HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

HELMHOLTZ MUNICH

CPC Comprehensive Pneumology Center

Graduate Centre Weihenstephan
TUM School of Life Sciences
Technical University of Munich

Single-cell transcriptomics driven exploration of pathogenesis and impaired regeneration in chronic lung diseases

Meshal Ansari

October 2021

TUM School of Life Sciences
Technische Universität München

TECHNISCHE UNIVERSITÄT MÜNCHEN

Single-cell transcriptomics driven exploration of pathogenesis and impaired regeneration in chronic lung diseases

Meshal Ansari

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung des akademischen Grades einer

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:

Prof. Dr. Martin Hrabě De Angelis

Prüfer der Dissertation:

1. Prof. Dr. Dr. Fabian J. Theis
2. Priv.-Doz. Dr. Claudia Staab-Weijnitz

Die Dissertation wurde am 15.10.2021 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 20.04.2022 angenommen.

Abstract

Cells are the building blocks of life, understanding their functions, cell fate decisions and interplay with other cells is crucial to pin-point mechanisms that might get dysregulated and promote disease development. This information is only accessible at the single-cell resolution, therefore it is not surprising that single-cell RNA sequencing studies have been on the rise in recent years. This technology empowers the profiling of millions of cells by now and allows to reveal cellular heterogeneity, novel cell types, signalling networks and provides a mean to study altered patterns in pathological conditions.

Throughout the presented work, three major diseases of the lung are studied via single cell RNA-sequencing technologies: interstitial lung disease ILD, chronic obstructive pulmonary disorder COPD, and the recently emerged coronavirus disease COVID-19. The obtained transcriptomic profiles can be leveraged to chart disease progression. However, as tissue samples are most often obtained from patients with end stage disease, the mechanisms that initiate disease development would be impossible to observe. For many years, the laboratory mouse (Mus Musculus) has remained the quintessential research animal of choice for studying human pathology. Accordingly, to increase the information recovery as much as possible, transcriptomic analyses were performed on cells from a variety of origins, ranging from lung organoid cultures, mouse models up to in vivo patient tissue.

I first explored the course of early lung development. The dense temporal resolution enabled a characterization of gene expression changes along the differentiation trajectory. Thereafter, scRNA-seq data obtained from mouse models was utilized to study the response of acute and chronic lung injury. As a result, the dynamics of mouse lung regeneration could be chartered for the first time at single-cell resolution. This led to the discovery of a transient cell state that precedes the regeneration of depleted cell populations. The hypotheses were extended to human patients suffering from lung fibrosis and further experimental validation. The comparison to the human setting pointed towards defective terminal differentiation of lung stem cells during regeneration and increased intercellular communication between pathological cell states, ultimately promoting pathogenesis. In a subsequent part the focus was shifted towards COPD, particularly to how the accumulation of infiltrating immune cells in response to harmful particle exposure leads to tissue damage. Pre-existing experimental data suggested the prevention of such tissue damage via inhibition of certain signalling pathways. An appropriate mouse model enabled the dissection of these mechanisms at the cellular level. The final part revolves around COVID-19 and its effect on the adaptive immune cells. Here I validated the correspondence of the derived antigen-induced T cell signatures from in vitro experiments in human patient data.

Overall, affected cell populations, cell state shifts and alterations in cellular communication following some sort of perturbations were of interest in all of these data sets. In each sub-project I focused on the most intriguing aspects, which will be laid out in greater detail throughout this thesis.

Zusammenfassung

Zellen bilden den Grundbaustein allen Lebens, ihre Funktionen, Entscheidungen während der Differenzierung und Zusammenspiel mit weiteren Zellen zu verstehen ist ein wesentlicher Punkt um die Mechanismen aufzufinden, die vermutlich fehlreguliert sind und letztendlich zu einer Krankheitsentwicklung führen. Diese Information ist nur auf der Einzelzell- ebene zugänglich, daher ist es nicht verwunderlich dass Einzellzell-RNA Sequenzierungs- methoden in den letzten Jahren an hoher Beliebheit gewannen. Diese Technologie befähigt uns heutzutage Millionen von Zellen zu messen und die zelluläre Heterogeneität, neue Zelltypen und Signalnetzwerke aufzuzeigen, und stellt damit Mittel zur Verfügung, um die veränderten Muster in pathologischen Bedingungen zu erforschen.

Im Verlauf dieser Arbeit werden drei bedeutende Lungenkrankheiten mittels Einzellzell-RNA Sequenzierung untersucht: interstitielle Lungenerkrankung ILD, chronisch obstruktive Lungenerkrankung COPD und die vor kurzem aufgetretene Coronavirus-Krankheit COVID-19. Die erhaltenen Transkriptome können genutzt werden, um den Krankheitsverlauf abzubilden. Jedoch sind die Gewebsproben meist aus Patienten, die sich im Endstadium der Krankheit befinden, daher wäre es nicht möglich die Mechanismen einzufangen, welche eine Krankheitsentwicklung anstoßen. Seit vielen Jahren wird die Labormaus (Mus Musculus) als essentielles Versuchstier eingesetzt, um die menschliche Pathologie zu erforschen. Dementsprechend wurden transkriptomische Analysen basierend auf Zellen aus unterschiedlichsten Quellen durchgeführt. Diese reichen von Organoidkulturen über Mausmodelle bis hin zu in vivo Gewebsproben aus Patienten.

Zunächst habe ich den Verlauf der frühen Lungenentwicklung erkundet. Die dichte Zeitauflösung ermöglichte eine Charakterisierung der Genexpressionsveränderungen entlang der Differenzierung. Anschließend wurden scRNA-seq Daten basierend auf einem Mausmodell genutzt, um die Reaktionsmuster in Bezug auf akute sowie chronische Lungenverletzungen zu studieren. In einem Teil konnte die Dynamik der Lungenregeneration in Mäusen erstmals in Einzelzell-Auflösung beschrieben werden. Dies führte zu dem Auffinden eines transienten Zellzustandes, der einer Regeneration der dezimierten Zellpopulationen vorausgeht. Die aufgestellten Hypothesen wurden auf Humanpatienten, die an Lungenfibrose erkrankt waren, ausgeweitet. Ein Vergleich zu der menschlichen Situation wies auf einen fehlerhaften Differenzierungsprozess der Lungenstammzellen während der Regeneration hin. Zudem wurde verstärkte interzelluläre Kommunikation zwischen pathologischen Zellzuständen aufgefunden, welche höchstwahrscheinlich zu einer Krankheitsentstehung beiträgt. In einem folgenden Teil wurde der Fokus auf COPD gelegt, insbesondere wie die Anhäufung von einwandernden Immunzellen als Reaktion auf schädliche Substanzen zu Gewebsschäden führt. Vorherige experimentelle Daten, basierend auf Inhibition gewisser Signalwege, wiesen auf eine Vorbeugung solcher Schäden hin. Durch angemessene Mausmodelle konnten die zugrundeliegenden Mechanismen auf zellulärer Ebene weiter aufgeschüsselt werden.

Der finale Teil behandelt die Auswirkung von COVID-19 auf die Zellen des adaptiven Immunsystems. Hierbei konnte ich abwägen, wie gut sich die Antigen-induzierten Signaturen in T Zellen, die aus in vitro Experimenten abgeleitet wurden, mit Transkriptomdaten aus Humanpatienten decken.

Contents

Now, here, you see, it takes all the running you can do, to keep in the same place.

The Red Queen Hypothesis
Lewis Caroll - Through the Looking Glass

Chapter 1

Introduction

During the Red Queen's race that takes place in Lewis Carroll's *Through the Looking-Glass*, Alice could not move forward despite constantly running. While panting, she points out "Well, in our country, you'd generally get to somewhere else - if you run very fast for a long time". The queen responds that in her country, things worked a little differently. If one would like to get somewhere, they would have to run at least twice as fast as that. This statement - although fitting for painting the picture of the bizarre world of wonderland that Alice ended up in - reflects an essential aspect of evolutionary biology. Leigh Van Valen was among those that recognized the applicability to their work and termed it the *Red Queen's Hypothesis*.[1] A species is under constant pressure to adapt and overcome challenges better and faster than competing species in order to survive and ensure the transfer of its genes to future generations.

However, this quote can not only be applied in an evolutionary sense but captures the trend that science has been undergoing in general. It has become increasingly easier to share methods, reagents, results and also failures across the globe. This sets the perfect environment for the fast development of new hypotheses and their adaptation based on rapid feedback from renowned peers. In recent years, technological advances have led to an explosion of data, especially in the biological field. These span from a few molecule sequences, that can be stored in text files, up to atlas-level measurements for a large number of samples, for which whole data bases have to be set up. Accompanying the rapid increase of data, an equally growing number of methods is necessary to make sense of the accumulation of information. Many methods described in this thesis have been developed in the last few years, are constantly being improved, and likely will be obsolete and replaced by superior approaches in the near future. Nevertheless, the consequences in this case are not as drastic as the extinction of an entire species. On the contrary, the speed at which biological research is progressing will not only enable deciphering mechanisms that have been inaccessible thus far, but will also have positive effects on society at large. As one of science's main objectives is to improve people's health and general well-being, the contemporary surge should and is being used for knowledge acquirement and propagation.

We will have to keep running to keep the pace.

1.1 Key technologies for the characterization of cells

Individual cells are the smallest unit of an organism and generally referred to as the building blocks of life. The term "cell" has been coined by the English physicist Robert Hooke (1635 – 1702). He described his observations with microscopes in his book Micrographia (1665) and was the first to visualize a microorganism, the microfungus Mucor.[3]
Interestingly, when Hooke was looking at a thin slice of cork under his microscope, the small box-like units he saw reminded him of "cells", the rooms in which Christian monks used to live and meditate in (from Latin *cella*, meaning *a small room*). Although Hooke used the word in a different context, the term *cell* still remains up to this date.[4,5]

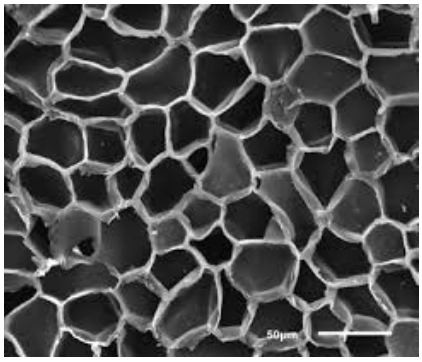


Figure 1.1: SEM of cork (tissue from bark of the cork oak tree), showing its cell structure[2].

The human body is estimated to be made up of 4 x 10^{13} cells,[6] each of them managing their gene expression program and maintaining intercellular interactions in order to ensure normal cellular function. Already small dysfunctions are sufficient to disrupt these processes, leading to flawed phenotypes or, in the worst case, manifest as human disease. As shown by cancer, the malfunctions generated in one single cell result in uncontrolled cell growth, potentially spreading throughout larger parts of the tissue and ultimately leading to the demise of an entire organism. In order to capture such dysfunctions, it remains a major challenge to understand biological processes and their interplay at the lowest level, e.g. pin-point the cells which play a role in certain diseases' progression. The process of classifying cells based on their properties has been ongoing since more than 150 years. Traditionally, it was mainly based on the cell's morphology, such as location, size or shape.[7]
Already in the 1930s Ernst Ruska, Reinhold Rüdenberg and many others sought to find a way to visualize micro-organisms. Rüdenberg for instance aimed to overcome limitation of contemporary low resolution capture of the submicroscopic polio virus, and to provide strategies for the diagnosis and treatment of cases of polio. With that goal, electron microscopy came into existence. By using a beam of accelerated electrons these microscopes provide a thousand-fold higher resolution than light microscopes. Not only viruses could be visualized, but also the structure of small biological specimen, such as micro-organisms, cells or crystals, enabling a distinction from each other based on their structural features.[8]

Only a decade later in 1941, Albert Coons first described immunostainings, a method that harnesses the principle of antibodies binding to specific antigens.[9] By coupling a fluorescent dye to antibodies specific to pneumococcal antigens and incubating with infected animal tissues, Coons showed that the antibodies remained specific for their target bacterial strains. Unlike uninfected tissues, antibodies in infected tissue agglutinated the bacteria and fluoresced brightly after excitation by ultraviolet light. The technique spread quickly across research disciplines, showing that it is possible to localize certain proteins of interest in biological tissues.[10]

In 1965, Fulwyler described a device that was capable of isolating biological cells of interest.[11] This was achieved by suspending the cells in a conducting medium and capturing them in droplets. A stream of droplets is generated, in which each droplet is electrically charged according to their measured volume. By entering an electric field, those droplets that contain cells that are to be separated from the mixture are deflected accordingly into a collection tube.
A few years later, Bonner et al. developed Fluorescence-Activated Cell Sorting (FACS) based on Fulwyler's separation technique.[12] Instead of an electric charge, the sorting parameter was cell fluorescence. The medium flows ideally one cell at a time through a laser beam, where a light signal is produced whenever a fluorescent cell crosses the laser beam. It became possible to tag cells by different fluorescent markers (e.g. by reactions with fluorescently labelled antibodies) which recognize specific surface markers. Additional physio-chemical features such as cell size, forward and sideward scatter can be measured and enabled sorting of the cells into distinct populations based on these characteristics.[13]

Given the technological progress at every step, an increasingly detailed description of each cell type was facilitated. Nonetheless, some limitations remained, for instance the techniques up until then could only be employed on easily dissociable tissues. By restricting fluorescent labelling to surface markers, a rather targeted approach was enforced. This relies heavily on prior knowledge of known cell type markers and does not allow the systematic discovery of yet undescribed cell types. Furthermore, it became clear that morphologically indistinguishable cells can still exhibit drastically different molecular functions. Characterizing cells based on superficial features such as morphology or the presence of a small number of markers alone does not reflect the whole picture.[14,15] Instead, over years efforts were made that tried to dig into the underlying roots that cause these outward features rather than merely looking at their manifestation.

The phenotype, i.e. the observable traits of cells, depends on the genetic information stored in their DNA. The transfer of that information happens via gene expression, a process that is carried out by all known organisms and shown in Fig. 1.2b. The information of the DNA is replicated into messenger RNA (mRNA), which is transported into the cytoplasm and gets translated to an amino acid sequence. After additional post-translational modifications that further influence the structure and properties, the final products in form of proteins are generated. Proteins are known as the molecular machinery of life as they maintain cellular functions.
Gene expression is regulated by the turnover of gene transcripts, indicated by the amount of copies of each expressed gene within in a cell. Measuring the expression levels reveals a snapshot of the set of RNA molecules present in a biological system, which ultimately dictates what cells are doing or what they are capable of at a given point in time. Scientist started discerning that altered patterns of gene expression reflect many cellular decisions regarding survival, growth and differentiation.[16] It became alluring to further determine factors that regulate the expression, be they nutritional, hormonal, environmental or due to certain pathological conditions that arise in aberrant environments, e.g. disease. This led to a rising interest in quantifying transcription and paved the way for a variety of methods. There were three techniques prevalently applied for measuring mRNA, which are still in common usage up to this day.

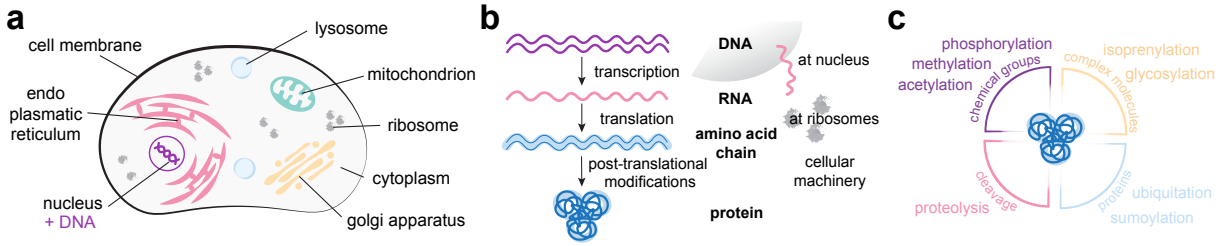


Figure 1.2: Conversion of genetic information to phenotype. **a** Structure and components of an eukaryotic cell. **b** Flow of genetic information in a biological system. DNA is transcribed in the cell's nucleus (*transcription*), resulting mRNA travels to ribosomes in cytoplasm where proteins are synthesized using mRNA as template (*translation*). **c** Additional post-translational modifications change protein structure and diversify their function. Partly adapted from Alberts (2008)[17] and Wang et al. (2014).[18]

The first, and most extensively used at that time, was Northern blotting, described by Alwine et al. in 1977.[19] Its underlying principle is the use of electrophoresis to separate RNA samples by size and detect it on a membrane. In an initial step total RNA is extracted from a tissue and immobilized on a solid membrane. A hybridization probe with a base sequence complementary to the sequence of the target mRNA is prepared. Hybridization signals can then be detected and allow a quantification of transcript levels of the gene of interest.[20] Although Northern blotting remains one of the key techniques in molecular biology, it does require relatively large amounts of input RNA.
The second method, the reverse transcription polymerase chain reaction (RT-PCR), provides considerable increase in sensitivity over Northern blotting and remains one of the most sensitive methods to detect (even low-abundance) gene transcripts. It is an in vitro method that combines amplification of defined mRNA sequences and their detection into one single step. After reverse transcribing mRNA into complementary DNA (cDNA), these sequences form the templates for exponential amplification using polymerase chain reaction (PCR). This reaction incorporates a heat stable DNA polymerase in order to amplify double-stranded DNA in an exponential manner. PCR consists of three main steps: heat denaturation, primer annealing and primer elongation.[21] Fluorophores are then added to monitor the amplification rates. These are excited and generate a light signal when bound to double-stranded DNA (PCR product), thus the PCR product concentration correlates with fluorescence intensity. After each amplification cycle, in which the amount of DNA sequences doubles, the fluorescent signal is measured and allows an accurate quantification of the mRNA levels.[22]

In contrast to the first two methods, fluorescence in situ hybridization (FISH) enables localization of transcripts to specific cells within a tissue. It circumvents the requirement of artificially amplifying the mRNA and averaging the signal across many cells from different locations. FISH is an adaptation of the approach presented by Joseph Gall and Mary Lou Pardue in 1969.[23] Molecular hybridization uses labelled DNA or RNA sequences as natural probes, as a way to localize their complementary counterpart in a biological sample. When the probe meets its target sequence, detectable hybridization occurs. Non-specifically paired sequences will be cleared away in the washing steps and only strongly bound strands remain hybridized. In the original study this was demonstrated using radioactive copies of a ribosomal DNA sequence and detecting biding events in tissue

autoradiographically.[24] Following this work, Rudkin and Stollar replaced the radioactive labels in hybridization probes by fluorescent labels. Briefly, after denaturation of both target and probe sequences, hydrogen bonds can form between complementary sequences in the subsequent hybridization step. The sites of hybridization can then be detected directly in their tissue context using a fluorescent microscope.[25]

Facilitated by technological advances, the simultaneous analysis of expression levels of thousands of genes was rapidly made possible. The DNA microarray hybridization system was introduced by Patrick O. Brown's group in 1995[26] and is conceptually similar to FISH. Thousands of microscopic DNA fragments are fixated on solid surfaces for hybridization analysis of corresponding, fluorescently labelled genes. As the signals are localized to spots on the array, it is easier to detect and quantify the hybridization signals. DNA microarrays provide a platform for screening the gene expression patterns in virtually any biologic sample in a more parallelized format.[27]
Furthermore, by using a two-colour fluorescence labelling approach or separate microarrays, it is possible to assess whether genes are differentially expressed. One could compare hybridization events of the same target genes, once in diseased and once in healthy tissue. As the expression response of thousands of genes could be monitored in a coordinated fashion by such arrays, their popularity in biomedical research did rise drastically: The amount of studies that employed this method increased exponentially in the 1990s. For instance, microarrays were widely employed for the cancer research, providing a tool to determine molecular differences between healthy and malignant cells but also to identify genes that are implicated in tumor formation or progression.[28] Despite their broad application, the methods were not perfect yet and faced certain limitations, such as reliance upon existing knowledge about genome sequences and distortion of signals due to high background levels of cross-hybridization.

In a more recent surge, the field of genomic research has been revolutionized by the advent of DNA sequencing. High-throughput sequencing technologies enabled massive-parallelization and further reduction in labour and cost to sequence millions of reads in a single run, making these methods accessible for whole-transcriptome analyses. Remarkably, by this method termed *RNA sequencing*, certain limitations encountered by microarrays could be overcome. It has been shown that the detection is not restricted to transcript-specific probes any longer, instead novel transcripts could be recovered. Furthermore, rare or low-abundance transcripts could be detected by simply increasing the sequencing coverage, allowing to find lowly expressed genes that might still show differential pattern in pathological conditions.[29,30]
Therefore, it is not surprising that over the years the measurement of transcriptome-wide gene expression has switched from microarrays to sequencing.[31] With these breakthroughs in genomics and transcriptomics, together with innumerable other key technologies, the complex regulatory networks present in biological samples could be assessed in a more systematic fashion - providing the means to identify new genes and pathways that play essential roles in disease progression, or to discover targets for novel therapies.

As RNA sequencing constitutes the foundation of this thesis, the next section is dedicated to outline the progression and key aspects of the method.

1.2 Increasing the resolution of transcriptomics

Scientists around the globe came together in an effort to completely map all genes present in the human genome, physically as well as functionally. The research program was based on two key insights that arose in recent decades:[32]

1. a global views on genomes could greatly improve biomedical research, by allowing researchers to tackle problems in a comprehensive and unbiased fashion
2. the creation of such global views would require a communal effort in infrastructure building, unlike anything previously attempted in biomedical research

This international effort laid the groundwork for accelerating biomedical research. It provided a reference for genomic sequences and enabled to pin-point anchoring points (e.g. proteins, genes, single nucleid polymorphism SNPs) that were significantly altered in disease. One level of mining the genome sequence for biological information would be by exploring the transcriptome at a given time, as a way to reflect the active genes that control the processes occurring in the cells.

As described in the previous section, RNA sequencing was a major breakthrough at the start of this century, replacing the widely used microarrays for measuring gene expression levels of a large number of input cells. A basic overview of the main steps in a standard RNA-seq protocol is given in Fig. 1.3. The first step is the extraction and purification of RNA from a sample of interest followed by an enrichment of target RNAs. Most commonly used is poly(A) capture, to select for polyadenylated RNAs. Next, the molecules are fragmented to appropriate size and reverse transcribed into double-stranded cDNA. The strands are flanked with adapters at their 3' and 5' ends and amplified by PCR, using the adapter sequences as primers. After sequencing the library (see section 2.3), the transcripts can be mapped against a reference genome to enable gene annotation and further downstream data analysis.
The capability to assess the transcriptome for a biological sample showed potential to be integrated into a variety of research areas. By progressive cost reductions and parallelization, the high throughput approach of RNA-seq became accessible to a larger number of scientists. Especially its main application, which is quantification of gene regulation, enabled uncovering the molecular processes and components in an organism.

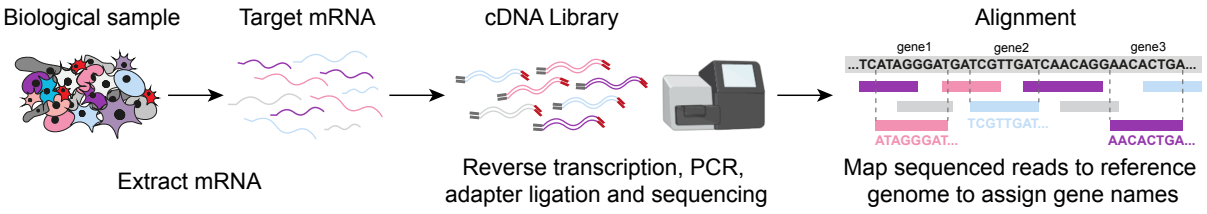


Figure 1.3: Scheme of the experimental steps in a RNAseq protocol. The cDNA library is built by isolating RNA targets from biological samples, reverse transcribing, amplifying and tagging mRNAs with sequencing adapters. After sequencing, the reads are mapped against a reference genome. Adapted from Wang et al. (2019).[31]

Of particular interest was the identification of gene sets that show alternating patterns across conditions. The settings to which the analyses could be applied to seemed endless,

comparing the gene expression between different tissues, genotypes, stimulations, time points, disease states etc. In clinical research for instance RNA-seq revealed the gene response to certain stimulations, providing a platform to assess the efficacy of therapeutic drugs.[31] Nonetheless, the quantification could be delineated even further, as was shown by recent developments in the transcriptomic field.

The gene expression measurement methods described above have retrospectively been titled *bulk* RNA-seq, as these were based on bulk tissue samples. Such samples contain ensemble measurements from a mixture of input cells. The major assumption is that the averaged assays catch the dominant biological mechanism of the individual cells belonging to the population. On the one hand, this assumption is necessary to create an expressive data set to allow further analyses. On the other hand, there are potentially cells whose behaviours are far from the mean.[33] Together with the rise of RNA-seq in the early 00's, opinions arose on how averaged data sets do not necessarily reflect the character of any single cell in them, obscuring cell-to-cell variations in a given population.[34] The averaged approach is appropriate enough for the study of homogeneous samples, however medical research would benefit greatly from an increased resolution.

Heterogeneity is a property of cell populations and implies the presence of cell-to-cell variability with respect to certain cellular traits, e.g. their morphology or their level of proteins. Cellular heterogeneity had often been conflated with noise before studies of gene expression or protein levels in mammalian tissues began to reveal that there is cell-to-cell variance, even within similar cell types in the same tissue.[35] Not only is this variance non-neglectable, stochastic fluctuations in levels of gene expression can play a critical biological role and trigger cell fate decisions. Such differences in expression explain how apparently identical cells in a homogeneous environment can exhibit strikingly diverse behaviour.[36] Defining cell types and states opens the door to assessing compositions in both homeostatic and perturbed conditions. Fig. 1.4 gives an illustrative example why in certain cases the analysis at single-cell level can be beneficial. Given a small farm with different species of animals with clearly distinguishable features, there are two routes for conducting studies on this population. The first one is in a *bulk* fashion (Fig. 1.4a), meaning one considers and averages the features across all species. The resulting signal would rather reflect a mythical creature, containing parts of every animal, incapable of resolving the heterogeneity in the population. In the second one each animal is considered separately prior to measurement of their features, which would yield a clearer picture of the population of interest (Fig. 1.4b). Relating back to transcriptomics, one measures the profiles of biological tissues (*farm*), for which the composition of cell types (*species*) is in many cases not known in advance.

Purifying cells into distinct populations on the basis of well-established markers is one way of resolving the transcriptomic signal. Nonetheless, many of such markers still obscure some subpopulations, e.g. $CD14^+$ monocytes do all share CD14 expression, but can be further divided into subtypes with distinct characteristics.[39] This approach will not work for yet undescribed cell states, and poses a fundamental limitation as the discovery of such condition-specific states is one of the main interests of tissue analysis.
The cells transition between multiple states particularly during developmental processes. Cell plasticity is commonly illustrated using the analogy of Waddington's landscape of

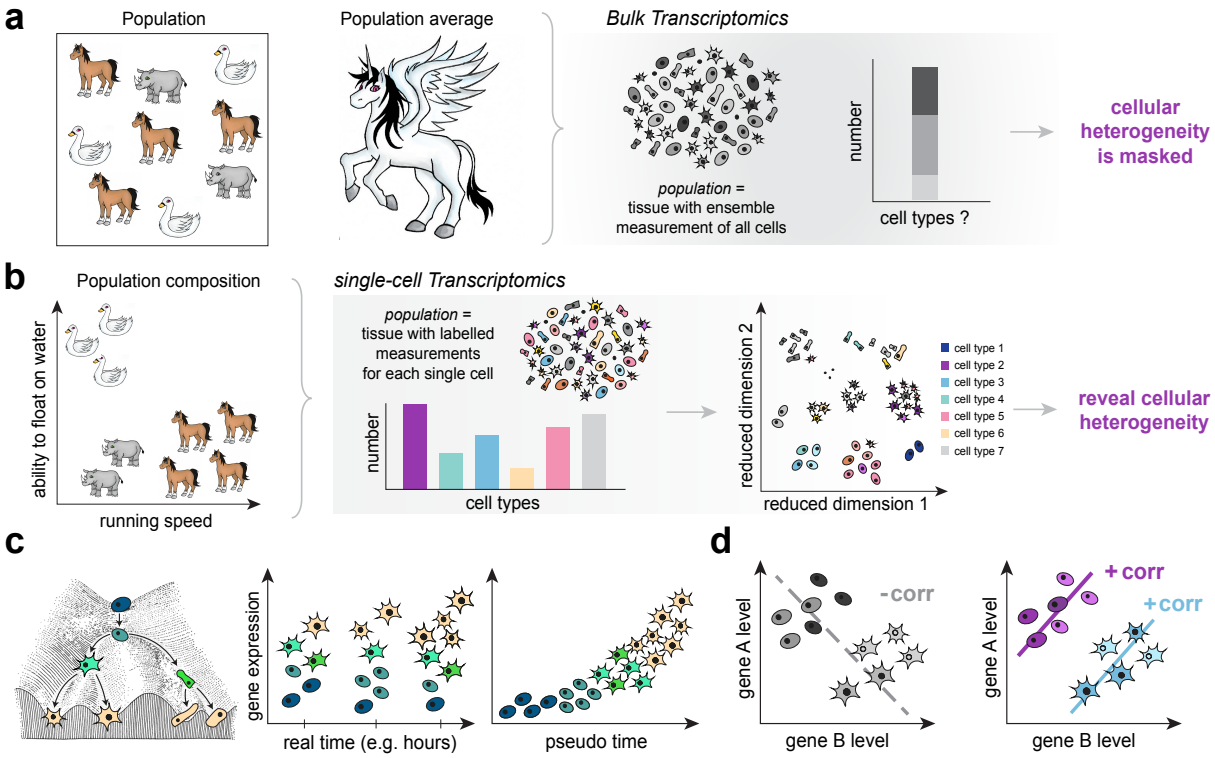


Figure 1.4: Comparison of well-established bulk transcriptomic approach with single-cell transcriptomic analysis. **a** Visual example demonstrating how information loss occurs after averaging over population. **b** Prior separation of animal species (figuratively *cell populations*) leads to a clearer representation, enabling more granular analyses. **c** Cell differentiation along Waddington's landscape. Uncovering temporal dynamics is hardly possible by averaging signal of all cells at a time point due to unsynchronized processes. Instead order cells by their underlying time. **d** Misleading conclusions arise if no proper separation of cell types is performed, e.g. correlation of gene A and B in multiple subpopulations. Partly adapted from 10x[37] and Trapnell (2015).[38]

possible states.[40] Waddington described an uneven landscape comprised of valleys, corresponding to likely cell states, and hills, reflecting unlikely cell states. A cell's developmental path is similar to a marble rolling down from the peak of this landscape, but instead of following a single trajectory, there are multiple potential paths influenced by the landscape's layout (Fig. 1.4c).

Cells often behave in an asynchronous manner along their gradual differentiation, making it difficult to put clear boundaries. Instead it might be more correct to catalogue cell types in a first layer and stratify their different flavors in a second layer, in form of so-called cell states which are induced in response to external factors. This characterization poses a challenge in itself and is hardly possible to tackle with bulk measurements, as these destroy the boundaries of individual cells. Single-cell technology on the other hand can not only determine a cell's location on this landscape, i.e. its stage in development, but furthermore assess the molecular mechanisms that shape the landscape itself.[38]
Apart from the decreased accuracy, averaging across many cells can lead to incorrect conclusions altogether. In a hypothetical example, if a population is a mixture of two subpopulations with different expression levels of a gene A and gene B, it will not be

possible to capture correlation trends without properly separating them by type beforehand (Fig. 1.4c). Additionally, one major pitfall is that it is unclear whether the changed expression pattern of a given gene arises due to up-/down-regulation in each cell, or can simply be attributed to the frequency change of a specific cell type population. As another example, a down-regulation could be indicated by measuring a gene's levels after certain drug treatment, whereas in reality the cell type expressing the gene of interest has been partially depleted instead.

Ensemble measurements are neither as labour intensive nor complex due to much lower levels of technical noise. Data acquired on single-cell level scales up to several thousands of cells and is intuitively harder to process. Technical errors and unwanted variability can be introduced to the data set, which could be misinterpreted as important biological heterogeneity where there is none.[33] Although certain computational approaches can be adapted from existing ones, more sophisticated quality control steps and normalization methods have to be developed tailored to the larger, more artefact-heavy data sets.
Nonetheless, putting in extra effort and establishing analyses at a much higher resolution allows to tackle biological problems that have been inaccessible this far. There are fundamental limitations in ensemble measurements which make them unsuited for the study of heterogeneous systems. Although this field is still relatively new, the power of single-cell RNA-seq (scRNA-seq) was highlighted in a vast number of studies already (Fig. 1.5). In the last decade considerable growth in the field of scRNA-seq methods has been made, both on the experimental as well as the computational side. This technology has been used to assess the transcriptomic profiles of a variety of organs, including regions of the brain,[41,42,43] the retina,[44,45,46] the pancreas[47,48,49] and for early embryonic development in model organisms.[50,51] Further potential applications range from systematic discovery of new cell types, to determining cell-fate decision points and key players during differentiation, up to resolving main pathways and mechanisms that are involved in pathogenesis at an unprecedented level of resolution.

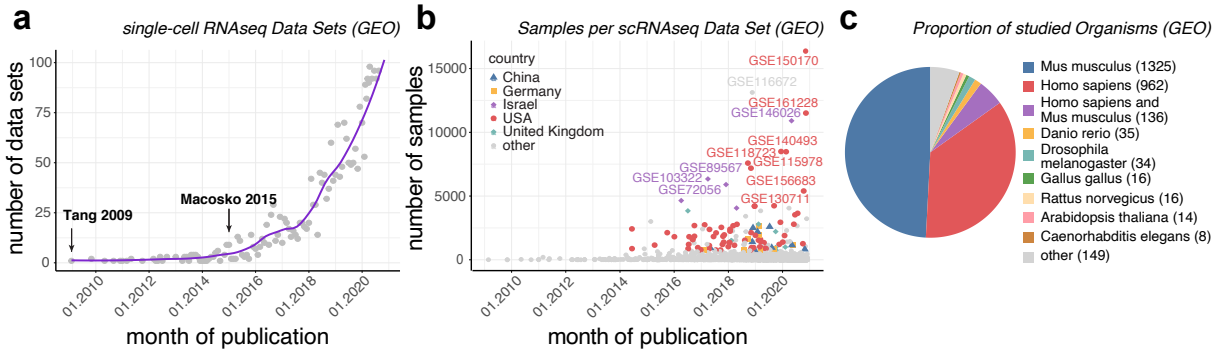


Figure 1.5: Growth of single-cell RNAseq data sets in Gene Expression Omnibus GEO from 2010 to 2020. **a** Number of scRNA-seq data sets and **b** number of samples per data set per month. First data set (Tang et al.)[52] and publication introducing droplet-based capture methods (Macosko et al.)[44] are highlighted. Upper 2 Percentile are coloured by country listed in GEO. **c** Pie chart showing proportions of organisms studied.

1.3 Recent advances in single-cell RNA-sequencing

In recent years numerous assays have been proposed and optimized to work at the level of single cells, getting more refined, less labour-intensive and incorporating high-throughput approaches. The next section will give an overview of the key technologies that have enabled such a drastic rise in popularity and applicability of scRNA-seq. The first attempt at sequencing an entire transcriptome at the level of one single cell dates back to 1992, when Eberwine and colleagues tried to understand how changes at the molecular level lead to different functional properties, even in a small number of morphologically similar cells.[53] The authors successfully dissociated cells from a defined region of a rat hippocampus, reverse transcribed the recovered mRNA of a selected few genes and finally increased the product using linear amplification in a first, and PCR amplification in a second step.

Several years later in 2009, a transcriptome-wide investigation of solely one manually picked cell was performed for the first time. Building on the concept of DNA microarrays, Tang et al. managed to incorporate the benefits of next-generation sequencing into single-cell transcriptome analysis.[52] In many conditions, especially early embryonic development, only a small number of cells is available in the biological system, whereas profiling techniques require microgram amounts of total RNA. Tang and colleagues could showcase that their modified workflow of whole-transcriptome amplification and gene expression analysis could reliably recover the expression of manually picked blastomeres from early mouse development and detect 75% more genes compared to the standard microarray. With these primary approaches, the perks of single-cell resolution were demonstrated and made generally accessible by commercial assays and high-throughput methods.

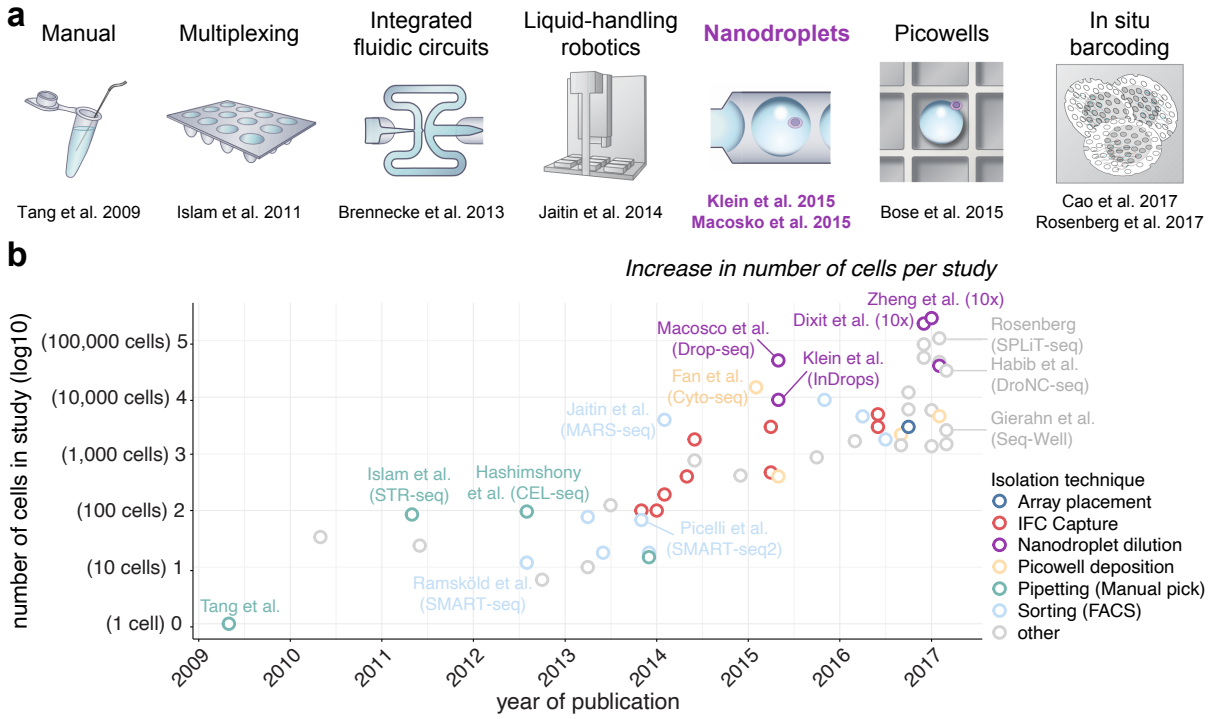


Figure 1.6: Exponential growth of single cells that can be sampled. **a** Key technologies that enabled the recent surge in reported cell numbers per study. **b** Number of cells reported in studies with key publications highlighted. Taken and adapted from Svensson et al. (2018).[54]

It is not surprising that the workflow has been continuously optimised in the following years and fueled a consistent increase in the number of cells that could be studied at once.[55,54] With the introduction of sample multiplexing, i.e. assaying many cells in parallel, microfluidic technologies and improved cell capture methods, the data set sizes jumped to several thousands of cells. These numbers continue to rise even further (Fig. 1.6). By now, a plethora of methods have been introduced for scRNA-seq analyses, each of which has different strengths and weaknesses. The choice of protocol should be guided by the specific research question, whether one would prefer profiling larger number of cells at lower transcript coverage, e.g. to profile cell type composition in general, or rather quantify condition effects at higher sensitivity in only selected populations of cells.[56]

The initial steps of scRNA-seq protocols are reminiscent of the steps in RNA-seq and shared across different platforms. This encompasses cell isolation, lysis, reverse transcription, amplification and finally sequencing as described previously. The main difference is the additional physical separation into smaller reaction chambers, or alternatively the labelling of gene transcripts such that they can be traced back to their cell of origin. The points in which scRNA-seq methods typically diverge from each other can be boiled down to how the following key challenges are addressed: automatic single-cell isolation, transcript amplification and sequencing method.[54] A broad overview of most commonly applied technologies is given in Fig. 1.7. For cell isolation, the protocols can be roughly divided into three categories. The first would be microfluidic devices that trap cells inside hydrogel droplets, such as inDrop,[57] Drop-seq[44] and the commercialized 10x Genomics Chromium.[45] The major advantage of the droplet strategy is the rapid compartmentalisation into single-cell reaction chambers at a frequency of several thousand droplets in seconds, enabling massive parallelization and reaction throughput at relatively low cost.[58,59] These techniques are heavily featured for data generation throughout this work, and are therefore described in more detail in section 2.1.
Although differing in details of sequence design and material, the droplets in all three assays are designed to simultaneously capture beads and cells. The on-bead primers contain a PCR handle, a cell barcode, an unique molecular identifier of 4-8 bp sequence (UMI) for amplification bias correction, and a poly-T tail.[59] However, it has been established that only up to 10% of transcripts will be retrieved and reverse transcribed,[60] making the detection rates of droplet-based methods relatively low compared to other capture methods. Still, many studies have shown that despite this low sequencing depth a robust identification of cell identity is possible. Their cost-efficient capture and library production of large number of cells make these methods attractive for certain scenarios, e.g. if the overall composition, or identification of rare subpopulations are the main interest.

The second isolation category is the physical separation of cells into 96-well plates. Coupled to cell-sorters, such as FACS or by using micro pipetting, cells are placed in individual wells containing lysis buffer. Up to 500 cells can be analyzed in a single experiment, each capturing 5 000 to 10 000 genes. However, due to this set-up, reverse transcription has to be carried out in each well separately, leading to numerous pipetting steps and potential technical noise and batch effects.[56] The increased sensitivity comes at a higher cost, power simulations illustrated how Drop-seq in particular is more cost-efficient for very large number of cells, whereas the plate-based methods SCRB-seq,[61] MARS-seq[62] and Smart-seq2[63] allowed superior characterization for fewer cells.[64]

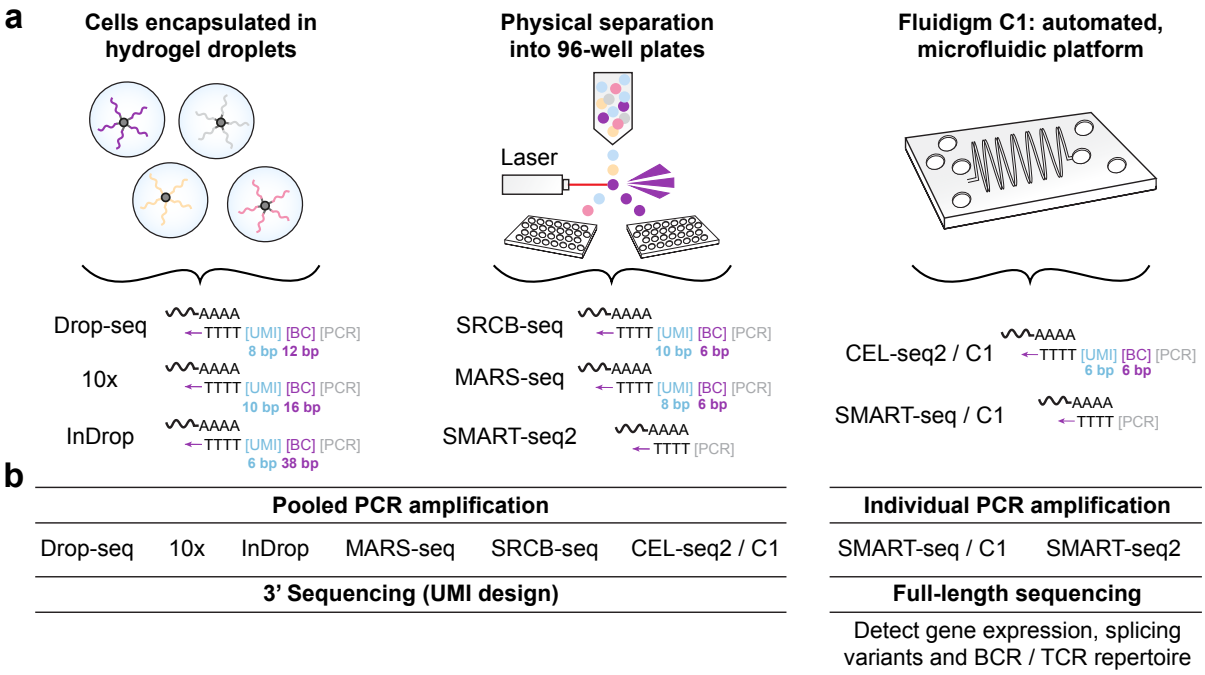


Figure 1.7: Key differences of prominent scRNA-seq technologies. **a** Cell capture can be performed using nano-liter droplets, separation into 96-well plate-based or by the automated C1 chips. **b** Droplet- and some plate-based methods enable pooled transcript amplification using cellular barcodes, allowing for only 3' sequencing. Other plate-based methods and Fluidigm C1 enable full-length sequencing due to PCR amplification per individual well. Adapted from Ziegenhain et al. (2017)[64] and Papalexi et al. (2018).[56]

Nonetheless, any cell that can be sorted can be analyzed, allowing greater generalizability. Coupled with their high sensitivity, plate-based methods are especially fitting for small-scale experiments with a focus on specifically sorted cells. The last isolation strategy is a commercial tool for passive cell capture released by Fluidigm. The C1 system is a microfluidic chip designed to load and separate cells into very small reaction chambers in an automated manner, combining the RNA extraction and library preparation step into one and decreasing manual labour tremendously. Its major drawback however is the amount of cells required as minimum input, which is more than 10 000. It is also advised to use rather homogeneous cells for the analysis, as cells will reach differently distant locations on the chip based on their size and introduce a location bias.[55] Again, the higher detection rates are bound to a higher cost, making this technology better suited if a selected population of cells is of interest.

The UMI design in the droplet-based methods, and some of the plate-based methods, makes it possible to distinguish between original transcripts and amplification duplicates after PCR. While this reduces the number of PCR reactions to one per experiment/plate, it also restricts the subsequent sequencing to 3' as the cell barcodes and PCR handles are added to only that end of the transcript. Full-length sequencing covering the full gene body is possible for certain plate-based methods and Fluidigm C1, as the amplification step is carried out on individual wells separately. This allows to recover not only gene expression, but also splicing variants and B/T cell receptor repertoire diversity. However, the number of PCR amplification reactions is equal to the number of cells that are being profiled, rendering this approach unsuited for studies encompassing large cell numbers.[56]

1.4 Anatomy and cell types of the lung

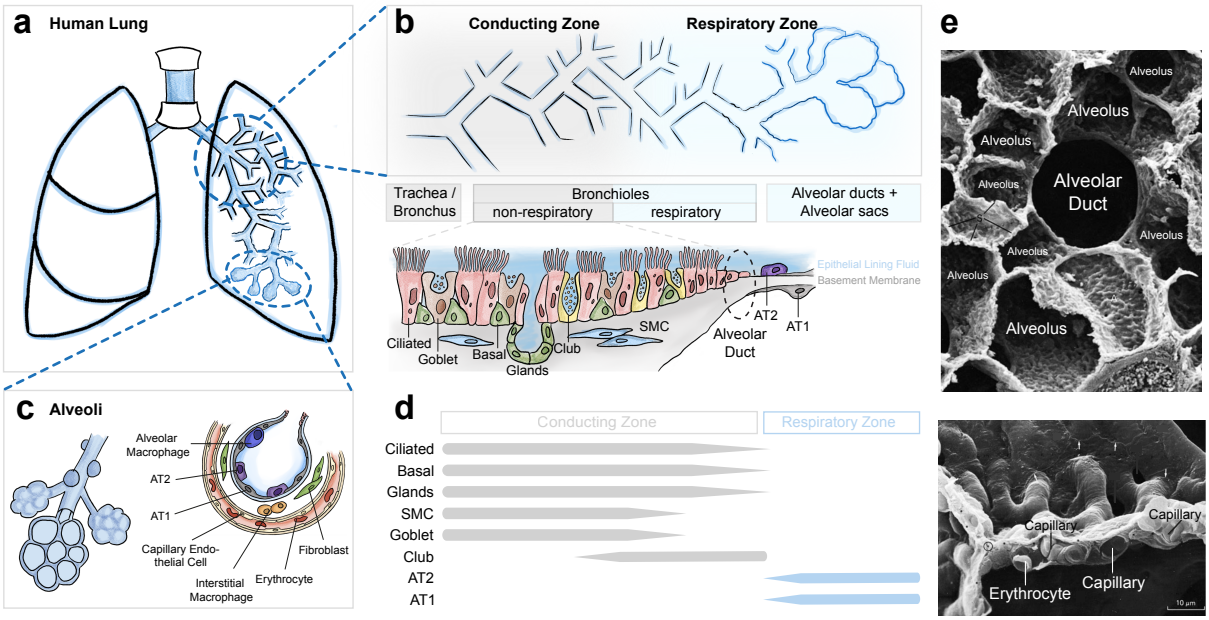


Figure 1.8: Overview of human lung architecture and main cell types. **a** Simplified scheme of the human lung, consisting of three anatomically distinct regions: Trachea, bronchioles, and alveoli. **b** Representation of airway branching, starting from trachea up to alveolar sacs present at the endpoints and cell types that constitute the epithelial layer. **c** Cell type composition in alveolar unit, interstitial space and capillary cells. **d** Different epithelial cell populations are present based on location in airway. **e** Scanning electron micrograph showing cross sections of the alveolar/capillary unit and lung parenchyma, taken and adapted from *Medical Physiology: A Systems Approach*.[65]

The human lung is a complex, paired organ found on either side of the heart with the main purpose of performing gas exchange. It is connected to the most proximal airway, the trachea, by its right and left bronchi, bordered by the concave-shaped diaphragm and enclosed by a membrane referred to as the pleura (mesothelium in mice). The lung can be divided into two smaller units, the so-called lobes. In humans, the right lung consists of three lobes (superior, middle, and inferior), whereas the left lung consists of two lobes (superior and inferior).[66]

Each of the two main bronchi divides further into progressively narrower airways (bronchioles). The main bronchi are reinforced with hyline cartilage, whereas the bronchioles are surrounded by smooth muscles for structural support. Starting with the trachea, the air passes through many branchings of airways before reaching the most distal part, the alveoli. Here the airways and blood vessels unite in form of thin alveolar epithelial cells and the fine capillary network that covers them.[67] Alveoli are the units facilitating gas exchange, as alveolar and capillary walls meet and enable gases to move across. These start appearing in the 17^{th} to 19^{th} generation of airways at the respiratory bronchioles. The first 16 generations of airways are referred to as the conducting zone, that proceeds via the transitional zone to the respiratory zone which starts at the 20^{th} generation[65] (Fig 1.8b).

By now 45 histological cell types of the adult human lung have been discovered, each with distinct location, structure and function. In a recent molecular profiling study this number increased further, defining 58 lung populations by their gene expression profiles and anatomical regions.[68]
Roughly, these lung cell types can be assembled into the main compartments epithelium, endothelium, stroma, pleura/mesothelium, neurons and immune cells. The cellular composition and structural organization varies depending on the location along the proximal-distal axis, adding another spatial layer to the cell type classification.

Under normal conditions, turnover in the lung is relatively slow compared to other organs, but sufficient to maintain homeostasis. In adults, stem cells are a pivotal point of proliferative hierarchies. They can, either directly or through a sequence of divisions, give rise to specialized cells with unique functional properties. In the context of injury response however, proliferation is enhanced in order to rapidly restore normal proportions of cell types and accommodate repair.[69] The heterogeneity in cellular composition, and therefore different stem cell populations, provides diverse response strategies depending on the spatial context at the site of the perturbation. With improved technologies, the potential cell hierarchies are becoming increasingly well-described.
The next section will provide a coarse overview of cell types in the lungs and their established differentiation hierarchies.

Epithelium

Epithelium refers to cells covering the surfaces of the body that are exposed to the outside world and lines the exterior of organs and therefore provide the first barrier of protection. In the case of the lung, the trachea, bronchi and bronchioles are lined by a pseudostratified epithelium, i.e. one that consists of a single layer or irregularly shaped cells, all connected to the basement membrane. Underlying the epithelium are blood vessels, smooth muscles/cartilage, stromal fibroblasts and nerves. The epithelial cells of the lung can be subdivided into airway and alveolar types.

