Strunz et al. 2019) for 15 to 25 min at 37°C in a thermomixer, shaking at 700 rpm. Enzyme activity was stopped by adding 10 ml of PBS supplemented with 10% fetal calf serum (FCS). Single cells were harvested by straining digested tissue through a 40 micron mesh and centrifugation at 300 x g for 5 min at 4°C. In case of a rather bloody suspension, red blood cell (RBC) lysis was performed by incubating the cells for 30 sec in RBC lysis buffer (Thermo Fisher) at room temperature (RT), which was stopped again with 10 ml of PBS supplemented with 10% FCS. Cell pellets were resuspended in PBS supplement with 1% FCS, singularized by thorough mixing through pipetting, and ultimately counted in a counting chamber (Strunz et al. 2019).

3.2.3 Magnetic-activated cell sorting (MACS)

Magnetic activated cell sorting (MACS) provides an easy and fast way to enrich or deplete specific cells out of a complex tissue sample. Like the more famous brother fluorescent activated cell sorting (FACS), it relies on specific antibody binding to cell-surface receptors. Antibodies can be directly coupled to magnetic microbeads, or contain a fluorescent dye that in a secondary incubation step then interacts with microbeads against the color. Once coupled to a microbead, cells are retained in a column while non-bound cells are washed out. Both positive and negative selection of distinct cell populations is possible: either the cells of interest are first retained and later eluted from the column, or undesired cells are retained in the column and the washed out cells are used for further experiments.

For the presented work, to enrich for the mesenchymal cell population, for which no common marker is known, negative selection was used. Single cell suspensions were first stained with antibodies, coupled to a fluorescent dye, against undesired cell populations, for 30 min at 4°C in the dark: CD326+ (Biolegend, 118212), CD31 (Invitrogen, 17-0311-82), CD45+ (Biolegend, 103112), Lyve1 (Invitrogen, 50-0443-82), Ter119 (Biolegend, 116218), and CD326+ cells (Biolegend, 118212).

3.2.4 Processing of WT and bleomycin mouse single cell data

For both the single cell data of the WT mice and the bleomycin time course data, the Dropseq computational pipeline was used (version 2.0) as previously described (Macosko et al. 2015). Briefly, STAR (version 2.5.2a) was used for mapping (Dobin et al. 2013). Reads were aligned to the mm10 reference genome (GSE63269). For barcode filtering, we excluded barcodes with less than 200 detected genes. For further filtering we kept the top barcodes based on UMI count per cell, guided by the number of estimated cells per sample. As we observed a certain degree of ambient RNA bias, we applied SoupX (Young et al. 2020) to lessen this effect. The pCut parameter was set to 0.3 within each sample before merging the count matrices together. The merged

3.2 Methods for the mouse lung mesenchyme project

expression table was then pre-processed further. 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 4000 UMIs were used in downstream analysis. Genes were only considered if they were expressed in at least 6 mice in the bleomycin time course data set (Dobin et al. 2013).

3.2.5 Analysis of mouse the WT and bleomycin single cell data

The downstream analysis of the mouse single cell data was performed using the Scanpy Package (Wolf et al. 2018). The expression matrices were normalized using *scrn*'s (Lun et al. 2016) normalization based on size factors which are calculated and used to scale the counts in each cell. Next log transformation was used via scanpy's function *pp.log1p()*. Highly variable genes were selected with the function *pp.highly_variable_genes(.)* A in the bleomycin timecourse gene was only considered as variable if it was labelled as such in at least 6 samples (which equals three consecutive time points with two mice each), resulting in a total of 10,475 genes which were further used for the principal component analysis. The majority of the cells were mesenchymal cells, although epithelial and immune cells were also captured. Further data filtering was achieved by clustering so that clusters expressing non-mesenchymal markers *Col1a2* and *Vim* could be excluded from the data set. The filtered object was scaled and the neighborhood graph constructed. With the batch-balanced k nearest neighbor tool (BBKNN) (Polanski et al. 2020), the different PCR cycles used in the experiment were accounted for by setting the *neighbors_within_batch=5* and *n_pcs=40*. Two dimensional visualization and clustering was carried out with the Scanpy functions *tl.louvain()* at resolution of 3 and *tl.umap()*. After manual curation of the markers, the clusters were combined, according to cell types, leading to 6 final meta cell types.

3.2.5.1 Cell-cell communication analysis

For the analysis of cell-cell communication networks, a list of annotated receptor-ligand pairs was downloaded (Ramilowski et al. 2015). The receptor-ligand information was integrated with the bleomycin timecourse cell type marker genes, followed by the generation of cell-cell communication networks. This is facilitated by the creation of edges between two cell types if both share a receptor-ligand pair as marker genes. The cell type interaction plots were generated based on the number of receptor ligand pairs between the cell types. Receptor-ligand expression plots were based on the expression of receptor and ligand genes, respectively.

3.2.5.1 MatchScore comparison

3.2 Methods for the mouse lung mesenchyme project

To better compare cell types across datasets in a quantitative way, *matchScore*, a Jaccard index based scoring system was used, to quantify clustering and marker accuracy in a combined score (Mereu et al. 2018). Thereby cluster identities were integrated across the human and mouse datasets by filtering the respective cell type marker tables with a cut-off for the fold-change at 2 and for the adj-p-pval at 0.0001.

3.2.6 SCRINSHOT staining

Single cell resolution in situ hybridization on tissues (SCRINSHOT), for highly multiplexed in situ mapping of gene expression, was performed as recently published (Sountoulidis et al. 2020). Multiplexed in situ hybridization was performed on 10µm fresh frozen OCT embedded lung slices. Cells in frozen sections were permeabilized with 0.1 M HCl in DEPC water at room temperature for 3 minutes and dehydrated in an ethanol series. An incubation chamber was mounted and nonspecific mRNA binding was blocked by adding oligo-dT fragments. Three to five gene-specific hybridization probes (padlock probes) were added for 5min at 55°C and for 90 min at 45°C. Ligation of the circular padlock probes was ensured by incubation with SplitR enzyme at room temperature overnight. Rolling circle amplification (RCA) of the padlock circle at 30°C overnight followed by a degradation step for non-specific products with uracil-N-glycosylase (UNG) for 45 minutes at 37°C, provided an amplified number of binding sites for labeled detection probes. One specific labeled probe per padlock probe, with probes for the same gene sharing one color, were hybridized to their binding sites for 60 min at room temperature in the dark. After washing, the tissue was hydrated, the incubation chamber removed and the slide was mounted with a coverslip for microscopy. Per hybridization cycle, three genes and DAPI could be visualized. After successful imaging, the coverslip and mounting media were removed, the tissue dehydrated again, and the procedure repeated from the UNG step onwards with different label probes.

3.2.7 SCRINSHOT data analysis

3.2.7.1 Processing of images

Microscope images of each hybridization cycle and region of the mouse lobe were processed in ZEN image processing software. An orthogonal projection was made to compress the image z-stacks. Image tiles were stitched with DAPI as a reference, enhanced and subsetting to obtain separate image files for each lobe region and hybridization. Hybridization images for a particular region were aligned using the images of DAPI stained nuclei and cropped to size in FIJI. Nuclei were individually drawn around using the freehand drawing tool in FIJI. An image with the nuclei shapes filled in was used for further processing. Using MATLAB, images were split into four tiles for the following steps in Cell Profiler. Nuclei shapes were enlarged by 2µm and fluorescent signal

3.2 Methods for the mouse lung mesenchyme project

(from labelled/stained mRNA (punctae)) was detected from each channel per cell shape and converted into individual pixels. The number of punctae per gene/channel per cell was measured in FIJI and analysed as a single-cell matrix in Perseus. (The number of punctae per gene/channel per cell was measured in FIJI and converted to a data matrix. The matrix was analysed analogous to single-cell mRNA sequencing data and was used for further visualization and quantification of cell types.)

3.2.7.2 ROI and cell filtering

Cells with lower than one visualized mRNA for one gene counted were filtered out. Cells with many punctae for many genes (minimum of 4 genes with greater than 5 punctae) were also filtered out due to the high likelihood of these being due to debris/dirt. The matrix was exported to Excel for further processing. The signal counts were totaled up for each cell type by using the 2 designated marker gene. If a cell was had 3 or more punctae in total, with at least 1 or more for each gene, it was counted as positive for that cell type. If a cell was positive for both cell types, the signal/cell was checked manually. ROIs positive for more than one cell type tended to be to debris/dirt. Positive ROIs with no signal were also manually removed. Remaining positive ROIs were transferred to compressed zip folders so they could be shown in FIJI. ROIs/cells were then colored based on cell type to show the locations of the cell type. For a more accurate picture, and for a smaller area such as part of the airway or vessel, the ROI of cells clearly positive for mRNA signals of a certain cell type were noted. The different cell types were then colored in FIJI.

3.2.7.3 Counting and quantification

Total numbers of specific cell types visualized by SCRINSHOT were manually counted by searching for SCRINSHOT signals using the image processing software ZEN. Six different imaged regions of the WT mouse lobe were used for quantification: Area 1 (destroyed region excluded), Area 4, Area 5, Area 7, Area 8, Area 12. AT2 cells were counted based on high Sftpc expression localized around a DAPI stained nucleus. Pericytes, Hhip cells, lipofibroblasts, matrix fibroblasts, and smooth muscle cells were counted if three or more fluorescent mRNA dots from two of the following respective 'marker genes' were localized at or closely around a DAPI stained nucleus; Gucy1a3 and Postn for pericytes, Hhip and Aspn for Hhip cells, Tcf21 and Npnt for lipofibroblasts, Serpinf1 and Clec3b for matrix fibroblasts, Acta2 and Myh11 for smooth muscle cells (with at least 1 dot visible each from both marker genes). Cell counts were normalized using the mean total area size of all regions measured, with blank spaces of airways or vessels removed.

The frequency of AT2 cells observed in close proximity to a mesenchymal subtype cells were also manually counted using the image processing software ZEN. An AT2 cell was designated as interacting with a cell type, if observed in very close proximity to one of the mesenchymal cell types. Spatial locations of cells were classified as being near an airway or vessel if positioned approximately 45µm away from the tissue structure. If a cell within the tissue was located further

3.2 Methods for the mouse lung mesenchyme project

than 45 μ m from a vessel or airway, it was classified as being in the alveolar space. Statistical analysis was carried out using GraphPad Prism 5.

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

4 Results

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

The work presented in this chapter has been drafted into a manuscript that was submitted to a scientific journal. By the time of submission of this thesis, the paper has not yet been peer-reviewed. However, a preprint version of the manuscript has already been published on *medRxiv*, a free online archive and distribution service for complete but unpublished manuscripts (preprints) in the medical, clinical, and related health sciences. MedRxiv is a spin-off of the more known *bioRxiv*, for life science in general, both being operated by Cold Spring Harbor Laboratory, as a not-for-profit research and educational institution. The manuscript file is accessible through the *medRxiv* website (<https://www.medrxiv.org>) or directly via the digital object identifier doi: (<https://doi.org/10.1101/2020.01.21.20018358>). In this thesis, the citation for this preprint will be: (Mayr et al. 2020).

The PhD student, as first author of this manuscript, was involved in all parts of this study, performed most of the practical work, executed programming code, drafted all the figure panels and was part in conceiving the manuscript. Credit goes to Dr. Simon Lukas, for providing the R code for deconvolution of protein data and DGE of the integrated data-set, Dr. Gabriela Leuschner for organization of the clinical and patient data, Meshal Ansari for support in single cell analysis, Dr. Philipp Geyer for assistance in MS measurements of plasma samples and Dr. Maximilian Strunz, Ilias Angelidis and Pawandeep Singh for collecting human single cell samples.

4.1.1 Introduction

Fibrotic diseases are one of the major health burdens of our time (Cox et al. 2011). In the lung, pulmonary fibrosis is represented by the very heterogeneous group of ILD, ultimately leading to respiratory failure due to destruction of the lung parenchyma, as introduced in Chapter 1.2. The heterogeneity of this group is given by several risk factors (Baumgartner et al. 1997, Fischer et al. 2012, Selman et al. 2016, Schiller et al. 2017, Allen et al. 2019, Sheng et al. 2019) but most importantly also by varying disease progression, diagnosis and the occurrence of acute exacerbations in the patients. New approaches are needed to dissect these heterogeneities on a molecular level to provide more personalized patient monitoring and care. With the surge of innovation in the field of single cell genomics, a novel cell type specific viewpoint on pathological

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

changes in disease becomes feasible (Mayr et al. 2020). Projects like the Human Cell Atlas focus at building comprehensive maps of all cells in the human body to understand fundamental biological processes with the aim of to improve monitoring, diagnosis and treatment of diseases (Schiller et al. 2019). Part of this initiative is the human Lung Cell Atlas that has been successfully launched with a first draft of cellular compositions of the human and mouse lung, as well as recent single cell studies profiling cellular and molecular changes associated with pulmonary fibrosis (Han et al. 2018, Reyfman et al. 2018, Adams et al. 2019, Angelidis et al. 2019, Habermann et al. 2019, Morse et al. 2019, Travaglini et al. 2019, Vieira Braga et al. 2019).

While powerful, single cell mRNA sequencing relies on a piece of lung tissue, which involves invasive procedures such as lung biopsy, resulting in most of the samples stemming of deceased or transplanted patients. Additionally, mRNA expression does not necessarily reflect expression of proteins that are the final biomolecules performing critical processes within cells in the organism (Angelidis et al. 2019). Proteomic analysis of body fluids, such as BALF or blood plasma, can be monitored longitudinal over time allowing for personalized projection of disease trajectories among the highly variable ILD patients (Neighbors et al. 2018, Maher et al. 2019). First proteomic changes and in end stage ILD patient lung tissue and BALF have already been analyzed and protein signatures have been identified (Foster et al. 2015, Prasse et al. 2019).

However, it is unclear which cellular and molecular processes in the lung correspond to these biomarker signatures, since they only represent a tissue or fluid average, which does not resolve cellular composition and disease specific cell states. But combining proteomics and single cell sequencing together with novel computational methods, can address this problem: signatures found in bronchoalveolar lavage and plasma, both of which are accessible for longitudinal monitoring of patients, can be used to predict pathological cell state changes in the lung (Mayr et al. 2020).

The following sections and the graphical abstract (Fig. 10) will summarize the results of the herein conducted investigations. Using single cell data from ILD patients from Munich, cell state changes in human fibrosis were derived at cellular resolution. To assess reproducibility across patient cohorts, and to establish a robust differential gene expression analysis for all major cell types in the lung, by increasing the number of patients, the data was integrated with two independent single cell datasets. Using mass spectrometry driven proteomics, bronchoalveolar lavage and plasma proteome compositions in patients from several independent large-scale ILD cohorts were quantified. A bioinformatics analysis strategy enabled the integration of the cell state descriptions on single cell level with these lavage and plasma proteomes, transcriptomic tissue bulk measurements and associated clinical meta-data. The analysis dissected lung fibrosis at the single cell level and the statistical power allowed the definition of robust differential gene expression profiles across multiple studies for ILD. This approach revealed that fluid proteome

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

signatures were predictive of specific cell state changes in the lung. Additionally, by using machine learning, several protein biomarker signatures that were associated with diagnosis, lung function, and smoking and injury status were discovered (Mayr et al. 2020).

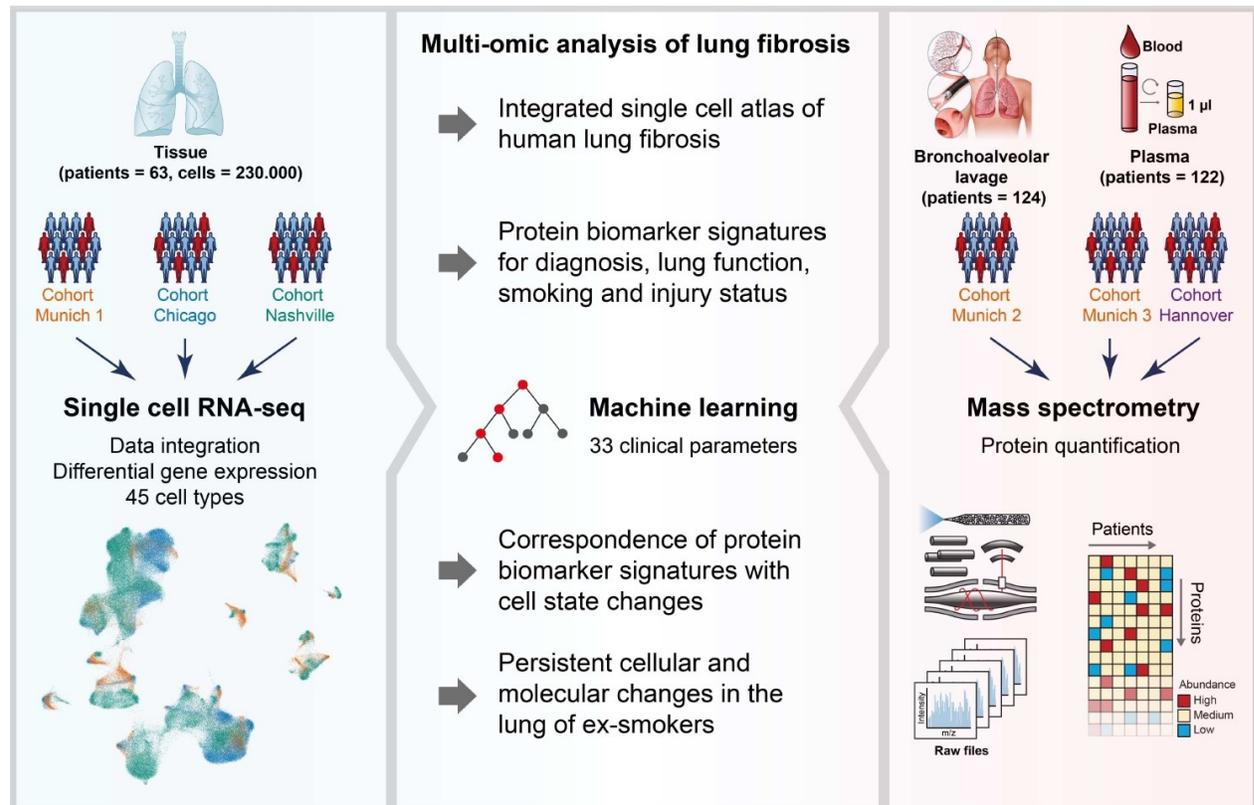


Figure 10. Graphical abstract for the project “Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids”. Figure and legend taken from (Mayr et al. 2020).

4.1.1 An integrated single cell atlas of human lung fibrosis

Transcriptional changes in lung fibrosis were studied using the Drop-seq platform for single cell sequencing (Macosko et al. 2015). Whole lung parenchyma was obtained from 11 non-fibrotic donors and three end stage lung fibrosis patients from the CPC-M bioArchive in Munich. Approx. 5000 cells per sample were sequenced, which finally, after filtering and quality control, resulted in a data set containing 41,888 cells (Fig. 11a). Single cell transcriptomes were subjected to data dimensionality reduction using the Uniform Manifold Approximation and Projection (UMAP) method (Becht et al. 2018). UMAPs were displayed with a color code for the health condition, either control or lung fibrosis (Fig.12a), or for individual patients (Fig. 12b) and assessment of overlaying of these groups proved both good accordance.

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

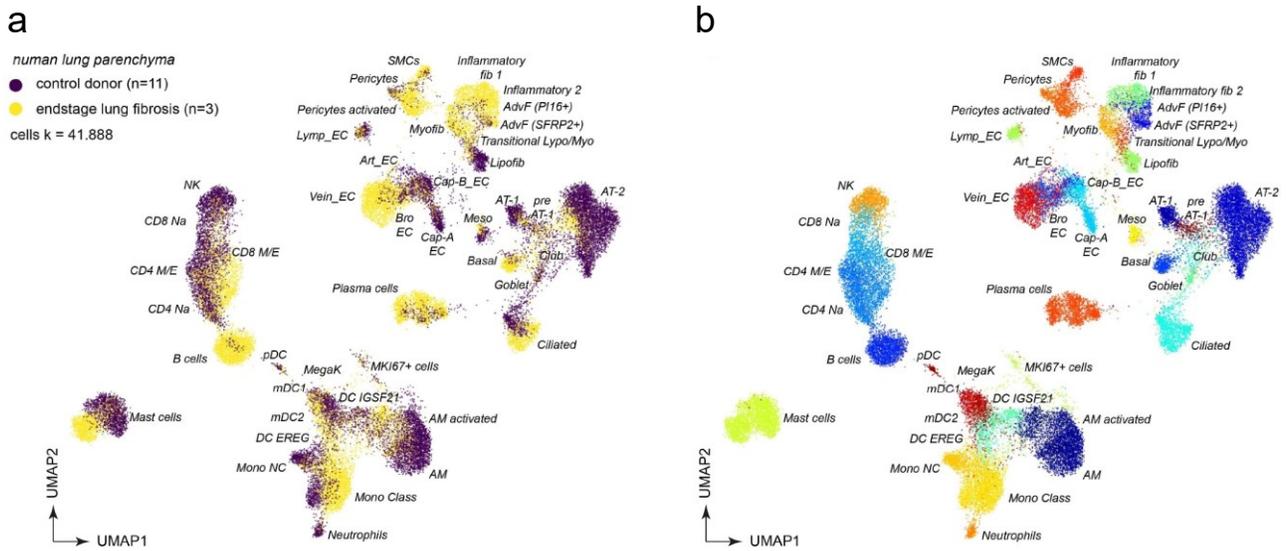


Figure 11. Single cell analysis of human lung parenchyma in health and disease reveals 45 distinct cell type identities and their marker genes. (a, b) Dimension reduced single cell transcriptomic data is visualized through Uniform Manifold Approximation and Projection (UMAP). The color code illustrates the disease status (a) and cell type identity (b). Figure and legend taken from (Mayr et al. 2020).

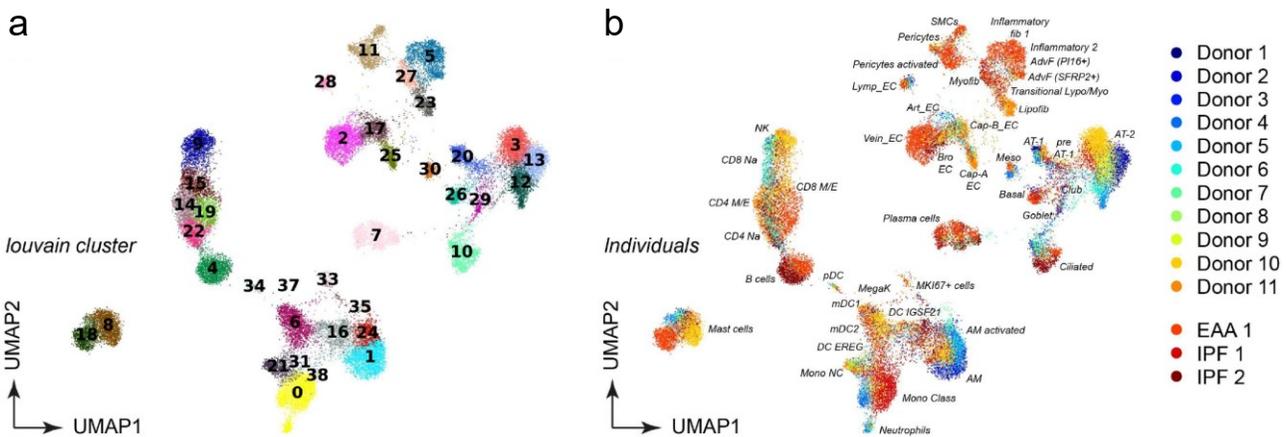


Figure 12. Clustering analysis and cell type annotation reveals 45 distinct cell type identities in human lung parenchyma. (a) UMAP embedding colored by Louvain clusters demonstrates separation of cells into major lineages. (b) UMAP embedding displays identified cell types, colored by individual patients. Figure and legend taken from (Mayr et al. 2020).

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

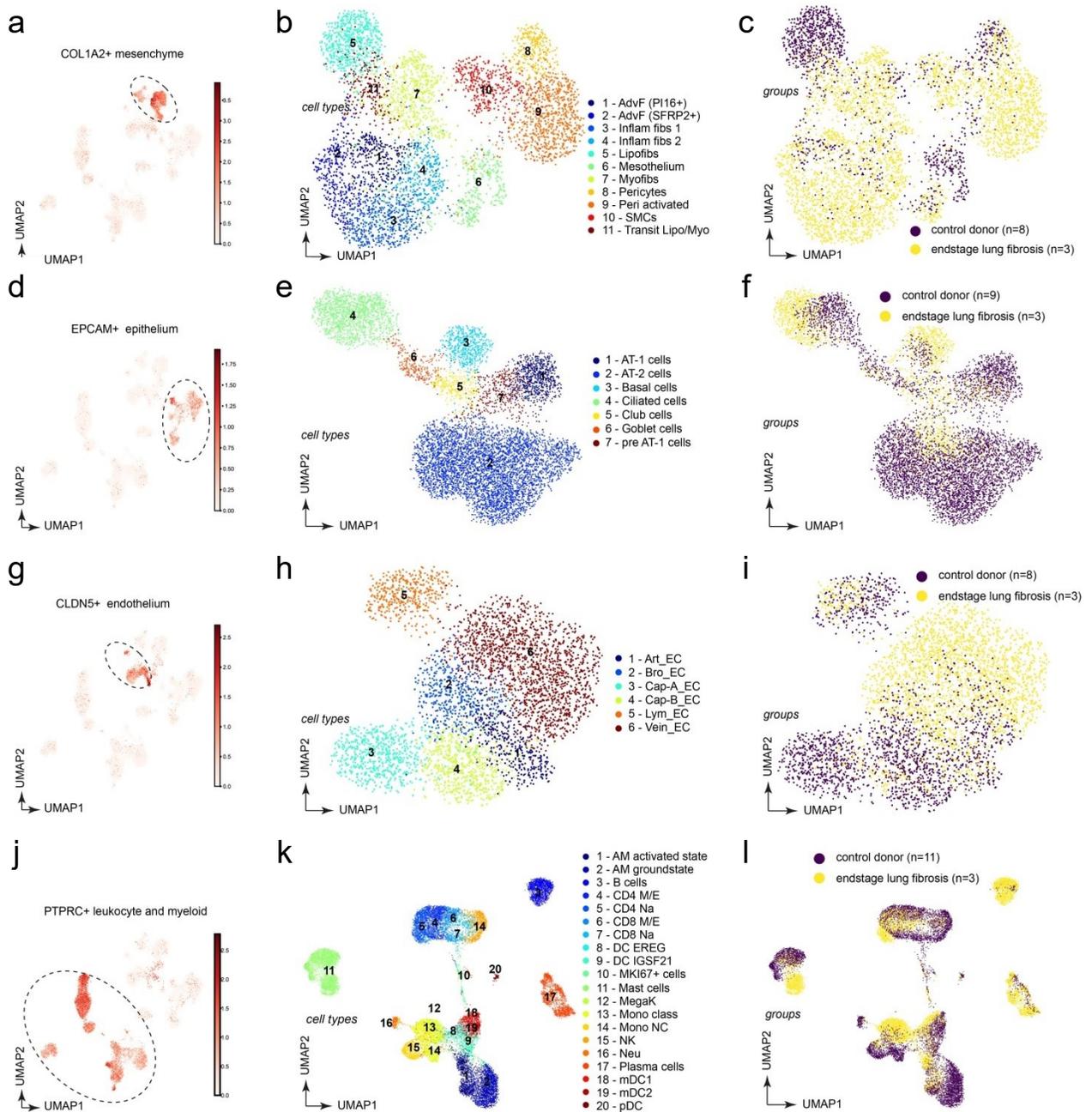


Figure 13. Clustering analysis and cell type annotation reveals 45 distinct cell type identities in human lung parenchyma. (a, d, g, j) The whole lung parenchymal dataset was split into subsets for (a) COL1A2+ mesenchymal cells, EPCAM+ epithelial cells (d), CLDN5+ endothelial cells (g) and CD45+ (gene name PTPRC) leukocytes (j). (b, e, h, k) New UMAP embeddings of the subsets demonstrate separation of cluster identities that allows for identification of cell states. (c, f, i, l) Cells colored in disease groups show origin of identified cell states. Figure and legend taken from (Mayr et al. 2020).

Unsupervised clustering of the data revealed 37 cell clusters (Fig. 12a). From this whole dataset, subsets based on the major lung cell lineages were generated and newly embedded in UMAPs,

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

to increase the resolution for individual cell types (Fig. 13). These subsets were COL1A2+ stromal cells (Fig 13a-c), EPCAM+ epithelial cells (Fig 13e-f), CLDN5+ endothelial cells (Fig 13g-i), and CD45+ leukocytes and myeloid cells (Fig 13j-l). The subsets were again quality controlled and samples, contributing less than 20 cells were excluded from the analysis. These major lung cell lineage subsets were used to derive cluster identities by manual annotation of these clusters with known canonical marker genes and signatures from published single cell RNA sequencing studies (Adams et al. 2019, Habermann et al. 2019, Travaglini et al. 2019). And indeed, with a total of 45, more lung cell type identities could be revealed from the sub-clustering (Fig. 11b). Notably, already by observation of the sub clusters colored-coded for the health-state (Fig. 13c, f, i, l) it became obvious, that some were mainly present in fibrosis and thus termed as “activated- “ cell types. All lung cell type identities were characterized by unique marker gene expression profiles, providing a useful tool for future cell type based analysis (Fig. 14).

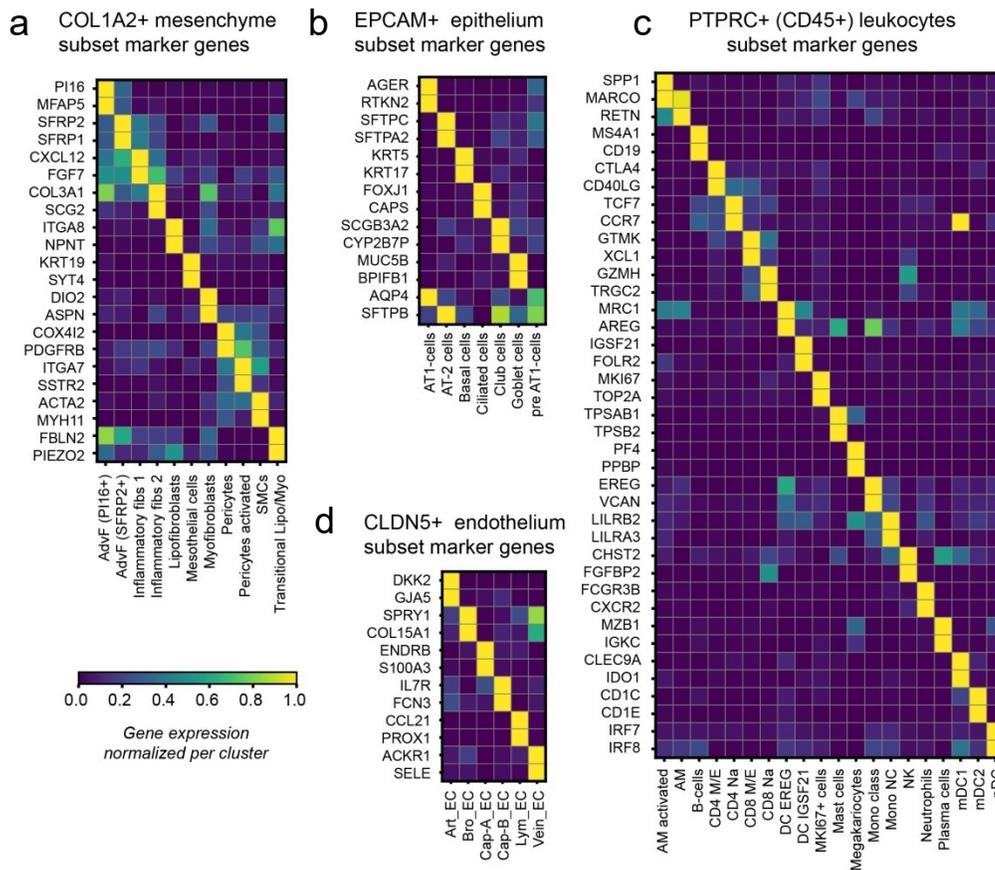


Figure 14. Distinct cell type identities are defined by specific marker genes. (a-d) The heatmaps show the relative gene expression levels for the indicated marker genes for the indicated stromal (a), epithelial (b), endothelial (c), and leukocyte (d) cell types. Figure and legend taken from (Mavr et al. 2020).

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

To increase statistical power and validate our results, we integrated our dataset with two large publicly available single cell RNA-seq datasets. The Chicago cohort from Reyfman and colleagues contains 9 ILD patients and 8 controls; the Nashville cohort from Habermann and colleagues provides 20 ILD patients and 10 controls. By bioinformatics integration of the three cohorts, Munich, Chicago and Nashville, we generated a data manifold that, after filtering and quality control, represents gene expression profiles of 233.638 single cells from 63 human individuals (ILD n=32, controls n=31) (Fig 15).

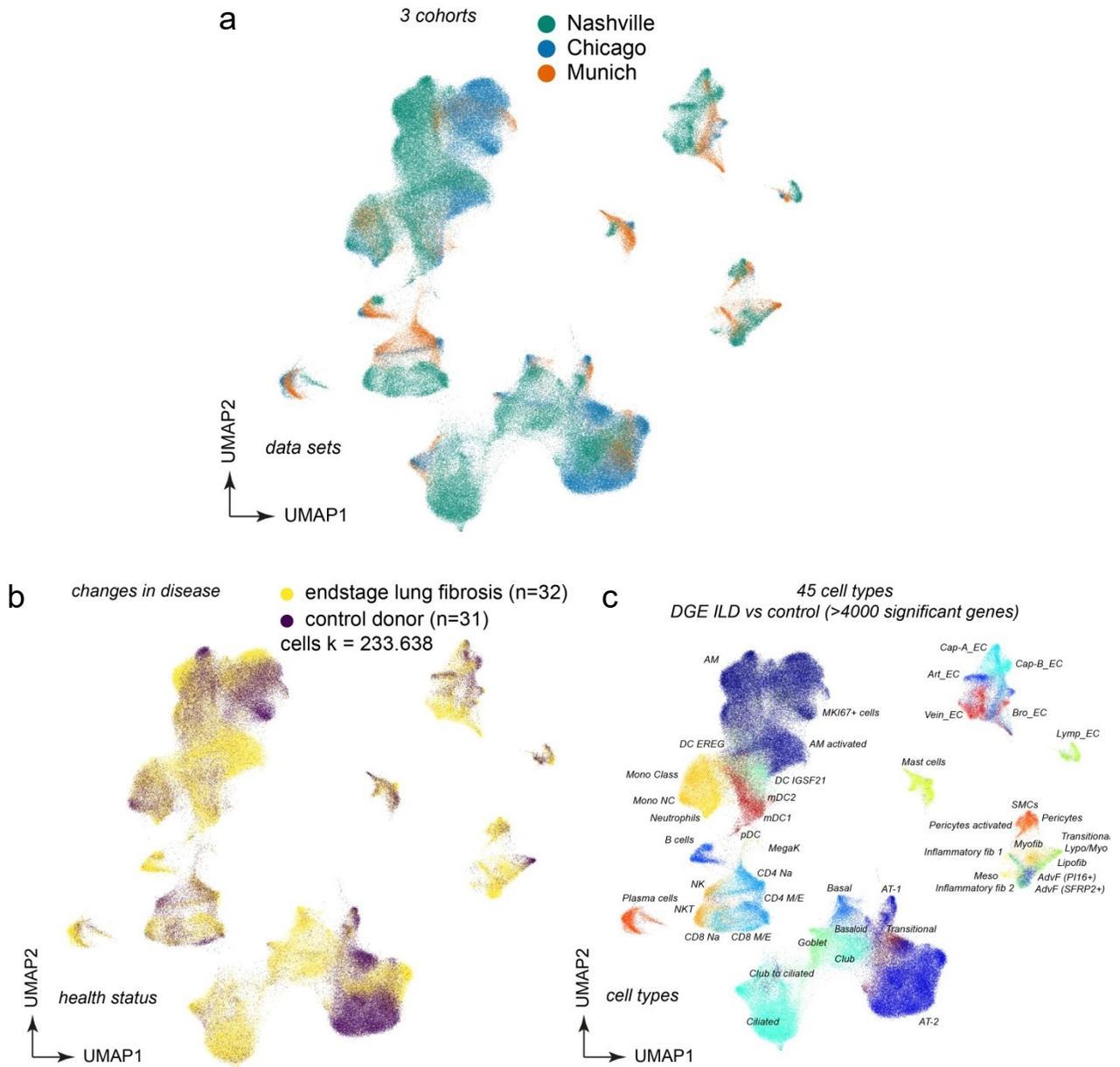


Figure 15. Multi-cohort single cell data integration. Dimension reduced single cell transcriptomic data is visualized through multiple UMAPs. The color code illustrates the patient cohort (a) disease status (b) and cell type identity. Figure and legend taken from (Mayr et al. 2020).

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

For validation of the degree of reproducibility with respect to technical quality across all three cohorts, the UMAP visualization was again assessed by the degree of overlay of the color-code according to cohorts (Fig. 15a.) and found as agreeing very well, considering that the Chicago and Nashville cohorts were sampled with a different single cell method. As described above for the Munich cohort, the object was sub-clustered to allow for cell type identification, this time using the already present classification of cells from the Munich cohort for easier assignment (Fig. 15c).

With the massive increase in statistical power of 63 patients by integrating all three cohorts, not only the marker signatures of cell state identities became much more distinct and unique. Importantly, the comparison of 31 end stage fibrosis with 32 control patients allowed for a, so far unseen, powerful differential gene expression (DGE) analysis for gene expression changes between fibrosis and controls, stratified by cell type. Gene expression changes in disease were more similar within the major lung cell compartments: goblet, ciliated, club and basal cells clustered together with alveolar epithelial cells, in which AT-1, AT-2 and differentiation intermediates were combined into “meta-cell-types” for better visualization, as were mesenchymal and leukocyte lineages. Accordingly, all subtypes of endothelial cells, marked by the “_EC” ending, showed a similar correlation among each other, as well as the lymphocytes and myeloid cells (Fig. 16a). In the top regulated genes for alveolar epithelium, fibroblasts and macrophages, cell types which all play an important role in fibrosis, known biomarkers and cell type specific genes were present, proving the efficient functioning of the DGE algorithm (Fig. 16b). Comparison of important marker genes, such as KRT17 for basal cells in the alveolar epithelium (Fig. 16c), DIO2 for myofibroblast (Fig. 16d) or SPP1 for macrophages (Fig. 16e), showed very good agreement between the three independent patient cohorts, with diseases patients acting alike and opposite to the control samples. Of note, these genes were chosen as representation as they are commonly known to be upregulated in diseases; however genes downregulated in ILD patients appeared correctly in the data as well, e.g. as seen in the heatmap in Fig. 16e.

The DGE between end stage lung fibrosis and donor patients, per cell type, provides a great resource to disentangle, or in other words, perform deconvolution on bulk data. Thereby proteomics or bulk mRNA sequencing data, which are more common than single cell data and exist for larger cohorts, can gain more value by assigning bulk changes of protein or gene signatures to specific cell types.