Human and mice airways have a very similar structure, however the cellular composition and structure diverges in some points. For instance, in humans the small airway changes to a more uniformly shaped cuboidal epithelium only at the most distal cells at the bronchioalveolar duct junctions, whereas in mice much of the small airway is composed of a cuboidal epithelium.[70]
Due to their constant exposure, the airway epithelial cells have evolved to provide a certain level of host protection. Atop the epithelium lies a mucuous layer, which traps any incoming micro-organisms or particles and will sweep those upwards via the so-called mucouciliary escalator. Involved in this process are secretory cells, which continuously produce mucins and antimicrobial peptides, and ciliated cells whose cilia beat in unison to remove debris.[70] Interspersed into the airway epithelium are the $KRT5^{+}TP63^{+}$ basal cells, which act as progenitor cells to self-renew or replenish secretory and ciliated cells during homeostasis and repair.[71] Additionally, submucousal glands are cellular arrangements containing mucous-secreting cells as well as basal cells. In humans both submucousal glands and basal cells can be found throughout the conducting airways, but are confined to the trachea and primary bronchi in mice lungs.[70]

A subtype of the secretory cells are the goblet cells, located predominantly in the larger airways, and gradually replaced towards the smaller airways by another secretory type, the club cells.[72] Club cells are conventionally characterized by their expression of secretoglobins (e.g. SCGB1A1, SCGB3A1, SCGB3A2)[73], whereas goblet cells continuously secrete mucins (e.g. MUC5AC, MUC5B).[74] Alongside the basal cells, club cells act as progenitors that on the one hand can produce new ciliated cells,[75] and on the other hand were shown to de-differentiate into a basal cell phenotype, persisting over long term. The latter was validated by lineage-tracing of Scgba1a1$^+$ or Atpv1b1$^+$ labelled secretory cells in a mouse trachea model in which specifically basal cells were depleted.[76] Under normal physiological conditions there are few goblet cells, however these may proliferate excessively in response to acute or chronic injuries. Such goblet hyperplasia results in exaggerated mucous production/accumulation and eventually leads to airway obstruction. While the mucous layer typically exhibits protective functions, mucous hypersecretion is a cardinal feature of many severe respiratory conditions, including asthma, chronic obstructive pulmonary disease and cystic fibrosis.[77] There are also some rare cell types scattered throughout the airways, such as pulmonary neuroendocrine cells (NEC) or ionocytes.[78] NEC make up less than 1% of the epithelium, but can be present in groups (neuroendocrine bodies) more distal of the airways.[79]

Alveolar ducts connect the respiratory bronchioles to a cluster of alveoli, which are many small grape-like sacs with elastic walls that can stretch during air intake. The alveoli consist of three main cell types, two types of alveolar cells and alveolar macrophages, phagocytic cells that roam the alveolar space and remove debris and pathogens that have reached the alveoli. The alveolar structure is supported by extracellular matrix proteins secreted by alveolar fibroblasts, and is surrounded by capillary cells.[67]
Alveolar type 1 cells (AT1) are highly permeable to gases and extremely thin, elongated cells which cover 95% of the alveolar surface area while making up only 8% of total cells in the human lung.[80] Interspersed among these are alveolar type 2 cells (AT2), which are much smaller and cuboidal in shape. AT2 cells secrete pulmonary surfactant, a substance composed of phospholipids and proteins (SFTPA, SFTPB, SFTPC) that reduces the surface tension of the alveoli and prevents alveolar collapse during breathing.[80]

Apart from maintaining homeostatic turnover of AT1 cells, AT2 cells have been shown to possess proliferative and stem cell properties.[70,72] Several decades ago, it was demonstrated how their stem cell function gets triggered in response to injury, causing AT2 cells to self-renew and to replenish the AT1 population.[81] This could be confirmed by recent lineage tracing analyses in mice, in which AT2 cells proliferated and contributed to alveolar renewal and repair after bleomycin injury (a chemotherapy drug causing transient disruption of alveolar structure) and hyperoxia.[82,83] With the increased resolution that can be achieved nowadays, even intermediate differentiation state of AT2 cell en route to AT1 cells could be described in more detail in vivo repair models, revealing an enrichment of genes associated with cellular senescence, DNA-damage response signalling and TP53, TGFβ pathways.[84]
Whether there is a distinct subpopulation that participates in alveolar repair, or whether the full population can potentially differentiate further remains an open question. Stem cell function is fueled by Wnt signals emanating from the niche for stem cells - not only in the lung, but multiple organs.[85] Interestingly, a Wnt-responsive AT2 cell subset marked by

Axin2 expression demonstrated enhanced organoid formation and increased proliferation, whereas Wnt-inhibition shunts their differentiation towards AT1 cell lineage. Although it is not clear how well the results obtained from injury models correspond to alveologenesis during early lung development, it is hypothesized that dysfunctional repair mechanisms contribute significantly to disease pathogenesis. The presence and accumulation of abnormal epithelial cells in particular was demonstrated in human disease (more details in section 1.5). Interestingly, some disease-specific cell population show high similarity to the transient AT2 population after injury, alluding to a potential defective repair and persistence of these transient states.[84,86]
Further research particularly on the alveolar compartment is necessary to understand the mechanisms of injury response and repair and will inform new therapy strategies.

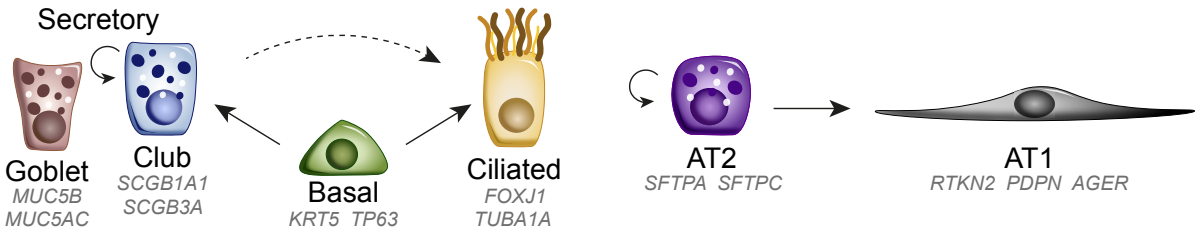


Figure 1.9: Lineage hierarchies in the lung epithelium. Basal cells act as stem cells in lung airway and can replenish secretory and ciliated cells. In the alveoli, AT2 cells adopt this role. Exemplary cell type markers are indicated. Adapted from Tata et al. (2017).[70]

Leukocytes

The majority of cell types that were captured in the presented whole lung specimen are of the leukocyte compartment. These are also known as white blood cells and are important components of the immune system. To perform their protective functions, or in response to chemical signals, leukocytes routinely leave the bloodstream to migrate to different locations in the body using the vascular system as a highway. They can cross physical barriers either by emigration, adhesive crawling or under extreme cellular deformation ("squeezing") through adjacent cells in blood vessel walls.[87] Leukocytes are derived from hematopoietic stem cells in the bone marrow and are classified broadly according to their structure into granulocytes, which contain abundant granules in their cytoplasm (neutrophils, eosinophils, basophils, mast cells) and agranulocytes, with far fewer granules (monocytes, macrophages, lymphocytes).[88]

Granulocytes are terminally differentiated cells. The most abundant leukocyte in humans are neutrophils. They are the first responders to the site of infection or invading pathogens, which they can eliminate via their granules containing antimicrobial agents. As effector cells of the innate immune system, neutrophils play a key role in the overall immune and inflammatory response.[89] Due to their pivotal role, dysfunctions in neutrophils can lead to the pathogenesis of a many human diseases, including, various lung diseases, autoimmune and inflammatory diseases and cancer.[90,91] Eosinophils on the other hand represent 2–4% of the total leukocyte count. The prevailing theory is that eosinophils participate in innate immune responses to parasites, in particular they secrete chemicals that destroy organisms that are too big for any white blood cell to phagocytize. The number of eosinophils in the blood and some tissues increases by 20-fold or more known during specific immune responses and in allergic diseases, including forms of asthma.[92,93]

The rarest leukocytes are Basophils, which comprise less than 1% of their total count. Because of their low abundance, it is hard to obtain these cells, thus their functions are less known. They generally intensify the inflammatory response and release histamines and heparins, a process which dilates blood vessels, counteracts blood clotting, and alleviates migration of recruited white blood cells.[94]

The two main types of agranulocytes can be distinguished by their cell lineage. They originate either from lymphoid stem cells (lymphocytes) or myeloid stem cells (mononuclear phagocytes) in the blood. Myeloid progenitors give rise to both dendritic cells (DC) and monocytes. DCs are antigen-presenting cells that trigger and regulate the adaptive immune response, while monocytes are described as cells that circulate the blood in order to scavenge dying cells or pathogens, and populate tissues as macrophages in the steady state.[95,96] Monocytes can be divided into three populations with distinct surface markers and functions: classical ($CD14^+CD16^-$, in mice: $Ly6C^{high}$ $CD43^{low}$), non-classical ($CD14^{low}CD16^+$, in mice $Ly6^{low}$ $CD43^{high}$), and intermediate ($CD14^+CD16^+$).[39] In steady-state conditions, non-classical monocytes patrol the resting vasculature and clear damaged endothelial cells, while classical monocytes circulate and survey tissues without differentiating.[97] In inflammatory milieu however, classical monocytes have been shown to differentiate into macrophages and monocyte-derived dendritic cells, thereby linking the innate defense to the adaptive immune responses. Monocytes display remarkable adaption to the challenged environment and ability to migrate to sites of need, actively shaping inflammation and its resolution in tissues.[98,99]

Likewise, macrophages show substantial heterogeneity in their phenotype and function as they occupy multiple tissue niches. They can have fixed locations or wander through tissues in order to respond to invading pathogens and support tissue homeostasis by removing dead cells and debris.
The major tissue-resident macrophage populations, such as liver kupffer cells, microglia in the brain and lung alveolar macrophages, are established prior to birth, derived from embryonic precursors that are either yolk sac macrophages or fetal liver monocytes.[100,101,102] Particularly in the lung, there are two classes of macrophages, merely separated by the thin alveolar wall: the more abundant alveolar macrophages (AM) residing within the lumen of the alveoli, and the interstitial macrophages (IM), arising from blood monocytes and residing within the lung parenchyma. Their location provokes different functions, apart from the phagocytosis of foreign particles. AMs catabolise the surfactant of the alveoli while IMs are assumed to be essential in tissue remodelling/maintenance as well as antigen presentation.[103,104] Recently, two distinct IM subtypes have been described, located adjacent to either nerve fibers ($Lyve^{low}$) or blood vessels ($Lyve^{high}$).[105]

These populations can maintain themselves during adulthood by self-renewal. However, macrophage number often expands dramatically in diseased tissues, mostly due to recruitment of circulating monocytes which then accumulate at the diseased sites.[97]
In response to substantial loss of embryonic-derived AM, monocytes are recruited to where the micro-environment shapes them into cells that closely resemble tissue-resident macrophages.[106] A direct link between the monocyte-derived AM and lung injury response has been established, as their depletion ameliorated disease severity, whereas similar depletion of tissue-resident AM did not have an effect on pathogenesis.

Interestingly, the monocyte-derived AM persist long-term in the lung even after injury resolution.[107]

The above described leukocytes mostly represent the innate arm of immunity, which is critical for protection and neutralization of pathogens. They recognize non-self cells, which can be cancer cells, pathogen-infected cells or other cells with abnormal surface proteins. The non-specific initial inflammatory response further informs and directs a more effective reaction. The adaptive arm of the immune response is mediated largely by lymphocytes, which arise from lymphoid stem cells in the bone marrow, and mature in lymphatic tissues. The three major types are natural killer cells NK, B cells and T cells.[108]

T cells migrate through the lymphoids in steady state and initiate the response. During maturation in the thymus, each T cell obtains an unique T cell receptor (TCR) with a variable chain, responsible for a T cell's specificity. Innate immune cells such as neutrophils, macrophages, DCs or NK cells are capable of "presenting" an antigenic peptide (epitope) derived from the pathogen via their major histocompatibility complex (MHC) class I or II, to which the epitope is then covalently bound. Upon interaction with the naive T cell, the antigen can be recognized by the TCR, T cell activation occurs and an immune response is triggered to defend against the infectious challenge. The subsequent expansion generates two different subtypes, either helper $CD4^+$ T cells, involved in cytokine and chemokine secretion to recruit new immune cells, or cytotoxic $CD8^+$ T cells, more streamlined to eliminate the infected host cell by the release of cytotoxins.[109]
B cells on the other hand mature in the bone-marrow and constitute the second component of the adaptive immune response. Less like the cell-mediated response of the T cells, B cells are primarily responsible for the antibody-mediated immunity, in which the produced antibodies bind and neutralize invading bacteria or viruses in response to antigen-presentation by other immune cells. After elimination of the pathogen, both T and B cells can differentiate to long-lived memory cells, which retain information on the encountered pathogen. Upon reinfection, the cells are able to mount a protective immune response tailored to the invading pathogen at a much faster pace (*acquired immunity*).[110]

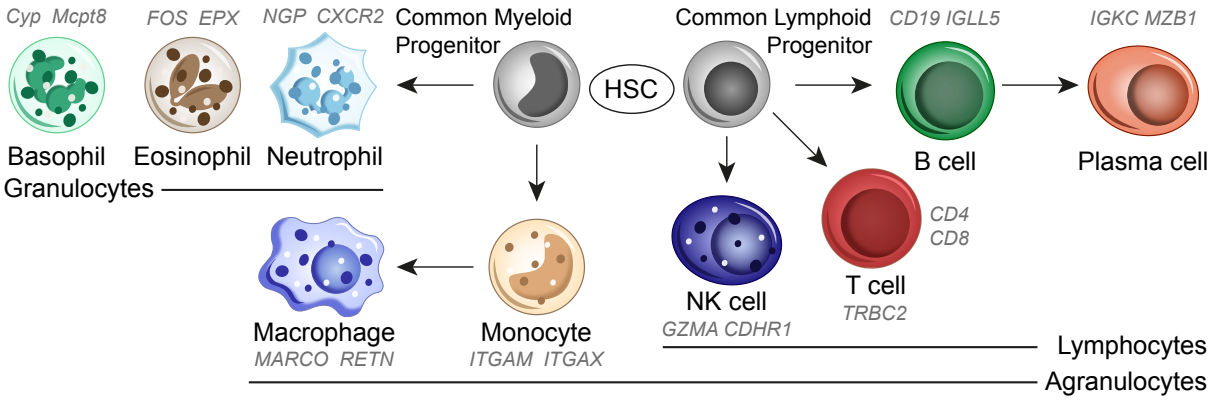


Figure 1.10: Hematopoietic stem cells HSC give rise to all blood cell types. Leukocytes originate either from myeloid (CMP) or lymphoid stem cells (CLP). CMP additionally generate erythrocytes, mast cells and megakaryocytes (not shown). Cell type markers are indicated. Adapted from *Hematology: Clinical Principles and Applications* (2020).[108]

Endothelium

Surrounding the alveoli at the distal parts of the lung lies a mesh of capillary vessels, providing the respiratory surface for the transfer of oxygen into the blood. The endothelium refers to structural cells that form the inner surface of these blood vessels. As this cell layer lies between the blood, airway and lung parenchyma, it acts as a selective barrier and enables leukocytes to migrate from the vessel to the tissue when necessary as well as normal blood flow.[108] Additional to the structural aspect, lung endothelial cells secrete cytokines, chemokines, interleukins, adhesion molecules, and growth factors to maintain tissue function.[111] In healthy lungs an intact endothelium prevents the aggregation and adhesion of platelets and leukocytes to vessel walls. Activated by changes in their cellular niche, e.g. by injury, endothelial cells can initiate a coagulation pathway, leading to the formation of a fibrin network and adhesion of platelets.[112]
Akin to the AT1 and AT2 cells that line the alveolar epithelium, the other side across the air-blood barrier has recently been described to consist of two interlaced populations: the $CAR4^{+}EDNRB^{+}$ aerocytes (aCap), mainly involved in leukocyte migration and in close contact with AT1 cells to enable gas exchange. The second type are the $FCN3^{+}EDN1^{+}$ general capillaries (gCap), proposed as a capillary stem cell in homeostasis and repair.[113]

Mesenchyme

The extracellular matrix (ECM) provides physical support to tissues, and actively influences cells' behavior in both health and disease. Alterations in composition, stiffness or injury initiate a reparative process to maintain tissue architecture and instruct cells to accumulate and replace damaged tissue.[114,115] The replacement occurs via secretion of ECM components by fibroblasts, which can produce the ECM's structural proteins (fibrous collagen, elastin) and adhesive proteins (laminin, fibronectin). They are involved in a variety of additional roles, such as ECM maintenance, wound healing, inflammation and tissue fibrosis, accompanied by the ability to produce and react to a broad array of cytokines and growth factors.[114]
The stromal cells in the lung show considerable heterogeneity already in healthy conditions, which has yet to be resolved. Each subtype has a distinct anatomical location and potential to respond to certain chemical signals that promote their activation.[116] It is assumed that the major collagen-producing cells in the lung are a subset of the fibroblasts, the myofibroblasts. These generate contractile forces to activate integrin-bound TGF-β and are characterized by their expression of the actin proteins ACTA2 and alpha-smooth muscle actin (α-SMA).[117] However, it is still poorly understood which cell population forms the source of these ECM-secreting cells in the context of disease. The most prominent model suggests that a sub-population of residual stromal cells are activated in response to inflammatory stimuli, start proliferating and secreting ECM. Other controversial models propose an external origin, either from peripheral blood (fibrocytes),[118] or even from alveolar epithelial cells, which lose their epithelial characteristics and gain mesenchymal properties, in a process called epithelial-mesenchymal transition (EMT).[119] Although, it should be noted that the literature on EMT is rather conflicted.[115,120]
As fibroblasts play a vital role in wound healing, their dysfunctions and uncontrolled matrix production can lead to a pathological scarring that is intrinsic to many fatal lung diseases (see section 1.5.2). Despite their key roles, the mechanisms and heterogeneity of these cells still have to be fully elucidated.

1.5 Pulmonary diseases

The lung is one of the organs most exposed to the external environment and therefore constantly subjected to harmful substances. These materials encompass a variety of agents, ranging from chemical substances (Nitrogen dioxide NO_2, ozone O_3, sulphur dioxide SO_2), particulate matter (tobacco smoke, exhaust fumes), biological components (e.g. allergens, derived from fungal spores and allergenic pollen) to bacteria or viruses in the atmosphere.[121]
Accordingly the lung requires the ability to protect itself from adverse effects caused by dangerous agents. There is a multitude of mechanisms in place to defend the organism. However, inhalation of such airway pollutants interferes with these mechanisms and can lead to airway injury, that can vary in both cause and effect. This further facilitates the development or exacerbation of pulmonary diseases.[122]

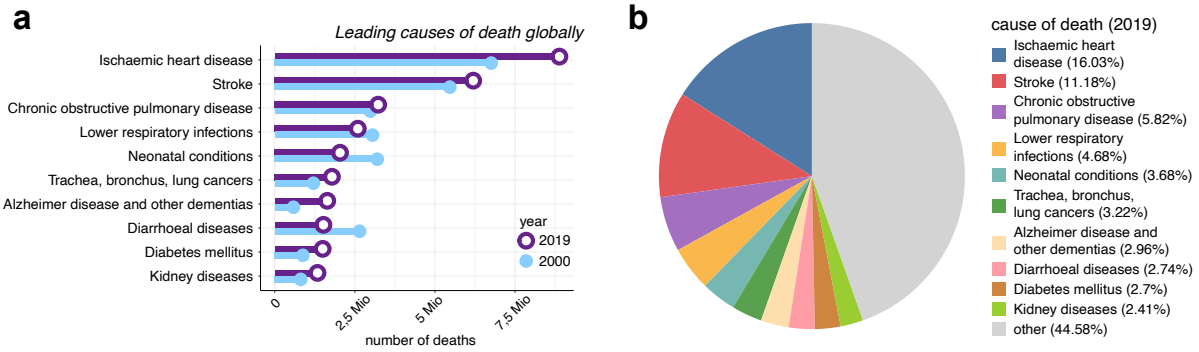


Figure 1.11: Leading causes of death worldwide. **a** Top 10 causes of death ranked by number of reported deaths in year 2000 and 2019. **b** Contribution of top 10 causes of death to global death numbers in percentage (year 2019). Raw data taken from the *World Health Organization*.[123]

It is not surprising that diseases affecting the lung are among the top global causes of death, in regard to the total number of lives lost. The World Health Organization (WHO) recently released a report *Global Health Estimates 2020*, in which they present health data and trends of 160 diseases and injuries annually from 2000 to 2019. The 10 leading causes of death can be roughly grouped into cardiovascular (ischaemic heart disease, stroke), respiratory (chronic obstructive pulmonary disease, lower respiratory infections) and neonatal conditions (Fig. 1.12a). In 2019, these accounted for 55% of the 55.4 million deaths worldwide, while chronic obstructive pulmonary disease (COPD) alone was responsible for 6% of total deaths and ranked third on this list (see Fig. 1.12b). The most deadly communicable diseases are lower respiratory infections. Although the number of deaths decreased, these still claimed at 2.6 million lives in 2019.[124] Lung cancers are among the most common and serious types of cancer. The major risk factor for lung cancer is smoking, which accounts for 75-80% of these deaths. This form of cancer is typically preventable, however it is usually diagnosed at an incurable stage.[125]

Respiratory conditions in general affect millions of people globally and make up a great portion of disease-related deaths. Even if the disease does not lead to the passing of an individual, the quality of their life will be heavily impaired. As is the case with many

health-related fields, respiratory research to improve treatment and possibly even the prevention of pathogenesis remains indispensable. This became gravely clear especially in 2020. The year shaped by the COVID-19 pandemic, caused by the novel coronavirus SARS-CoV-2, affected not only diseased patients but heavily restricted the lives of non-infected individuals as well. There was a pressing need to understand its pathogenesis, mobilizing large numbers of researchers worldwide in search of coping mechanisms and vaccines. As this outbreak occurred recently, it was not considered yet in the latest reports provided by WHO and is covered separately in section 1.5.3.

There are many additional types of respiratory diseases with varying symptoms and causes. For this thesis the focus will be on three common lung condition areas by the means of one specific example respectively. The first would be obstructive lung diseases characterized by increased resistance to airflow (e.g. COPD, Asthma). The second would be restrictive conditions, in which the expansion of the lung is confined (e.g. ILD). Many of these diseases, if severe enough, can lead to respiratory failure, which would be the last covered condition (e.g. ARDS).[126]

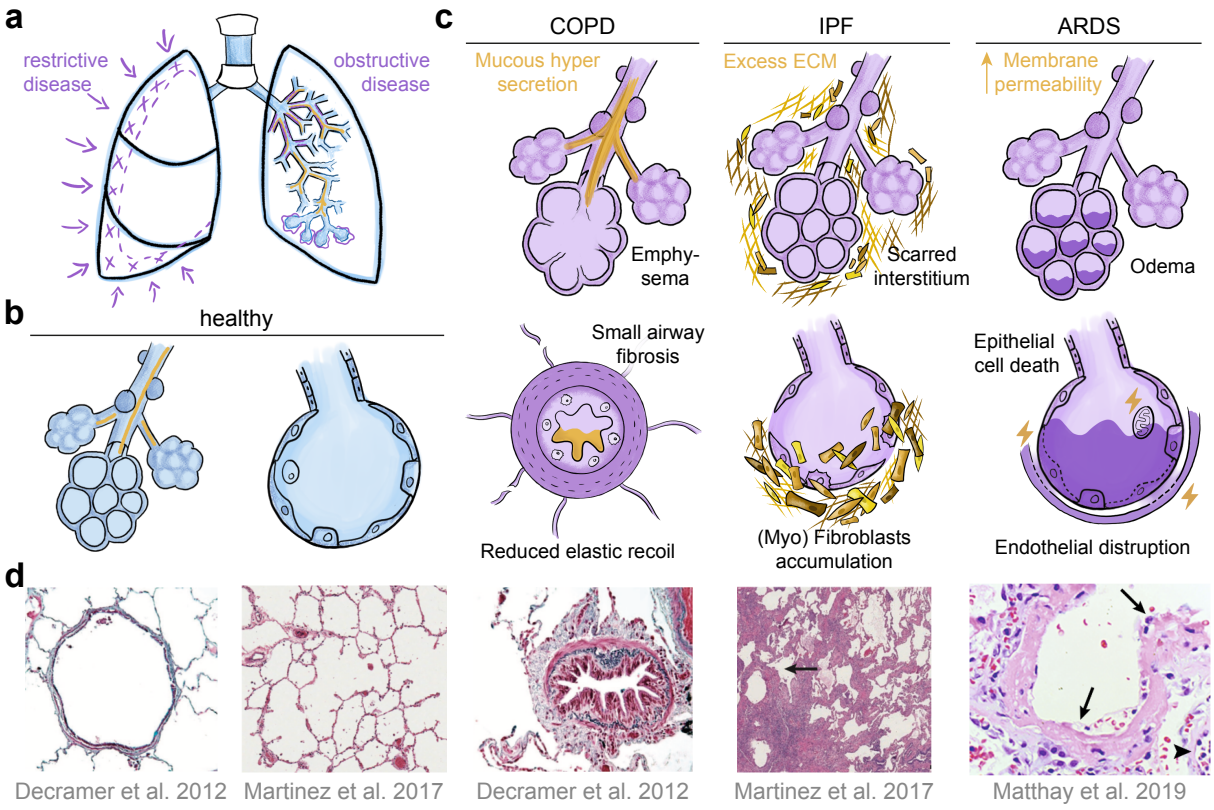


Figure 1.12: Overview of main lung diseases and their characteristics. **a** Comparison of manifestations in restrictive (lung cannot fully expand with air) to obstructive lung conditions (lung cannot fully exhale air). Both cases lead to shortness of breath. **b** Scheme of airway and alveolar structure in healthy lungs. **c** Pathologies of three major lung diseases. In COPD, the small airways are thickened by inflammation, fibrosis and mucous, leading to emphysema and disruption of alveolar attachments. In IPF, the activated epithelium secretes mediators that recruit and activate fibroblasts, which grow resistant to apoptosis and persistently secrete ECM. In ARDS, oedema fluid builds up, first in the interstitium and later in the alveoli. **d** Histology of lung sections from healthy patients and patients with pulmonary diseases.

1.5.1 Chronic obstructive pulmonary disease COPD

The Global Initiative for Chronic Obstructive Lung Disease GOLD describes COPD as[127]

> *"a common, preventable and treatable disease that is characterized by persistent respiratory symptoms and airflow limitation that is due to airway and/or alveolar abnormalities usually caused by significant exposure to noxious particles or gases and influenced by host factors including abnormal lung development."*

Among the key drivers for COPD are cigarette smoking and old age and the likelihood of an outbreak is higher in patients above 40 years of age, with a peak prevalence at 65 years.[128] Due to the population growth and its increased average age, the number of global COPD deaths has increased further in recent years.[129] Especially in developing countries the prevalence has risen strikingly, partially due to the exposure to other forms of air pollutants additional to cigarette smoke.[130,131] Interestingly, not all smokers and exposed individuals develop an airway obstruction. There is evidence showing that non-smokers may develop the disease as well, indicating that there is a genetic susceptibility to the disease.[132] As the name implies, this disease is characterized by progressive airflow obstruction that is only partly reversible, inflammation in the airways, and systemic effects or comorbities. Patients typically have chronic cough, impaired exercise tolerance and shortness of breath over several years due to difficulty exhaling all air from the lungs, also known as air trapping.

COPD can be classified via the GOLD guidelines (see Tab. 1.1), based on the forced expiratory volume in 1 second (FEV1, air volume that can be forcefully exhaled in 1 second) and forced vital capacity (FVC, air volume of air that can forcibly be blown out after full inspiration). In healthy individuals typically above 70% of the vital capacity can be exhaled in the first second. Due to the airflow limitation, this ratio can be heavily decreased along disease progression. The disease burden is often aggravated in patients with other severe conditions, which can be due to bacterial or viral infections.

	GOLD1 (mild)	**GOLD2 (moderate)**	**GOLD3 (severe)**	**GOLD4 (very severe)**
FEV_1:FVC	< 0.70	< 0.70	< 0.70	< 0.70
FEV_1	$\geqslant 80\%$	79–50%	40-30%	$< 30\%$

Table 1.1: GOLD guideline for classification of COPD stages.[127]

Pathogenesis

The progressive airflow limitation is caused by two major pathologic processes: Destruction of the lung parenchyma (emphysema) and remodeling/narrowing of small airways (chronic inflammation). Excessive mucous production by epithelial cells can also contribute to the airway obstruction. The relative contributions of these processes vary from person to person. Emphysema is characterized by the destruction of the gas-exchanging surfaces of the lung (alveoli), consequently leading to the loss of alveolar attachments of the small airways. In healthy individuals elastin fibers are involved in keeping the airways

open, enabling alveolar air to be expired. In COPD the narrowed airways together with emphysema decrease lung elastic recoil and diminish the ability of the airways to remain open during expiration.[133,126]
The inflammation is caused as response to the inhalation of irritants (cigarette smoke, biomass fuel, air pollutants) and involves both innate and adaptive immunity. As a first line of defense increased numbers of neutrophils and macrophages are recruited into the lung as acute inflammatory response, supported by activation of airway epithelial cells and mucous secretion. There is a correlation between numbers of neutrophils, macrophages and lymphocytes in the parenchyma and the degree of inflammation as the disease progresses.[134] Particularly the alveolar macrophages play a key role in orchestrating the inflammatory response. They release inflammatory mediators after activation, which in turn attract other proinflammatory cells, such as circulating neutrophils, monocytes, and lymphocytes.[133,135]

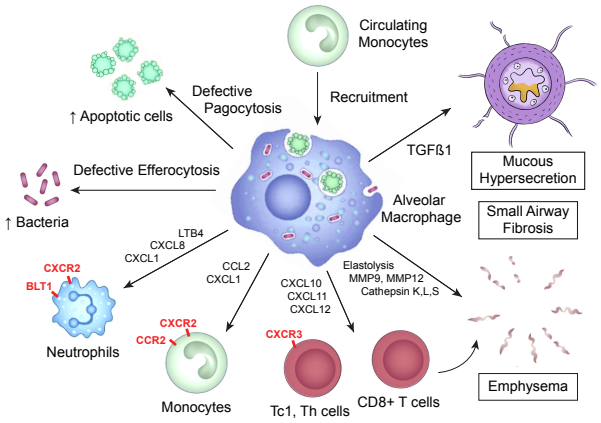


Figure 1.13: Central role of alveolar macrophages in COPD. AM, either tissue resident or derived from circulating monocytes, secrete inflammatory mediators (TNF-α, CXCL1, CXCL8, CCL2, LTB4) that orchestrate the inflammatory process. Release of elastolytic enzymes (matrix metalloproteinases MMP, cathepsins) causes elastolysis, which contributes to emphysema together with cytotoxic T cells. Release of TGF-β1 may induce fibrosis of small airways. Taken and adapted from Barnes et al. (2014)[133] and (2015).[130]

The repair process aims to remodel damaged tissue and return it to its previous state. However, the regenerative mechanisms of the lung are severely compromised in COPD. AM have shown reduced phagocytic uptake of bacteria as well as defects in clearance of apoptotic cells in COPD patients.[136,137] This accumulation of apoptotic cells and bacterial colonization may contribute to the failure to resolve inflammation in COPD. Due to increased numbers of T and B lymphocytes in the lungs, these cells can be organized as lymphoid follicles in COPD. The number of airways containing such lymphoid follicles has been shown to increase with disease progression.[138] Furthermore, excessive release of the enzyme lysosomal elastase from both neutrophils and alveolar macrophages after smoke exposure is hypothesized be involved in the pathogenesis of emphysema. This results in the destruction of elastin (a structural protein of the lung) and cleavage of type IV collagen (involved in integrity of the alveolar wall).[139,126]

Although being a lung disease, COPD is often associated with other chronic conditions. The most common comorbidities are ischaemic heart disease, diabetes, skeletal muscle wasting, osteoporosis, depression, and lung cancer. The frequent association of these severe diseases suggests common risk factors and pathways. As example, tobacco smoke was shown to be a major risk not only for COPD but also for cardiovascular disease, osteoporosis, and lung cancer as well.[140]

1.5.2 Interstitial lung disease ILD

The American Thoracic Society and European Respiratory Society ATS/ERS published an international consensus statement in 2000 on the diagnosis of IPF as[141]

"a specific form of chronic, progressive fibrosing interstitial pneumonia of unknown cause, occurring primarily in older adults, limited to the lungs, and associated with the histopathologic pattern of Usual Interstitial Pneumonia."

The interstitium of the lung fills the space between the alveolar epithelium and the capillary endothelium. It is a lace-like network of tissue that extends throughout the lung area and structurally supports the alveoli. One side of the capillary consists of the fused basement membranes of the epithelial and endothelial layers. On the other side the interstitium is usually wider and is involved in fluid exchange across the endothelium, whereas the thin side is responsible for most of the gas exchange.[126]
Interstitial lung disease (ILD) is an umbrella term for disorders affecting the interstitium and is typically characterized by inflammation and lung fibrosis. The symptoms include dyspnea, which typically worsens on exercise, shallow breathing and irritating cough. The development of lung fibrosis is associated with underlying disorders, such as sarcoidosis, involving granulomatous tissue, i.e. collections of inflammatory cells that form lumps, chronic occupational exposures (silicosis, asbestosis), or hypersensitivity pneumonitis, which develops as reaction to inhaled organic dust or fumes. Pulmonary fibrosis can also arise due to unknown causes and thus be idiopathic in nature (IPF), which is the most common form of ILD.[142] IPF has a median survival of 3 to 5 years after its diagnosis and is generally viewed as a disease of aging, as its prevalence increases drastically with age, mostly affecting adults after their fifth decade of life. In patients older than 65 years, the estimated amount of cases is as high as 400 per 100,000 people.[143] The key feature of IPF is progressive scarring and thickening of the interstitium due to accumulation of extracellular matrix in the distal lung, rendering the lung stiff and compromising its main function of facilitating gas exchange. For the diagnosis of IPF other ILDs with known causes have to be excluded and the presence of usual interstitial pneumonia UIP on surgical lung biopsy is required, which is characterized by patchy chronic inflammation (alveolitis), small aggregates of proliferating fibroblasts (fibroblastic foci) and cystic spaces with thickened walls composed of dense collagen and fibrous tissue (honeycombs, formed by dilated bronchioles, which lead to the destruction of alveolar architecture).[144,145]

Pathogenesis

The lung exhibits remarkable repair mechanisms that allows the well-orchestrated replacement of dead or damaged cells after injury. This includes[146]

1. Initial response: Epi-/endothelial cells release inflammatory mediators at injury site
2. Platelets: Coagulate to prevent blood loss and increase permeability of blood vessels
3. Inflammatory phase: Neutrophils and macrophages are recruited to clear dead cells
4. Proliferation and migration of fibroblasts: Deposition of extracellular matrix, which in turn provides structural and biochemical support to neighbouring cells
5. Final remodelling and resolution: Restores normal tissue architecture

The causes that lead to the development of IPF are not completely understood, still there is rising evidence proposing that repair mechanisms might be dysregulated. IPF potentially arises when repetitive epithelial injury to the lung triggers an abnormal regeneration response.[126,145] Macrophages, as crucial regulators of fibrosis, have been at the center of several recent publications in the field. They are often in close proximity to fibroblastic foci and have been shown to be involved in ECM processing through secretion of matrix metalloproteases or by direct uptake of collagen. During the aberrant regeneration cascade in IPF their numbers increase and they produce profibrotic cytokines, contributing to pulmonary fibrosis.[147,103] In a recent publication Aran et al. (2019)[148] could identify a profibrotic macrophage subtype that appears after bleomycin injury in mouse lungs. The cell type localized to the fibrotic niche and exerted a transitional profile between alveolar and monocyte-derived macrophages. While macrophages are well studied in mouse models, the knowledge about their role in human lung fibrosis is still incomplete. There has been evidence that proinflammatory cytokines (e.g. macrophage inflammatory protein CCL18, Chitinase CHI3L1), produced by AM during aberrant wound-healing, are elevated in BAL fluids of IPF patients.

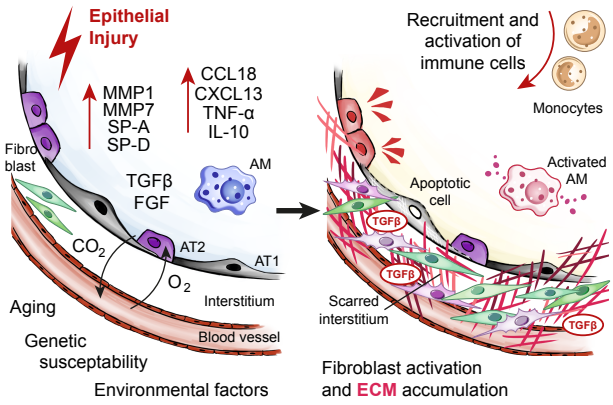


Figure 1.14: Maladaptive responses to injury in IPF underlying the fibrotic process. Activated epithelium releases growth factors, matrix metalloproteinases MMPs and further signals that induce activation of fibroblasts (TGF-β1, FGF), which grow resistant to apoptosis and persistently produce ECM. Resident AM are activated and circulating monocytes are recruited to modulate fibrotic responses and secrete inflammatory mediators (CCL18, CXCL13, IL-10, TNF-α). Adapted from Martinez et al. (2017)[145] and Desai et al. (2018).[149]

The role of lymphocytes in fibrosis is still controversial. It is known that lymphocytes are not required for the experimental induction of fibrosis in mice,[150] nonetheless regulatory T cells for instance can produce the Interleukin IL-10 and platelet-derived growth factor TGF-β, having the potential to both promote or suppress fibrosis depending on the context.[149]

Although increase of inflammatory cells suggested that IPF is a principally inflammatory disease, many studies have shown that the important aspect lies rather in the complex cross talk between the alveolar epithelium and mesenchymal cell types.[142] There are findings identifying hyperplastic AT2 cells that overlie fibroblastic foci. Such AT2 cells can be aberrantly active and secrete elevated levels of mediators that promote the migration and expansion of fibroblasts. These include TGF-β (primary inductor of fibroblast to myofibroblast differentiation), connective tissue growth factor CTGF, several matrix metalloproteinases (e.g. MMP1, MMP7, MMP19) and a number of chemokines (e.g. the immune cell-attracting chemokine CCL18 and certain interkeukins).[151] The activated fibroblast can grow resistant to apoptosis and continuously secrete ECM components, an accumulation thereof leading to an unresolvable fibrotic scar.

1.5.3 Acute respiratory distress syndrome ARDS

The 1994 American European Consensus Conference defined the following pathological findings as the gold standard for diagnosis of ARDS:[152]

"ARDS requires the presence of an acute onset, persistent dyspnea, bilateral infiltrates on chest radiograph consistent with pulmonary edema, and the lack of evidence of cardiogenic pulmonary edema or a pulmonary artery occlusion."

In 2011 this standard was updated to the Berlin definition, which furthermore established that the respiratory failure has to occur within one week of a known insult or new or worsening respiratory symptoms and that the respiratory failure could not be explained by cardiac function/volume overload. Depending on the level of blood oxygenation, the diagnosis of ARDS can be categorized into "mild", "moderate" and "severe". Patients typically encounter difficulty in breathing, progressive respiratory failure from pulmonary edema and often require mechanical ventilation because of severe arterial hypoxemia.[153]

ARDS commonly develops as an end result of a variety of insults to the lung. Critical to its pathogenesis is epithelial injury, the extent of which is also indicative of the severity of ARDS. The injury can be due to pneumonia by bacterial and viral products, hyperoxia or extrinsic factors, including sepsis from non-pulmonary sources, pancreatitis and major trauma (blunt or penetrating injuries, burns). As a result of the initial injury, proinflammatory cytokines are released for the activation and recruitment of neutrophils. These neutrophils subsequently release reactive oxygen species, proteases and cytokines that damage AT1 cells and capillary endothelial cells, leading to a disruption of the tight barrier properties and increased permeability of alveolar endothelial and epithelial barriers to liquid and proteins.[126]
In healthy conditions both AT1 and AT2 cells have the ability to absorb excess fluid from the airspaces via ion transport channels. Once the oedematous fluid is absorbed into the lung interstitium, it can be removed by lymphatics and the lung microcirculation. The increased permeability during pathogenesis of ARDS however leads to an extravascular accumulation of oedematous fluid rich in proteins, neutrophils and red blood cells into the alveolar space and interstitium, as well as in the appearance of hyaline membranes, which are composed of proteins and dead cells that line the alveoli.[154,155]

The recent surge in coronavirus disease 2019 (COVID-19) had far-reaching effects on a global scale. In its most severe form, COVID-19 manifests as ARDS and is associated with prolonged ventilator dependence in intensive care units and high mortality.[156] The next section will give a brief introduction to this type of acute respiratory syndrome.

Coronavirus disease 2019 COVID-19

Coronaviruses are a diverse group of viruses infecting many different animals, and can cause mild to severe respiratory infections in humans. COVID-19 is caused by the highly transmissible and pathogenic severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).[157] Starting with reports of several cases of pneumonia of unknown cause in Wuhan back in late December 2019, the number of confirmed cases increased explosively to several thousands. Mere weeks later on 30 January, the WHO declared the novel coronavirus outbreak a public health emergency of international concern.[158]

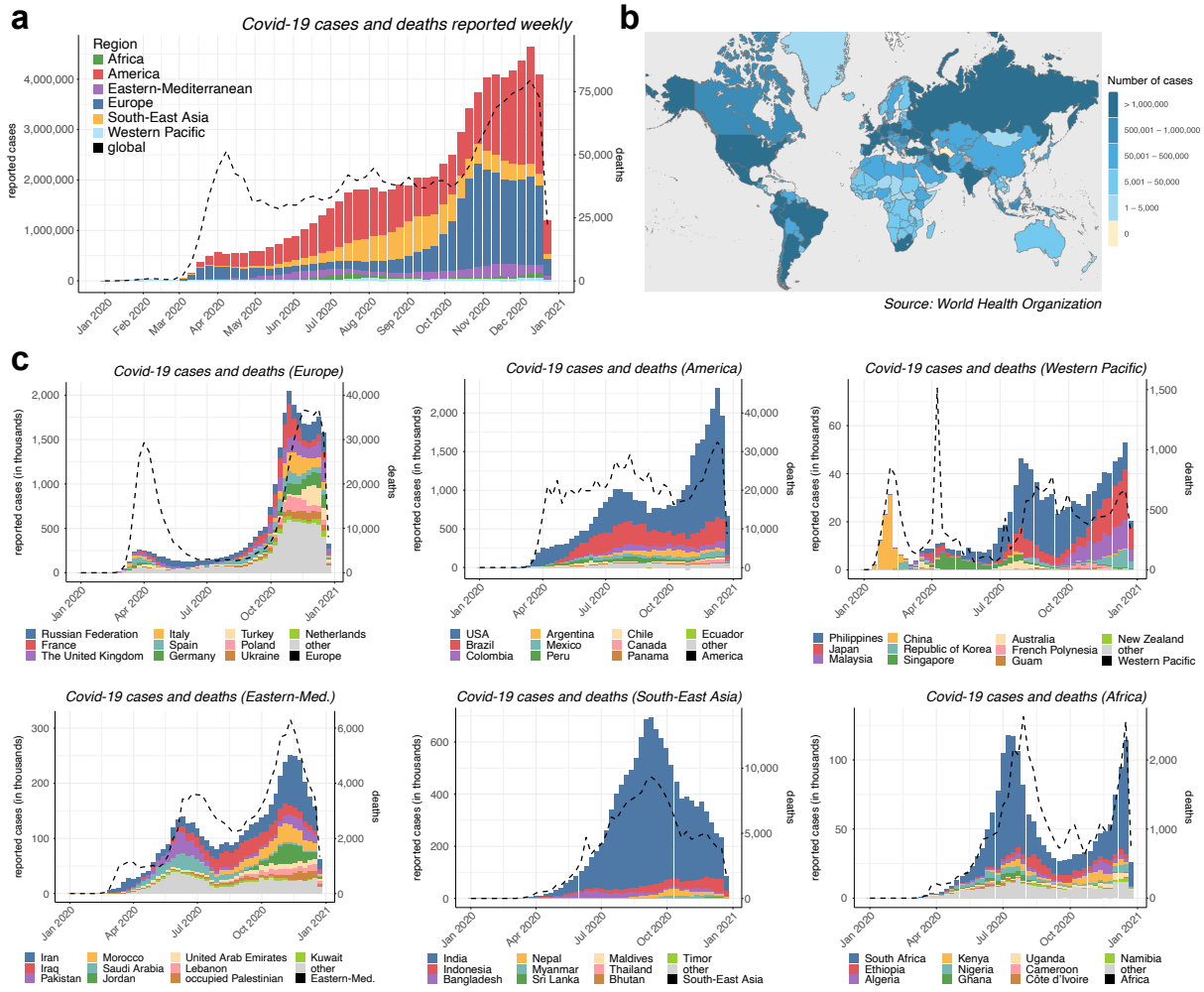


Figure 1.15: Global impact of COVID-19 in 2020 a Number of globally reported COVID-19 cases and deaths by region. X-axis shows week, y-axis shows number of cases as bars (left) and number of global deaths (right) as dashed line. **b** Choropleth map of countries coloured by total number of reported COVID-19 cases. **c** COVID-19 cases and deaths split by region as defined by WHO. For each region, the 10 countries with highest number of total deaths are highlighted. Raw data taken from *World Health Organization* (retrieved: 30.Dec.2020).[159]

Despite heavy containment measures and travel restrictions over months, the ongoing outbreak of COVID-19 poses an extraordinary threat to the global public health (Fig. 1.15) and has claimed over 4,5 Mio lives (source: Johns Hopkins University, as of September 2021). COVID-19 manifestation in patients ranges from mild symptoms (fever, fatigue, dry cough) up to severe respiratory failure. The vast majority of young people show only mild disease or are asymptomatic, whereas older people above 60 years, in particular men with comorbidities, have a greater risk of developing a severe respiratory disease.[160]

The viral envelope is coated by spike glycoprotein (S), envelope (E), and membrane (M) proteins. Viral cell entry is facilitated by binding to the human host factor Angiotensin-Converting Enzyme 2 (ACE2) as the target receptor for SARS-CoV[161] (Fig. 1.16). Certain host proteases are needed for cleavage of the virus' spike protein and mediating the viral

entry, such as the Type II Transmembrane Serine Protease TMPRSS2,[162] Cathepsin L and Furin.[163] The ACE2 receptor is found on the epithelium of a number of organs such as the intestine, endothelial cells in the kidney and blood vessels. The main sites of ACE2 and TMPRSS2 co-expressing cells were rigorously investigated and shown to be mainly in nasal secretory cells, bronchial branches and AT2 cells in the lung and ileal absorptive enterocytes, explaining some of the tissue tropism of SARS-CoV-2.[164,165]

Upon binding to epithelial cells, SARS-CoV-2 starts replicating and migrating further down to alveolar epithelial cells in the lungs, potentially triggering a strong immune response due to the rapid expansion of the virus. The initial inflammatory response attracts T cells to the site of infection in order to eliminate infected cells, facilitating recovery in most people. In some patients however, the host immune response is dysfunctional and clearance of the virus is not achieved. Instead, the cytokine storm involving proinflammatory cyto- and chemokines (TNF-α, IL-1, IL-6, IL-8) induces respiratory distress which may progress to acute lung injury or ARDS during the incubation period of 1 to 14 days. Post-mortem lung tissue histology of deceased COVID-19 patients confirmed characteristics of ARDS, including diffuse alveolar damage, hyaline-membrane formation and interstitial mononuclear infiltrates, reflecting the inflammatory nature of the injury.[166]

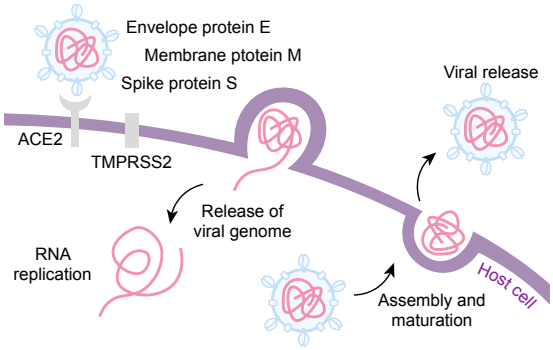


Figure 1.16: SARS-CoV-2 entry mechanism. The virus binds to ACE2 as the host target cell receptor in synergy with the host's transmembrane serine protease TMPRSS2. Upon membrane fusion, the viral genome is released into the host cytoplasm where it replicates and maturates. After viral assembly and maturation, the virus is released. Adapted from Hu et al. (2021)[157] and Cevik et al. (2020).[166]

1.6 Human Cell Atlas HCA

The 150 year old effort of characterizing all cell types present in an organism is still unfulfilled, but far from unreachable. On the contrary, the advances as described in the first section, and the recent progress in methods, especially high throughput single-cell profiling, have accelerated the potential completion of this endeavour.
Akin to the Human Genome Project, which aimed to build a reference map for all human genes, a comparable reference map of the molecular phenotypes of cells in healthy human tissue would pave the way for systematic studies of physiological states, developmental trajectories, regulatory circuits and cell-cell interactions. Just a few years ago, the idea of the Human Cell Atlas emerged, an international effort incorporating diverse scientific expertise in order to provide a harmonized framework and description of all cells in the human body. This will not be limited to transcriptomic profiles alone, but is planned to be connected with classical cellular descriptions like spatial location, morphology and function as well in the near future.[14]

Notably, the lung was the central focus of biological research in recent months, and of great interest for medical research long before the pandemic. The Lung Cell Atlas forms the flagship-project of the HCA, as the lung was among the 12 priority organs within the consortium. Although more than 40 discrete cell types have been reported, new findings and novel cell states continue to be identified, making it evident that our understanding is still incomplete.[78]

Among many other things, the year 2020 has show-cased how fast researchers across the globe join forces and are willing to share data in order to tackle a contemporary crisis, accompanied by a pressing need for immediate insights.
Large-scale collaborations provide statistical power that is required to uncover underlying patterns. For instance, in a recent study by Muus and colleagues (2021)[167] the integration of over 200 donors across different ethnicities, the majority of which were part of unpublished data sets at that time, enabled to find associations of COVID-19 to age, sex and smoking status. Many more studies have already proven the convenience of this international effort, identifying virus-affected cell populations across tissues at the necessary resolution.[165,164,168]

1.7 Outline of this thesis

In the following, the overall structure of this thesis will be outlined briefly. A graphical overview is given in Fig. 1.17. After having introduced the main experimental protocols and relevant parts of lung biology as a solid groundwork, chapter 2 provides further details on the workflow, pre-processing steps and tools that were frequently used during data analysis. The results section is divided into 4 sub-chapters, which give a brief context and rational for each project and then list the relevant outcome of analysis. Their order mirrors the chronological appearance during my PhD period.

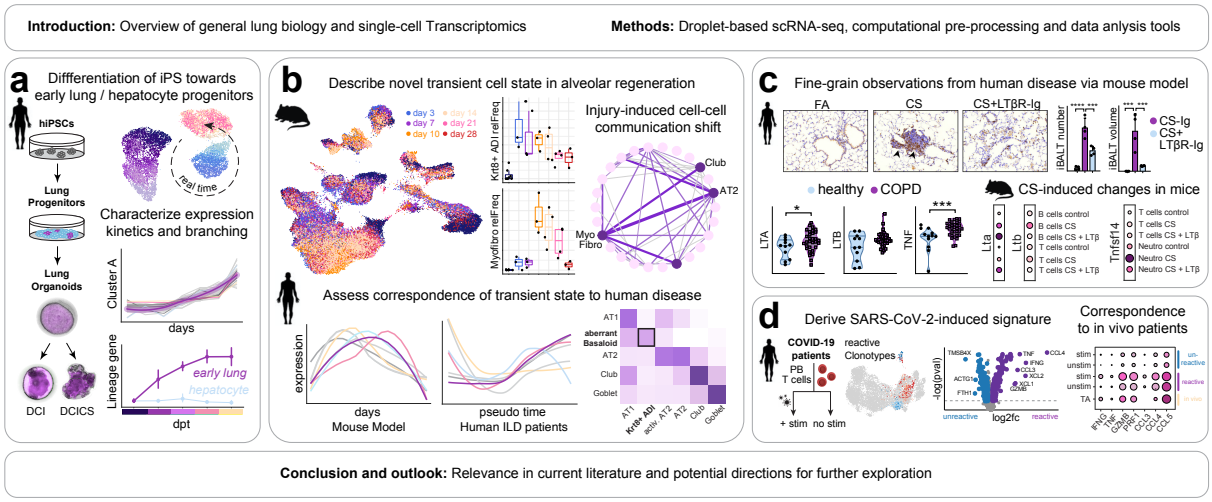


Figure 1.17: Outline of this thesis' results chapter. **a** Chapter 3.1 Introduction of longitudinal scRNA-seq data sets in the form of a differentiation trajectory starting from iPS. **b** Chapter 3.2 Bleomycin-induced lung injury and its correspondence to human ILD. **c** Chapter 3.3 Inhibition of LTβR signalling and its cellular characterization via CS-exposed mice. **d** Chapter 3.4 Derive ex vivo signatures of T cells from severely affected COVID-19 patients.

A general overview of the cell populations, affected cell types and frequency shifts is added as a common starting point for most analyses, followed by a more specific description of the relevant aspects of each project individually.

Chapter 3.1 revolves around an early project that describes the lung development starting from human induced pluripotent stem cells up to lung progenitors. Due to the dense time points of sampling with Drop-seq, a detailed description of gene expression and their temporal patterns along the differentiation was possible. Already prior to accessing the transcriptomes, the co-existence of lung progenitors and hepatocytes during the differentiation protocol has been perceivable. The genes potentially driving this lineage specification are therefore additionally explored on the single-cell level.

Chapter 3.2 embodies the main project, as most of the time was devoted to the exploration of the corresponding data sets. The mechanisms involved in lung regeneration were characterized using the bleomycin model to mirror acute lung injury. The mouse model captures aspects of human lung fibrosis, therefore the focus shifted towards the main affected cell types: the epithelial compartment and their interaction with stromal populations. In the second part of this chapter, the transferability to human disease, particularly pulmonary fibrosis, is assessed. This generated hypotheses regarding dysregulated regeneration that could drive disease pathology, particularly the accumulation of transient cell states that persist in human disease.

Chapter 3.3 inverts the approach and starts out with observation from human patients with obstructive pulmonary disease. Particularly the formation of lymphoid follicles in disease was of interest, and how these could be dispersed following a novel therapeutic approach. The cellular mechanisms for its efficacy however were not entirely clear, and were explored using a cigarette smoke exposure mouse model. Furthermore, given the developments during the writing of this thesis, this mouse data was suited for another unrelated research question. The status of smoking impacted the development of the coronavirus disease 2019, and could be linked to an increased expression of the main entry factor in some epithelial populations from smoke-exposed mice.

Chapter 3.4 is the final chapter and was in fact not anticipated as part of this work. It focuses on the recently emerged SARS-CoV-2 pathogen and particularly its effects on T cells, which are essential in the host's adaptive immune response. Antigen-specific gene expression changes were derived from an targeted in vitro experiment, in which T cells from severely diseased patients were either stimulated with the virus antigen or left unstimulated. The derived signatures were validated in a number of COVID-19 patient cohorts. Finally, the induced changes were set into broader context by exploring cellular communication with other immune cells.

The pre-processing and quality control for each of these projects is structured according to the overriding workflow which will be introduced in chapter 2. Following a section-wise description of the projects, the discussion will build an overarching theme by connecting the key messages and relate these to recent movements in the field. To streamline the chapters and avoid overcrowding, the exact filtering criteria and final parameters for visualization of each of the data sets are listed in the appendix.

1.8 Scientific publications

During the course of my PhD, I was involved in the preparation and publication of several articles. The following list shows submitted manuscripts and peer-reviewed publications relevant for each chapter of this thesis:

Chapter 1 - Differentiation trajectory of human pluripotent stem cells

- Ori O*, **Ansari M***, Angelidis I, Theis FJ, Schiller HB and Drukker M. Single cell trajectory analysis of human pluripotent stem cells differentiating towards lung and hepatocyte progenitors. *Manuscript submitted.*

Chapter 2 - Bleomycin-induced lung injury and human ILD

- Strunz M*, Simon LM*, **Ansari M**, Kathiriya JJ, Angelidis I, Mayr CH, Tsidiridis G, Lange M, Mattner LF et al. and Theis FJ, Schiller HB. Alveolar regeneration through a Krt8^{+} transitional stem cell state that persists in human lung fibrosis. *Nat Commun.* 2020 Jul 16; 11(1):3559.

- Mayr CH*, Simon LM*, Leuschner G, **Ansari M**, Schniering J, Geyer PE, Angelidis I, Strunz M, Singh P, Kneidinger N, et al. and Theis FJ, Schiller HB. Integrative analysis of cell state changes in lung fibrosis with peripheral protein biomarkers. *EMBO Mol Med.* 2021 Apr; 13(4):e12871.

Chapter 3 - Cigarette smoke-exposed mice and human COPD, COVID-19

- Conlon TM*, John-Schuster G*, Heide D, Pfister D, Lehmann M, Hu Y, Ertüz Z, Lopez MA, **Ansari M**, Strunz M, Mayr C, Angelidis I, et al. and Königshoff M, Heikenwalder M, Yildirim AÖ. Inhibition of LTβR signalling activates WNT-induced regeneration in lung. *Nature.* 2020 Dec; 588(7836):151-156.

- Muus C*, Luecken MD*, Eraslan G*, Sikkema L*, Waghray A*, Heimberg G*, Kobayashi Y*, Vaishnav ED*, Subramanian A*, Smillie C*, Jagadeesh KA*, Duong ET*, Fiskin E*, Triglia ET*, **Ansari M***, Cai P*, Lin B*, Buchanan J*, Chen S*, Shu J*, Haber AL*, Chung H*, Montoro DT*, et al. and Human Cell Atlas Lung Biological Network. Single-cell meta-analysis of SARS-CoV-2 entry genes across tissues and demographics. *Nat Med.* 2021 Mar; 27(3):546-559.

Chapter 4 - Ex vivo signatures of SARS-CoV-2-reactive T cells

- Fischer DS*, **Ansari M***, Wagner KI*, Jarosch S, Huang Y, Mayr CH, Strunz M, Lang NJ, D'Ippolito E, et al. and Theis FJ, Busch DH, Schiller HB, Schober K. Single-cell RNA sequencing reveals ex vivo signatures of SARS-CoV-2-reactive T cells through 'reverse phenotyping'. *Nat Commun.* 2021 Jul; 12(1):4515.

* indicates equal contribution.

Further publications

Within collaborations, I was involved in further projects which are not specifically discussed in this thesis. The contributions were mostly scRNA-seq data analyses and have resulted in the following co-authorships:

- Angelidis I*, Simon LM*, Fernandez IE, Strunz M, Mayr CH, Greiffo FR, Tsitsiridis G, **Ansari M**, Graf E, Strom TM, and Theis FJ, Schiller HB. An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nat Commun.* 2019 Feb; 10(1):963.

- Fischer A*, Koopmans T*, Ramesh P, Christ S, Strunz M, Wannemacher J, Aichler M, Feuchtinger A, Walch A, **Ansari M**, et al. and Schiller HB, Rinkevich Y. Post-surgical adhesions are triggered by calcium-dependent membrane bridges between mesothelial surfaces. *Nat Commun.* 2020 Jun; 11(1):3068.

- Hadrup N, Zhernovkov V, Jacobsen NR, Voss C, Strunz M, **Ansari M**, Schiller HB, Stoeger T, et al. and Saber AT, Vogel U. Acute Phase Response as a Biological Mechanism-of-Action of (Nano)particle-Induced Cardiovascular Disease. *Small.* 2020 May; 16(21):e1907476.

- Ziegler CGK*, Allon SJ*, Nyquist SK*, Mbano IM*, Miao VN, Tzouanas CN, Cao Y, Yousif AS, Bals J, Hauser BM, Feldman J, Muus C, et al. and Shalek AK, Ordovas-Montanes J, HCA Lung Biological Network. SARS-CoV-2 Receptor ACE2 Is an Interferon-Stimulated Gene in Human Airway Epithelial Cells and Is Detected in Specific Cell Subsets across Tissues. *Cell.* 2020 May; 181(5):1016-1035.

- **Ansari M**, Fischer DS and Theis FJ. Learning Tn5 Sequence Bias from ATAC-seq on Naked Chromatin. *ICANN 2020. Lecture Notes in Computer Science*, 2020 Oct;

- Lange M, Bergen V, Klein M, Setty M, Reuter B, Bakhti M, Lickert H, **Ansari M**, Schniering J, Schiller HB, Pe'er D, and Theis FJ. CellRank for directed single-cell fate mapping. *Nat Methods.* 2022 Feb;19(2):159-170.

- Spix B, Butz ES, Chen CC, Rosato AS, Tang R, Jeridi A, Kudrina V, Plesch E, Wartenberg P, Arlt E, Briukhovetska D, **Ansari M**, Günsel GG et al. and Grimm C. Lung emphysema and impaired macrophage elastase clearance in mucolipin 3 deficient mice. *Nat Commun.* 2022 Jan;13(1):318.

- Chakraborty A, Mastalerz M, **Ansari M**, Schiller HB and Staab-Weijnitz CA. Emerging Roles of Airway Epithelial Cells in Idiopathic Pulmonary Fibrosis. *Cells.* 2022 Mar;11(6):1050.

- Günsel GG*, Conlon T*, Jeridi A*, Kim R, Ertüz Z, Lang NJ, **Ansari M**, Novikova M, Jiang D, Strunz M et al. and Yildirim AÖ. The Arginine Methyltransferase PRMT7 Promotes Extravasation of Monocytes resulting in tissue injury in COPD. *Nat Commun.* 2022 Mar 14;13(1):1303

- Wu X, Bos IST, Conlon TM, **Ansari M**, Verschut V et al. and Gosens R. A transcriptomics-guided drug target discovery strategy identifies receptor ligands for lung regeneration. *Sci Adv.* 2022 Mar 25;8(12):eabj9949.

Submitted manuscripts

Finally, there are several projects at our institute that are still under active compilation, or submitted to peer-reviewed journals. Author listings and titles are subject to change.

- Kalgudde Gopal S*, Stefanska AM*, **Ansari M**, Jiang D, Ramesh P, Bagnoli JW, Correa-Gallegos D, Christ S, Angelidis I, et al. and Schiller HB, Rinkevich Y. Fate restricted stromal fibroblasts and adipocytes demonstrate multi-modal responses to tissue injury. *Manuscript submitted and under review.*

- Leuschner G*, Mayr CH*, **Ansari M**, Seeliger B, Frankenberger M, Kneidinger N, Hatz RA, Anne Hilgendorff A, Prasse A, Behr B, Mann M and Schiller HB. A proteomics workflow reveals predictive autoantigens in idiopathic pulmonary fibrosis. *Manuscript submitted.*

Work in Progress

- Voss C*, **Ansari M***, Strunz M, Angelidis I, et al. and Theis FJ, Schiller HB, Stoeger T. Cracking nanotoxicology's bottleneck: single-cell transcriptomics to decipher nanomaterial specific pulmonary cellular response patterns.
Manuscript in preparation.

- Stoleriu MG*, **Ansari M**, Heydarian M, Strunz M, Voss C, Schamberger AC, Schneider JJ, Gerckens M, Burgstaller G, Castelblanco A, et al. and Stoeger T, Schiller HB, Hilgendorff A. Altered basal cell state in end stage COPD propagates to differentiated progeny and causes secretory to ciliated cell imbalance.
Manuscript in preparation.

- Schniering J*, Mayr CH*, Ogar P, Strunz M, Angelidis I, Lang NJ, **Ansari M** et al. and Schiller HB. Diversity and dynamics of stromal-parenchymal cell crosstalk in alveolar lung regeneration. *Manuscript in preparation.*

- Yang L, Heumos L, Angelidis I, Strunz M, **Ansari M**, Zhou S, Mayr CH, Simon LM, Theis FJ. and Adler H, Schiller HB. Single-cell transcriptomic dissection of virus induced immunopathology in interferon gamma receptor null mice.
Manuscript in preparation.

- Yan H*, **Ansari M**, Lehmann M, et al. and Schiller HB, Koenigshoff M. Dissecting the cellular and molecular abnormalities of distal lung epithelial progenitor cells in COPD. *Manuscript in preparation.*

* indicates equal contribution.

Chapter 2

Methods

2.1 Droplet-based capture methods

ScRNA-seq was the method of choice to tackle the objectives throughout this thesis. In particular droplet-based capture methods were used to generate the data sets, due to their cost efficiency and high throughput as described in section 1.3. Drop-seq has been introduced by Macosko et al. in 2015[44] and is based on the encapsulation of cells into nanoliter-sized droplets by the means of microfluidic devices and specialized microbeads. A barcoding strategy was developed in order to retain a molecular memory of the cell identity a mRNA was isolated from (Fig. 2.1). Additional to a cell and molecule specific sequence, the primers on a beads contain a PCR handle to enable amplification in later steps. A cell's mRNA is instantly released upon lysis when it is encapsulated in a droplet, and is captured by barcoded oligonucleotides that are attached on the ideally co-captured beads. All droplets are collected and broken to release the STAMPs, PCR and reverse transcription are carried out for the capture and amplify the transcripts and finally tagmented, which is the random cut of the transcripts and the addition of sequencing adapters, resulting in single cell libraries that are ready for sequencing. The beads contain more than 108 individual primers sharing the same cell barcode but differ in unique molecular identifiers, which enables mRNA transcripts to be digitally counted and traced back to the cell they originated from later on.

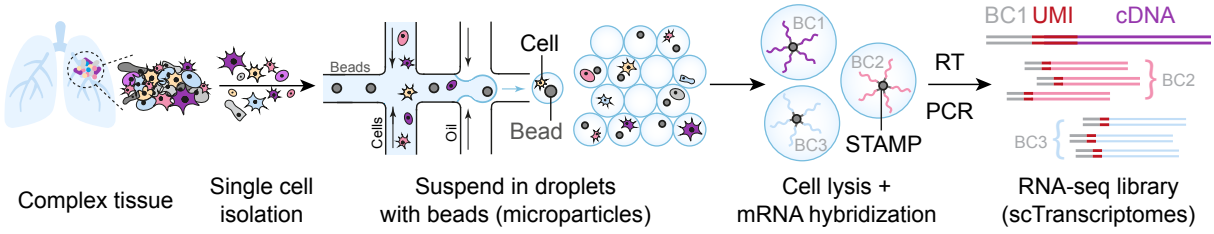


Figure 2.1: Drop-seq: capture single cells along with sets of uniquely barcoded beads. Scheme of single-cell mRNA-seq library preparation with Drop-seq. The tissue of interest is dissolved into individual cells. A microfluidic device enables capturing these cells in nanoliter droplets containing lysis buffer. Each droplet ideally compartmentalizes one cell and one microparticle (bead, with barcoded primers on its surface). The primers on each bead contain an unique cell barcode (BC) and an additional unique molecular identifier (UMI). Upon lysis of each cell within a droplet, its mRNAs bind to the primers. The mRNAs are reverse transcribed (RT) to cDNAs and amplified using PCR, forming the RNA-seq library containing the transcriptomes of each single cell. Adapted from Macosko et al. (2015).[44]

2.2 Experimental methods

The purchases of reagents and experimental procedures listed in this section were all performed by colleagues as indicated in the corresponding chapter. Nonetheless, the essential steps in generating the scRNA-seq data sets that were used in the presented studies are outlined briefly.

2.2.1 Differentiation of iPSCs to NKX2-1$^+$ lung progenitors [section 3.1]

Stem cell maintenance. Human iPSC line NKX2-1^{eGFP+} (Hannover Medical School)[169] was maintained in feeder-free conditions (StemMACS iPS-Brew XF, Miltenyi Biotech) and passaged with Accutase (Sigma-Aldrich) on tissue cell culture plates pre-coated with 1:100 dilution of Geltrex Basement Membrane Matrix in DMEM/F-12 (ThermoFisher).

Differentiation protocol. To induce definitive endoderm differentiation, hiPSCs were maintained in iPSC-Brew and when reached 80% confluency (day 0), the cells were rinsed with DPBS and incubated in Accutase (Sigma-Aldrich) for 10 minutes, at 37°C. The detached cells were triturated into single-cell suspensions and seeded onto 24-well plates pre-coated with 1:40 Growth Factor Reduced (GFR) Matrigel (Corning), in a density of 2 x 10 cells/cm. Cells were immediately treated with 100ng/ml activin-A, 1µM CHIR99021, and 10µM Y-27632 (RD Systems), in Definitive Endoderm Basal Media (DE-BM). On days 1-6 the DE-BM was supplemented with 100ng/ml activin-A, 1µM CHIR99021 (RD Systems) and 0.25mM (day 1) and 0.125mM (days 2-6) sodium butyrate (Sigma-Aldrich). For the foregut endoderm stage the Basal Mediums (FE-BM1 and FE-BM2) were prepared as follows, FE-BM1: DMEM/F12, 1 x GlutaMAX, 1 x B-27 and N-2 supplements, 50U/ml of penicillin/streptomycin, 0.05mg/ml of L-ascorbic acid and 0.4mM of monothioglycerol or FE-BM2: 75% IMDM, 25% Ham's F-12, B27 supplement and N2 supplement, 0.05% bovine serum albumin, 1 x GlutaMAX, 50U/ml of penicillin/streptomycin, and 0.05mg/ml of L-ascorbic acid, 0.4mM of 1-thioglycerol. On day 6 DE cells were collected with Accutase for 10 minutes at 37°C and re-plated in a density of 1:2-1:4 onto GFR Matrigel-coated plates. The cells were treated for 2 days (days 6, 7) with 50ng/ml SHH (RD Systems), 2µM dorsomorphin (Tocris) and 10µM SB431542 (Miltenyi Biotec), supplemented with 10µM Y-27632 (RD Systems) for the first 24 hours. On day 8 the medium was changed to FE-BM1.
To induce the lung progenitor stage on day 10, the medium was switched to FE-BM1 or BM2 containing 20ng/ml recombinant human BMP4 (RD Systems), 50nM retinoic acid (Sigma-Aldrich), 3µM CHIR99021 and 20ng/ml rhFGF10 (RD Systems). For the inhibition of Notch and TGF-β pathways, 10µM SB431542 and 100µM DAPT (TOCRIS Bioscience) were additionally used at the LP induction stage of differentiation respectively. [Taken from Ori et al. (2021)[170]]

2.2.2 Animal handling [section 3.2 and 3.3]

All mice used in the presented studies were purchased from Charles River, Germany, and maintained at the animal husbandry of the Helmholtz Zentrum München, Munich, Germany. Pathogen-free female C57BL/6J mice were purchased from Charles River Germany and maintained at the appropriate biosafety level at constant temperature and humidity with a 12h light cycle. Animals were allowed food and water ad libitum.

Mice were randomly allocated into experimental groups with no statistical methods used to predetermine sample size. Sample sizes were chosen based on similar studies from the literature and sufficient to detect statistically significant differences between groups. All animal experiments were approved by the ethics committee for animal welfare of the local government for the administrative region of Upper Bavaria (Regierungspräsidium Oberbayern) and were conducted under strict governmental and international guidelines in accordance with EU Directive 2010/63/EU.

Bleomycin treatment. Mice were divided randomly into two groups and treated with either saline (PBS) or bleomycin (Bleo). Lung injury and pulmonary fibrosis were induced by single-dose administration of bleomycin hydrochloride (Sigma Aldrich), which was dissolved in sterile PBS and given at 2U/kg (oropharyngeal instillation) and 3U/kg (intratracheal instillation) bodyweight. Mice were sacrificed at designated time points after instillation. Animals were under strict observation with respect to phenotypic changes, abnormal behaviour and body weight loss. [Taken from Strunz et al. (2020)[171]]

Cigarette smoke exposure and LTβR–Ig treatment. CS was generated from 3R4F Research Cigarettes (Tobacco Research Institute, University of Kentucky), with the filters removed. Mice were whole-body-exposed to active 100% mainstream cigarette smoke of 500mg m^{-3} total particulate matter for 50 min twice per day for 4 and 6 months in a manner mimicking natural human smoking habits. CO concentrations in the exposure chamber were constantly monitored by using a GCO 100 CO Meter (Greisinger Electronic) and reached values of 288 ± 74 ppm. All mice tolerated CS-mediated CO concentrations without any sign of toxicity, with CO-Hb levels of 12.2 ± 2.4%. In two parallel experiments, mice were treated with an LTβR–Ig fusion protein (80µg intraperitoneally, weekly) (muLTβR-muIgG, Biogen Idec) or control-Ig (MOPC, Biogen Idec) for 2 months, starting from 2 and 4 months of CS exposure. Control mice were kept in a filtered air environment, but exposed to the same stress as CS-exposed animals. 24 hours after the last exposure, mice were sacrificed. [Taken from Conlon et al. (2020)[172]]

2.2.3 Human tissue handling [section 3.2 and 3.4]

Human tissue of the Munich cohorts were obtained from the bioArchive of the Comprehensive Pneumology Center Munich (CPC-M). Written informed consent was received from all patients, and the study was approved by the local ethics committee of the Ludwig-Maximilians University of Munich, Germany (EK 333-10 and 382-10). ILD lung tissue for single-cell analysis was freshly obtained after lung transplantation at the University Hospital Munich and compared to lung tissue of non-CLD patients as tumor free, uninvolved lung tissue freshly obtained during tumor resections, performed at the lung specialist clinic "Asklepios Fachkliniken Munich-Gauting". All participants gave written informed consent. The study was approved by the local ethics committee of the Ludwig Maximilians University (LMU), Munich, Germany. [Taken from Mayr et al. (2021)[173]]

T cells and antigen-specific stimulation. Munich cohort patients were PCR-confirmed SARS-CoV-2 positive, admitted to the ICU in the University Hospital of the Ludwig-Maximillian's University, Munich (n = 5), or the Asklepios Lung Clinic Munich-Gauting, Gauting (n = 4), for treatment of severe COVID-19 requiring invasive, mechanical ventilation. PBMCs and TA samples were taken at the end of April 2020.

Written informed consent was obtained from the donors or their care-givers, usage of the blood samples was approved according to national law by the local Institutional Review Board (Ethikkommission der Medizinischen Fakultät der LMU München) and samples were used according to legal provisions defined by the German Infection Protection Act. [Taken from Fischer et al. (2021)[174]]

2.2.4 Generation of single-cell suspensions for whole-lung tissue

Mouse lungs. Single-cell suspensions were generated as previously described.[175] Briefly, lung tissue was perfused with sterile saline through the heart and the right lung was tied off at the main bronchus. The left lung lobe was subsequently filled with 4% paraformaldehyde for histologic analysis. Right lung lobes were removed, minced, and transferred for mild enzymatic digestion for 20–30 min at 37 °C in an enzymatic mix containing dispase (50 caseinolytic U/ml), collagenase (2mg/ml), elastase (1mg/ml), and DNase (30µg/ml). Cells were harvested by straining the digested tissue suspension through a 40 micron mesh. [Taken from Strunz et al. (2020)[171], Conlon et al. (2020)[172]]

Human lungs. Lung tissue was processed as previously described.[173] Briefly, around 1.5g of tissue per sample was manually homogenized into smaller pieces and cleared by washing excessive blood through a 40-lm strainer with ice-cold PBS before tissue digestion. The tissue was transferred into enzyme mix consisting of dispase, collagenase, elastase, and DNase for mild enzymatic digestion for 1h at 37°C while shaking. Enzyme activity was inhibited by adding PBS supplemented with 10% FCS. Dissociated cells in suspension were passed through a 70-lm strainer and pelleted. The cell pellet was resuspended in red blood cell lysis buffer and incubated shortly at room temperature to lyse remaining red blood cells. After incubation, PBS supplemented with 10% FCS was added to the suspension and the cells were pelleted. [Taken from Mayr et al. (2020)[173]]

Peripheral blood. PBMC were isolated from whole blood by gradient density centrifugation (Biocoll) and frozen in FCS + 10% DMSO (Merck) for liquid nitrogen storage. T cells were cultured in RPMI 1640 (Gibco) supplemented with 5% human serum, 0.025% l-glutamine, 0.1% HEPES, 0.001% gentamycin, and 0.002% streptomycin- PBMCs were stimulated with 0.6 nmol of SARS-CoV-2 spike protein-peptide mix (PepTivator SARS-CoV-2 Prot_S, Miltenyi). $CD3^+CD4^+$ and $CD8^+$ T cells were sorted by flow cytometry, centrifuged and the supernatant was carefully removed. Cells were resuspended in the Mastermix + 37.8µl of water before 70µl of the cell suspension were transferred to the chip. After each step, the integrity of the pellet was checked under the microscope. From here on, 10x experiments have been performed according to the manufacturer's protocol (Chromium next GEM Single Cell VDJ V1.1, Rev D). Quality control has been performed with a high sensitivity DNA Kit (Agilent 5067-4626) on a Bioanalyzer 2100 and libraries were quantified with the Qubit dsDNA hs assay kit.

Tracheal aspirates. TAs were digested with 4ml dispase (50 units/ml) (Corning) and 25µl DNase (30µg/ml) (Qiagen) at 37°C for 10 min with occasional shaking. The digestion was then stopped with 10ml of ice-cold 10% FCS/PBS. To obtain single-cell suspensions, the digestion mix was passed through a 70µm cell strainer. Red blood cell lysis was performed only when necessary by incubating the cells with 3ml RBL buffer at RT for 1 min. [Taken from Fischer et al. (2021)[174]]

2.2.5 Single-cell RNA-sequencing

Transcriptomic profiling with scRNA-seq was performed according to the original protocols for Drop-seq[44] and 10x,[45] respectively. Drop-seq was used for the data sets shown in chapter 3.1 (iPS differentiation), chapter 3.2 (bleomycin-treated mice, Munich patient cohort) and chapter 3.3 (smoke exposed mice). 10x Chromium was used for the data set in chapter 3.4 (PB T cells and TA from Munich COVID-19 patients).

Drop-seq The collected cells were taken up in PBS supplemented with 10% FCS, counted using a Neubauer chamber, and critically assessed for single-cell separation and viability. Cells were aliquoted in PBS supplemented with 0.04% of bovine serum albumin and loaded for Drop-seq at a final concentration of 100 cells/ll. Dropseq experiments were performed according to the original protocols.[44,64] Briefly, using the microfluidic device (Nanoshift), single cells (100/µl) were co-encapsulated in droplets with barcoded beads (120/µl, ChemGenes) at rates of 4000µl/h. Droplet emulsions were collected for 10–20 min/each prior to droplet breakage by perfluorooctanol (Sigma-Aldrich). After breakage, beads were harvested and the hybridized mRNA transcripts reverse transcribed. Unused primers were removed by the addition of exonuclease I (New England Biolabs), following which beads were washed, counted, and aliquoted for pre-amplification using a total of either 10 (IPS) or 12 PCR cycles (bleomycin, human ILD, CS-exposed mice). PCR details: Smart PCR primer AAGCAGTGGTATCAACGCAGAGT (100 µM), 2× KAPA HiFi Hotstart Readymix (KAPA Biosystems), cycle conditions: 3 min 95 °C, 4 cycles of 20 s 98 °C, 45 s 65 °C, 3 min 72 °C, followed by 8 cycles of 20 s 98 °C, 20 s 67 °C, 3 min 72 °C, then 5 min at 72 °C). For the CS-exposed mice study, the quality of the single cell transcripts and later the sequencing recovery was improved by subjecting beads to Klenow enzyme treatment, as described for the Seq-Well single cell protocol.[176]
PCR products were pooled sample-wise and purified twice by 0.6x clean-up beads (Clean NA). Prior to tagmentation, cDNA samples were loaded on a DNA High Sensitivity Chip on the 2100 Bioanalyzer (Agilent) to ensure transcript integrity, purity, and amount. For each sample, 1ng of pre-amplified cDNA from an estimated 1000 cells was tagmented by Nextera XT (Illumina) with a custom P5-primer (Integrated DNA Technologies; primer: AATGATACGGCGACCACCGAGATCTACACGCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGT*A*C (10 µM). Single-cell libraries were sequenced in a 100 bp paired-end run on the Illumina HiSeq4000 using 0.2 nM denatured sample and 5% PhiX spike-in. For priming of read 1, 0.5 µM Read1CustSeqB (primer: GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTAC) was used.

10x Chromium. The cells were counted, diluted to 1000 cells/µl, and loaded on the 10x Chromium Next GEM Chip G with a targeted cell recovery of 10,000. The following steps were completed according to the manufacturer's protocol (Chromium Next GEM sc 3' Reagent Kits v3.1). Libraries have been pooled according to their minimal required read counts (35,000 or 50,000 reads/cell for 3' gene expression libraries, 20,000 reads/cell for 5' gene expression libraries, and 5000 reads/cell for TCR libraries). Illumina paired-end sequencing was performed with 150 or 200 (3' gene expression) and 100 cycles (5' gene expression and TCR libraries) on a NovaSeq 6000. [Taken from Fischer et al. (2021)[174]]

2.3 Sequencing transcripts with Illumina platform

Compared to the first sequencing approaches, which sequenced one DNA fragment at a time, next generation sequencing extends this process across millions of fragments in a massively parallel fashion. Currently the Illumina platform remains as the most dominant platform with respect to sequencing of biological molecules.[177]

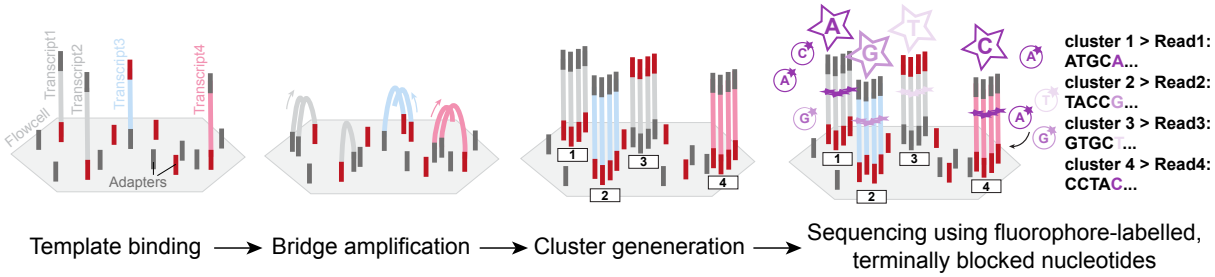


Figure 2.2: Sequencing by synthesis used by Illumina platforms. Fragmented, double-stranded target sequences are loaded onto a flow cell and bind to oligos on the surface which are complementary to the library adapters. One end is fixated on the surface, while the free end can interact with nearby primers, forming a bridge structure. A second strand is produced by PCR. After several cycles, clonal clusters of the initial sequences are generated through bridge amplification. Primer, DNA polymerase and modified nucleotides are supplemented. In each cycle, only one nucleotide will be added to all fragments in one cluster, as nucleotides have additional 3-O-azidomethyl group that prevents further base additions. After washing away unbound bases, the new base for each cluster can be detected by the fluorescent signal. Adapted from Goodwin et al. (2016)[178] and Illumina.[179]

Illumina's HiSeq 4000 and NovaSeq were the instruments used for generating the raw data presented in this thesis. The next generation sequencing workflow of both relies on 4 basic steps (see Fig. 2.2 for details). First, the library is prepared by fragmenting the target molecules to appropriate sizes. As Illumina works with DNA as input, the mRNA is reverse transcribed to cDNA. Second, the library is loaded onto flow cells in which the fragments are fixated on the surface. In the third step each fragment is amplified via PCR and forms a distinct clonal cluster. In a last step, the sequence of each cluster is determined by a process called sequencing by synthesis. In each cycle a fluorescently labelled nucleotide is added. These are modified versions of the natural occurring nucleotides, as their ribose 3-OH group is blocked in order to prevent further elongation by the DNA polymerase. The emissions of each flow cell are recorded and the incorporated base in each cluster is called. The fluorophore, blocking group and unbound molecules are washed away and the next cycles of DNA synthesis for the next position can start. To generate reads of length n, the cycle is repeated n times.[179,178]
The raw outputs from the sequencer are stored in so-called *binary base call* (BCL) files. These are generated in real time and for every cycle the base call and quality information is saved in binary format. For further processing, this format is typically converted to FASTQ, a text file that stores both raw sequence data and quality scores for each detected read:

```
>ReadID
READ SEQUENCE
SEQUENCING QUALITY SCORES
```

The scRNA-seq protocols are sequenced with paired-end sequencing. In the droplet-based methods of choice, two reads per transcripts are generated and gathered by the sequencer.

- **Read1 (barcoded read):** contains cell and molecular barcode used to trace from which transcript in which cell the tagged read originated
- **Read2 (biological read):** contains portion of the transcript's genomic sequence

2.4 Alignment of sequenced reads to reference genome

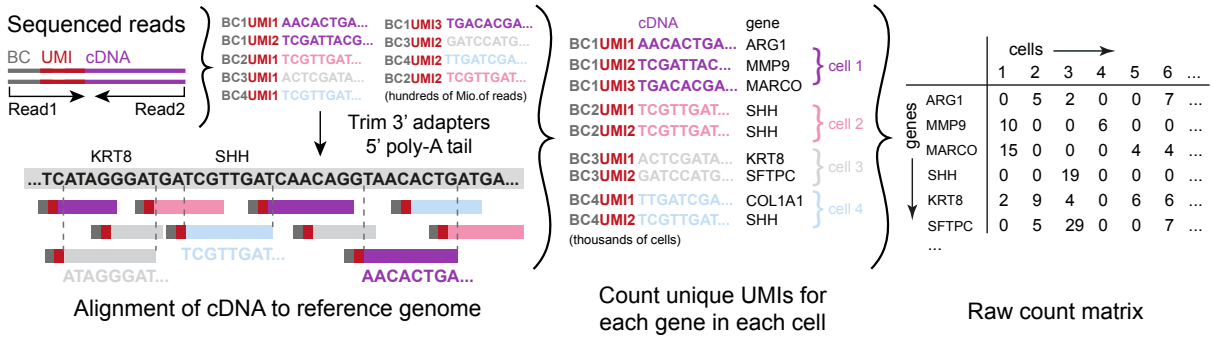


Figure 2.3: Extraction of single-cell transcriptomes from droplet-based platforms. After sequencing and trimming the RNA-seq library, the cDNA sequences (typically 50bp) are aligned to a reference genome in silico, in order to infer the encoded genes. Thanks to the cell barcodes and UMIs attached to each transcript, the number of transcripts for each gene can be counted and linked back to their cells of origin. Finally, the integer numbers of detected transcripts are arranged into *raw count matrices*, each column corresponding to a cell and each row to a gene. Adapted from Macosko et al. (2015).[44]

The FASTQ files are the raw format of the scRNA-seq data that can be accessed computationally. Raw reads can also be stored as a BAM file, which is a more efficient, highly compressed file format. Here, Illumina's `bcl2fastq` and Picard's `FastqToSam` are employed for the conversion to unmapped BAM files. The processing steps that produce the raw count matrices are summarized into pipelines and provided as Drop-seq-tools[180] (Broad Institute, v2.1.0) for Drop-seq experiments, as well as CellRanger[181] for 10x experiments (10x Genomics, v3.1.0). These are largely overlapping, thus the following section will sketch the main procedure shared by the two workflows, instructed by their documentations.

Bases from the cell- and molecular barcode of the barcoded reads are transferred over to the genomic reads and converted to one combined BAM containing single-ended reads. At this point, reads which show low quality are already discarded. Reads are trimmed such that the adapter at the 5' as well as the 3' poly-A tails are removed from the final genomic reads. For Drop-seq, the cell barcode is extracted from bases 1-12 and the molecular barcode from bases 13-20 of the barcode read. For 10x the ranges are slightly different, 1-16 for the cell barcode and 17-26 for the molecular barcode.

To determine from where on the genome the reads originated from, the alignment is done using STAR[182] (Spliced Transcripts Alignment to a Reference, v2.5.2), an aligner

that performs splicing-aware alignment of the short, non-continuous reads to a reference genome. The reference genome is figuratively the photograph on a puzzle box, it is used as template to annotate the transcripts' raw sequences with their encoded genes. The genomes are provided as combination of two input files:

1. **Fasta file:** includes a identifier line beginning with ">unique identifier" followed by the nucleotide sequence of the corresponding gene

2. **Genome annotation file GTF:** includes information about gene structure e.g. annotations, transcripts, exons, start and end position on chromosome

The genomes of frequently used model organisms can be downloaded from main genomic data bases, in this case the Human Genome Build 38 (GRCh38) and Mouse Build 10 (GRCm10). In certain cases custom genomes had to be generated, e.g. for the inclusion of the transcript from the SARS-CoV-2 virus (NCBI Reference Sequence: NC_045512.2) in order to detect those additionally to the human genes.
During the alignment, reads are classified based on whether they are sense, antisense, exonic, intronic or whether their splicing pattern is compatible with transcript annotations associated with that gene. The output from STAR alignment are aligned BAM files, which specify where in the genome the reads mapped to, however the information given by the cell and molecular barcode is lost. These annotations are recovered in a step which merges data from the previous unmapped BAMs with the gene and exon information from the new aligned BAMs. In an attempt to clean the BAM files, the UMI sequences are scanned for sequencing errors and repaired if possible.

To count the gene transcripts digitally, the reads that were confidently mapped are placed into groups that share the same barcode, UMI, and gene annotation. To account for sequencing or PCR errors, the tools allow the UMIs to differ by a single base (Hamming distance of 1) to be merged. Attributed to the cell barcodes, each observed barcode, UMI, gene combination can be counted and returned in form of the count matrices. Each matrix for one sample has the dimension number of cell barcodes x number of genes, and each entry denotes how many molecules were assigned to a particular gene in a cell.

The final steps of the workflow are cell-calling algorithms which scan the count matrices and remove low quality cells. Drop-seq-tools provides a variety of thresholding parameters, such as the number of expected cells in the sample, minimum number of transcripts per cell or even a pre-defined list of barcodes that should be retained. For the runs presented here at least 200 detected genes were required to keep a cell. CellRanger does this in a two-step process. First, a cutoff based on total molecules per cell defines good-quality cells. The cutoff is defined as 0.1 * 99th percentile of top n barcodes (ranked by UMI counts), where n is the number of expected cells. Second, cells which may have lower RNA content due to biological reasons are recovered. The RNA profile of each barcode not called cell in the first step is compared to the background model that represents *empty* cells. Barcodes whose RNA profile strongly disagrees with the background model are added to the set of positive cell calls. The output of both pipelines is given as both full raw and filtered count matrices, the latter include only cells that remained after cell-calling. The filtered versions compose the final input which is then subjected to the pre-processing and analysis methods detailed in the next sections.

2.5 Computational single-cell RNA-seq data analysis

The field of scRNA-seq is relatively immature. Nevertheless, or rather because of its recency, scRNA-seq data analysis is a rapidly moving field and every year the number of methods developed increases drastically.[183] Experimental bulk RNA-seq methods are similar to the single-cell counterparts, likewise some approaches for single-cell data analysis can be adapted from already existing tools. Despite its popularity, there is currently no standardized analysis pipeline available yet.
The workflow maintained throughout this thesis is customly put together based on the standard steps proposed in popular single-cell analysis packages, particularly `Seurat` (R library)[184] and `scanpy` (python package).[185]

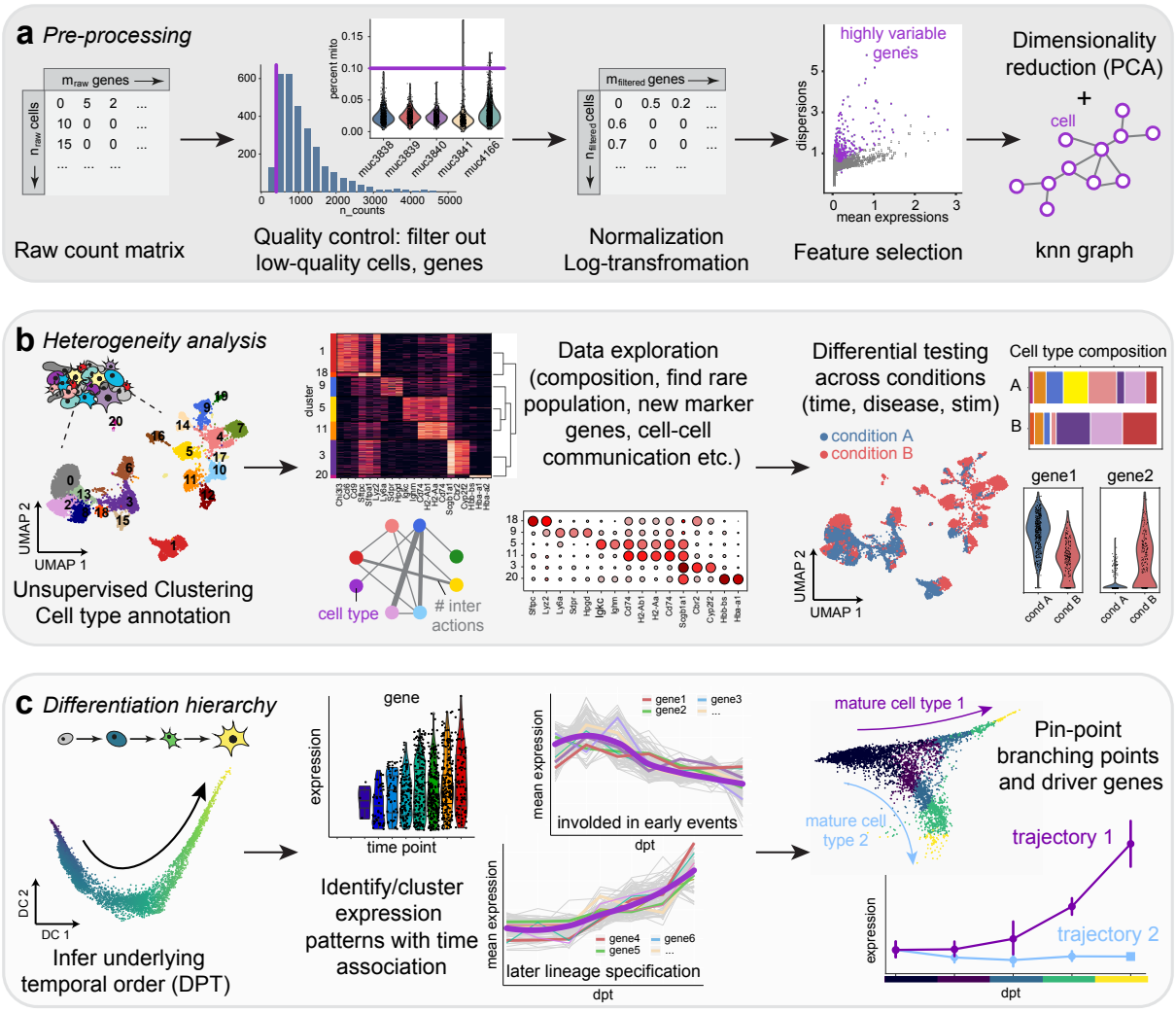


Figure 2.4: Standard workflow of single-cell data analysis and potential applications. **a** Pre-processing steps in order to clean raw input data and reduce dimensionality employed in all analyses. **b** Potential explorations for heterogeneous cell populations. Visualize data set in 2 dimensional map (t-SNE, UMAP), identify cluster (typically correspond to cell types) and eventually implement more detailed analyses as laid out. **c** Potential explorations for delineating differentiation hierarchies. Diffusion maps are commonly used to highlight trajectories. Ordering of cells along DPT enables association of expression patterns to underlying time.

Prior to describing the individual data sets and results, the next section will outline the pre-processing of the raw count matrices that is fundamental to each analysis presented in this work. A graphical depiction of the quality control and potential downstream analyses is given in Fig. 2.4. The individual steps are heavily inspired by vignettes from Seurat, which is currently the leading tool with respect to single-cell data analysis developed by the Satija Lab,[186,187] as well as the best-practice recommendations described in Lücken et al. (2019).[188] The starting point are the sample-wise integer count matrices which are obtained after alignment using CellRanger (10x) or Drop-seq-tools (Drop-seq), with respect to the experimental platform. In the first step the matrices for each sample included in a project are concatenated into one large count matrix and each cell is annotated with the meta data information that is available for the biological sample it originated from, e.g. sample name, health state, time point, clinical information.

2.5.1 Quality control

As the effort that has to be put into cleaning the data set scales with its complexity, the task of establishing a good quality base line is more tedious for single-cell data sets compared to bulk methods. For downstream analysis it is crucial that effects of technical artefacts and low quality features are removed. At the first level, only those cells that are deemed "real" cells are retained. Notably, incomplete cell lysis or failures during library preparation can result in low-quality cells. In a few cases barcodes can mistakenly label multiple cells, a phenomenon which occurs if two cells are encapsulated by one droplet (doublets, or multiplets for multiple cells), or do not assign any cell at all, in case there was no cell captured in the original droplet. At this point a light first cell filtering has occurred already, as the alignment pipelines have inbuilt methods to discard potentially damaged cells. However, the remaining cells should still be manually inspected and filtered based on additional parameters.
The criteria typically used to assess the cell quality are the number of genes expressed, the number of total reads detected and the percentage of expression contributed by mitochondrial genes.[189] To examine the distributions of these criteria, the library sizes of each cell, i.e. its total number of transcripts detected, is visualized via bar and violin plots. Outliers might be low-quality cells or empty droplets, which typically have very few genes and should be excluded. An average value of mitochondria-encoded genes in a cell is estimated at 10% to 20%. Cells with low number of counts but high fraction of mitochondria-encoded genes are indicative of dying cells whose mRNA leaked through broken membranes, retaining only mitochondrial RNA. On the other side of the spectrum, particularly high number of counts and genes may mark doublets, and should be removed as well, as a doublet reflects a hybrid transcriptome which cannot be computationally assigned uniquely to one cell. The algorithm used in `scrublet` is a more elegant solution to detect doublet cells. Based on the raw count matrix it simulates artificial multiplets and builds a nearest neighbour classifier.[190] Additionally to the harsh thresholds, `scrublet` was run on top to retrieve a *predicted doublet score* for each cell.

Corresponding to removing cells with poor quality, genes whose expression level is considered undetectable should be discarded as well. Genes are retained if their corresponding transcripts are detected in a reasonable minimum number of cells, else they are not substantially contributing to the cellular heterogeneity. These characteristics are highly dependent on the experimental workflow and data set specific, therefore the fil-

tering thresholds should be chosen after manual exploration of the data quality. Also, during downstream analyses certain suspicions could arise, e.g. whether an expected cell population is missing or some clusters are formed dominated by relatively low counts etc. It can therefore be beneficial to re-evaluate the chosen cut-offs once a more tangible view of the data is established.

2.5.2 Normalization and log-transformation

The number of counts for each cell is influenced by factors such as cell size, sampling efficiency, bias during PCR amplification etc. Cells that are larger in size are expected to have higher count levels, whereas quiescent cells have temporarily decreased levels. These differences in numbers do not account for biological effects and vary from cell to cell, preventing the direct comparison of transcripts across cells. Normalization of single-cell RNA sequencing data is critical to eliminate such cell-specific biases and obtain a common scale of relative expression abundances instead.[191] Normalization methods that are well-established for bulk RNA are not suited for the single-cell resolution. Particularly the sparsity of the data, i.e. the high amount of zero counts, and the dropout effect, i.e. genes are expressed but not detected, are common pitfalls in these data sets.[192]

One widely used strategy adopted from bulk RNA is global-scaling, which normalizes the expression within each cell by a constant scaling factor. For example, the counts per million (CPM) normalization divides each cell by its total number of reads (library size) and multiplies it by 10^6. This approach is simple, yet not robust to the presence of differentially expressed genes, as a small number of highly expressed genes can dominate the transcriptome.[193] By normalization, either differential effect can be masked, or be falsely induced in remaining genes. This approach further assumes that all differences in library size arise only due to artefacts and discard the biological component, which is not given in most real scRNA-seq data sets. Popular variants of these method are reads per kilobase (RPK), reads per kilobase million (RPKM) or fragments per kilobase million (FPKM). These additionally adjust counts by the length of the respective gene.[186]
The method frequently used in the presented analyses is taken from `scran`.[191] Instead of calculating the size factors on a single-cell basis, it pools multiple cells for a more robust cell-specific size factor estimation. The summation across cells leads to fewer zero entries, circumventing the problem of high amounts of zeroes. These size factors are then deconvolved to yield cell-based factors. This approach allows for more cellular heterogeneity and has shown consistently good performances.[188,192] It should be noted that a normalization step could be performed to correspondingly weight all genes equally for downstream analyses. This is done by scaling gene counts to have zero mean and unit variance. It is still discussed whether the expression level is indicative of the genes importance, therefore in order to rather retain this information, the scaling will be skipped in the workflow.

Many downstream tools operate on normally distributed data, which is typically not the distribution underlying the gene expression. Consequently, the matrices are log transformed after normalization to ensure that the downstream procedures operate on relative, rather than absolute expression differences. As log transformation is not possible for zero values, it is common practice to add 1 as a pseudo-count prior to taking the log.

2.5.3 Feature selection

Single-cell data sets are capable of capturing more than 20 000 genes for human or mouse studies. However, large amounts of genes can be neglected depending on the research question. For instance, genes that are variable across different organs would be rather uninformative when studying only the heart. Apart from these, there would still be many genes that do not exhibit substantial variation across samples.[194] Rather, genes of interest are those that contribute to cellular heterogeneity, or show specific expression patterns only in response to certain stimuli. On the one hand as a mean to purify those biological signals of interest, on the other as necessary step to mitigate computational burden, the count matrices are reduced to only the most informative genes.

There are different approaches for the identification of meaningful genes in the field, here the subset of highly variable genes (hvgs) are considered. Notably, while the hvgs will be the basis for the subsequent computationally more intense dimensionality reduction and visualization stages, the non-informative genes will not be discarded and still available for downstream analyses. The number of genes can vary and is commonly chosen depending on the data set. The hvgs will be identified as described by Satija et al. (2015),[195] for which genes are divided into 20 bins based on their mean expression. Their dispersion is calculated, which is defined as the variance divided by the mean expression of a gene across all cells. As input the normalized and log-transformed matrix is used, although it is also possible to use the raw count data. Within each bin of similar average expression, the dispersion measure is z-normalized and compared to all other genes, returning the bin-wise hvgs. As an additional step to avoid selecting genes that are primarily variable due to expression differences between experimental batches, this selection was modified slightly. The binning is performed on each experimental batch separately prior to the dispersion measure calculation. Only those gene that were marked as highly variable in at least a certain number of samples will be deemed as overall highly variable. This step can be seen as a *light batch correction*, as the data space that will be generated in the next stages will be less dominated by biologically unrepresentative batch genes. Furthermore, to avoid distortion of expression signals due to proliferation state, curated genes associated to cell cycle are also removed from the hvgs list.

2.5.4 Louvain clustering

A crucial step in exploratory analyses is to condense information and group cells based on their similarities. The collective of cells can be viewed as a big network that can be decomposed into sub-units. Such sub-communities are sets of highly interconnected nodes and typically represent cell types, or even cell states that get induced after perturbations. As cell type identification and discovery are the basis for all of the downstream analyses, the community structure needs to be uncovered.
The algorithm of choice in this thesis has been the Louvain algorithm, which optimizes the modularity in a heuristic fashion. Modularity reflects the relative density of edges inside a community with respect to the edges outside. The algorithm starts with optimizing locally on smaller communities, before grouping them into bigger nodes and repeating the procedure.[196] The parameter *resolution* controls the number of resulting clusters. Finally, clusters can be annotated with cell types based on the expression levels of known marker genes, or be declared novel in case they do not resemble any described cell types.

2.5.5 Dimensionality reduction

The technological advances in recent decades have led to the era of *Big Data*, information processing and storing became feasible in increasingly lesser time. More data means more raw material for data exploration, however new challenges were introduced along the way. Single-cell data sets can encompass more than 20 000 genes and the number of cells nowadays easily scales up to several tens of thousands. When computationally analyzing such a wealth of data, one problem that is encountered is referred to as the *curse of dimensionality*. With increasing dimensionality the required volume to contain that data increases drastically.[197]
Fortunately, biological systems show the natural property of being low-dimensional at their core. The cell's machinery and transcription programs are tightly controlled in a spatial and temporal manner, allowing genes to be grouped together with other genes based on their highly correlated expression patterns. Instead of describing the feature space with all genes, it would be sufficient to represent it as a summarization of these modules.[198] Heimberg et al. developed a mathematical framework to evaluate the trade-off between sequencing depth and the amount of biological information that can be extracted reliably. Remarkably, the study found that dominant transcriptional programs are highly noise-tolerant and could still be identified at 1% of conventional read depths.[199]

Down-scaling does not only enable downstream analyses in the first place, but also visualizes the data in a way that lets humans process it intuitively at one glance. As the human eye is known for its abilities to detect visual structures, the high-dimensional data has to be projected into lower dimensional spaces, ideally into 2D or 3D maps that still preserve the underlying biological manifold.
A first minor reduction has already taken place when focusing on the potentially most interesting genes during feature selection. Removing likely uninformative genes further reduces the noise and makes calculations more memory and time efficient. The process of characterizing the data and the relationship between individual data points using fewer features is known as *dimensionality reduction*. There are a number of methods that achieve this, the ones further used in this work are briefly introduced in the next section.

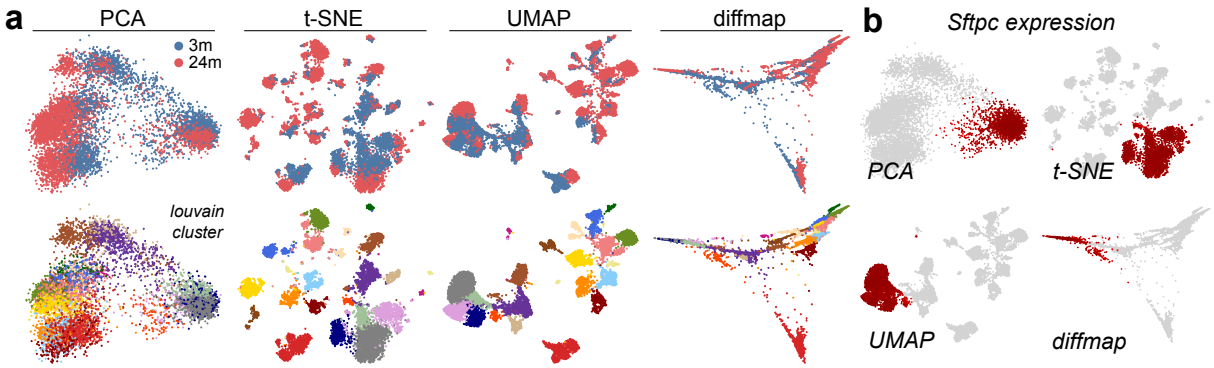


Figure 2.5: Comparison of established visualization techniques. **a** Linear (PCA) and non-linear methods (t-SNE, UMAP, diffmap). The latter are known to preserve global structures, while UMAP tendentially excels in this regard. **b** Sftpc expression demonstrates how finer populations structures of AT2 cells and their global context are captured by the methods.

Principal Component Analysis PCA

The oldest and probably best known linear dimensionality reduction method is principal component analysis (PCA), which was first introduced by Karl Pearson (1901)[200] and developed independently by Hotelling (1933).[201] The central idea is to reduce the dimensionality of the data to a smaller number of components while conserving the variation present as much as possible.
Mathematically, PCA is an orthogonal linear transformation of the data, containing potentially correlated observations, to a new coordinate system. This is achieved by finding a small number of uncorrelated variables, the so-called *principal components* (PC). The PCs are ordered, such that the greatest variance lies along the first component, the second greatest on the second component and so on. The new subspace is specified by the PCs in form of orthogonal vectors, each succeeding component being orthogonal to the previous one.[202]

The covariance matrix C to vector x contains the covariance between the i^{th} and j^{th} element of x. For any $k = 1, 2, ...n$ the k^{th} PC is given by the k^{th} largest eigenvalue of C λ_k and α_k is the corresponding eigenvector. Hence α_1 is the eigenvector corresponding to the largest eigenvalue λ_1 of C. After establishing the first principal component, the next one is calculated by looking for a linear function uncorrelated with $\alpha_1^T x$ while having maximum variance. This is continued such that at the k^{th} step a linear function $\alpha_k^T x$ is found which maximizes variance while being uncorrelated with all previous linear functions $\alpha_1^T, ...alpha_{k-1}^T$. Consider the original data as a matrix, each of the rows representing a sample x_i and each column a certain feature n_j. The transformation is defined as a linear combination of the original parameters. Each row vector of the original data set is then mapped to a new vector via the linear function

$$PCk = \alpha_k^T x = \alpha_{k1}x_1 + \alpha_{k2}x_2 + ... + \alpha_{kn}x_n = \sum_{i=1}^{n} \alpha_{ki}x_i$$

in which $^\mathrm{T}$ denotes the transpose and α the vector of n constants which are called the *coefficients* of the linear transformation ($-1 \leq \alpha_{ki} \leq 1$).[202]

There are as many components as features in the data, but typically the first few components are sufficient for further calculations, as they already account for most of the total variation in the original values. In many cases visualizing the first two components as scatter plot gives a good first overview, e.g. for detection of outliers or sub-structures in the data. PCA has been used in a vast number of single-cell genomics studies and is incorporated into many standardized workflows. The PCs effectively capture dominant biological trends in the data by representing the highly covariant gene modules in form of fewer vectors than the original gene expression space.

However, non-linear dimensionality reductions have gained popularity in the last years, as they are able to avoid overcrowding in the visual representation and appear superior in capturing the underlying clusters.[203] Still, PCA remains a widely-used pre-processing step to summarize the data, and the first n PCs are often propagated to other dimensionality reduction approaches as their initial starting point.

t-Stochastic Neighbour Embedding t-SNE

Linear methods such as PCA focus on maintaining the distance of dissimiliar points in the reduced space. It finds directions of maximal variance and tends to discard variation along other directions, obscuring finer patterns of the population structure.[204] As the gene expression space is inherently low-dimensional as described above, it is typically more essential to preserve proximity of similar data point in the reduced space. One non-linear approach to overcome this limitation would be *t-SNE*, introduced by van der Maaten and Hinton in 2008.[205] The t-SNE approach visualizes the data by mapping each data point to a location in a 2 or 3 dimensional map, aiming to find a representation of those points in the lower dimensional plane that corresponds to the similarity in the original space.