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

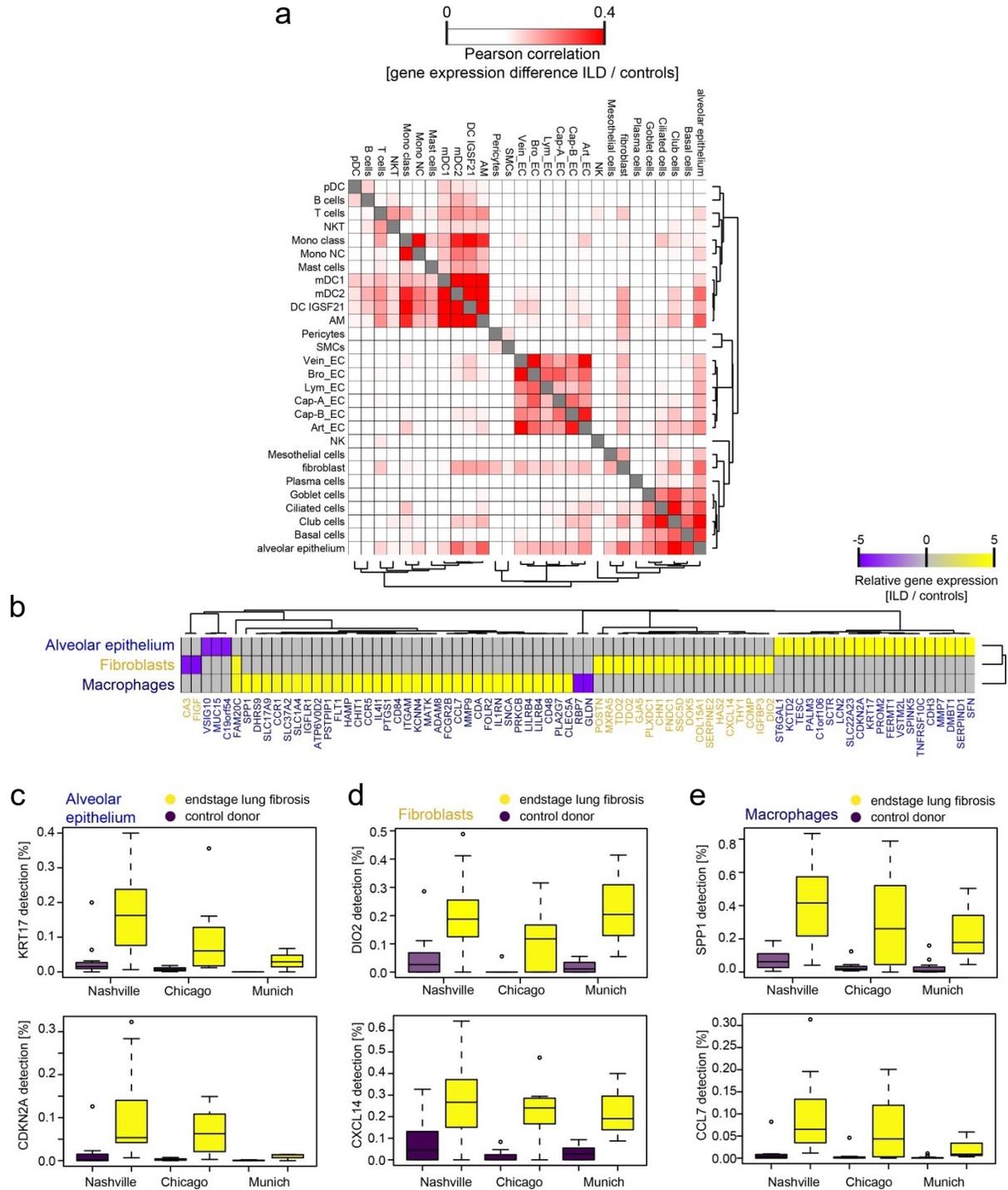


Figure 16. Transcriptional changes in >40 cell types and altered cell type frequencies in disease progression. (a) Differential gene expression between endstage lung disease patients and controls across cohorts was compared for the indicated cell identities. The color code illustrates similarities of gene expression changes calculated by Pearson correlation of the t-value coefficient, which represents differences in likelihood of detection for any gene between health and disease. (b) The heatmap illustrates the top 79 genes differentially expressed in the indicated cell identities. (c-e) The box plots illustrate differences in mRNA detection for the indicated genes between control donors and fibrosis patients in (c) alveolar epithelial cells, (d) fibroblasts and (e) macrophages. Figure and legend taken from (Mayr et al. 2020).

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

4.1.2 Human lung bronchoalveolar lavage fluid (BALF) proteomes reflect changes in disease activity

With the cell state changes from the single cell analysis at hand, the question arose, whether they might be reflected in the proteomic composition of the luminal epithelial lining fluid (ELF), which is accessible for sampling during bronchoscopic examination of patients. Besides the higher clinical relevance that BALF analysis could bring compared to single cell analysis, the comparison with the protein level is of interest, since changes on a transcriptional level are not necessarily correlating with protein abundance, which holds true especially for secreted proteins (Angelidis et al. 2019). State of the art MS was used for an in depth study of ELF proteins, sampled from cell free BALF of a large ILD and non-ILD patient cohort (Fig 18a). BALF proteomes were measured from eight groups of patients that were diagnosed with different form of ILD (Fig. 18b), including patients with idiopathic pulmonary fibrosis (IPF, n=16), hypersensitivity pneumonitis (HP/EAA, n=8), cryptogenic organizing pneumonia (COP, n=11), idiopathic non-specific interstitial pneumonia (NSIP, n=10), smoking associated respiratory bronchiolitis ILD (RB-ILD, n=3), sarcoidosis (n=22), unclassifiable ILDs (other ILDs, n=25), and non-ILD controls (non-ILD, n=29) (see Table 1 for clinical features of the ILD cohort). Of note, the majority of lavage fluids from patients in this cohort was collected during evaluation of initial diagnosis of ILD, and thus rather represents early disease states.

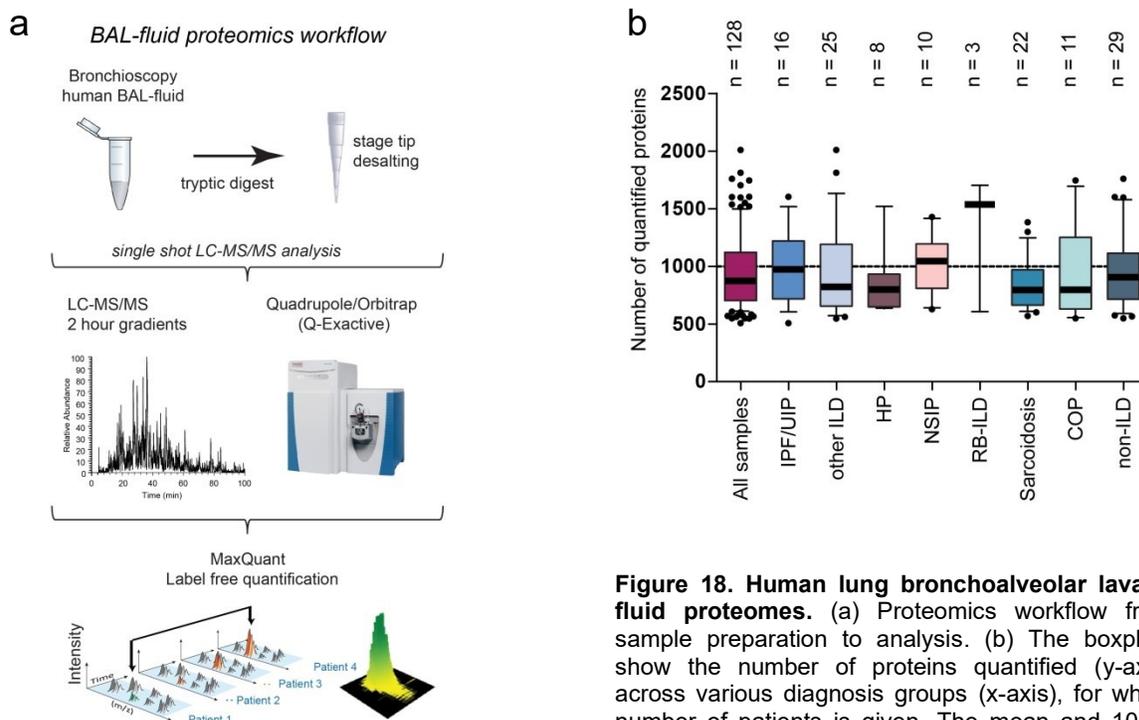


Figure 18. Human lung bronchoalveolar lavage fluid proteomes. (a) Proteomics workflow from sample preparation to analysis. (b) The boxplots show the number of proteins quantified (y-axis) across various diagnosis groups (x-axis), for which number of patients is given. The mean and 10-90 percentiles are shown. Figure and legend taken from (Mayr et al. 2020).

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

	All (n=124)	IPF (n=16)	ILD (n=25)	HP (n=8)	NSIP (n=10)	RB-ILD (n=3)	Sarcoidosis (n=22)	COP (n=11)	Non-ILD
Age, years	60.3±14.5	68.0±8.3	68.0±8.3	61.3±16.1	63.5±8.6	42.7±11.0	47.3±15.0	67.0±9.9	60.8±14.4
Female gender, n (%)	55 (44.4)	4 (25.0)	13 (52)	3 (37.5)	7 (70)	1 (33.3)	6 (27.3)	7 (63.6)	14 (48.3)
History of smoking	73 (65.8) n= 111	12 (85.7) n=14	15 (75.0) n=20	3 (50.0) n=6	5 (10.0) n=10	3 (100) n=3	11 (52.4) n=21	6 (66.7) n=9	18 (64.3) n=28
Pack years	22.3±23.1	40.5±24.9	17.5±15.8	20.0±28.3	35.7±32.2	25.0±8.7	6.8±10.2	27.0±20.2	22.2±24.6
Lung function									
FVC [l] (n=112)	2.9±1.0	2.5±0.7	2.7±0.9	2.7±0.8	2.3±0.5	3.1±0.4	3.8±1.2	2.5±0.6	3.1±1.0
FVC [%soll] (n=112)	80.0±19.8	71.5±17.1	75.8±16.6	77.9±21.6	79.5±20.6	71.4±4.5	88.1±19.3	80.9±22.0	83.9±22.4
FEV1 [l] (n=112)	2.3±0.9	2.1±0.7	2.1±0.7	2.2±0.7	1.9±0.4	2.4±0.8	3.1±1.0	2.0±0.6	2.4±1.0
FEV1 %/ FVC (Tiffeneau %) (n=112)	81.9±11.9	86.7±8.6	81.3±8.4	82.9±12.4	88.5±14.5	75.0±15.8	84.3±7.7	80.1±17.3	77.0±13.1
TLC [l] (n=110)	5.2±1.4	4.5±1.1	5.0±1.6	5.2±0.6	4.1±0.9	4.9±0.4	5.8±1.5	4.8±1.0	5.9±1.4
RV [l] (n=111)	2.3±0.9	1.9±0.6	2.3±1.0	2.4±0.4	1.8±0.7	1.8±0.1	2.0±0.6	2.3±0.4	2.8±1.0
DLC _o (SB) [% Soll] (n=94)	58.3±21.8	42.6±20.9	51.6±16.8	52.9±13.7	46.8±9.1	48.6±0.4	73.5±22.5	61.2±23.9	67.4±21.8
Laboratory findings									
LDH U/l (n=83)	249.1±87.3	266.1±93.3	241.1±84	242.9±90.2	340.4±100.3	190.5±30.4	212.7±60.5	270.7±102.9	234.8±73.1
CRP [mg/l] (n=85)	16.6±27.4	7.3±6.1	21.3±37.2	3.0±1.4	8.8±8.4	3.4±0.6	10.0±20.0	39.8±33.6	20.1±30.1
BAL cytopsin (n=36)									
BAL alveolar makrophages [%]	66.9±23.0	71.9±24.5	66.1±20.6	38.6±19.7	84.3±13.1	92.8±2.9	60.4±27.5	49.3±27.2	72.8±13.5
BAL lymphocytes [%]	24.4±19.7	20.8±24.3	32.3±16.3	55.0±15.8	8.4±8.1	3.0±1.1	30.8±14.4	28.2±25.4	16.1±11.5
BAL neutrophiles [%]	8.2±15.4	2.2±1.3	3.6±3.1	3.0±0.3	7.7±6.6	3.5	9.2±16.0	30.3±43.3	8.7±8.1
BAL eosinophiles [%]	2.6±3.4	4.3±3.6	2.9±4.9	3.3±4.2	1.8±1.9	1.9±0.1	0.5±0.5	2.5±2.6	2.2±3.9
CD4/CD8 quotient	2.6±2.9	2.3±3.1	2.4±3.2	3.9±4.7	0.7±0.9	1.4±0.8	5.0±5.1	1.5±1.2	3.3±2.4

Table 1. Clinical characteristics of bronchoalveolar lavage fluid study cohort.

Abbreviations: idiopathic pulmonary fibrosis (IPF), interstitial lung disease (ILD), hypersensitivity pneumonitis (HP), non-specific idiopathic pneumonitis (NSIP), respiratory bronchiolitis interstitial lung disease (RB-ILD), cryptogenic organizing pneumonia (COP), bronchoalveolar lavage (BAL). Table and legend taken from (Mayr et al. 2020).

BALF in general is difficult to analyze by MS, due its high dynamic range of protein copy numbers present, that not only stem from the lung. The composition of the ELF of the airways and the respiratory unit is determined by a combination of truly secreted proteins, plasma and tissue

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

leakage proteins, leaking into the damaged alveolar space from the blood and plasma, proteolytically shredded receptors, and intracellular proteins released from dying cells. Depending on the integrity of the epithelial barrier and the degree of cell death, the relative proportion of these protein sources can change. Therefore, some quality criteria, such as at least 500 proteins detected per sample and one protein had to be identified in at least 20 patients, were installed, resulting in a cohort of 124 patients, consisting of 95 ILDs and 29 non-ILDs. A median of 835 proteins per individual patient were quantified, leading to a total of 1513 unique proteins that were detected in at least 20 patients (Fig. 18b), which represents a very good depth.

To better define the proteins that are true constituents of the ELF, rather than the mentioned tissue leakage proteins, a quantitative comparison of BALF content and total tissue proteomes from 11 end-stage ILD tissue biopsies was done (Schiller et al. 2017). Proteins detected in both tissue and fluid proteomes, were scored as either a 'tissue leakage' protein or true 'epithelial lining fluid' protein based on their enrichments in the respective compartment by a t-test statistic score (Fig 19a). 110 proteins with significant enrichment in the BALF proteome were identified, indicating that these proteins are released into the airspace lumen in a directed manner. Upon category enrichment, indeed, proteins specific for secretory epithelial cells such as club and AT-2 cells had a significantly higher ELF enrichment score as proteins specific to non-secretory tissue resident AT-1 cells. Similarly, we found that proteins with the Uniprot keywords 'secreted' had a higher score than 'transmembrane' proteins and 'cytoplasmic' proteins (Fig. 19b). To identify proteins, with high heterogeneity between patients, the coefficient of variation (CV) was calculated, resulting in 285 proteins with a high coefficient above noise (Fig. 19c). A Fisher's exact test was performed on those top 285 proteins, to test for enriched protein categories across patients. They were significantly enriched for gene annotations such as 'secreted', 'plasma lipoprotein', 'antimicrobial', 'nucleosome', 'intermediate filament' and 'extracellular matrix', indicating that these categories are regulated across patient groups (Fig. 19d).

Interestingly, principal component analysis of patients, did not reflect overall clinical diagnosis, but revealed heterogeneous patient clusters that were either significantly enriched in 'complement', 'coagulation proteins' and 'plasma lipid transport' proteins, or showed higher levels of 'antimicrobial' proteins and 'histones'. Thus, these clusters were pointing towards the involvement of an inflammatory response driven by neutrophil extracellular traps (NETs) in these patients (Fig. 19d, e) and might therefore rather reflect molecular and cellular alterations that are a consequence of disease progression or acute illness.

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

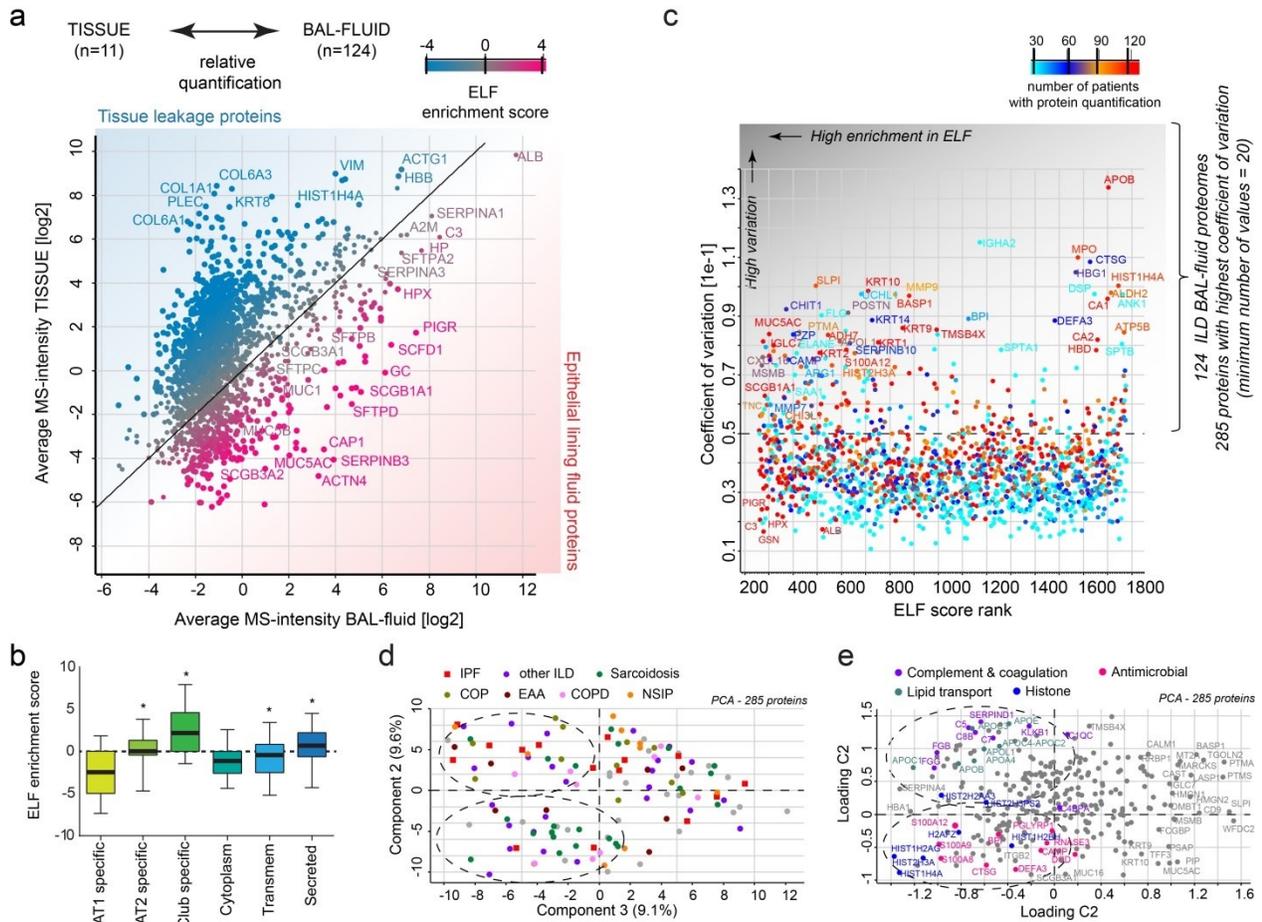


Figure 19. Heterogeneity of the epithelial lining fluid proteome in the interstitial lung disease patient cohort. (a) Comparison of BAL fluid proteome (n=128) and ILD lung tissue proteome (n=11) allows the separation of true constituents of the epithelial lining fluid from tissue leakage proteins and identifies 199 proteins with significant enrichment in the BAL-fluid proteome. The color code shows the relative enrichment of proteins in fluid versus tissue (ELF score). (b) The box plot shows distributions of ELF enrichment scores for the indicated gene categories. (c) Identification of 285 proteins with highest coefficient of variation and heterogeneity between patients enriched in human BAL-fluid. (d, e) Principal component analysis reveals groups of patients with similar enrichment for the indicated gene categories, irrespective of the indicated clinical diagnosis. Figure and legend taken from (Mayr et al. 2020).

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

Correlation analysis of the protein expression profiles revealed, that in fact, many proteins were co-expressed across patients, revealing co-regulated protein modules that were enriched for distinct signatures. The co-regulated protein modules can be either derived from the same cell type and reflect changes in relative cell type frequencies, or be the consequence of a common transcriptional program of cells to their environment (Fig. 20). Module A was enriched for transmembrane receptors and contained many macrophage specific proteins such as the Scavenger receptor cysteine-rich type 1 protein M130 (CD163) and the Complement C1q subcomponent subunit C (C1QC). Module B showed a positive correlation with module A and was composed of proteins involved in wound healing, such as the ECM proteins Tenascin-C (TNC), Fibronectin (FN1), Collagen type 6 (COL6) and Periostin (POSTN), as well as lipid transport, complement and coagulation proteins such as Apolipoprotein B-100 (APOB) and Complement component C7 (C7). Module C was significantly enriched for Keratin family proteins and contained the basal cell markers Keratin-5 (KRT5) and Keratin-14 (KRT14). Module D was enriched for cytoplasmic and cytoskeletal components, as well as focal adhesion and lysosomal proteins such as Talin-1 (TLN1) and Cathepsin Z (CTSZ). Module F was enriched for oxidoreductases and contained the Aldo-keto reductase family 1 member B10 (AKR1B10). Finally, module G was enriched in proteins involved in antimicrobial defense and neutrophil chemotaxis, including granulocyte specific proteins such as S100-A8, S100-A9, Cathelicidin antimicrobial peptide (CAMP) and Myeloperoxidase (MPO) (Fig. 20).

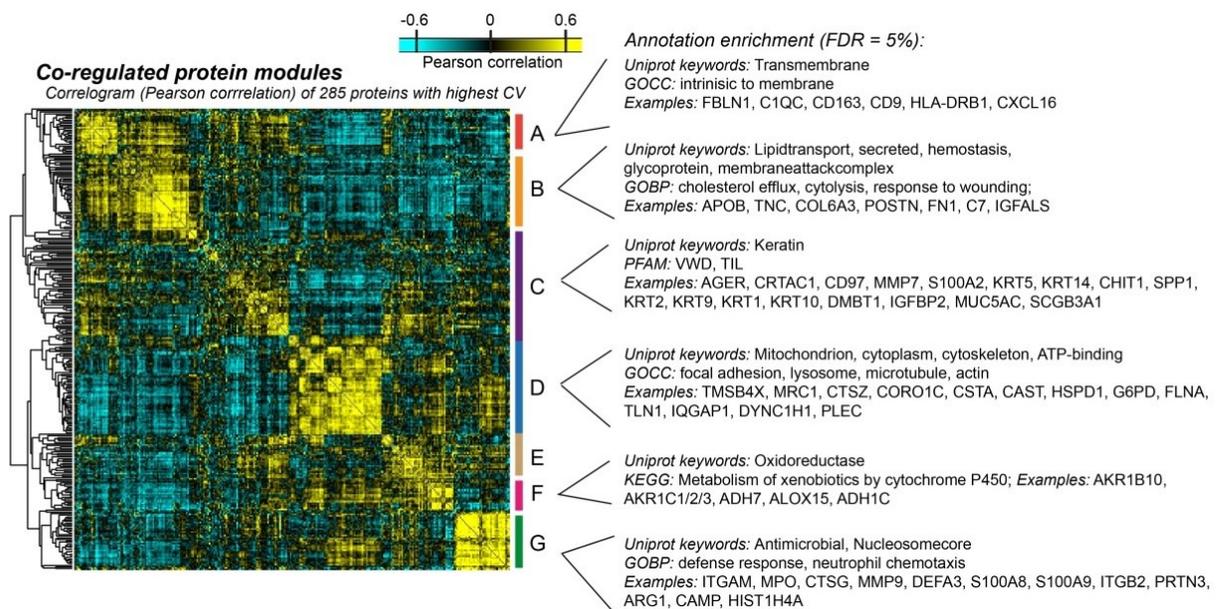


Figure 20. Co-regulated protein modules in the epithelial lining fluid proteome of the interstitial lung disease patient cohort. The correlogram shows the Pearson correlation of the 285 proteins with highest CV across individual patients. Enriched gene categories and examples of co-regulated proteins are shown. Figure and legend taken from (Mayr et al. 2020).

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

Unsupervised analysis of the BALF proteome data revealed molecular signatures that did not correspond to clinical diagnosis. However, the clinical cohort of BALF samples, was characterized by clinical parameters consisting of 33 individual clinical measurements per patient, including various lung function parameters and plasma lactate dehydrogenase (LDH), that were available for analysis. To identify associations of the identified distinct protein signatures with these clinical parameters, Pearson correlation analysis with subsequent hierarchical clustering was performed (Fig. 21e). Many clinical parameters were significantly correlated with each other (Fig. 22a). This is exemplified by the expected decline of the diffusing capacity of the lung for carbon monoxide (DLCO) with age observed in our cohort, illustrating the high quality of clinical data used for this biomarker study (Fig. 22b). In sarcoidosis, it has been reported, that DLCO and CD4/CD8 ratio decrease with disease severity (from sarcoidosis stage I to III) (Capelli et al. 2002). Interestingly, in this cohort, a significant positive correlation of CD4/CD8 cell count ratio in BAL of ILD with DLCO was observed (Fig. 22c). The 33 clinical parameters were then employed to identify biomarkers by Pearson correlation analysis to all 1513 proteins that were quantified in BALF in at least 20 patients.

Thereby biomarker signatures were identified, revealing highly significant correlations of distinct sets of proteins with lung function parameters, including DLCO (Fig. 21a), or plasma levels for lactate dehydrogenase (pLDH) (Fig. 21b). Of note, top outliers remained significant even after accounting for patient age (Fig. 23b-e). For instance, the cartilage acidic protein 1 (CRTAC1) showed a very robust negative correlation (Fig. 21c), while the ECM protein Fibulin-1 (FBLN1) was positively correlated to pLDH (Fig. 21d). In total, 72 proteins with significant correlation (<10% FDR) with at least one clinical parameter were identified. Hierarchical clustering of these proteins identified two main clusters of proteins, mainly driven by correlations with lung function but also demographics, laboratory values and cytopsin results (Fig. 21e). Proteins of the first cluster included A2M, the chemokine CCL18 and the ECM protein FBLN1, and showed a strong negative correlation to lung function parameters such as DLCO, and a positive correlation to age and plasma LDH. The second cluster, including proteins like the intercellular adhesion molecule 1 (ICAM1), tetranectin (CLEC3B) and the cartilage acidic protein 1 (CRTAC1) showed the opposite pattern.

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

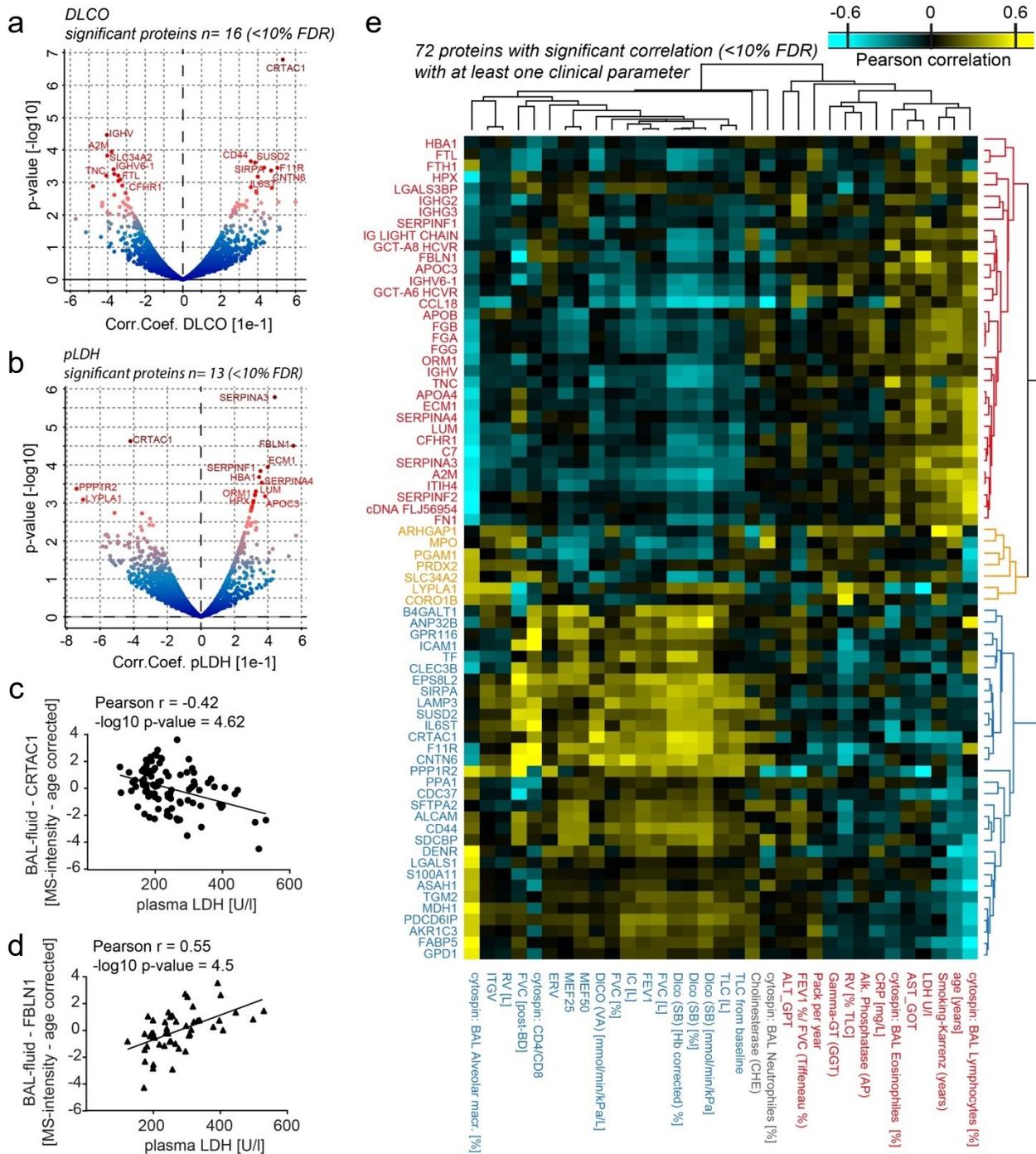


Figure 21. Human lung bronchoalveolar lavage fluid proteome changes correlate with clinical parameters.

(a, b) The volcano plot shows Pearson correlation values (x-axis) and the $-\log_{10}$ p-value (y-axis) for BALF protein abundance and (a) DLCO and (b) plasma LDH (pLDH) values, respectively. (c, d) The scatter plots show (c) the negative correlation of CRTAC1 abundance in BALF (MS-intensity) with plasma LDH levels, and (d) the positive correlation of FBLN1 abundance in BALF (MS-intensity) with plasma LDH levels. (e) The Pearson correlation values for 72 significantly regulated proteins ($<10\%$ FDR) with indicated clinical parameters were grouped by hierarchical clustering into protein negatively and positively correlated with lung function (red and blue font respectively for negative and positive associations). Figure and legend taken from (Mayr et al. 2020).

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

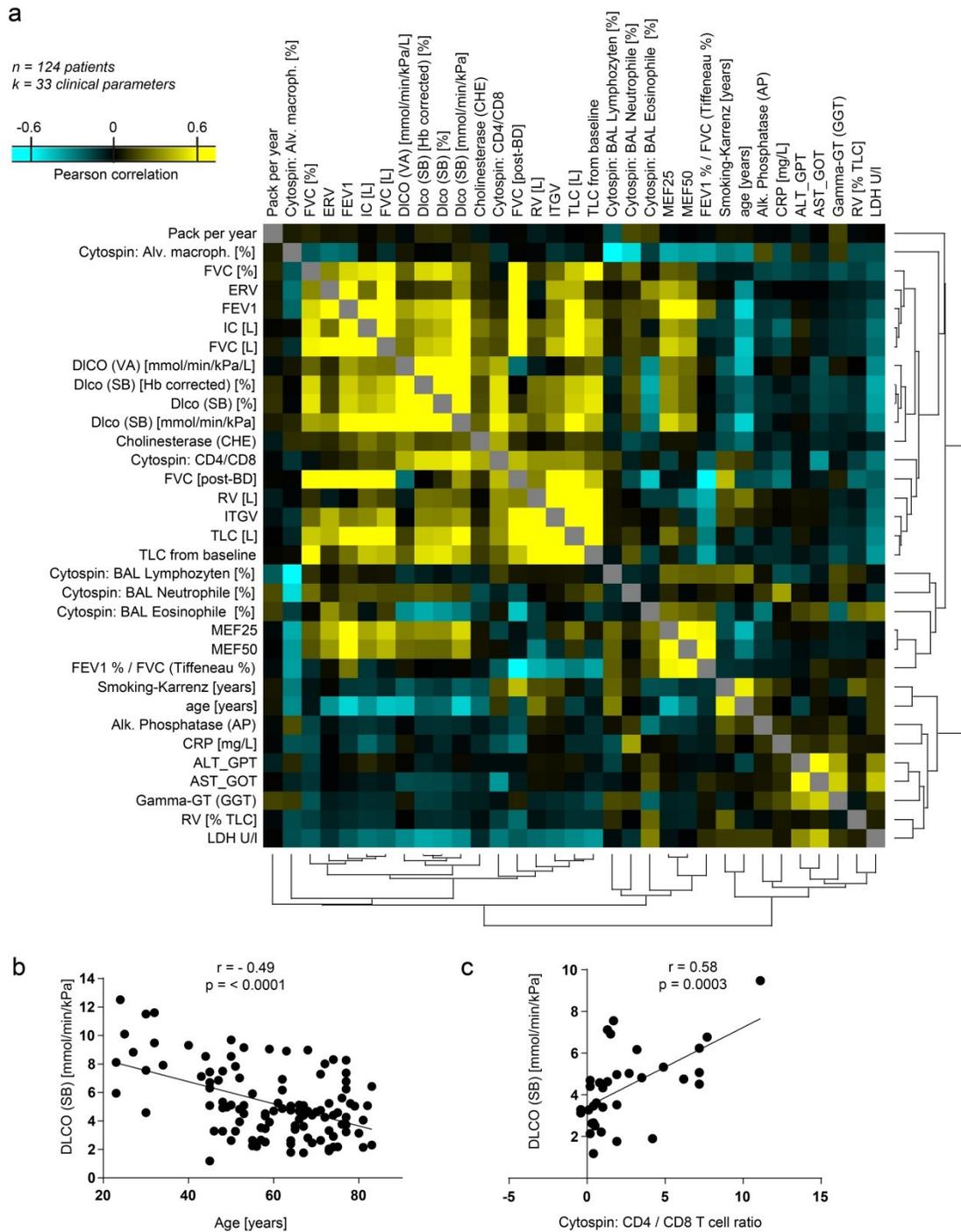


Figure 22. Correlation patterns between 33 clinical parameters in the ILD BALF patient cohort. (a) Pairwise Pearson correlation values of 33 clinical parameters were grouped by hierarchical cluster analysis. (b) DLCO shows negative correlation with age in the study cohort ($p < 0.0001$). (c) Positive correlation of CD4/CD8 ratio in BAL fluids with DLCO. Figure and legend taken from (Mayr et al. 2020).

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

Most proteins that we found increased in BALF of patients with high pLDH, were also associated with lower lung function (Fig. 23a). LDH in blood plasma is routinely measured in the clinics. With LDH being released during tissue damage and transpired to the blood, its levels in blood plasma are clinically used as a marker of common ongoing injuries and diseases such as heart failure. One hypothesis coming up was that BALF proteins with correlation to pLDH in human patients diagnosed with ILD, represent much likely a lung injury signature.

To test this, a comparison of the human pLDH signature with BALF proteomes from mice after bleomycin treatment, representing acute lung injury, was performed (Schiller et al. 2015). And indeed, similar outlier proteins across species were revealed, including the injury marker Tenascin-C (TNC), mostly produced by myofibroblasts in the lung and other ECM proteins that occur predominantly upon injury, such as Fibulin-1 (FBLN1), Collagen alpha-3(VI) chain (COL6A3) and Periostin (POSTN) (Fig. 23f). Using 1D annotation enrichment analysis on the compared orthologous mouse and human proteins, the similarity became even more striking. Categories such as `Fibrosis`, `Collagen` and `Extracellularmatrix` as well as `Acuteinflammation` and `Immunity`, which are associated with an upregulation upon injury, showed analogous positive enrichments (Fig. 23g). The comparison confirmed the hypothesis, that the pLDH correlation revealed protein changes in human patient BALF proteomes that are similar with the ones that can be observed upon a defined acute lung injury in the bleomycin mouse model.

The analysis of BALF proteomes from 124 patients and the correlated protein expression with an extensive set of clinical parameters represents a comprehensive characterization of human pulmonary epithelial lining fluid composition. The identified co-regulated protein modules that were associated with patients lung function and current injury status, provide interesting protein signatures of which some of could be used to monitor acute or subclinical exacerbations of ILD patients.

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

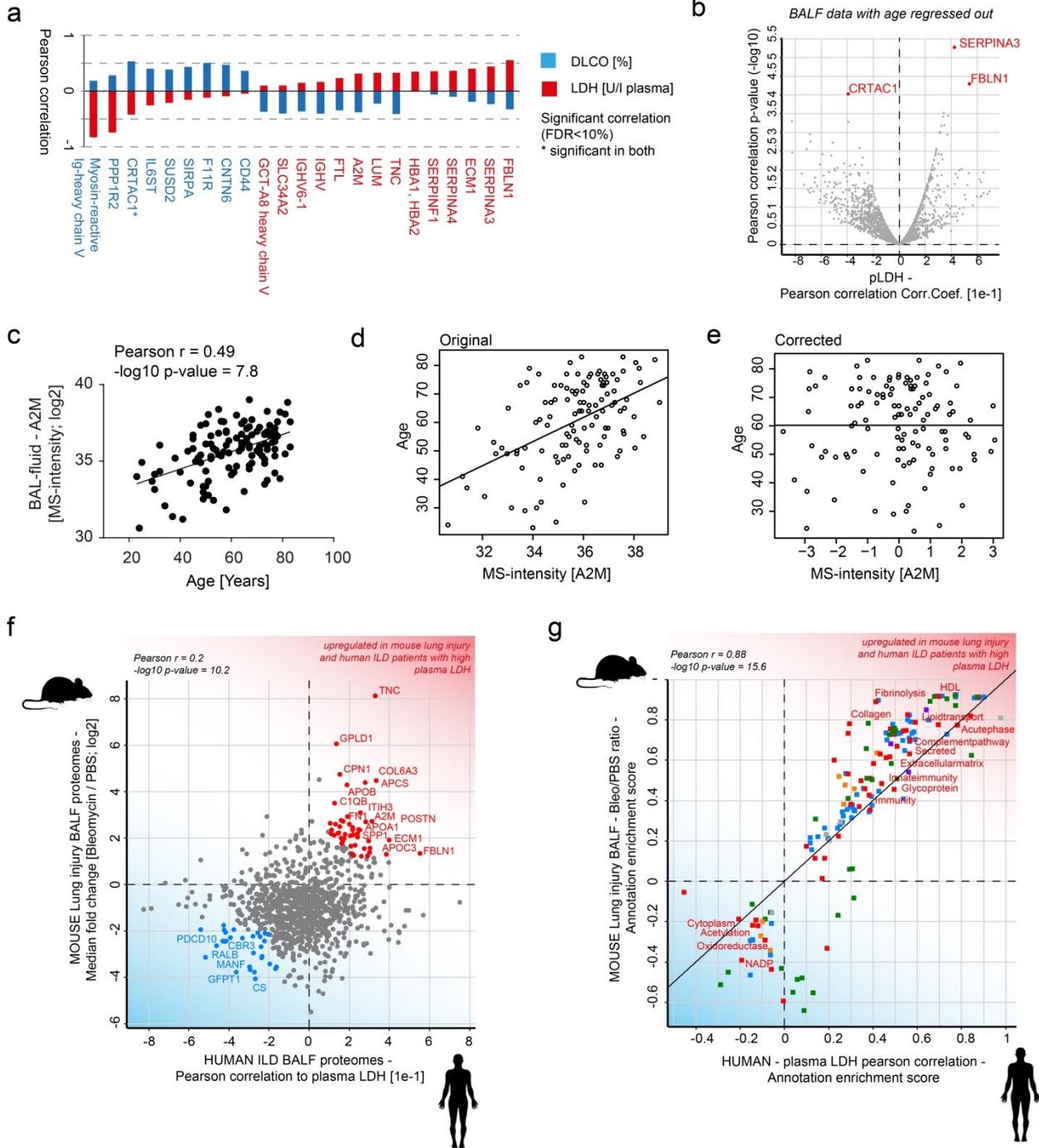


Figure 23. BALF proteins correlating with plasma LDH represent a human lung injury signature. (a) The bar graph shows the Pearson correlation values of the indicated proteins for DLCO [%] (blue) and plasma LDH [U/L] (red). (c) The scatter plot shows significant correlation of alpha 2 macroglobulin (A2M) abundance (MS-intensity) in BALF with patient age. (d, e) Original (d) and age-corrected (e) correlation of A2M in BAL fluids with age. (f) The scatter plot shows the Pearson correlation of individual proteins from the human BAL fluid proteome with plasma LDH (x-axis) and the fold changes (y-axis) of the orthologous proteins in mouse lung after bleomycin injury. (g) The annotation enrichment score shows a common upregulation of gene categories like acute phase, ECM, complement and innate immunity in the BAL fluids of bleomycin mice and human ILD with high plasma LDH. Figure and legend taken from (Mayr et al. 2020).

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

4.1.3 Correspondence of fluid proteins with transcriptional changes in specific cell types

Having the analysis of bulk BALF proteomes from 124 patients and the correlations of protein expression with 33 clinical parameters at hand, that were associated with patients lung function and current injury status, the next step was to make again use of the single cell DGE data between healthy and ILD. The aim was to explain quantitative changes in BALF protein signatures with the cell state changes analyzed by single cell RNA-seq. Therefore the diagnosis specific protein biomarker signatures in the BALF proteomes were deconvoluted and evaluated for the relative contribution of cell types. This was done by testing the mean intensity z-scores of protein expression across different clinical diagnosis groups, for an enrichment of transcriptional signatures that are cell type specific (Fig. 24a).

Markers of several pro-fibrotic cell types including the fibroblast subsets of pericytes, plasma cells and mesothelial cells, were strongly increased in protein measurements of COP, NSIP, HP/EAA, and IPF compared to non-ILD controls, confirming the power of BALF proteomics to correctly score fibrogenic remodeling in the patients. Also plasma cells showed a higher correlation with a fibrotic diagnosis, confirming data that show increased MZB1-positive plasma cells in ILD (Schiller et al. 2017) (Fig. 24a). Interestingly, RB-ILD and Sarcoidosis samples behaved more similar to non-ILD controls for this fibroblast signature, which is consistent with their distinct histopathology that does not involve strong interstitial fibrosis. While RB-ILD protein analysis featured very strong enrichment for proteins specific to airway basal, ciliated and goblet cells, the same airway protein signature was depleted in patients with COP, NSIP and HP/EAA but not IPF (Fig. 24a).

Similar to this analysis the correlations of the 33 clinical parameters were tested for correspondence with specific cell types. Also here strong associations of cell type signatures with the Pearson correlation of most clinical parameters with protein measurements were found (Fig. 24b). For instance, the myofibroblast specific proteins quantified in patient BALF tended to be negatively correlated with lung function (DLCO), in agreement with the expectation that myofibroblast mostly occur during injury (Fig. 24c). Similarly, the number of alveolar macrophages in BAL-cytospins tended to be negatively correlated with proteins (mostly antibodies) secreted by plasma cells into the ELF (Fig. 24d). Also the positive correlation of Critical Response Protein (CRP) value, an indicator for inflammation in the body, with monocytes and lymphocytes, proofed the quality of the analysis. Thus, deconvolution with cell type signatures could successfully reveal diagnosis and disease state specific biomarker fingerprints in protein measurements.

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

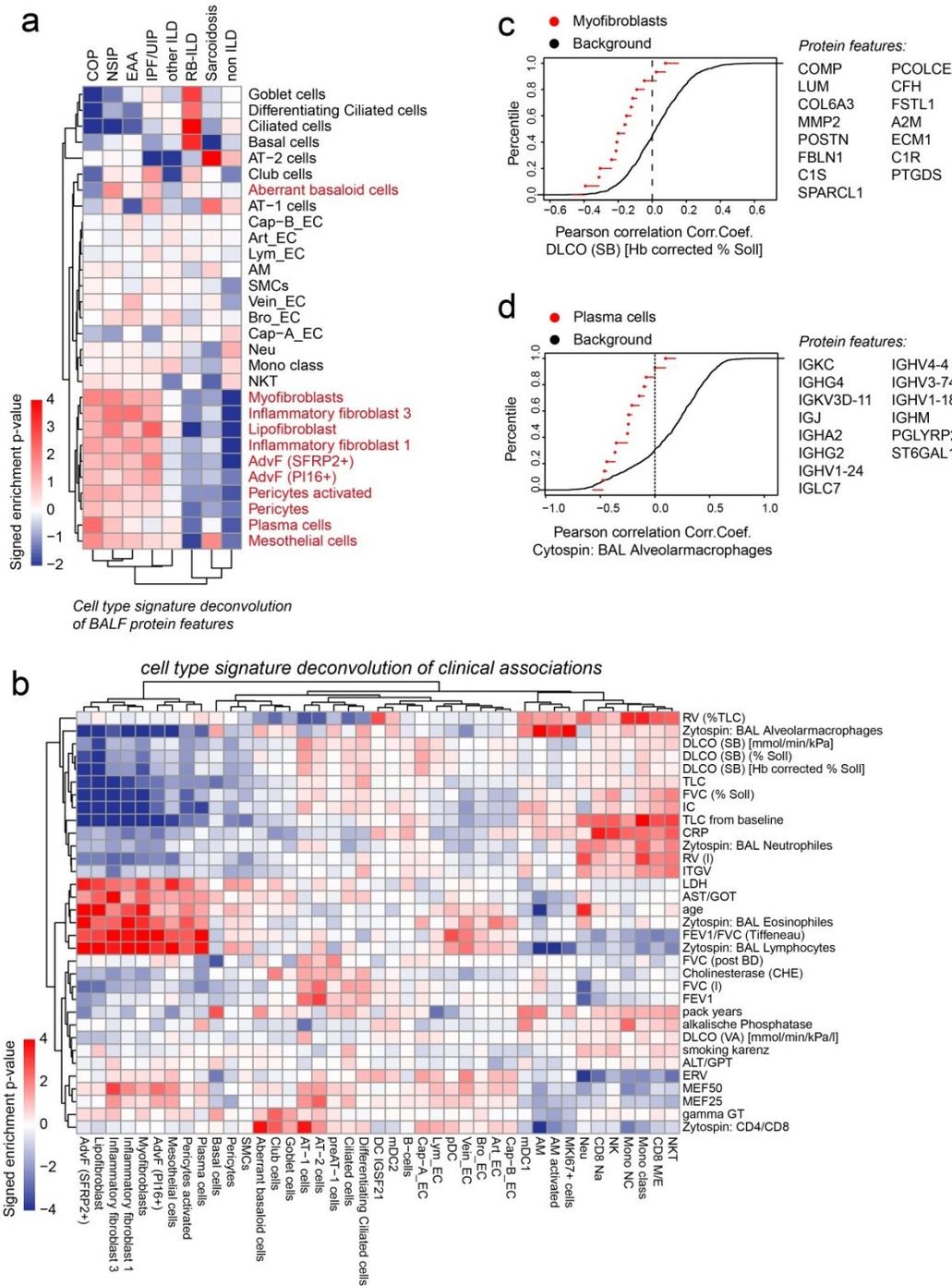


Figure 24. Protein signatures in BALF predict changes in cell types. (a) The heatmap shows relative contribution of cell types to the diagnosis specific protein biomarker signatures in epithelial lining fluid (ELF). (b) The heatmap shows relative contribution of cell types to the association of protein biomarker signatures in ELF with the indicated clinical parameters. (c, d) Empirical cumulative density plots show the distribution of correlation coefficients for (c) myfibroblast markers (red points) with DLCO and (d) Plasma cell markers (red points) with % alveolar macrophages in BAL and the background proteins (black line). Figure and legend taken from (Mayr et al. 2020).

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

While so far only the proteome data was tested with the single cell DGE, the next was to apply information from the BALF proteomics and the clinical data onto the single cell patients. To test if such a transfer of information from the proteomics modality into the scRNAseq data modality is possible, artificial intelligence and machine learning were applied. A random forest was trained on the protein quantification data to predict lung function (DLCO) using a set of proteins (Fig. 25a). These proteins had to fulfil the criteria of showing a high correlation with lung function (DLCO) and having their corresponding transcript detected in the scRNAseq data. The scRNAseq data was transformed into an *in-silico* bulk data, since instead of the single cell information the whole transcriptome per patient was needed for this analysis. Next, the trained model was applied to this bulk scRNAseq data and the mRNA expression mapped to proteins (Fig. 25b). This model then correctly predicted the direction of lung function changes in the three single cell RNA-seq cohorts, with endstage lung fibrosis patients being predicted, to have a significantly (in two out of three cohorts) lower lung function (Fig. 25b). Looking at the most important protein features, driving the random forest model, indeed, the most of them appeared to be regulated in a cell type specific manner at the mRNA level (Fig. 25c).

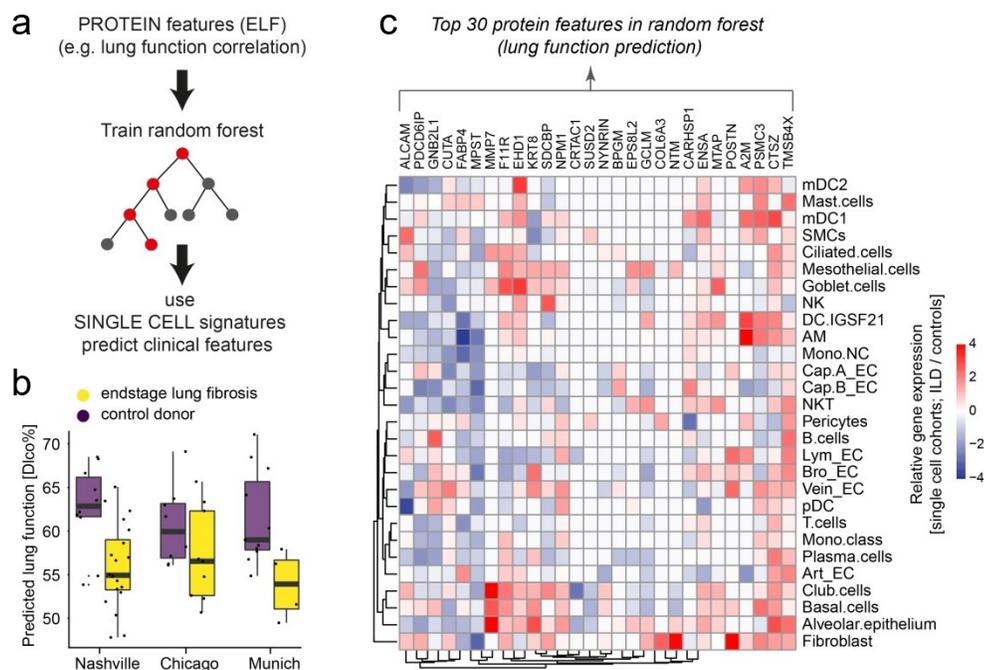


Figure 25. Protein signatures in BALF predict lung function decline and the corresponding cellular changes. (a) The Pearson correlation of protein features in ELF was used to train a random forest algorithm. Training data was used on transcriptional signatures in single cell RNA-seq data to correctly predict reduced lung function in endstage lung fibrosis compared to control donors. (b) Box plots show predicted lung function changes (DLCO %) in the three single cell RNA-seq cohorts. (c) Top protein features in the random forest training data are shown with their relative gene expression changes in the different indicated cell types, illustrating cell type specific changes in lung fibrosis for these BALF biomarkers. Figure and legend taken from (Mayr et al. 2020).

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

4.1.4 Smoking induces highly persistent cell state changes in the lung

Part of the 33 clinical parameters of the BALF cohort, was not only the information if the patients were smokers or not, but also if they had previously been smoking, as well as the pack years index for all smokers; used for COPD risk prediction and calculated by multiplying the number of cigarettes per day with the number of years they were smoking. Clinical characteristics of patients included in the smoking analysis are shown in Table 2. Given the critical impact of smoking on lung health, being the major risk factor for six of the eight leading causes of deaths in the world, including respiratory and cardiovascular diseases, stroke and several malignant diseases, the analysis of smoking induced changes was the next aim.

A robust smoking protein signature was identified in the BALF proteomes of the ILD cohort. To look at changes that stem from active smoking, a t-test between active smokers (n = 19) and never smokers (n = 36) identified 422 significantly regulated proteins (FDR<10%) (Fig. 26a). The most significantly regulated proteins included AKR1B10 and the ubiquitin carboxyl-terminal hydrolase isozyme L1 (UCHL1), both of which have been shown to be elevated in epithelium of 'healthy' smokers (non-ILD smoker) (Carolan et al. 2006, Wang et al. 2010). Comparing active smokers with ex-smokers (n = 49) identified 137 significantly regulated proteins (FDR<10%) (Fig. 26b). Hereby, only some of the top outliers were similar to the one from the comparison of active against never smokers. Interestingly, the test between ex-smokers against never smokers showed 36 significantly altered proteins (FDR<10%) (Fig. 26c). This indicates that while, some of the alterations induced by smoking are acute and vanish over time, there are persisting changes even after smoking cessation. Indeed, the fold changes between active/never and ex/never smokers were significantly correlated (Pearson $r = 0.4785$, $-\log_{10}$ p-value = 14.6536) (Fig. 27a). However, while surprisingly many changes in smokers persisted after cessation even over decades, a number of oxidoreductases, such as AKR1B10, were highly specific markers of active smoking; AKR1B10 showed significantly higher MS-values in active smokers than in never smokers or ex-smokers (Fig. 27a-c). The chemokine CCL18, on the other hand, previously reported as downregulated in alveolar macrophages (Kollert et al. 2009), continued to show lower MS-values also after smoking cessation, suggesting that alveolar macrophages permanently change their phenotype upon smoking (Fig. 27d, e). The scavenger receptor CD163 was also downregulated in both active and ex-smokers compared to never smokers (Fig. 24f). The decline in CCL18 and CD163 was permanent and independent of time after cessation (Fig. 27e and 27g, respectively). Contrary to never smokers, the intensity distribution of CD163 was bimodal in smokers (Fig. 27h), indicating that only a subset of patients is affected in this manner by smoke exposure and suggesting involvement of additional factors.

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

characteristics	active smokers n=19	Ex smokers n=49	Never smokers n=36	p-value
Age	46.9±14.5	65.9±10.9	58.8±13.6	<0.0001 ^{1,2,3}
Pack years	30.3±21.4	30.3±23.9	0	<0.0001 ^{2,3}
FVC (%)	76.8±21.9	81.1±20.7	79.6±18.0	0.76
DLCO (%)	59.9±21.8	53.6±21.3	63.5±21.3	0.16

Table 2. Clinical parameters of smoking cohort. (1) p-value active vs ex smokers <0.05; (2) p-value active vs never smokers <0.05; (3) p-value ex vs never smokers <0.05; Table and legend taken from (Mayr et al. 2020).

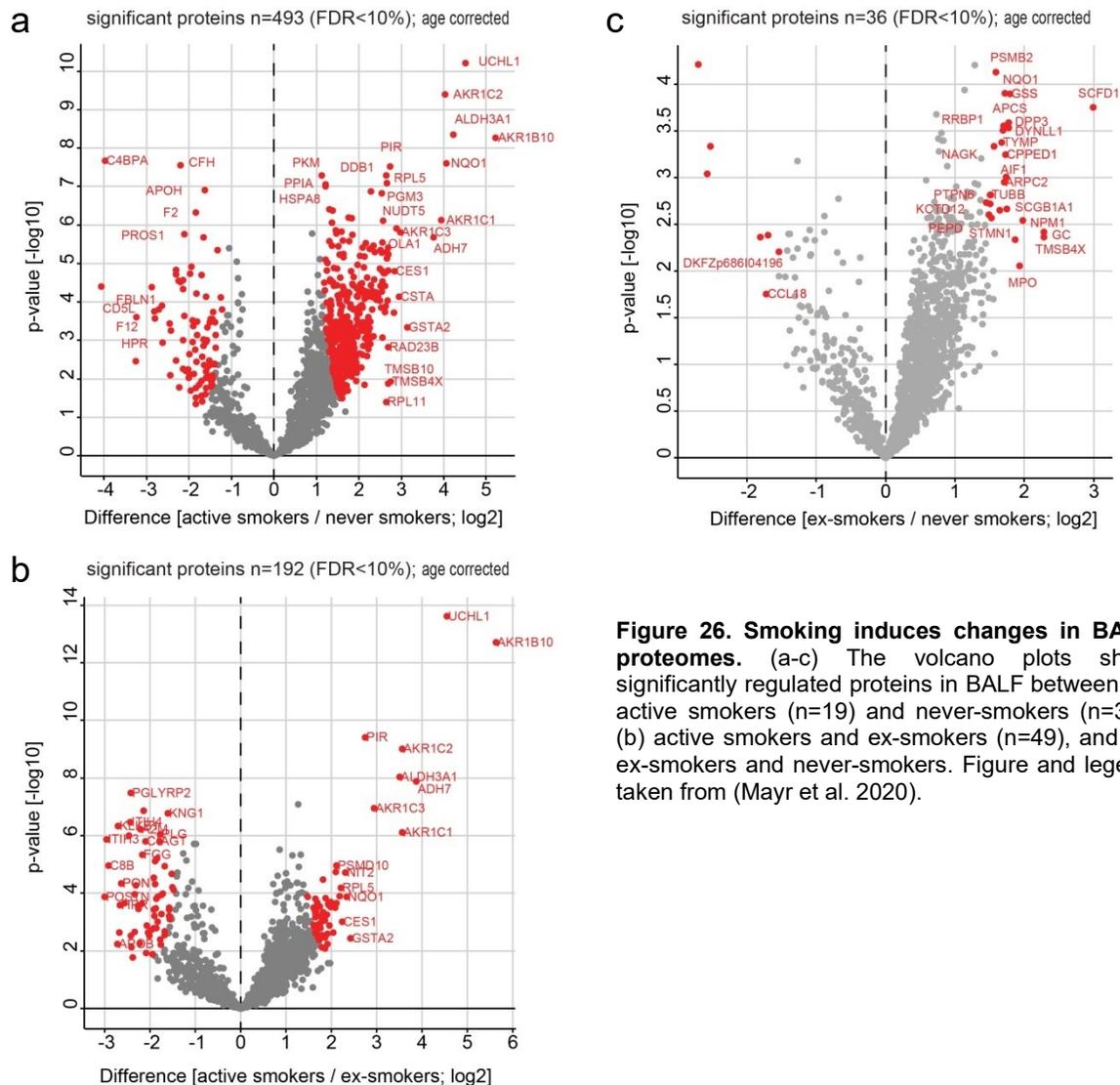


Figure 26. Smoking induces changes in BALF proteomes. (a-c) The volcano plots show significantly regulated proteins in BALF between (a) active smokers (n=19) and never-smokers (n=36), (b) active smokers and ex-smokers (n=49), and (c) ex-smokers and never-smokers. Figure and legend taken from (Mayr et al. 2020).

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

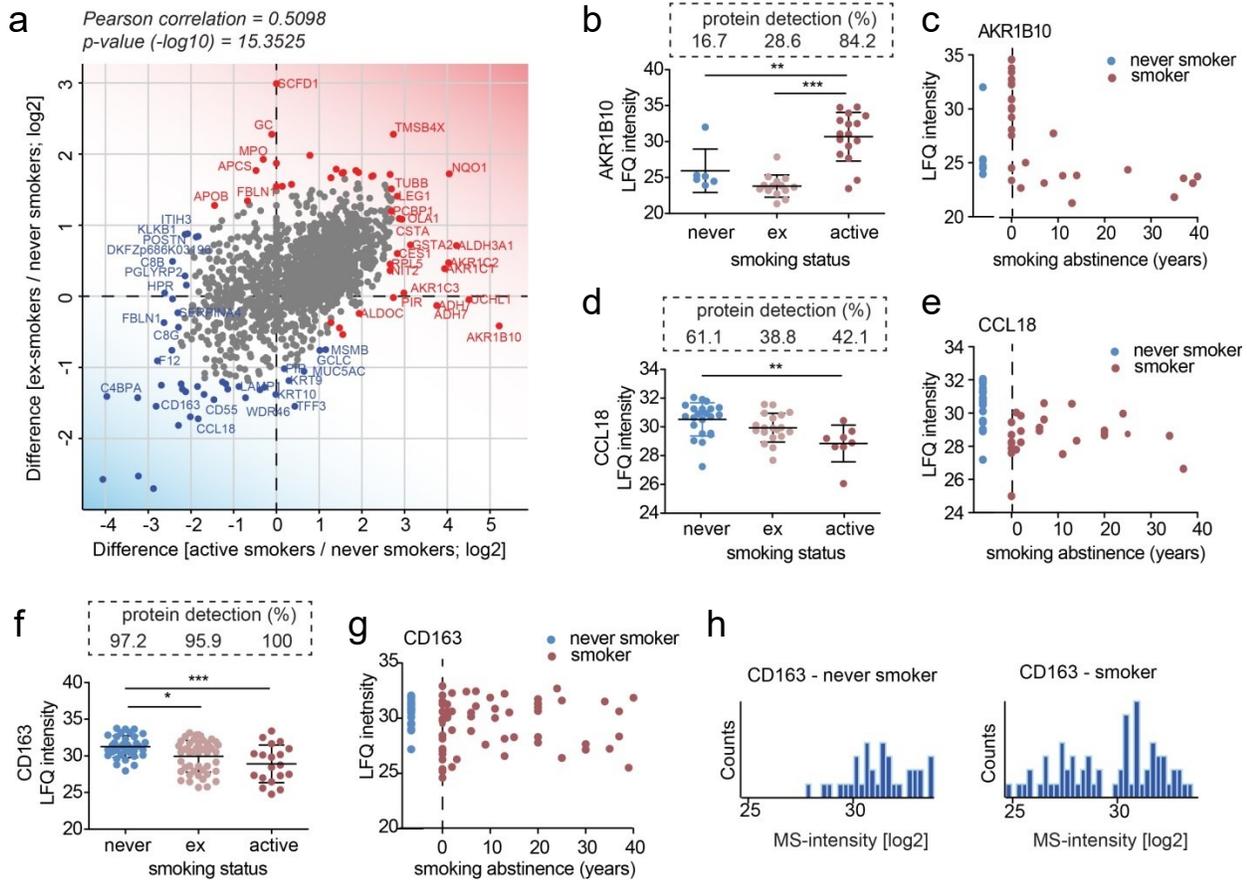


Figure 27. Smoking, active and former, induces specific and reliable transient and persistent changes reflected in BALF proteomes. (a) The scatter plot compares the fold changes between active smokers vs. never-smokers (x-axis) and ex-smokers vs. never-smokers (y-axis). (b) AKR1B10 is significantly higher and more often detected in BAL fluid of active smokers in comparison to ex- ($p < 0.0001$) and never smokers ($p < 0.0001$). (c) Analysis of AKR1B10 and years of smoking abstinence show a reversible expression pattern of AKR1B10 after smoking cessation. (d) CCL18 is detected significantly more often in never vs ex-smokers ($p = 0.0417$) and with significantly higher intensity in never vs active smokers. (e) The reduced levels of CCL18 do not go back to baseline after smoking cessation. (f) While CD163 was detected in almost all patients irrespectively of the smoking status, ex- and active smokers showed significantly lower protein levels in comparison to never smokers. (g) CD163 levels are changed in smokers and do not recover after smoking cessation. (h) The histograms show the distribution of MS-intensity values for CD163 in ELF across patients in never smokers (left panel) and smokers (ex- and active smokers – right panel). Figure and legend taken from (Mayr et al. 2020).

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

With the analysis of bulk BALF proteomes from smoking patients and the correlations of protein expression with the smoking status at hand, the next step was to make again use of the single cell DGE data. The relative contribution of cell types to the protein biomarker signatures in BALF with patient smoking status was tested by using deconvolution (Fig. 28a). Proteins derived from airway epithelial cells, especially ciliated, basal and goblet cells were strongly enriched in BALF of smokers compared to non-smokers. Interestingly, the increase of proteins from basal and goblet was reversible, while increased abundance of proteins derived from ciliated cells was persistent (Fig. 28a). Thus, the scRNAseq data allowed the assignment of the persisting and reversible effects that could be observed in the BALF protein data, to three distinct cell types in the lung. Interestingly, especially fibroblasts showed little contribution to the specific smoking signatures. Of note, the top proteins regulated with smoking status in BALF, which were also driving the deconvolution, were also regulated on transcriptional level in the three single cell RNA-seq cohorts in at least one cell type (Fig. 28b).

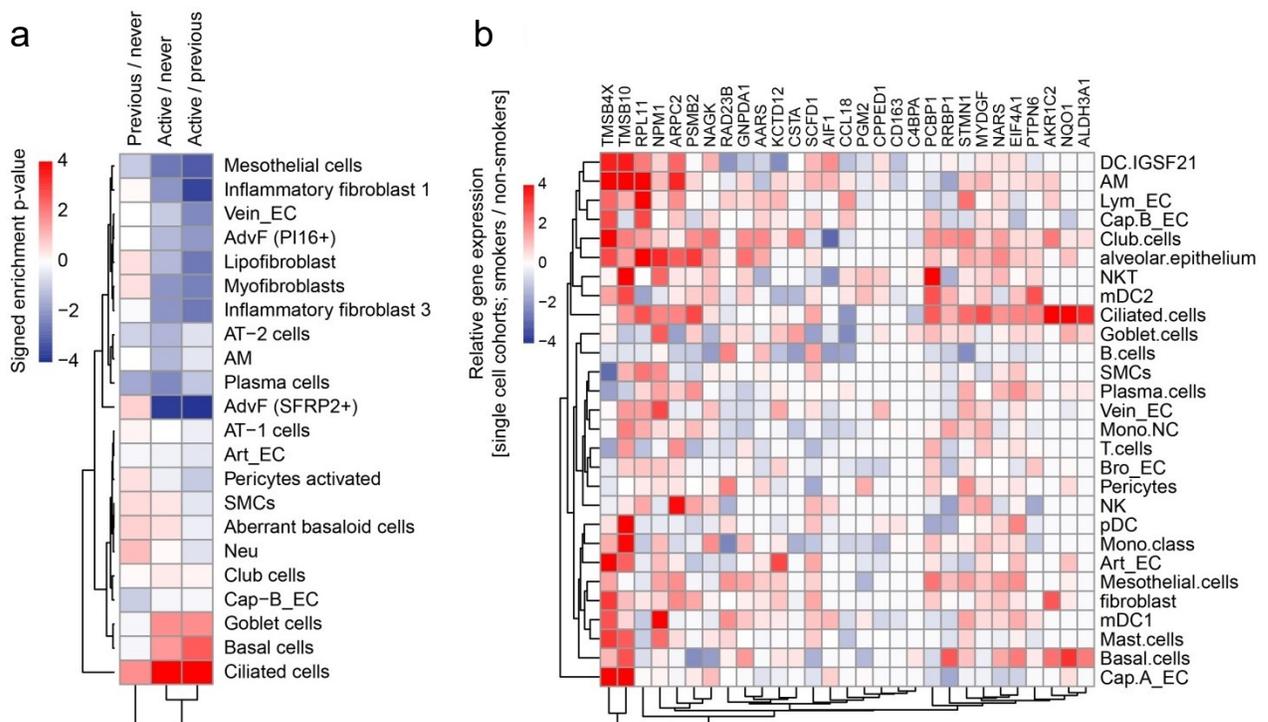


Figure 28. Smoking changes reflected in BALF proteomes can be assigned to certain cell types. (a) The heatmap shows relative contribution of cell types to the specific protein biomarker signatures in epithelial lining fluid (ELF) of active, former and never smokers. (b) Top protein features in the smoking correlation data are shown with their relative gene expression changes in the different indicated cell types. Figure and legend taken from (Mayr et al. 2020).

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

4.1.5 Biomarkers of lung health in plasma proteomes

As introduced earlier, body fluids in general, but most importantly blood, are used in the clinics on a regular basis to detect disease specific biomarkers and infer diagnosis or treatment. Having analyzed BAL fluid extensively the next goal was to extend the analysis to the blood, more specifically the plasma proteome. Using an established high throughput plasma MS-based proteomics workflow (Geyer et al. 2017, Geyer et al. 2019, Niu et al. 2019) (Fig. 29a), plasma proteomes from two independent cohorts of ILD patients were generated (Munich, n=30 and Hannover, n=81; healthy age matched controls, n=30; see Table 3 for clinical characteristics; due to ethics no information about the healthy individuals was available). Due to the specialization of the hospital in Hannover, the cohort included more patients with better lung function on average, with samples taken mainly at time of initial diagnosis. On the other hand, the cohort from the Munich hospital specialized on transplantation, contained patients closer to end stage disease. Thus, by default, no perfect match of these two cohorts was expected, but given the large numbers of patients, a validation cohort was much appreciated. Correlating the MS-intensities with FVC %, a shared panel of proteins in both cohorts was identified, that were either positively or negatively associated with the lung function outcome (Fig. 29b).

The secreted protein cartilage acidic protein 1 (CRTAC1) showed the most significant positive correlation with lung function, with higher levels being measured in the plasma of patients with better lung function in both cohorts. CRTAC1 was also positively correlated with lung function in the BALF proteome analysis (Fig. 29c, Fig. 21c), and was robustly detected by mass spectrometry in >80% of the plasma samples (Fig. 29d). Interestingly, CRTAC1 correlated negatively with pLDH, which correlated with a lung injury signature. Re-analysis of published bulk transcriptomes confirmed a highly significant downregulation of CRTAC1 mRNA in the lung of ILD patients compared to healthy controls and COPD patients (Fig. 29e).

	All n=111	Cohort Munich 3 n=30	Cohort Hannover n=81
Age, years	64.8±11.7	62.4±7.3	65.7±12.9
FVC [%soil]	65.7±19.2	55.4±16.9	69.3±18.7
DLC_o (SB) [% Soil]	46.6±21.4	27.5±17.2	53.8±18.3
LDH U/l	172.5±96.4	286.4±113.5	130.3±37.9

Table 3. Comparison of plasma proteome cohort Munich 3 and cohort Hannover. Table and legend taken from (Mayr et al. 2020).

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

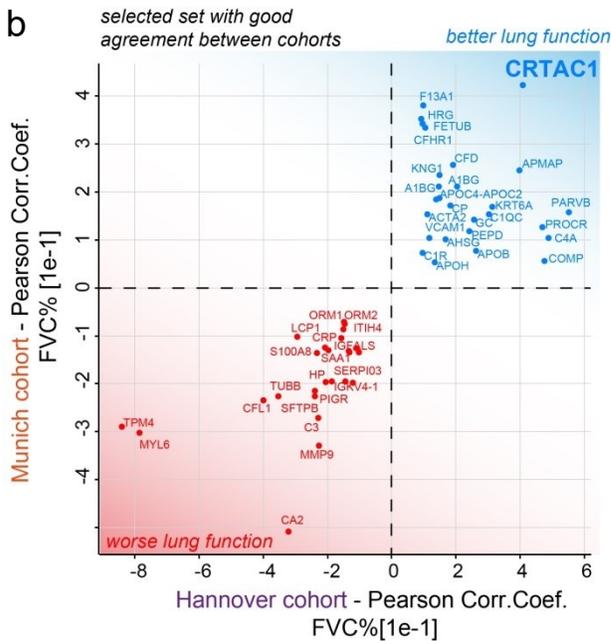
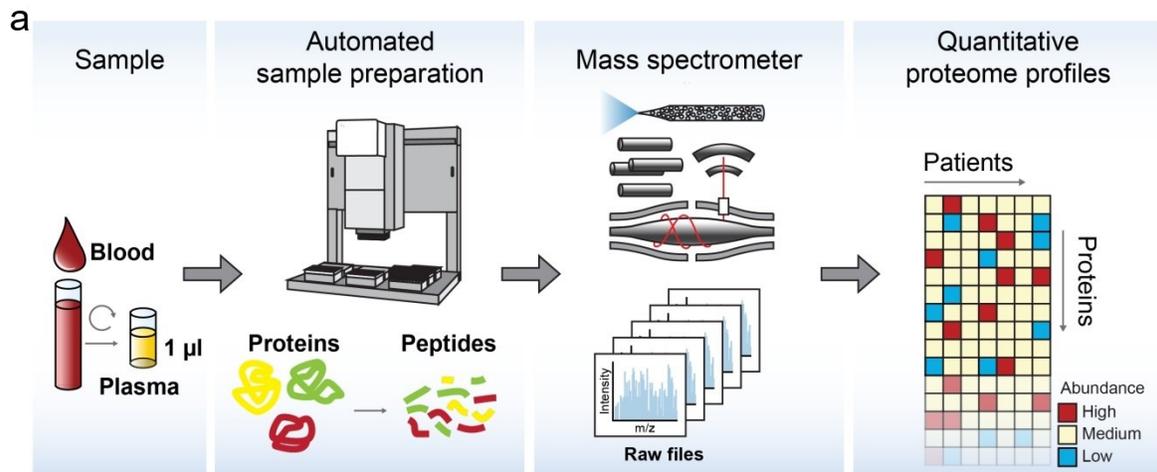
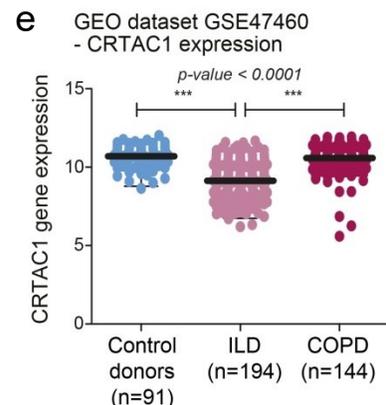
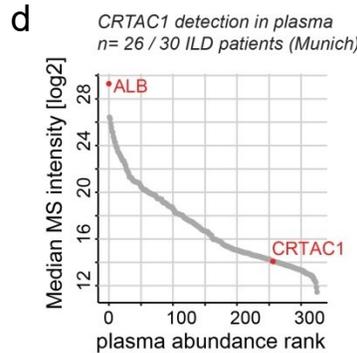
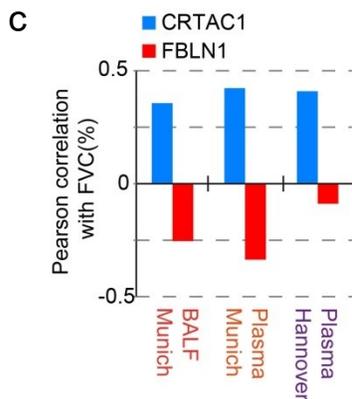


Figure 29. Multi-cohort plasma proteome analysis identifies the novel peripheral biomarker of lung health CRTAC1. (a) A high throughput experimental workflow for plasma proteomics³². (b) The indicated proteins were selected based on their common direction of correlation with patient lung function in two independent patient cohorts with distinct clinical characteristics. (c) The bar graph shows Pearson correlation coefficients of the indicated proteins with the lung function parameter forced vital capacity (FVC %) in the bronchoalveolar lavage fluid proteome cohort (BALF Munich), and two plasma proteome cohorts (plasma Munich, plasma Hannover). (d) All proteins quantified in plasma ranked by their abundance measured by mass spectrometry (MS-intensity). (e) Relative gene expression levels of *CRTAC1* in GSE47460. Dots represent average expression in tissue of individual patients. The line represents the mean. *CRTAC1* is significantly downregulated in ILD but not COPD patients (One-way ANOVA). Figure and legend taken from (Mayr et al. 2020).



4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

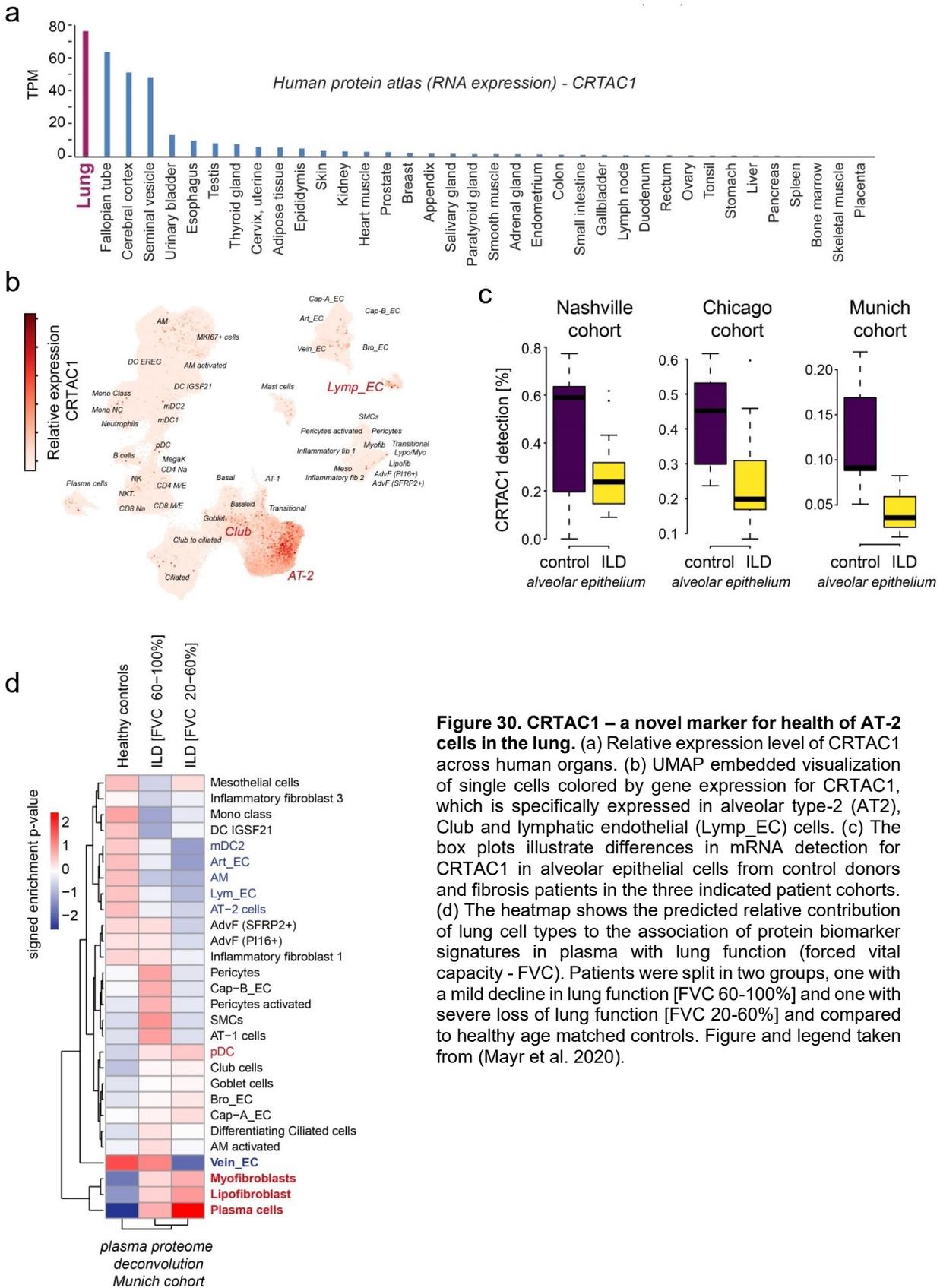


Figure 30. CRTAC1 – a novel marker for health of AT-2 cells in the lung. (a) Relative expression level of CRTAC1 across human organs. (b) UMAP embedded visualization of single cells colored by gene expression for CRTAC1, which is specifically expressed in alveolar type-2 (AT2), Club and lymphatic endothelial (Lymp_EC) cells. (c) The box plots illustrate differences in mRNA detection for CRTAC1 in alveolar epithelial cells from control donors and fibrosis patients in the three indicated patient cohorts. (d) The heatmap shows the predicted relative contribution of lung cell types to the association of protein biomarker signatures in plasma with lung function (forced vital capacity - FVC). Patients were split in two groups, one with a mild decline in lung function [FVC 60-100%] and one with severe loss of lung function [FVC 20-60%] and compared to healthy age matched controls. Figure and legend taken from (Mayr et al. 2020).