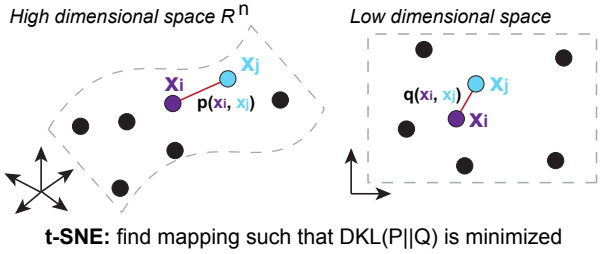


Figure 2.6: Scheme of t-SNE algorithm. Distance between two data points is given by $p(x_i, x_j)$ in high dimension and $q(x_i, x_j)$ in low dimension. Data points in low dimensional space are embedded such that *Kullback Leibler divergence* D_{KL} is minimized.

The basis for this algorithm is *Stochastic Neighbour Embedding* SNE, which converts Euclidean distances to conditional probabilities for the representation of similarities between data points. An important input parameter to set is *perplexity*, which is an estimate of the size of the neighbourhood. It puts a border between the local and global aspects of the data and influences the bandwidth of the Gaussian kernels σ_i = variance of the Gaussian centred on x_i. One should evaluate different values for perplexity, with low perplexities local structures dominate, but for too large values, the algorithm shows a poor performance.[206]

Given data point x_i and x_j, the conditional probability $p_{i,j}$ is "the conditional probability that x_i would pick x_j as its neighbour, if neighbours were picked in proportion to their probability density under a Gaussian centred at x_i". It is given by

$$p_{ij} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/2\sigma_i^2)}$$

The bottom part of the fraction normalizes over all pairs of points involving x_i. Following this, picking a particular pair of points will be proportional to the similarity of the points, i.e. similar points (close together in the high dimensional space) will have a large $p_{i,j}$. A similar conditional probability $q_{i,j}$ can be constructed for the low dimensional counterparts y_i and y_j. In both cases the probability of the event $x_i = x_j$ is set to 0.

However, data sampled from a very high dimensional space cannot be accurately represented in a two dimensional map. As example, in a 10 dimensional plane it is possible for ten data points to mutually have the same distances, whereas there is no way to show this for the 10 point in a two dimensional map. The Student-t distribution circumvents that problem. This distribution is much more heavy tailed than a Gaussian one, allowing moderate distances in the higher dimension to be modelled by much larger

distances in the map. Essentially, points that are far apart from one another would have little effect on the joint probabilities, thus such dissimilar point are modelled far apart in the map. Therefore, when computing the similarity between data points y_i and y_j in the low dimension map, t-SNE uses a Student-t distribution.

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k \neq i}(1 + \|\mathbf{y}_i - \mathbf{y}_k\|^2)^{-1}}$$

For the calculation of p_{ij} in high dimensional space, Gaussian distributions are maintained. The aim of t-SNE is to arrange the n points in a low dimensional space such that the similarities q_{ij} between low dimensional points match p_{ij} as closely as possible. Intuitively, p_{ij} and q_{ij} being equal indicates a successful low dimensional data representation of the higher dimensional data.
The *Kullback Leibler divergence* is a measure for divergence between probability distributions. The points in the low dimensional map should be laid out in a way that minimizes this divergence between the two conditional probabilities $p_{i,j}$ and $q_{i,j}$.

$$C = D_{\mathrm{KL}}(P\|Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

where *P (Q)* denotes a joint conditional probability distribution in the *high (low)* dimensional space. Large p_{ij} modelled by small q_{ij} equals a large penalty in the *Kullback Leibler divergence*. It leads to a big effect on the cost function, as one multiplies by the log of a large value. In contrast, small p_{ij} which are modelled by large q_{ij} lead to a smaller penalty. Essentially, the focus lies on preserving local structures without considering dissimilar data points in an equal amount.[205,207]

Uniform Manifold Approximation and Projection UMAP

Non-linear dimensionality reduction algorithms avoid overcrowding of data points on the lower dimensional representation. Particularly t-SNE remained one of the most used methods in the community. Still, due to the rising sample sizes in contemporary biological data, it became apparent that its computation time and scaling to larger data sets are not optimal. Furthermore, while it is successful in revealing local similarities, it often disregards larger global structures.[203] Recently another non-linear dimensionality reduction technique titled *Uniform Manifold Approximation and Projection* UMAP was proposed by McInnes and colleagues,[208] which is highly competitive with the widely used t-SNE for the visualization of biological heterogeneity. It is growing in popularity due to its superior run time performance and its claim to preserve more of the global topology from the high dimensional feature space.[203]

Mathematically it builds upon Laplacian eigenmaps from Belkin and Niyogi,[209] a technique which presumes that the data in the high-dimensional space is essentially embedded on a low dimensional manifold. Following assumptions are made:[208]

1. There exists a manifold on which the data would be uniformly distributed
2. The underlying manifold of interest is locally connected
3. Preserving the topological structure of this manifold is the primary goal

The underlying structure is represented by a *Riemannian manifold* M, i.e. a space that locally resembles an Euclidean space with well-defined notions of distances, angles, and volumes. Computationally, the *Riemannian manifold* here can be described as a weighted graph, built from information of the k nearest neighbours for each data point (k-nearest-neighbour graph (knn)). The main input parameters for UMAP are the number of nearest neighbours that define neighbourhood boundaries during graph construction and the minimum distance.

The algorithm can be divided into three phases. First, a weighted knn graph is constructed in high dimensions, in which the nearest neighbours are weighted more heavily. This focus on preserving neighbourhood structure rather than absolute distances allows for densely populated regions to be "strechted out", circumventing the overcrowding problem in the lower dimension as noted earlier.[204]

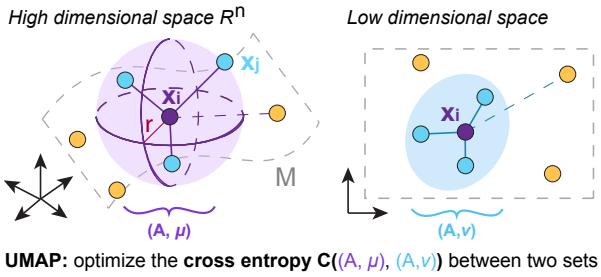


Figure 2.7: Scheme of UMAP algorithm. The input data points from $X = (x_1, x_2, \ldots, x_n)$ are uniformly drawn from a *Riemannian manifold M*, then mapped into R^n. Local topological structures (A, μ) and (A, ν) for k = 3 are constructed and compared using *cross entropy C*.

These topological constructions are based on a measure that approximates the distance of any two data points x_i and $x_j \in X = x_1, x_2, \ldots, x_n$ in the high dimensional manifold. A sphere centered on data point x_j with fixed radius r_i contains its k nearest neighbours. Because the data is uniformly distributed, such a sphere can be centered for any other data point x_j and should have approximately fixed volume and capture its k nearest neighbours. The distance $d(x_i, x_j)$ is then defined such that the value for any data point within this sphere is close to 1, while the values for points outside get exponentially smaller the farther they are away from the local neighbourhood of x_i (see dashed lines in Fig. 2.7). To approximate the manifold, these spheres (or better, local metric spaces) are represented by fuzzy simplicial sets. In this subtype of mathematical sets the membership of elements is not given via yes or no classifications, but instead each element has a certain degree of membership. These are denoted by a reference set A and a membership strength function μ that maps the elements of A to a value between 0 and 1, $\mu : A \mapsto [0, 1]$. Conceptually, each element in a reference set A corresponds to a cell and μ quantifies the neighbourhood relation within A.

In the second phase a low dimensional layout of this graph, which preserves the established neighbourhood structure as much as possible, is computed using the same approach. In a final step the layout of the low dimensional map is optimized, such that it minimizes the error between the two topological representations. For this. The cross-entropy C between the two sets (A, μ) and (A, ν) is optimized using stochastic gradient descent.[208]

$$C = ((A, \mu), (A, \nu)) = \sum_{a \in A} \left(\mu(a) \log \left(\frac{\mu(a)}{\nu(a)} \right) + (1 - \mu(a)) \log \left(\frac{1 - \mu(a)}{1 - \nu(a)} \right) \right)$$

Diffusion Maps and Diffusion Pseudotime

The last dimensionality reduction method used in this thesis are *diffusion maps*. These are non-linear methods and were introduced by Coifman and Lafon (2005).[210] In contrast to the previous methods, diffusion maps emphasize transitions in the data and are preferably used for the study of continuous processes such as differentiation or cell state shifts.[188]

Given by the nature of scRNA-seq capture protocols, cells have to be destroyed in order to retrieve their transcriptomic profiles at a given time point. Consequently, gene expression measurements reflect snapshots of discrete cell states at various stages, rather than a continuous read-out of the cells' gene dynamics over time. Nonetheless, the temporal profile of a single cell is intrinsically encoded in its gene expression, making it possible to reconstruct this information computationally. One method that attempts to order cells based on their underlying cellular time adopted diffusion maps into its algorithm and made this dimensionality reduction method more accessible to the biological field.[211]

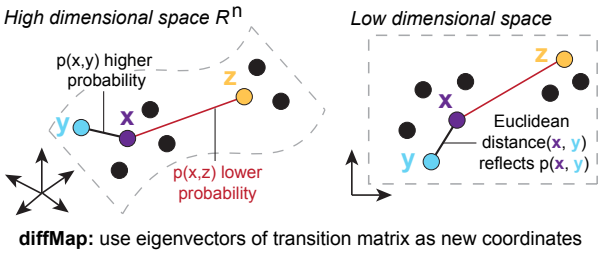


Figure 2.8: Scheme of diffusion maps. Construction of a transition matrix that describes random walks between data points. Transition probabilities p are higher for points that are close (well-connected in the manifold). Embed data such that Euclidean distance approximates diffusion distances.

Briefly, diffusion maps utilize eigenvectors of Markov matrices as coordinates for their embedding in a low dimensional Euclidean space. A measure of connectivity is defined between two data points x and y as the probability to jump from x to y in one step of a random walk over the knn graph of the data. Intuitively, the connectivity is higher for points that are nearby than for those far apart. A Gaussian kernel $k(x, y)$ is used to specify these probability in terms of a likelihood function.

$$connectivity(x, y) = p(x, y) = \frac{k(x, y)}{\alpha} \propto k(x, y)$$

with α being a normalization factor for sampling density. This kernel has to be symmetric $k(x, y) = k(y, x)$ and also preserve positivity $k(x, y) \geq 0$. Inside a given neighbourhood, this kernel is assumed to be an accurate measure of similarity, whereas outside the measure is less reliable and quickly goes to zero. Based on this, a row-normalized diffusion matrix P is constructed, where each entry gives the connectivity between two data points p_{xy}.

$$Diffusion\ Matrix\ P = \begin{pmatrix} p(x,x) & p(x,y) \\ p(y,x) & p(y,y) \end{pmatrix} = \begin{pmatrix} p_{xx} & p_{xy} \\ p_{yx} & p_{yy} \end{pmatrix}$$

To consider more than one step, the power of the diffusion matrix is taken. This can be clarified with following construction, let t be 2 in a 2 x 2 diffusion matrix P.

$$Diffusion\ Matrix\ P^2 = \begin{pmatrix} p_{xx}p_{xx} + p_{xy}p_{yx} & p_{xy}p_{yy} + p_{xx}p_{xy} \\ p_{yx}p_{xy} + p_{yy}p_{yx} & p_{yy}p_{yy} + p_{yx}p_{xy} \end{pmatrix} = \begin{pmatrix} p_2(x,x) & p_2(x,y) \\ p_2(y,x) & p_2(y,y) \end{pmatrix}$$

$p_2(x,x)$, which is the probability to start at x and end up in x within two steps, is $p_{xx}p_{xx} + p_{xy}p_{yx}$. This encompasses the two probabilities to simply stay at x and to move to y and back to x again.
During the diffusion process the probabilities of P^t are calculated for larger values of t. Paths that do follow the underlying structure of the data have a higher probability. If the points are dense, i.e. highly connected along the geometric structure, the paths consist of short high probability steps. Whereas paths that do not follow it have longer jumps, lowering the paths overall probability.[197]

By separating highly likely paths from others, this process has an important role in revealing structure and reducing noise in the data. The similarity of two points x and y is established by summing up the probabilities of all possible paths of length t connecting them. It is given as *diffusion distance* $D_t(x,y)$

$$D_t(x,y)^2 = \sum_{z \in X} |p_t(x,z) - p_t(y,z)|^2 = \sum_{z \in X} |P^t_{xz} - P^t_{zy}|^2$$

where z is any other data point in the data set. This metric is in accordance with the concept of clusters: For the diffusion distance $D_t(x,y)$ to be small, $p_t(x,z)$ and $p_t(y,z)$ have to be roughly equal. This is achieved if there are many high probability paths of length t between x and y. In essence, if x and y are close and in a highly connected sub-part of the data, they are well-connected via z. Since the geometry information is accumulated and propagated, D_t captures the similarity of points along the underlying structure.

Coifman and Lafon showed that this metric can be calculated based on the eigenvectors ψ_k and corresponding eigenvalues λ_k up to a certain accuracy δ[212]

$$D_t(x,y) = \left(\sum_{l=1}^{s(\delta,t)} \lambda_l^{2t} (\psi_l(x) - \psi_l(y)) \right)^{\frac{1}{2}}$$

where $s(\delta,t)$ is the maximal value for l such that $|\lambda_l|^t > \delta |\lambda_l|^t$. Finally, dimensionality reduction is performed by mapping the data points into a new lower dimensional Euclidean space (*diffusion space*) formed by the eigenvectors. This embedding should be such that the Euclidean distance between data points is equal to the diffusion distance, ensuring that distance in the new space reflects the relationship between data points in terms of their connectivity.

$$Diffusion\ Map\ \Psi_t(x) = (\lambda_1^t \psi_1(x), \lambda_2^t \psi_2(x), \ldots, \lambda^t_{s(\delta,t)} \psi_{s(\delta,t)}(x))$$

where ψ_k is the k^{th} eigenvector of the normalized matrix P and λ_k is the associated eigenvalue, indicating its importance. The components of $\Psi_t(x)$ are called *diffusion coordinates* (DC). The reduction is done by retaining those DCs associated with the dominant eigenvectors, as these approximate diffusion distance best.
In summary, by iterating the Markov transition matrix, i.e. running the random walk forward, *diffusion coordinates* are constructed. Embedding data points along these coordinates allows to represent the data's intrinsic geometry in a lower dimensional space.

Diffusion Pseudotime dpt

With respect to scRNA-seq, the data set is given as a matrix of the dimension cells x genes. The transition matrix can be interpreted as the cells' probabilities to differentiate towards potential fates. The pseudo-temporal ordering of cells along such an hypothetical trajectory can be established by measuring cell similarity in gene expression, assuming that developmental processes dominate their transcriptomes. One method accomplishes that by directly building upon the concept of diffusion maps and estimating a *diffusion pseudotime* (dpt). For that case, the diffusion map algorithm applied in this thesis includes slight alterations from the original approach as proposed in Haghverdi et al. (2016). The major change would be that instead of using a fixed global Gaussian kernel width, here instead a local Gaussian kernel width for each cell is estimated, based on a cell's distance to its k nearest neighbours. Effectively, a weighted knn graph of the data is constructed.[211]

For the dpt calculation a modified version of the transition matrix T is used. M is defined as T without the first eigenspace. As the first eigenvalue λ_0 with corresponding eigenvector ψ_0 is associated with the steady state, this modification excludes the possibility to stay in the current state. M shares the same eigenvectors with T except for ψ_0. Further, the connectivity measure *diffusion distance* introduced by Coifman et al. relies on the input parameter t, i.e. the fixed length of the random walks. In contrast to the original metric, dpt is independent of t as the transitioning probabilities for any two cells are computed by summing over all random walk of arbitrary lengths.

$$M = \sum_{t=1}^{\infty} \tilde{T}^t = \sum_{t=1}^{\infty} (T \; - \; \psi_0 \psi_0^T)^t$$

Given a predefined root cell, the accumulated transition matrix M is calculated. The transition probabilities for a cell x are stored in a vector $M(x, .)$, represented as a row of M. Finally, the *diffusion pseudotime* for two cells x and y is calculated as the Euclidian distance between their two vectors.

$$dpt(x, y)^2 = ||M(x, .) - M(y, .)||^2 = \sum_z (M(x, z) - M(y, z))^2$$

$$dpt(x, y) = \left(\sum_{i=1}^{n-1} \left(\frac{\lambda_i}{1 - \lambda_i} \right) \; - (\psi_i(x) \; - \; \psi_i(y)) \right)^{\frac{1}{2}}$$

2.5.6 K-nearest neighbour graph knn

Complex systems can be represented as networks $G = (V, E)$, in which the fundamental compartments form the vertices V, and a measure of similarity the corresponding edges E. Compartments share more edges and a higher density of internal links, whereas links to other compartments occur more sparsely.[213] In the context of biological tissues, the cells represent the elementary units (nodes) and the edges connect cells with similar transcriptomic profile. Identifying the different compartments present in the data enables to assess cell populations of different sizes and densities. Therefore, after dimensionality reduction the cells are further abstracted into a k-nearest neighbour (knn) graph, in which cells will be connected to its k nearest neighbours.

Partition-based graph abstraction PAGA

PAGA[214] is based on the knn graph. Instead of presenting the graph at single-cell resolution, the PAGA graph shows the connectivity structure of the data at a more coarse level. Partitions are typically the Louvain cluster at an appropriate resolution, as these correspond to cell types in the data, but can be experimentally validated labels as well.

The task of connecting different cell types by their underlying biological mechanisms is commonly not achieved by looking at isolated paths of single cells. With the partitioning approach, PAGA follows a group of similar cells that pass through several partitions, achieving a higher confidence level. The connectivity of an edge between two partitions can be quantified based on a statistical model comparing the real number of inter-edges to the number of inter-edges expected under random assignment. This connection strength reflects the confidence in the presence of an actual connection and can be used to discard noise-related, spurious edges. Finally, a denoised lower dimensional topology with connected and disconnected regions is obtained, preserving the underlying topology of the data and enhancing its interpretability. PAGA was part of a recent trajectory inference comparison study[215] which evaluated 45 tools on the following criteria: prediction accuracy, scalability, usability and prediction robustness. PAGA was one of the few methods next to `Slingshot`[216] and `SCORPIUS`,[217] performing well across the board.

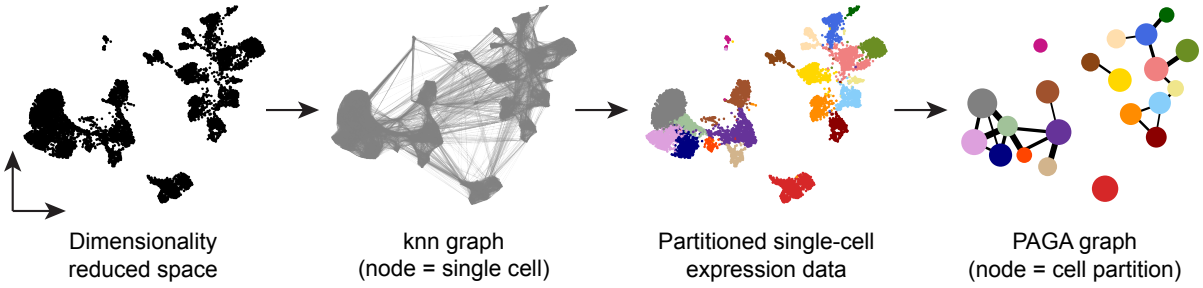


Figure 2.9: Outline of the PAGA algorithm. Knn graph together with coarse clustering of cells build the input for this method. The edges between any two partitions are weighted by a statistical measure of connectivity. After excluding low-weight edges, underlying coarse topology is revealed in form of a 2D map. Adapted from Wolf et al. (2019).[214]

2.5.7 Batch correction using BBKNN

Through global efforts such as the Human Cell Atlas, the generation and sharing of large collections of scRNA-seq data sets has been accelerated. Nowadays the atlas-level studies often do not only include samples that span conditions and spatial locations, but also data generated from multiple labs or across multiple technologies.
While this is a great way of gaining statistical robustness and finding effects that might have been overlooked in individual data sets, unwanted side-effects can be introduced. Batch effects arise when biologically equivalent samples are handled in different experimental runs. Slight variations in the procedure can already introduce variation that does not reflect the underlying biology, such as different individuals handling the material or external factors (e.g. temperature, unknown sources of contamination). Naturally, this hampers meaningful interpretation of the data.

Biological variation, e.g. inherent to cell types/states or induced after perturbation, should be kept separate from variation that is not of interest in the scope of a study, e.g. naturally given between individuals or technical artefacts. Ideally experiments should be designed in a way in which the distinction of these types is easily possible. As this is not always the case, batch effects should be accounted for in downstream analyses while retaining biological variation. Many common approaches assume that all data sets share at least one cell type or overall exhibit the same expression structures across all data sets. With such strong assumptions these methods are prone to over-correction, especially when the data sets have considerable differences in cellular composition. Much like the choice of experimental technology for a study, the method for batch correction should be guided by the specific goals.

A recent large-scale benchmarking study[218] evaluated the performances of several available integration methods, determining `BBKNN`,[219] `Scanorama`[220] and `scVI`[221] as best performers, particularly on complex integration tasks.
Batch variations can distort the knn graph, leading to major connections only in between rather than across batches. This disconnected structure is not representative for samples with similar cell type populations. `BBKNN's` core assumption is that cells of each cell type are present in all of the user-defined groups and any cell type variation across batches is only due to technical artefacts. Thus the algorithm motivates inter-batch edges and constructs a batch-balanced graph by identifying a cell's nearest neighbours in each user-defined group independently. Similar cell types across batches will be grouped together while unrelated cell types stay unconnected.[219] However, it should be noted that this approach will mask rare populations that are only present in some batches. Owing to the favourable run time and with the main objective of avoiding over-correction in mind, this graph-based methods was incorporated. Particularly because it returns a batch-balanced knn graph and does not alter the underlying gene expression values, batch effects will be reduced via `BBKNN` in the following analyses. The corrected neighbourhood graph serves as the basis for the downstream methods whenever appropriate.

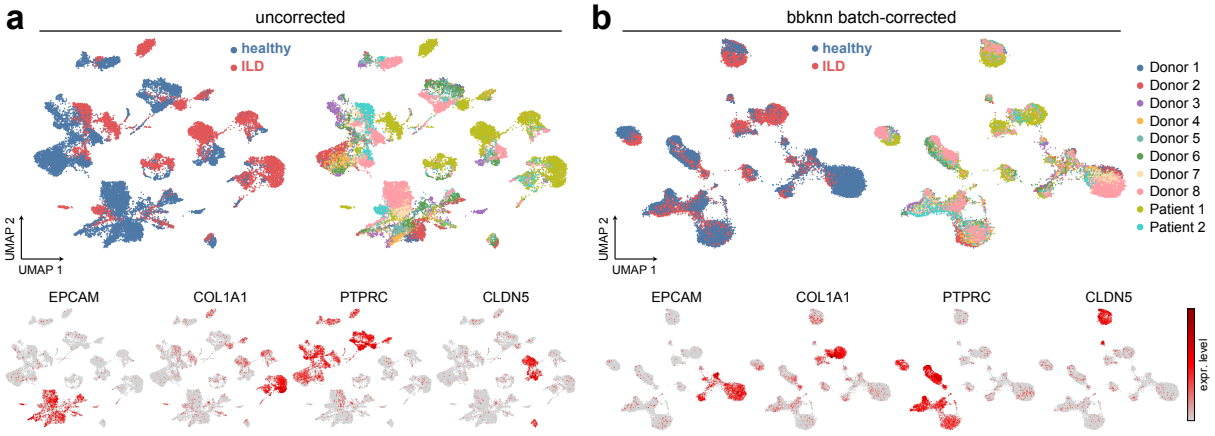


Figure 2.10: UMAP using uncorrected vs. batch-corrected knn graph. a Uncorrected knn graph results in patient-specific cluster, well-recognizable in EpCAM$^+$ epithelium.
b After `BBKNN` correction on patients, biologically meaningful variation appears conserved in the embedding. Cells of individuals visually overlay while disease-specific shifts are still discernible.

2.5.8 Ambient gene correction using SoupX

Droplet-based RNA capture methods enable rapid processing of a large number of cells in parallel. During that process however certain phenomena can arise that distort the interpretation of data. Next to doublets or empty cells, non-endogenous mRNA transcripts from damaged droplets can be present in the input solution. Such ambient background contamination with cell-free RNA (soup) is present in even high-quality data sets and will muddle the profiles of each cell.
This confounding is non-negligible and should ideally be removed before computational analysis. The R package `SoupX`[222] demonstrates a reasonable approach to quantify and purify ambient contamination and was therefore incorporated into the workflow.

Briefly, `SoupX` attempts to recover the true molecular abundance of the genes in each cell by estimating the ambient mRNA expression profile. This approach is based on cells that are considered empty, i.e. droplets with less UMIs than a defined threshold detected and assumes that background contamination does not differ between cells.
The fraction of UMIs originating from the soup can be identified in an automated approach, in which the markers of each cell cluster are identified first. For the estimation of the contamination fraction, the true number of count for these is assumed to be 0 in clusters in which the gene is not a marker. This is repeated across all strong marker genes and provides a set of estimates, from which the most common value will be chosen as the final estimate of the contamination fraction. The additional step of looking at the count distributions from a cluster perspective helps to overcome the sparsity of scRNA-seq data. Alternatively, the contamination fraction can also be set manually. In the following analyses, this parameter was always set to 0.3 as it reflects a reasonable contamination fraction for the data sets presented.
In a final step the expression of each cell is modified using both the ambient mRNA expression profile and contamination fraction, producing a corrected matrix of counts, which was used in place of the original count matrix in downstream analyses.

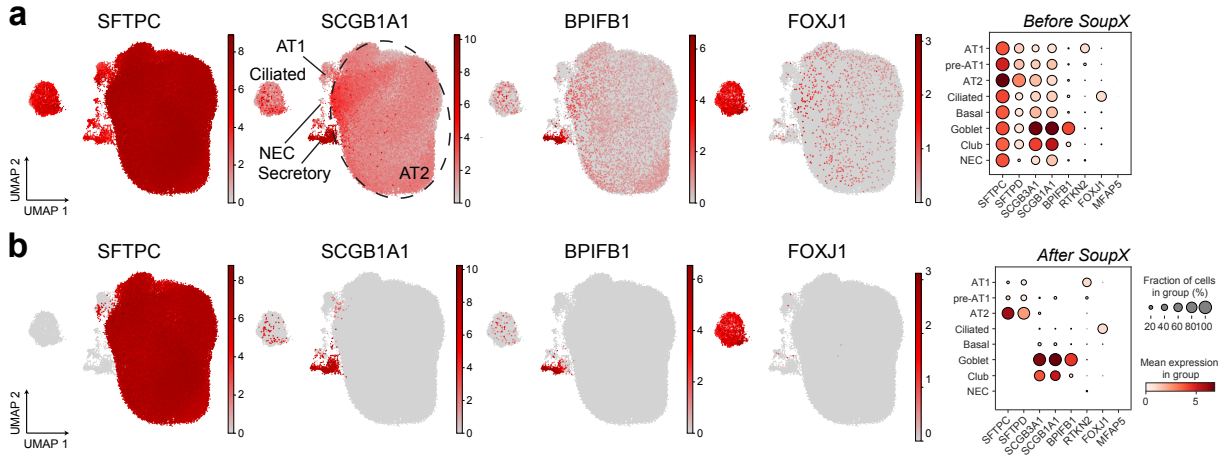


Figure 2.11: Comparison of gene expression before and after ambient count removal. **a** Gene expression of common ambient genes before correction on scRNA-seq data of human lung epithelium, cell types as indicated. **b** After SoupX correction, ambient counts are removed in cell types, in which corresponding genes are not expected to be expressed.

2.5.9 Differential gene expression analysis

Once the data set contains only those cells with sufficient quality and is available in a clean version, the foundation for downstream analysis to extract and describe the underlying biological mechanism is set.

A first step is typically to put biological meaning into the previously detected subcommunities using prior knowledge. The clusters group cells that are highly similar to each other and can be annotated with cell types based on the expression levels of known marker genes, or be declared novel in case they do not resemble any described cell types or states at that time. The marker genes are identified by the detection of differentially expressed genes, i.e. focus on genes that are higher expressed in a given cluster compared to all other clusters. As marker genes are driving the cluster separation, a strong difference in expression between these groups is expected, which can be assessed and ranked with simple statistical tests such as Wilcoxon rank-sum test or the t-test.[188] There are methods available to automatically annotate the cell types based on the top marker genes, using overlaps to known signatures or other enrichment approaches. However, such approaches are restricted to cell types present in the reference data bases, and manual curation might be needed to refine the labels based on data-driven exploration.
After the rough structuring of the data into a biologically interpretable framework, the exploration can be focused on relevant compartments with most interesting patterns. For instance, many research questions are about uncovering significant changes across conditions, most commonly in healthy controls compared to diseased individuals. One significant effort in this area is the detection of differentially expressed genes on the single-cell scale. This analysis employs a group of statistical tests to establish whether there exists a significant variation across a set of tested conditions for each gene, ultimately uncovering driver genes or even full cellular programs that play important parts in pathogenesis.

One well-known pit-fall in scRNA-seq data would be the lower capture efficiency compared to bulk measurements, because of which many transcripts tend to be missed during the reverse transcription. As a result, some transcripts are highly expressed in one cell but are missed in another, a phenomenon known as a drop-out event. Therefore, in comparison to bulk RNA-seq data, scRNA-seq data sets are inherently much more heterogeneous and exhibit large portions of zero counts, which requires appropriate models to handle such sparse, complex input.[223] One of the first differential expression models for single cell data was Single-cell differential expression (`SCDE`),[224] which models the expression as a mixture of two probabilistic processes. The first process models the rate at which the transcript is amplified and detected at a level correlating with its abundance (with a negative binomial distribution) and the second considers the drop-out events (with a low-magnitude Poisson process). This approach was advanced to a hurdle model in the Model-based Analysis of Single-cell Transcriptomics (`MAST`),[225] based on a two-part generalized linear model. Here, the mean expression follows a normal distribution while the dropout component follows a binomial model.
`Monocle2` is another popular tool that converts the relative expression in single cells to *consensus counts*, a measure of relative transcript counts that aims to eliminate parts of the technical variability in single cell experiments. These *consensus counts* are then easier to model with standard regression techniques compared to the conventional normalized transcript counts.[226]

Throughout this work, differential gene expression analysis will be performed with `diffxpy`,[227] a python package that covers a wide range of differential expression analysis scenarios encountered in scRNA-seq and integrates easily into `scanpy` workflows. In the most basic application, the model is applied to test for the difference between two groups. As the Wald test allows for adaptive assumptions on the noise model and for the testing of more complex effect, e.g. account for different study cohorts or include potential confounding factors in the model, it will be the preferred method.

Model differentially expressed genes across time-series

Since many biological systems are dynamic in their character, the longitudinal sampling of their gene expression can provide clearer insights about how expression levels evolve in time and which particular genes are driving a biological process. Time-series expression data is a valuable source of information in order to understand the unfolding of a biological process in response to perturbations, and is heavily featured in this work.

Methods for differential gene expression rely on static expression states. However, due to differences in sampling rates and time variations in biological data in general, such methods cannot be applied directly to time-series expression data. One promising approach to obtain a continuous time formulation is by the use of cubic splines, i. e. a set of piece-wise cubic polynomials to represent gene expression curves.[228] During spline interpolation, low-degree polynomials are fitted to small subsets of the values instead of fitting a single, high-degree polynomial. This allows for increased control given a set of regulation points (knots) and generally reduces the interpolation error compared to polynomial interpolation. It is possible to include time as a continuous covariates in the model generation with `diffxpy`, which uses such a spline basis space to represent the smooth expression trends.

2.5.10 Gene set enrichment analysis GSEA

The meaningful interpretation of the gene lists with condition-specific behaviour is the next challenge. The genes can be ordered in a ranked list, e.g. based on the amount of difference across the conditions (log fold change). Instead of focusing on a handful of genes at the top and bottom of these lists, it is essential to find a way to group genes that share common biological function or regulation patterns. The Gene Ontology (GO) consortium has the goal to produce a structured, precisely defined and common vocabulary to describe the roles of genes and gene products in any organism. With this a comprehensive and complete source of information on the functions of genes, their cellular localization and biological processes should be established.[229]
Gene Set Enrichment Analysis (GSEA) builds on this resource among others. To reveal their biological relevance, it tests whether certain gene ontology terms are over-represented in a list of genes. In the following chapters, GSEA will be performed with the python package `GOAtools`,[230] which uses the latest annotations and organizes results using a novel GOATOOLS GO grouping method. It is based on the assessment, whether selected gene sets contain over- or under-representation of certain functional classes, by comparing the frequency of genes for a particular GO term in the sample to the frequency in the background. A p-value is then computed, on the basis of Fisher's exact test and adjusted with Bonferroni correction.

2.5.11 Inference of intercellular communication

Finally, to take a step back from the single-cell resolution, it is important to put the discovered expression patterns into a broader context. Cells react to extracellular signals produced by neighbouring cells in their microenvironment and complex signalling cascades are initiated upon the binding of ligands to their cognate receptors. To understand which processes occur in response to certain stimuli, it is essential to understand the cross-talk between cells within their respective niches.

Thanks to the measurement of ligands and target receptors across a magnitude of interacting cell types, it is now possible to decode such intercellular communication networks. Whole data bases have been established to chart ligand–receptor relationships. This not only sheds light on tissue function, but also allows to detect their alterations in disease. `CellPhoneDB`[231] for instance tries to identify biologically relevant interacting pairs based on a manually curated data base. The algorithm considers the expression levels of ligands and receptors within each cell state and performs permutation tests to predict which molecular interactions show significant cell-state specificity. `CellPhoneDB` emphasizes specificity of the ligand–receptor interaction, arguing that some pairs might be ubiquitously expressed by the cells in a tissue and therefore will not be informative regarding communication between particular cell states.

Another recent tool is `NicheNet`,[232] that - much like `CellPhoneDB` - integrates prior knowledge on ligand-to-target signalling paths. Contrary to similar approaches, it goes beyond describing extracellular signals by which cells are capable of communicating. `NicheNet` incorporates a predefined target gene list and infers the effects of sender-cell ligands on receiver-cell expression. The basis are its weighted networks, data sources which are integrated and optimized such that the most informative sources contribute more to the final model. The regulatory potential is introduced as a quantitative measure that is calculated between all pairs of ligands and target genes, denoting how strongly existing knowledge supports that a given ligand may regulate the expression of a target gene. This value will be high if the regulators of the respective gene are downstream of the signalling network of the corresponding ligand. The signal from a ligand is propagated further starting from its target receptors, factoring in signalling proteins and transcription factors that are involved in the pathways leading up to the genes at the tail of the cascades.
For the calculation of ligand activity scores, two gene sets are defined: (1) potentially active ligands in sender cells and (2) a set of affected genes of interest in the receiving cells. To quantify the ability of each ligand to predict the target genes, the authors chose the Pearson correlation coefficient. The correlation is calculated between regulatory potential scores of each ligand and the target indicator vector, which reflects whether a gene belongs to the gene set of interest or not. Eventually, ligands are ranked according to how well their prior target gene predictions correspond to the observed gene expression changes resulting from communication with sender cells.

This approach is more advanced than simply matching against a ligand-receptor data base, and proved to be useful when trying to explain shifts in gene expression, making `NicheNet` especially applicable to trace potential ligands that explain induction of condition-specific gene expression.

Chapter 3

Results

3.1 Differentiation trajectory of human pluripotent stem cells to lung and hepatocyte progenitors

The first chapter of this thesis revolves around a project exploring early lung development at a high temporal resolution with Drop-seq data. Due to the dense time points this data set has been ideal for the establishment of methods that will be used throughout this work. As understanding the development of human respiratory tissues is crucial for modelling and treating lung disorders, much prior knowledge has accumulated over the years and enabled the validation of the employed analysis angles and results. The work has been summarized in a manuscript. On 24th February 2021, a not yet peer-reviewed version has been uploaded to bioRxiv, a free online platform preprints in biological research. The file is accessible via the link `https://www.biorxiv.org/content/10.1101/2021.02.23.432413v2` or the doi `10.1101/2021.02.23.432413`.
Experimental data planning and collection such as organoid cultures, scRNA-seq, FACS-sorting, and respective statistical analysis were performed by Chaido Ori and Ilias Angelidis. In this project, my contribution encompassed the single-cell data analysis, including sequence alignment, pre-processing, differential gene expression and trajectory modelling. Thus, parts of the results in this chapter have been used in the preprint.

Introduction

One of the main questions in stem cell biology is how the cascades of gene regulatory networks and signalling pathways give rise to the structure and function of organs. Resolving these mechanisms enables researchers to characterize adult stem cells and treat disorders, for example by manipulating their behaviour for therapeutic purposes in humans.

Particularly the foregut endoderm (FE) formation is of interest, as the FE gives rise to several organs, including the lung and liver. To better understand the aspects of human development, appropriate experimental models are necessary. Induced pluripotent stem cells (iPS) have proven to be a powerful tool to study molecular details of the differentiation from definite endoderm (DE) progenitors over FE up to early lung progenitors (ELP). IPS cells are derived from adult somatic cells and can be reprogrammed into a pluripotent state. By maintaining embryonic stem cell conditions, their stem cell morphology and growth properties as well as expression of stem cell marker genes can be maintained.[233] Such iPS cells hold promise not only in basic research but also regenerative medicine, as they can propagate indefinitely and give rise to every other cell type in the body.[234]

In the following analysis, the formation of lung foregut precursors in human development was investigated by the means of such iPS. The combination of the experimentally directed differentiation with scRNA-seq provides a strong basis to compose a roadmap governing human development in vitro. Here, the transcriptomic profiles were assessed via Drop-seq at a high temporal resolution along the differentiation of human iPS cells. An established differentiation protocol was optimized and monitored on the tissue and the transcriptomic level. Briefly, step-wise activation and inhibition of Activin/Nodal and Wnt/β-catenin signalling leads to appearance of a NKX2-1$^+$ endodermal population at day 8.[235] The expression of this transcription factor specifies definitive endoderm to respiratory endoderm commitment and is the earliest indication of the establishment of both respiratory progenitors and thyroid epithelium in the FE. NKX2-1$^+$ cells are highly enriched for undifferentiated progenitors, which are competent at expressing a variety of pulmonary specific genes, including surfactant proteins (SP-A, SP-B, SP-C, CC-10).[236]

Studies have shown that the transcription factor Sonic Hedgehog (SHH) and the Fibroblast Growth Factor 10 (FGF10) give cues essential in foregut and lung development. Mice with a targeted deletion of Shh have foregut defects as early as embryonic day 9.5, showing anomalies similar to those observed in humans with foregut defects. Notably, their lung mesenchyme displays enhanced cell death and down-regulation of Shh target genes.[237] The distal mesenchyme regulates the growth and branching of the endoderm, Fgf10 for example is expressed in the mesenchyme adjacent to the distal buds.[238] Therefore, SHH and FGF10 were supplemented at the second stage of the protocol. Simultaneously, the transcriptional states underlying the differentiation were charted by daily single-cell suspension processing and Drop-seq. This 16 day time-series scRNA-seq experiment resulted in a total of 10,667 cells used later in downstream analyses.

3.1.1 IPS differentiate towards lung and hepatocyte progenitors in parallel

Prior to data analysis, the validity of the adapted differentiation protocol had to be assessed. The appearance of lung progenitors was tracked via a human iPS cell line integrated with the eGFP downstream to the endogenous NKX2-1 promoter. Starting between days 13 to 15, the formation of eGFP$^+$ progenitors became apparent and their numbers were significantly increased by supplementation of SHH and FGF10, in both tested types of basal media BM1 = DMEM/F12 and BM2 = IMDM (Fig. 3.1c).
To evaluate the developmental potential of the appearing eGFP$^+$ lung progenitors, spheroid 3D culture assays were used. Clusters of eGFP$^+$ cells from day 15 of the differentiated cells were picked and embedded in matrigel with supplementary media in order to promote the proliferation of lung progenitors in suspension culture. This led to the outgrowth of the spherical structures, which tripled in size within 7 days and maintained expression of eGFP. Further treatment by dexamethasone, cAMP and IBMX, which promote maturation of the fetal lung,[239] induced expression of proteins characteristic for club and goblet cells in the proximal region of the lung (SCGB1A1, MUC5AC) referred to as DCI. CHIR addition lead to Wnt/β-catenin pathway activation, promoting branch development in the spheroids, while inhibition of TGF-β lead to the spheroids growing substantially larger to an average size of 1.6 mm by day 35 (DCICS). They also exhibited branches and markers of AT2 cells (SFTPB, SFTPC) and lower expression of SCGB1A1 and MUC5AC compared to DCI (Fig. 3.1b,d). These results confirmed that the selected cells were indeed capable of generating lung progenitors.

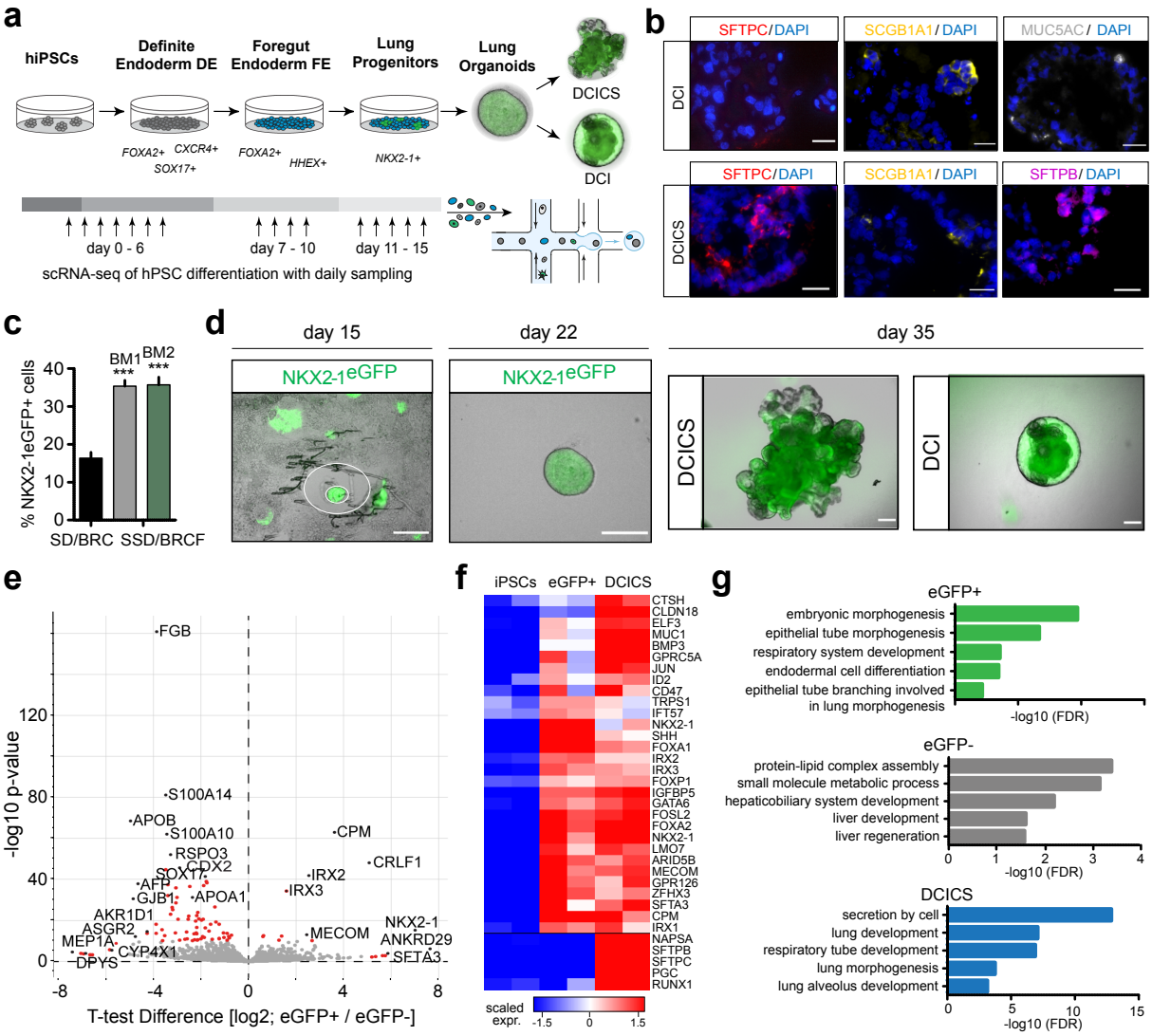


Figure 3.1: Differentiation of lung progenitors from human induced pluripotent stem cells. **a** Schematic illustration of daily sampling for scRNA-seq on to the first 16 days of the lung progenitor differentiation protocol to produce eGFP$^+$ lung progenitor cells, spheroids and organoids. **b** Immunofluorescence staining of day 35 DCI and DCICS spheroid-organoids for SFTPC, SFTPB, SCGB1A1 and MUC5AC (scale bar = 20µm). **c** Quantification of the amount of eGFP$^+$ cells by flow cytometry on day 15 of differentiation, comparing condition with and without FGF10+SHH treatment in defined basal media DMEM/F12 (FE-BM1) and IMDM (BM2). Bars represent mean ± SD, n = 3 biological replicates, ***p $\leq$ 0.0001 unpaired two-tailed t-test. **d** Fluorescence microscopy of day 15 NKX2-1$^+$/eGFP$^+$ cell sectors and a representative day 22 spheroid produced by embedding a colony in matrigel, and the further growth of spheroids (Scale bar = 200µm) observed on day 35 with treatments. **e** Volcano plot showing differentially expressed genes comparing the eGFP$^+$ and eGFP$^-$ sorted populations on day 15, and the corresponding GO terms of the eGFP$^-$ population. **f** Heatmap displaying the expression of developmental markers and genes important for lung function based on the bulk RNA-seq analysis of the indicated conditions (p-value $<$ 0.05, scale displays normalized log2 expression). **g** Gene Ontology term enrichment analysis of differentially expressed genes in eGFP$^+$ and DCICS spheroid-organoids relative to undifferentiated NKX2-1 cells (p-value $<$ 0.05, GO term FDR $<$ 0.05).

The mRNAs of undifferentiated cells was sequenced in a bulk fashion, in order to characterize the differences across the populations at day 15. Prior to sequencing, the cells were sorted into DCICS organoids, eGFP$^+$ and eGFP$^-$ cells. Expression analysis revealed FGB up-regulation was the main feature differentiating eGFP$^-$from eGFP$^+$ cells and will be instructive later on during single-cell analysis (Fig 3.1e). Genes that were higher expressed in eGFP$^+$ cells and DCICS organoids compared to baseline stem cells have already been implicated in the formation of respiratory epithelial cells in the lung (FOXA1, FOXA2, FOXP1, and NKX2-1) as well as branching morphogenesis and differentiation of the distal lung (RUNX1, MUC1, SFTPC, SFTPB, CLDN18 and NAPSA)[240] (Fig 3.1f).

Other genes that were expressed in the negative population were primarily fetal liver genes, including apolipoproteins (APOA1, APOB) and the plasma protein Alpha Fetoprotein (AFP).[241] Further GO term analysis showed enrichment of genes involved in embryonic respiratory lung morphogenesis and alveolar development in eGFP$^+$ progenitors and DCICS organoids, while processes related to liver development were prominent in eGFP$^-$ cells. The co-existence of lung and hepatocyte progenitors on day 15 of the differentiation protocol raised the question of what mechanisms drive the exclusive specification of these lineages from the FE stage.

3.1.2 Time-resolved single-cell characterization of early lung development

To assess the transcriptomic changes on a daily basis, Drop-seq was performed on the first 16 days of the differentiation protocol. For pre-processing, the main procedure as described in section 1.3 was applied, for which the final parameters and thresholds are listed in the appendix. As 1000 cells were expected per sample, the first 1200 cells with the highest number of transcripts per cell were included further. This resulted in a total of 10,667 cells for which the UMAP showed a sequential arrangement of cells, agreeing with the temporal order of sampling, particularly show-casing the harsh perturbation in gene expression induced by the medium change after the DE (day 0 to day 6) and FE stage (day 7 to day 10) (Fig. 3.2a). PAGA was performed to assess the connectivity of the Louvain clusters (resolution 1) and further corroborated the three major domains in the high dimensional manifold corresponding to the three stages of the differentiation protocol (Fig. 3.2b). Next, the the dynamics of gene expression along the differentiation trajectory from day 0 to 15 was modelled. Louvain cluster marker reflected the temporal heterogeneity, which was further delineated by inclusion of real time points. For this step, genes that show significantly altered expression patterns across time were of interest. As the patterns towards early lung progenitors were of interest, only those cells that were positive for either eGFP or NKX2-1 from days 11 to 15 were selected for the trajectory inference. Ribosomal derived genes were excluded from this analysis.

A regression model based on splines was used for this time-course data to model non-linear effects of continuous variables. For each gene a natural cubic spline with 4 knots was fit while using the time points of extraction as explanatory variables. UMI counts of each cell were included in the model as a covariate to account for differences in library size. As it is non-trivial to interpret p-values across time, the adjusted p-value solely served as a ranking for the genes, of which the top 200 are depicted in Fig. 3.2f. Upon closer inspection, well-described transcription factors in undifferentiated stem cells (POU5F1, NANOG) as well as genes defining the DE stage (SOX17, MIXL1, EOMES,

CER1, CXCR4 and LEFTY1)[240] were indeed among the top ranked genes at earlier time points, whereas genes important for the formation of lung progenitors (NKX2-1, IRX3)[236] came up during later stages, when the commitment towards lung lineage is expected.

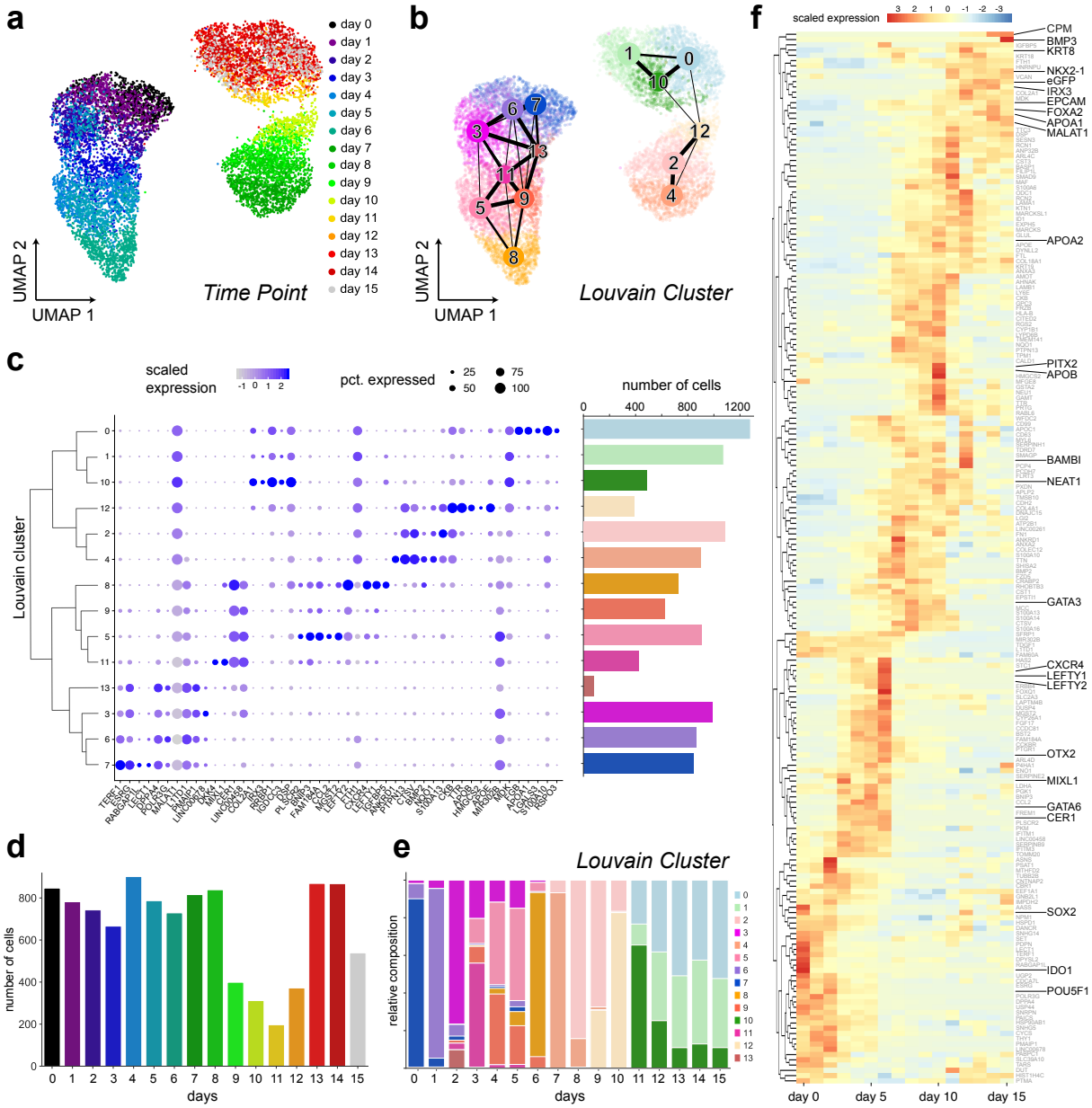


Figure 3.2: Time-resolved analysis of iPSC differentiation using scRNA-seq shows distinct hierarchy of gene expression changes. **a** Time point of sampling is colour coded on the UMAP projection of scRNA-seq transcriptomes. **b** The connectivity of distinct Louvain clusters as determined by graph abstraction (PAGA) overlaid onto the UMAP projection. **c** Top 5 genes per cluster shown in a dendrogram-sorted dotplot. Normalized expression levels are reflected by dot colour and the percentage of cells expressing the respective gene by dot size. **d** Barplots displaying the number of cells included in downstream analysis after quality control per time point of sampling. **e** Barplot of the Louvain cluster composition according to day. **f** Heatmap illustrating peaks of gene expression ordered by time point. Selected developmental markers of the lung and the liver are highlighted.