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

Total tissue RNA-seq analysis of CRTAC1 (Human Protein Atlas, (Uhlen et al. 2015)) showed the strongest expression in lung compared to all other organs (Fig. 30a). To identify the so far uncharacterized cellular source of CRTAC1 protein expression, the single cell data revealed specific expression in lung lymphatic endothelium, airway club cells, and most prominently in alveolar type-2 epithelial cells (Fig. 30b). Expression of CRTAC1 in alveolar epithelial cells was consistently downregulated in ILD samples compared to controls in all three patient cohorts analyzed by single cell RNA-seq (Fig. 30c). Therefore, CRTAC1 was found as a particular robust example for a marker that reports lung health, which can be monitored with mass spectrometry workflows from easy accessible patient blood.

Finally, the relative contribution of cell types to the protein biomarker signatures found in plasma, which correlated with lung function, was analyzed (Fig. 30d). Therefore, the patients were divided into two groups representing mild (FVC 60-100%) and severe disease (FVC 20-60%) based on lung function (FVC %) and compared these two groups with healthy controls (no FVC % data available). Interestingly, a gradual increase of proteins potentially derived from lung fibroblast subsets, plasma cells and plasmacytoid dendritic cells (pDC) was observed, which is in agreement with the scRNAseq and BALF data, highlighting the pivotal role of fibroblast in fibrotic diseases. Similarly a gradual reduction of proteins potentially derived from lung endothelial cells, alveolar macrophages, AT-2 cells, and mDC2 was found, demonstrating that lung damage can be seen in blood plasma (Fig. 30d).

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

4.1.6 Discussion

With the rise of single cell genomics and its fast evolution, making cell atlases for almost every organ available to everyone, the field is now moving on, towards the mechanistic characterization of pathogenesis and disease progression. Therefore, larger cohorts can help the concept, that variance between individuals within patient cohorts can help to model the diseases on a cellular level. Patients at different stages of a disease progression trajectory will have their cells and tissues in different characteristic states. Body fluids of those patients potentially contain a composite representation of these disease stages with specific differences in the form of proteins and possibly cell free DNA. To be able to make predictions about cell and tissue level changes in the patient, and help with diagnosis, these composite signatures need to be deconvoluted. Ultimately, training of machine learning algorithms with large datasets of matched single cell genomic and fluid proteomic or sequencing readouts will enable new automated tools for clinical decision making (Walsh et al. 2019) and drug monitoring (Maher et al. 2019) thereby and strengthen personal medicine approaches (Mayr et al. 2020).

Along those lines, the presented work describes the integration of multiple lung fibrosis scRNAseq datasets, gaining statistical power, to resolve cellular origins of predictive protein signatures in body fluids of large ILD patient cohorts, gained from mass spectrometry-based proteomics.

As introduced earlier, injuries to alveolar epithelial cells (AEC) paired with an aberrant repair response, cause pathological interactions of AEC with fibroblasts and subsequent accumulation of scar tissue, leading to pulmonary fibrosis in ILD patients. In this study, correlation of ELF protein abundances with levels of LDH in patient plasma identified a human lung injury signature that could be verified in the bleomycin mouse model of acute lung injury. The protein injury signature contained known biomarkers of IPF, such as MMP7 and CCL18, which correlated with lung function decline (Mayr et al. 2020). The injury signature furthermore featured regulated proteins, derived from various cell types: fibroblast subsets, plasma cells and mesothelial cells were increased, whereas proteins derived from several other cell types including alveolar macrophages, airway and alveolar epithelial and endothelial cells showed reduced levels (Mayr et al. 2020). Thereby, the aim using the human lung IPF cell atlas to deconvolute the bulk protein data was successfully reached.

CRTAC1 was the most robustly detected protein across all modalities, with its function in the lung unknown. CRTAC1 levels in plasma and BALF correlated not only with lung function but also pLDH, inferring an injury status, and the highest expression in the human body could be traced back to the AT-2 cells in the lung (Mayr et al. 2020). Interestingly, others have shown that in isolated human AT-2 cells, the levels of CRTAC1 were increased upon differentiation with glucocorticoids (Ballard et al. 2010), which are essential for alveolar maturation in lung

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

development (Gerber 2015). The function remains obscure but with the downregulation of CRTAC1 in AT-2 cells of ILD patients, the data might hint at uncharacterized changes in glucocorticoid signaling upon injury in these cells (Mayr et al. 2020).

ILD patients experience highly diverse clinical courses, with progression often accelerated due to acute exacerbations (AE), associated with a high mortality (Collard et al. 2016). As introduced earlier, efforts have been made to find predictive biomarkers for AE and disease outcome (Collard et al. 2010, Neighbors et al. 2018). For instance elevated serum levels of AT-2 derived surfactant protein A (SP-A) and D (SP-D) are associated with an increased risk of mortality in IPF (Greene et al. 2002, Kinder et al. 2009). Nevertheless, such biomarkers are currently not clinically established and often it is unclear which cellular changes they represent, hindering more precise treatment. While early detection of AE is currently of major interest, there are also patients who present with clinical worsening without meeting the criteria for AE. A daily home spirometry study resulted in highly diverse lung function trajectories in IPF (Russell et al. 2016), suggesting that lung function diversity could also reflect different stages after epithelial lung injury, that could be monitored with BALF. Thereby, phases of decreased lung function (potentially being a phase of subclinical injury/exacerbation) would be followed by phases of slightly increased lung function (potentially being a phase of successful tissue repair). Therefore, the human lung injury signature discovered in this study can help in the design of follow up studies, to evaluate the use of the signature in monitoring acute or subclinical exacerbations of ILD patients (Mayr et al. 2020).

Unexpectedly, and not defined as an aim beforehand, this study presents a surprisingly large set of changes between ex-smokers and never smokers. Notably, some of those changes can still be found even after decades of smoking cessation, suggesting permanent epigenetic alterations. Indeed, in genome-wide DNA methylation studies, signatures of smoking induced changes have been shown to persist over many years (Joehanes et al. 2016). Deconvolution of BALF proteomes with cell type signatures, showed a strong increase in contributions from airway club and goblet cells as well as ciliated cells in active smokers, which was most prominent in smoking associated RB-ILD patients. Interestingly, the increase in club and goblet cell derived proteins, which could stem from smoking induced goblet cell hyperplasia (Duclos et al. 2019), was fully reversible upon smoking cessation. The number of proteins derived from ciliated cells remained upregulated in ex-smokers, though. This indicates a permanent shift in airway epithelial cell composition in ex-smokers. Top hits in the differential gene expression analysis, between smokers and non-smokers in the three single cell cohorts, also showed a bias towards airway epithelial cells. However, as expected also alveolar macrophages (AM) were severely affected. For instance, a negative correlation of AM derived CCL18 with lung function was observed accompanied by an irreversible downregulation of the protein in BALF, once the patient ever smoked. CCL18 has been reported to be elevated in the serum and BALF in patients with interstitial lung disease (Prasse et al. 2007). A pooled post-hoc analysis of the CAPACITY and ASCEND studies identified CCL-18 as the most

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

robust blood marker for disease progression in IPF (Neighbors et al. 2018). Given the new data in this study, that CCL18 levels are permanently reduced in ex-smokers, and with the majority of IPF patients having a smoking history, it will be necessary to re-evaluate previous reports on association of CCL18 with IPF progression, with respect to the smoking status (Mayr et al. 2020).

Also the predefined aim of using the scRNAseq and MS data for clinical predictions was reached. The machine learning analysis shows, that correspondence of fluid proteomes and single cell transcriptomes can be used to correctly predict the direction of lung function changes across modalities. Therefore, further development of this concept will reach the stage where it can contribute to future clinical decision making. The presented work has several limitations that prohibit the full completion of this task, though. With the cross-modal analysis having been done on non-matched patient cohorts, it is currently difficult to assess the specificity of fluid proteome signatures for tissue level cell state changes, and to go beyond associative signatures. Thus, carefully designed longitudinal multi-modal analysis of animal models and patient cohorts will be required to train machine learning algorithms, in particular causal inference models, for future applications in predictive personalized medicine (Mayr et al. 2020).

Regarding the aims of this study it should be noted that the integration of multiple single cell data sets was an important step. Previously published single cell datasets on fibrosis, also already comparing health and disease, like the Chicago or Nashville cohort alone, consisted of a small number of samples. Taking into account the heterogeneity of patients and diseases trajectories, this could have an impact on conclusion drawn out of the analysis. Not only was the statistical power missing, both in sample as well in cell number to perform significant analysis, but also cell type annotations are different, with no regulations in place, for defining names for cell states. In the worst case scenario this can lead to different names for the same cell type, rendering comparison between studies useless. The successful integration of datasets in this study therefore provides a valuable tool for the community. As for limitations of the study, small differences that can still be observed between cohorts, most likely stem from different single cell sampling techniques. This includes not only differences in microfluidic platforms to capture the mRNA from single cells, where the Chicago and Nashville cells were sampled with the Chromium platform from 10x Genomics, while Drop-seq was used in Munich. Also alternating compositions of enzyme mixes used for tissue digestion to release cells and create single cell suspensions can account for differences. This can be the reason for varying cell type frequencies between the studies, as seen in Fig 13.

The aim of the proteomic analysis of BALF and plasma was also reached, demonstrating however the still existing limitations in the sensitivity of MS measurement. While the BALF data provided significant insights in the smoking changes, the proteome at the sensitivity of the analysis did not

4.1 Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids

reflect overall clinical diagnosis. Rather it might have reflected molecular and cellular alterations that are a consequence of disease progression or acute illness. The MS analysis of two plasma cohorts from Munich and Hannover showed only a weak correlation between cohorts. This can be attributed to the difference in sample source, with the Munich blood sampled in plasma tubes, while in Hannover serum was extracted. However, finding lung specific proteins in blood plasma, that correlate with clinical data sampled because of lung injury, is an important milestone, when keeping in mind that blood is the foremost body fluid that is circulating in the whole human body and could pick up changes in every organ.

4.2 Mesenchymal heterogeneity in the lung

The work presented in this chapter summarizes the findings in this project up to the completion of the thesis. The project is still ongoing to delineate the heterogeneity of stromal cell subtypes and their specific localization first at the healthy baseline, as well as their changes and roles in injury and repair in the bleomycin mouse model of lung fibrosis and how this compares to the human disease.

The PhD student, was involved in all parts of this study, performed most of the practical work, executed programming code and drafted all the figure panels. Credit goes to Paulina Ogar, for help in setting up the SCRINSHOT technique and analyzing the *in situ* localization data as part of her master thesis under supervision of the PhD candidate, Meshal Ansari for support in single cell analysis, Dr. Maximilian Strunz and Ilias Angelidis for help with mouse work and with collecting the mouse single cell samples and to Dr. Andreas Liotons for introduction to the SCRINSHOT technique as part of a lab exchange for two weeks in Stockholm.

4.2.1 Introduction

Epithelial cells receive multiple signals from the surrounding mesenchymal niche, which controls homeostasis and activation of stem cells upon injury. Along the airway tree and the respiratory unit numerous specialized epithelial cell types are surrounded by a poorly defined heterogeneous mesenchyme. The mesenchyme in the lung consist of several distinct cell lineages with major functions in the respiratory system, including poorly defined populations of different fibroblast subtypes, stromal cells such as pericytes, endothelial cells, mesothelial cells, airway smooth muscle cells (ASMCs) and vascular smooth muscle cells (VSMCs). Some of these cell lineages are defined by common and known markers, such as platelet endothelial cell adhesion molecule/CD31 (PECAM1) for endothelial cells, platelet-derived growth factor receptor- β (PDGFRB) for pericytes, α -2-smooth muscle actin (ACTA2) in pan-smooth muscle cells or mesothelin (MSLN) in mesothelial cells (Danopoulos et al. 2020). However, markers to identify each cell type are missing or nonspecific for the majority. This is especially the case for fibroblasts in general, and also for VSMCs in comparison to ASMCs, where there are no definitive markers allowing to distinguish between them (Danopoulos et al. 2020). Thus, defining cellular identity, and delineating which specific markers characterize these specific subpopulations, have been significant hurdles in understanding the role and function of certain mesenchymal cells in the lung (Sucre et al. 2020). Along with the unknown function, also the identification and nomenclature of subpopulations of mesenchymal cells is poorly defined yet and has been based largely on the phenotype or location of these cells, if known e.g. for lipofibroblasts and myofibroblasts (Sucre et al. 2020).

4.2 Mesenchymal heterogeneity in the lung

scRNA-seq again comes to the rescue to answer these fundamental questions of cellular identity and heterogeneity, and to generate new hypotheses about lung homeostasis and disease pathogenesis. Use of scRNA-seq allows not only for the analysis of mesenchymal subtypes at single cell resolution, but also within the context of transcriptomic differences between them.

Along with the rise of scRNA-seq techniques, novel advancements in spatially resolved transcriptomics have developed at a similarly fast pace. These technologies are centered on providing omics data from specimens while keeping positional information (Crosetto et al. 2015). Among those, spatially resolving methods are available that evolved from single-molecule fluorescence *in situ* hybridization (smFISH) which has already been widely used to detect single RNA molecules (Femino et al. 1998). Newer iterations allow multiplexing of several colors and up to 3000 cells and 80 genes at once (Lee et al. 2015, Nagendran et al. 2018, Sountoulidis et al. 2020). Therefore, not only newly found marker gene signatures can be validated in a histological context in the tissue *in situ*, but the location of a cell can also give an indication for its function, which is currently an urgent need in the mesenchymal compartment (Stuart et al. 2019).

While the crosstalk of epithelium and mesenchyme is important during hemostasis, it becomes even more critical during injury and during repair. As introduced earlier, factors such as Tgf β -1 or sonic hedgehog (Shh) signaling, coming from the mesenchyme, are essential to promote AT-2 cell self-renewal and differentiation into AT-1 cells, with errors during this process or repeated injury leading to fibrosis (Barkauskas et al. 2013, Peng et al. 2015). Different epithelial – mesenchymal interactions are likely to occur during this process, however identity and precise location of these different niches are unknown. Consensus is achieved, though, over the critical role of myofibroblast during fibrotic injury in human, as well as, nicely reproduced in mouse model of bleomycin lung injury. Activated fibroblasts and myofibroblasts are regarded as the key cell types that lead to an increased deposition of extracellular matrix and fibroblastic foci in lungs of patients with idiopathic pulmonary fibrosis. The origin of myofibroblasts however remains elusive, with multiple sources suggested in different studies over the years (Phillips et al. 2004, Kim et al. 2006, Hoyles et al. 2011, Rock et al. 2011).

The following sections will summarize the results of the herein conducted investigations. Single cell transcriptomics on whole lung cell suspensions of healthy adult wild-type mice was performed. The cells were therefore first negatively depleted for epithelial and endothelial as well as leukocytes, to use an unbiased approach for classification of mesenchymal cell subtypes. Massively parallel scRNA-seq of mesenchymal cells identified 6 distinct cell types, including mesothelial cells, smooth muscle cells, and 5 distinct stromal cell types with features of fibroblast phenotypes. Multiplexed *in situ* FISH, using the SCRINSHOT technique, enabled the definition of distinct locations and interaction partners of these subtypes, indicating and corresponding with their diverse function. Having identified mesenchymal heterogeneity at base-line, an 18 time point

4.2 Mesenchymal heterogeneity in the lung

scRNAseq experiment with over 30 mice was performed using the bleomycin mouse model of lung fibrosis and regeneration. The analysis dissected mesenchymal cell types during injury at single cell level and identified activated states with a changed mRNA expression profile for all mesenchymal subtypes upon injury. Importantly, the scRNA-seq data revealed a clear progenitor of myofibroblasts among the cell types and allowed to define myofibroblasts by an extensive gene signature. The comprehensive analysis of the mouse lung mesenchymal compartment at baseline and in lung injury was finally compared to the situation in the human that was introduced in the first chapter, to identify common and species specific changes.

4.2.1 Fibroblast heterogeneity in the adult wild-type mouse lung

Cell type diversity in the mesenchymal compartment of the mouse lung was studied using the Drop-seq platform for single cell sequencing (Macosko et al. 2015). Whole lung parenchyma cell suspension was obtained from two adult (p70) wild-type C57BL/6 mice. Negative selection was performed using MACS based on the indicated marker proteins, to remove endothelial (Pecam1), epithelial (Epcam), hematopoietic (Cd45), lymphatic (Lyve1) and erythrocyte (Ter119) cell populations. Approximately 2500 cells per sample were sequenced, which finally, after filtering (mesothelial cells were taken out of the data to focus on fibroblasts) and quality control, resulted in a data set containing 1267 cells (Fig. 31a). The single cell dataset was then reduced in the data dimensionality, using the UMAP method (Becht et al. 2018). The UMAP was displayed with a color code for the two replicates, and assessment of overlaying of the cells, and all clusters consisting of both replicates, both proved good accordance (Fig. 31d). All cells correctly displayed expression of mesenchymal markers Vimentin (Vim) and Collagen 1 alpha chain 2 (Col1a2), marking them as mesenchymal cells (Fig. 31c). Unsupervised clustering and Partition-based graph abstraction (PAGA) of the data revealed 6 cell clusters and their connectivity (Fig. 31f). Next the cell state identities were derived by manual annotation of the clusters with known canonical marker genes and signatures from published single cell RNA sequencing studies (Angelidis et al. 2019, Strunz et al. 2019).

And indeed, with lipofibroblasts, peribronchiolar fibroblasts, adventitial fibroblasts, pericytes and smooth muscle cells (SMCs) and the new cell type Hhip+ fibroblasts, six mesenchymal cell types could be revealed (Fig. 31b,f). Notably, already by observation of the location of the clusters as nodes in the PAGA map, and their intensity of connectivity, it became obvious, that some were more similar to each other than others, such as the SMCs and pericytes, both showing high alpha smooth muscle expression (Acta2) (Fig. 31e). Interestingly, Hhip-cells, defined by their characteristic expression of the hedgehog interacting gene (Hhip), of which the function is unknown seem to connect the lipofibroblasts with the more Acta2 expressing cells.

4.2 Mesenchymal heterogeneity in the lung

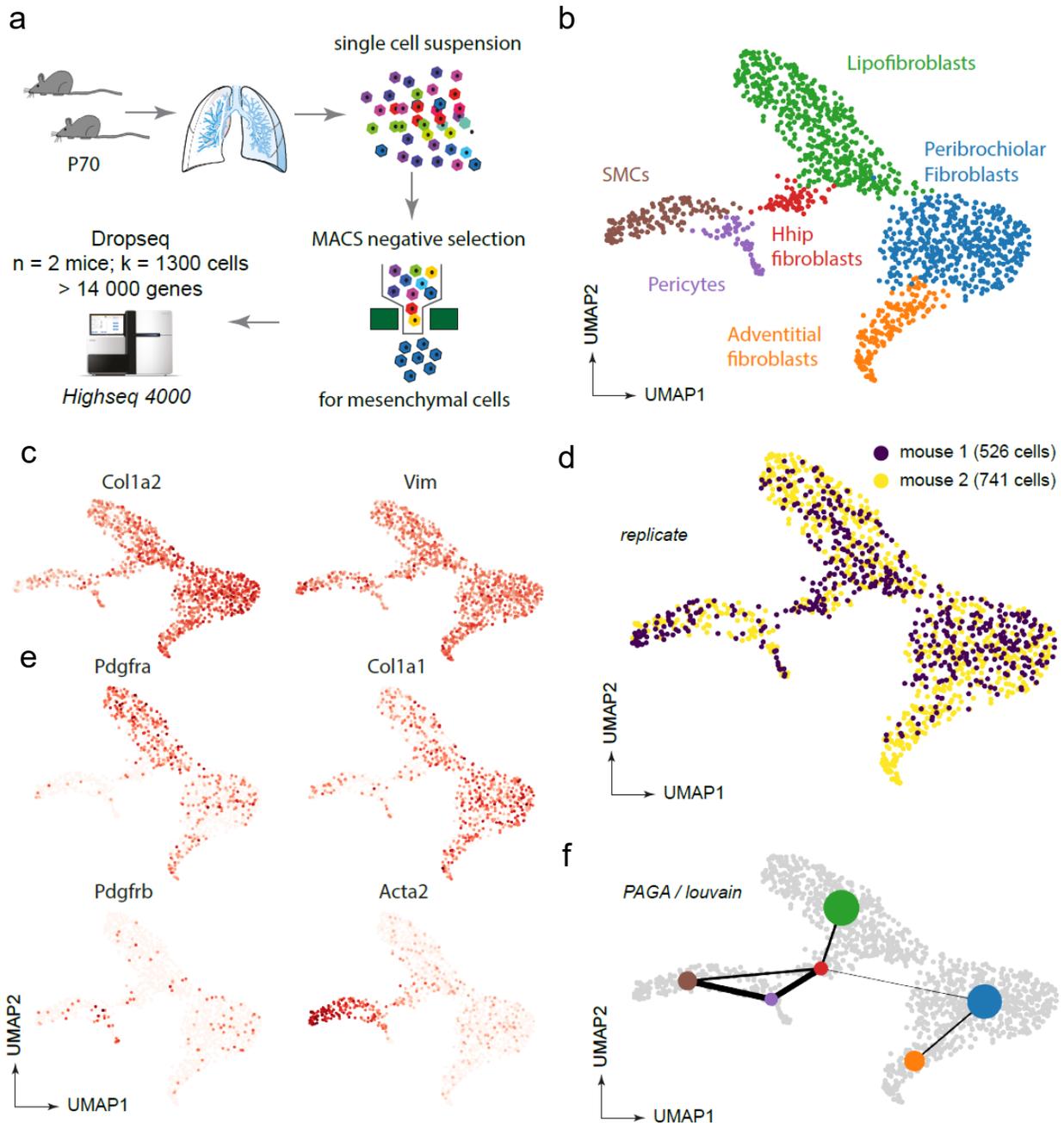
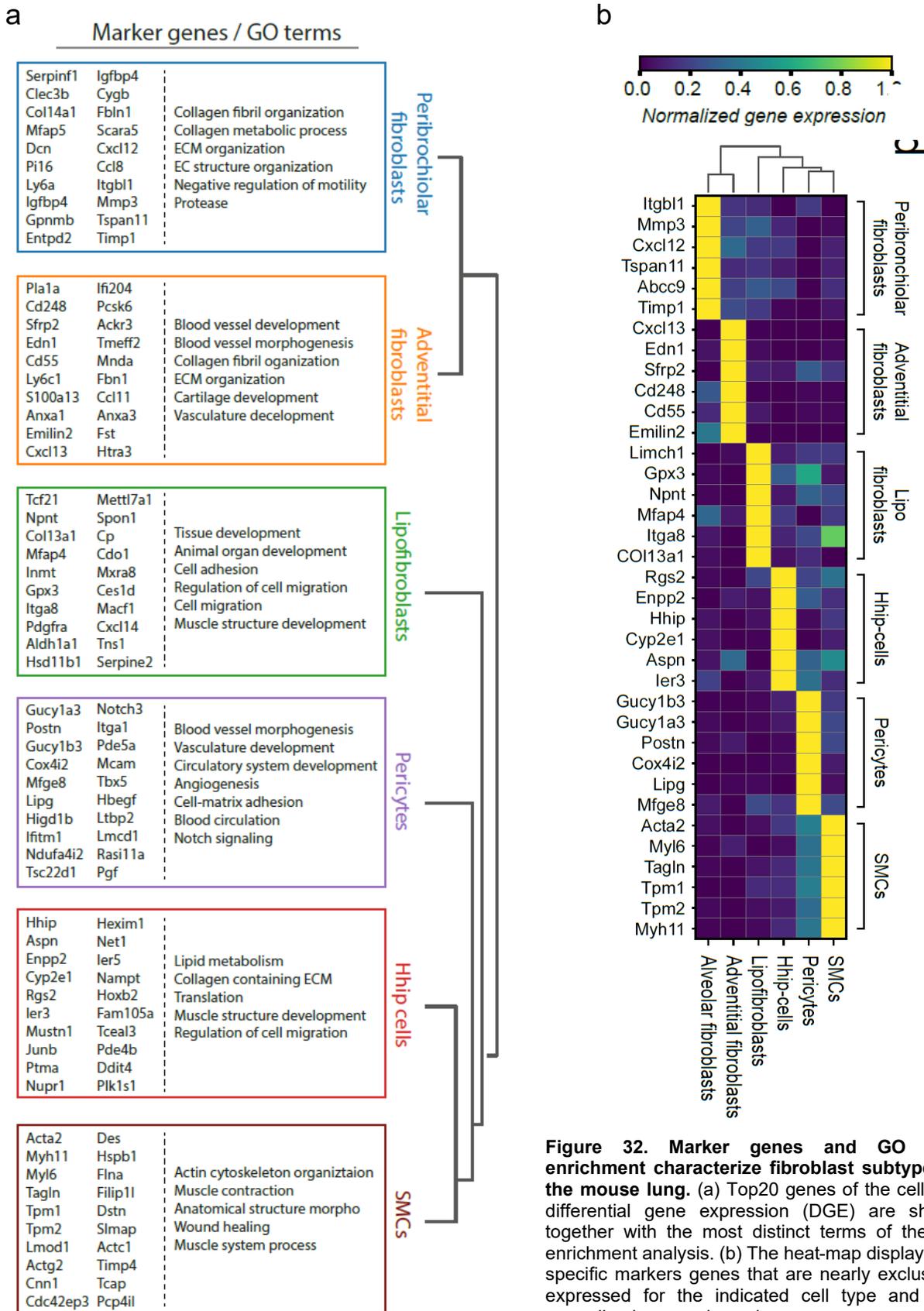


Figure 31. Single cell analysis of mouse lung mesenchymal compartment reveals 6 distinct cell type identities. (a) scRNAseq experimental design. (b-e) Dimension reduced single cell transcriptomic data is visualized through Uniform Manifold Approximation and Projection (UMAP). The color code illustrates the cell types (a), replicates (d), or marker genes (c,e). (f) Partition-based graph abstraction (PAGA) of the data shows 6 cell clusters as nodes and their connectivity/commonness as thickness of the edges.

4.2 Mesenchymal heterogeneity in the lung



4.2 Mesenchymal heterogeneity in the lung

Platelet-derived growth factor receptor A and B (Pdgfra and Pdgfrb) expression among the cells, both often used as generic marker for fibroblasts together with Col1a1, was rather distinct and more pronounced in certain cell types, demonstrating nicely the need for better sub type markers (Fig. 31e). All fibroblast cell type identities were characterized by unique marker gene expression profiles, after performing differential gene expression (DGE) between and against all clusters. While the heatmap provides five genes per cell type and proves their specific expression, the dendrogram shows the Top20 genes of the DGE, providing a useful tool for future cell type based analysis (Fig. 32a,b). Additionally, gene ontology (GO) enrichment analysis (Consortium 2019, Mi et al. 2019) was performed on the fibroblast subsets. SMCs were rightfully characterized by the terms “muscle contraction” and “wound healing”, while pericytes were marked by “blood circulation” and “vasculature development”. Lipofibroblast contained more general terms such as “tissue development” or “cell adhesion”, while the Hhip+ cells, so far uncharacterized seemed to have the best of both worlds with “muscle structure development” and “collagen containing ECM”. Interestingly, adventitial fibroblasts were marked by terms like “blood vessel morphogenesis” or “vasculature development”, emphasizing their function as cells, providing structure and ECM around the vessels. Terms such as “ECM organization” and “collagen metabolic process”, hinting at clear matrix protein production, were mainly found in the peribronchiolar fibroblasts (Fig 32a).

Fibroblasts are known to be the major source of extracellular matrix and structural components in the tissue. Therefore, the fibroblast cell types were analyzed for their specific protein expression that can be attributed to the “Matrisome” in the mouse lung, as the ensemble of ECM proteins and associated factors, as defined by MS-based proteomics (Naba et al. 2012) (Fig. 33). While Hhip+ cells show nearly no distinct Matrisome protein expression, pericytes and SMCs share a few. Lipofibroblasts, with 25 proteins, express the second most specific proteins, which is in concordance with their appearance as cluster with the second most cells of all fibroblast types in the mesenchyme of mouse lungs. Far the most distinct Matrisome proteins are expressed by the peribronchiolar and adventitial fibroblasts, that were since put together into one group termed “Matrix fibroblasts” for further analysis. Over 70 specific ECM proteins, together with the GO term of “ECM organization” suggest a major role in structural integrity for the matrix fibroblast, be it around the vessels or in the alveolar space.

In summary, an in-depth analysis of the mesenchymal compartment in the adult and healthy mouse lung was generated on a single cell level. UMAP clustering revealed six fibroblast subtypes, among them a new cell type of Hhip+ cells. Differential gene expression identified specific marker genes, especially for adventitial and peribronchiolar fibroblasts and GO term enrichment analysis characterized each cell type. Comparison of Matrisome proteins, revealed nearly exclusive expression of ECM protein by each fibroblast subtype, indicating their specialized individual role and function.

4.2 Mesenchymal heterogeneity in the lung

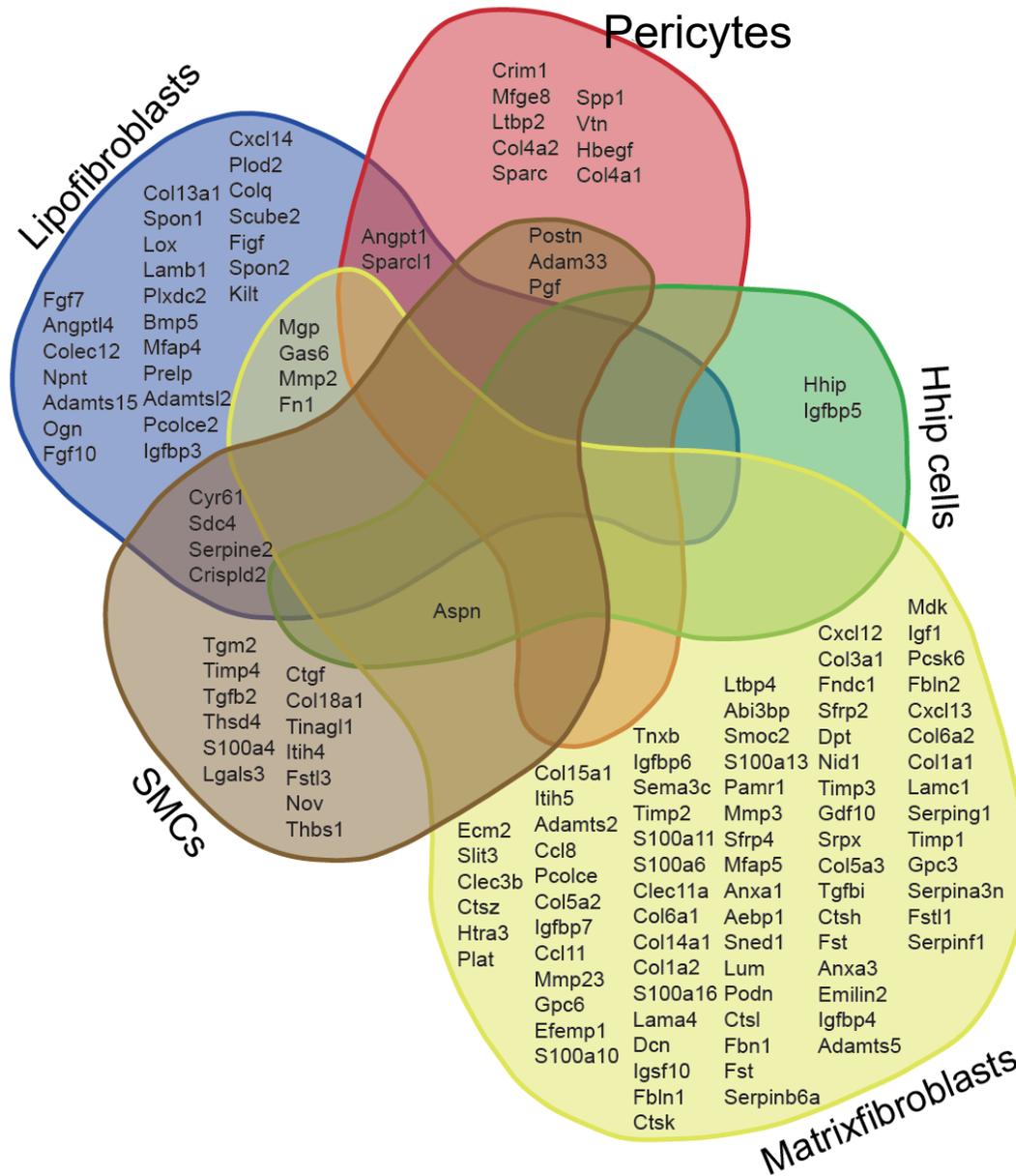


Figure 33. Fibroblasts subtypes in the mouse lung are characterized by specific expression of Matrisome proteins. The top200 proteins per cluster, out of the differential gene expression between cell types, were used to extract gene names. Genes were filtered for an entry in the “Matrisome” list. Matrisome protein gene names were applied to a Venn diagram analysis. [Matrix fibroblasts is a merged cell type annotation with clusters of alveolar and adventitial fibroblasts combined]

4.2 Mesenchymal heterogeneity in the lung

4.2.2 Spatial mapping of mouse lung fibroblast subtypes

While scRNA-seq enables transcriptome analysis of individual cells and the identification of new cellular states, these methods fail to capture the spatial organization in the organ due to tissue dissociation. Cell types may display similar gene expression patterns, but be located in different spatial positions or the location of a cell can simply give an indication to its function. Therefore the combination of scRNA-seq methods and new spatial coordinate mapping methods are a valuable match (Stuart et al. 2019).

To localize fibroblasts in the tissue context, a recently published method for sensitive multiplexed RNA mapping, called SCRINSHOT (Single Cell Resolution IN Situ Hybridization On Tissues) was adopted (Sountoulidis et al. 2020). In brief, direct hybridization of gene specific padlock probes on mRNA is followed by circularization and rolling circle amplification (RCA) on the hybridized padlock probes. Sequential detection of RCA-products using sequence specific fluorophore-labeled oligonucleotides, enables the visualization of up to thousand cells in frozen tissue sections in multiple areas per section (Fig. 34a,b). Automated image analysis allows to transfer the fluorescence info into pixel data, allowing for easy quantification and analysis of the data (Fig. 34c) (Sountoulidis et al. 2020).

For analysis of fibroblast subtypes in the adult wild type mouse lung as baseline, SCRINSHOT was performed on 13 regions of a right mouse lobe, spanning all areas of interest, from distal to proximal and including vessels and airways (Fig. 34b). Padlock probes, 54 in total and 3 per gene with 18 genes selected, were designed against well-known marker genes of lung cell types, such as *Sftpc* for AT-2 cells or *Scgb1a1* for secretory airway cells, and markers of fibroblast subpopulations. Those markers, two per subpopulation, were taken from the DGE of the scRNAseq data of wild type mouse, focusing on high specificity for one subtype, as well as on a certain level of expression, to allow for sensitive detection by SCRINSHOT (Fig 35, scRNAseq feature plots). The genes used as markers to detect the other different cell types, were *Postn* and *Gucy1a3* for pericytes, *Hhip* and *Aspn* for Hhip-cells, *Tcf21* and *Npnt* for lipofibroblasts, *Serpinf1* and *Clec3b* for general matrixfibroblasts and *Acta2* and *Myh11* for smooth muscle cells (Fig. 35). While AT-2 cells were easily visible due to the high expression and therefore good detection of *Sftpc* by the SCRINSHOT technique, the number of fluorescent dots was lower for fibroblast marker genes. Therefore, a threshold was set and a cell was only determined to belong to one subpopulation, if both of the marker genes were expressed, with three dots in total to be observed in one cell (Fig. 35, zoomed inlets). The cells were hereby defined by DAPI in all iterations of SCRINSHOT multiplexing rounds, with the circumference of the DAPI-marked nucleus enlarged for the analysis by the software to fit the size of a cell.

4.2 Mesenchymal heterogeneity in the lung

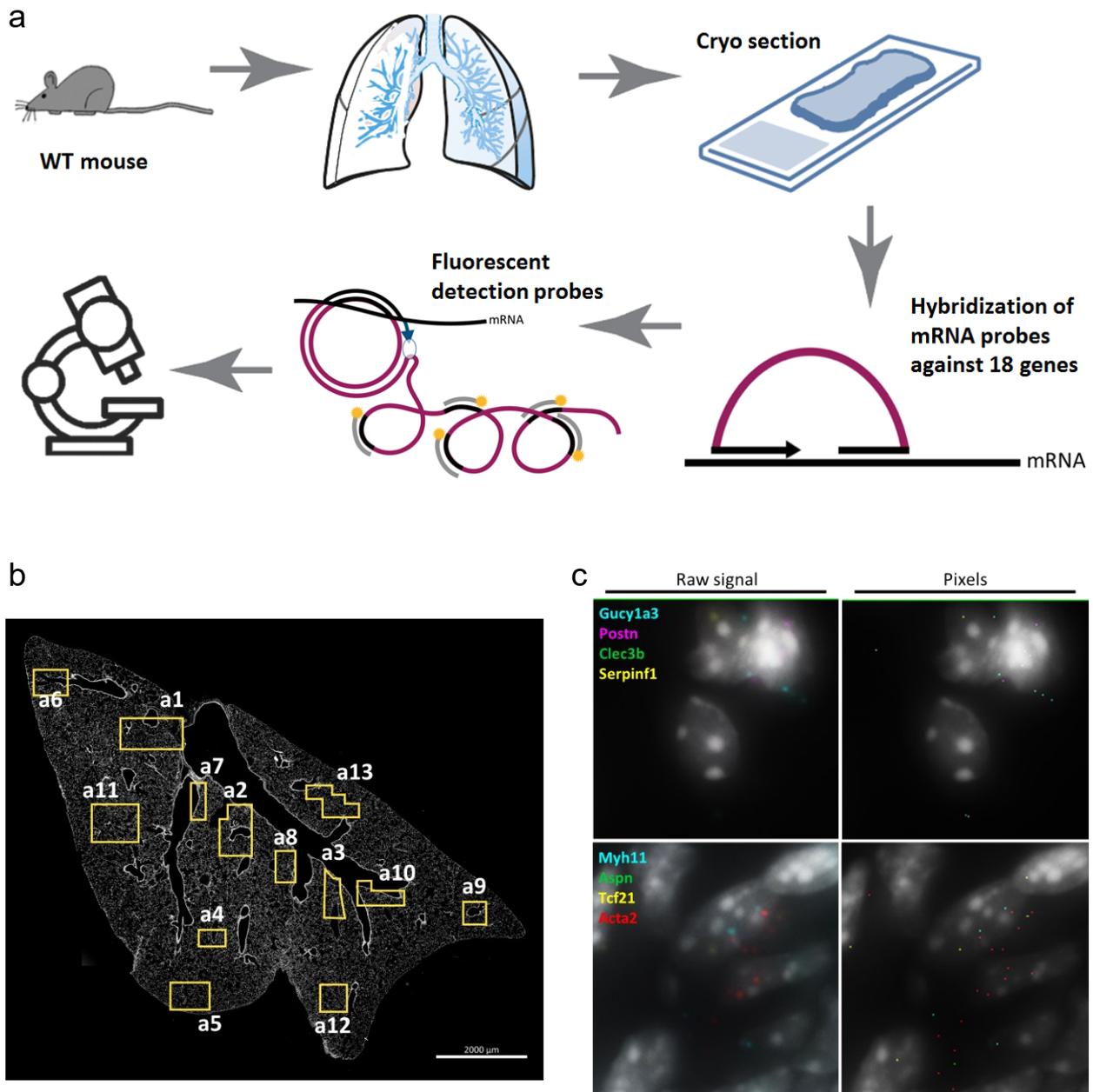


Figure 34. SCRINSHOT, a multiplexed sensitive RNA mapping approach. (a) SCRINSHOT workflow: Mouse lung lobes cryo-sections were used for hybridization of padlock probe to mRNA. After ligation, rolling circle amplification (RCA) of the hybridized padlock ensures multiplication of fluorescent probes binding sites. RCA products are visualized by hybridization of 3'-fluorophore-labelled detection oligos. (b) All 12 areas on one mouse lung lobe that were imaged by SCRINSHOT. (c) Example picture of how the raw signal looks like when imaged and after transformation into pixel data that can then be analyzed in an automated way.

4.2 Mesenchymal heterogeneity in the lung

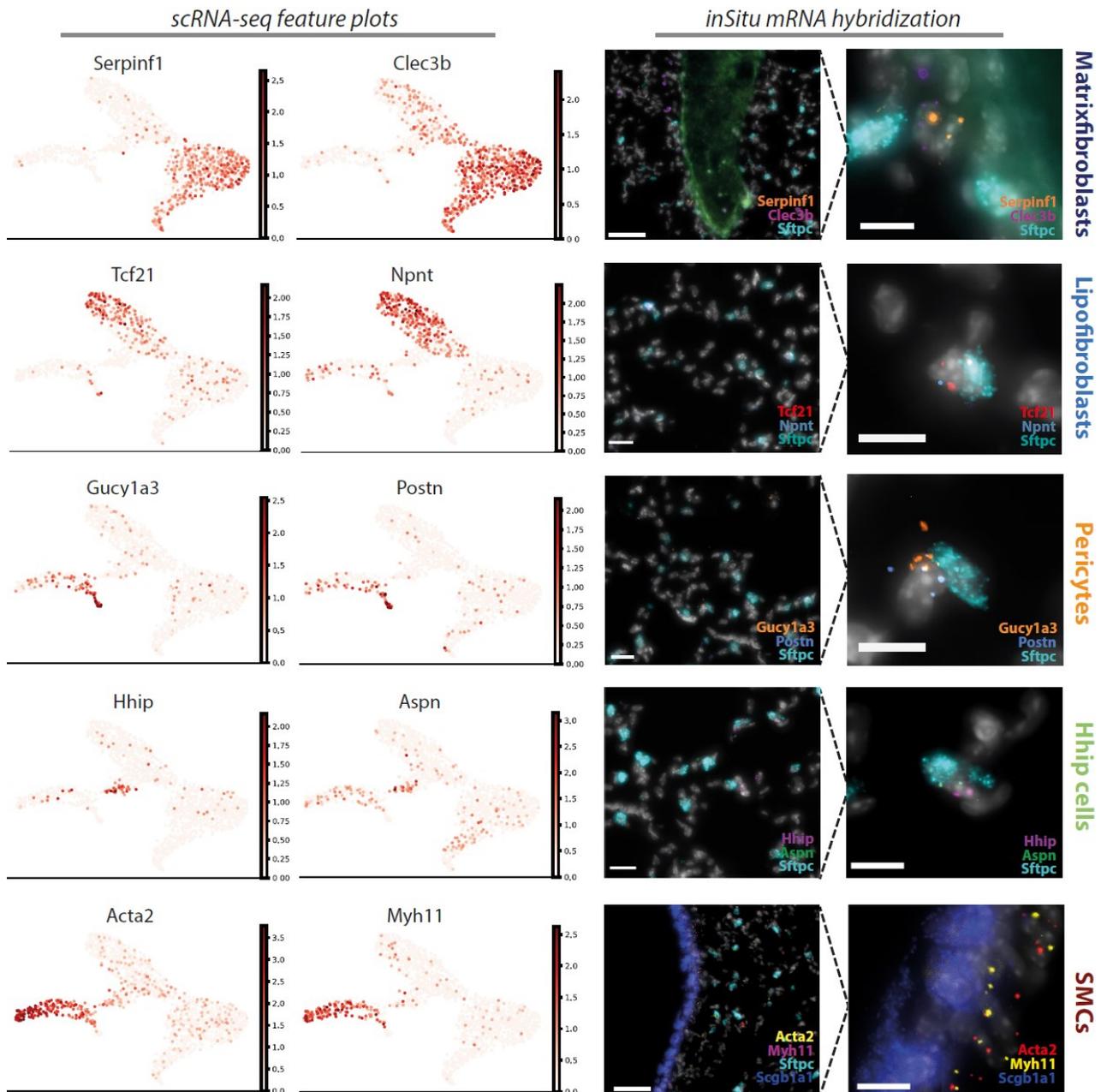


Figure 35. Marker genes for SCRINSHOT probe selection. scRNA-seq feature plots show expression of marker genes in the wild type adult mouse scRNA data. Two marker genes per fibroblast subtype were chosen, based on specific expression in one cluster and high expression in that cluster. *inSitu* mRNA hybridization pictures how SCRINSHOT raw image data of the selected marker genes. Zoom-in inlets show hybridization dots in cell shape. DAPI nucleus single in grey.

4.2 Mesenchymal heterogeneity in the lung

For Area 1, all fibroblast subtypes could be observed in the image raw data. Processing and filtering the pixel data resulted in a single cell matrix table, with marker gene expression per cell in the whole image area. Curating those cells with the threshold for cell type specific expression allowed to map back the identified cell types onto the tissue. Thereby already first observations could be made: matrix fibroblasts were rather rare, and seemed to localize more at regions close to vessels and airways. SMCs marker expressions were observed mostly near airways and also vessels, as expected by their known contractile function. Marker genes denoting lipofibroblasts and pericytes were observed quite frequently in cells in the tissue, all throughout the alveolar space, with pericytes being more in numbers. Additionally, upon initial observation, rare Hhip-cells seemed distributed throughout the tissue, not revealing their unknown function (Figure 36).

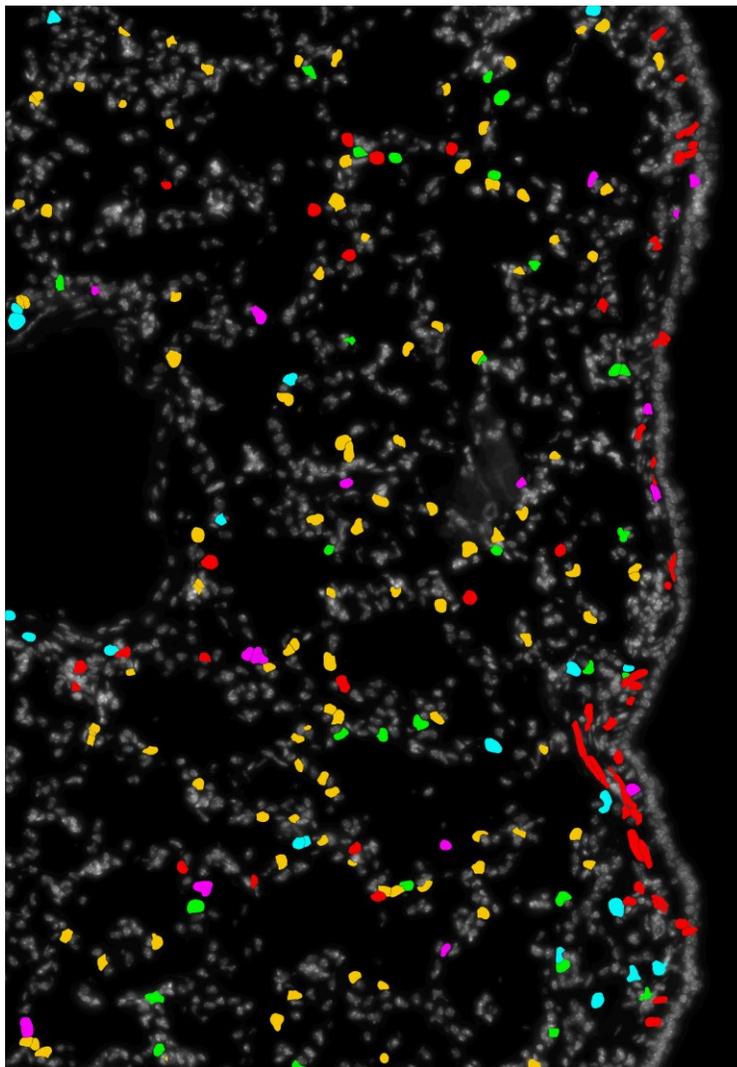


Figure 36. Fibroblast subtypes spatial distribution in the tissue. Detected SCRINSHOT signal was filtered, curated and assigned to whole cells and cell types. Those were then mapped back into a representative region to visualize fibroblast subtype frequencies and distribution in the tissue context.

4.2 Mesenchymal heterogeneity in the lung

For a more quantitative way of analyzing the images, with Area 1, Area 4, Area 5, Area 7, Area 8 and Area 12, six of the regions of the lobe that were imaged, were processed and analyzed. Thereby, a total of 2245 cells from the above cell types were identified over the regions, with 1269 AT-2 cells, 508 pericytes, 70 Hhip-cells, 206 lipofibroblasts, 41 matrix fibroblasts and 151 smooth muscle cells detected (Figure 37a). The total numbers of cell types quantified per area were additionally normalized according to the mean area size, as the imaged regions of the lung lobe section were of differing sizes, and contained different stereotypical tissue structures such as airways or vessels, as well as blank space with no cells. In general, after accounting for the area size difference, there was a similar distribution across the areas, within cell types (Fig. 37a). As expected, the most abundant cell type was AT-2 cells. The most abundant mesenchymal cell subtype was pericytes, followed by lipofibroblasts, smooth muscle cells, Hhip- cells, and matrix fibroblasts, the rarest mesenchymal cell subtype observed. Pericytes abundance is known to be high in the lung, with pericytes being closely related to vascular smooth muscle cells and covering the vasculature within the capillary basement membrane (Armulik et al. 2011). Of note, these numbers differ from those observed in the scRNAseq data. While relative frequencies and tissue extraction can play a role for the scRNA-seq data, tissue accessibility, marker gene expression as well as sensitivity of the SCRINSHOT method can influence the imaging data, giving neither method the edge for a perfect quantification.

Signaling between epithelial and mesenchymal cells in the alveolar niches plays a critical role in tissue homeostasis. To elucidate the interactions that distinct mesenchymal cell types may have with the epithelium, the frequency of the mesenchymal cell types in close proximity to AT-2 cells was quantified. If a mesenchymal cell type was observed to be near to an AT-2 cell, identified by *Sftpc* signal, as shown in Fig. 37b, the AT-2 cell was counted as having an interaction partner. In general, the percentage of a mesenchymal cell type with AT-2 partners was fairly consistent within the areas chosen. Pericytes were the cell type found most commonly in close proximity to an AT-2 cell, with an average of 55% of AT-2 cells being partnered. Following this, AT-2 cells were observed to be paired with lipofibroblasts 33% of the time. Across the areas, the average percentage of AT-2 cells paired with a Hhip-cell cell was 7%, pairing with a SMC was 3% and the lowest interaction percentage of AT-2 cells was with matrix fibroblasts at 0.6% (Fig. 37c). Lipofibroblasts, as adipocyte-like cells, are known to assist AT-2 cells in surfactant production by transfer of lipids between them, providing a reasonable explanation for the overserved high interaction rate (Schultz et al. 2002). For pericytes, the reason might simply be their high abundance, with consistently being found as the most prevalent stromal cell type across all areas, increasing the chance of interactions.

4.2 Mesenchymal heterogeneity in the lung

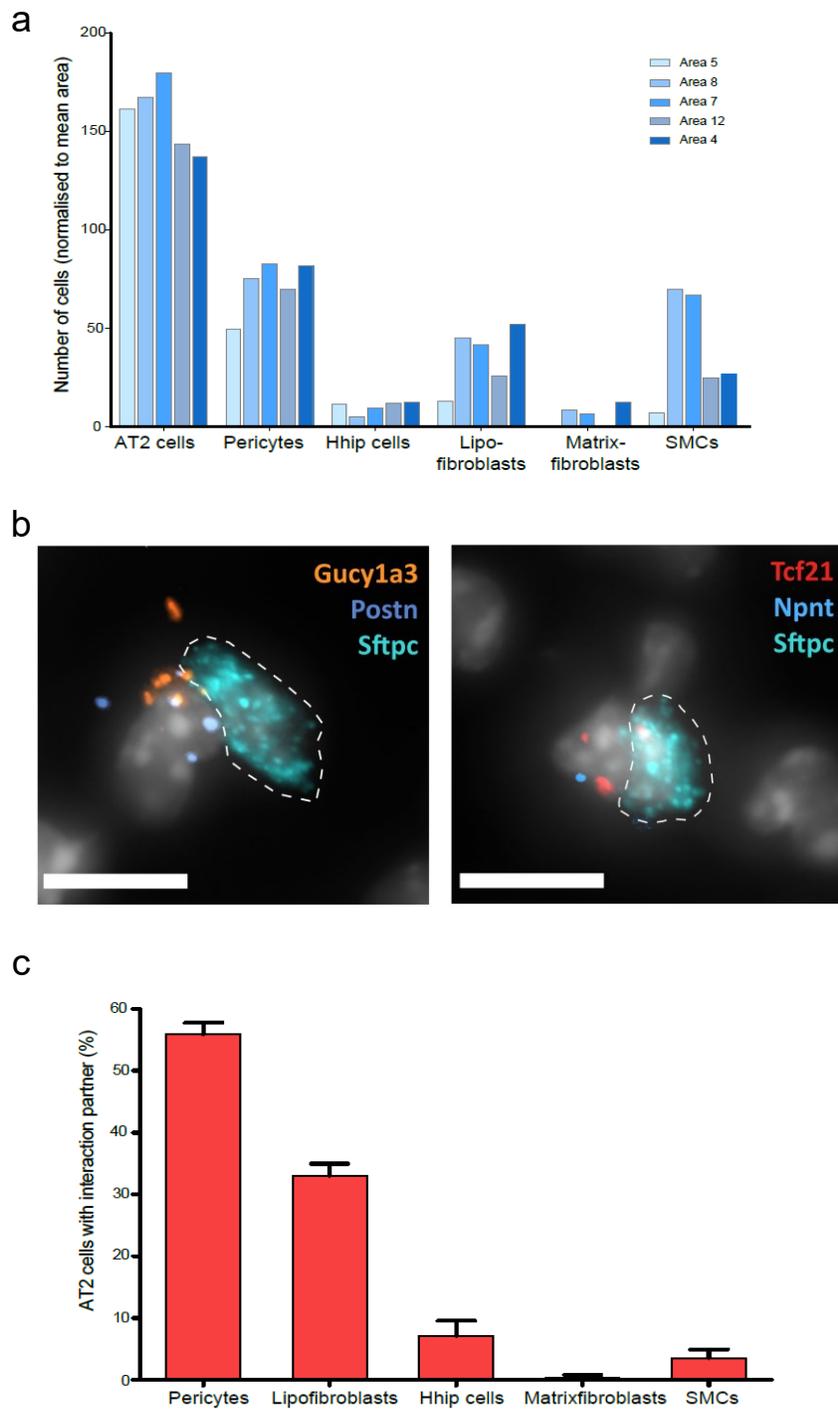


Figure 37. Fibroblast subtype prevalence and interaction with epithelium in the tissue shown by SCRINSHOT. (a) Total number of cells quantified by area. (b) The number of AT-2 cells that were observed in close proximity to fibroblasts, was analyzed by the frequency of interacting fibroblast subtype. (c) The frequency of the fibroblast cell types in close proximity to AT-2 cells was quantified and normalized to all AT-2 cells found. Analysis was performed across all six areas.

4.2 Mesenchymal heterogeneity in the lung

The spatial distributions of the mesenchymal cell types were further analyzed, by assessing the cell type occurrence compared to the total number of cells sampled in specific tissue regions, such as near airways, near vessels or in the alveolar space. This cell type density was determined by the percentage of the cell type, of all DAPI positive nuclei counted in the respective region, across all six areas. Apart from the visual conception, this quantitative analysis gave a good impression of the specific spatial heterogeneity all fibroblast subtypes (Fig. 38).

This analysis showed that pericytes had a significantly higher cell density in both the alveolar space and near vessels, at 3.65% compared to the airway regions, where the pericytes comprised of 1.45% of the total cells. The cell density of lipofibroblasts was significantly higher in the alveolar space at 1.62% compared to around airways at 0.31%. The average cell density for this cell type in the vessel regions was 1.11%, but not significantly different to either the alveolar space or airway region values. Hhip-cells were significantly more prevalent in the airway regions at 1.38%, in comparison to 0.4% in the alveolar space, and 0.1% by the vessels. The cell density of smooth muscle cells was significantly higher by the airways than in the alveolar space or near the vessels. Smooth muscle cells comprised 4.2% of total cells in the airway compared to 1.69% in vessel regions and 0.3% in the alveolar space. Lastly, matrixfibroblasts were 0.89% of the cells in the airways, 0.81% of the cells by the vessel and 0.03% of the cells in the alveolar space, though the differences here in density were not significant (Fig. 38).

The analysis split the cell types in certain groups. Pericytes and lipofibroblasts were the most prevalent fibroblasts in the alveolar space, with pericytes also being present around vessels. In contrast, matrixfibroblasts were never found in the alveolar space but were equally distributed around airways and vessels. Of note, in the scRNA-seq data, the matrixfibroblasts could be split up in in so called adventitial fibroblast, referring to a possible function in covering vessels, as inferred by the GO term enrichment analysis, and the peribronchiolar fibroblasts, alluding to a possible function around airways. While the markers, *Serpinf1* and *Clec3b* admittedly could not distinguish between those two subsets (Fig. 35), the SCRINSHOT analysis strongly confirms the scRNA-seq data, by indicating no alveolar localization of matrixfibroblasts. As expected, SMCs were found mostly around vessels and airways. Interestingly, Hhip-cells were almost never found around vessels, but significantly more around airways (Fig. 38). Referring again back to the scRNA-seq data, Hhip-cells in the UMAP space were placed in close proximity to SMCs and pericytes, suggest similar function. Potentially, the SCRINSHOT localization could thus indicate newly discovered cell type of Hhip-cells as the long missed airway smooth muscle cells (ASMCs), but further analysis is required.

4.2 Mesenchymal heterogeneity in the lung

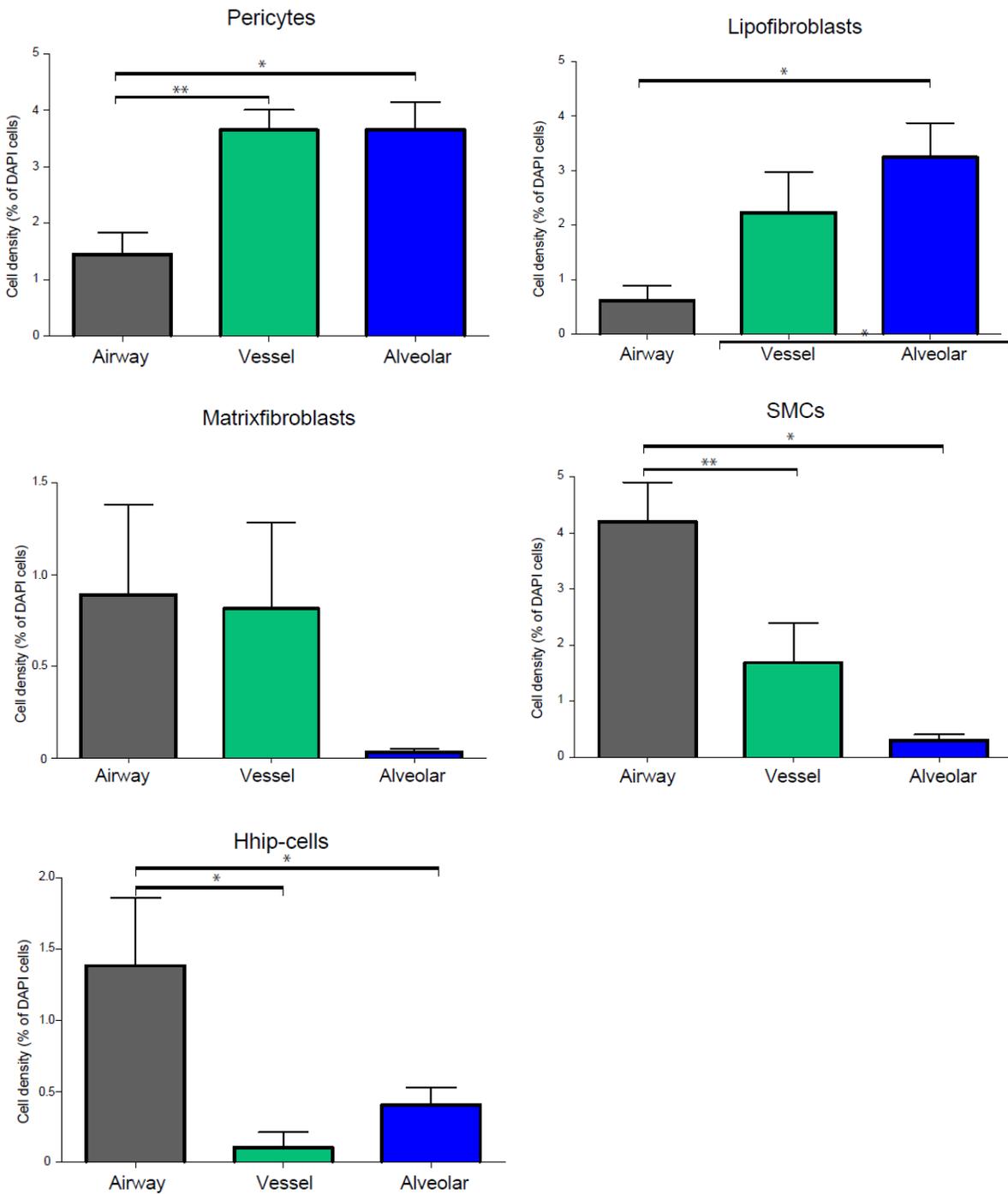


Figure 38. Cell density of fibroblast subtypes. The cell density of mesenchymal subtype cells out of total cell numbers as counted by DAPI-positive nuclei, in the different microscopic tissue regions: airway, vessel, alveolar space. Airways (n=4), vessels (n=3), alveolar space (n=6). Unpaired t-test: Smooth muscle cells airway vs smooth muscle cells alveolar ***p=0.0001. Smooth muscle cell airway vs smooth muscle cell vessel *p=0.0421. Hhip cell airway vs. Hhip cell vessel *p=0.0126, Hhip cell alveolar vs. Hhip cell airway *p=0.0426; lipofibroblast airway vs. lipofibroblast alveolar *p=0.0127; pericyte airway vs. pericyte alveolar *p=0.0115, pericyte vessel vs. pericyte airway **p=0.0035

4.2 Mesenchymal heterogeneity in the lung

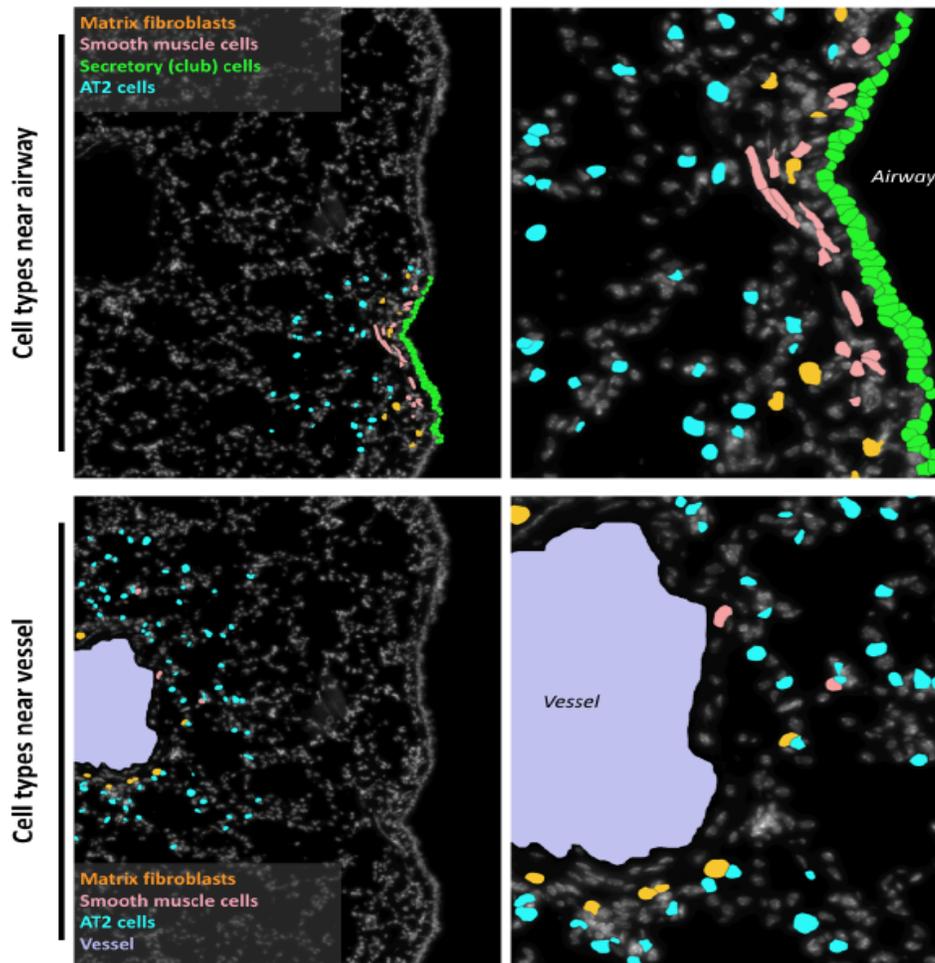


Figure 39. Spatial map of manually assigned cell types, as detected by SCRINSHOT signals. Spatial distribution of matrix fibroblasts (orange), smooth muscle cells (pink), AT2 cells (cyan) and secretory club cells (green) near an airway region and near a vessel region.

To illustrate these findings in the tissue context, the detected matrixfibroblasts and cells smooth muscle were manually matched back in to the representative Area 1, where both a big vessel, determined by the absence of *Scgb1a1* airways cells, and a large airway stretch, determined by the presence of *Scgb1a1* airway cells could be shown (Fig. 39). This schematic nicely summarizes the analysis by showing the exclusive localization of matrixfibroblasts round vessels and airways, referring to adventitial and peribronchiolar fibroblasts, respectively.

In summary, SCRINSHOT was able to provide a comprehensive overview of fibroblast subtype localization, prevalent and epithelial interaction frequencies in tissue in the healthy wild type mouse lung. Thereby, indications from the scRNAseq data such as role and localization of the matrixfibroblasts-constituents, adventitial and peribronchiolar fibroblasts were confirmed. A first hint at the role of Hhip-cells was provided by an exclusive localization around airways.

4.2 Mesenchymal heterogeneity in the lung

4.2.3 Longitudinal changes in fibroblast subtypes in bleomycin time course

The hallmark feature of IPF is the accumulation of activated myofibroblasts that excessively deposit ECM proteins, compromising lung architecture and function and hindering gas exchange. Despite immense interest in fibroblasts in the lung, the origin of the harmful myofibroblasts but also the role of other fibroblast subtypes during injury remains unsolved. While it is impossible to study longitudinal behavior in the human lung, the fibroblast features of IPF are nicely reproduced in the mouse model of bleomycin lung injury.

A change in the mesenchymal compartment of the mouse lung over time during fibrotic injury was studied using again the Drop-seq platform for single cell sequencing (Macosko et al. 2015). Fibrotic lung injury was induced by oropharyngeal instillation of adult (p84) wild-type C57BL/6 mice with a single dose 2U/kg bleomycin (Barbayianni et al. 2018). Whole lung parenchyma cell suspensions were obtained from two mice every day from day 2 to day 14 after bleomycin induction. Additional time points with two mice were taken at day 21, day 28, day 36 and day 54 after bleomycin injury. While the high resolution daily sampling details the early changes upon acute lung injury, the later days represent repair and recovery of the lung, a characteristic of the bleomycin model that differs from the human disease (Schiller et al. 2015). Negative selection was performed using MACS, based on the indicated marker proteins, to remove endothelial (Pecam1), epithelial (Epcam), hematopoietic (Cd45), lymphatic (Lyve1) and erythrocyte (Ter119) cell populations. Of note, the epithelial cells from this time course experiment were used in the manuscript “Longitudinal single cell transcriptomics reveals Krt8+ alveolar epithelial progenitors in lung regeneration” (Strunz et al. 2019).

In total, over 60.000 mesenchymal cells from over 30 mice at 18 time points were sequenced. After filtering and quality control, the high resolution mesenchymal dataset contained 23.232 cells (Fig. 40a). The single cell data was then reduced in the data dimensionality, using the UMAP method (Becht et al. 2018) and unsupervised clustering provided clusters with cells of the same likelihood (Fig. 40c). The cell state identities were derived by manual annotation of the clusters with cell type specific marker genes, gained from the mouse experiment of healthy adult mice in the previous chapter (Fig. 40). Indeed, the same six fibroblast subtypes as in the healthy state were found, with the myofibroblast in addition and displayed in an UMAP colored by cell type (Fig. 40b). Several clusters per cell type in matrixfibroblasts, lipofibroblasts and myofibroblasts, however, were already hinting at more heterogeneity within cell types (Fig. 40c).

4.2 Mesenchymal heterogeneity in the lung

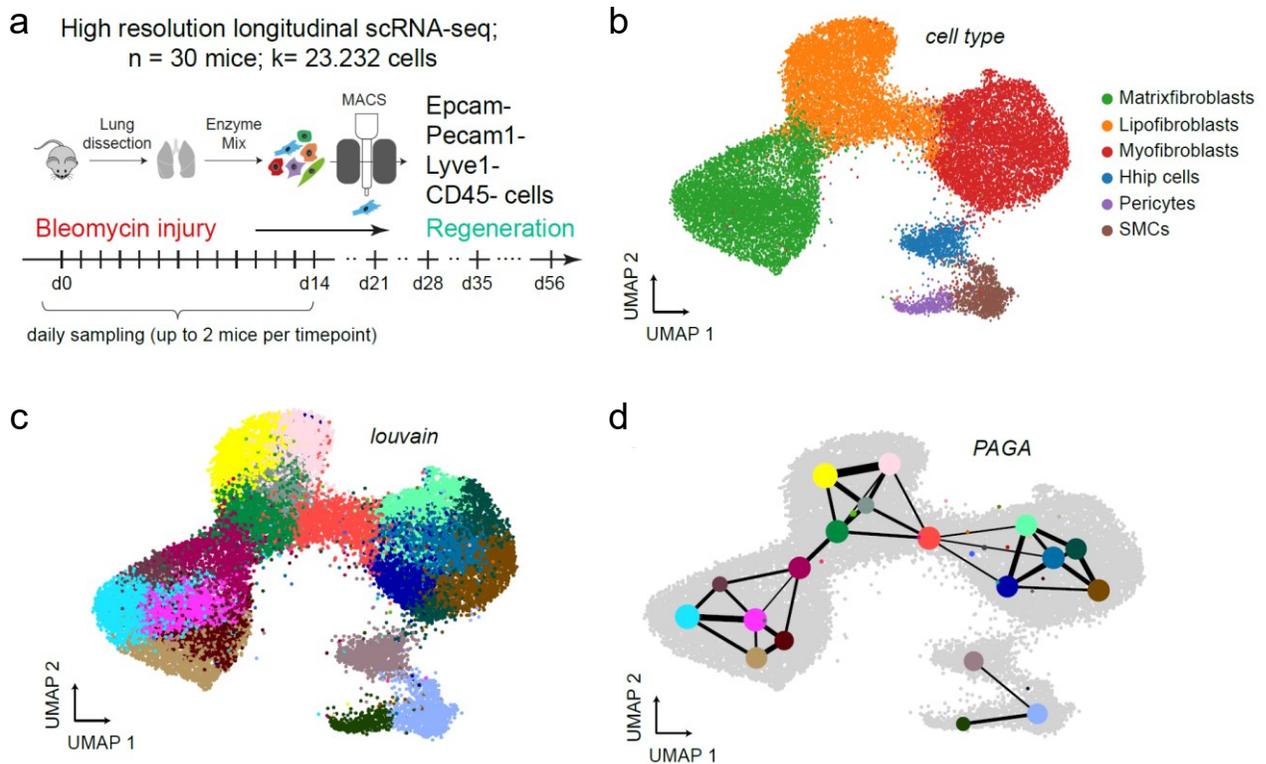


Figure 40. ongitudinal high resolution single cell analysis of mouse lung mesenchymal compartment from a bleomycin injury time course. (a) scRNAseq experimental design. Mesenchymal cells were sampled from up to 2 mice per day from day 2 to day 14 and days 21, 28, 35 and 56 after bleomycin incution. (b-e) Dimension reduced single cell transcriptomic data is visualized through Uniform Manifold Approximation and Projection (UMAP). The color code illustrates the cell types (b) and louvain clusters (c). (d) Partition-based graph abstraction (PAGA) of the data shows 6 main cell clusters with Louvain clusters as nodes and their commonness as thickness of the edges.

Partition-based graph abstraction (PAGA) of the data gave a first glance at the relationship of the cell types, by revealing the connectivity between the fibroblast subtypes, which could infer potential cell trajectories (Fig. 40f). Interestingly, judging from the UMAP and PAGA plot, the cell types presented themselves rather distinct. Pericytes, SMCs and Hhip-cells were more similar to each other and had not connection to the remaining cell types, which is in agreement with the healthy scRNAseq data. Lipofibroblasts and Matrixfibroblasts posed the two largest cell clusters, with matrixfibroblasts not splitting up in adventitial and alveolar fibroblast. Both major cell types seem connected, with the biggest connectivity between clusters within the representative cell types, though.

Importantly, myofibroblasts were only connected to lipofibroblasts (Fig. 40c,d). The connection occurred over one “transition” cluster that was attributed to the lipofibroblasts due to great overlap of gene expression profiles. This transition cluster was also only connected to other lipofibroblast clusters, hinting at lipofibroblasts as the major source of injury induced myofibroblasts in the mouse (Fig. 40d).

4.2 Mesenchymal heterogeneity in the lung

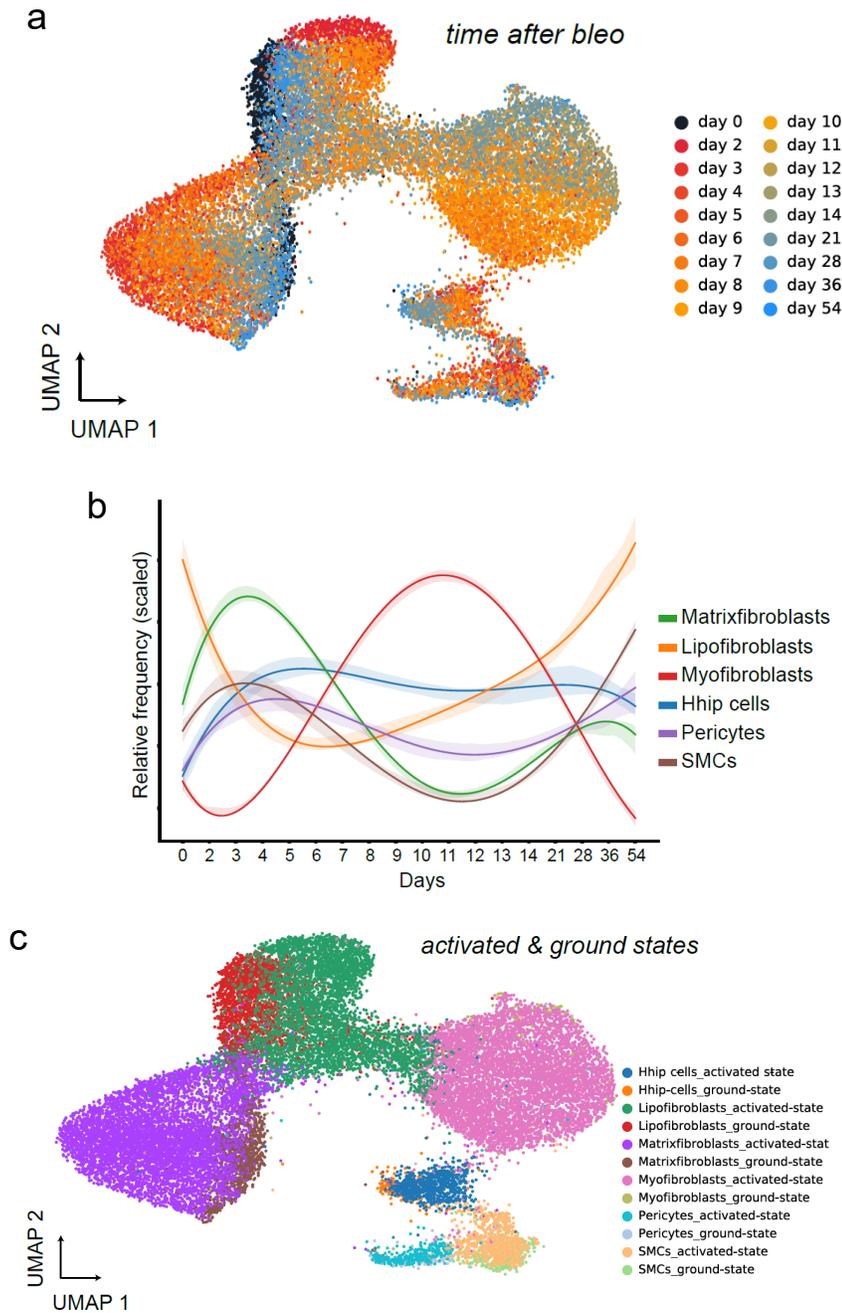


Figure 41. A high-resolution longitudinal data set was generated by subjecting sorted cells from the mesenchymal compartment to scRNAseq from the 18 indicated time points. UMAP embedding displays cells colored by cell type identity (a) and time point (c). (b) The scaled lines represent smoothed relative frequencies of fibroblast subtypes over the time course. Confidence interval derived from the smoothing fit is shown.