To capture overarching trends and avoid focusing on selected markers only, genes were grouped according to their average expression per day. For the DE and FE stage separately, the genes with significant association to time were selected using the spline model (adjusted p-value < 0.005) and categorized by hierarchical clustering. The dendrogram tree returned by the `hdist()` function of the R package `stats` was cut into 10 clusters, which were manually re-annotated to 6 final groups for each stage. The average expressions of the 100 genes with the lowest adjusted p-value per cluster are displayed in Fig. 3.3. This clustering approach revealed temporal patterns of stage specific markers, that corresponded well to literature-based expectations and were consistent with consecutive expression of DE, FE and lung progenitor markers.
SOX2 is a transcription factor that is essential for maintaining pluripotency of undifferentiated embryonic stem cells and was used as first validation point during day 0 to 6. Accordingly, cluster A1 showed decreasing levels as the differentiation progressed.

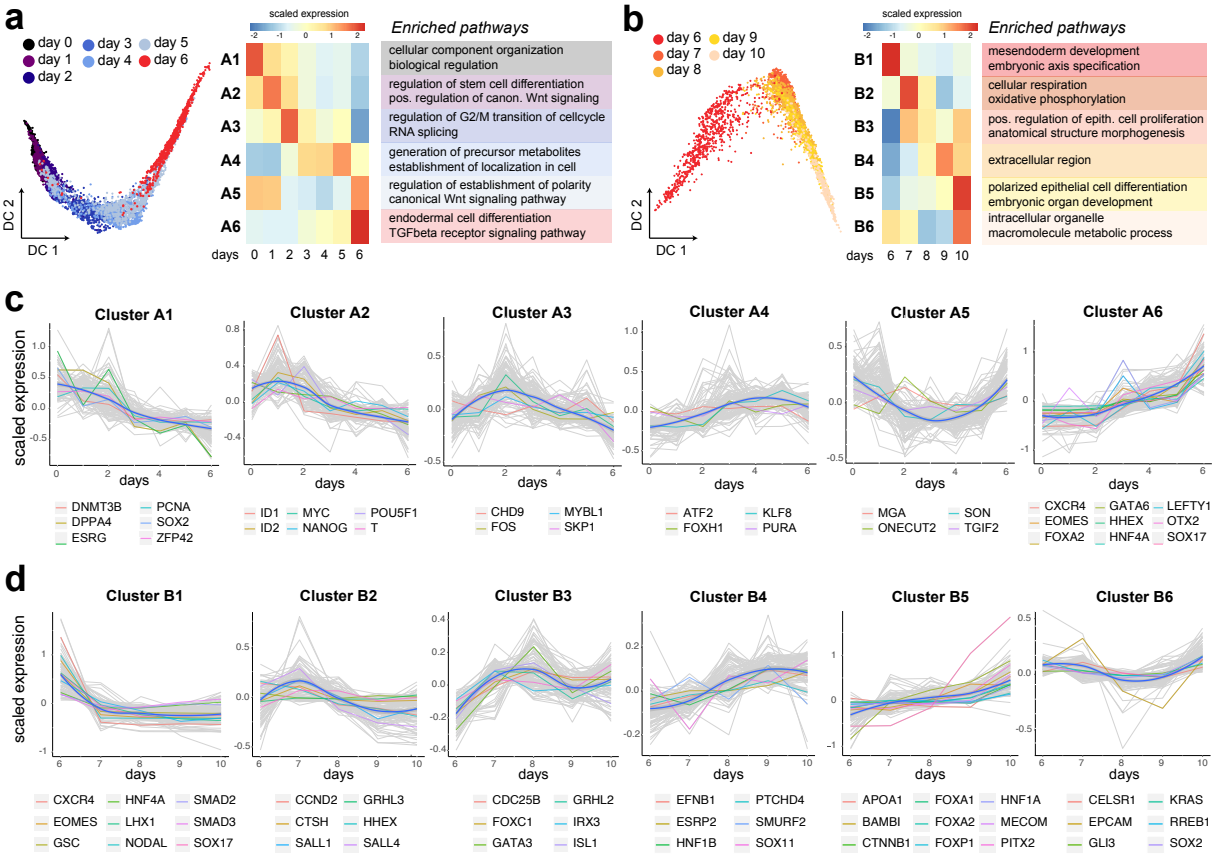


Figure 3.3: Dissection of the kinetic patterns of genes underlying the differentiation of lung progenitors and hepatoblasts. **a, b** Diffusion map of single cells for days 0-6 showing the transition from pluripotency towards DE (a) and FE (b). Scaled mean expression per gene cluster and associated pathways identified by hierarchical clustering of the genes with significant associations to days 0-6 (adjusted p-value < 0.005, 6101 genes) and days 6-10 (adjusted p-value < 0.005, 9352 genes). **c, d** Scaled mean expression of each gene cluster at the indicated time points, DE stage in (c) and FE stage including day 6 in (d). Grey lines correspond to the top 100 genes per cluster, ranked by adjusted p-value. Blue lines show the median expression within each cluster and the kinetics of genes of interest are highlighted.

Cluster A6 was of particular interest, as major DE genes like EOMES, LHX1, OTX2, CXCR4, LEFTY1, SOX17 showed gradual up-regulation, peaking at day 6. Further, A6 was enriched in pathways involved in endoderm differentiation and regulation of SMADs and TGF-β signalling.
During day 6 to day 10 the emergence of lung progenitors is anticipated, and potentially coincides with the maturation of hepatoblasts. Cluster B1 displays sharp down-regulation of genes linked to the induction of DE (EOMES, LHX1, GSC, OTX2, SOX1) and GO pathways associated with meso-endoderm development, which is consistent with the experimental dual-SMAD inhibition at this stage. Cluster B3 was enriched for epithelial cell proliferation and tube closure, indicating that the lung progenitor program has been initiated between days 6 to 10.

Isl1 has recently been shown to regulate the development of lung lobes and trachea-esophagus tube separation by the activation of Nkx2-1 in mice[242] while Irx3 promotes the proliferation of branched epithelium during lung formation,[243] both genes which were assigned to B3. Cluster B5 on the other hand showed increase in expression over time, coinciding with the fact that other included genes (FOXA2, FOXP1 and PITX2) are crucial for the lung morphogenesis and asymmetry in the mouse.[244] The GO terms lung morphogenesis, embryonic organ development and epithelial cell differentiation were also enriched in this cluster. The desired activation of the SHH pathway was further apparent, as target genes GLI4 and GLI3[245] were assigned to clusters with an up-regulation trend.

3.1.3 Recover gene kinetics during cell fate trajectory branching

In certain clusters with upwards expression pattern, hepatoblast/hepatocyte characteristic genes emerged, including transcription factors specific to the liver (HNF1A, HNF1B, TBX3) as well as first indications of genes, which are secreted by the liver (APOB, APOA2, TTR). APOB for instance appeared as early as the second stage of the protocol (Fig. 3.4a). Particularly at the last stage, there was a strong heterogeneity which was also reflected in the Louvain clustering. Contrary to cluster 10, cluster 0 and 12 did not express lung markers, but instead showed higher levels of hepatocyte markers in general (Fig. 3.4b,c). The expression pattern of FGB, which was the most up-regulated gene in the eGFP$^-$ population during bulk transcriptomics at day 15 (Fig. 3.1e), substantiated this separation further.

As this data indicated that the lung and liver lineages start separating during the second stage, it would be interesting to see which genes are associated with either differentiation lineage. Thus, a trajectory analysis was performed on differentiation day 7 to 15, covering the second and third stage of the protocol. To neglect the harsh effect introduced by the media change after day 10 in the embedding, genes that were differentially expressed between the FE and LP stage (logFC $>$ 1 and $<$ -1) were excluded from the highly variable gene list. Further, to guide the dimensionality reduction, the signatures from the initial bulk experiment on day 15 were taken into account. Genes that were significantly regulated between eGFP$^-$ versus eGFP$^+$ cells (logFC $>$ 1 and $<$ -1, 1294 genes) were used as input for the subsequent PCA. Following diffusion map calculation, the high dimensional manifold displayed several branching events. The expression of eGFP and NKX2-1 was restricted to one of the trajectories, while FGB expressing cells, as proxy for hepatocyte progenitors, were located on a different trajectory (Fig. 3.4d).

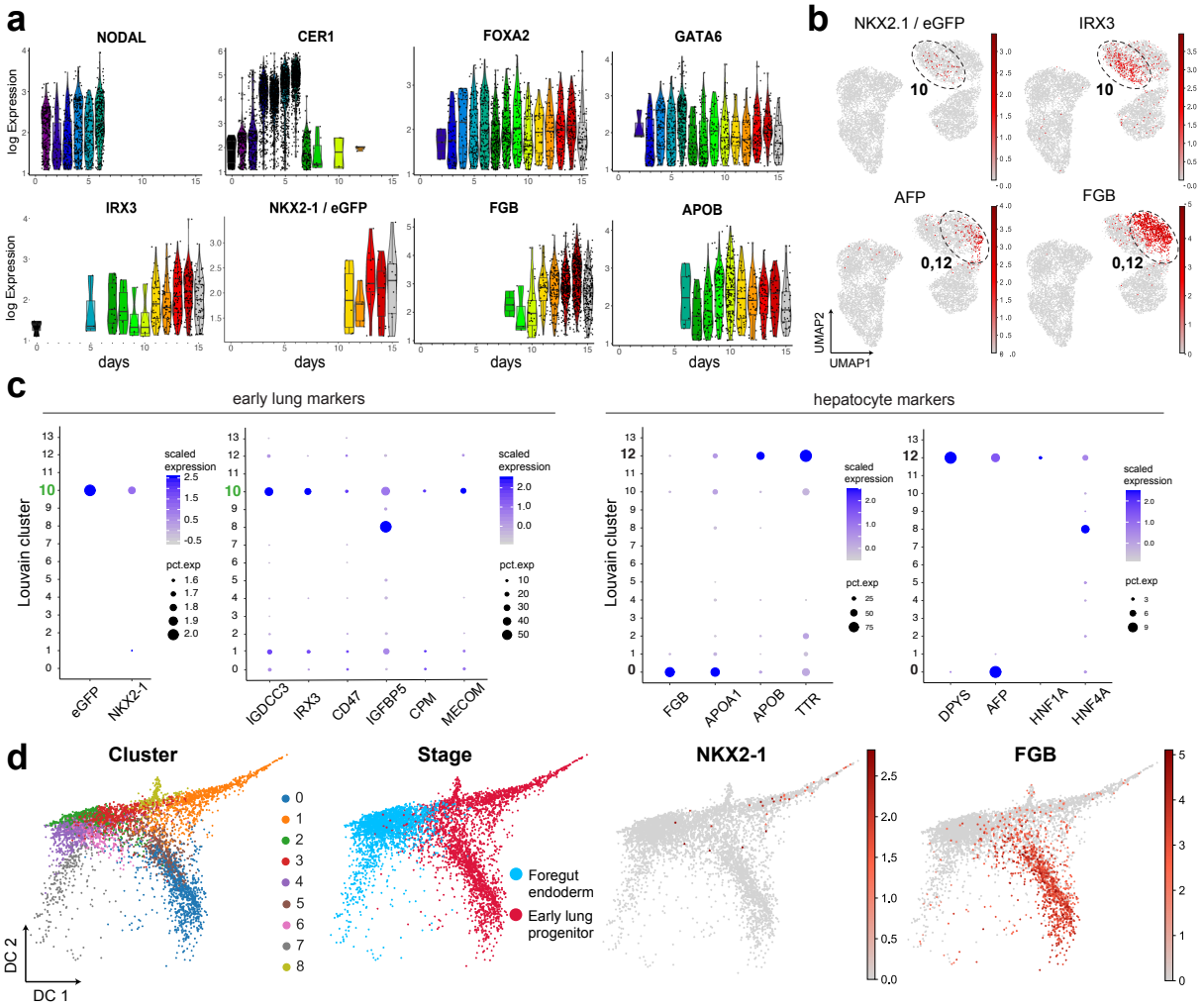


Figure 3.4: Overview of the single cell data set after filtering. **a** Violinplots of stage specific marker genes are shown for the day of sampling. **b** UMAPs overlaid with expression of lung progenitor markers NKX2-1 and IRX3, and liver markers AFP and FGB, showing the visual separation of the two lineages. **c** Dot plot indicating that markers of the early lung show higher expression levels in Louvain clusters 10 and 1, whereas markers of hepatocytes dominate cluster 0 and 12. **d** Diffusion maps showing that the lung markers NKX2-1 and eGFP were restricted to one sub branch. The hepatocyte marker FGB was enriched in another branch encompassing clusters 0 and 5. Diffusion map of cells from day 7 to 15 coloured by Louvain cluster, stage and NKX2-1 as lung lineage and FGB as hepatocyte lineage marker.

Additional branches (namely cluster 4, 6, 8) showed high expression of ribosomal/mitochondrial genes, or non-lung/non-hepatocyte lineage in general, and were therefore excluded from the analysis. Consequently, the two branches of interest remained, which originated in the FE stage and had their endpoint in either the lung or hepatocyte progenitor trajectory.

As a proof of principle, the retained cells were scored based on their similarity to the bulk signature. The cells scoring high for either lung or hepatocyte signature were indeed enriched in only one of the two branches (Fig. 3.5b). In the last step of the analysis, the branching event was better characterized to derive a roadmap of genes that drive the separation of lung progenitors from the liver fate. For this purpose, the differential gene

expression analysis was again performed with `diffxpy`[227] using the spline basis. To factor in the asynchronous behaviour of the cells, the diffusion pseudo time was calculated for the lung (cluster 1, 2, 3) and hepatocyte branch (cluster 2, 3, 5, 0) separately, instead of relying on the real time point which might not reflect the correct state of all of its cells.

The model should capture genes that show temporally altered expression pattern, and are differentially expressed across these two branches as well. The input consisted of the pseudo time as a continuous covariate and a categorical annotation *trajectory*, implying which branch each cell was assigned to, as factor to test for. For visualization purposes, the trajectory-wise pseudo times were manually binned into one shared source (cells from the FE stage) and 4 additional groups each. Average expression of cells per bin is shown for top 100 genes ranked by adjusted p-value in Fig. 3.5a. With this model, some key differences in the lung and hepatocyte differentiation branches could be revealed. As proof of concept, the expression of lineage specific markers such as NKX2-1, IRX3, and HNF1B, and FGB was exclusively increased along the expected trajectory (Fig. 3.4c). Other expression patterns validated the central roles of SHH and Wnt/β-catenin in early lung development, as key components of the pathway (DKK1, WNT5A, SP5, AXIN2) exhibited considerably higher expression in the lung branch.
Further, the exclusive expression of SOX2 highlighted the lineage specific activity of this pathway, as SOX2, canonical Wnt signalling, and FGFs often intersect in the regulation of self-renewal in development.[246] Another pathway that plays a role in the regulation of embryonic development would be the Notch pathway. HES1 is one of its canonical transcription factors, while DLK1 is a pathway modulator that is known to be involved in lung branching and morphogenesis.[247] While these two were specifically expressed in the lung trajectory, another activating canonical ligand (DLL1) showed expression in the opposite branch. This may indicate that hepatoblasts promote Notch signalling in lung progenitors by paracrine signalling. Key players of the TGF-β pathway (TGFB2 and THSB1) show a comparable upregulation in the lung branch as well. It was interesting to see that the exogenous treatment by SHH and CHIR did apparently not activate these pathways in the neighbouring hepatoblasts. To mechanistically test the exclusiveness to the lung trajectory, the Notch and TGF-β pathway were inhibited with the treatment of the γ-secretase inhibitor DAPT and SB431542 from day 11 onwards, respectively. Quantification of $eGFP^+NKX2\text{-}1^+$ cells yielded a significant decrease of the lung progenitors and decrease of NKX2-1 expression overall (Fig. 3.4d,e). These observations highlighted the involvement of the Notch and TGF-β pathways during lung specification towards lung progenitors as well as the important role of cross-talk between them.

Conclusion

In summary, the high temporal resolution in this project allowed for a detailed exploration of the mechanisms and their timing. The analysis of the single-cell data proved to be a powerful tool to delineate potential drivers that regulate the specification of lung progenitors in the foregut, as the derived patterns could be validated within the biological setting. Overall, the calculated pseudo time agreed well with the highly resolved time series of sampling, and the top genes ranked by their association to time reflected much of what has already been described in the field. The application of `diffxpy` based on splines in particular showed to be promising in extracting the genes with interesting temporal expression patterns, and will therefore be used the chapters to come as well.

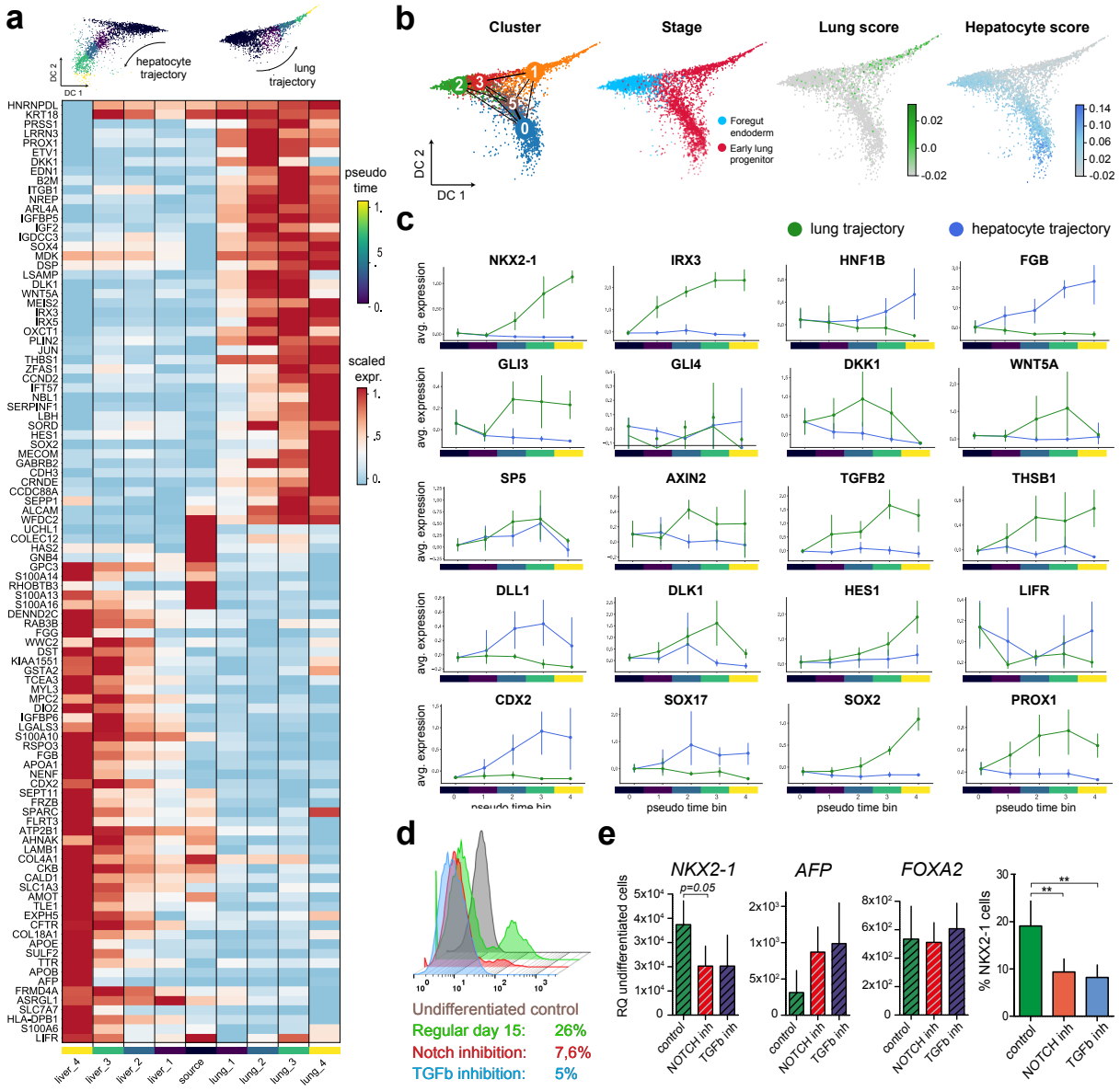


Figure 3.5: Reconstructing the transcriptional transitions from pluripotency to lung progenitors and hepatoblasts. **a** Heatmap showing results of unsupervised clustering of top differentially expressed genes along the lung and liver pseudo time trajectory, respectively. **b** Diffusion maps and Louvain clustering representing single cell transcriptomes of days 7-10 (blue) and 11-15 (red). Colour coded scores indicate similarity of single cell transcriptomes to bulk mRNA-seq data of NKX2.1^{+}eGFP^{+} and eGFP^{-} populations. **c** Line plots showing the expression dynamics of the indicated genes in the respective branches in accordance to the binned pseudo time ordering (vertical lines represent confidence intervals of 95%). **d** Representative flow cytometry and the quantification of NKX2-1^{+}eGFP^{+} lung progenitors on day 15 of differentiation, with or without the treatment by DAPT or SB431542 as indicated in days 11-15 (bars represent mean $\pm$ SD, n $=$ 4 biological replicates, ** denotes p-value $\leq$ 0.01 by unpaired, one-tailed t-test). **e** Respective fold change of NKX2-1, AFP and FOXA2 on day 15 relative to the parental undifferentiated cells (quantified by qRT-PCR, bars represent mean $\pm$ SD, n $=$ 3 biological replicates, unpaired, one-tailed t-test).

3.2 Bleomycin-induced lung injury leads to transient cell state that may persist in human lung fibrosis

During the last 3.5 years, the majority of time and effort went into the project described in the next section. Again, transcriptomic analysis was performed on longitudinal Drop-seq data, this time in order to characterize the mechanisms that occur in response to acute lung injury. Focus of this study were the shifts in gene expression space across the different cell types leading to tissue regeneration, or contrarily, to pin-point mechanisms that are dysregulated and drive disease manifestation rather than resolution. In the first part of this chapter, the transcriptomes of whole lungs from mice are chartered for the purpose of generating hypotheses and gaining an understanding of key mechanisms. To then evaluate their transferability to human disease, the second part will utilize data of human ILD patients across several cohorts and assess whether certain insights gained from the mouse model can indeed be translated to the in vivo reference.
Parts of this chapter address analyses that are part of concluded publications. Therefore, there will be overlap with the results presented in Strunz et al. 2019[171] and Mayr et al. 2021,[173] albeit the quality control and annotations have been improved in the data sets. Additional study cohorts became available in the meantime and have been incorporated as well, further expanding on some key messages from the listed publications.

Experimental data planning and sample collection, such as scRNA-seq, protein quantification, FACS-sorting, immunohistochemical stainings and respective statistical analyses were performed by Maximilian Strunz, Christoph Mayr and colleagues. Gabriela Leuschner was responsible for organization of the patient data. Some of the analyses were inspired by established code from Lukas Simon and heavily incorporated his input.
My contributions in both sub-project include great parts of the computational analysis, the pre-processing in general (alignment, quality control, visualization), integration across cohorts regarding the human data set, in silico exploration of cell state shifts and particularly the characterization of the impaired mechanisms and intercellular communication in the alveolar epithelium in the mouse and human lungs.

Introduction

As briefly introduced in section 1.4, the lung is made up of a variety of cell types that are necessary for proper function and displays remarkable potential to regenerate. In response to injury, certain cell populations become activated and adjust to changes in the micro-environment. Particularly the stem cell and progenitor populations have the ability to proliferate and replenish damaged cells. The exact mechanisms driving this repair have been the subject of a number of recent studies, and will be explored in this chapter.
The factors that lead to the manifestation of ILD remain largely unknown, and due to its progressive and chronic nature it is difficult to model in animals. Nonetheless, bleomycin has been shown to elevate proinflammatory cytokines (Interleukin-1 IL-1, Tumor Necrosis Factor-α TNF, Interleukin-6 IL6, Interferon-γ IFNG) and to cause fibrotic reactions within a short period of time after intratracheal instillation. Around day 9 the switch from an inflammatory response to a fibrotic one occurs.[248] Certain histological hallmarks of IPF patients, such as the deposition of collagen or obliteration of alveolar space, are well reflected in treated animals, allowing the study of some aspects of human disease.

The bleomycin-mediated lung injury mouse model was used to simulate early stages of lung fibrosis, and the induced transcriptomic changes were profiled over a 4 week time-course (Fig. 3.6a). Drop-seq was performed on day 3, 7, 10, 14, 21 and 28, leading to a total of 54,786 cells from 28 mice. The time course made it possible to quantify gene expression shifts, changes in cell type composition and altered cell-cell communication that arise after injury during the fibrotic phase. Many known mechanisms could be captured, such as the recruitment and activation of expected cell types. During the first exploratory analysis a peculiar alveolar intermediate state arose, mainly consisting of cells from time points after the lung injury. Therefore, the study was extended to a second more stream-lined experiment. The time points were upscaled to daily sampling for two weeks and EpCam$^+$ cells were specifically sorted and enriched for using Magnetic Activated Cell Sorting (MACS), allowing a higher temporal and cellular resolution of the processes leading to regeneration in the alveolar compartment.

The use of a mouse model enabled a detailed study on the mechanism involved in fibrogenesis, but does not necessarily reflect human disease in its entirety. In the latter half of this chapter, some striking aspects are then compared to single-cell data on human ILD patients. To increase generalizability, a number of recently published ILD scRNA-seq cohorts were considered in the process. Although this will not unveil the mechanisms that trigger disease development, it will provide some new insights and potential dysregulated mechanisms that help shedding light on the pathogenesis.

3.2.1 A time resolved single-cell picture of lung regeneration

gene	gene name	UNIPROT summary for the encoded protein
ACTA2	Actin Alpha 2	Involved in cell motility, vascular contractility and blood pressure homeostasis.
ARG1	Arginase 1	Arginine metabolism is a critical regulator of innate and adaptive immune responses. Upon release from granulocytes it accumulates extracellularly during inflammation and suppressese T cell proliferation and cytokine synthesis.
CCL2, 7	C-C Chemokine Ligand 2, 7	Chemokines involved in immunoregulatory and inflammatory processes, binding to chemokine receptors CCR. Chemotactic activity for monocytes and basophils.
CCR1, 2, 5	C-C Chemokine receptor 1, 2, 5	Chemokine receptor family, whose mediated signal transduction are critical for the recruitment of effector immune cells to the site of inflammation.
CDKN1A	Cyclin dep. Kinase Inhibitor	Regulator of cell cycle progression at G1, tightly controlled by the tumor suppressor p53 in response to stress stimuli.
KRT8	Keratin 8	Keratins heteropolymerize to form filaments in the cytoplasm of epithelial cells. Plays a role in maintaining cellular structural integrity and cellular differentiation.
LCN2	Lipocalin 2	Involved in innate immunity, limits bacterial proliferation by sequestering iron bound to microbial siderophores.
LGALS3	Galectin 3	Localizes to the ECM, cytoplasm and the nucleus, plays a role in apoptosis, innate immunity, cell adhesion and T cell regulation.
MFGE8	Lactadherin	Promotes phagocytosis of apoptotic cells and has been implicated in wound healing, autoimmune disease and cancer.
SPRR1A	Cornifin-A	Envelope protein of keratinocytes, becomes cross-linked to membrane proteins, resulting in formation of insoluble envelope beneath the plasma membrane.
SPP1	Secreted Phosphoprotein 1	A cytokine that up-regulates expression of IFNG and IL12. Among its related pathways is degradation of the extracellular matrix.

Table 3.1: Marker genes for certain injury-induced cell states in whole lung mice data after bleomycin treatment. Retrieved and shortened gene descriptions from UniProt.[249]

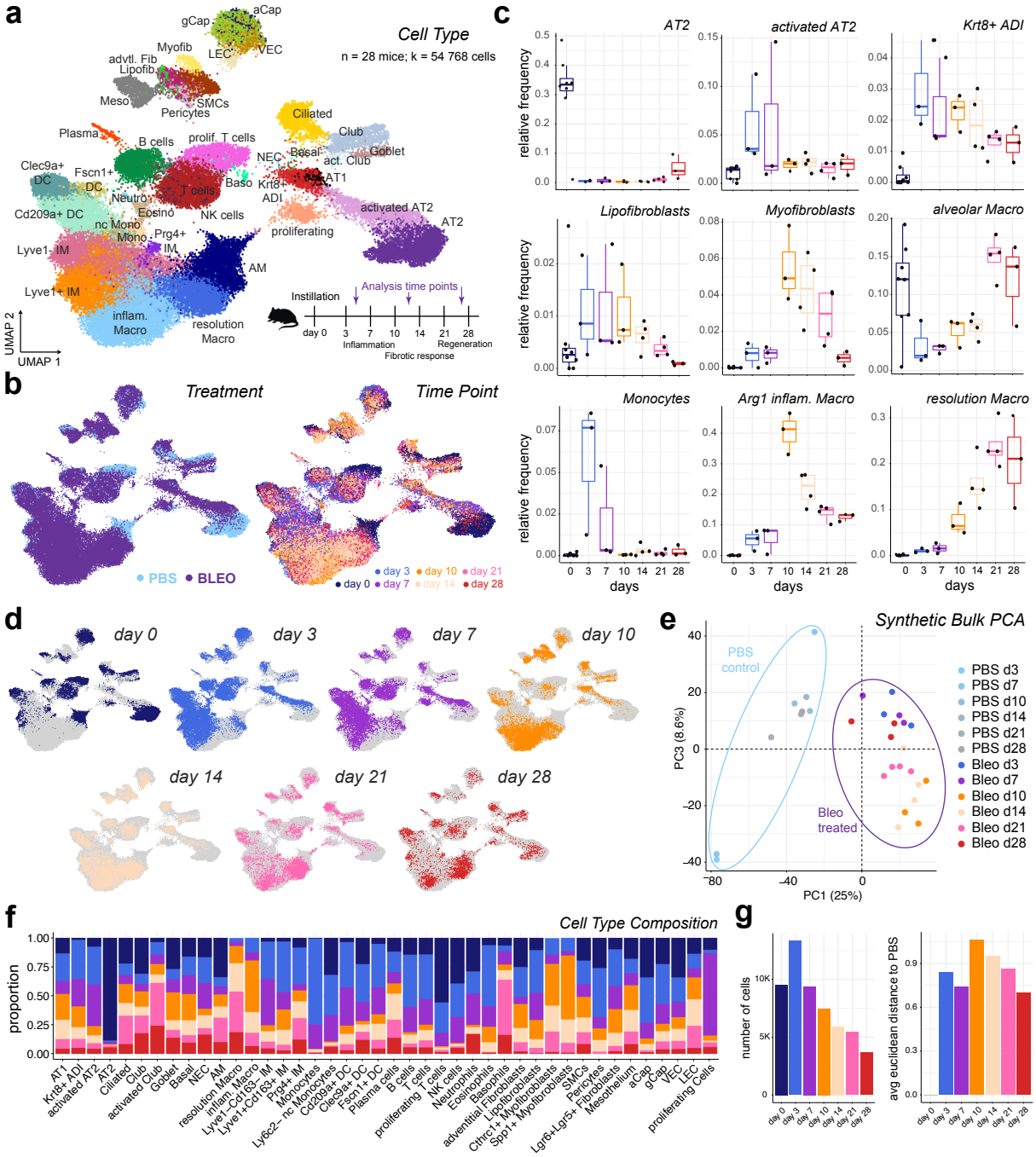


Figure 3.6: Longitudinal scRNA-seq analysis reveals injury induced cell type composition changes. **a** Transcriptomes of single cell suspension from murine whole lungs were analyzed using Drop-seq at indicated time points following bleomycin-mediated lung injury and projected using the UMAP algorithm. **b** UMAP with cells colour-coded by treatment and time point of extraction. **c** Boxplots of relative frequency of indicated cell type with respect to all other cell types per individual mouse. PBS mice (n = 8) are treated as day 0 control. **d** Separate UMAPs highlighting cells of each time point show-case temporal shifts in the lungs during injury response. **e** Scatter plot of PCA results based on synthetic bulks by sample-wise summation of counts for each gene across cells. The first component separates bleomycin treated mice from controls, whereas the third component corresponds to the temporal pattern of the injury response. **f** Relative composition across all cell types. **g** Overview of number of cells (left) and euclidean distance to PBS (right) for each cell type averaged per time point.

Pre-processing and quality control were carried out as described in the methods chapter, the exact filtering criteria and parameters for the analyses can be found in section 5. A 2D representation of the bleomycin mouse experiment after manual cell type annotation is displayed in Fig. 3.6a,b,d. PBS mice were used as uninjured lung reference and are denoted as day 0. To test reproducibility across replicates, gene expression was manually summed over cells for each sample in order to generate synthetic bulks, for which PCA confirmed good agreement of the replicates per time point (Fig. 3.6e). The 5 main lineages were separated and subsequently annotated using canonical marker genes and previously published scRNA-seq data sets (Fig. 3.7), resulting in 38 final cell type identities. Most clusters, especially cell types that are present in baseline, contained cells from both conditions, while some bleomycin-specific cell states were enriched in mice from later time points. Cell frequency dynamics reflected many well-described processes in early inflammation (Fig. 3.6c,f). Fibroblast populations are primarily responsible for the tissue scarring as regular response after injury due to their increased expression of extracellular matrix proteins, which are necessary to promote proper healing. Activated myofibroblasts are assumed to undergo apoptotic clearance afterwards, also through interaction with activated macrophages during scar resolution.[250] Acta2$^+$ Myofibroblasts in this data set reflected the prior knowledge, as their numbers were increased during the fibrotic peak at day 10 to 14 (Fig. 3.7e).
Ly6c2$^+$ classical monocytes were recruited from blood within days after injury and likely contributed to the appearance of Arg1$^+$ profibrotic macrophages peaking at day 10. After initial response, Mfge8$^+$ resolution macrophages started appearing and replaced the inflammatory macrophages from day 14 onwards.

The increase of cells after induced injury was strikingly apparent in the alveolar epithelial compartment, already in the visual representation. AT1 and AT2 cells were connected by cells mainly derived from intermediate time points. Subclustering of epithelial cells separated the alveolar types into four distinct clusters (Fig. 3.7b). AT1 and AT2 cells could be quickly identified by Sftpc and Rtkn2 expression, respectively. An activated AT2 state showed up-regulation of injury-induced genes, such as Lipocalin 2 (Lcn2) and Interleukin-33 (Il33), additional to AT2 markers. Another cluster was characterized by decreased expression of AT1 and AT2 markers, and displayed an unique gene signature. Marked by their up-regulation of Keratin-8 (Krt8), a gene encoding a fibrous structural protein, this cell state was titled Krt8$^+$ alveolar differentiation intermediate (ADI). This cell state further had higher expression levels of Sppr1a, encoding Cornifin-A that functions as a component of the cross-linked envelope in squamous differentiating cells,[251] and Lgals3, encoding Galectin-3 which plays a role in cell-cell adhesion, macrophage activation and apoptosis.[252]

3.2.2 Injury-induced shifts in cellular communication across time

A first exploratory analysis was performed to gain a more detailed understanding of the altered transcriptomic profiles. The gene expression shifts induced by bleomycin treatment were calculated with `diffxpy`[227] for each cell type at each time point separately, using the scaled number of counts as covariates and the treatment as factor to be tested for. It should be noted that for differential expression analysis certain populations were combined to meta cell types, as some cell states were only present at intermediate time points and not in the baseline. In order to ensure a condition to compare to, Krt8$^+$

ADI and activated AT2 were added to AT2, and resolution/inflammatory macrophages to AM. For myofibroblasts there is no obvious single origin, therefore all myofibroblasts were compared to a merged population of baseline (PBS) fibroblasts.

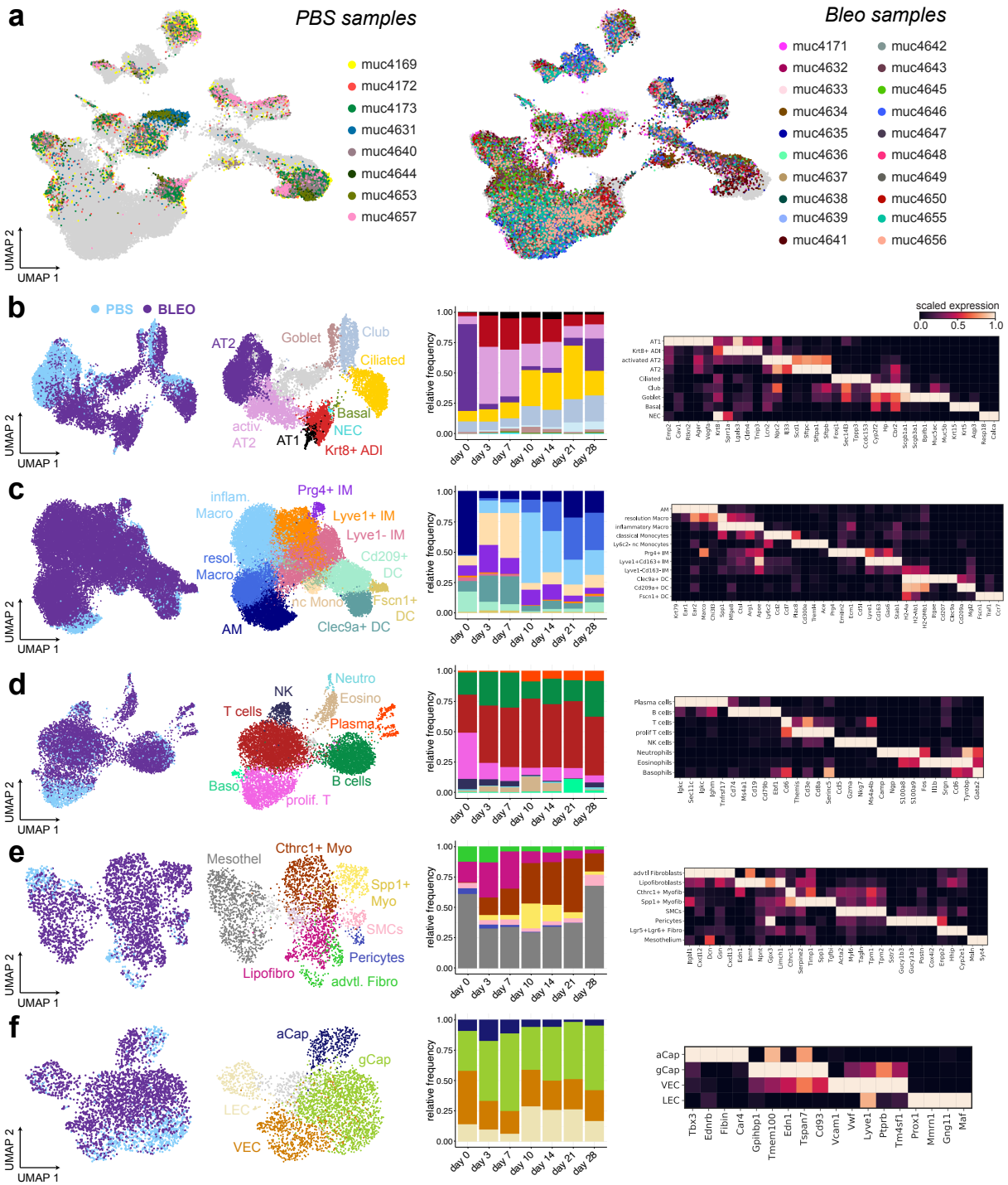


Figure 3.7: Compartment-wise annotation of cells from bleomycin exposed mice. **a** Split view of UMAP, separated by samples exposed to PBS control (left) or bleomycin (right). **b-f** Cell type annotation, relative composition within compartment and literature-derived marker genes of the epithelium (b), mononuclear phagocytes (c), lymphocytes and granulocytes (d), mesenchyme (e) and endothelium (f). Light gray cells were excluded.

To quantify the alterations, the amount of up- or down-regulated genes are shown in Fig. 3.8a. Apart from the neutrophil population, whose cell numbers were too low to be considered in the test, well-known response patterns could be discerned. Immune cells such as alveolar/interstitial macrophages, classical monocytes and T cells showed strongly altered regulations in both directions throughout the time course. Especially the macrophage populations reacted early to the stimulus and responded by up-regulation of chemoattractants for several leukocytes (monocyte chemoattractant proteins Ccl2 and Ccl7, Chemokine ligand Cxcl16, Fig. 3.8c). The corresponding receptors (Chemokine receptor Ccr1, Ccr2 and Ccr5, Complement C3a Receptor C3ar1, a receptor that stimulates chemotaxis and granule enzyme release) were not only expressed but also up-regulated in populations that are recruited to the site of inflammation, particularly monocytes from the blood and macrophages from the interstitial parts of the lung.
The up-regulated genes (logFC > 0.5 and percentage of expressing cells in the relevant group $> 10\%$) were then mapped against known receptor-ligand pairs to create communication networks. Because only induced genes with respect to day 0 cells were considered, the edges represent the number of pairs for which both the receptor and ligand levels were increased after bleomycin, not factoring in the communication that happens during homeostatic conditions. The strong initial communication shifted towards mesenchymal and endothelial populations at later time points, while AT2 states maintained strong interactions up until resolution of the fibrotic state (Fig. 3.8d). For day 10, the top ranked receptor ligand pairs are listed in Fig. 3.8e. Overall, the communication died down dramatically at day 28, a time point at which the cell populations start resembling their baseline counterparts.

Although the strong intercellular communication between these immune cell types is caught instantly by the eye, another strong edge was drawn towards AT2 cells as well. This population showed the strongest transcriptomic shift after injury, peaking with almost 500 up-regulated genes during day 10, which is not unexpected as there was a striking increase of cell number of the intermediate injury-induced cell state. As bleomycin reaches down to the alveoli after intratracheal administration, AT2 cells are directly hit by this reagent. They either get depleted for large parts, or enter another cell state promoting an inflammatory response. The induced gene program suggested the latter, many of the up-regulated ligands found corresponding receptors on macrophage populations, such as the previously listed Lgals3, Cxcl16, and also Annexin A1 Anxa1, which promotes resolution of inflammation and wound healing.
Additional to the immune cell interactions, signalling towards the gCap population was also occuring. For instance, the Cell Surface Death Receptor Fas plays a central role in the physiological regulation of programmed cell death, and had corresponding ligands in the AT2 population (Anxa1, Lgals3, Calm1, Amphiregulin Areg, an epidermal growth factor.)

Many of the up-regulated ligands in myofibroblasts on the other hand are involved in cell adhesion or cell-matrix interactions (e.g. Integrinβ-1 Itgb1, Serpin Family E Member Serpine1, Osteopontin Spp1, Fibronectin Fn1, Tenascin C Tnc, Thrombospondin Thbs1). For many ligands the corresponding receptor lies on other myofibroblast cells as well, suggesting heavy autocrine signalling. Still, the strong edges towards epithelial and stromal populations motivated a more detailed exploration with focus on these compartments.

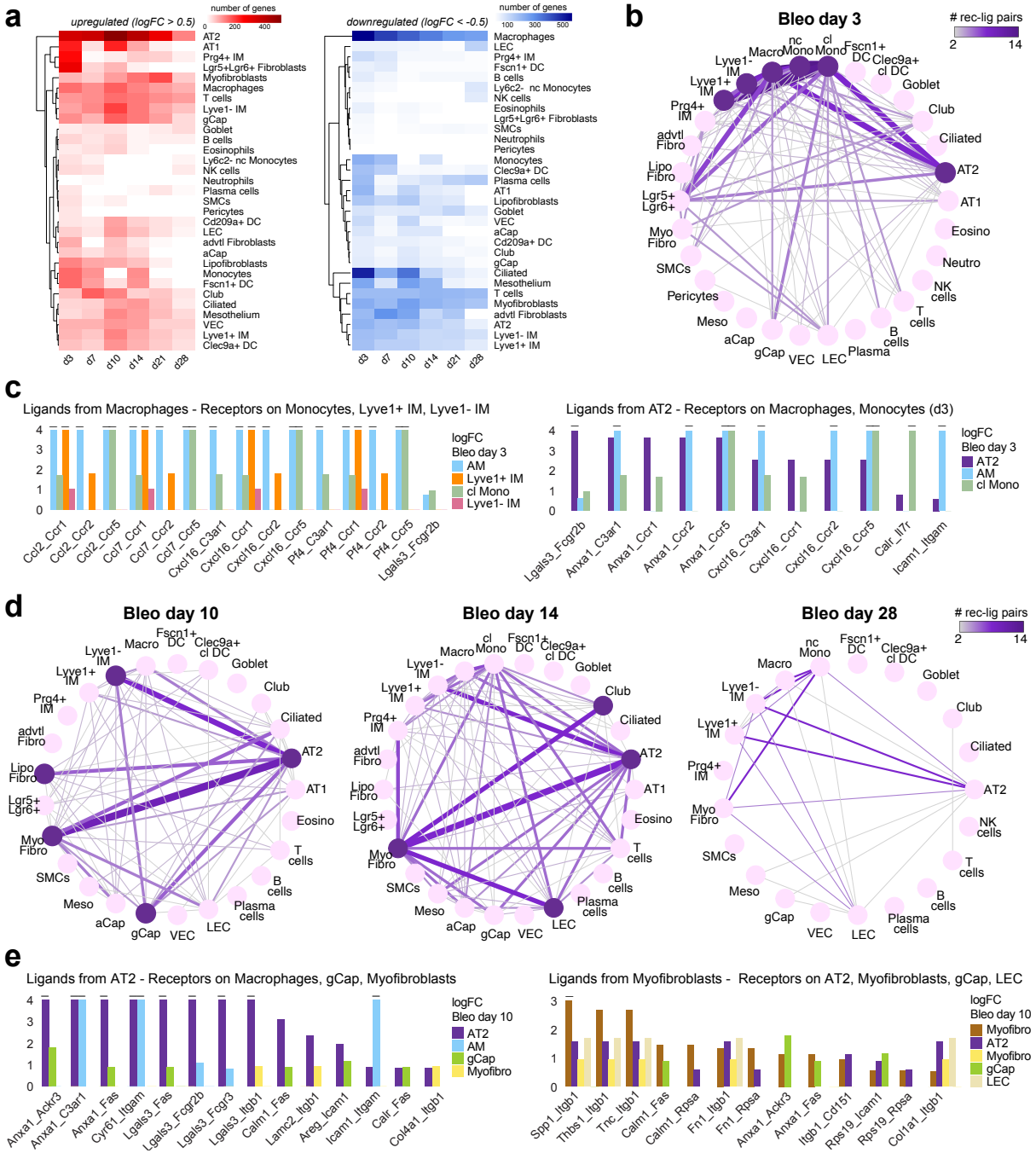


Figure 3.8: Altered pattern of cellular communication across time. **a** Heatmaps summarizing differential gene expression results for each time point and cell type combination. Colour indicates number of up-regulated (left) and down-regulated genes (right). **b** Connectome highlighting contribution of immune cell population during early stage of inflammation at day 3. Edge colour and width reflects the number of receptor-ligand pairs between the cell types. **c** Barplots of logFC values for the receptor-ligand pairs in cell types of interest that constitute the edges in the network at day 3. **d** Connectome plot as in (b) separately for day 10, 14 and 28 to show changes in communication during regenerative response and shift from immune cells towards mesenchymal compartment. **e** Barplots of logFC values for the receptor-ligand pairs in indicated cell types during the peak of fibrosis at day 10.

Specifically, `NicheNet`[232] was used to predict which ligands from other cell populations (*sender*) are most likely to affect target gene expression in a cell population of interest (*receiver*) and what their target genes would be. As sender populations all cell types were chosen, for the final visualization however only the prioritized ligands expressed in selected populations are displayed, to avoid overcrowded panels. In a first analysis, the gene signature of classical monocytes and alveolar macrophages at day 3 after bleomycin treatment was used as target gene signature, respectively (Fig. 3.9). The proposed upstream ligands are not only expressed but also up-regulated after bleomycin, particularly in endothelial populations (e.g. Col4a1, Jam2, Ccl7), fibroblasts (e.g. Thbs1, Tnc, Fn1, Spp1) and also in the transient alveolar state (e.g. Lgals3, Ceacam1, F11r).

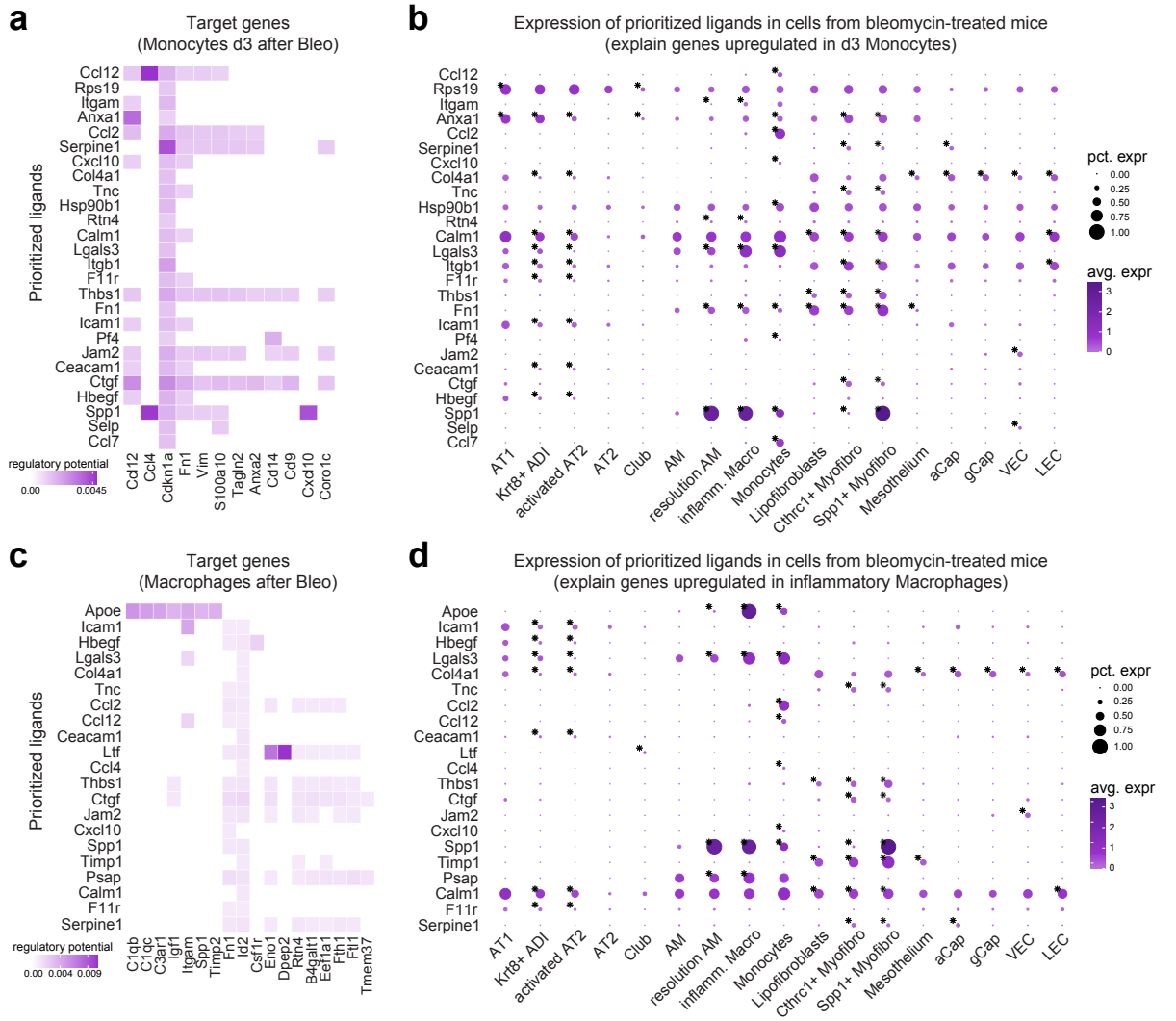


Figure 3.9: Immune cells react to cellular cues from surrounding cell populations. **a, c** Potential ligands explaining gene up-regulation at day 3 in recruited monocytes (a) and macrophages (c). Regulatory potential of top 30 ligands based on NicheNet's pearson correlation, restricting to ligands that have been up-regulated in at least one epithelial, mesenchymal or endothelial cell type upon bleomycin injury. **b, d** Dotplots visualizing the expression level and percentage of prioritized ligands implicated in gene expression shifts in monocytes at day 3 (b) and macrophages after bleo (d). Asterisk indicates significant up-regulation of the gene after bleomycin treatment. For clarity only relevant cell types are shown.

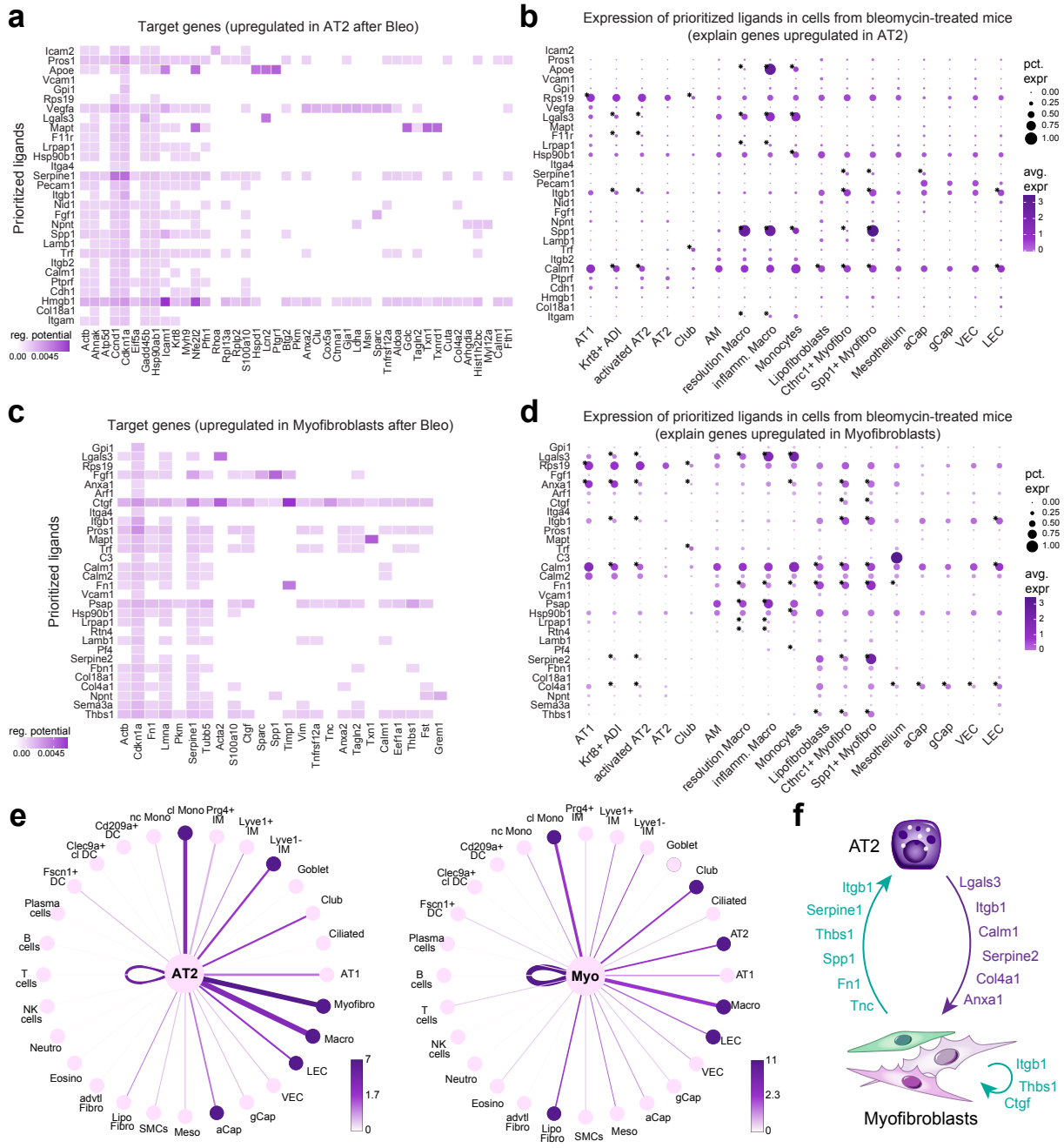


Figure 3.10: Zoom into inter-cellular signalling between injury-induced cell states. **a, c** Potential ligands explaining gene signature in AT2 cells (a) and myofibroblasts (c) after bleomycin injury. During differential testing, injury-induced alveolar cell states have been added to the AT2 meta cell type, while baseline fibroblasts from PBS mice were chosen as reference for myofibroblasts. **b, d** Dotplots visualizing the expression level and percentage of prioritized ligands implicated in gene expression shifts in AT2 cells (b) and myofibroblasts after bleomycin (d). Asterisk indicates significant up-regulation of the gene after bleomycin treatment. For clarity only relevant cell types are shown. **e** Starplot reflecting contribution of each cell type to changes is AT2 (left) and myofibroblast population (right). Edge width shows score which quantifies the predicted contribution of each cell type to the changes in the respective receiver cell population. **f** Schematic of prominent up-regulated ligands involved in AT2 cell-myofibroblast communication based on `NicheNet's` output.

Cdkn1a appeared as dominant affected gene downstream of many pathways, its encoded protein is likely involved in p53 mediated inhibition of cellular proliferation in response to DNA damage.[253] Additional to the increased signalling across compartments, the proinflammatory response was enhanced by strong autocrine signalling of macrophage populations, for instance by the up-regulation of inflammatory cytokines such as Ccl2, Ccl12, Cxcl10.

Following the temporal pattern, the second `NicheNet` analysis focused on injury-specific cell states and the potential niche cues responsible for their induction. Fig. 3.10 displays the top ranked ligands explaining gene expression shifts in AT2 cells, notably reflecting the Krt8$^+$ ADI signature, and in fibroblast populations likewise the myofibroblast signature. Expanding on the results in Fig. 3.8e, macrophage populations may interact with AT2 cells via Lgals3, Anxa1, the LDR Receptor Related Protein Associated Protein Lrpap1, encoding a receptor-related proteins that might regulate ligand binding activity along the secretory pathway, Osteopontin (Spp1), which promotes cell-mediated immune responses, and plays a role in chronic inflammatory and autoimmune diseases,[254] and the Integrin Itgam, implicated in various adhesive interactions of macrophages and their uptake of coated particles and pathogens. Some of these genes up-regulated in macrophages also found a corresponding receptor in myofibroblasts (Lgals3, Lrpap1, Fn1). Interestingly, the Connective Tissue Growth Factor CTGF was associated with a large portion of genes in the myofibroblast signature, which was not surprising due to its association with TGF-β and its already proven role in the pathophysiology of many fibrotic disorders.[255] Along with Itgb1, Thbs1, and many other ligands, the myofibroblasts were indicated to be capable of generating positive feedback loops.

During cell-cell communication exploration, the proposed interaction routes in between transient alveolar cell states and the myofibroblasts were of high interest, as they showed the greatest number of receptor-ligand pairs. Certain pairs indeed underpinned the hypothesized interactions (Fig. 3.10f), most prominently Lgals3, Itgb1, Serpine2, Col4a1 and Anxa1, all up-regulated in the Krt8$^+$ ADIs and with the potential to affect the myofibroblast gene signature. Likewise, eminent myofibroblast marker such as Serpine1, Itgb1, and Tnc, Spp1, Fn1 from Fig. 3.8e, have corresponding receptors in AT2 cell populations.

3.2.3 Transient squamous Krt8$^+$ cell state in alveolar regeneration

Driven by the rise of the intermediate alveolar state, their striking up-regulation of known inflammatory ligands and their consequent interaction with other activated cell types, the second transcriptomic profiling experiment aimed at a deep-dive into the epithelial compartment for a clearer picture. Lung epithelial cells were specifically selected by sorting EpCam$^+$ cells using Magnetic Activated Cell Sorting (MACS) prior to the transcriptomic profiling. Drop-seq was carried out daily up to day 13, and also at separate later time points up to day 54 after injury (18 time points in total, 2 replicates each, n = 36 mice) to capture the recovery of the system back to baseline with fully regenerated AT1 cells. This increase in cellular and temporal resolution should ideally model the rise of Krt8$^+$ ADI and the associated gene programs. After pre-processing and exclusion of remaining, non-epithelial cells using the filtering criteria as listed in section 5, the final UMAP of this high resolution epithelial data set is depicted in Fig. 3.11a.

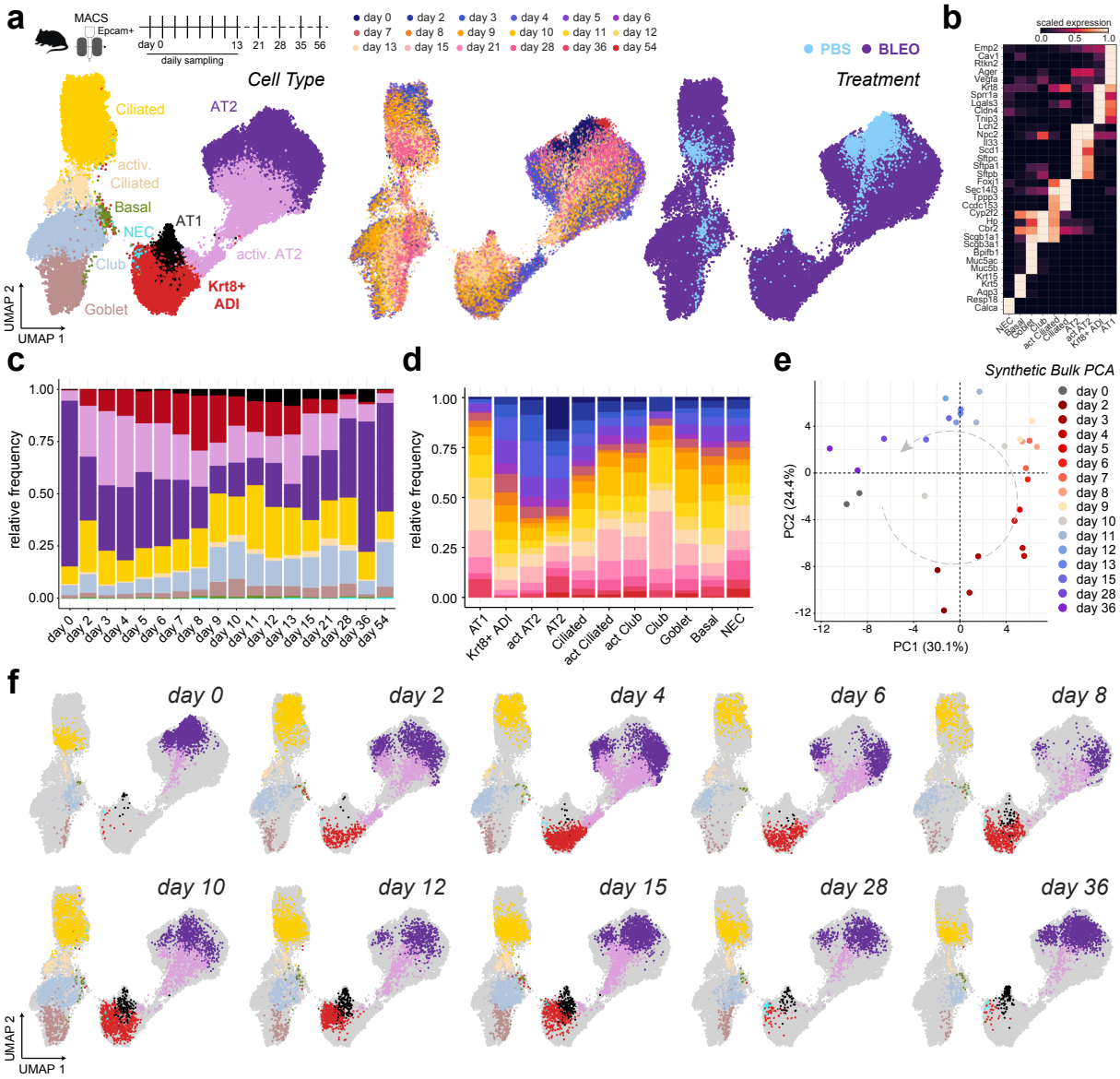


Figure 3.11: Sampling with higher temporal and cellular resolution affirms a transient cell state during alveolar regeneration. **a** Generation of a high resolution data set by MACS-sorting for EpCam$^+$ cells and profiling with Drop-seq at 18 indicated time points after bleomycin injury. UMAP of epithelial compartment coloured by cell type, time point and treatment (n = 36 mice). **b** Common marker genes for epithelial cell types and the novel transient cell state characterized by Krt8 expression. **c** Relative proportions of cell types along temporal axis highlighting appearance and disappearance of cell states after injury. **d** Relative proportions of time points within each cell type. **e** Scatter plot of PCA results on synthetic bulks by sample-wise summation of counts for each gene across cells. Replicates appear to follow a circular pattern aligning with the temporal trend. **f** UMAP highlighting cells of selected time point. Note the time dependent movement of cells within the data manifold.

Cell type identities were consistent with the whole-lung experiment described earlier, capturing the same epithelial populations at much higher cell numbers. The injury-specific alveolar cell states, which appeared at the earliest measured time point (day 3) in the whole lung data, was also induced in the epithelial counterpart already as early as day 2.

Most reassuring was the phenomenon how baseline AT2 cells again switched their expression towards an activated state with increased expression of inflammatory genes such as Lcn2, Npc2, Il-33 (Fig. 3.11b). Relative composition analysis revealed the continuing rise of the Krt8$^+$ ADI state up to day 8-10, from which onwards their presence decreased over time and disappeared almost entirely at the later days (Fig. 3.11c,d). PCA on manually sample-wise summed expression data (synthetic bulk) and trends in the UMAP reinforced the temporal changes in the transcriptomic space, displaying a progressive distinction to the baseline cells during fibrotic phases and returning to closer resemblance to non-perturbed lungs after three weeks during the regenerative phase (Fig. 3.11e,f).

In the coming sections, the linear trajectory of AT2 cells towards AT1 cells upon lung injury will be the main focus. Therefore, as means to model AT1 cell regeneration, both the whole lung and the high resolution epithelial data set were subset to only the alveolar compartment including the transient cell states. The two data sets were combined, encompassing 46,264 alveolar cells, and the PCs, knn graph and 2D embeddings were re-calculated. The resulting UMAP reflected the temporal patterns (Fig. 3.12a,b), while the diffusion map overlaid with the diffusion pseudo time allowed for a reasonable ordering of the cells. Scoring cells for gene programs of interest revealed that the Krt8$^+$ ADI cells highly expressed genes involved in cell senescence and pathways associated with stress-responses and secretion of profibrotic mediators such as p53, MYC, TNF-α via NF-κB, oxidative phosphorylation and epithelial-mesenchymal transition EMT (Fig. 3.12c), all pathways previously shown to be crucial for lung regeneration.[256,257,258] GSEA with `GOA Tools`[230] further confirmed the significant enrichment of genes known to be associated with wound healing, cell migration, ECM interaction and apoptosis in general (Fig. 3.12e).

The differentiation trajectory and gene expression profiles of activated cells over time proposed the lineage hierarchy of AT2 cells towards alveolar intermediates to AT1 cells. To gauge which genes and transcriptional regulators drive this differentiation, spline regression was employed via `diffxpy` to capture genes that show a significant time-association. Owing to the dense temporal sampling, the real time points of extraction were used as continuous covariate to test for, along with the scaled number of counts to account for differences in cell sizes. Only genes expressed in more than 5% of cells in at least one of the 4 cell types were considered, as a mean to decrease the multiple testing burden. Finally, the gene expression analysis resulted in 3,082 significantly regulated genes (adjusted p-value < 0.05). To counteract the drop-out effect and display the results in a more defined manner, cells adjacent in pseudo time were taken together and their averaged expression across the resulting 500 dpt bins is shown in Fig. 3.12d. The differentiation trajectory can further be split into 4 phases, for each the top 20 genes are displayed. The inital phase was marked by well-described AT2 cell type markers, such as Sftpa1/2, Sftpb, Sftpc. Following the pseudo time, markers of the activated AT2 show increased expression leading up to the Krt8$^+$ ADI state, whose previously described markers (Sprr1a, Krt8, Lgals3, Areg) also peak temporally prior to the terminally differentiated AT1 cells. S100 Calcium Binding proteins are involved in cellular processes like cell cycle progression and differentiation, some of which show increased expression in the differentiation intermediate as well (S100A6, S100A10 and S100A11). Other striking examples were already encountered in the previous receptor-ligand analysis, such as Clu, Anxa1, Areg, Cdkn1a, Calm1 among

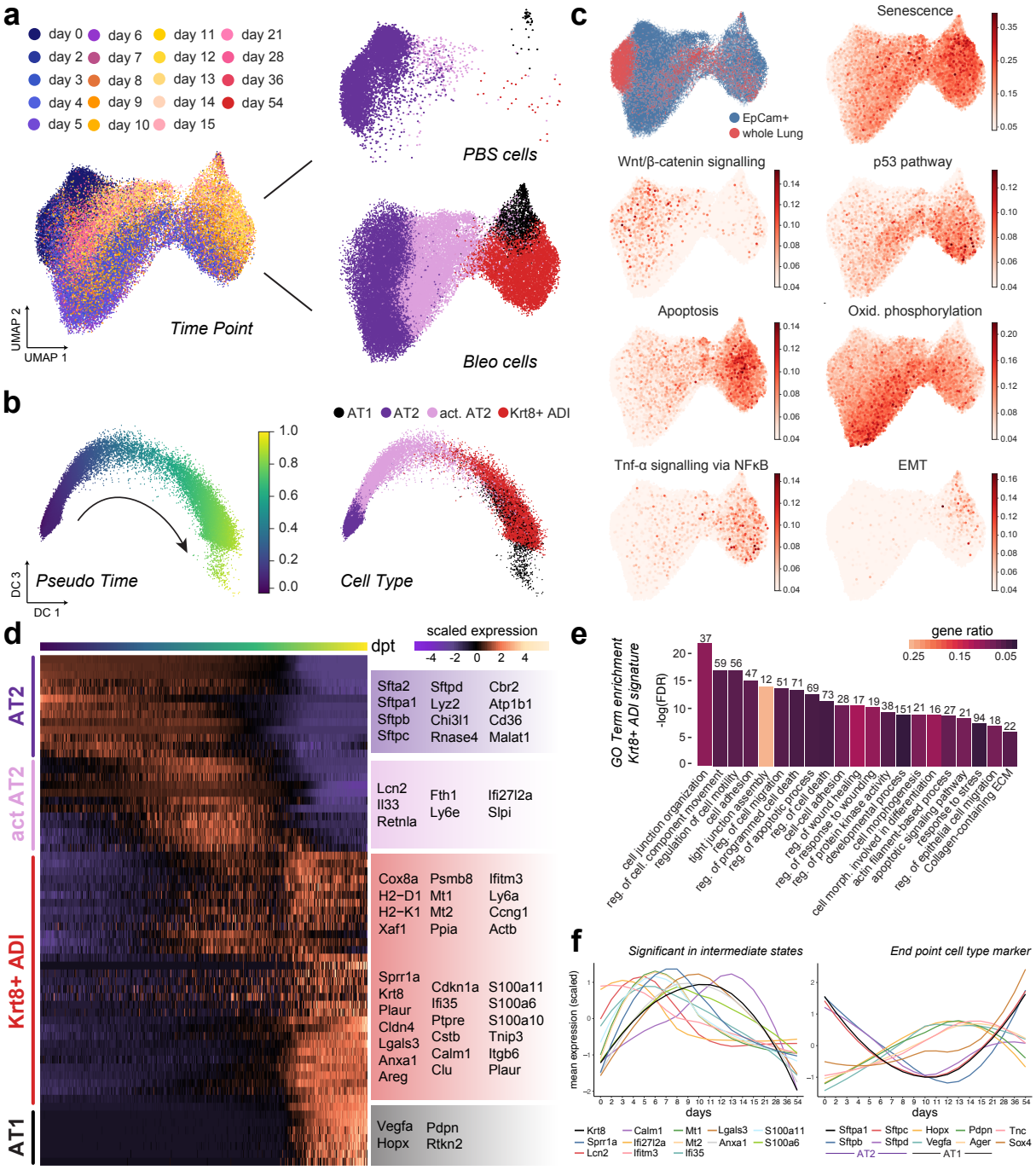


Figure 3.12: Differentiation trajectory modelling of AT2 towards AT1 during alveolar regeneration. **a** Integrated overview of alveolar epithelial cells from whole lung and EpCam$^+$ enriched data set, coloured by time point and cell type. **b** Diffusion map visualization of the alveolar epithelium, highlighting differentiation trajectory over pseudo time. **c** UMAP coloured by gene expression scores for indicated signatures (retrieved from MSigDB Hallmark gene sets). **d** Heatmap showing scaled expression across the differentiation trajectory for most significant genes ranked by association to real time. The top gene list was clustered (k = 4) and top 20 genes per cluster are displayed by binning expression values across nearest cells. **e** A selection of terms that were significantly (FDR < 0.05) enriched in the Krt8$^+$ ADI signature compared to all other epithelial cell types. **f** Line plots illustrating scaled expression levels of selected intermediate state markers (left) and AT1/AT2 cell type markers (right). Expression values are averaged per real time point and smoothed for clarity.

others, demonstrating that many genes that influence the gene expression profiles in other cell types get increasingly higher expressed during the regenerative response in the alveolar compartment. This further highlighted the manifestation of strong inter-cellular communication across niches in response to perturbation.

Finally, terminally differentiated AT1 cells were characterized by high expression of corresponding marker genes such as Vegfa, Pdpn, Hopx. To link the induced gene expression changes back to the tangible real time points, the scaled expression in alveolar cells of manually selected genes is shown in 3.12f. The transient appearance of the Krt8$^+$ ADI cells was further corroborated by the increase of its marker genes during fibrotic phases and their decrease thereafter. The AT2 cell signature, marked particularly by the expression of surfactant genes, was drastically decreased upon injury in the alveolar epithelial compartment, however it started recovering at later days after two weeks. The regenerative process in AT1 cells was mirrored by the increase of AT1 cell type markers during regeneration as well. In a recent study the binding of Tp53 to Sox4 and Nupr1 has been observed,[84] which are genes implicated in DNA-damage and cellular senescence pathways and known to regulate cytoskeletal genes.[259] The enrichment of these two genes in the transient population and their persistence of Sox4 expression in AT1 cells suggests that these play a role in AT2 de-differentiation.

The presence of cytoskeletal regulators such as Sprr1a and Sox4 hinted at cytoskeletal rearrangements and increased cell contractility in the Krt8$^+$ ADI cells during the final steps of maturation towards AT1 cells. To capture this injury-induced cell state in its natural environment and validate its transient emergence, immunostainings of Krt8 were performed in tissue sections of the lung parenchyma. Indeed, quantification of mean fluorescence intensity in alveolar space confirmed the transient burst of Krt8 expression during day 10 to 14 after injury, while Krt8 expression in uninjured control lung and fully regenerated lungs at eight weeks after injury was mostly restricted to the airways (Fig. 3.13a,b). Notably, there were rare Krt8$^+$ cells also found in the alveolar space of control lungs, raising the possibility that the observed cell state also represents natural homeostatic turnover. Additionally to their appearance, their morphology was examined via morphometric analysis on 300 micron-thick precision cut lung slices. Cuboidal AT2 cells from control lungs exhibited very low levels of Krt8, whereas in bleomycin-injured lungs the expression was increased in both Sftpc^{high} cuboidal cells and Sftpc^{neg} cells with starting squamous shape. As a means to quantify the mporphological changes, the spericity factor of 21 cells per condition is listed in Fig. 3.13c, describing a significantly flatter shape of the Krt8$^+$ ADI compared to the baseline AT2 cells.

Intermission

As a first intermediate result the transient appearance of this squamous alveolar differentiation cell state shall be recorded at this point. The transcriptomic profiling using model organisms was a suitable platform to study their rise. It was possible to describe the gene programs associated with their appearance and disappearance as they likely proceed to differentiate towards AT1 cell fate. Notably, these cells displayed morphological changes, enrichment of senescence genes and appeared to be a potent cytokine-releasing population majorly involved in cellular signalling. Furthermore, the pronounced role of macrophage populations during early inflammatory stages, as well as the amplification and transdiffer-

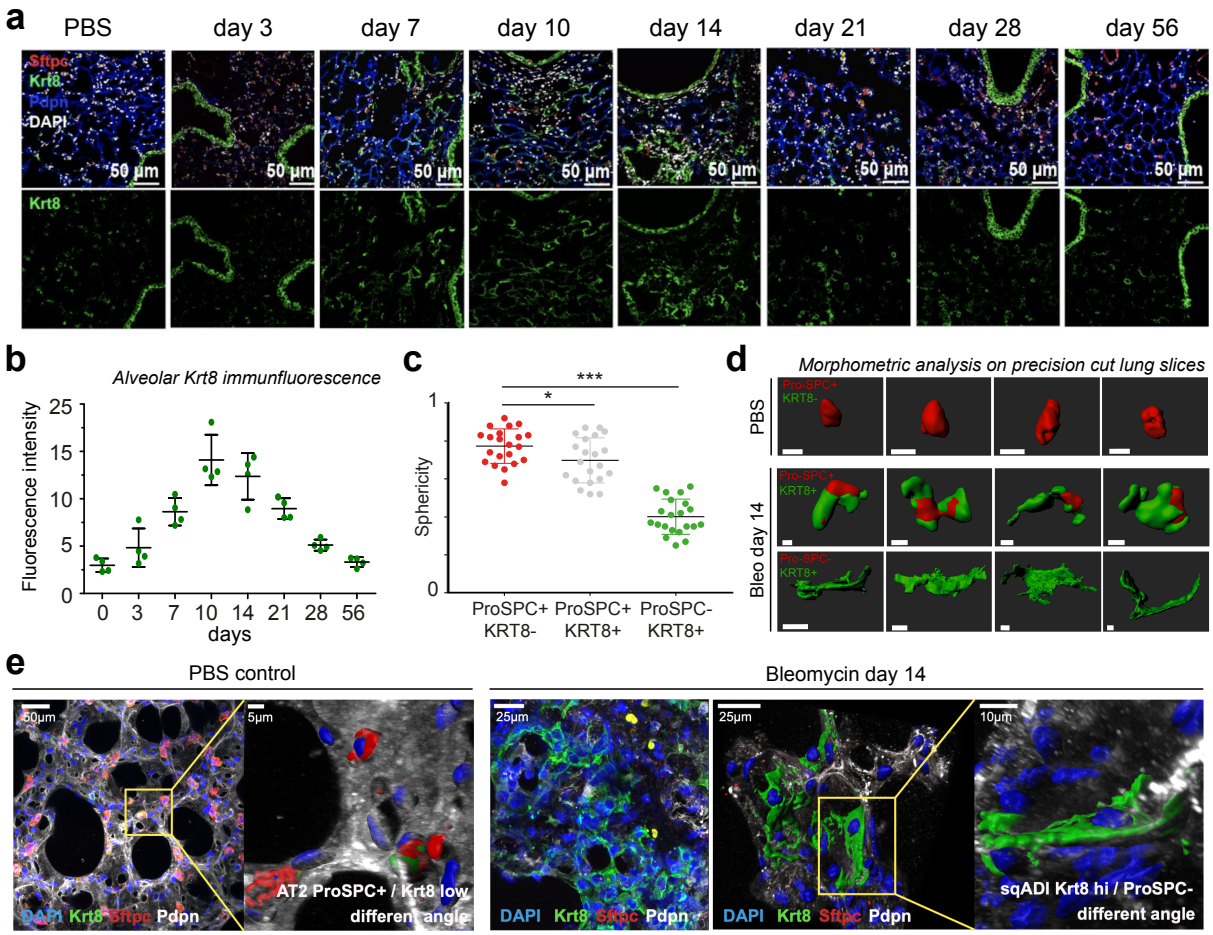


Figure 3.13: Validation of transient squamous cell state marked by Krt8 expression in mouse lung tissue. **a** Fluorescent immunostainings across analysis time points show nuclei, Krt8 and Sftpc (AT2) and Pdpn (AT1) (scale bar 100 µm). **b** Quantification of Krt8 mean fluorescence intensity in alveolar space (n = 4 per time point). **c** Alveolar cell sphericity analysis of 21 cells per condition reveals elongated cell shape for alveolar Krt8$^+$ cells in precision cut lung slices. Sphericity of 1 corresponds to round, cuboidal cells, 0 to flat cells. (PBS n = 2, Bleo n = 2, one-way ANOVA with Dunnett's post testing: * p = 0.0376, *** p < 0.0001). **d** Single cell morphometric analysis on immunostained 300 µm-thick precision cut lung slices (PCLS) highlight elongation in shape of transient Krt8$^+$ cells (scale bar = 15 µm). **e** Maximum projections of confocal z-stacks taken from PCLS are shown for a representative PBS control mouse and a mouse at day 14 after bleomycin injury.

entiation of fibroblast populations to the activated, ECM-producing myofibroblasts was properly captured and conveys the validity of the model and results. Intriguingly, it is well established in the field that epithelial cells that line fibrotic foci in IPF exhibit similar features of senescence, growth arrest and differentiation blockade.[260] The pathological milieu in lung tissue from patients might correspond in parts to the short-lived, pro-inflammatory and profibrotic environment after acute injury. Curious to assess how these observations relate to the respective human disease, similar computational analyses as just demonstrated are undertaken next by making use of in vivo patient data.

3.2.4 Multi-cohort integration of single-cell human lung fibrosis data

The field of single-cell genomics is continuously moving and evolving, providing more and more mechanistic characterizations of pathogenesis, disease progression and cell atlases that allow their description. During the course of data exploration of the mouse lungs, pre-prints based on transcriptomic analysis of human ILD patient lungs started appearing. While interesting ILD-induced shifts and aberrant cell populations were catalogued in a number of recent studies, a peculiar disease-specific epithelial cell state was of particular interest. It was first described in the studies from Haberman et al. (2020)[86] and Adams et al. (2020),[261] who titled the population *aberrant basaloid cells*. This population was entirely absent from healthy donor lungs and co-expresses basal epithelial, mesenchymal and senescence markers. The published aberrant basaloid signature included many genes that were encountered during analysis of alveolar regeneration in mice. As the bleomycin experiment sought to deduce mechanisms in human fibrosis, the targeted comparison of the induced cell states in both bleomycin-injured mouse and human ILD patient lungs has been carried out and is described in this second part of the chapter.

For an unrelated project the transcriptional changes in lung fibrosis were profiled on site, generating whole lung parenchyma single-cell suspensions of end-stage lung fibrosis tissues (n = 8) as well as uninvolved lung tissue freshly obtained during tumor resections, denoted as non-fibrotic control tissues (n = 13). A first exploratory analysis revealed many disease-specific patterns and covered the majority of cell types expected in the lung (Fig. 5.5). However, potentially due to their fragility and the lower coverage inherent to the Drop-seq method, aberrant basaloid cells could not be captured in the Munich cohort. To enable the targeted comparison to mouse alveolar epithelium and furthermore to increase the generalizability of the results, the on-site data set was combined with recently published human ILD cohorts, namely the priorly mentioned data from Haberman et al.[86] (*Nashville cohort*) and from Adams et al.[261] (*Newhaven cohort*). Reyfman et al. published one of the first large-scale ILD patient data sets in 2019,[262] which was incorporated here as well (*Chicago cohort*). Finally, a yet unpublished data set has been generated as part of an ongoing collaboration with the Königshoff Lab, in which EpCAM$^+$ cells from ILD and donor patients were enriched prior to scRNA-seq (*Denver cohort*). An overview of number of patients, cells and data source for these cohort is given in Tab. 3.2.
Apart from the enhancement of variability and robustness, one major advantage of these cohorts is their higher sequencing depth due to transcriptomic profiling with 10x. The baseline for integration were the filtered count matrices as they were made available by the authors.

Cohort	#control patients	#ILD patients	#cells	data retrieved from
Munich[173]	13	8	66,343	generated on site
Chicago[262]	8	9	73,237	https://metadataplus.biothings.io/geo/GSE122960
Nashville[86]	12	31	146,348	https://ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE135893
Newhaven[261]	30	31	202,688	https://ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE136831
Denver	4	3	79,523	generated in collaboration on site

Table 3.2: Overview and origin of data sets considered during integrative analysis.

The data sets were pre-processed separately according to the general work flow as outlined in section 2.5. The final filtering criteria, cell type annotations and corresponding marker signatures can be examined in Fig. 5.3-5.8. The following analyses are based on the ensemble of data sets, therefore the results are presented for the integrated data set, neglecting cohort-specific effects after their initial description. The batch correction was carried out in a two-step process. At the first level, a list of variable genes was defined for each of the cohorts separately as follows: The variable genes were calculated for each patient individually, setting the top 4000 variable genes as "highly variable" (hvgs) for the corresponding patient. For each cohort the intersection of these genes, annotated as highly variable across a minimum number of patients, was then defined as cohort-specific hvgs list. The thresholds were motivated by sample size of each cohort (Munich 4 patients 8268 hvgs, Chicago 5 patients 4763 hvgs, Nashville 8 patients 6783 hvgs, Nashville 10 patients 7306 genes). Finally, the intersection of the hvgs list yielded 1311 final genes after removal of cell cycle genes. To account for variation in number of transcripts owing to the different extraction protocols, the count matrices were scaled data set wise to unit variance and zero mean. The scaled count matrices were concatenated and PCA was performed using the defined hvgs list that was conserved across the cohorts. As a second level of batch correction, the neighbourhood graph was calculated using `BBKNN`,[219] setting `n_pcs = 50, neighbors_within_batch = 20` and `batch_key = "data_set"`. Cell type labels were retained as established in the cohort-wise annotation process.

The resulting visualization is shown in Fig. 3.14, demonstrating a general good agreement of annotations across the embedded space. Apart from the Denver cohort, which encompasses only epithelial cells, the majority of cell types were represented by all data sets. Nonetheless, variations in cell population frequencies remained, which can be caused by true biological differences, as well as differences in cell isolation protocols. Some noticeable disparities would be the higher number of captured granulocytes in the Munich cohort, with comparably low numbers of dendritic cells. The Chicago cohort displayed overall lower proportions of mesenchymal cells and Newhaven cohort mostly contributed to the leukocyte compartment, while the Nashvile cohort dominated the endothelial compartment of the data (Fig. 3.14b).
The induced compositional changes in the profiled patient lungs were in line with the epithelial remodelling and ECM expansion known to occur in ILD patients. As laid out in section 1.5.2, the hallmarks of this disease contain patchy chronic inflammation (alveolitis), small aggregates of proliferating fibroblasts (fibroblastic foci) and the phenomenon of honeycombs, spaces with thickened walls composed of fibrous tissue. Especially the "bronchiolization", i.e. a histologic abnormality in which metaplastic epithelial cells that are thought to be derived from bronchiolar epithelial cells, are typically observed in areas of the alveolar ducts and alveoli of ILD lungs. The increase of airway epithelial cells was apparent across all data sets, as their frequency was much higher in the disease condition. At the same time the numbers of alveolar type cells, which form a substantial fraction of epithelial cells in the distal healthy lung, were decreased or potentially replaced by airway cells usually confined to the proximal lung (Fig. 3.14e). Another eye-catching aspect of the epithelial compartment was the appearance of the previously described aberrant basaloid cells, restricted to ILD lungs only. These cells were captured by 3 of the considered data sets after manual re-annotation, and completely missing from the Munich and $EpCAM^+$ enriched Denver cohorts.

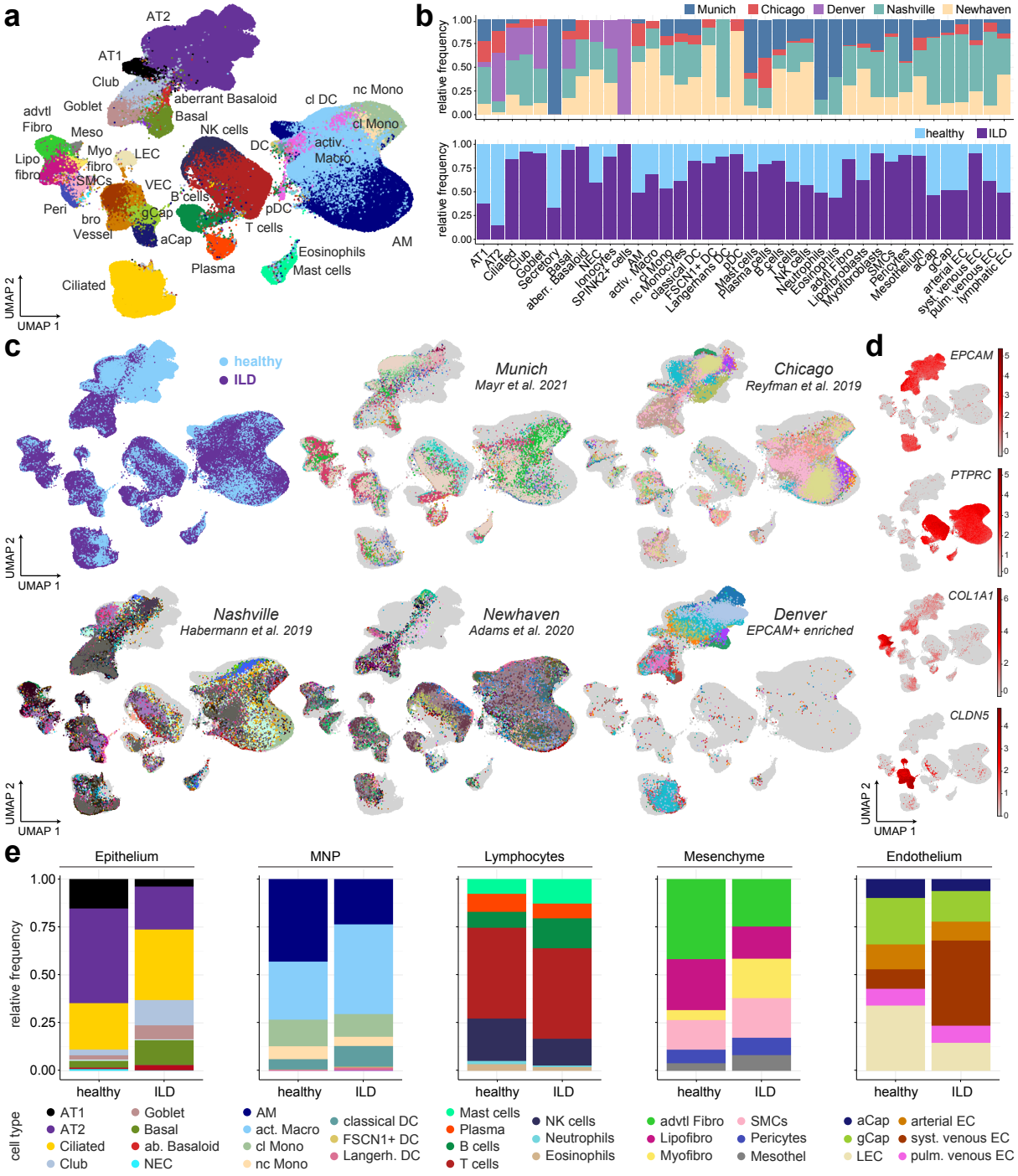


Figure 3.14: Integrated single-cell atlas of human lung fibrosis reveals disease-specific cell state and composition changes. **a** UMAP after integration of 5 patient cohorts and `BBKNN` correction of the knn graph, colour-coded by cell type annotation. Annotation was performed data set wise using an established list of marker genes, while considering published annotations for orientation. **b** Relative composition for each of the 40 cell types by data set (upper) and health state (lower), emphasizing disease-specific cell states. **c** Split view on UMAPs highlighting health state (first) and patients for each cohort separately, demonstrating a generally good agreement. **d** Indicated marker genes that were used to select clusters for subsetting into epithelial cells (EPCAM), leukocytes (PRTPRC) stromal cells (COL1A1) and endothelial cells (CLDN5). **e** Relative composition of cell types within compartment for integrated data set separated by health state.

The fibrosis observed in ILD is caused by the ECM expansion and regions of myofibroblast foci, defined by accumulation of matrix underneath epithelial cells undergoing injury and apoptosis, which eventually leads to progressive scarring of the tissue.[263] Matching these descriptions, the fibroblast cell states were all consistently increased in the disease condition. Apart from the myofibroblasts, which were largely absent in donor lungs, the overall composition of the stromal department was not shifted towards one particular cell type, suggesting that many of the distinct fibroblast populations play a role during scar formation. ILD is a primarily fibrotic disease, although some inflammatory responses can be observed, especially in areas of such fibrotic foci. In the course of this study, the focus has been on the epithelial and mesenchymal compartments. Nonetheless, immune cell types displayed interesting phenotypic changes in ILD lungs as well. Macrophages for instance are typically marked by high FABP4 expression, which is decreased in ILD macrophages. Instead, chemoattractants like CCR2, CCL7 or Osteopontin SPP1, all of which were also up-regulated in alveolar macrophages in mice lungs upon injury, marked the macrophages in ILD and were therefore titled *activated macrophages*. These made up the vast majority of mononuclear phagocytes in disease condition.

Vascular density, i.e. the ration of capillary area to surface area of alveolar walls, has been found to be gradually decreased at increasing stages of fibrosis.[264] In the integrated object, most of the endothelial populations showed unchanged proportions in health and disease. One subset however had strongly increased number in ILD lungs. Interestingly, in the original publication by Adams et al. (2020) the authors described such an expanded vascular endothelial population. These cells expressed COL15A1 and could be found in affected regions in the distal parenchyma of IPF lungs. However, these cells were transcriptomically indistinguishable from systemic bronchial vascular cells from control patients, which in turn were restricted to the peribronchial vasculature and were never seen in the lung parenchyma. The expansion of such vascular endothelial cells was evident after integration as well, and might reflect the expansion of the bronchial vascular network throughout the IPF lung.[261] Such additional spatial information is currently not reflected by the composition panels, but has to be kept in mind.

3.2.5 Disease progression alters cell type signatures and compositions

Cell type proportions are known to be skewed due to the dissociation bias intrinsic to single-cell tissue experiments, and hampers an exact description of cell frequency changes across conditions. As a mean to validate the compositions, a recent bulk RNA-seq study on lung tissues from 6 control and 10 IPF patients was leveraged (GEO GSE124685).[263] To approximate the progression of pathological changes, the authors sampled differentially affected regions in the same lungs multiple times, resulting in 95 samples. The extent of fibrosis for each sample was determined using quantitative microCT imaging and tissue histology and resulted in 3 categories: IPF stage 1, 2, and 3 corresponding to samples with no, moderate and advanced fibrosis, respectively.

For each of the 3 stages a signature was derived by establishing the differential gene expression changes against the control lungs. To infer cell type frequency changes from the bulk transcriptome, the cell type signatures were calculated via `scanpy's tl.rank_genes_groups` in ILD patients only and the top 50 genes with highest log fold changes for each cell type were considered further. The enrichment of each signature in a ranked list of

fold changes was statistically evaluated by the Kolmogorov-Smirnov test, which returns a p-value score signed by the effect size. Negative and positive values represent depletion and enrichment of the given signature in the ranked list, respectively. As shown in Fig. 3.15a, significant changes of many cell types were already discernible in early stage IPF 1, which still harboured more alveolar and capillary cell identities compared to the more advanced stages IPF 2 and IPF 3. At the other end of the spectrum, immune and mesenchymal cell population were increasingly more enriched across the fibrotic stages. The most prominent increases in cell numbers were marked by the goblet and ciliated cell types, again reflecting the bronchiolization that occurs during pathogenesis. Log fold changes in the bulk RNA-seq of the top 40 cell type markers derived from the single-cell data further corroborated the gradual increase of airway epithelial cells, and simultaneous decrease of alveolar epithelial as well as capillary cells along the disease progression (Fig. 3.15b).

Consistent with the increase in frequency, the myofibroblast signature along with the aberrant basaloid cell signature was clearly up-regulated in early stages, indicating that these represent early events in disease progression. The myofibroblast signature features various collagens and ECM proteins (COL1A1, COL1A2, COL3A1, COL5A1, COL10A1). Several genes have been up-regulated in the myofibroblasts in bleomycin-injured mice lungs as well, such as the Collagen Triple Helix Repeat Containing CTHRC1, involved in vascular remodelling and collagen matrix deposition,[265] the Iodothyronine Deiodinase DIO2, an enzyme shown to be correlated with ILD severity,[266] or the Secreted Frizzled-related Protein SFRP4, that acts as modulator of Wnt signalling.[267] Conversely, the aberrant basaloid signature displayed how these cells express the basal cell markers TP63 and KRT17 in addition to epithelial markers, but were negative for other established basal markers such as KRT5 and KRT15. Much like the alveolar differentiation intermediate in the mice lungs, genes associated with senescence (Cyclin Dependent Kinase Inhibitors CDKN2A, CDKN2B, Cyclins CCDN1, CCDN2) were up-regulated in this population. Further, genes involved in degradation of extracellular matrix (Cadherin CDH1, Serine Protease PRSS2), cell adhesion (Cadherins CDH2, CDH3) and several integrins and laminins (ITGA2, ITGB6, LAMB3, LAMC2) were included in the signature as well.

Instead of manually checking a small number of genes, a systematic cross-species comparison of the respective cell type marker signatures was carried out via the Jaccard index based `matchScore`[268] (Fig. 3.15d).
Although bleomycin administration in mice causes an inflammatory response and does not reflect the irreversible progression of ILD in human patients, other hallmarks can be captured in the mouse model, especially during the fibrotic phases.[248] As the switch to a fibrotic response happened during day 10 to 14 in the mouse data set, the cell type signatures were established with `rank_genes_groups` only on mice from day 10 after bleomycin treatment. Genes with logFC > 1 were considered in the final signature. Upon inspection of the correlation heatmap in Fig. 3.15d, the strong similarity of mouse fibroblast and endothelial populations with the human counterparts immediately catches the eye. Stromal populations from both mice and human have been shown to up-regulate ECM and adhesion proteins in previous sections, so the high correlation to the human disease was not surprising and reflected the validity of the model.

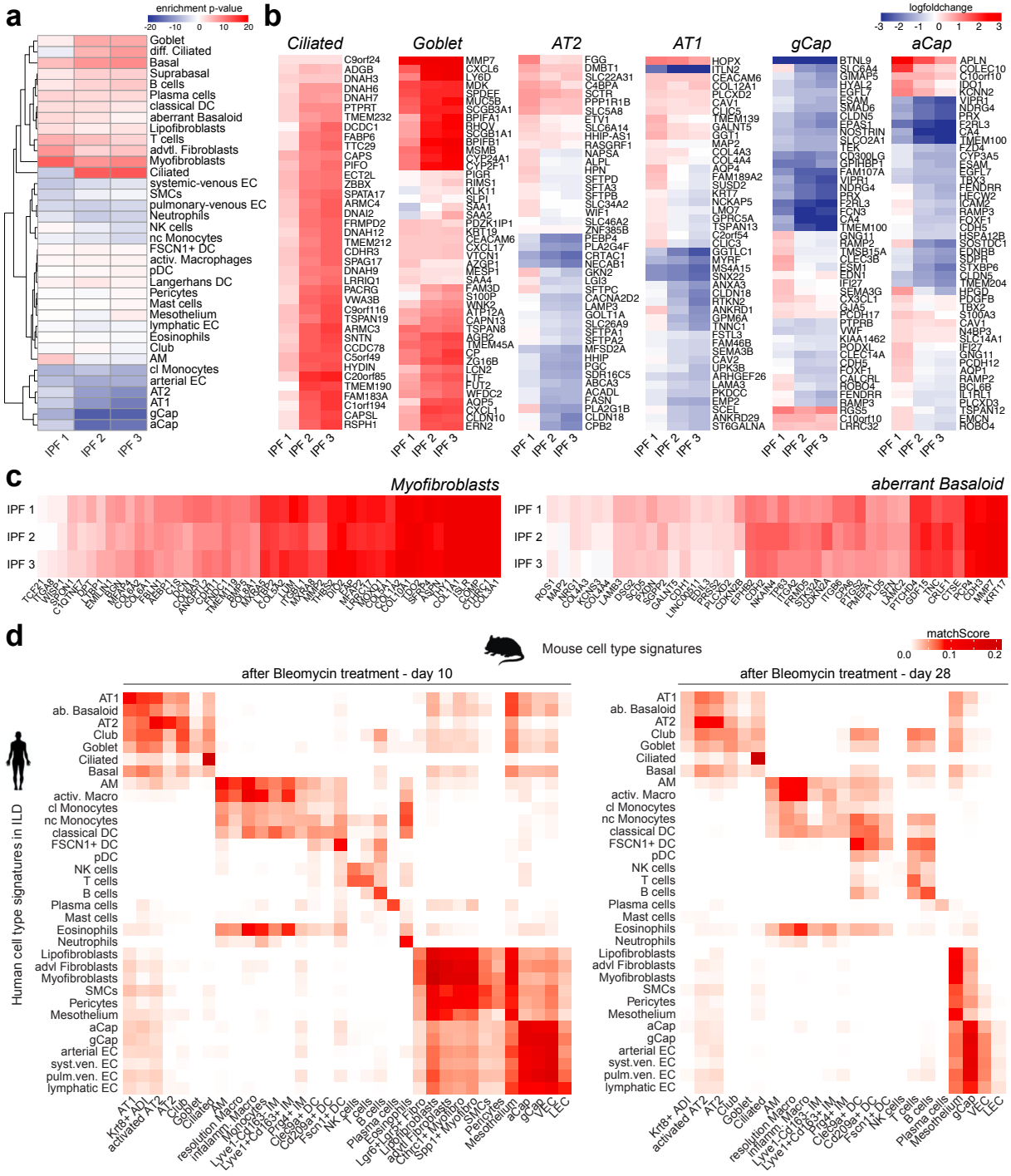


Figure 3.15: Disease progression alters cell type frequencies. **a** Cell type marker signatures from integrated ILD data set were used to deconvolve bulk transcriptome from a published data set across different histopathological stages, representing increasing extent of fibrosis (GEO GSE124685).[263] **b** Row-scaled heatmaps showing log fold changes of indicated cell type markers across IPF stages with respect to control patients. Cell type signatures were derieved from the integrated single-cell data. For the most affected cell types, the top 40 marker genes ranked by logFC are shown. **c** Heatmaps as in (b), with focus on disease-induced cell state in fibroblast (myofibroblasts) and epithelial populations (aberrant basaloid). Log fold chages from bulk IPF data confirm consistently increased expression across all IPF stages compared to control. **d** Quantification of signature overlap between ILD patients and bleomycin-treated mice at different time points using `matchScore`.[268] Human signatures were established on ILD patients only.

The cell type signatures were largely conserved across species, as most cell types displayed highest resemblance to the corresponding type in the other organism, while the transcriptomes of human macrophages were slightly shifted towards inflammatory subtypes from the mice. Another interesting aspect would be the heightened similarity of the epithelial compartment from human ILD patients to the mesenchymal and endothelial populations of the mouse at day 10, which was not as apparent the other way around. Of key interest was the signature overlap in the alveolar epithelial compartment. Akin to the macrophage phenotypes, the AT2 cells in human displayed slightly higher correlation to the activated AT2 cells in mouse, likely due to their up-regulation of inflammatory cytokines and senescence markers. Particularly the correspondence of the injury-induced $Krt8^+$ ADI to the ILD-specific aberrant basaloid cells served as a first validation of the hypothesis posed at the beginning of this chapter. Motivated by their similarity, a more fine-grained comparison was performed and is detailed in the latter part of this section.

In the mouse model the injury is at least partially reversible and does not require independent intervention, which is a major divergence from the human disease. The data also reflected lower cell type signature similarity during the resolution phases after 3 weeks of instillation. It has to be noted that the mouse data was retrieved with much lower transcript coverage, and many smaller populations could have been under-represented or missed. For the cell type signature establishment only populations with at least 20 cells at the time point were considered, to comment on the fact why some populations are not represented in the heatmaps. Especially fibroblast cells were heavily diminished towards the end of the experiment, concurring with the decrease of ECM deposition and regeneration at later time points.

3.2.6 Shift in cellular communication towards ILD-induced populations

Similar to the mouse data set, a cell type specific differential gene expression analysis should reveal that not only cell type frequencies are altered in disease, but also their gene expression is affected at variable extents across the cell states. Again, `diffxpy` was used to test for differences between end-stage lung fibrosis and control tissue, while accounting for study cohort and difference in sequencing depths. The results were quantified by displaying the number of up- and down-regulated genes in the heatmaps in Fig. 3.16a, stratified by cell type and effect size. Overall, the higher sequencing depth achieved by 10x and greater sample size, owing to the integration across multiple cohorts, resulted in a much more granular and detailed description of the differentially regulated genes compared to the previous Drop-seq mouse experiment. This level of resolution was carried over to the cellular interaction analysis presented at later stages.
The effect sizes of the induced signatures resembled the frequency changes as established during bulk deconvolution, showing major up-regulation of genes in fibroblast and airway epithelial populations, and down-regulations most prominently in alveolar epithelial and capillary cell types. The induced gene expression changes in ILD were most similar in cell types within the respective epithelial, mesenchymal, and leukocyte lineages, meaning that the up-regulated genes in fibroblasts for instance are more likely to be also regulated in other mesenchymal cells, rather than in leukocytes and vice versa (Fig. 3.16b). Full names and annotated functions for a selection of genes can be looked up in the appendix (Table 5.1). Up-regulation patterns in the most affected cell types generally confirmed many shifts as they are described in the field.

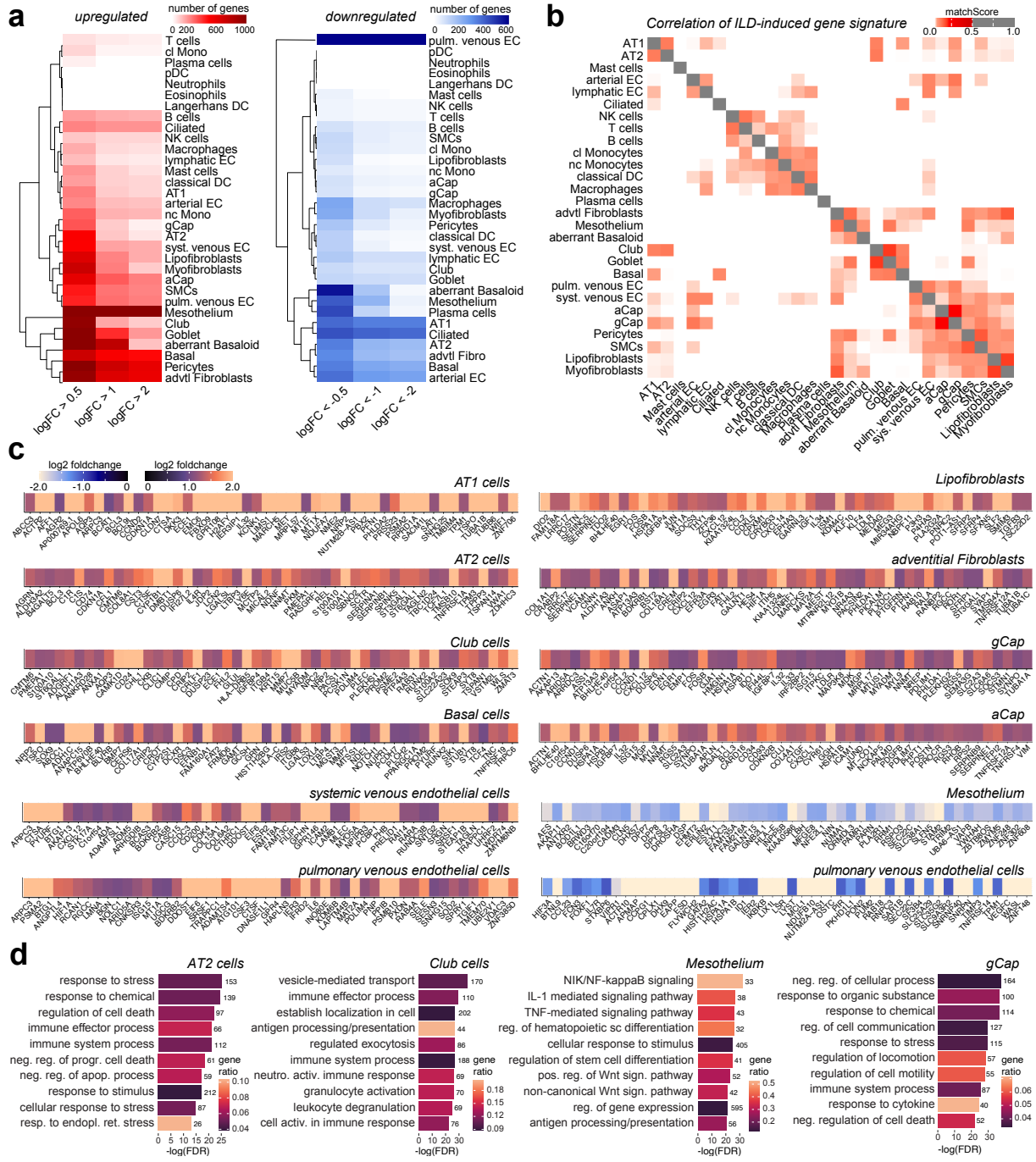


Figure 3.16: Disease induces cell type specific shifts in the transcriptomic space. **a** Heatmaps summarizing differential gene expression results for each cell type to control patients. Colour indicates number of up-regulated (left) and down-regulated genes (right). Columns correspond to number of regulated genes at indicated logFC cut-offs to reflect effect size of differential expression. **b** Heatmap demonstrates similarities of gene expression changes calculated by `matchScore` on the log fold changes, considering genes up-regulated in ILD with logFC > 0.5 in indicated cell type. **c** ILD-induced gene signatures, displaying log fold changes for top 50 genes in most affected cell types. **d** Selection of terms that were significantly enriched after GSEA on the ILD-induced signatures (adj. p-value < 0.05, logFC > 0.5) in indicated cell types. Number of up-regulated genes used as input were 508 (AT2), 912 (Club), 1897 (Mesothelium) and 337 (gCap).