4.2 Mesenchymal heterogeneity in the lung

As a next step, the high resolution time component of the data was analyzed. A UMAP was color coded for the different time points of the experiment (Fig. 41a). To understand how the fibroblast lineages change over time, relative cell frequency over time was assessed (Fig. 41b). Myofibroblast showed a gradual increase starting from day 2 until day 11, when they reach their highest numbers, before diminishing again gradually, with no more myofibroblast observed at day 54. The question remained which fibroblast cell type might give rise to the myofibroblasts. Hhip-cells and pericytes stayed at almost constant levels throughout the time course, indicating no differentiation. Matrixfibroblasts, Lipofibroblasts and SMCs on the other hand demonstrated an inverse behavior to the myofibroblasts, with numbers gradually decreasing until day 11 and rising again. Hereby, SMCs and lipofibroblasts reached the same levels as initially, but matrixfibroblasts remained lower in numbers at day 54 (Fig. 41b). While it offers a way to see the development over time, relative frequency is based on the total cell number present at that time point. This can lead to effects as observed here, that when one cell type rises in numbers others must decrease. Cell numbers can also vary due to differences in tissue extraction efficiency at different days, with the tissue becoming more stiff and hard to access while fibrosis develops.

Looking back at the UMAP with color code for time points, interestingly, black and blue colors, marking cells from day 0 and the late regenerative time points of day 21, 28, 36 and 54, respectively, were showing nice overlap among them, in all cell types. This highlights the regenerative feature of the bleomycin model (Fig. 41a). More so, with the time points in between showing a shifted location in the UMAP again for all cell types, it seems as all fibroblast subtypes get activated upon injury and show a different gene expression profile. To account for this remarkable transcriptional heterogeneity during the time course, the cluster of the cells from time points day 2 to day 14 were classified as “activated-states”, while the remaining day 0 and later regenerative time points were named “ground-state” for future analysis (Fig. 41c).

To further analyze fibroblast subtype gene expression kinetics over time, splines were computed. They fit smooth curves between known expression values, by computing polynomial regression between those values (Perperoglou et al. 2019). Accounting for a possible differentiation from lipofibroblasts to myofibroblasts, the two cell types were combined for this analysis. The result is large data matrix, showing changes over time for every gene in each cell type. From this table, the most significant genes were picked to illustrate how different the expression changes were in each cell type (Fig. 42). The mesenchyme marker and highly abundant ECM protein *Col1a2* had an increasing expression until day 11 that goes back to baseline until day 54 in all fibroblast subtypes. On the other hand, the marker for intermediate repair ECM *Tnc*, went only up in the combined fraction of lipofibroblasts and myofibroblasts, as well as Hhip-cells. *Cxcl13* and *Grem1* are two other examples, with *Cxcl13* going up early only in Matrixfibroblats, while *Grem1* expression peaked late in lipofibroblasts and myofibroblast (Fig. 42).

4.2 Mesenchymal heterogeneity in the lung

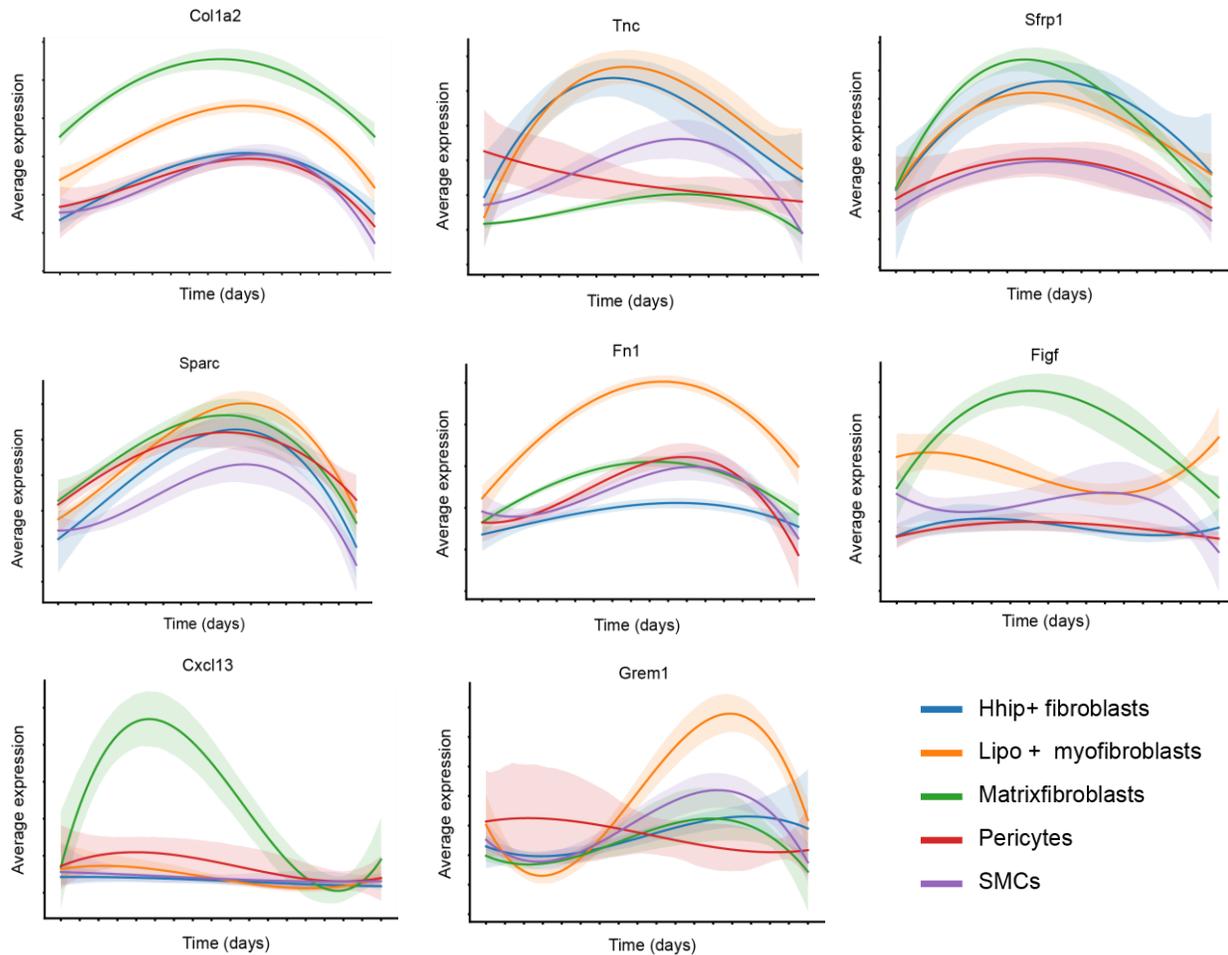


Figure 42. Gene expression kinetics over time. Several genes from the spline table were picked to illustrate distinct expression profile changes of all fibroblast subunits over time such. Average expression ranges from 0 to 1. Time ranges from day 0 to day 14 daily and including the days 21, 28, 36, and 54 as late time points.

This data nicely demonstrated that all of the stromal lineages are contributing to fibrogenic remodeling, in different, and each subtype in their own way. Moreover, heterogeneity of mesenchymal subtypes that could be observed in the healthy state seemed to be conserved during tissue repair. Not multiple fibroblast types, but only a distinct fraction of the lipofibroblasts appeared to be the source of myofibroblasts, while the remaining part differentiated to their own individual activated state during injury.

With multiple activated states being present in the injury scenario, the question arose, how cell interactions and communication between the mesenchymal subtypes and the epithelium in the alveolar niche might change during disease progression. Using the scRNAseq data from the epithelium of the same mice (Strunz et al. 2019), a cell communication network was established by pairing receptor and ligand expressions in cell types to receptor-ligand interactions between cell types (Fig. 43).

4.2 Mesenchymal heterogeneity in the lung

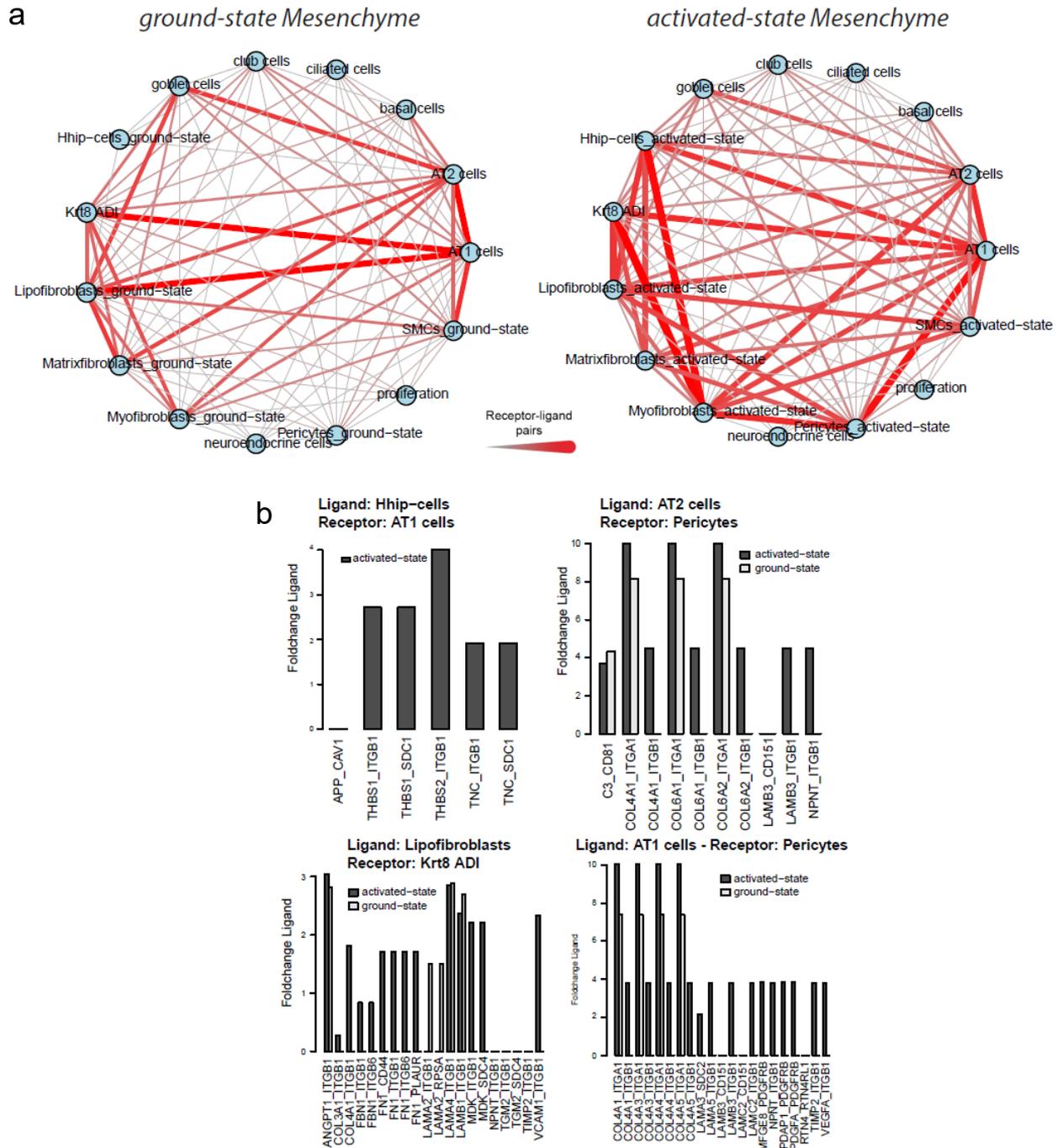


Figure 43. Receptor ligand interactions between epithelium and mesenchyme during injury. (a) The cell-cell communication network displays the number of receptor-ligand pairs between the mesenchymal subtypes in their ground as well as activated state and epithelial cell types. (b) The bar graphs show the average log₂ fold change of the respective receptor-ligand pairs. The first named gene is always the receptor, the second one the ligand. Cell-types expressing the ligand or the receptor are specified. Black bars show activated-states, white bars show ground-state of the respective mesenchymal subtype.

4.2 Mesenchymal heterogeneity in the lung

To visualize the changes upon injury, the interactions were calculated separately for ground-state and activated states (Fig 43a). In the ground-state most interactions were found between lipofibroblasts and AT-1 cells as well as goblet cells. Also Matrixfibroblasts interacted with AT-2 cells. Contrary, Hhip-cells showed very little to no interactions at all (Fig. 43a,b). This changed dramatically during injury in the activated-states. Notably, Hhip-cells now had a lot of interactions with AT-1 cells that were not present before. Also between Myofibroblasts (mostly only present in the activated-state) and AT-1 and AT-2 cells many interaction pairs were found and also pericytes in the activated state seemed to be able to interact more. The bar plots demonstrate, that there was a dramatic change in the number of receptor-ligand pairs, between the ground- and activated-states. But at the same time, also expression of already present receptor and ligand genes increased or decreased upon injury. Interestingly, as demonstrated for the healthy cell states before, the receptor and ligand genes expressed by different subtypes were mostly specific and distinct to one cell type, highlighting the important role of the heterogeneity of fibroblast also during injury.

In summary, the high-resolution temporal scRNAseq data on mesenchymal cell types during bleomycin lung injury in mice provided important insights. The data demonstrated that multiple stromal lineages are contributing to fibrogenic remodeling, each subtype in their own way. All subtypes become activated after injury, with changed gene expression profiles that manifest in altered cell communication with differential expression of receptor-ligand pairs. Moreover, heterogeneity of mesenchymal subtypes that could be observed in the healthy state seems to be conserved during tissue repair. Importantly, not multiple fibroblast types converge together, but only the lipofibroblasts appeared to be the source of myofibroblasts. However, this finding still needs further validation.

4.2.4 Comparison of mesenchymal cell states in mouse and human lung injury on a single cell level

The mouse scRNA-seq experiments over the bleomycin time course provided information about the activation of fibroblasts upon injury, with multiple activated states being present apart from the pathogenic myofibroblasts. To gain translational insights from the analysis, the mouse data was compared with the human scRNA-seq data of healthy donor and ILD patients from the first chapter.

Sub-clustering and calculating a new UMAP embedding of human *COL1A2* expressing mesenchymal cells revealed 11 distinct cell states (Fig. 44a). Those were annotated based on known marker genes, published datasets and comparison with the mouse data (Fig. 45).

4.2 Mesenchymal heterogeneity in the lung

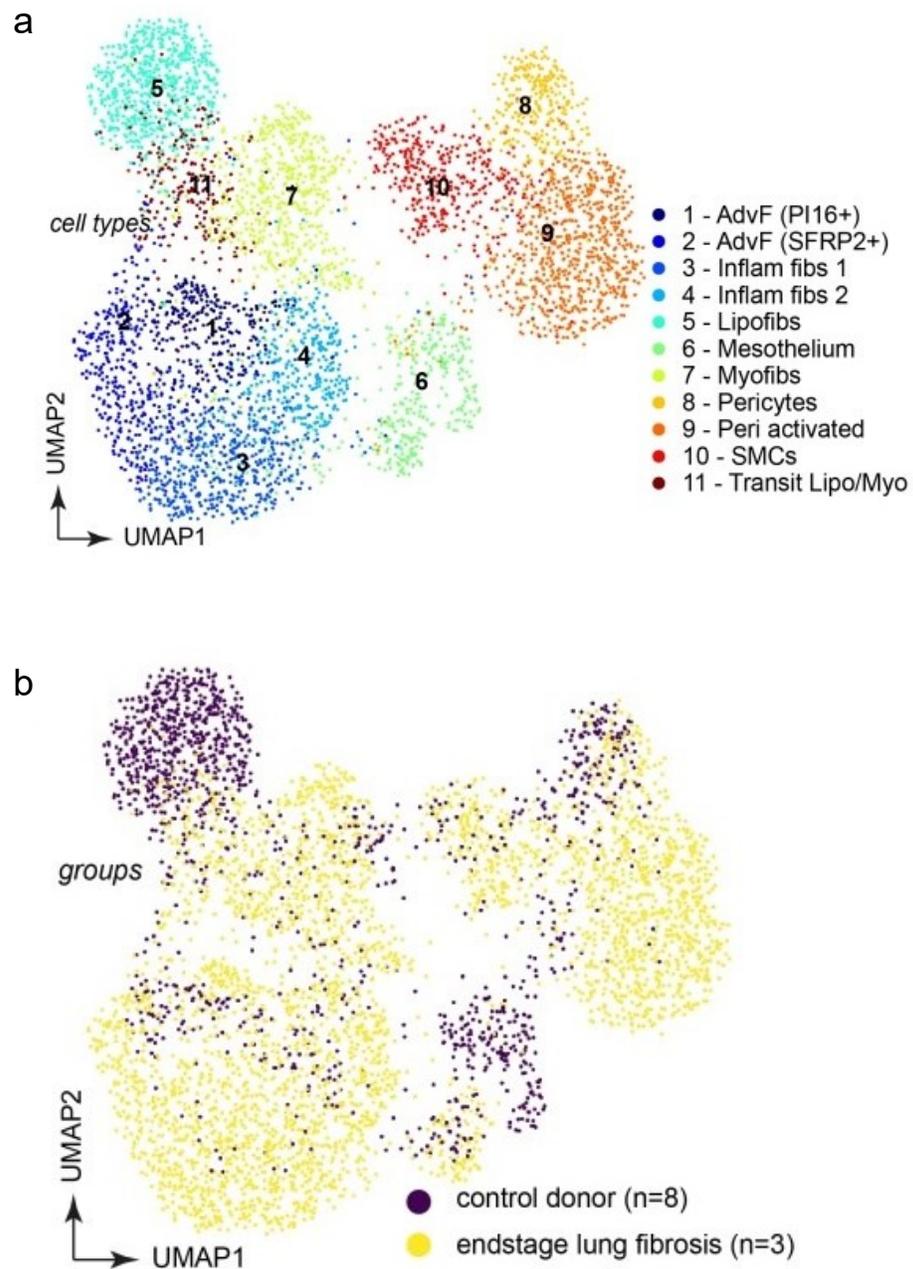


Figure 44. Clustering analysis and cell type annotation of stromal subtypes in human lung parenchyma. (a) COL1A2+ mesenchymal cells, from the whole human dataset, were embedded in a new UMAP to allow for separation of cluster identities and identification of cell states. (b) Cells colored in disease groups show origin of identified cell states. Figure and legend taken from (Mayr et al. 2020).

4.2 Mesenchymal heterogeneity in the lung

They were called cell states, since some of the cell types e.g. the pericytes, could be split up in two clusters of which one was only present in diseased ILD patients, and was therefore termed “activated” pericytes (Fig. 44a).

In general the resulting UMAP looked similar to the mouse data. SMCs and pericytes were located away from the fibroblasts populations. Different from the mouse data, the human mesothelial cells were still present. The fibroblasts split up into three main groups, the Myofibroblasts, the Lipofibroblasts and a population that was called Matrixfibroblasts in the mouse, consisting now of four different subtypes. Due to the expression of cytokines and chemokines such as *CXCL12*, two of the sub-clusters were termed “inflammatory” fibroblasts 1 and 2. Additionally, they were only present in ILD patients, hinting at a potential role in diseases. Two adventitial fibroblast populations could be defined, that were similar but differed in the expression of the ECM proteins secreted frizzled-related protein (SFRP1 and 2), containing a cysteine-rich domain homologous to the putative Wnt-binding site.

Interestingly, located closest to the Myofibroblasts were again the Lipofibroblasts, with only a transitional cluster, containing cells with marker genes from both populations, separating them (Fig 44a). While no direct proof, the data in human seems to suggest the same principle as the mouse fibrosis data, that lipofibroblasts might be the only predecessor of myofibroblasts in human ILD.

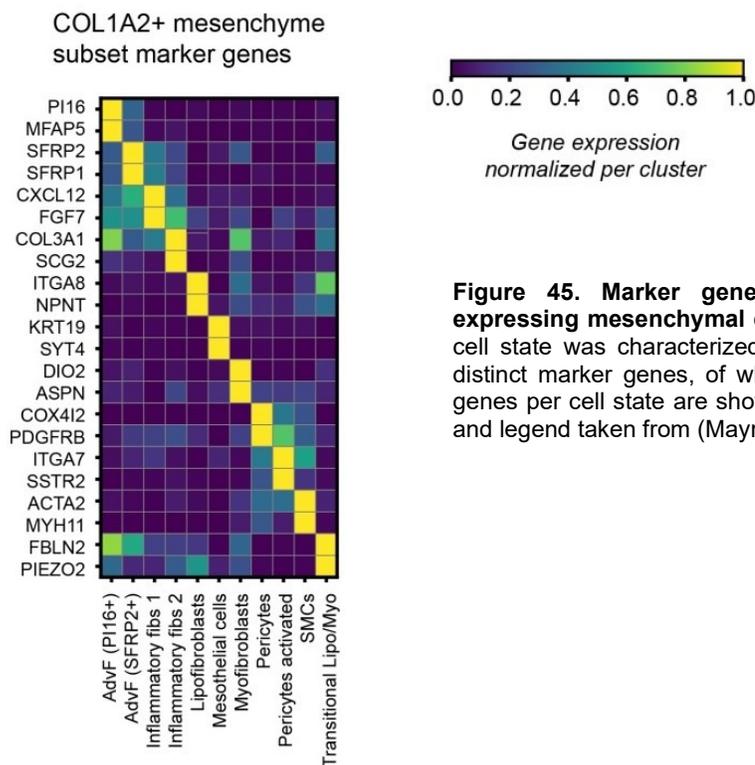


Figure 45. Marker genes for human COL1A2 expressing mesenchymal cell types. Each identified cell state was characterized by a whole signature of distinct marker genes, of which the two most distinct genes per cell state are shown in the heatmap. Figure and legend taken from (Mayr et al. 2020).

4.2 Mesenchymal heterogeneity in the lung

While clustering single cells into a UMAP provides a good overview for that sample, data-driven results often exceed prior knowledge about tissue composition and novel cell clusters are not straightforward to characterize. Additionally, data interpretation between datasets is extremely challenging. Therefore, to better compare mouse and human data in a more quantitative way, *matchScore*, a Jaccard index based scoring system was used, to quantify clustering and marker accuracy in a combined score and to integrate cluster identities across different data sets (Mereu et al. 2018).

First, the human ILD mesenchymal sub types were compared to the healthy mouse baseline data, based on the distinct marker signatures of the different clusters (Fig. 46). Both SMCs and the two pericytes cell states in human, correlated best with their equivalents identified in mouse. Of note, lipofibroblasts showed a good overlap between mouse and human, defying some discussion in the field that lipofibroblasts do not exist in human. Interestingly, human myofibroblast, of course not present in the healthy mouse, also correlated best with the mouse lipofibroblasts, giving further indication that those might be the source for myofibroblasts both in human and mouse. The discrimination between adventitial and peribronchiolar, that was clearly visible in mouse, could not be disentangled in human. All four human cell states of inflammatory and adventitial fibroblasts showed similar correspondence with the mouse peribronchiolar or adventitial fibroblasts. Thereby, the overlap with the peribronchiolar fibroblast was higher (Fig. 46). Surprisingly, the newly identified mouse cell type of Hhip-cells did not correspond at all with any human mesenchymal cell type, leading to more questions about their function in the mouse.

Next, the human ILD mesenchymal sub types were compared to the cell types found in the mouse data of bleomycin induced fibrosis, to analyze how similar the disease specific transcriptional alternations might correspond between species (Fig. 47). Interestingly, apart from SMCs, all all activated states in mouse, corresponding to the time points day 2 until day 14 in the bleomycin time course, showed a higher overlap to the human equivalent cell states than the ground state. Surprisingly, human and mouse myofibroblast showed nearly no similarities. Instead, the Hhip-cell activated state in the mouse correlated best with human myofibroblast signatures, followed by matrixfibroblasts. The Hhip-cells groundstate however, seemed again not to be present in human. Since the mouse timecourse data did not allow to distinguish between adventitial and peribronchiolar fibroblast, they were summed up in the matrixfibroblasts, to which now the human adventitial and inflammatory fibroblast corresponded best (Fig. 47).

In summary, the comparison of human and mouse mesenchymal cell types demonstrated that although the human mesenchyme seems to split up in more subtypes, cell types and their marker genes are similar. Quantitative analysis revealed, that activated cell states upon bleomycin induced fibrosis in mice, are more similar to the human subtypes than the ground state and revealed the activated Hhip-cells as most similar cell type to human myofibroblasts.

4.2 Mesenchymal heterogeneity in the lung

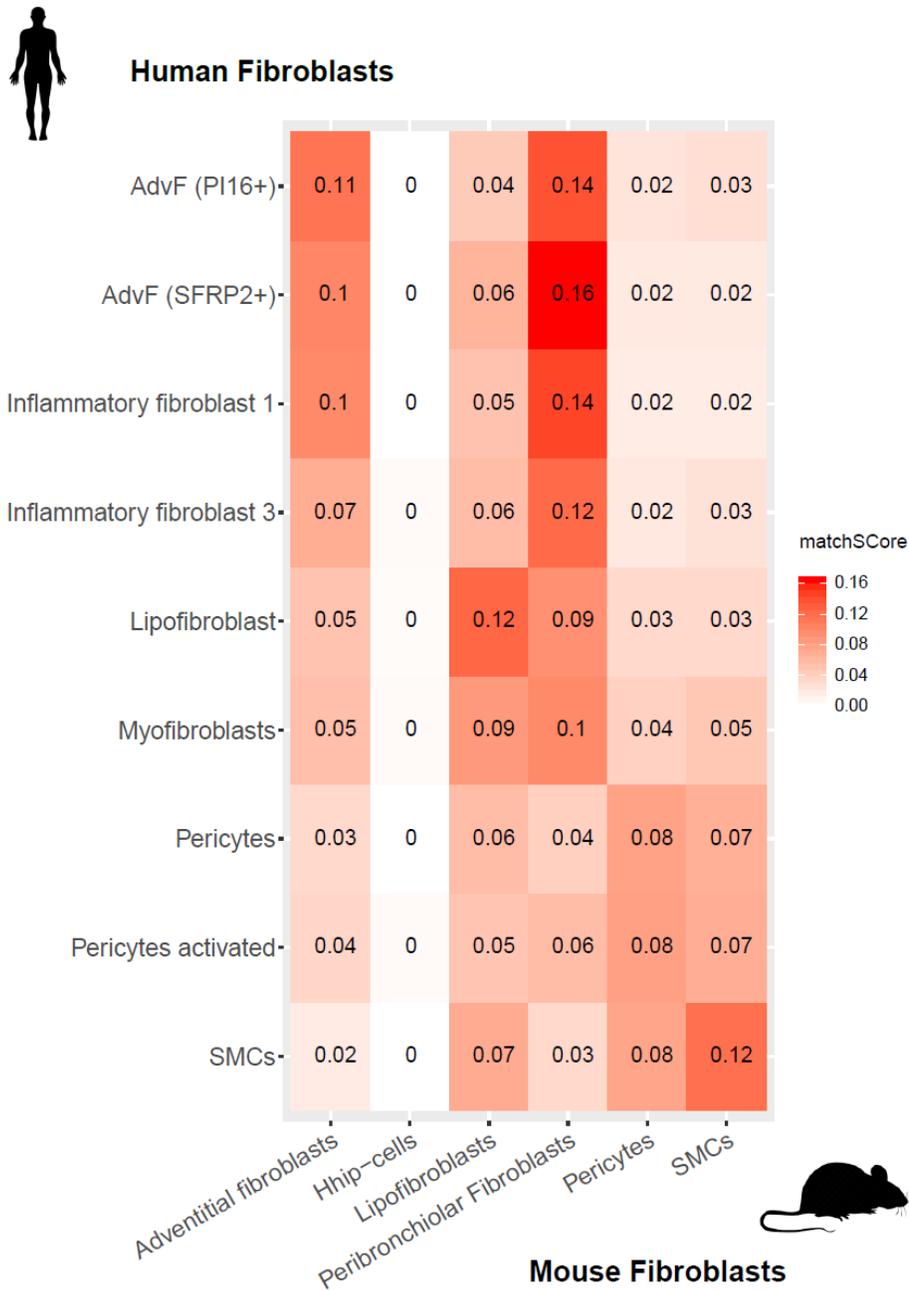


Figure 46. Comparison of human and mouse mesenchymal subsets. *matchScore* comparison values were calculated, showing the overlap between marker gene signatures with a defined logfoldchange and adjusted p value cut off, from the human and healthy mouse scRNAseq data.

4.2 Mesenchymal heterogeneity in the lung

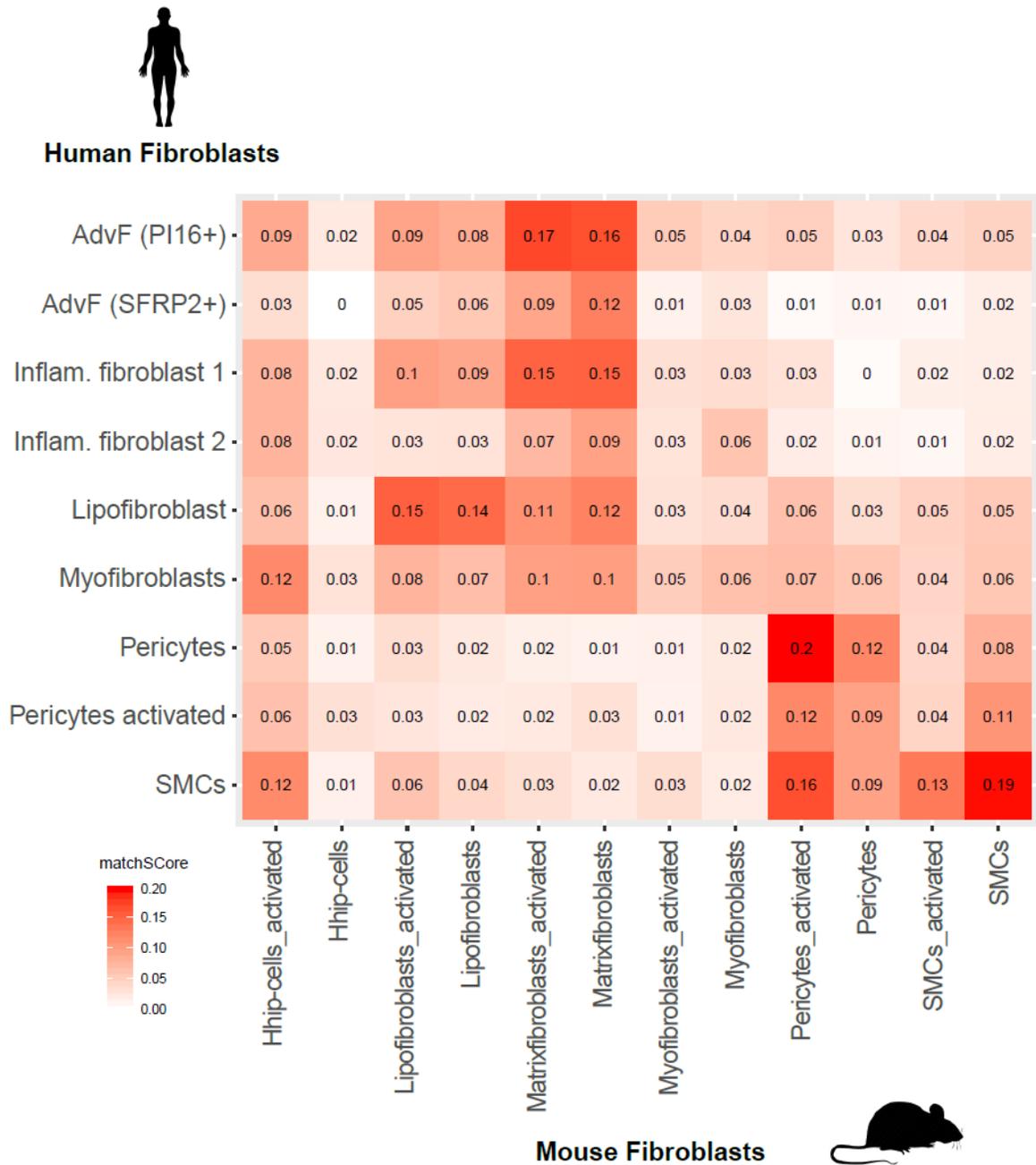


Figure 47. Comparison of human and mouse mesenchymal subsets. *matchScore* comparison values were calculated, showing the overlap between marker gene signatures with a defined logfoldchange and adjusted p value cut off, from the human and bleomycin induced fibrosis mouse timecourse scRNAseq data.

4.2 Mesenchymal heterogeneity in the lung

4.2.5 Discussion

The field of single cell genomics is moving on to characterize pathogenesis and disease progression, after having defined the baseline for nearly every organ with a cell atlas. Monitoring human lung fibrosis progression is difficult due to the need of invasive surgery of already suffering patients, to get a piece of tissue to suspend for scRNA-seq. Therefore, studies in the mouse lung are indispensable for defining disease progression trajectories and characteristic cell states that can then be translated back and integrated into human data to infer potential treatment options. While atlases give the impression of dissecting the whole organ comprehensively, sub-compartments are often overlooked in the surge to cover the whole organ. In both mouse and human, the mesenchymal compartment is such a poorly defined example. The first chapter of this study demonstrated on human single data in fibrosis, that only sub clustering into compartments allows for accurate definition of cell types and states. The aim of this second chapter of this thesis was to better define the baseline as well as the disease progression during fibrosis of the mesenchymal compartment in the mouse lung on a single cell level, with special focus on fibroblasts.

Using MACS depletion, mesenchymal cells from healthy adult mouse lungs were enriched from whole lung single cell suspensions and sequenced, revealing 6 distinct stromal cell types, including Lipofibroblasts, Peribronchiolar and adventitial fibroblasts, SMCs, Pericytes and the new cell type of Hhip-cells. Although, only two mice were used, the cell types clusters were robust, consisted of cells from both mice and could be defined with marker signatures. While other studies, in the early days of scRNAseq had already looked at the mesenchyme, none was able to dissect the cellular composition in that resolution. Additionally, most other studies so far, used a lineage label to enrich for mesenchymal cells, potentially inferring a pronounced bias, based on a limited number of markers, thereby masking variability within a subpopulation of cells that express the selected marker genes (Peng et al. 2015, El Agha et al. 2017, Lee et al. 2017, Zepp et al. 2017). This approach might therefore give a biased view on organ complexity (Kester et al. 2018). In contrast, the presented study used an unbiased approach with cells from the whole lung and only enriching the mesenchymal populations by negative depletion of other well-known lineages, which could still be found as contaminants in the data before filtering. For the aim of identifying cell types, this approach seemed more appropriate and worked convincingly well, as demonstrated.

Defining cellular identity and delineating which specific markers characterize these specific subpopulations have been significant hurdles in understanding the role and function of certain mesenchymal cells in the lung (Sucre et al. 2020). In this study, especially the GO term enrichment was of great help to overcome some of these obstacles. While not obvious from the marker list, the GO enrichment identified a subset of genes that marked the whole cluster of adventitial fibroblasts as to be involved in “blood vessel morphogenesis”, helping dramatically in

4.2 Mesenchymal heterogeneity in the lung

distinguishing it from the closely related subset next to them (Mi et al. 2019). It remains to be proven and requires further experiments to rightfully address those neighboring cells as peribronchiolar cells.

Novel advancements in spatially resolved transcriptomics that have developed at a similarly fast pace as scRNAseq methods, can help to verify such hypotheses. In this study the recently published SCRINSHOT method was employed to identify cell type specific locations in the histological tissue context for each of the mesenchymal subtypes (Sountoulidis et al. 2020). The marker genes used, were chosen before the scRNAseq dataset was fully analyzed and could therefore not distinguish between adventitial and peribronchiolar fibroblasts. However, density analysis revealed that those cells (termed Matrixfibroblasts in the SCRINSHOT analysis) were significantly enriched around airways and vessels compared to the alveolar space, giving further evidence that the initial classification holds true. This nicely demonstrates the power of mRNA localization techniques, by showing that the location of a cell can give an indication to its function (Stuart et al. 2019).

Proliferation and repair of different epithelial cells in the lung is dependent on specific cell interaction partners from the mesenchyme. To elucidate the interactions that distinct mesenchymal cell types may have with the epithelium, the frequency of the mesenchymal cell types in close proximity to AT-2 cells was quantified. Lipofibroblasts, as adipocyte-like cells, are known to assist AT-2 cells in surfactant production by transfer of lipids between them, and were also shown to support their congenic growth *in vitro*, providing a reasonable explanation for the overserved high interaction rate (Schultz et al. 2002). Pericytes were the cell type consistently identified the most across all areas. Surprisingly, they were significantly enriched around vessels but also in the alveolar space, compared to airway regions. Pericytes, being highly abundant in the lung, share the basement membrane with endothelial cells. They are therefore structurally in close contact with epithelial cells that facilitate the gas exchange with the blood, potentially explaining their high interaction rate with epithelial type 2 cells (Armulik et al. 2005). A limitation of the study in this context was the lack of marker genes that could identify vessels, as it was the case for *Scgb1a1* for airways. While major vessels could be identified by their structure and lack of the airway markers, smaller vessels amid the alveolar space were potentially not recognized, giving one possible explanation for the high localization of pericytes in the alveolar space.

Former studies also looked at mesenchymal cell types, their interactions and signaling. Peng and colleagues could trace back sonic hedgehog (Ssh) signaling to club and AT-2 cells in the adult lung, but *Gli1*, downstream transcriptional effector and target of hedgehog, was demonstrated to be expressed in *Col1a1*, *Vim*, *Pdgfra* and *Pdgfrb* positive mesenchymal cells, specifically located mostly adjacent to proximal airway and pulmonary arteries, but belonging to smooth muscle cells. They were able to show that the epithelium actively maintains quiescence of the mesenchyme

4.2 Mesenchymal heterogeneity in the lung

through hedgehog (Hh) signaling, which in a feedback loop also regulates epithelial quiescence. During injury and repair, hedgehog activation is inversely correlated to the proliferation of mesenchymal cells (Peng et al. 2015).

Lee and colleagues found four mesenchymal cell populations with a total of 182 cells, by sorting lineage labeled *Lgr5* and *Lgr6* cells. Thereby, they identified *Lgr6*-expressing cells to surround bronchiolar epithelia and in the alveolar space, whereas *Lgr5*-expressing cells were largely alveolar (Lee et al. 2017). *Lgr6* cells, a subtype of airway smooth muscle cells (ASMCs) contribute to quiescence and with *Gli1*, express the same *Ssh* target that Peng and colleagues identified, suggesting it is the same cell type. *Lgr5* cells on the other hand, promoted alveolar differentiation by activation of the Wnt pathway but were found to be different from *Pdgfra* expressing cells, which were already known to promote AT-2 cell renewal (Barkauskas et al. 2013, Lee et al. 2017).