Specifically, chronic diseases feature a cellular senescence signature, which has been increasingly recognized as an important contributor to aging-related diseases like ILD.[260] Up-regulated senescence markers in alveolar cells were the kinase inhibitors CDKN1A, CDKN2A, encoding for p21 and p16, as well as the Cyclins D1 and 2 (CCND1, CCND2). Furthermore, an increased expression of Wnt pathway ligands has been observed in lung tissues of IPF patients compared to donor lungs. The Wnt pathway regulates stem and progenitor cell function and most likely contributes to lung injury and repair. A partial reactivation in adult tissues following injury might contribute to pathogenesis of chronic lung diseases.[269] Differential expression analysis reflected the literature as certain players of the Wnt pathway, such as the transcription factor TCF7L1 and the Wnt Ligand Secretion Mediator WLS, were among the most up-regulated genes in the epithelial cell types (Fig. 3.16c).
Interestingly, further mediators of the Wnt pathway were listed in the AT1 signature, such as the Wnt Inhibitory Factor WIF1 and the Dickkopf Wnt Signalling Pathway Inhibitor DKK3. DKK proteins have already be shown to be higher expressed in IPF lung epithelial cells and are hypothesized to affect impaired epithelial injury and repair processes.[270]

Fibroblast populations were contributing to this pathway as well, marked by their up-regulation of soluble frizzled-related proteins (SFRP1, SFRP2, SFRP4), which function as modulators of Wnt signalling, or by the increase of ROR1 transcripts, a known receptor for WNT5A.[271] Much like the activated stromal populations in injured mouse lungs, fibroblasts from ILD patients displayed significant enrichment of genes associated with collagen (CTHRC1, COL1A1, COL5A1, COL6A3, COL8A1), ECM (MMP7, TGFB3, POSTN) and cell adhesion (EPHA3, VCAM1, FAT1, TNFRSF12A). The induced gene signatures in both aberrant basaloid cells and myofibroblasts overlap largely with their cell type signatures as described in 3.15c. As these ILD-induced cell states are not present in control lungs, epithelial and fibroblast populations from control donors were considered as baseline during differential expression testing. This procedure yielded similar results as cell type marker signatures, for which all remaining cell types in the lung are commonly chosen as reference group.
Capillary cells appeared to mirror the pathological processes in the alveoli, as certain senescence markers (CDKN2B, CCND1, MAP3K8) were higher expressed as well. Additionally, many genes involved in cell adhesion (CX3CL1, RHOB, CD34, CD93, ICAM1, IGFBP7, POSTN, TNFRSF12A) and cytokines for inflammatory or immune responses (CXCL12, CCL2, IL32, IL33) were up-regulated in diseased patients, implicating a communication route towards immune cells. Interleukin-32 for example is known to be elevated in various inflammatory autoimmune diseases and interacts with other inflammatory cytokines such as TNF-α, IL-1β and IFN-γ.[272] Looking at the complete induced gene signatures from a higher perspective, genes involved in the immune system, stress response and apoptotic processes were among the significantly enriched terms (Fig. 3.16d).

In the next analysis the functional consequences of the induced expression patterns were augmented with prior information in the context of all cell types in the lung. Analogous to the mouse data, the up-regulated genes (logFC > 0.5 and percentage of expressing cells in the relevant cell type $> 10\%$) were mapped against a known receptor-ligand data base. The resulting communication patterns are expected to be driven by the number of genes that were differentially expressed. Nevertheless, a higher number of induced genes

is equated with a greater transcriptomic shift in the respective cell type upon disease, and was therefore not eliminated prior to the matching. Owing to the increased sequencing depth, the total number of interaction pairs across cell types was considerably higher compared to the Drop-seq mouse experiment.

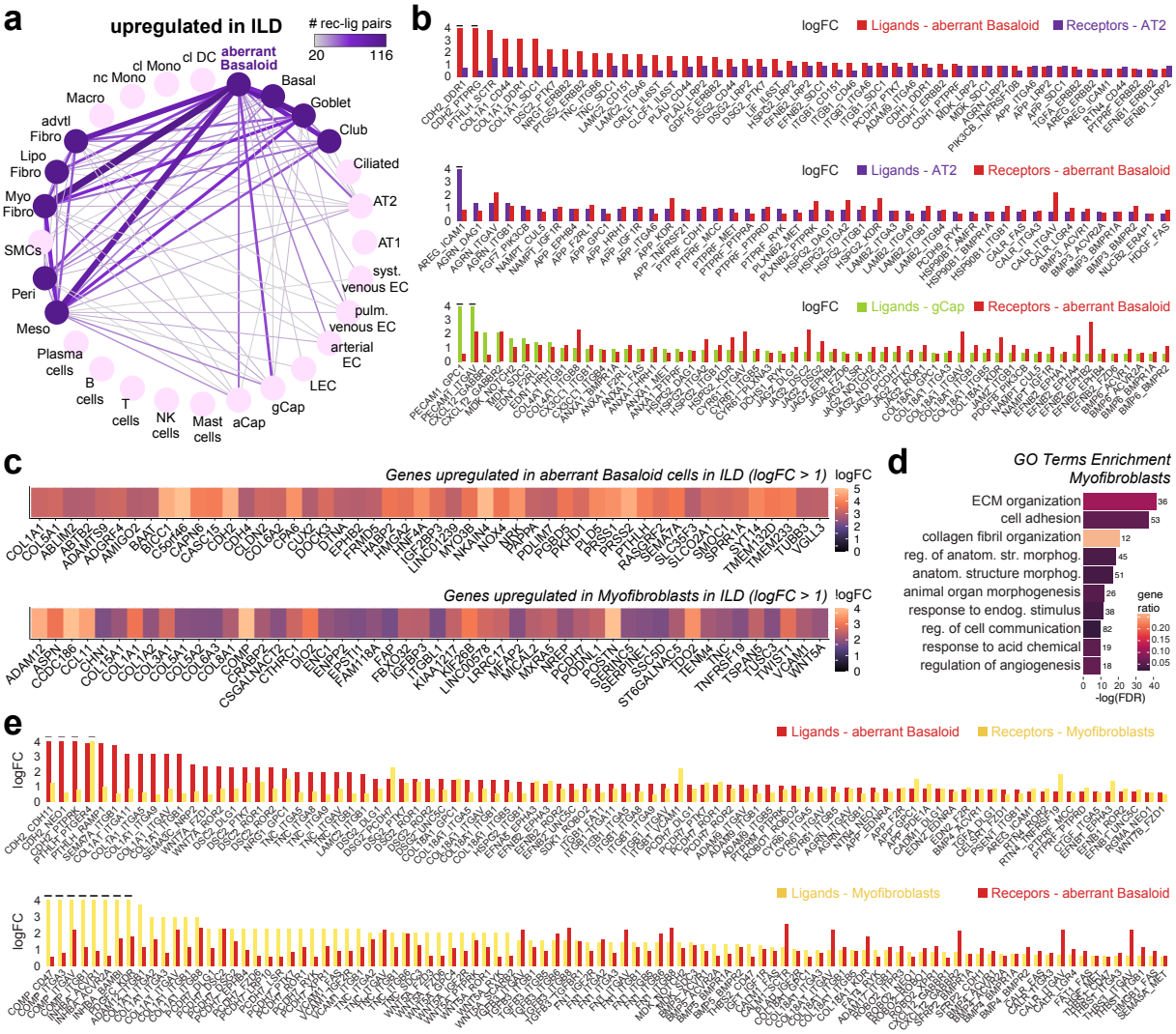


Figure 3.17: Cellular communication drastically shifts towards aberrant disease-specific cell populations. **a** Connectome highlighting induced signalling routes in ILD. Edge colour and width reflects number of receptor-ligand pairs between the cell types. Notably, pairs are only considered if both the receptor and ligand are up-regulated in ILD (logFC $>$ 0.5). **b** Barplots of logFC values derived from differential expression analysis control vs. ILD. Receptor-ligand pairs in cell types of interest for the most prominent edges in the network in (a) are shown. **c** ILD-induced gene signatures, displaying logFCs for the top 50 genes in ILD-induced cell states. During differential gene expression testing in aberrant basaloid cells, remaining epithelial cells from control patients were used as background. Likewise for myofibroblast signature control pericytes, adventitial- and lipofibroblasts. **d** Selection of terms that were significantly enriched in the myofibroblast signature (adj. p-value $<$ 0.05, logFC $>$ 1, 249 up-regulated genes as input). **e** LogFC barplots of receptor-ligand pairs as in (b), focusing on the pronounced communication between aberrant basaloid and myofibroblast cells.

Nonetheless, some interesting patterns seemed to be conserved in the human in vivo reference. The most striking observation was the pronounced communication between epithelial and stromal populations, particularly among disease-specific cell states (Fig. 3.17a). Intriguingly, this mirrored the picture of bleomycin treated mice at day 14 (Fig. 3.8d). The highest number of interaction pairs was between myofibroblasts and AT2 cells, which included the intermediate populations during differential testing and mostly reflected the $Krt8^+$ ADI signature at this time point. The display of multiple edges from and to human fibroblast populations could be attributed to the fact that fibroblast activation does resemble a sliding scale of activation rather than an instant switch. This makes it difficult to set boundaries as fibroblasts proliferate and transdifferentiate towards myofibroblasts and also plays into the overlap between their disease-induced expression changes and the myofibrotic signature.

The molecular mechanisms and origin of these myofibroblasts have been the focus of many studies, but no definite answer has been reached yet. Additional to the many fibroblast populations that are proposed as source population, studies have shown that pleural mesothelial cells have the potential to migrate into the pulmonary parenchyma in IPF and transition into myofibroblasts. Specifically, in a mouse model of lung fibrosis the mesothelial cells underwent phenotypic transition to myofibroblasts in response to stimulation with the profibrotic mediator TGF-β in vitro.[273] Moreover, pleural mesothelial cells could be found in parenchymal cells of explanted lung tissues from 16 IPF patients, providing further evidence for their potential to traffic into the lung and contribute to the myofibroblast population during lung fibrosis.[274]
The pleural mesothelium is derived from the embryonic mesoderm, whose interactions with the endoderm by paracrine signals such as Wnt/β-catenin, Bone Morphogenetic Protein BMP4, Sonic Hedgehog SHH, and the Fibroblast Growth Factor FGF10 are essential during lung development. It has been proposed that the mesothelium may be involved in lung injury-repair by reactivation of such developmental programs in adult, which seems to be dysfunctional in diseased individuals.[275] The integrated data also pointed towards an up-regulation of genes associated with developmental programs such as the NIK/NF-κB or the Wnt signalling pathway in pathological mesothelial cells (Fig. 3.16d) and an overall strongly increased expression of potential ligands towards the epithelial compartment in ILD.

To supplement the edges in the connectome plots, the specific interaction pairs for cell types of interest are listed in Fig. 3.17b,e and their description can be looked up in the appendix (Table 5.1). The AT2 cells acted as a source of ligands for the aberrant populations, expressing for instance the profibrotic Amphiregulin AREG, the Amyloid Beta Precursor Protein APP and Nicotinamide Phosphoribosyltransferase NAMPT, involved in stress response and aging.
Many integrin types (ITGA3, ITGA6, ITGB1), membrane receptors known to be involved in cell adhesion and tissue repair, were up-regulated in basaloid cells and found corresponding binding partners in AT2 (Laminins LAMB2, LAMC2, transmembrane proteins CD46 CD151, Syndecan SDC1), general capillaries (collagens COL4A1, COL18A1) and most of all in myofibroblasts (TNC, COL1A1, VCAM1, FN1, Thrombospondin THBS1, Cartilage Oligomeric Matrix Protein COMP, and other Integrins such as ITGA8, ITGA9 and ITGAV).

Aberrant basaloid cells further displayed expression of the Notch signalling receptors NOTCH2, NOTCH3, for which corresponding ligands are expressed on gCap cells in particular (Midkine MDK, Jagged Canonical Notch Ligand JAG2). Among the highest up-regulated receptor and ligand pairs of myofibroblast and aberrant basaloid populations were many genes associated with the Wnt signalling pathway, for instance the receptors FZD1, FZD3, FZD6, ROR2, GPC4, RYK. TGF-β is the key player driving myofibroblast differentiation, thus it is not surprising that many genes of the TGF-β superfamily are part of the communication between myofibroblasts (Inhibin Subunit Beta INHBA, Transforming Growth Factor Beta TGFB3).

In the next analysis the increased communication between these injury induced cell states was further explored with `NicheNet`.[232] Ligands were ranked based on their potential to induce the transcriptomic signature of either aberrant basaloid or myofibroblast populations, and then linked back to their cell type origins. For this, only cells from ILD patients were used as input. Motivated by the connectome plots, the following cell types were chosen as potential sending populations: AT1, AT2, ciliated, club, goblet, basal, aberrant basaloid, adventitial fibroblasts, lipofibroblasts, myofibroblasts, pericytes, mesothelium, aCap and gCap. Ligands are shown if they were significantly increased in ILD in at least one of these cell types.

The general picture painted by the up-regulated receptor-ligands pairs was further solidified. For instance, myofibroblasts and aberrant basaloid cells greatly affected the transcriptomic signatures in both directions, while also featuring heavy autocrine signalling (Fig. 3.18). Many of their induced genes appeared to lie downstream of the Wnt pathway, or are directly affected by it. In both populations certain ligands of this pathway, namely WNT7A, WNT7B, WNT9A, WNT10A, SFRP2, DKK1, influenced a remarkable portion of the target genes. Many of these ligands were induced by the disease in aberrant basaloid cells, myofibroblasts, and also the mesothelium. The latter further up-regulated the Fibroblast Growth Factors FGF1 and FGF2, which are among the top prioritized ligands predicting the myofibroblast signature. Other notable ligands regarding aberrant basaloid cells would be SFRP2, TGFB3 and TNC, major markers of the myofibroblast population. Likewise TGFB2, Cell Adhesion Molecule CADM1, Prostaglandin-Endoperoxide Synthase PTGS2, related to NF-κB signalling, were prominently up-regulated in the basaloid cell state and were associated with the transcriptomic shifts in myofibroblasts. Additionally, other fibroblast populations and myofibroblasts themselves provided potent ligands (ITGFB3, SFRP2, Insulin Like Growth Factor IGF1, Calreticulin CALR).

On a final note regarding cell-cell communication, these signalling patterns are predictions based on prior knowledge on interacting pairs and induced transcriptomic shifts stratified by cell type. They appeared to be supported by the established knowledge in the field, but were not experimentally validated at this point. As cellular signalling is highly condition and cell type specific, further experiments should ideally be conducted to provide final functional proof for these processes, but were unfortunately beyond the temporal scope of this work.

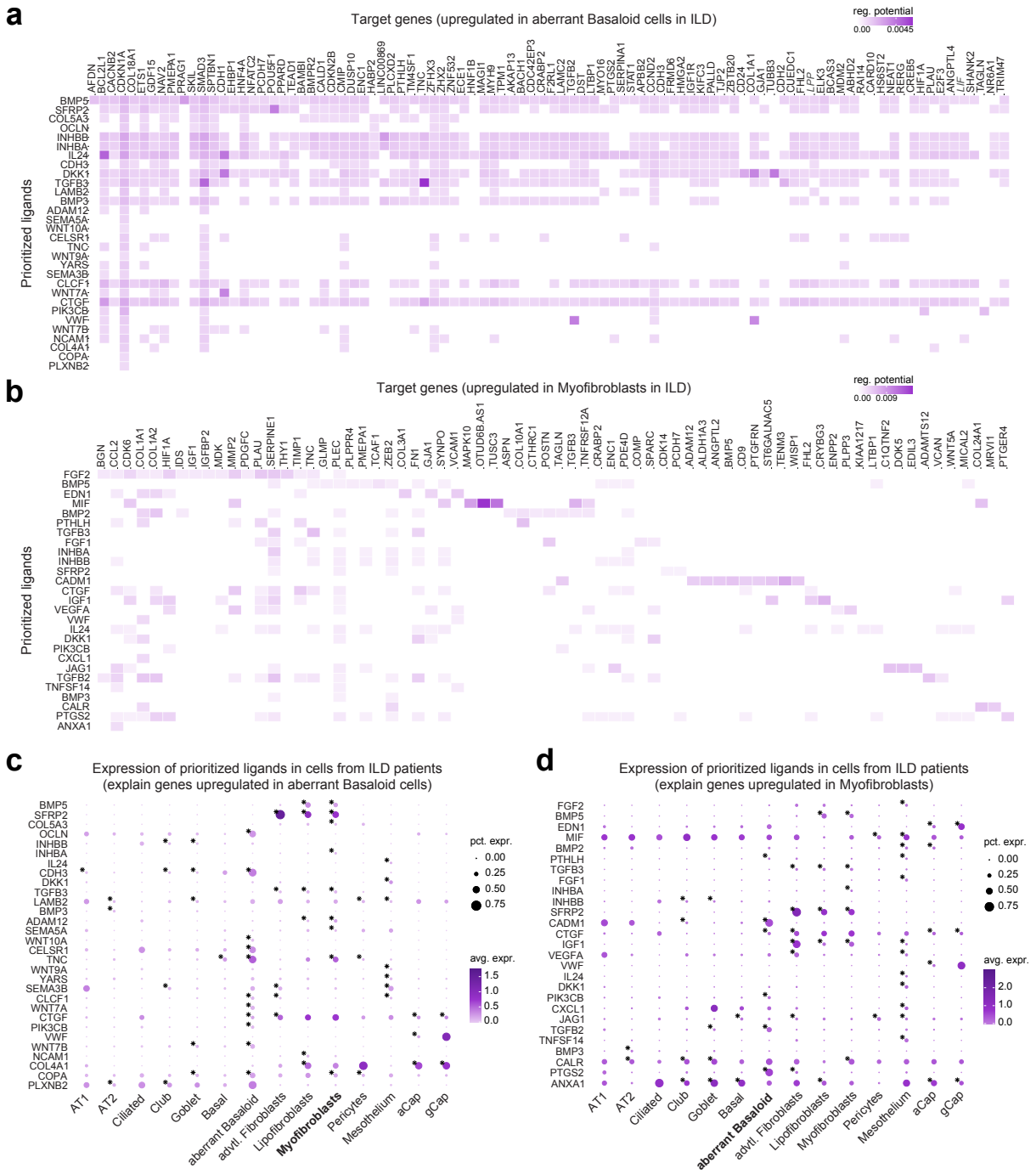


Figure 3.18: Zoom into inter-cellular signalling between ILD-induced cell states. a, b Regulatory potential of ligands explaining up-regulated gene signature in disease-specific aberrant basaloid cells (a) and myofibroblasts (b). During differential testing injury-induced cells were compared to healthy epithelial and fibroblast populations, respectively. Prioritized ligands are pre-selected for those up-regulated in at least one of the epithelial, mesenchymal or endothelial cell types. Top 40 of these ligands ranked by `NicheNet's` Pearson correlation are displayed. **c, d** Dotplots visualizing the expression level and percentage of prioritized ligands implicated in gene expression shifts in aberrant basaloid (c) and myofibroblasts (d) in pathological conditions. Asterisk indicates significant up-regulation of the gene in ILD patients compared to healthy donors. For clarity only relevant cell types are shown.

3.2.7 Correspondence of human pathogenesis to regeneration in mouse model

The similarity of the disease-specific cell state to the transient populations in the regenerating mouse epithelium has been teased throughout the chapter. Therefore, to conclude this main passage, the phenotype of the described basaloid cells was put to direct comparison with the Krt8$^+$ ADI. It is not clear how the aberrant state arises in disease, and could potentially be derived from cell populations serving as progenitors for depleted AT1 and AT2 cells. These could be either remaining alveolar cells, or even bronchoalveolar stem cells, which are known to generate both AT1 and AT2 cells in response to bleomycin-induced lung injury.[115,171] Echoing the trajectory modelling approach in the mouse lungs, the integrated data was subset to the alveolar epithelial compartment including the aberrant basaloid cells. Put more precisely, the possible route of de-differentiation of human AT2 cells in ILD was the focus of the next analysis.

Aberrant basaloid cells from the different patient cohorts overlapped in the `BBKNN`-corrected space and their enrichment in ILD patients was visually pronounced in the diffusion map (Fig. 3.19a). These terminally differentiated cells were clearly distinguishable by their respective cell type markers. Unlike in the mouse experiment, no underlying temporal ordering was available for the patients, as the cells could hardly be labelled by progression of the disease. To quantify the gradual decrease in similarity of these cell types, a pseudo time trajectory was derived using AT2 cells as root population. The genes' association to the health state was tested with `diffxpy`, along with the dpt values, data set label and the scaled number of counts as covariates. To decrease the multiple testing burden, only genes expressed in more than 10% of cells in at least one of the cell types were considered. The expression analysis resulted in 5,997 differentially regulated genes (adjusted p-value < 0.05). The gene list was divided into 3 main groups: AT1, AT2 and aberrant basaloid, depending on which cell type showed the highest percentage of expression for each regulated gene. To counteract the differences in sequencing depth and drop-out effects during visualization, the expression values of cells adjacent in pseudo time were averaged. This resulted in 500 dpt bins for AT1, AT2 cells and 20 bins for basaloid cells due to their much smaller cell number. Expression patterns are displayed equally across the spectrum by selecting the most significantly regulated genes separately for each cell type in the heatmaps in Fig. 3.19d. For AT2 the top 40 down- and for the aberrant basaloid cells the top 40 up-regulated genes were selected, while for AT1 cells both the top 25 up- and down-regulated genes are shown, all ranked by adjusted p-values.

The de-differentiation of AT2 cells is best reflected by the down-regulation of surfactant proteins (SFTPA1, SFTPA2, SFTPB, SFTPC, SFTPD) and simultaneous up-regulation of AT1 cell markers (CAV1, AQP4, RTKN2, EMP2) in the corresponding cell types. As anticipated, the association to pseudo time also returned many aberrant basaloid cell state marker (CDH1, CDH2, KRT17), and genes encountered during the previous communication analysis (CDKN2A, DKK3, WNT10A).
For clarity, the temporal expression patterns of selected markers are displayed as separate line plots in Fig. 3.19e as well, putting the up- and down-regulations of the respective gene programs into context with the other alveolar cell types. To emphasize the similarity of the signatures, the markers of the murine intermediate state as selected in Fig. 3.12f were inspected as well. The human homologues of the most prominently used markers Krt8, Sprr1a and Lcn2 showed an interesting behaviour in the AT2 and basaloid subset.

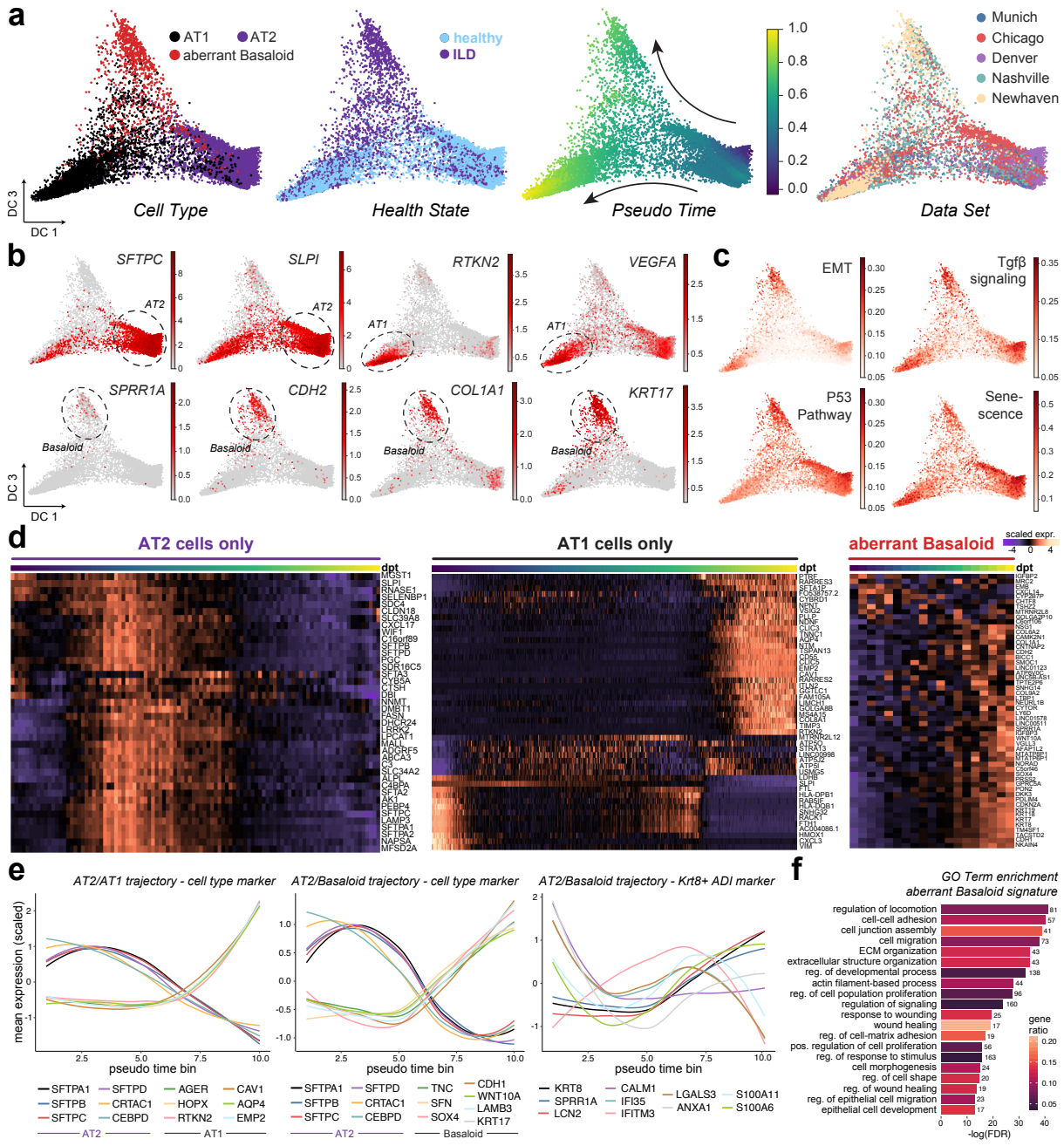


Figure 3.19: Differentiation trajectory modelling of AT2 either AT1 or aberrant basaloid state. **a** Integrated diffusion map of alveolar epithelial cells and aberrant basaloid cells from 5 disease cohorts, coloured by cell type, health state, pseudo time and cohort. **b, c** Diffusion map coloured by prominent cell type marker (b) and gene expression scores for indicated signatures (c, retrieved from MSigDB Hallmark gene sets). **d** Heatmap showing scaled expression in selected cell types across pseudo time. Most significant genes were selected by ranking their association to pseudo time. The top gene list was clustered ($k = 3$) and binned expression values of top 40 genes per cluster are displayed in corresponding cells only. **e** Line plots illustrating scaled expression levels of AT1/AT2 cell type markers (right), AT2/aberrant basaloid markers (middle) and selected markers from the intermediate state in the mouse trajectory (Fig. 3.12f, right). Pseudo time is binned ($n = 10$) to average expression values of adjacent cells and smoothed for clarity. **f** Selected terms that were significantly (adj. p-value < 0.05) enriched in the aberrant basaloid signature compared to all other lung cell types.

While the expression of these genes was generally low in AT2 cells, it gradually increased along the pseudo time and persistently stayed high in aberrant basaloid cells. Not all genes that peaked during the $Krt8^+$ ADI state showed this pattern, but many key genes did follow the trend, notably SOX4, S100A6, TNC and SFN.

Scoring the alveolar trajectory in mice for the induced gene signature in aberrant basaloid cells (111 genes with logFC $>$ 2 in expression test healthy vs. ILD) further demonstrated their phenotypic similarities, as the correlation score was highest for cells belonging to the $Krt8^+$ ADI cluster. Conversely, aberrant basaloid cells displayed the highest correlation to the $Krt8^+$ ADI signature from the mouse data (62 genes from cell type marker table with logFC $>$ 3). A comparison of epithelial cell type marker in both organisms in general confirmed that the transient state from mice was most similar to the disease-induced basaloid cells in human ILD patients (Fig. 3.20a,b). Here, the cell type marker tables were calculated on only ILD patients and bleomycin-treated mice after day 10, as the fibrotic phenotype was expected to align best at this time point.
Finally, it was validated whether $KRT8^+$ alveolar cells can also be observed in human acute and chronic lung disease independent of the scRNA-seq data. Human tissue sections were co-stained for SFTPC and KRT8. No expression of KRT8 was detected in the alveolar space of non-fibrotic control lungs (n = 9). In sharp contrast, strong alveolar expression of KRT8 was measurable in lung sections from Influenza-A induced ARDS (n = 2) and IPF patients (n = 5). In a second round of immunofluorescence analysis, the specific marker for both $Krt8^+$ ADI and basaloid cells SPRR1A was included. Indeed, $KRT8^+SPRR1A^+SFTPC^-$ cells were located in close proximity to $SFTPC^+$ cells in fibrotic areas exclusive to IPF lungs, potentially undergoing fibrotic remodelling and representing early stages of the disease (Fig. 3.20c,d).

Conclusion

In summary, the herein collected scRNA-seq data allowed for the modelling of cellular dynamics over a four week time-course after bleomycin lung injury and, together with independent external data resources and experimental validation, proved the transient appearance of a squamous $Krt8^+$ alveolar differentiation intermediate en route to regeneration. Gene expression analyses based on real time points revealed transcriptional regulators that drive these expression dynamics and implicated the transient population as a specific source of profibrotic mediators such as Ctgf, Itgb6, Areg, Edn1 and Lgals3. Still, their functional role in the context of alveolar regeneration remains unclear. Corresponding human transcriptomes of patients with the very disease the bleomycin mouse model aims to reflects were searched for a comparable population. The recently described aberrant basaloid cells showed remarkable similarity, most pronounced by a number of overlapping cell type marker as well as their shared communication pattern in the form of strongly increased interaction between epithelial cell states with the ILD-induced myofibroblast populations.
Together with the spatial location, these observations indicated that aberrant basaloid cells might be the result of AT2 cells undergoing regenerative repair but failing along the way, owing to the pathological milieu that promotes defective cell differentiation, which in addition to increased cellular signalling could lead to the aberrant persistence of the typically transient regenerative cell states. Further exploration of functional implications in the context of contemporary findings will be continued in the discussion section.

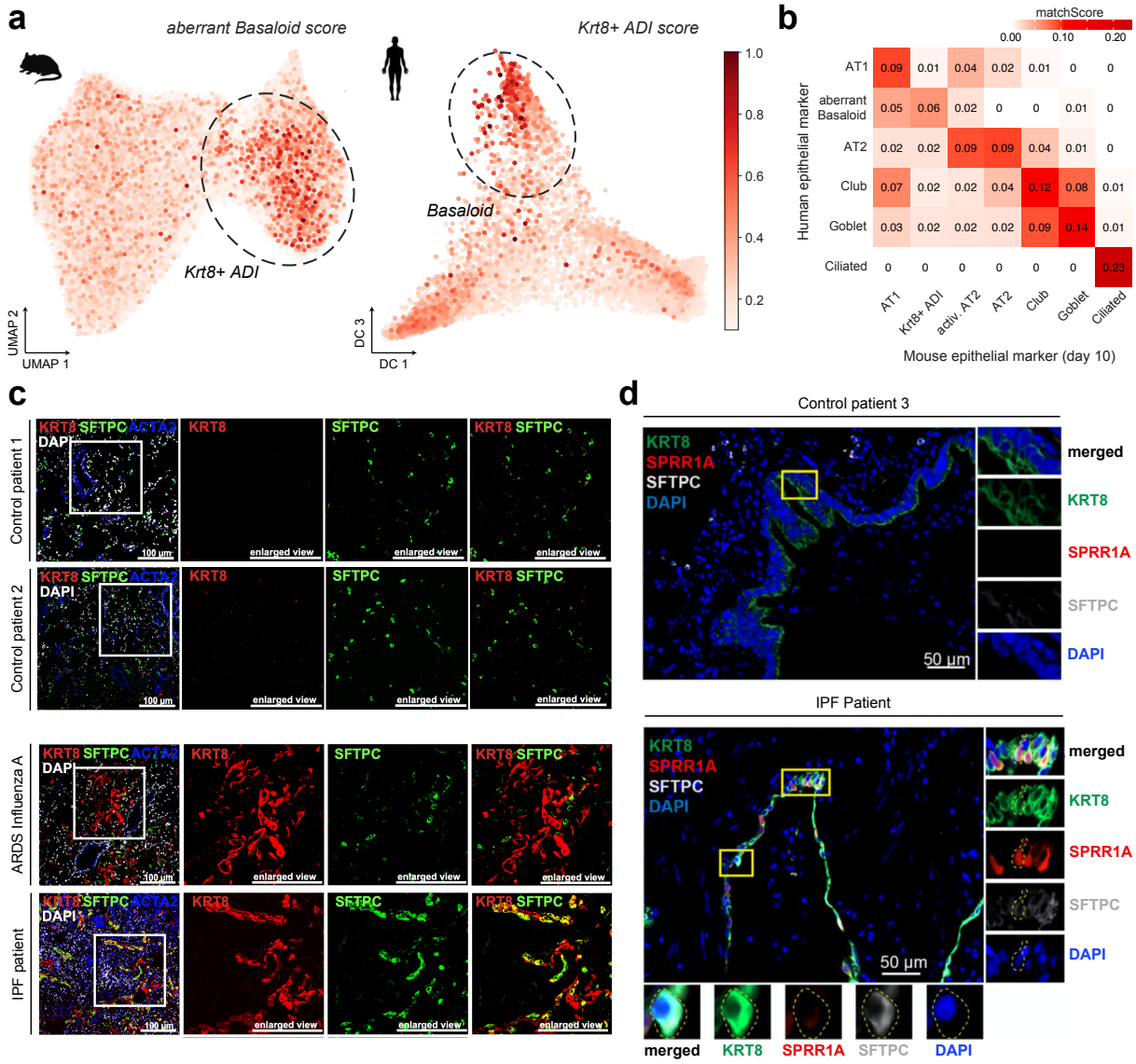


Figure 3.20: Cells similar to Krt8$^+$ ADI in mouse lungs persist in human ILD patients. **a** UMAP of alveolar mouse epithelium, overlaid with correlation score based on the gene signature of aberrant basaloid cells in human (derived from differential expression control vs ILD, logFC $>$ 4). Diffusion map of integrated human alveolar cells overlaid with correlation score to Krt8$^+$ signature (derived from differential expression PBS vs bleomycin, logFC $>$ 2). **b** Quantification of marker signature similarity across species using the Jaccard index-based `matchScore`.[268] For the human counterpart, the cell type signatures derived from ILD patients only was used, whereas the signatures for mice was calculated on bleomycin-injured mice on day 10 after injury. For both signatures, genes with logFC $>$ 4 were considered in the comparison. **c** Human lung tissue sections from non-fibrotic controls as well as ILD patients (fibrotic regions of ARDS, IPF diagnosis) were stained against indicated proteins (scale bar = 100 µm). Pronounced KRT8 expression at the site of acutely injured lesions is apparent (scale bar = 100µm). **d** Immunofluorescence analysis of SPRR1A, KRT8 as well as SFTPC in IPF and control samples. KRT8$^+$SFTPC$^-$ cells are located in close proximity to areas of limited fibrotic remodelling.

3.3 Exploration of pathological patterns found in COPD and COVID-19 lungs via mouse model

Two smaller side-projects spiralled out of another Drop-seq experiment based on a time-series mouse lung data set. As both of these projects demonstrated the importance of collaborative efforts and addressed contemporary issues, and more importantly fit with the general trends presented in this work, I wish to include these as a shorter chapter.

In this study, already established experimental data indicated an important role of the LTβR signalling in the pathogenesis of COPD. Measurements of bronchoalveolar lavage (BAL) fluids from a corresponding mouse model and quantitative PCR (qPCR) on lung biopsies from human COPD patients showed an increased expression of signalling molecules and ligands associated with that pathway. However, the cellular origins cannot be delineated by such ensemble measurements. ScRNA-seq was performed in order to increase the cellular resolution and explore the cell type involvements in greater detail. Due to unforeseen circumstances, the data set at hand provided a suitable resource for another project. After large-scale integration of gene expression data from more than 30 human data sets, a cell type-specific associations to smoking status was predicted for the three SARS-CoV-2 entry factors ACE2, TMPRSS2 and CTSL. Correspondence of the gene expression shifts in smoke exposed mice from the initial project to the observed patterns in humans is assessed in the second part of this chapter. The results reported in this chapter have contributed to the following publications: Conlon et al. accepted and published in Nature in December 2020[172] and Muus C et al. accepted and published in Nature Medicine in March 2021[167] within the framework of the Human Cell Atlas Lung Biological Network.

Experimental data planning and sample collection, mouse handling, proteomics quantification, FACS-sorting, immunohistochemical stainings, respective statistical analysis among others were performed by Thomas Conlon and colleagues. Experimental scRNA-seq profiling was performed by members of the Schiller Lab, while the computational analysis of the collected data was within my area of responsibility. The original project has been ongoing for years before my arrival. To narrow down the results, only relevant parts of the experimental analyses in line with my computational contributions will be listed in the following chapter.

Introduction

The immune response primarily aims to eradicate dangerous pathogens invading the organism. However, if the pathogen is constantly replenished as in case of continuous smoking, the sustained immune response can ultimately result in tissue damage and chronic inflammation. COPD is characterized by such an enhanced inflammatory response, mostly attributed to long-term exposure to toxic gases and particles. Features of its pathology are the destruction of the gas-exchange areas in the lung, chronic inflammation leading to remodelling of the small airways, and excessive mucous production that contributes to airway obstruction. The disease progression has been linked to the extent of the inflammatory response, as reflected by the number of acute inflammatory cells and lymphocytes that infiltrate the airways.[138] COPD lungs have increased numbers of

T and B cells, which are often organized into tertiary lymphoid follicles referred to as inducible bronchus-associated lymphoid tissue (iBALT). They appear both in severe human COPD[276] and its animal models based on chronic smoke exposure.[277] The formation of iBALT is also known as lymphoid neogenesis, and is promoted by Lymphotoxin β-receptor signalling. Activated lymphocytes express the TNF superfamily members Lymphotoxin-α LTA and -β LTB, which in turn interact with the Lymphotoxin β-receptor LTβR on stromal organizer cells during chronic inflammation.[278]

The interaction of LTA, LTB and TNF with their respective receptors triggers downstream non-canonical NF-signalling. Members of the NF-κB family of transcription factors regulate expression of genes crucial to immune responses, cell growth, and apoptosis, and activation of this pathway results in up-regulation of genes involved in inflammation and lymphoid organogenesis.[279] Specifically, the stimulation of LTβR triggers the release of chemokines such as CXCL13, which for example attracts B cells into the lymphoid follicles and creates a positive feedback loop maintaining LTA, LTB expression on lymphocytes.[277] The signalling cascade further induces the expression of adhesion molecules in endothelial cells (e.g. Intercellular Adhesion Molecule ICAM1, Vascular Cell-Adhesion Molecule VCAM1, Peripheral Node Addressin PNAD), that likely initiate mononuclear accumulation.[280]
Due to its prominent role in the development of tissue injury, the question arises whether therapeutic inhibition of LTβR signalling could hamper or even disrupt the formation of iBALT structures. To understand the gradual formation of these follicles, an appropriate mouse model is necessary. The abnormal inflammatory response in the lungs during disease pathogenesis can be induced via cigarette smoke exposure of mice.
In the early acute reaction during the first week of exposure, predominantly neutrophils and macrophages show a strong influx into the lung. After one month the response shifts towards a progressive inflammation, characterized by the additional recruitment of lymphocytes. Long term exposure, up to 4 months or longer, has been shown to recapitulate pathological features present in COPD, such as small airway remodelling and tissue damage in form of emphysema.[281]

In the following section, the increased expression of LTβR ligands in adaptive and innate immune cells, enhanced non-canonical NF-κB signalling, and enriched LTβR target gene expression in lung epithelial cells from COPD patients as well as in mice chronically exposed to cigarette smoke, is demonstrated. Finally, the effects of LTβR signalling inhibition are explored.

3.3.1 Inhibition of LTβR-signalling disrupts iBALT formation

In order to study COPD pathogenesis, mice were whole-body exposed to mainstream cigarette smoke (CS). To mimic natural human smoking habits, the exposure lasted 50 minutes per day, for up to 6 months. Baseline control mice were kept separately in a filtered air environment, but were exposed to the same stress. To explore the therapeutic implications of LTβR signalling inhibition, a sub-group of smoke-exposed mice was treated with either LT-Ig fusion protein or control-Ig (MOCPC21) for two months, starting from 2 and 4 months after CS. For each time point, 3 control and 5 CS exposed mice were sacrificed. After 6 months, 5 additional LTβR-treated mice were added and the transcriptomic profiles of a total of 28 samples were measured with Drop-seq.

Regarding the computational analysis, the pre-processing and quality control was carried out as described in the methods section, the exact filtering criteria and parameters for the analyses can be found in section 5. One striking observation unique to this data set was that even after excluding cells with a mitochondrial fraction of more than 20%, one cluster highly enriched for cells with relatively high number of mitochondrial transcripts still remained (Fig. 3.21). Cells belonging to this cluster featured markers of every main lineage, thus this cluster was designated as *low quality mixture cluster*. However, it is interesting to note that the composition of this cluster appeared to be driven by the duration of cigarette smoke exposure (Fig. 3.21b). The longer the exposure period of a given sample, the higher the fraction of cells assigned to the mitochondrial enriched cluster. Reactive oxygen species (ROS) are well-known in the field and are used as an important determinant of cancer risk. Tobacco smoke induces oxidative stress by creating such ROS, to which mitchondria are highly susceptible. Mitochondrial DNA lacks protective histones and has low repair capacity, instead they increase their copy number to compensate for damages.[282] Interestingly, increased mitochondrial copy number in circulating blood mononuclear cells of smokers has been described, which additionally correlated with the number of cigarettes smoked per day.[283]

Here, the increased mitochondrial percentages could reflect the stress-induced cellular damage with increased exposure periods. As this cluster contained potentially damaged cells and would have been difficult to delineate, it was removed from further analysis. Before exploring the single cell data, the experimental observations that motivated the transcriptomic analysis are summarized. To evaluate the iBALT formation and the effects of LTβR signalling inhibition in response to CS exposure, sections from mouse lung were stained for $B220^+$ B cells and $CD3^+$ T cells, the main constituents of the lymphoid follicles.

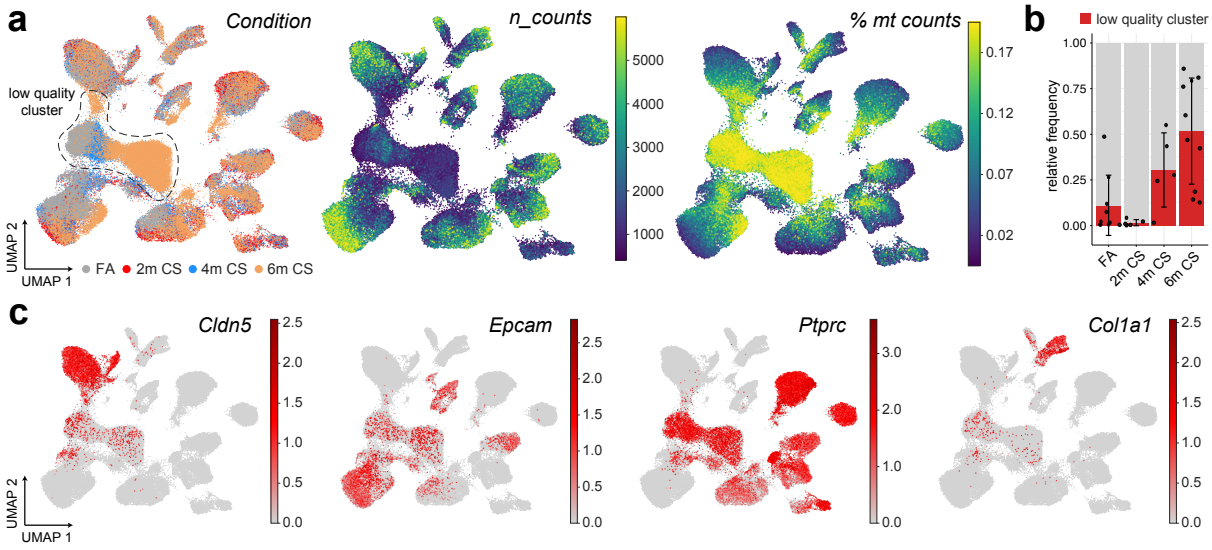


Figure 3.21: Low quality cluster exhibits expression of a mixture of lineage markers in smoke exposed mice. **a** UMAP coloured by exposure condition, number of counts and percentage of mitochondrial reads after initial filtering of low quality cells. The low quality cluster is defined based on their higher mitochondrial fraction. **b** Relative amount of cells from each mouse being assigned to the low quality cluster, with error bars. **c** Feature plot of common lineage markers, demonstrating mixture of all lineages in the low quality cluster.

While representative images in control mice did not display any cell aggregates positive for the two markers, exposure to CS resulted in the development of iBALT as early as 4 months (Fig. 3.22a,b). After 6 months of exposure, the volume of these aggregates was drastically enhanced. Treatment of the CS exposed mice with the LTβR-Ig fusion protein in both the prophylactic (from 2 to 4 months) and therapeutic setting (from 4 to 6 months) led to considerably reduced iBALT formations and even caused dispersion of the immune cells. Specifically, the number of incidence and volume of iBALT in the airways was significantly reduced in the LTβR-Ig treated mice compared to their MOCP treated counter parts.
The tissues structure was further assessed by haematoxylin and eosin (HE) stainings to skim the lung sections for emphysema formation. Additionally, the extent of airspace damage was quantified in form of airspace enlargement as mean chord length and the alveolar surface area in the lung sections. After 4 months of CS, emphysema was fully established in the mouse lungs. Reduction in LTβR signalling greatly affected the lung pathogenesis. Prophylactic treatment starting at 2 months prevented emphysema formation, and therapeutic treatment starting at 4 months even led to a full restoration of lung tissue (Fig. 3.22c,d). Finally, the amount of airway remodelling can be determined by a quantification of the accumulated collagen around the airways. Indeed, the airway collagen deposition in the CS exposed mice was increased. LTβR-Ig treatment in turn revealed a regression in the smoke-mediated airway remodelling compared to the MOCP-treated mice. The combination of these results strongly suggests a protective or even regenerative role LTβR signalling blockade during iBALT-mediated pathogenesis.

Having established the consequences of LTβR inhibition on lung tissue level, the cellular and molecular key players involved in these mechanisms were of interest. Following quality control, the cells were divided into the 5 main lineages and annotated using marker genes derived from literature and previously published scRNA-seq data sets (Fig. 3.23), resulting in over 62,635 cells from 29 mice across 4 conditions (Fig. 3.22f,g). Mice within the same condition showed good agreement in UMAP and PCA space (Fig. 3.23a, Fig. 3.22h), therefore no further batch correction methods were employed.
Consistent with knowledge in the field, certain cell types exhibited stronger changes in their transcriptomic profile after CS exposure. Noteworthy observations would be the shift towards a more activated state in AT2 cells, alveolar macrophages and neutrophils. Comparable to the bleomycin-injury model, the activated AT2 cell state was marked by up-regulation of Lcn2 and Il-33, and the activated macrophage state likewise by its increased expression of Spp1 and Ccl6 among other secreted cytokines. A recent publication has explored single-cell transcriptomic profiles of sorted neutrophil populations, and established a distinct signature during inflammation in lung tissue.[284] Using this signature as reference, neutrophils were separated into a baseline and an activated state as well.

It should be noted that the gene expression differences for the activated states were not as drastic as in the previous mouse model, due to the smaller magnitude of introduced perturbation here. Frequency analysis validated an enrichment of the activated cell states in the CS condition compared to FA controls. In the lymphocyte compartment, LTβR-Ig treated mice after 6 months of CS displayed a smaller proportion of neutrophils, compared to the MOCP counterparts. Apart from these biologically reasonable exceptions, most of the cell types are represented by cells across all conditions.

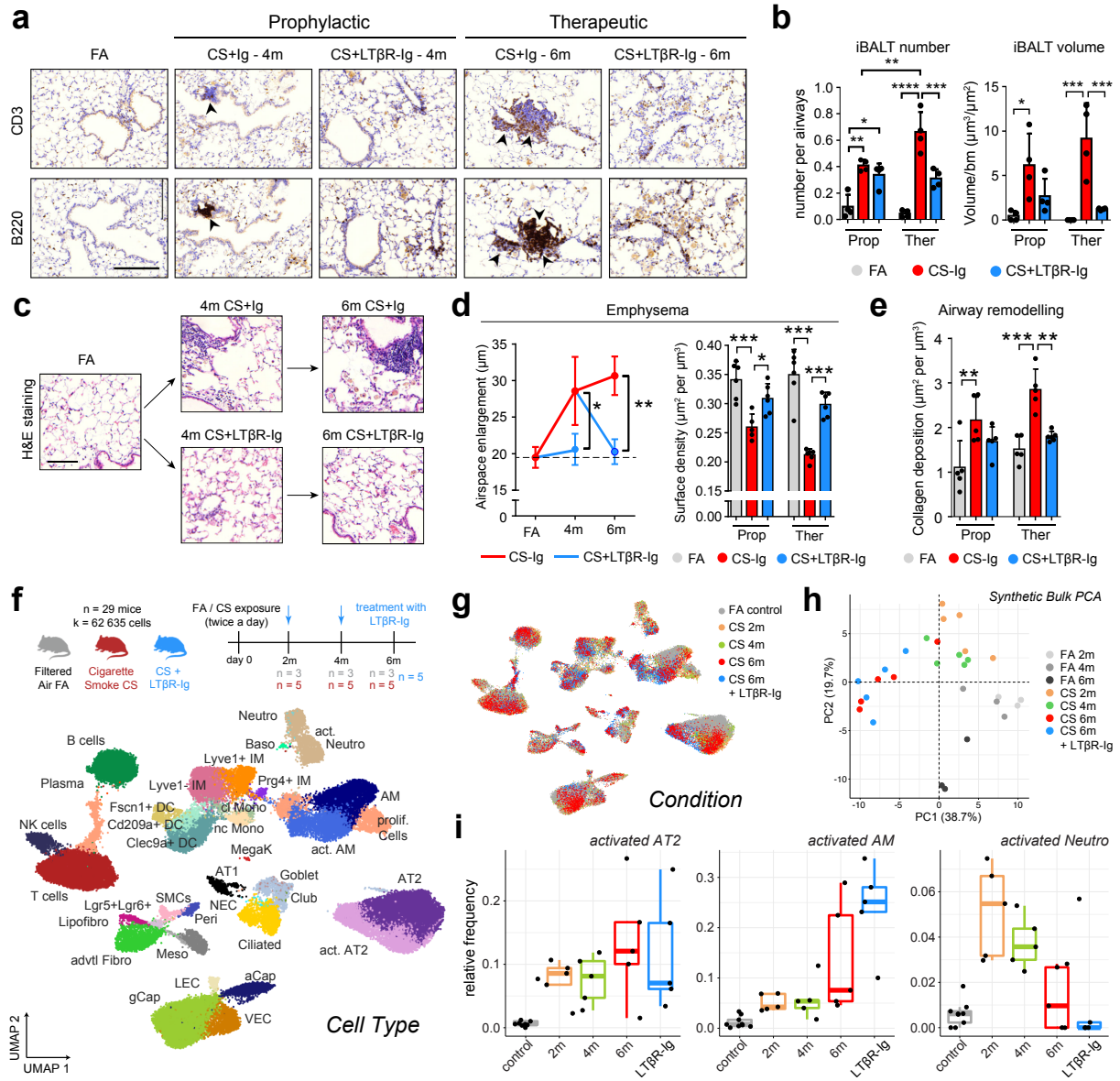


Figure 3.22: Inhibition of LTβR-signalling disrupts iBALT formation in the lungs of mice exposed to cigarette smoke. **a** Representative images of immunohistochemical analysis for B220$^+$ B cells and CD3$^+$ T cells (brown signal) in lung sections from mice exposed to filtered air (FA) or cigarette smoke (CS) for 4 or 6 months. Mice were additionally treated with LTβR-Ig or control from 2 to 4 months (prophylactic) and from 4 to 6 months (therapeutic) (scale bar = 200 µm). **b** Quantification of lung iBALT as mean iBALT number per airway and volume of iBALT normalized to surface area of airway basement membrane. **c** Representative images of haematoxylin and eosin (HE) stained lung sections from mice, split by exposure and treatment group (scale bar = 100 µm) **d** Quantification of airspace enlargement as mean chord length and alveolar surface area in lung sections from mice. **e** Quantification of airway collagen deposition normalized to surface area of airway basement membrane from sections in c. **f** Experimental scheme and UMAP embedding of cells from whole lung suspensions of mice exposed to FA (n = 9) or CS for 2, 4 or 6 months, plus LTβR–Ig or control Ig at 6 months (n = 5 per condition). **g** UMAP coloured by exposure condition. **h** Scatter plot of PCA results on synthetic bulks. The first component corresponds to time, whereas the second component separates the smoke exposed from the control mice. **i** Box plots of relative frequency of activated cell types with respect to all other cell types per individual mouse.