The Hhip-cells, are a newly discovered mesenchymal subtype in this study that was found in both the baseline and time course experiment, characterized by the distinct expression of hedgehog interacting protein (Hhip) as well as Asporin (*Aspn*). While neither the marker gene list nor the GO term enrichment indicated any specific function, SCRINSHOT showed a specific location to mostly airways, followed by the alveolar region. Interestingly, by comparing with the literature, the Hhip-cells were the only cells in this study that expressed both *Lgr5* and *Lgr6* at similar levels. While the locations were fitting with observations from Lee and colleagues, the two populations could not be resolved in the dataset of this study (Lee et al. 2017). However, the embedding close to the SMCs and pericytes could suggest the role of ASMCs, at least for part of the Hhip-cells cell population, while the AT-2 supporting *Lgr5* cells remain elusive.

Zepp and colleagues have found two mesenchymal subtypes, characterized by the expression of *Axin2*, of which one also supports alveolar self-renewal (Zepp et al. 2017). *Axin2* is a gene in the Wnt pathway and its expression reports responsiveness of the expressing cell to this critical signaling for maintaining homeostasis (Clevers et al. 2014, Frank et al. 2016). The cell type that also supports alveolar differentiation expresses *Pdgfra* in addition to *Axin2*, was found as primary interaction partner of AT-2 cells, stimulated organoid growth in vitro and was thus termed Mesenchymal alveolar niche cells (MANCs) (Zepp et al. 2017). Based on the data of this study, the MANCs corresponded best with the lipofibroblasts, that were the primarily *Pdgfra* expressing cells and showed the second most interactions with AT-2 cells, confirming nicely the data. The other cell type Zepp and colleagues have found was missing *Pdgfra* expression and had lower levels of *Axin2*, indicating lower Wnt responsiveness. At day 21 after bleomycin lung injury those *Axin2* positive cells had been contributing to over 40% of myofibroblasts, leading to their name as *Axin2*⁺ myofibrogenic progenitors (AMP) (Zepp et al. 2017).

Comparing these findings to the datasets in this study leads to the aims that wanted to be addressed with the high-resolution bleomycin time course. Activated fibroblasts and

4.2 Mesenchymal heterogeneity in the lung

myofibroblasts are regarded as the key cell types that lead to an increased deposition of extracellular matrix and fibroblastic foci in lungs of patients with idiopathic pulmonary fibrosis. The origin of myofibroblasts however remains elusive, with multiple sources suggested in different studies over the years (Phillips et al. 2004, Kim et al. 2006, Hoyles et al. 2011, Rock et al. 2011). Based on the marker panels in the publications that defined the AMP lineage, it became obvious that Zepp and colleagues had identified pericytes, which are commonly known to express *Pdgfrb*, *Notch3* and *Gucy1a3* (Zepp et al. 2017). Of note, they did not distinguish between SMCs and AMPs, giving a possible explanation for the mix of marker genes. Their data on the origin of myofibroblasts with the approach of lineage tracing several cell types, was well done and could help the cause. However, the selection of day 21 post bleomycin injury to assess the number *Acta2* expressing myofibroblasts was poorly chosen, since myofibroblast numbers are already decreasing again at that time point (Zepp et al. 2017).

Xie and colleagues also worked on deconvolution of fibroblast heterogeneity in mouse pulmonary fibrosis, using the bleomycin model at day 0 and day 21 (Xie et al. 2018). While they found up to five similar populations as in this study there were also major differences and points of critics. They identified a substantial myofibroblast population already at day 0, which in fact corresponded to the Hhip-cells in this study; SMCs were missing at all. Their lipofibroblasts contained a “macrophage-signature” which clearly reflected contamination during their mesenchyme enrichment. Interestingly however, their model of cell state trajectories based on SOM machine learning reported a directed hierarchical relation of lipofibroblasts with myofibroblasts ((Xie et al. 2018).

The data in this thesis suggest that myofibroblast numbers were highest at day 11 and decreased to only half of the maximum at day 21. Additionally, based on the UMAP embedding, PAGA connectivity and gene expression kinetics, this study proposes lipofibroblasts as the only source of myofibroblasts in the bleomycin model. Further experiments, such as lineage tracing of the lipofibroblasts and also data analysis approaches such as RNA velocity and pseudo time analysis, to assess differentiation trajectories of the lipofibroblasts based on their expression of spliced and unspliced mRNA are needed to support this hypothesis. It is important to stress, that the data in this study showed that the distinct subtypes did not converge to one canonical activated fibroblasts / myofibroblast cell state. Instead, all mesenchymal subtypes got activated during injury and contribute to fibrogenic remodeling in different ways. While the receptor-ligand analysis already suggested interesting interaction pairs, more analysis is needed to assess which molecular pathways are involved in both the activation upon injury as well as during dedifferentiation when the damage is cleared.

5 Conclusions and future directions

4.2 Mesenchymal heterogeneity in the lung

In conclusion, the presented work described in two chapters how an integrated single cell analysis of human lung fibrosis can resolve cellular origins of predictive protein signatures in body fluids and contributed to a better understanding of mesenchymal heterogeneity in health and in temporally resolved fibrotic lung injury in mice.

In the first chapter, an ILD patient cohort from Munich analyzed by scRNAseq was integrated with similar public available datasets to establish a robust differential gene expression analysis for all major cell types in the lung, with high statistical power due to increased number of patients (Mayr et al. 2020). Such larger cohorts can help the concept, that inter-individual variance within patient cohorts can help to better model the diseases, with various patients at different stages of a disease progression trajectory possessing cells of different characteristic states (Mayr et al. 2020). Body fluids of those patients potentially contain a composite representation of these disease stages. Therefore, bronchoalveolar lavage (BALF) and plasma proteomes in patients from several independent ILD cohorts were quantified, using mass spectrometry driven proteomics (Mayr et al. 2020). A bioinformatics analysis strategy enabled the integration of the cell state descriptions on single cell level with these lavage and plasma proteomes and associated clinical meta-data to reveal that fluid proteome signatures were predictive of specific cell state changes in the lung (Mayr et al. 2020). While one of the first positive examples in the field this study shows how, training of machine learning algorithms with large datasets of matched single cell genomic and fluid proteomic or sequencing readouts can enable new automated tools. These can be used for clinical decision making (Walsh et al. 2019) and drug monitoring (Maher et al. 2019) and strengthen personalized medicine approaches. However, certain weaknesses still need to be assessed, such as the fact that single cell and proteomics data were not derived from the same patients but stem from various sources with their own variabilities and from different types of ILDs, which are also highly diverse (Mayr et al. 2020).

For the situation in human, be it for the healthy base line or in chronic lung disease, researchers team up for current endeavors for the Human Cell Atlas (Regev et al. 2017) and the Human Lung Cell Atlas in particular (Schiller et al. 2019). This project will lay basis to understand lung development, lung homeostasis, and lung regeneration. Thereby, guidelines and more accepted nomenclatures will be provided that are of immense value for the community. However, such efforts are only feasible by the generation of longitudinal data derived from many individuals and by highly functional data integration methods (Stuart et al. 2019). Combining experimental models of pulmonary fibrosis with longitudinal sequencing and novel computational approaches (Strunz et al. 2019, Mayr et al. 2020) as used in this presented work, complex dynamic processes can be identified, even across samples and species empowering predictions for future translational work.

In the second chapter, single cell transcriptomics on mesenchyme enriched cell suspensions of healthy adult wild-type mice was performed. This unbiased approach for classification of

4.2 Mesenchymal heterogeneity in the lung

mesenchymal cell subtypes is new in a field that relied on lineage tracing of potential cell type markers in the past. In total, 7 distinct cell types were identified, including mesothelial cells, smooth muscle cells, and 5 distinct stromal cell types with features of fibroblast phenotypes. Multiplexed in situ FISH, using the SCRINSHOT technique, enabled the definition of distinct locations and interaction partners of these subtypes in the tissue context (Sountoulidis et al. 2020). Thereby, VSMCs and ASMCs, could be identified, allowing to distinguish between them (Danopoulos et al. 2020). Therefore this study contains the most comprehensive analysis of mesenchymal subtypes in the mouse lung so far and provides marker signatures for all cell types, some of which have been missing or were nonspecific so far. A temporal high resolution experiment using the bleomycin mouse model of lung fibrosis and regeneration, dissected mesenchymal cell types during injury at single cell level and identified activated states with a changed mRNA expression profile for all mesenchymal subtypes upon injury. Although this work is still ongoing, the scRNA-seq data suggest Lipofibroblasts as the only progenitor of myofibroblasts. Thereby, this work adds to the field where activated fibroblasts and myofibroblasts are regarded as the key cell types in fibrosis, with the origin of myofibroblasts however being elusive and multiple sources critically discussed in over the years (Phillips et al. 2004, Kim et al. 2006, Hoyles et al. 2011, Rock et al. 2011).

All in all, the comparison with literature and similar projects of other groups as discussed, demonstrates that there is still an urgent need for a better understanding of the mesenchymal subtypes in the lung. While the interest is high, and progress is made with the rise of new technologies such as scRNAseq or spatial methods, one of the major obstacles remains the nomenclature of cell types (Sucre et al. 2020). As this study tried to point, several of groups have found similar cell types which only become obvious with a deeper understanding of the matter, detailed knowledge about marker genes and experience in pitfalls in e.g. variations in cell extraction protocols painting different UMAPs of cell types. The presented work on the heterogeneity of fibroblasts will continue in this direction and will try to provide an integration of cell types identified by other groups, as to provide a comprehensive overview of mesenchymal subtypes, their location, and putative function in health and fibrotic lung diseases in the mouse for the field.

On an experimental level, the work on the mouse mesenchyme could benefit from further validation using more conventional biological assays. Isolating the distinct subtypes by FACS and growing in them separately in organoids with epithelial cells will give further information about their function in the alveolar niche to provide signals that maintain epithelial homeostasis (Barkauskas et al. 2017). Performing the same experiment with activated disease states will help to understand better the mesenchymal and epithelial interactions in fibrosis. Ultimately, combining the gained scRNAseq data with classical lineage tracing experiments on lipofibroblasts markers in fibrosis will validate the hypothesis that they give rise to myofibroblasts.

4.2 Mesenchymal heterogeneity in the lung

Overall, the herein presented data added new aspects of lung fibrosis to the research field by eliciting how protein signatures in body fluids can be assigned to various cell types that respond to lung injury. The powerful combination of single cell RNA sequencing and novel computational analysis methods, plus the characterization of the heterogeneity in the mesenchyme that is heavily involved in fibrotic processes, will aid the ongoing research efforts to better understand the molecular events involved in lung tissue fibrosis and regeneration.

6 Bibliography

Adams, T. S., Schupp, J. C., Poli, S., Ayaub, E. A., Neumark, N., Ahangari, F., Chu, S. G., Raby, B. A., Deluliis, G., Januszyk, M., Duan, Q., Arnett, H. A., Siddiqui, A., Washko, G. R., Homer, R., Yan, X., Rosas, I. O. and Kaminski, N. (2019). "Single Cell RNA-seq reveals ectopic and aberrant lung resident cell populations in Idiopathic Pulmonary Fibrosis." [bioRxiv](#): 759902.

Aebersold, R. and Mann, M. (2016). "Mass-spectrometric exploration of proteome structure and function." [Nature](#) 537: 347-355.

Allen, R. J., Guillen-Guio, B., Oldham, J. M., Ma, S. F., Dressen, A., Paynton, M. L., Kraven, L. M., Obeidat, M., Li, X., Ng, M., Braybrooke, R., Molina-Molina, M., Hobbs, B. D., Putman, R. K., Sakornsakolpat, P., Booth, H. L., Fahy, W. A., Hart, S. P., Hill, M. R., Hirani, N., Hubbard, R. B., McAnulty, R. J., Millar, A. B., Navaratnam, V., Oballa, E., Parfrey, H., Saini, G., Whyte, M. K. B., Zhang, Y., Kaminski, N., Adegunsoye, A., Strek, M. E., Neighbors, M., Sheng, X. R., Gudmundsson, G., Gudnason, V., Hatabu, H., Lederer, D. J., Manichaikul, A., Newell, J. D., Jr., O'Connor, G. T., Ortega, V. E., Xu, H., Fingerlin, T. E., Bosse, Y., Hao, K., Joubert, P., Nickle, D. C., Sin, D. D., Timens, W., Furniss, D., Morris, A. P., Zondervan, K., Hall, I. P., Sayers, I., Tobin, M. D., Maher, T. M., Cho, M. H., Hunninghake, G. M., Schwartz, D. A., Yaspan, B. L., Molyneaux, P. L., Flores, C., Noth, I., Jenkins, R. G. and Wain, L. V. (2019). "Genome-Wide Association Study of Susceptibility to Idiopathic Pulmonary Fibrosis." [Am J Respir Crit Care Med](#).

Altorki, N. K., Markowitz, G. J., Gao, D., Port, J. L., Saxena, A., Stiles, B., McGraw, T. and Mittal, V. (2019). "The lung microenvironment: an important regulator of tumour growth and metastasis." [Nat Rev Cancer](#) 19: 9-31.

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., Nagendran, M., Desai, T., Eickelberg, O., Mann, M., Theis, F. J. and Schiller, H. B. (2019). "An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics." [Nat Commun](#) 10: 963.

Annesi-Maesano, I., Norback, D., Zielinski, J., Bernard, A., Gratiou, C., Sigsgaard, T., Sestini, P. and Viegi, G. (2013). "Geriatric study in Europe on health effects of air quality in nursing homes (GERIE study) profile: objectives, study protocol and descriptive data." [Multidiscip Respir Med](#) 8: 71.

Armulik, A., Genove, G. and Betsholtz, C. (2011). "Pericytes: developmental, physiological, and pathological perspectives, problems, and promises." [Dev Cell](#) 21: 193-215.

Bale, S., Sunkoju, M., Reddy, S. S., Swamy, V. and Godugu, C. (2016). "Oropharyngeal aspiration of bleomycin: An alternative experimental model of pulmonary fibrosis developed in Swiss mice." [Indian J Pharmacol](#) 48: 643-648.

Balhara, J. and Gounni, A. S. (2012). "The alveolar macrophages in asthma: a double-edged sword." [Mucosal Immunol](#) 5: 605-609.

Ballard, P. L., Lee, J. W., Fang, X., Chapin, C., Allen, L., Segal, M. R., Fischer, H., Illek, B., Gonzales, L. W., Kolla, V. and Matthay, M. A. (2010). "Regulated gene expression in cultured type II cells of adult human lung." [Am J Physiol Lung Cell Mol Physiol](#) 299: L36-50.

Barbayanni, I., Ninou, I., Tzouveleakis, A. and Aidinis, V. (2018). "Bleomycin Revisited: A Direct Comparison of the Intratracheal Micro-Spraying and the Oropharyngeal Aspiration Routes of Bleomycin Administration in Mice." [Front Med \(Lausanne\)](#) 5: 269.

Barkauskas, C. E., Chung, M. I., Fioret, B., Gao, X., Katsura, H. and Hogan, B. L. (2017). "Lung organoids: current uses and future promise." [Development](#) 144: 986-997.

Barkauskas, C. E., Crouce, M. J., Rackley, C. R., Bowie, E. J., Keene, D. R., Stripp, B. R., Randell, S. H., Noble, P. W. and Hogan, B. L. (2013). "Type 2 alveolar cells are stem cells in adult lung." [J Clin Invest](#) 123: 3025-3036.

4.2 Mesenchymal heterogeneity in the lung

Barron, L., Gharib, S. A. and Duffield, J. S. (2016). "Lung Pericytes and Resident Fibroblasts: Busy Multitaskers." Am J Pathol 186: 2519-2531.

Baumgartner, K. B., Samet, J. M., Stidley, C. A., Colby, T. V. and Waldron, J. A. (1997). "Cigarette smoking: a risk factor for idiopathic pulmonary fibrosis." Am J Respir Crit Care Med 155: 242-248.

Becht, E., McInnes, L., Healy, J., Dutertre, C. A., Kwok, I. W. H., Ng, L. G., Ginhoux, F. and Newell, E. W. (2018). "Dimensionality reduction for visualizing single-cell data using UMAP." Nat Biotechnol.

Benoit, M., Desnues, B. and Mege, J. L. (2008). "Macrophage polarization in bacterial infections." J Immunol 181: 3733-3739.

Bivas-Benita, M., Zwier, R., Junginger, H. E. and Borchard, G. (2005). "Non-invasive pulmonary aerosol delivery in mice by the endotracheal route." Eur J Pharm Biopharm 61: 214-218.

Bruderer, R., Bernhardt, O. M., Gandhi, T., Miladinovic, S. M., Cheng, L. Y., Messner, S., Ehrenberger, T., Zanotelli, V., Butscheid, Y., Escher, C., Vitek, O., Rinner, O. and Reiter, L. (2015). "Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues." Mol Cell Proteomics 14: 1400-1410.

Cao, Z., Lis, R., Ginsberg, M., Chavez, D., Shido, K., Rabbany, S. Y., Fong, G. H., Sakmar, T. P., Rafii, S. and Ding, B. S. (2016). "Targeting of the pulmonary capillary vascular niche promotes lung alveolar repair and ameliorates fibrosis." Nat Med 22: 154-162.

Capelli, A., Di Stefano, A., Lusuardi, M., Gnemmi, I. and Donner, C. F. (2002). "Increased macrophage inflammatory protein-1alpha and macrophage inflammatory protein-1beta levels in bronchoalveolar lavage fluid of patients affected by different stages of pulmonary sarcoidosis." Am J Respir Crit Care Med 165: 236-241.

Carolan, B. J., Heguy, A., Harvey, B. G., Leopold, P. L., Ferris, B. and Crystal, R. G. (2006). "Up-regulation of expression of the ubiquitin carboxyl-terminal hydrolase L1 gene in human airway epithelium of cigarette smokers." Cancer Res 66: 10729-10740.

Chandler, D. B. (1990). "Possible mechanisms of bleomycin-induced fibrosis." Clin Chest Med 11: 21-30.

Clevers, H., Loh, K. M. and Nusse, R. (2014). "Stem cell signaling. An integral program for tissue renewal and regeneration: Wnt signaling and stem cell control." Science 346: 1248012.

Collaborators, G. S. (2016). "Measuring the health-related Sustainable Development Goals in 188 countries: a baseline analysis from the Global Burden of Disease Study 2015." Lancet 388: 1813-1850.

Collard, H. R., Calfee, C. S., Wolters, P. J., Song, J. W., Hong, S. B., Brady, S., Ishizaka, A., Jones, K. D., King, T. E., Jr., Matthay, M. A. and Kim, D. S. (2010). "Plasma biomarker profiles in acute exacerbation of idiopathic pulmonary fibrosis." Am J Physiol Lung Cell Mol Physiol 299: L3-7.

Collard, H. R., Ryerson, C. J., Corte, T. J., Jenkins, G., Kondoh, Y., Lederer, D. J., Lee, J. S., Maher, T. M., Wells, A. U., Antoniou, K. M., Behr, J., Brown, K. K., Cottin, V., Flaherty, K. R., Fukuoka, J., Hansell, D. M., Johkoh, T., Kaminski, N., Kim, D. S., Kolb, M., Lynch, D. A., Myers, J. L., Raghu, G., Richeldi, L., Taniguchi, H. and Martinez, F. J. (2016). "Acute Exacerbation of Idiopathic Pulmonary Fibrosis. An International Working Group Report." Am J Respir Crit Care Med 194: 265-275.

Collard, H. R., Ward, A. J., Lanes, S., Courtney Hayflinger, D., Rosenberg, D. M. and Hunsche, E. (2012). "Burden of illness in idiopathic pulmonary fibrosis." J Med Econ 15: 829-835.

Comrie, W. A., Li, S., Boyle, S. and Burkhardt, J. K. (2015). "The dendritic cell cytoskeleton promotes T cell adhesion and activation by constraining ICAM-1 mobility." J Cell Biol 208: 457-473.

Consortium, T. G. O. (2019). "The Gene Ontology Resource: 20 years and still GOing strong." Nucleic Acids Res 47: D330-D338.

- Corfield, A. P. (2015). "Mucins: a biologically relevant glycan barrier in mucosal protection." *Biochim Biophys Acta* 1850: 236-252.
- Costa-Almeida, R., Gomez-Lazaro, M., Ramalho, C., Granja, P. L., Soares, R. and Guerreiro, S. G. (2015). "Fibroblast-endothelial partners for vascularization strategies in tissue engineering." *Tissue Eng Part A* 21: 1055-1065.
- Cottin, V., Hirani, N. A., Hotchkiss, D. L., Nambiar, A. M., Ogura, T., Otaola, M., Skowasch, D., Park, J. S., Poonyagariyagorn, H. K., Wuyts, W. and Wells, A. U. (2018). "Presentation, diagnosis and clinical course of the spectrum of progressive-fibrosing interstitial lung diseases." *Eur Respir Rev* 27.
- Cox, J. and Mann, M. (2008). "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification." *Nat Biotechnol* 26: 1367-1372.
- Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V. and Mann, M. (2011). "Andromeda: a peptide search engine integrated into the MaxQuant environment." *J Proteome Res* 10: 1794-1805.
- Cox, T. R. and Erler, J. T. (2011). "Remodeling and homeostasis of the extracellular matrix: implications for fibrotic diseases and cancer." *Dis Model Mech* 4: 165-178.
- Crapo RO, C. E. (1998). Aging of the respiratory system. *Pulmonary Diseases and Disorders*. F. AP. New York, McGraw-Hill: 251-264.
- Crystal, R. G., Randell, S. H., Engelhardt, J. F., Voynow, J. and Sunday, M. E. (2008). "Airway epithelial cells: current concepts and challenges." *Proc Am Thorac Soc* 5: 772-777.
- Danopoulos, S., Bhattacharya, S., Mariani, T. J. and Al Alam, D. (2020). "Transcriptional characterisation of human lung cells identifies novel mesenchymal lineage markers." *Eur Respir J* 55.
- Degryse, A. L., Tanjore, H., Xu, X. C., Polosukhin, V. V., Jones, B. R., Boomersshine, C. S., Ortiz, C., Sherrill, T. P., McMahon, F. B., Gleaves, L. A., Blackwell, T. S. and Lawson, W. E. (2011). "TGFbeta signaling in lung epithelium regulates bleomycin-induced alveolar injury and fibroblast recruitment." *Am J Physiol Lung Cell Mol Physiol* 300: L887-897.
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T. R. (2013). "STAR: ultrafast universal RNA-seq aligner." *Bioinformatics* 29: 15-21.
- Duclos, G. E., Teixeira, V. H., Autissier, P., Gesthalter, Y. B., Reinders-Luinge, M. A., Terrano, R., Dumas, Y. M., Liu, G., Mazzilli, S. A., Brandsma, C. A., van den Berge, M., Janes, S. M., Timens, W., Lenburg, M. E., Spira, A., Campbell, J. D. and Beane, J. (2019). "Characterizing smoking-induced transcriptional heterogeneity in the human bronchial epithelium at single-cell resolution." *Sci Adv* 5: eaaw3413.
- Egger, C., Cannel, C., Gerard, C., Jarman, E., Jarai, G., Feige, A., Suply, T., Micard, A., Dunbar, A., Tigani, B. and Beckmann, N. (2013). "Administration of bleomycin via the oropharyngeal aspiration route leads to sustained lung fibrosis in mice and rats as quantified by UTE-MRI and histology." *PLoS One* 8: e63432.
- El Agha, E., Moiseenko, A., Kheirollahi, V., De Langhe, S., Crnkovic, S., Kwapiszewska, G., Szibor, M., Kosanovic, D., Schwind, F., Schermuly, R. T., Henneke, I., MacKenzie, B., Quantius, J., Herold, S., Ntokou, A., Ahlbrecht, K., Braun, T., Morty, R. E., Gunther, A., Seeger, W. and Belluscio, S. (2017). "Two-Way Conversion between Lipogenic and Myogenic Fibroblastic Phenotypes Marks the Progression and Resolution of Lung Fibrosis." *Cell Stem Cell* 20: 571.
- Englert, J. M., Hanford, L. E., Kaminski, N., Tobolewski, J. M., Tan, R. J., Fattman, C. L., Ramsgaard, L., Richards, T. J., Loutaev, I., Nawroth, P. P., Kasper, M., Bierhaus, A. and Oury, T. D. (2008). "A role for the receptor for advanced glycation end products in idiopathic pulmonary fibrosis." *Am J Pathol* 172: 583-591.
- Evans, M. J., Cabral, L. J., Stephens, R. J. and Freeman, G. (1975). "Transformation of alveolar type 2 cells to type 1 cells following exposure to NO₂." *Exp Mol Pathol* 22: 142-150.

- Femino, A. M., Fay, F. S., Fogarty, K. and Singer, R. H. (1998). "Visualization of single RNA transcripts in situ." Science 280: 585-590.
- Fernandez, I. E. and Eickelberg, O. (2012). "New cellular and molecular mechanisms of lung injury and fibrosis in idiopathic pulmonary fibrosis." Lancet 380: 680-688.
- Fischer, A. and du Bois, R. (2012). "Interstitial lung disease in connective tissue disorders." Lancet 380: 689-698.
- Foster, M. W., Morrison, L. D., Todd, J. L., Snyder, L. D., Thompson, J. W., Soderblom, E. J., Plonk, K., Weinhold, K. J., Townsend, R., Minnich, A. and Moseley, M. A. (2015). "Quantitative proteomics of bronchoalveolar lavage fluid in idiopathic pulmonary fibrosis." J Proteome Res 14: 1238-1249.
- Frank, D. B., Peng, T., Zepp, J. A., Snitow, M., Vincent, T. L., Penkala, I. J., Cui, Z., Herriges, M. J., Morley, M. P., Zhou, S., Lu, M. M. and Morrissey, E. E. (2016). "Emergence of a Wave of Wnt Signaling that Regulates Lung Alveologenesis by Controlling Epithelial Self-Renewal and Differentiation." Cell Rep 17: 2312-2325.
- Frankel, S. K., Cosgrove, G. P., Cha, S. I., Cool, C. D., Wynes, M. W., Edelman, B. L., Brown, K. K. and Riches, D. W. (2006). "TNF-alpha sensitizes normal and fibrotic human lung fibroblasts to Fas-induced apoptosis." Am J Respir Cell Mol Biol 34: 293-304.
- Franks, T. J., Colby, T. V., Travis, W. D., Tuder, R. M., Reynolds, H. Y., Brody, A. R., Cardoso, W. V., Crystal, R. G., Drake, C. J., Engelhardt, J., Frid, M., Herzog, E., Mason, R., Phan, S. H., Randell, S. H., Rose, M. C., Stevens, T., Serge, J., Sunday, M. E., Voynow, J. A., Weinstein, B. M., Whitsett, J. and Williams, M. C. (2008). "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.
- Gerber, A. N. (2015). "Glucocorticoids and the Lung." Adv Exp Med Biol 872: 279-298.
- Geyer, P. E., Holdt, L. M., Teupser, D. and Mann, M. (2017). "Revisiting biomarker discovery by plasma proteomics." Mol Syst Biol 13: 942.
- Geyer, P. E., Kulak, N. A., Pichler, G., Holdt, L. M., Teupser, D. and Mann, M. (2016). "Plasma Proteome Profiling to Assess Human Health and Disease." Cell Syst 2: 185-195.
- Geyer, P. E., Voytik, E., Treit, P. V., Doll, S., Kleinhempel, A., Niu, L., Muller, J. B., Buchholtz, M. L., Bader, J. M., Teupser, D., Holdt, L. M. and Mann, M. (2019). "Plasma Proteome Profiling to detect and avoid sample-related biases in biomarker studies." EMBO Mol Med 11: e10427.
- Greene, K. E., King, T. E., Jr., Kuroki, Y., Bucher-Bartelson, B., Hunninghake, G. W., Newman, L. S., Nagae, H. and Mason, R. J. (2002). "Serum surfactant proteins-A and -D as biomarkers in idiopathic pulmonary fibrosis." Eur Respir J 19: 439-446.
- Griffin, M., Bhandari, R., Hamilton, G., Chan, Y. C. and Powell, J. T. (1993). "Alveolar type II cell-fibroblast interactions, synthesis and secretion of surfactant and type I collagen." J Cell Sci 105 (Pt 2): 423-432.
- Group, B. D. W. (2001). "Biomarkers and surrogate endpoints: preferred definitions and conceptual framework." Clin Pharmacol Ther 69: 89-95.
- Guler, S. A. and Ryerson, C. J. (2018). "Unclassifiable interstitial lung disease: from phenotyping to possible treatments." Curr Opin Pulm Med 24: 461-468.
- Gurtner, G. C., Werner, S., Barrandon, Y. and Longaker, M. T. (2008). "Wound repair and regeneration." Nature 453: 314-321.
- Haagsman, H. P. and Diemel, R. V. (2001). "Surfactant-associated proteins: functions and structural variation." Comp Biochem Physiol A Mol Integr Physiol 129: 91-108.
- Habermann, A. C., Gutierrez, A. J., Bui, L. T., Yahn, S. L., Winters, N. I., Calvi, C. L., Peter, L., Chung, M.-I., Taylor, C. J., Jetter, C., Raju, L., Roberson, J., Ding, G., Wood, L., Sucre, J. M. S., Richmond, B. W., Serezani, A. P., McDonnell, W. J., Mallal, S. B., Bacchetta,

- M. J., Loyd, J. E., Shaver, C. M., Ware, L. B., Bremner, R., Walia, R., Blackwell, T. S., Banovich, N. E. and Kropski, J. A. (2019). "Single-cell RNA-sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis." [bioRxiv: 753806](#).
- Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., Huang, D., Xu, Y., Huang, W., Jiang, M., Jiang, X., Mao, J., Chen, Y., Lu, C., Xie, J., Fang, Q., Wang, Y., Yue, R., Li, T., Huang, H., Orkin, S. H., Yuan, G. C., Chen, M. and Guo, G. (2018). "Mapping the Mouse Cell Atlas by Microwell-Seq." [Cell](#) 173: 1307.
- He, L., Vanlandewijck, M., Mae, M. A., Andrae, J., Ando, K., Del Gaudio, F., Nahar, K., Lebouvier, T., Lavina, B., Gouveia, L., Sun, Y., Raschperger, E., Segerstolpe, A., Liu, J., Gustafsson, S., Rasanen, M., Zarb, Y., Mochizuki, N., Keller, A., Lendahl, U. and Betsholtz, C. (2018). "Single-cell RNA sequencing of mouse brain and lung vascular and vessel-associated cell types." [Sci Data](#) 5: 180160.
- Hecker, L. and Thannickal, V. J. (2011). "Nonresolving fibrotic disorders: idiopathic pulmonary fibrosis as a paradigm of impaired tissue regeneration." [Am J Med Sci](#) 341: 431-434.
- Hein, M. S., Kirti, Cox, Juergen; Mann, Matthias (2013). Chapter 1 - Proteomic Analysis of Cellular Systems, [Handbook of Systems Biology](#).
- Hinz, B., Celetta, G., Tomasek, J. J., Gabbiani, G. and Chaponnier, C. (2001). "Alpha-smooth muscle actin expression upregulates fibroblast contractile activity." [Mol Biol Cell](#) 12: 2730-2741.
- Hochreiter-Hufford, A. and Ravichandran, K. S. (2013). "Clearing the dead: apoptotic cell sensing, recognition, engulfment, and digestion." [Cold Spring Harb Perspect Biol](#) 5: a008748.
- Hoyle, R. K., Derrett-Smith, E. C., Khan, K., Shiwen, X., Howat, S. L., Wells, A. U., Abraham, D. J. and Denton, C. P. (2011). "An essential role for resident fibroblasts in experimental lung fibrosis is defined by lineage-specific deletion of high-affinity type II transforming growth factor beta receptor." [Am J Respir Crit Care Med](#) 183: 249-261.
- Hsu, Y.-C. and Fuchs, E. (2012). "A family business: stem cell progeny join the niche to regulate homeostasis." [Nature reviews. Molecular cell biology](#) 13: 103-114.
- Hsu, Y.-C., Pasolli, H. A. and Fuchs, E. (2011). "Dynamics between stem cells, niche, and progeny in the hair follicle." [Cell](#) 144: 92-105.
- Huang, S. (2009). "Non-genetic heterogeneity of cells in development: more than just noise." [Development](#) 136: 3853-3862.
- Hughes, C. C. (2008). "Endothelial-stromal interactions in angiogenesis." [Curr Opin Hematol](#) 15: 204-209.
- Hung, C., Linn, G., Chow, Y. H., Kobayashi, A., Mittelsteadt, K., Altemeier, W. A., Gharib, S. A., Schnapp, L. M. and Duffield, J. S. (2013). "Role of lung pericytes and resident fibroblasts in the pathogenesis of pulmonary fibrosis." [Am J Respir Crit Care Med](#) 188: 820-830.
- Hyltdgaard, C., Hilberg, O. and Bendstrup, E. (2014). "How does comorbidity influence survival in idiopathic pulmonary fibrosis?" [Respir Med](#) 108: 647-653.
- Jo, H. E., Troy, L. K., Keir, G., Chambers, D. C., Holland, A., Goh, N., Wilsher, M., de Boer, S., Moodley, Y., Grainge, C., Whitford, H., Chapman, S., Reynolds, P. N., Glaspole, I., Beatson, D., Jones, L., Hopkins, P. and Corte, T. J. (2017). "Treatment of idiopathic pulmonary fibrosis in Australia and New Zealand: A position statement from the Thoracic Society of Australia and New Zealand and the Lung Foundation Australia." [Respirology](#) 22: 1436-1458.
- Joeannes, R., Just, A. C., Marioni, R. E., Pilling, L. C., Reynolds, L. M., Mandaviya, P. R., Guan, W., Xu, T., Elks, C. E., Aslibekyan, S., Moreno-Macias, H., Smith, J. A., Brody, J. A., Dhingra, R., Yousefi, P., Pankow, J. S., Kunze, S., Shah, S. H., McRae, A. F., Lohman, K., Sha, J., Absher, D. M., Ferrucci, L., Zhao, W., Demerath, E. W., Bressler, J., Grove, M. L., Huan, T., Liu, C., Mendelson, M. M., Yao, C., Kiel, D. P., Peters, A., Wang-Sattler, R., Visscher, P. M., Wray, N. R., Starr, J. M., Ding, J., Rodriguez, C. J., Wareham, N. J., Irvin, M. R., Zhi, D., Barrdahl, M., Vineis, P., Ambatipudi, S., Uitterlinden, A. G., Hofman, A., Schwartz, J., Colicino, E., Hou, L., Vokonas, P. S., Hernandez, D. G., Singleton, A. B., Bandinelli, S., Turner, S. T., Ware, E. B., Smith, A. K., Klengel, T., Binder, E. B., Psaty, B. M., Taylor, K. D., Gharib, S. A., Swenson, B. R., Liang, L., DeMeo, D. L., O'Connor, G. T., Herceg, Z., Ressler, K. J.,

- Conneely, K. N., Sotoodehnia, N., Kardia, S. L., Melzer, D., Baccarelli, A. A., van Meurs, J. B., Romieu, I., Arnett, D. K., Ong, K. K., Liu, Y., Waldenberger, M., Deary, I. J., Fornage, M., Levy, D. and London, S. J. (2016). "Epigenetic Signatures of Cigarette Smoking." Circ Cardiovasc Genet 9: 436-447.
- Jordana, M., Sarnstrand, B., Sime, P. J. and Ramis, I. (1994). "Immune-inflammatory functions of fibroblasts." Eur Respir J 7: 2212-2222.
- Kammerl, I. E., Caniard, A., Merl-Pham, J., Ben-Nissan, G., Mayr, C. H., Mossina, A., Geerlof, A., Eickelberg, O., Hauck, S. M., Sharon, M. and Meiners, S. (2019). "Dissecting the molecular effects of cigarette smoke on proteasome function." J Proteomics 193: 1-9.
- Kester, L. and van Oudenaarden, A. (2018). "Single-Cell Transcriptomics Meets Lineage Tracing." Cell Stem Cell 23: 166-179.
- Kim, K. K., Kugler, M. C., Wolters, P. J., Robillard, L., Galvez, M. G., Brumwell, A. N., Sheppard, D. and Chapman, H. A. (2006). "Alveolar epithelial cell mesenchymal transition develops in vivo during pulmonary fibrosis and is regulated by the extracellular matrix." Proc Natl Acad Sci U S A 103: 13180-13185.
- Kinder, B. W., Brown, K. K., McCormack, F. X., Ix, J. H., Kervitsky, A., Schwarz, M. I. and King, T. E., Jr. (2009). "Serum surfactant protein-A is a strong predictor of early mortality in idiopathic pulmonary fibrosis." Chest 135: 1557-1563.
- King, C. S. and Nathan, S. D. (2017). "Idiopathic pulmonary fibrosis: effects and optimal management of comorbidities." Lancet Respir Med 5: 72-84.
- King, T. E., Jr., Bradford, W. Z., Castro-Bernardini, S., Fagan, E. A., Glaspole, I., Glassberg, M. K., Gorina, E., Hopkins, P. M., Kardatzke, D., Lancaster, L., Lederer, D. J., Nathan, S. D., Pereira, C. A., Sahn, S. A., Sussman, R., Swigris, J. J. and Noble, P. W. (2014). "A phase 3 trial of pirfenidone in patients with idiopathic pulmonary fibrosis." N Engl J Med 370: 2083-2092.
- King, T. E., Jr., Tooze, J. A., Schwarz, M. I., Brown, K. R. and Cherniack, R. M. (2001). "Predicting survival in idiopathic pulmonary fibrosis: scoring system and survival model." Am J Respir Crit Care Med 164: 1171-1181.
- Kisseleva, T. and Brenner, D. A. (2008). "Fibrogenesis of parenchymal organs." Proc Am Thorac Soc 5: 338-342.
- Kokosi, M. A., Margaritopoulos, G. A. and Wells, A. U. (2018). "Personalised medicine in interstitial lung diseases: Number 6 in the Series "Personalised medicine in respiratory diseases" Edited by Renaud Louis and Nicolas Roche." Eur Respir Rev 27.
- Kollert, F., Probst, C., Muller-Quernheim, J., Zissel, G. and Prasse, A. (2009). "CCL18 production is decreased in alveolar macrophages from cigarette smokers." Inflammation 32: 163-168.
- Kreuter, M., Ehlers-Tenenbaum, S., Palmowski, K., Bruhwiler, J., Oltmanns, U., Muley, T., Heussel, C. P., Warth, A., Kolb, M. and Herth, F. J. (2016). "Impact of Comorbidities on Mortality in Patients with Idiopathic Pulmonary Fibrosis." PLoS One 11: e0151425.
- Kreuter, M., Ehlers-Tenenbaum, S., Schaaf, M., Oltmanns, U., Palmowski, K., Hoffmann, H., Schnabel, P. A., Heussel, C. P., Puderbach, M., Herth, F. J. and Warth, A. (2015). "Treatment and outcome of lung cancer in idiopathic interstitial pneumonias." Sarcoidosis Vasc Diffuse Lung Dis 31: 266-274.
- Kulak, N. A., Geyer, P. E. and Mann, M. (2017). "Loss-less Nano-fractionator for High Sensitivity, High Coverage Proteomics." Mol Cell Proteomics 16: 694-705.
- Laitinen, A. and Laitinen, L. A. (1994). "Airway morphology: epithelium/basement membrane." Am J Respir Crit Care Med 150: S14-17.
- Laniado-Laborin, R. (2009). "Smoking and chronic obstructive pulmonary disease (COPD). Parallel epidemics of the 21 century." International journal of environmental research and public health 6: 209-224.
- Lee, J. H., Daugharthy, E. R., Scheiman, J., Kalhor, R., Ferrante, T. C., Terry, R., Turczyk, B. M., Yang, J. L., Lee, H. S., Aach, J., Zhang, K. and Church, G. M. (2015). "Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and

tissues." [Nat Protoc](#) 10: 442-458.

Lee, J. H., Tammela, T., Hofree, M., Choi, J., Marjanovic, N. D., Han, S., Canner, D., Wu, K., Paschini, M., Bhang, D. H., Jacks, T., Regev, A. and Kim, C. F. (2017). "Anatomically and Functionally Distinct Lung Mesenchymal Populations Marked by Lgr5 and Lgr6." [Cell](#) 170: 1149-1163 e1112.

Lehmann, M., Korfei, M., Mutze, K., Klee, S., Skronska-Wasek, W., Alsafadi, H. N., Ota, C., Costa, R., Schiller, H. B., Lindner, M., Wagner, D. E., Gunther, A. and Konigshoff, M. (2017). "Senolytic drugs target alveolar epithelial cell function and attenuate experimental lung fibrosis ex vivo." [Eur Respir J](#) 50.

Ley, B., Swigris, J., Day, B. M., Stauffer, J. L., Raimundo, K., Chou, W. and Collard, H. R. (2017). "Pirfenidone Reduces Respiratory-related Hospitalizations in Idiopathic Pulmonary Fibrosis." [Am J Respir Crit Care Med](#) 196: 756-761.

Lun, A. T., McCarthy, D. J. and Marioni, J. C. (2016). "A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor." [F1000Res](#) 5: 2122.

Macosko, E. Z., Basu, A., Satija, R., Nemes, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A. and McCarroll, S. A. (2015). "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets." [Cell](#) 161: 1202-1214.

Maher, T. M., Evans, I. C., Bottoms, S. E., Mercer, P. F., Thorley, A. J., Nicholson, A. G., Laurent, G. J., Tetley, T. D., Chambers, R. C. and McAnulty, R. J. (2010). "Diminished prostaglandin E2 contributes to the apoptosis paradox in idiopathic pulmonary fibrosis." [Am J Respir Crit Care Med](#) 182: 73-82.

Maher, T. M., Stowasser, S., Nishioka, Y., White, E. S., Cottin, V., Noth, I., Selman, M., Rohr, K. B., Michael, A., Ittrich, C., Diefenbach, C. and Jenkins, R. G. (2019). "Biomarkers of extracellular matrix turnover in patients with idiopathic pulmonary fibrosis given nintedanib (INMARK study): a randomised, placebo-controlled study." [Lancet Respir Med](#) 7: 771-779.

Martinez, F. J., Collard, H. R., Pardo, A., Raghu, G., Richeldi, L., Selman, M., Swigris, J. J., Taniguchi, H. and Wells, A. U. (2017). "Idiopathic pulmonary fibrosis." [Nat Rev Dis Primers](#) 3: 17074.

Mayr, C. H., Simon, L. M., Leuschner, G., Ansari, M., Geyer, P. E., Angelidis, I., Strunz, M., Singh, P., Kneidinger, N., Reichenberger, F., Silbernagel, E., Boehm, S., Adler, H., Hilgendorff, A., Lindner, M., Prasse, A., Behr, J., Mann, M., Eickelberg, O., Theis, F. J. and Schiller, H. B. (2020). "Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids." [medRxiv](#): 2020.2001.2021.20018358.

McDonough, J. E., Ahangari, F., Li, Q., Jain, S., Verleden, S. E., Herazo-Maya, J., Vukmirovic, M., Deluili, G., Tzouveleakis, A., Tanabe, N., Chu, F., Yan, X., Verschakelen, J., Homer, R. J., Manatakis, D. V., Zhang, J., Ding, J., Maes, K., De Sadeleer, L., Vos, R., Neyrinck, A., Benos, P. V., Bar-Joseph, Z., Tantin, D., Hogg, J. C., Vanaudenaerde, B. M., Wuyts, W. A. and Kaminski, N. (2019). "Transcriptional regulatory model of fibrosis progression in the human lung." [JCI Insight](#) 4.

Meadors, M., Floyd, J. and Perry, M. C. (2006). "Pulmonary toxicity of chemotherapy." [Semin Oncol](#) 33: 98-105.

Mereu, E., Iacono, G., Guillaumet-Adkins, A., Moutinho, C., Lunazzi, G., Santos, C. P., Miguel-Escalada, I., Ferrer, J., Real, F. X., Gut, I. and Heyn, H. (2018). "matchScore: Matching Single-Cell Phenotypes Across Tools and Experiments." [bioRxiv](#): 314831.

Mi, H., Muruganujan, A., Ebert, D., Huang, X. and Thomas, P. D. (2019). "PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools." [Nucleic Acids Res](#) 47: D419-D426.

Moeller, A., Ask, K., Warburton, D., Gauldie, J. and Kolb, M. (2008). "The bleomycin animal model: a useful tool to investigate treatment options for idiopathic pulmonary fibrosis?" [Int J Biochem Cell Biol](#) 40: 362-382.

- Moodley, Y. P., Misso, N. L., Scaffidi, A. K., Fogel-Petrovic, M., McAnulty, R. J., Laurent, G. J., Thompson, P. J. and Knight, D. A. (2003). "Inverse effects of interleukin-6 on apoptosis of fibroblasts from pulmonary fibrosis and normal lungs." Am J Respir Cell Mol Biol 29: 490-498.
- Morse, C., Tabib, T., Sembrat, J., Buschur, K. L., Bittar, H. T., Valenzi, E., Jiang, Y., Kass, D. J., Gibson, K., Chen, W., Mora, A., Benos, P. V., Rojas, M. and Lafyatis, R. (2019). "Proliferating SPP1/MERTK-expressing macrophages in idiopathic pulmonary fibrosis." Eur Respir J 54.
- Mouratis, M. A. and Aidinis, V. (2011). "Modeling pulmonary fibrosis with bleomycin." Curr Opin Pulm Med 17: 355-361.
- Naba, A., Clauser, K. R., Hoersch, S., Liu, H., Carr, S. A. and Hynes, R. O. (2012). "The matrisome: in silico definition and in vivo characterization by proteomics of normal and tumor extracellular matrices." Mol Cell Proteomics 11: M111 014647.
- Nagendran, M., Riordan, D. P., Harbury, P. B. and Desai, T. J. (2018). "Automated cell-type classification in intact tissues by single-cell molecular profiling." Elife 7.
- Nathan, S. D., Albera, C., Bradford, W. Z., Costabel, U., Glaspole, I., Glassberg, M. K., Kardatzke, D. R., Daigl, M., Kirchgaessler, K. U., Lancaster, L. H., Lederer, D. J., Pereira, C. A., Swigris, J. J., Valeyre, D. and Noble, P. W. (2017). "Effect of pirfenidone on mortality: pooled analyses and meta-analyses of clinical trials in idiopathic pulmonary fibrosis." Lancet Respir Med 5: 33-41.
- Neighbors, M., Cabanski, C. R., Ramalingam, T. R., Sheng, X. R., Tew, G. W., Gu, C., Jia, G., Peng, K., Ray, J. M., Ley, B., Wolters, P. J., Collard, H. R. and Arron, J. R. (2018). "Prognostic and predictive biomarkers for patients with idiopathic pulmonary fibrosis treated with pirfenidone: post-hoc assessment of the CAPACITY and ASCEND trials." Lancet Respir Med 6: 615-626.
- Neurohr, C. and Behr, J. (2015). "Changes in the current classification of IIP: A critical review." Respirology 20: 699-704.
- Ng, A. W., Bidani, A. and Heming, T. A. (2004). "Innate host defense of the lung: effects of lung-lining fluid pH." Lung 182: 297-317.
- Niu, L., Geyer, P. E., Wewer Albrechtsen, N. J., Gluud, L. L., Santos, A., Doll, S., Treit, P. V., Holst, J. J., Knop, F. K., Vilsboll, T., Junker, A., Sachs, S., Stemmer, K., Muller, T. D., Tschop, M. H., Hofmann, S. M. and Mann, M. (2019). "Plasma proteome profiling discovers novel proteins associated with non-alcoholic fatty liver disease." Mol Syst Biol 15: e8793.
- Noble, P. W., Albera, C., Bradford, W. Z., Costabel, U., du Bois, R. M., Fagan, E. A., Fishman, R. S., Glaspole, I., Glassberg, M. K., Lancaster, L., Lederer, D. J., Leff, J. A., Nathan, S. D., Pereira, C. A., Swigris, J. J., Valeyre, D. and King, T. E., Jr. (2016). "Pirfenidone for idiopathic pulmonary fibrosis: analysis of pooled data from three multinational phase 3 trials." Eur Respir J 47: 243-253.
- Noble, P. W., Albera, C., Bradford, W. Z., Costabel, U., Glassberg, M. K., Kardatzke, D., King, T. E., Jr., Lancaster, L., Sahn, S. A., Szwarzberg, J., Valeyre, D. and du Bois, R. M. (2011). "Pirfenidone in patients with idiopathic pulmonary fibrosis (CAPACITY): two randomised trials." Lancet 377: 1760-1769.
- O'Sullivan, J. M., Huddart, R. A., Norman, A. R., Nicholls, J., Dearnaley, D. P. and Horwich, A. (2003). "Predicting the risk of bleomycin lung toxicity in patients with germ-cell tumours." Ann Oncol 14: 91-96.
- Patterson, K. C., Shah, R. J., Porteous, M. K., Christie, J. D., D'Errico, C. A., Chadwick, M., Triano, M. J., Deshpande, C., Rossman, M. D., Litzky, L. A., Kreider, M. and Miller, W. T., Jr. (2017). "Interstitial Lung Disease in the Elderly." Chest 151: 838-844.
- Pedersen-Bjergaard, J., Daugaard, G., Hansen, S. W., Philip, P., Larsen, S. O. and Rorth, M. (1991). "Increased risk of myelodysplasia and leukaemia after etoposide, cisplatin, and bleomycin for germ-cell tumours." Lancet 338: 359-363.
- Peng, T., Frank, D. B., Kadzik, R. S., Morley, M. P., Rathi, K. S., Wang, T., Zhou, S., Cheng, L., Lu, M. M. and Morrissey, E. E. (2015). "Hedgehog actively maintains adult lung quiescence and regulates repair and regeneration." Nature 526: 578-582.

Perperoglou, A., Sauerbrei, W., Abrahamowicz, M. and Schmid, M. (2019). "A review of spline function procedures in R." BMC Med Res Methodol 19: 46.

Petzold, T., Thienel, M., Dannenberg, L., Mourikis, P., Helten, C., Ayhan, A., M'Pembete, R., Achilles, A., Trojovky, K., Konsek, D., Zhang, Z., Regenauer, R., Pircher, J., Ehrlich, A., Lusebrink, E., Nicolai, L., Stocker, T. J., Brandl, R., Roschenthaler, F., Strecker, J., Saleh, I., Spannagl, M., Mayr, C. H., Schiller, H. B., Jung, C., Gerdes, N., Hoffmann, T., Levkau, B., Hohlfeld, T., Zeus, T., Schulz, C., Kelm, M. and Polzin, A. (2020). "Rivaroxaban Reduces Arterial Thrombosis by Inhibition of FXa-Driven Platelet Activation via Protease Activated Receptor-1." Circ Res 126: 486-500.

Phillips, R. J., Burdick, M. D., Hong, K., Lutz, M. A., Murray, L. A., Xue, Y. Y., Belperio, J. A., Keane, M. P. and Strieter, R. M. (2004). "Circulating fibrocytes traffic to the lungs in response to CXCL12 and mediate fibrosis." J Clin Invest 114: 438-446.

Polanski, K., Young, M. D., Miao, Z., Meyer, K. B., Teichmann, S. A. and Park, J. E. (2020). "BBKNN: fast batch alignment of single cell transcriptomes." Bioinformatics 36: 964-965.

Prasse, A., Binder, H., Schupp, J. C., Kayser, G., Bargagli, E., Jaeger, B., Hess, M., Rittinghausen, S., Vuga, L., Lynn, H., Violette, S., Jung, B., Quast, K., Vanaudenaerde, B., Xu, Y., Hohlfeld, J. M., Krug, N., Herazo-Maya, J. D., Rottoli, P., Wuyts, W. A. and Kaminski, N. (2019). "BAL Cell Gene Expression Is Indicative of Outcome and Airway Basal Cell Involvement in Idiopathic Pulmonary Fibrosis." Am J Respir Crit Care Med 199: 622-630.

Prasse, A., Pechkovsky, D. V., Toews, G. B., Schafer, M., Eggeling, S., Ludwig, C., Germann, M., Kollert, F., Zissel, G. and Muller-Quernheim, J. (2007). "CCL18 as an indicator of pulmonary fibrotic activity in idiopathic interstitial pneumonias and systemic sclerosis." Arthritis Rheum 56: 1685-1693.

Rafii, R., Juarez, M. M., Albertson, T. E. and Chan, A. L. (2013). "A review of current and novel therapies for idiopathic pulmonary fibrosis." J Thorac Dis 5: 48-73.

Raghu, G., Collard, H. R., Egan, J. J., Martinez, F. J., Behr, J., Brown, K. K., Colby, T. V., Cordier, J. F., Flaherty, K. R., Lasky, J. A., Lynch, D. A., Ryu, J. H., Swigris, J. J., Wells, A. U., Ancochea, J., Bouros, D., Carvalho, C., Costabel, U., Ebina, M., Hansell, D. M., Johkoh, T., Kim, D. S., King, T. E., Jr., Kondoh, Y., Myers, J., Muller, N. L., Nicholson, A. G., Richeldi, L., Selman, M., Dudden, R. F., Griss, B. S., Protzko, S. L. and Schunemann, H. J. (2011). "An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management." Am J Respir Crit Care Med 183: 788-824.

Raghu, G., Collard, H. R., Egan, J. J., Martinez, F. J., Behr, J., Brown, K. K., Colby, T. V., Cordier, J. F., Flaherty, K. R., Lasky, J. A., Lynch, D. A., Ryu, J. H., Swigris, J. J., Wells, A. U., Ancochea, J., Bouros, D., Carvalho, C., Costabel, U., Ebina, M., Hansell, D. M., Johkoh, T., Kim, D. S., King, T. E., Jr., Kondoh, Y., Myers, J., Muller, N. L., Nicholson, A. G., Richeldi, L., Selman, M., Dudden, R. F., Griss, B. S., Protzko, S. L., Schunemann, H. J. and Fibrosis, A. E. J. A. C. o. I. P. (2011). "An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management." Am J Respir Crit Care Med 183: 788-824.

Raghu, G., Rochweg, B., Zhang, Y., Garcia, C. A., Azuma, A., Behr, J., Brozek, J. L., Collard, H. R., Cunningham, W., Homma, S., Johkoh, T., Martinez, F. J., Myers, J., Protzko, S. L., Richeldi, L., Rind, D., Selman, M., Theodore, A., Wells, A. U., Hoogsteden, H. and Schunemann, H. J. (2015). "An Official ATS/ERS/JRS/ALAT Clinical Practice Guideline: Treatment of Idiopathic Pulmonary Fibrosis. An Update of the 2011 Clinical Practice Guideline." Am J Respir Crit Care Med 192: e3-19.

Ramilowski, J. A., Goldberg, T., Harshbarger, J., Kloppmann, E., Lizio, M., Satagopam, V. P., Itoh, M., Kawaji, H., Carninci, P., Rost, B. and Forrest, A. R. (2015). "A draft network of ligand-receptor-mediated multicellular signalling in human." Nat Commun 6: 7866.

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Gottgens, B., Hacohen, N., Haniffa, M., Hemberg, M., Kim, S., Klenerman, P., Kriegstein, A., Lein, E., Linnarsson, S., Lundberg, E., Lundeberg, J., Majumder, P., Marioni, J. C., Merad, M., Mhlanga, M., Nawijn, M., Netea, M., Nolan, G., Pe'er, D., Phillipakis, A., Ponting, C. P., Quake, S., Reik, W., Rozenblatt-Rosen, O., Sanes, J., Satija, R., Schumacher, T. N., Shalek, A., Shapiro, E., Sharma, P., Shin, J. W., Stegle, O., Stratton, M., Stubbington, M. J. T., Theis, F. J., Uhlen, M., van Oudenaarden, A., Wagner, A., Watt, F., Weissman, J., Wold, B., Xavier, R. and

Yosef, N. (2017). "The Human Cell Atlas." Elife 6.

Renkin, E. M. (1992). "Cellular and intercellular transport pathways in exchange vessels." Am Rev Respir Dis 146: S28-31.

Reyfman, P. A., Walter, J. M., Joshi, N., Anekalla, K. R., McQuattie-Pimentel, A. C., Chiu, S., Fernandez, R., Akbarpour, M., Chen, C.-I., Ren, Z., Verma, R., Abdala-Valencia, H., Nam, K., Chi, M., Han, S., Gonzalez-Gonzalez, F. J., Soberanes, S., Watanabe, S., Williams, K. J. N., Flozak, A. S., Nicholson, T. T., Morgan, V. K., Hrusch, C. L., Guzy, R. D., Bonham, C. A., Sperling, A. I., Bag, R., Hamanaka, R. B., Mutlu, G. M., Yeldandi, A. V., Marshall, S. A., Shilatifard, A., Amaral, L. A. N., Perlman, H., Sznajder, J. I., Winter, D. R., Hinchcliff, M., Argento, A. C., Gillespie, C. T., D'Amico Dematte, J., Jain, M., Singer, B. D., Ridge, K. M., Gottardi, C. J., Lam, A. P., Bharat, A., Bhorade, S. M., Budinger, G. R. S. and Misharin, A. V. (2018). "Single-Cell Transcriptomic Analysis of Human Lung Reveals Complex Multicellular Changes During Pulmonary Fibrosis." bioRxiv: 296608.

Richeldi, L., Costabel, U., Selman, M., Kim, D. S., Hansell, D. M., Nicholson, A. G., Brown, K. K., Flaherty, K. R., Noble, P. W., Raghu, G., Brun, M., Gupta, A., Juhel, N., Kluglich, M. and du Bois, R. M. (2011). "Efficacy of a tyrosine kinase inhibitor in idiopathic pulmonary fibrosis." N Engl J Med 365: 1079-1087.

Richeldi, L., du Bois, R. M., Raghu, G., Azuma, A., Brown, K. K., Costabel, U., Cottin, V., Flaherty, K. R., Hansell, D. M., Inoue, Y., Kim, D. S., Kolb, M., Nicholson, A. G., Noble, P. W., Selman, M., Taniguchi, H., Brun, M., Le Maulf, F., Girard, M., Stowasser, S., Schlenker-Herceg, R., Disse, B. and Collard, H. R. (2014). "Efficacy and safety of nintedanib in idiopathic pulmonary fibrosis." N Engl J Med 370: 2071-2082.

Rock, J. R., Barkauskas, C. E., Cronic, M. J., Xue, Y., Harris, J. R., Liang, J., Noble, P. W. and Hogan, B. L. (2011). "Multiple stromal populations contribute to pulmonary fibrosis without evidence for epithelial to mesenchymal transition." Proc Natl Acad Sci U S A 108: E1475-1483.

Rockey, D. C., Bell, P. D. and Hill, J. A. (2015). "Fibrosis--A Common Pathway to Organ Injury and Failure." N Engl J Med 373: 96.

Russell, A. M., Adamali, H., Molyneaux, P. L., Lukey, P. T., Marshall, R. P., Renzoni, E. A., Wells, A. U. and Maher, T. M. (2016). "Daily Home Spirometry: An Effective Tool for Detecting Progression in Idiopathic Pulmonary Fibrosis." Am J Respir Crit Care Med 194: 989-997.

Ryerson, C. J., Urbania, T. H., Richeldi, L., Mooney, J. J., Lee, J. S., Jones, K. D., Elicker, B. M., Koth, L. L., King, T. E., Jr., Wolters, P. J. and Collard, H. R. (2013). "Prevalence and prognosis of unclassifiable interstitial lung disease." Eur Respir J 42: 750-757.

Saxman, S. B., Nichols, C. R. and Einhorn, L. H. (1997). "Pulmonary toxicity in patients with advanced-stage germ cell tumors receiving bleomycin with and without granulocyte colony stimulating factor." Chest 111: 657-660.

Schiller, H. B., Fernandez, I. E., Burgstaller, G., Schaab, C., Scheltema, R. A., Schwarzmayr, T., Strom, T. M., Eickelberg, O. and Mann, M. (2015). "Time- and compartment-resolved proteome profiling of the extracellular niche in lung injury and repair." Mol Syst Biol 11: 819.

Schiller, H. B., Mayr, C. H., Leuschner, G., Strunz, M., Staab-Weijnitz, C., Preisendorfer, S., Eckes, B., Moinzadeh, P., Krieg, T., Schwartz, D. A., Hatz, R.

A., Behr, J., Mann, M. and Eickelberg, O. (2017). "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.

Schiller, H. B., Montoro, D. T., Simon, L. M., Rawlins, E. L., Meyer, K. B., Strunz, M., Vieira Braga, F. A., Timens, W., Koppelman, G. H., Budinger, G. R. S., Burgess, J. K., Waghay, A., van den Berge, M., Theis, F. J., Regev, A., Kaminski, N., Rajagopal, J., Teichmann, S. A., Misharin, A. V. and Nawijn, M. C. (2019). "The Human Lung Cell Atlas: A High-Resolution Reference Map of the Human Lung in Health and Disease." Am J Respir Cell Mol Biol 61: 31-41.

Schultz, C. J., Torres, E., Londos, C. and Torday, J. S. (2002). "Role of adipocyte differentiation-related protein in surfactant phospholipid synthesis by type II cells." Am J Physiol Lung Cell Mol Physiol 283: L288-296.

Schwartz, D. A., Helmers, R. A., Galvin, J. R., Van Fossen, D. S., Frees, K. L., Dayton, C. S., Burmeister, L. F. and Hunninghake, G. W. (1994). "Determinants of survival in idiopathic pulmonary fibrosis." Am J Respir Crit Care Med 149: 450-454.

Selman, M., Lopez-Otin, C. and Pardo, A. (2016). "Age-driven developmental drift in the pathogenesis of idiopathic pulmonary fibrosis." Eur Respir J 48: 538-552.

Selman, M. and Pardo, A. (2002). "Idiopathic pulmonary fibrosis: an epithelial/fibroblastic cross-talk disorder." Respir Res 3: 3.

Selman, M. and Pardo, A. (2014). "Revealing the pathogenic and aging-related mechanisms of the enigmatic idiopathic pulmonary fibrosis. an integral model." Am J Respir Crit Care Med 189: 1161-1172.

Sheng, G., Chen, P., Wei, Y., Yue, H., Chu, J., Zhao, J., Wang, Y., Zhang, W. and Zhang, H. L. (2019). "Viral Infection Increases the Risk of Idiopathic Pulmonary Fibrosis: A Meta-Analysis." Chest.

Sikic, B. I. (1986). "Biochemical and cellular determinants of bleomycin cytotoxicity." Cancer Surv 5: 81-91.

Sirianni, F. E., Chu, F. S. and Walker, D. C. (2003). "Human alveolar wall fibroblasts directly link epithelial type 2 cells to capillary endothelium." Am J Respir Crit Care Med 168: 1532-1537.

Sleijfer, S. (2001). "Bleomycin-induced pneumonitis." Chest 120: 617-624.

Somogyi, V., Chaudhuri, N., Torrisi, S. E., Kahn, N., Muller, V. and Kreuter, M. (2019). "The therapy of idiopathic pulmonary fibrosis: what is next?" Eur Respir Rev 28.

Sountoulidis, A., Lontos, A., Nguyen, H. P., Firsova, A. B., Fysikopoulos, A., Qian, X., Seeger, W., Sundström, E., Nilsson, M. and Samakovlis, C. (2020). "SCRINSHOT, a spatial method for single-cell resolution mapping of cell states in tissue sections." bioRxiv: 2020.2002.2007.938571.

Stripp, B. R. (2008). "Hierarchical organization of lung progenitor cells: is there an adult lung tissue stem cell?" Proc Am Thorac Soc 5: 695-698.

Strunz, M., Simon, L. M., Ansari, M., Mattner, L. F., Angelidis, I., Mayr, C. H., Kathiriya, J., Yee, M., Ogar, P., Sengupta, A., Kukhtevich, I., Schneider, R., Zhao, Z., Neumann, J. H. L., Behr, J., Voss, C., Stöger, T., Lehmann, M., Königshoff, M., Burgstaller, G., O'Reilly, M., Chapman, H. A., Theis, F. J. and Schiller, H. B. (2019). "Longitudinal single cell transcriptomics reveals Krt8+ alveolar epithelial progenitors in lung regeneration." bioRxiv: 705244.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., 3rd, Hao, Y., Stoeckius, M., Smibert, P. and Satija, R. (2019). "Comprehensive Integration of Single-Cell Data." Cell 177: 1888-1902 e1821.

Stuart, T. and Satija, R. (2019). "Integrative single-cell analysis." Nat Rev Genet 20: 257-272.

Sucre, J. M. S. and Hagood, J. (2020). "Single cell analysis of human lung development: knowing what mesenchymal cells are and what they may be." Eur Respir J 55.

Thannickal, V. J., Flaherty, K. R., Martinez, F. J. and Lynch, J. P., 3rd (2004). "Idiopathic pulmonary fibrosis: emerging concepts on pharmacotherapy." Expert Opin Pharmacother 5: 1671-1686.

Thomeer, M. J., Costabe, U., Rizzato, G., Poletti, V. and Demedts, M. (2001). "Comparison of registries of interstitial lung diseases in three European countries." Eur Respir J Suppl 32: 114s-118s.

Tomashefski J.F., F. C. F. (2008). Anatomy and Histology of the Lung. Dail and Hammar's Pulmonary Pathology. Springer, New York, NY

- Torday, J. S. and Rehan, V. K. (2016). "On the evolution of the pulmonary alveolar lipofibroblast." Exp Cell Res 340: 215-219.
- Travaglini, K. J., Nabhan, A. N., Penland, L., Sinha, R., Gillich, A., Sit, R. V., Chang, S., Conley, S. D., Mori, Y., Seita, J., Berry, G. J., Shrager, J. B., Metzger, R. J., Kuo, C. S., Neff, N., Weissman, I. L., Quake, S. R. and Krasnow, M. A. (2019). "A molecular cell atlas of the human lung from single cell RNA sequencing." bioRxiv: 742320.
- Travis, W. D., Costabel, U., Hansell, D. M., King, T. E., Jr., Lynch, D. A., Nicholson, A. G., Ryerson, C. J., Ryu, J. H., Selman, M., Wells, A. U., Behr, J., Bouros, D., Brown, K. K., Colby, T. V., Collard, H. R., Cordeiro, C. R., Cottin, V., Crestani, B., Drent, M., Dudden, R. F., Egan, J., Flaherty, K., Hogaboam, C., Inoue, Y., Johkoh, T., Kim, D. S., Kitaichi, M., Loyd, J., Martinez, F. J., Myers, J., Protzko, S., Raghu, G., Richeldi, L., Sverzellati, N., Swigris, J. and Valeyre, D. (2013). "An official American Thoracic Society/European Respiratory Society statement: Update of the international multidisciplinary classification of the idiopathic interstitial pneumonias." Am J Respir Crit Care Med 188: 733-748.
- Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M. and Cox, J. (2016). "The Perseus computational platform for comprehensive analysis of (prote)omics data." Nat Methods 13: 731-740.
- Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigartyo, C. A., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P. H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J. and Ponten, F. (2015). "Proteomics. Tissue-based map of the human proteome." Science 347: 1260419.
- Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. and Luscombe, N. M. (2009). "A census of human transcription factors: function, expression and evolution." Nat Rev Genet 10: 252-263.
- Vieira Braga, F. A., Kar, G., Berg, M., Carpaij, O. A., Polanski, K., Simon, L. M., Brouwer, S., Gomes, T., Hesse, L., Jiang, J., Fasouli, E. S., Efremova, M., Vento-Tormo, R., Talavera-Lopez, C., Jonker, M. R., Affleck, K., Palit, S., Strzelecka, P. M., Firth, H. V., Mahbubani, K. T., Cvejic, A., Meyer, K. B., Saeb-Parsy, K., Luinge, M., Brandsma, C. A., Timens, W., Angelidis, I., Strunz, M., Koppelman, G. H., van Oosterhout, A. J., Schiller, H. B., Theis, F. J., van den Berge, M., Nawijn, M. C. and Teichmann, S. A. (2019). "A cellular census of human lungs identifies novel cell states in health and in asthma." Nat Med 25: 1153-1163.
- Volckaert, T. and De Langhe, S. (2014). "Lung epithelial stem cells and their niches: Fgf10 takes center stage." Fibrogenesis Tissue Repair 7: 8.
- Walsh, S. L. F., Lederer, D. J., Ryerson, C. J., Kolb, M., Maher, T. M., Nusser, R., Poletti, V., Richeldi, L., Vancheri, C., Wilsher, M. L., Antoniou, K. M., Behr, J., Bendstrup, E., Brown, K. K., Corte, T. J., Cottin, V., Crestani, B., Flaherty, K. R., Glaspole, I. N., Grutters, J., Inoue, Y., Kondoh, Y., Kreuter, M., Johannson, K. A., Ley, B., Martinez, F. J., Molina-Molina, M., Morais, A., Nunes, H., Raghu, G., Selman, M., Spagnolo, P., Taniguchi, H., Tomassetti, S., Valeyre, D., Wijsenbeek, M., Wuyts, W. A. and Wells, A. U. (2019). "Diagnostic Likelihood Thresholds That Define a Working Diagnosis of Idiopathic Pulmonary Fibrosis." Am J Respir Crit Care Med 200: 1146-1153.
- Wang, R., Wang, G., Ricard, M. J., Ferris, B., Strulovici-Barel, Y., Salit, J., Hackett, N. R., Gudas, L. J. and Crystal, R. G. (2010). "Smoking-induced upregulation of AKR1B10 expression in the airway epithelium of healthy individuals." Chest 138: 1402-1410.
- Wells, A. U. and Denton, C. P. (2014). "Interstitial lung disease in connective tissue disease—mechanisms and management." Nat Rev Rheumatol 10: 728-739.
- Whitsett, J. A. (2018). "Airway Epithelial Differentiation and Mucociliary Clearance." Ann Am Thorac Soc 15: S143-S148.
- Wolf, F. A., Angerer, P. and Theis, F. J. (2018). "SCANPY: large-scale single-cell gene expression data analysis." Genome Biol 19: 15.
- Wuyts, W. A., Dooms, C. and Verleden, G. M. (2013). "The clinical utility of bronchoalveolar lavage cellular analysis in interstitial lung disease." Am J Respir Crit Care Med 187: 777.

Wynn, T. A. (2011). "Integrating mechanisms of pulmonary fibrosis." J Exp Med 208: 1339-1350.

Wynn, T. A. and Vannella, K. M. (2016). "Macrophages in Tissue Repair, Regeneration, and Fibrosis." Immunity 44: 450-462.

Xie, T., Wang, Y., Deng, N., Huang, G., Taghavifar, F., Geng, Y., Liu, N., Kulur, V., Yao, C., Chen, P., Liu, Z., Stripp, B., Tang, J., Liang, J., Noble, P. W. and Jiang, D. (2018). "Single-Cell Deconvolution of Fibroblast Heterogeneity in Mouse Pulmonary Fibrosis." Cell Rep 22: 3625-3640.

Young, M. D. and Behjati, S. (2020). "SoupX removes ambient RNA contamination from droplet based single-cell RNA sequencing data." bioRxiv: 303727.

Zepp, J. A., Zacharias, W. J., Frank, D. B., Cavanaugh, C. A., Zhou, S., Morley, M. P. and Morrisey, E. E. (2017). "Distinct Mesenchymal Lineages and Niches Promote Epithelial Self-Renewal and Myofibrogenesis in the Lung." Cell 170: 1134-1148 e1110.

Zhang, Y., Alexander, P. B. and Wang, X.-F. (2017). "TGF- β Family Signaling in the Control of Cell Proliferation and Survival." Cold Spring Harbor perspectives in biology 9: a022145.

Publications

In chronologic order:

- Schiller, H. B., **Mayr, C. H.**, Leuschner, G., Strunz, M., Staab-Weijnitz, C., Preisendorfer, S., Eckes, B., Moinzadeh, P., Krieg, T., Schwartz, D. A., Hatz, R. A., Behr, J., Mann, M. and Eickelberg, O. (2017). "Deep Proteome Profiling Reveals Common Prevalence of MZB1-Positive Plasma B Cells in Human Lung and Skin Fibrosis." *American Journal of Respiratory and Critical Care Medicine* 196: 1298-1310.
- 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., Nagendran, M., Desai, T., Eickelberg, O., Mann, M., Theis, F. J. and Schiller, H. B. (2019). "An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics." *Nature Communications* 10: 963.
- Kammerl, I. E., Caniard, A., Merl-Pham, J., Ben-Nissan, G., **Mayr, C. H.**, Mossina, A., Geerlof, A., Eickelberg, O., Hauck, S. M., Sharon, M. and Meiners, S. (2019). "Dissecting the molecular effects of cigarette smoke on proteasome function." *Journal of Proteomics* 193: 1-9.
- Strunz, M., Simon, L. M., Ansari, M., Mattner, L. F., Angelidis, I., **Mayr, C. H.**, Kathiriya, J., Yee, M., Ogar, P., Sengupta, A., Kukhtevich, I., Schneider, R., Zhao, Z., Neumann, J. H. L., Behr, J., Voss, C., Stöger, T., Lehmann, M., Königshoff, M., Burgstaller, G., O'Reilly, M., Chapman, H. A., Theis, F. J. and Schiller, H. B. (2019). "Longitudinal single cell transcriptomics reveals Krt8+ alveolar epithelial progenitors in lung regeneration." *bioRxiv*: 705244. *Manuscript submitted, preprint available online* .
- **Mayr, C. H.**, Simon, L. M., Leuschner, G., Ansari, M., Geyer, P. E., Angelidis, I., Strunz, M., Singh, P., Kneidinger, N., Reichenberger, F., Silbernagel, E., Boehm, S., Adler, H., Hilgendorff, A., Lindner, M., Prasse, A., Behr, J., Mann, M., Eickelberg, O., Theis, F. J. and Schiller, H. B. (2020). "Integrated single cell analysis of human lung fibrosis resolves cellular origins of predictive protein signatures in body fluids." *medRxiv*: 2020.2001.2021.20018358. *Manuscript submitted, preprint available online*. #

- Jiang D., Christ S., Correa-Gallegos D., Ramesh P., Gopal S.K., Wannemacher J., Yu Q., Mayr C.H., Lupperger V., Mück-Häusl M., Rajendran V., Wan L., Liu J., Mirastschijski U., Volz T., Marr C., Schiller H.B., Rinkevich Y. (2020). "Swarms of specialized fibroblasts drive scarring through N-cadherin". *Manuscript submitted*.
- Meul T., Berschneider K., Schmitt S., Mayr C.H., Mattner L., Schiller H.B., Yazgili A.S., Wang X., Lukas C., Prehn C., Adamski J., Graf E., Schwarzmayr T., Perocchi F., Kukat A., Trifunovic A., Kremer L., Prokisch H., Popper B., von Toerne C., Hauck S.M., Zischka H., Meiners S. (2020). "Mitochondrial regulation of proteasome function". *Manuscript submitted and under review*.
- Petzold, T., Thienel, M., Dannenberg, L., Mourikis, P., Helten, C., Ayhan, A., M'Pembele, R., Achilles, A., Trojovky, K., Konsek, D., Zhang, Z., Regenauer, R., Pircher, J., Ehrlich, A., Lusebrink, E., Nicolai, L., Stocker, T. J., Brandl, R., Roschenthaler, F., Strecker, J., Saleh, I., Spannagl, M., **Mayr, C. H.**, Schiller, H. B., Jung, C., Gerdes, N., Hoffmann, T., Levkau, B., Hohfeld, T., Zeus, T., Schulz, C., Kelm, M. and Polzin, A. (2020). "Rivaroxaban Reduces Arterial Thrombosis by Inhibition of FXa-Driven Platelet Activation via Protease Activated Receptor-1." *Circulation Research* 126: 486-500.
- Leuschner G.*, **Mayr C.H.***, Ansari M., Angelidis I., Adler H., Hilgendorff H., Prasse H., Behr J., Mann M., Schiller H.B. (2020). "Mass spectrometry driven unbiased identification of interstitial lung disease associated autoantigens reveals similar break of immune tolerance in IPF and CTD-ILD". *Manuscript in preparation*.
- Mayr C.H., Ansari M., Simon L.M., Ogar P., Strunz M., Angelidis I., Lontos A., Samakovlis C., Morty R., Ahlbrecht K., Theis F.J., Schiller HB. (2019). "Single cell profiling reveals diversity of mouse and human lung mesenchymal cell types and their distinct activation states in fibrogenesis". *Manuscript in preparation*.

this publication resulted from this thesis.

Acknowledgments

First of all I want to thank and express my gratitude to my PhD supervisor and mentor Dr. Herbert Schiller, who provided full scientific support throughout the time at the CPC. I especially would like to thank you for trusting in my abilities and for the chances you gave me to develop them further. Thank you for having always an open door in case of sudden questions and for the long discussions and ideas that arose from these meetings. I am very grateful for all the fascinating projects I could be part of, especially for the fancy and next generation methods we established in the lab. I greatly appreciate the numerous opportunities you gave me to present my data at conferences and learn more about science at the lab exchanges in Utrecht and Stockholm.

I would like to acknowledge the Comprehensive Pneumology Center and the Helmholtz Zentrum München for providing me an outstanding environment to pursue my PhD. I also thank the members of my thesis committee Prof. Dr. Jürgen Behr and Prof. Dr. Silke Meiners as well as Prof. Dr. Richard Scheltema, for fruitful discussions and very helpful advices.

I am also very grateful for the training and support I received from the CPC Research School „Lung Biology and Disease“. In particular I want to thank Dr. Doreen Franke, Dr. Claudia Staab-Weijnitz and Dr. Darcy Wagner for putting together a great and successful program.

It has been my good fortune to work with the best colleagues one can ever imagine. The team spirit in the schiller lab was always outstanding and, in any situation, provided helping hands, fruitful discussions, great laughter, dancing and singing, practical and mental support and overall the best atmosphere possible to work in.

Greetings of day go to the “Coffee Team”, the “Schiller Boys”, the “Mouse Team”, the “Milky Way Gang” and the “Daiquiri Team”! Thank you Max, Tom, Ilias, Laura, Gabi, Vanessa and Doreen for the amazing times we have spent together. It is peers, friends and colleagues like you that made the PhD a memorable time in my life, I will always cherish. All the best to you!!!

Last but not least, I am extremely grateful for the support I received from my parents and my sister from the start of my studies, throughout the PhD and beyond, in every aspect of my life.

Finally, I would like to thank Sabine for being the most amazing part of life in my work-life-balance over all those years.