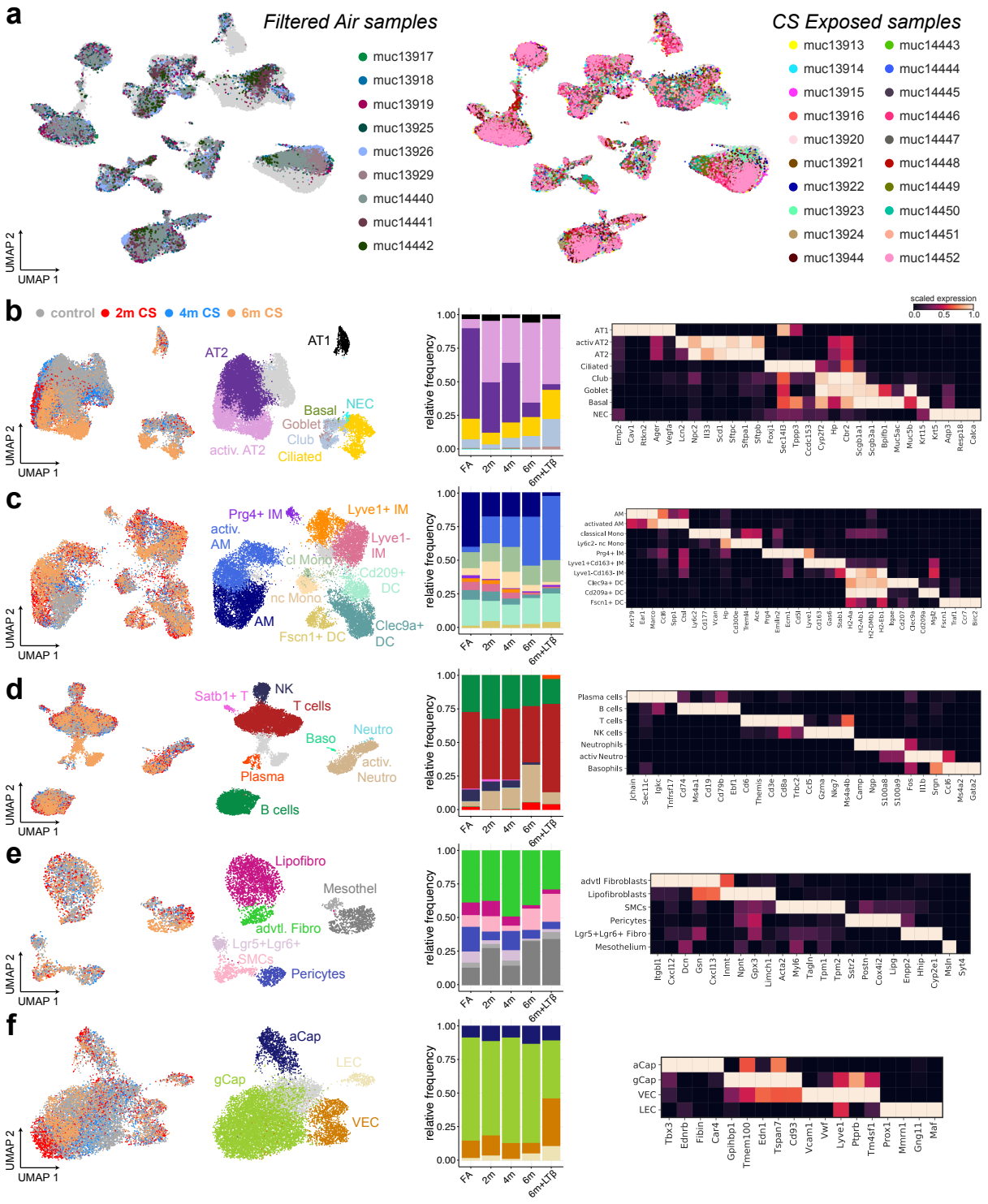


Figure 3.23: Compartment-wise annotation of cells from smoke exposed mice. a Split view of UMAP, separated by samples exposed to filtered air (left) or cigarette smoke (right). **b-f** Cell type annotation, relative composition within compartment and literature-derived marker genes of the epithelium (b), mononuclear phagocytes (c), lymphocytes and granulocytes (d), mesenchyme (e) and endothelium (f). Light gray cells from the cell type coloured UMAPs were estimated to be of low-quality and were excluded from further analysis.

3.3.2 Effect of chronic smoke exposure on epithelial and immune cells

Expanding on the insights from compositional cell type changes, differential gene expression painted a similar picture. The expression shifts induced by smoke exposure were calculated with `diffxpy`[227] and stratified by cell type. The input parameters consisted of the scaled number of counts as covariate and the exposure status as factor to test for. The effect sizes are visualized in form of number of up-regulated genes per cell type in Fig. 3.24a. Most cell types displayed a time dependant up-regulation of increasingly more genes. However, the cell types with most differentially regulated genes were the alveolar macrophages, neutrophils, T cells and airway epithelial cell types. These observations further corroborated the results from the compositional analysis by linking the increase of cells from activated cell types to changes in their transcriptomic spaces. As pointed out earlier, neutrophil and macrophage numbers are increased early on in the CS-exposure model. Lymphocytes are then additionally recruited once the milieu changes towards progressive inflammation in response to the sustained smoke exposure. This pattern of delayed activation was also evident regarding alterations in gene expression space, as T cells, ciliated and club cells show a striking increase in numbers of up-regulated gene in mice during later time points of sustained exposure.

At this level of inspection it appeared that LTβR signalling inhibition returned the transcriptomes of endothelial cells closer to baseline, as the number of up-regulated genes resembled the results from 4 month mice more closely than the ones from 6 month mice. Gene set enrichment analysis with `GOAtools`[230] allowed to assess the functional relevance of the induced gene lists at the last measured time point of CS exposure.

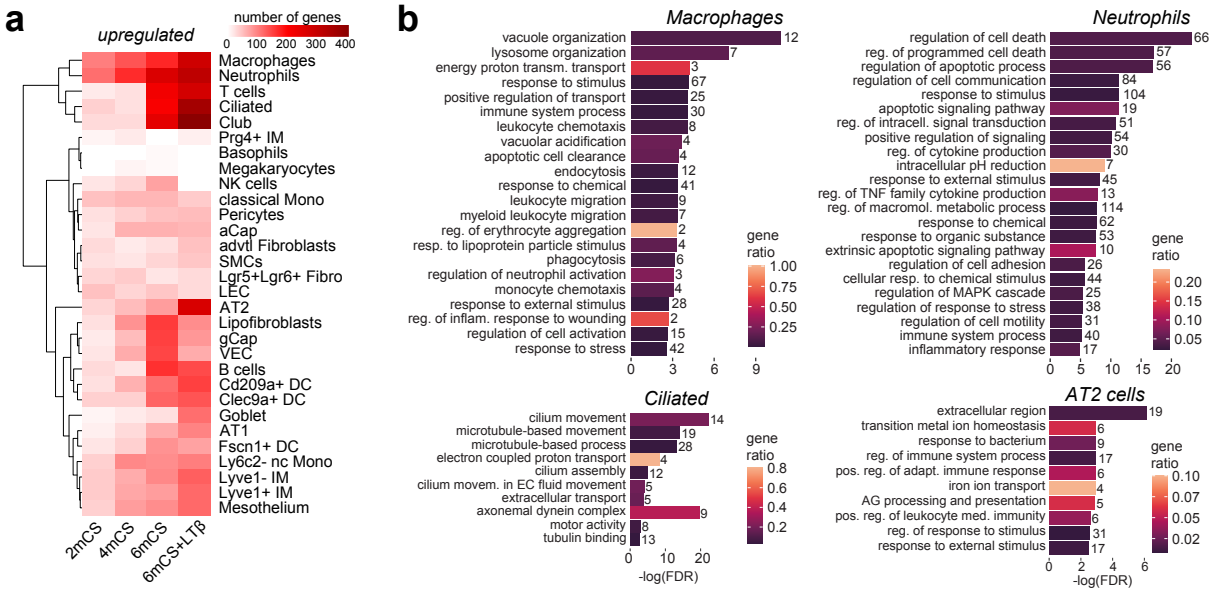


Figure 3.24: Chronic smoke exposure drastically affects immune and epithelial cells in the lung. **a** Heatmaps showing effect quantification of chronic smoke CS exposed mice. Colour indicates number of up-regulated genes, columns correspond to duration of CS exposure in months. **b** Selection of terms that were enriched in significantly up-regulated genes after 6 months of CS exposure (adj. p-value < 0.05, logFC > 0.5) in indicated cell types. Number of input genes were 185 (macrophages), 256 (neutrophils) 190 (ciliated) and 77 (AT2).

Consistent with expectations, terms associated with phagocytosis, apoptosis, response to pathogens, inflammation and general stress response were among the most significantly enriched terms for alveolar macrophages, neutrophils and AT2 cells. Ciliated cells followed a different pattern. Due to their essential role in mucouciliary clearance, they also seemed to enter an activated state after 6 months of exposure (see drastic shift in UMAP space Fig. 3.23b). To respond to the continuous supply of pathogens, they up-regulated many genes associated with microtubules and cilium movement.
Smoking has been recognized to suppress mucouciliary clearance in most smokers and can be improved by smoking cessation. Examination of their airways showed patches of atypical nuclei and missing cilia, confirming that constant particle exposure induces detrimental effects on ciliated cell structure and function.[285]

In the final part of exploration of LTβR inhibition, expression of its signalling molecules in adaptive and innate immune cells and LTβR target gene expression in cells from COPD patients as well as in the mouse model was evaluated.
Quantification of mRNA levels determined by qPCR in lung core biopsies from healthy individuals and patients with COPD revealed an increased expression of the LTβR ligands LTA, LTB and TNFSF14 (also known as LIGHT) and TNF. Furthermore, its downstream targets, namely the chemokines CCL2, CXCL8 and CXCL13 displayed significantly higher levels in COPD patients as well (Fig. 3.25a). To delineate the cellular sources of these signalling molecules, the corresponding expression is shown for the CS exposure model in Fig. 3.25b. Expression of the main ligands Lta and Ltb localized mostly to B and T cells. An alternative LTβR ligand Tnfsf14 was expressed on T cells, NK cells and granulocytes, whereas Tnf was expressed by all leukocytes in the mouse lungs. Mimicking the human settings, expression of these genes was significantly increased upon smoke exposure.

Neutrophils are the first line of defense against pathogens, as they phagocytize invading microorganisms and destroy them by the internal generation of reactive oxygen species and the action of proteases such as elastase or cathepsins.[286] This subset is examined separately in Fig. 3.25c, in which a clear shift correlating with the duration of smoke-exposure was discernible in the UMAP space. A recent study explored the driver genes during the transition from immature to mature neutrophils and provided signatures of neutrophils migrating into inflamed mouse lung.[284] Some example genes that were significantly higher during inflammation also corresponded to the CS exposure time in the given data set. These were for instance Interleukin 1 Receptor Antagonist (Il1rn), which modulates Interleukin 1 related immune and inflammatory responses, particularly in the acute phase of inflammation. Other examples are the Ferritin Heavy Chain Fth1 and the Macrophage Inflammatory Protein Ccl3, which is involved in the inflammatory response through binding to the receptors Ccr1, Ccr4 and Ccr5. The temporal pattern of up-regulation was particularly apparent for the LTβR ligand TNF. Interestingly, the Nuclear Factor Interleukin 3 Regulated Nfil3 was also listed in the inflammatory signature, and has been implicated in the control of IL-1β and TNF production by myeloid cells.[287]

The gradual expression increase of the listed genes towards longer periods of smoke exposure reflected the sustained inflammatory milieu. These observations suggested that neutrophils stay in an activated state, keep releasing inflammatory cytokines, which in turn influence other cell types and may contribute to their inability to return to baseline. Literature supports this notion, as neutrophils are the most abundant inflammatory cells

present in the bronchial wall and lumen of COPD patients[288] and evidence from cell culture, mouse models to human patients implicates neutrophil-derived proteases as key mediators of the tissue damage and associated decline in lung function.[286]
RNA quantification revealed increased expression of genes associated with non-canonical NF-κB signalling in COPD lung biopsies. Following a scoring of the whole GO term signature, it became clear that smoke induced the expression of the corresponding genes in alveolar epithelial cells. In the mice treated with LTβR-Ig, the average expression of these genes was reduced and resembled the value distribution of control mice (Fig. 3.25e). These results indicated that disruption of the LTβR-signalling pathway reverses cigarette-smoke-induced iBALT formation in lung tissue. The subsequent decrease of non-canonical activation of NF-κB-signalling in alveolar progenitor cells appeared to be a key element in initiating the ensuing lung regeneration.

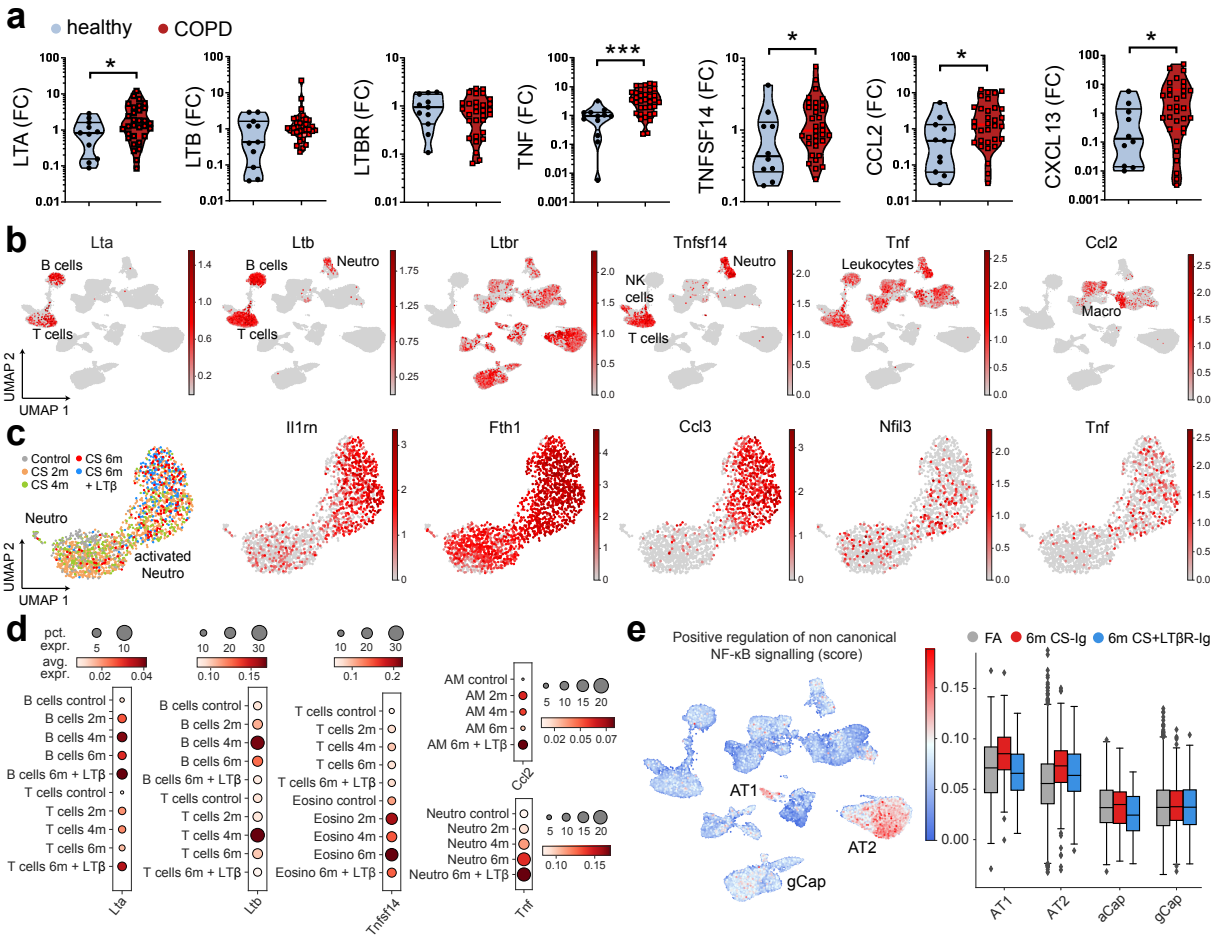


Figure 3.25: LTβR-signalling is activated in lungs of COPD patients and mice exposed to cigarette smoke. **a** Fold change of mRNA expression determined by qPCR in lung core biopsies from healthy participants (n = 11) and patients with COPD (n = 32) for signalling molecules of the LTβR signalling pathway. **b** UMAP highlighting cellular origins of genes in (a). **c** UMAP of neutrophil subset coloured by exposure condition and treatment, as well as expression of selected genes involved in inflammatory response. **d** Dotplot split by condition of ligands and downstream targets of the LTβR signalling pathway. **e** UMAP and boxplot visualizing enrichment score for genes associated with positive regulation of non-canonical NF-κB signalling pathway (GO:1901224) in the indicated cell types.

3.3.3 Evaluate association of ACE2 expression to smoking habits

As a brief excursion, the focus will be shifted towards a different, equally devastating pulmonary disease, that has marked the years following 2019, COVID-19.

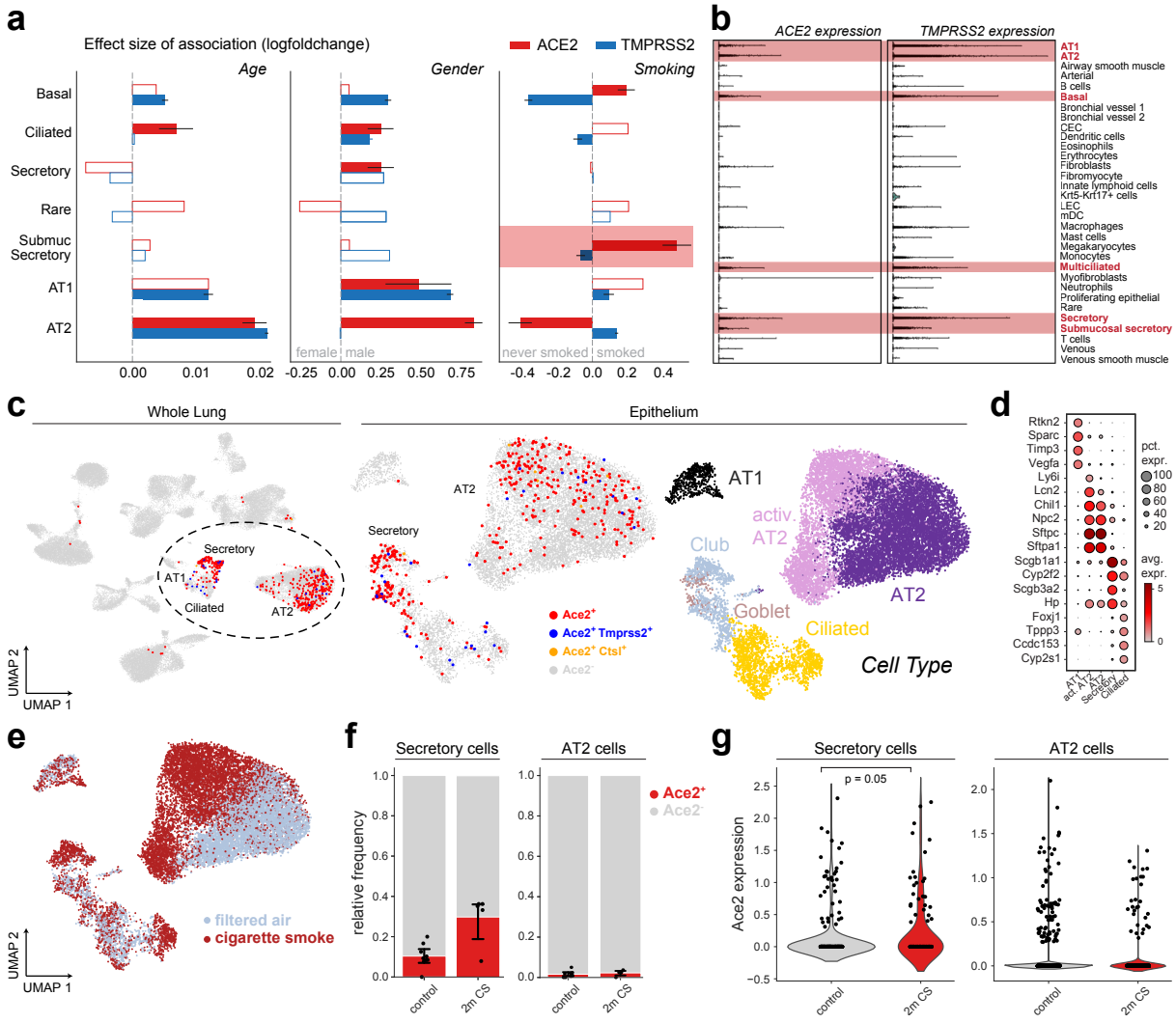


Figure 3.26: Increase of Ace2 expression in secretory cells with smoking. **a** Association of ACE2 and TMPRSS2 in donor lungs with age, gender and smoking status. The effect size is given as logFC (gender, smoking status) or slope of log expression per year with age. Coloured bars indicate significant association in respective cell type (adj. p-value < 0.05, one-sided Wald test on regression model coefficients). **b** Distribution of ACE2 and TMPRSS2 expression in healhy lungs per cell type after harmonization of published annotations. Main cell types are indicated by red shading. Figures in a, b are taken and adapted from Muus et al. (2021).[167] **c** UMAP of FA and CS exposed mice, highlighting $Ace2^+$ and $Ace2^+Tmprss2^+$/ $Ace2^+Ctsl^+$ double-positive cells in the whole lung (left) or epithelial cell type subset (middle). UMAP of epithelial subset coloured by cell type label (right). **d** Dotplot of selected epithelial marker genes that were used for cell type annotation. **e** UMAP of epithelial subset colour-coded by exposure condition. **f** Proportion of $Ace2^+$ cells in AT2 cells per sample as barplot show-case smoke-induced expansion in airway secretory cells, but not in AT2 cells (95% confidence intervals). **g** Violinplot reflecting increase of Ace2 expression after 2 months of smoke exposure in secretory cells, but not in AT2 cells (p-value derived from wilcoxon rank-sum test).

This large-scale meta study has been enabled within the global framework of the Human Cell Atlas consortium. Christoph Muus, Malte Luecken and collegues collected, integrated and harmonized cell type specific expression of the three SARS-CoV-2 entry factors ACE2, TMPRSS2 and CTSL from 107 (partially un-published) scRNA-seq studies from different tissues, along with experimental validation of the key messages.[167] In this study I could contribute by evaluating conserved expression pattern of the entry factors in the corresponding mouse model.
$ACE2^{+}TMPRSS2^{+}$ double positive cells were found across the human body, most prominently in the epithelial cells within the ileum, liver, lung, nasal mucousa, testis, prostate and kidneys.[167] Zooming into lung-related locations pin-pointed the secretory goblet and multiciliated cells in nose and airway and AT2 cells in the distal lung as the main doublet-positive populations (Fig. 3.26b). Concomitant with the expression patterns in non-diseased human lungs, the $Ace2^{+}Tmprss2^{+}$ and $Ace2^{+}Ctsl^{+}$ double-positive cells were present primarily in club and ciliated cells in the mice lungs (Fig. 3.26c).

The prevalence of COVID-19 is greater in older people, in particular men with additional co-morbidities have an increased risk of developing a severe disease.[160] To build a bridge to the virus' entry factors, the association of expression levels of ACE2 and TMPRSS2 to the clinical factors age, sex and smoking habits were modelled using a generalized linear model, while accounting for technical variation arising from study cohort and potential covariate interactions, put more precisely:[167]

$$Y \sim age + sex + age : sex + smoking + sex : smoking + age : smoking + dataset$$

The integrative analysis revealed their cell type specific associations (Fig. 3.26a), ACE2 expression increased with age in AT2 cells, and was elevated in males in airway secretory, AT1 and AT2 cells. Further, the levels were higher in past or current smokers in basal and submucousal secretory cells, and lower in AT2 cells.[167] As models that approximate human physiology are essential during pre-clinical studies, it is important to confirm their correspondence before-hand. Especially the trends regarding the smoking status were intriguing and motivated a comparison to the patterns in the mouse data at hand.
Upon 2 months of smoke exposure, the proportion of $Ace2^{+}$ cells as well as Ace2 expression levels substantially increased in secretory cell populations, but not in AT2 cells compared to the air exposed controls (Fig. 3.26f,g). The expression profiles of Ace2 in airway epithelial cells were in agreement with the pattern in the respective human counterparts, and served as indication to why smokers are more likely to develop severe SARS-CoV-2.

Conclusion

Taken together, by the analysis of lung tissue from patients with COPD coupled to scRNA-seq of the smoke-exposure mouse model, the induction of LTβR signalling mediators and their cellular origins could be delineated. Inhibition of this very signalling cascade dispersed and even prevented the formation of lymphoid follicle structures, potentially affected by the decrease in non-canonical NF-κB signalling through non-canonical NIK in alveolar stem cells. The presented projects benefited from the use of a mouse model, as aspects from both COPD and COVID-19 with regards to human smoking habits could be mirrored to a substantial degree. These results again show how appropriate models that approximate human pathology are crucial for pre-clinical studies, as they enable the identification of key pathways and drug targets that aid in improving human health.

3.4 Reveal ex vivo signatures of SARS-CoV-2-reactive T cells through reverse phenotyping

The work presented in the this last chapter will remain within the frame of the lung disease explored moments ago, the coronavirus disease COVID-19. The focus is shifted away from stem cells in the lung towards immune cells, particularly the T cells and the different effects an infection induces in them are explored.
T cells from COVID-19 patients were extracted and stimulated with the SARS-Cov-2 antigen. This allowed to categorize them based on whether they changed their transcriptomic profile in reaction to the virus or remained non-reactive. The induced transcriptional signatures of currently and previously activated T cells could be characterized and set into context by comparing them to unperturbed phenotypes of T cells from the respiratory tract of diseased patients. Finally, the consequences of the gene expression alterations were assessed by the means of intercellular communication analyses, facilitated by embedding the gained results into larger, recently published data sets.

The herein described results have been drafted into a manuscript, which was officially accepted for publication in *Nature Communications* as of 17$^{\text{th}}$ June 2021.[174] A non-peer-reviewed version has been uploaded to the medRxiv server, which is a free online platform for health science preprints. The file is accessible via the link https://www.medrxiv.org/content/10.1101/2020.12.07.20245274v1 or the doi `10.1101/2020.12.07.20245274`. Experimental data collection such as scRNA-seq, FACS-sorting, antigen-stimulation was performed by Karolin Wagner and co-authors. Quality control and computational preprocessing of the PBMC data was performed by David Fischer, and communication analysis via `NicheNet` in the latter half by Niklas Lang. My contribution encompass great parts of the single-cell analysis, in particular processing and analysis of tracheal aspirate data, merging it with the PBMC data, exploration of antigen-induced gene expression shift and finally integration and validation of the results using additional patient cohorts. Therefore, sections of the following chapter have been used in parts in the final manuscript.

Introduction

T cells are integral in the host's adaptive immune response and clearance of virus-infected cells of the respiratory system.[289] Memory T cells can provide lifelong protection against pathogens and contribute to long-lasting immunity, an attribute which is commonly exploited during vaccination against infectious diseases. Previous studies have demonstrated that SARS-CoV-specific T cells for instance can still be detected many years after infection in patients who have recovered from SARS.[290] Due to the urgency throughout 2020, the interest in better characterizing the in vivo phenotype of T cells reactive to the SARS-CoV-2 antigens has increased drastically. Immunodominant SARS-CoV-2-antigen specificities have been identified for this emerging pathogen, and many studies performed phenotypic characterization of antigen-reactive T cells. The methodologies to assess antigen-reactivity differ across study, still there is general agreement that SARS-CoV-2-reactive T cells are activated and differentiate during the course of the immune response. It has been demonstrated recently that scRNA seq can be used to reveal activation-induced phenotypic profiles of antigen-reactive T cells.[291]

A T cell receptor's (TCR) clonotype is a unique nucleotide sequence that arises during its maturation and provides the cell's specificity to antigenic peptides. Additional to the global profile, scRNA-seq enables the cell-specific capture of TCRs and thereby the identification of the sequences of such antigen-reactive clonotypes.

Here, T cells of severely diseased COVID-19 patients were isolated from the peripheral blood mononuclear cells (PBMC, n = 2) and tracheal aspirates (TA, n = 9) which will be referred to as *Munich cohort*. Consistent with the disease's risk profile, seven of nine patients were male and age ranged from 51-82 years. All patients were treated in an intensive care unit (ICU) and had been on a respirator for 8-38 days at the time of sampling, from which ultimately 2 patients deceased and 7 patients recovered. ScRNA-seq was performed for all TA samples (3' transcriptomics), while PBMC samples were taken from two patients (GT_3, GT_2) and split into two groups. To isolate antigen-specific cells later on, the first group was stimulated with a SARS-CoV-2 spike protein peptide mix, while the second control group was left unstimulated. After extraction of more than 10.000 CD4 and CD8 T cells with flow cytometry-assisted cell sorting (FACS), scRNA seq (10x 5' transcriptomics and VDJ) was performed (Fig. 3.27a). The general results were consistent in both samples, however will only be shown for patient GT_3 for clarity hereafter.
Upon in vitro re-stimulation with antigens, the T cells showed a transcriptomic shift and could be categorized into *antigen-reactive* or *non-reactive* based on their up-regulation of well-known activation markers. The TCR sequence can thereby serve as a natural barcode to link T cells of the stimulated to the ones from the unstimulated condition, by matching their in vivo expanded clonotypes with common antigen specificities. This process was titled *reverse phenotyping*.

3.4.1 SARS-CoV-2-antigen-induced shifts in PBMC T cells

To assess whether the antigen-stimulation successfully induced transcriptional shifts in the PBMC cells, the data set was pre-processed following the general workflow described in the methods section. Briefly, genes that were expressed in at least 10 cells were retained before cell-wise scaling the expression vectors to a total count of 10,000 and logp1-transforming the data. Variable genes were selected with `scanpy's pp.highly_variable_genes(flavor = "seurat")` and used as input for the PCA and subsequent knn graph construction `pp.neighbors(n_neighbors = 50)`. Cell type assignment was done in a two-stage process. First, unsupervised clustering returned 32 Leiden groups which were assigned to either CD4 or CD8 T cells based on their relative mean expression in the group. Second, cells from clonotypes, which contained both CD4 and CD8 T cells, were assigned to the major cell type found in the respective clonotype.

One group of cells was visually distinct from all other cells and encompassed mostly CD4 and CD8 T cells from the stimulated condition (Fig. 3.27a). 2.88% of all stimulated cells were assigned to this cluster.
After assessing the expression of the gene encoding Interferon-γ (IFNG), a prominent effector cytokine that is released as a product of antigen-stimulated lymphocytes, it became apparent that this distinct stimulation-induced cluster represented antigen-reactive T cells. To anchor this hypothesis on more than a single gene, scoring of unbiased activation signatures for CD4 and CD8 T cells as previously identified by scRNA-seq was

performed. While CD4 T cells have been described to show gradually enriched IFN response (early activation) and proliferation (late activation) scores, CD8 T cells undergo sequential transcriptional states reflected by high cytotoxic (early activation) or cytokine secretion scores (late activation).[292] Accordingly, the CD4 T cells showing high proliferation scores and CD8 cells with high cytokine scores were exclusively present upon stimulation in cluster 29 (Fig. 3.27b). In line with CD8 T cell activation, differential gene expression analysis confirmed that cluster 29 CD8 T cells showed up-regulation of genes such as IFNG, TNF, IL2, CCL3, CCL4 or GZMB.

TCR variation is primarily associated with the third complementarity determining region (CDR3) that interacts specifically with the MHC-bound peptide. Differences in these regions are generated by somatic rearrangements within the variable (V) and joining (J) gene segments of the TCRα chain and within the V, D (diversity), and J gene segments of the TCRβ chain.[293] This VDJ recombination can be exploited via VDJ sequencing, in which VDJ regions are purified and amplified before sequencing, enabling the assessment and comparison of the clonal heterogeneity of T cells.
CDR3$\alpha\beta$ sequences were detected in 92% of the analyzed T cells (69.9% fully paired CDR3$\alpha\beta$; 3.4% CDR3α only; 26.7% CDR3β only), from which only clonotypes with identical fully paired CDR3$\alpha\beta$ sequences were retained in the analysis.

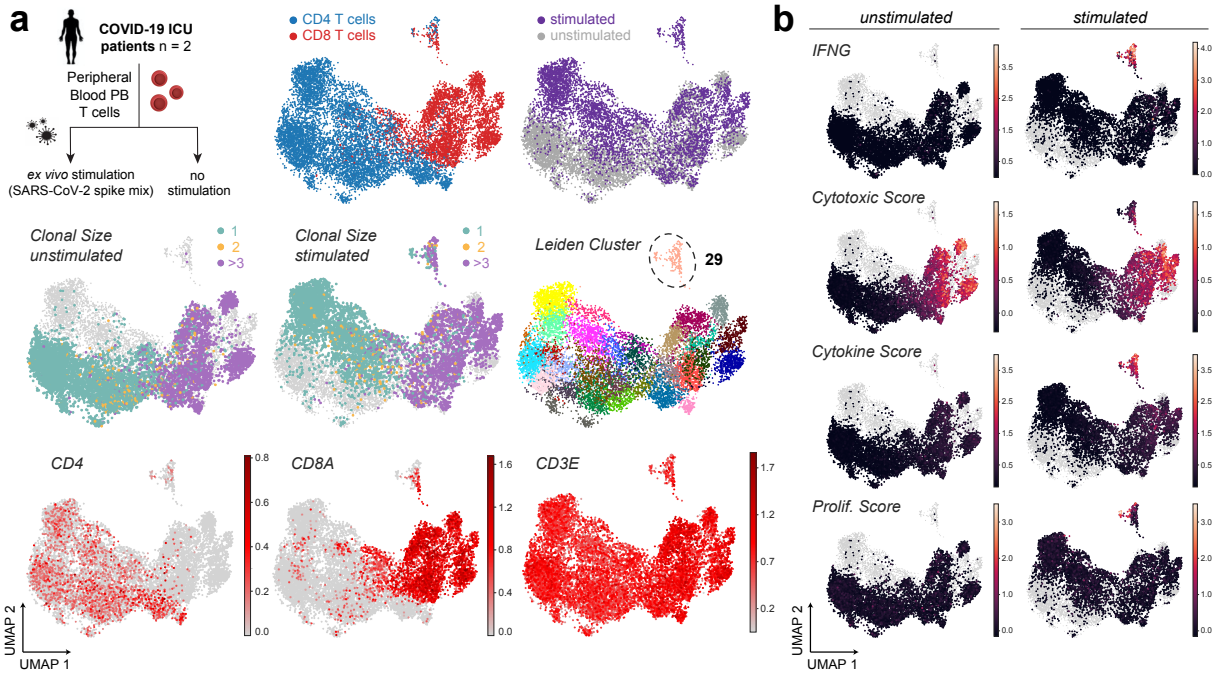


Figure 3.27: scRNA-seq reveals ex vivo signatures of SARS-CoV-2-reactive T cells. **a** Scheme of experimental setup. T cells from flow cytometry-sorted peripheral blood (PB) of intensive care unit (ICU) patients with COVID-19 were profiled with 10x. Before scRNA-seq, PB T cells were stimulated with SARS-CoV-2 spike protein peptide mix or left untreated. For the following analysis, cells from one patients are used. UMAP colour coded by cell type, stimulation state, Leiden clustering results and common T cell marker genes (n = 11,460 cells in total). **b** UMAP showing IFNG expression, cytotoxic score, cytokine score and proliferation score, highlighted in either unstimulated (left panels) or stimulated (right panels) T cells. Signature genes were derived from literature.

Subsequently, clonotypes that underwent transcriptional shifts upon antigenic stimulation were detected by their enhanced IFNG expression. Among clonotypes with at least three cells per stimulation condition, those with a statistically significant up-regulation of IFNG after antigen-specific stimulation (two-way ANOVA with Sidak's multiple comparisons test) were defined as antigen-reactive, resulting in each 5 clonotypes for CD4 and CD8 T cells (Fig. 3.28a, b). Apart from clonotype 574, reactive clonotypes were larger in size, consistent with previous observations in activation and clonal expansion. As expected, a high fraction of the stimulated cells characterized as antigen-reactive were assigned to the stimulation-induced Leiden cluster 29. The distinct transcriptional shifts in specific clonotypes confirmed the presence of antigen-reactive T cells.
Targeted re-stimulation of the cells with SARS-CoV-2 antigens introduces a major bias compared to the unperturbed cells from diseased patients. Instead, inspecting the clonotypes without re-stimulation would be closer to the in vivo setting and give a more relevant read-out. Using the TCR sequence as natural barcode, it was possible to link the reacting clonotypes back to those from the non-stimulated condition. As non-stimulated cells reflect the phenotype that reactive cells would have had if they had not been stimulated, it was possible to explore the unperturbed ex vivo phenotype of antigen-reactive T cells.

It should be noted that the individual clonotypes have different phenotypes a priori, explaining why some display unique reactivity upon stimulation. For example, after stimulation all reactive clonotypes up-regulated PDCD1, which encodes a surface protein of B and T cells that regulates the immune response, or down-regulated CXCR4, which is known to occur during differentiation into T cells with an effector phenotype.[294] Contrary to these synchronized shifts, stimulation also induced clonotype-specific changes in CD4 T cells, as clonotype 138 up-regulated TBX21 expression while clonotype 256 down-regulated it (Fig. 3.28f). Such intra-clonal variability adds an additional layer of complexity which should be kept in mind.

Next, it was attempted to experimentally validate the antigen reactivity of the selected clonotypes. With the advent of genomic engineering possibilities through tools such as CRISPR/Cas9, it became possible to replace a T cell's endogenous receptor by a transgenic one. In this process called orthotopic TCR replacement (OTR), transgenic TCRs can be inserted into specific endogenous gene loci using homology-directed repair, which places the transgenic TCR under physiological transcriptional control, while simultaneously eliminating the endogenous TCR.[295]
OTR was used to generate TCR-transgenic T cells, by equipping healthy donor T cells with identified CD4 TCRs selected by the just described screening (reactive TCRs 138, 19, 256 and 574) as well as the TCR of the largest CD4 clonotype (TCR 48). Remarkably, the T cells equipped with TCRs defined as reactive all showed SARS-CoV-2 spike antigen-dependent reactivity, whereas TCR 48 knocked-in T cells did not (Fig. 3.28i) As clonotypes which had less than three cells in each condition were excluded from the analysis, small clonotypes (e.g. 1373 and 1904) were not considered during the definition of reactive clonotypes, although TCR 1904 showed reactivity and two of the cells in the stimulated condition had a transcriptional shift into cluster with IFNG up-regulation. Overall, this data functionally validated the SARS-CoV-2 reactivity of the selected TCRs and reinforced the approach to use IFNG up-regulation after stimulation as read-out to detect antigen-reactive clonotypes.

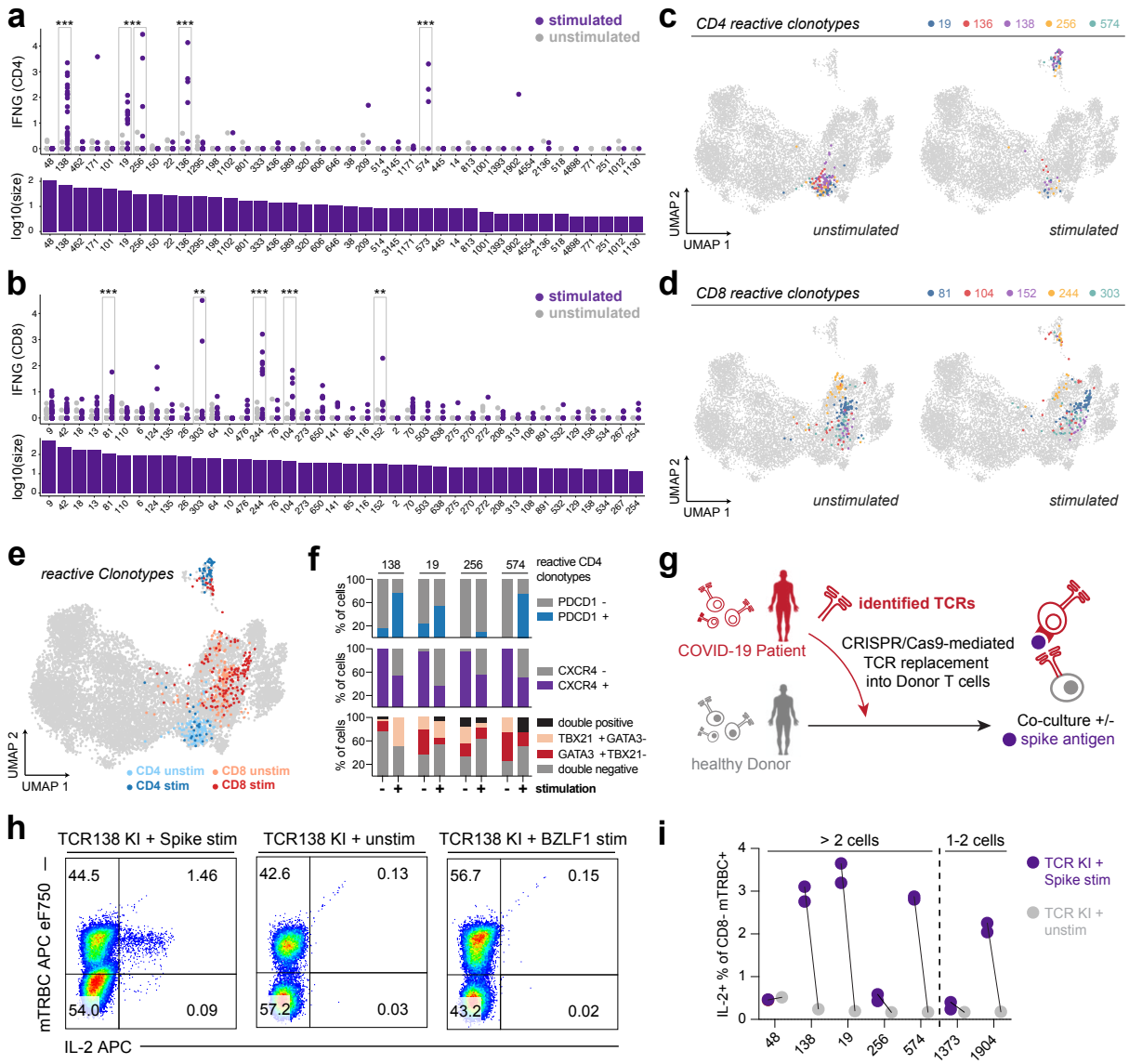


Figure 3.28: Identification and validation of SARS-CoV-2 antigen-reactive T cell receptors. **a, b** Barplot of IFNG expression in CD4 T cells (a) and CD8 T cells (b) per clonotype, ordered by clonotype size (*** $p < 0.0001$, ** $p < 0.005$). Clonotypes are defined as antigen-reactive if they show significant upregulation of IFNG upon stimulation. **c, d** Reactive clonotypes are highlighted in UMAP for CD4 (c) and CD8 T cells (d). **e** Reactive clonotypes are highlighted in UMAP for both CD4 and CD8 T cells. **f** Fraction of cells positive for indicated genes. **g** Donor T cells were equipped with TCRs identified from COVID-19 patients by CRISPR/Cas9-mediated orthotopic TCR replacement (OTR). Transgenic T cells were co-incubated with antigen-loaded patient PBMCs. **h** Flow-cytometry of stimulated TCR-engineered T cells, 1 week after OTR. mTRBC: murine constant region of the TCR beta chain incorporated into transgenic TCRs. Shown gates are pre-gated for $CD3^{+}CD8^{-}$ living lymphocytes. **i** Quantification of spike antigen-specific reactivity for selected clonotypes tested in (a) as well as two additional smaller ones.

3.4.2 Matching the phenotypes of antigen-reactive T cells to the ones from the respiratory tract of COVID-19 patients

To investigate the unperturbed phenotypes of antigen-reactive T cells and assess whether T cells with antigen-reactive signatures can be found in the respiratory tract as well, the PB data set from one patient was integrated with the previously described TA samples from an identical as well as eight additional patients.
Already in the UMAP space the TA T cells clustered in-between stimulated and unstimulated reactive clonotypes from peripheral blood and nestled towards the respective T cell lineage (Fig. 3.29b). Corresponding to the PB CD8 T cells, the TA counterpart showed high cytokine, cytotoxic scores and increased IFNG expression, while a similar activation pattern was not detected for CD4 TA cells (Fig. 3.29c).

Many induced molecules that showed distinct patterns in the two stimulation groups were known to be differentially regulated after re-stimulation in vitro. Antigen-reactive CD4 T cells strongly up-regulated TNFRSF9 and the effector cytokines IFNG, TNF, XCL1, XCL2 or CCL3 with respect to unstimulated reactive CD4 T cells. While GZMB or CCL4 were expressed in both reactive and non-reactive CD4 T cells in the unstimulated condition already, their expression was boosted by stimulation for reactive CD4 T cells only. CXCR4 is known to be more expressed in less differentiated T cells, and was down-regulated upon T cell activation after stimulation. T cells express co-inhibitory receptors that regulate T cell function, which negatively modulate TCR signalling and are essential in maintaining the balance between tolerance and autoimmunity. Notably, the in vitro stimulated cells reflect T cell activation rather than exhaustion, as such co-inhibitory molecules (PD-1, LAG3, TIGIT) were almost absent on antigen-reactive cells before stimulation, and only expressed thereafter.

To further corroborate the comparison, and to better describe the reactive clonotype signatures, gene expression analyses were performed via `diffxpy`[227] across the different reactivity groups for CD4 and CD8 T cells separately. The phenotypes of T cells from the unstimulated condition reflected the in vivo setting more closely and were therefore subjected to differential expression analysis comparison of SARS-CoV-2-reactive to non-reactive cells. Antigen-reactive CD8 T cells showed high expression of KLRB1, granzymes, CCL5 and the cytotoxic marker NKG7, while down-regulating TYROBP, KIR2DL3 and KLRC3. KLRB1 encodes CD161 which is part of cytotoxic/Th 1 anti-viral T cells and was identified to be most significantly up-regulated in reactive CD4 T cells, consistent with a recent study describing the up-regulation after SARS-CoV-2 re-stimulation in vitro.[296] TYROBP encodes DAP12 and has known activating as well as inhibitory immune cell signalling roles when paired with killer inhibitory receptors (KIR) or family members of the killer lectin like receptors (KLR). Down-regulation of TYROBP and KLR upon CD8 T cells activation has been shown previously as well.[297]

Regarding respiratory T cells, it was apparent that their transcriptomic profile resembles the antigen-reactive T cell signatures for great parts, for instance the genes up-regulated in reactive clonotypes with respect to un-reactive clonotypes (IFNG, TNF, PRF1, NKG7, CCL5, TGFB1, CST7, KLRD1, GZMA, GZMB, GZMH, GZMK) featured higher expression in the T cells from the respiratory tract than in the non-reactive T cells from peripheral blood (Fig. 3.29h).

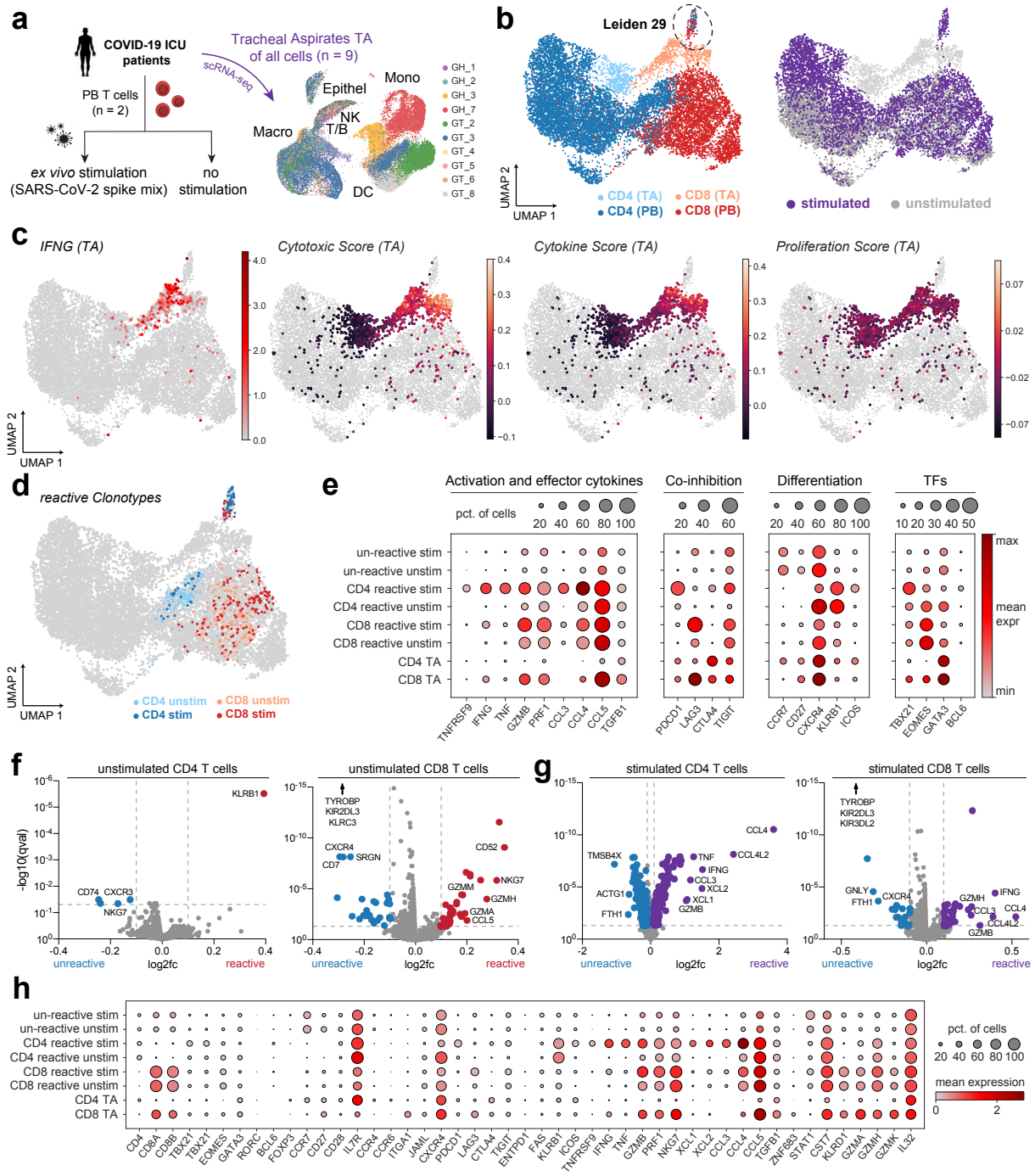


Figure 3.29: Transcriptional shift of in vitro stimulated reactive T cells is also present in ex vivo T cells of diseased patients. **a** Experimental setup, additionally to the PB T cells from Fig.3.27, cells from tracheal aspirates (TA) of intensive care unit (ICU) patients with COVID-19 were profiled with 10x. **b** Integrated UMAP of T cells from both PB and TA, overlaid with cell type and stimulation status. **c** UMAPs showing IFNG expression and enrichment scores of genes associated with cytotoxicity, cytokines and proliferation. **d** *Reactive* clonotypes are highlighted in UMAP for both CD4 and CD8 T cells. **e** Dotplot of genes associated with cytokines, per cell type and reactivity combination. **f** Volcano plots showing differential expression results of PB T cells from non-reactive versus reactive clonotypes, calculated for each cell type and condition combination. **h** Dotplot of PB and TA cells for selected genes from expression testing results, grouped per cell type and reactivity combination.

Certain transcription factors mediate programs specifically in tissue-resident cell types. ZNF683 for instance encodes the Tissue Resident Memory T Cell-associated transcription factor Hobit, which is expressed in tissue-resident memory T cells after initial antigen exposure, providing localized protection against pathogens upon reinfection.[298] ITGA1 encodes CD49a, another tissue resident memory marker, shown to persist at sites of previous infection.[299] Both of these markers were expressed in TA CD8 T cells and distinguished them from antigen-reactive CD8 T cells from the blood. TA CD8 T cells were generally very similar to reactive PB T cells - especially from the stimulated condition. This data indicated that the stimulation induced broad changes in transcriptional profiles, particularly the up-regulation of some expected activation markers, and could be matched to the phenotypes of T cells from the respiratory tract of COVID-19 patients.

To test the generalizability of the results on independent cohorts and in order to include varying levels of severity and more healthy controls, the single cell data from Munich was extended by recently published samples from bronchoalveolar lavage fluid (BALF) of COVID-19 patients with severe and mild disease as well as healthy controls. Raw count matrices from the publicly available cohorts Shenzhen,[300] Chicago[301] and Berlin[302] were re-processed separately following the standard pre-processing workflow and the extracted T cells were combined into one integrated data set. The subset of T cells encompassed 30,033 cells from 28 patients in total. For visualization of the concatenated data sets, a UMAP and a batch-corrected neighbourhood graph was constructed via `BBKNN (neighbors_within_batch = 10, batch key = cohort)`.

For some of the severely diseased patients, SARS-CoV-2 transcripts were still detectable in the cells, thus the patients were further categorized accordingly into *severe virus-positive* and *severe virus-negative* (Fig. 3.30a, b).
The integrated T cells were subjected to Louvain clustering in order to gain a first unbiased look. Stimulated CD4 or CD8 T cells from peripheral blood clustered in distinct niches that also contained respiratory T cells from severely diseased patients. Cell type-dependent fractions within CD8 or CD4 T cells from peripheral blood revealed that Louvain cluster 11 was enriched not only for reactive clonotypes from the stimulated condition, but also contained the majority of T cells from in the stimulation induced Leiden cluster 29 as seen in Fig. 3.27a. At the same time, CD8 and CD4 from respiratory samples that were located in this *integration Louvain cluster 11* were mostly from severely diseased patients across the cohorts (Fig. 3.30c, d). Intriguingly, PB CD8 reactive T cells clustered together with respiratory CD8 T cells from severely diseased patients, particularly from Chicago and Shenzhen, which in contrast to the Berlin cohort, encompassed a substantial number of virus-positive patients very early after entering the ICU. This led to the hypothesis that the phenotypic signatures of the stimulated reactive clonotypes could reflect active virus replication, while the unperturbed counterpart from the unstimulated cells mirrors virus cleared respiratory tract environments of severely diseased patients, in which virus was either still detectable or not detectable anymore. Looking at a specific gene, CCL4 (Macrophage Inflammatory protein-1β), which encodes a chemoattractant for a variety of other immune cells,[303] was markedly up-regulated in respiratory T cells from patients with severe disease, even more so in virus-positive patients. This is mirrored by higher expression in reactive T cells from peripheral blood, with particularly pronounced expression in the stimulated condition.

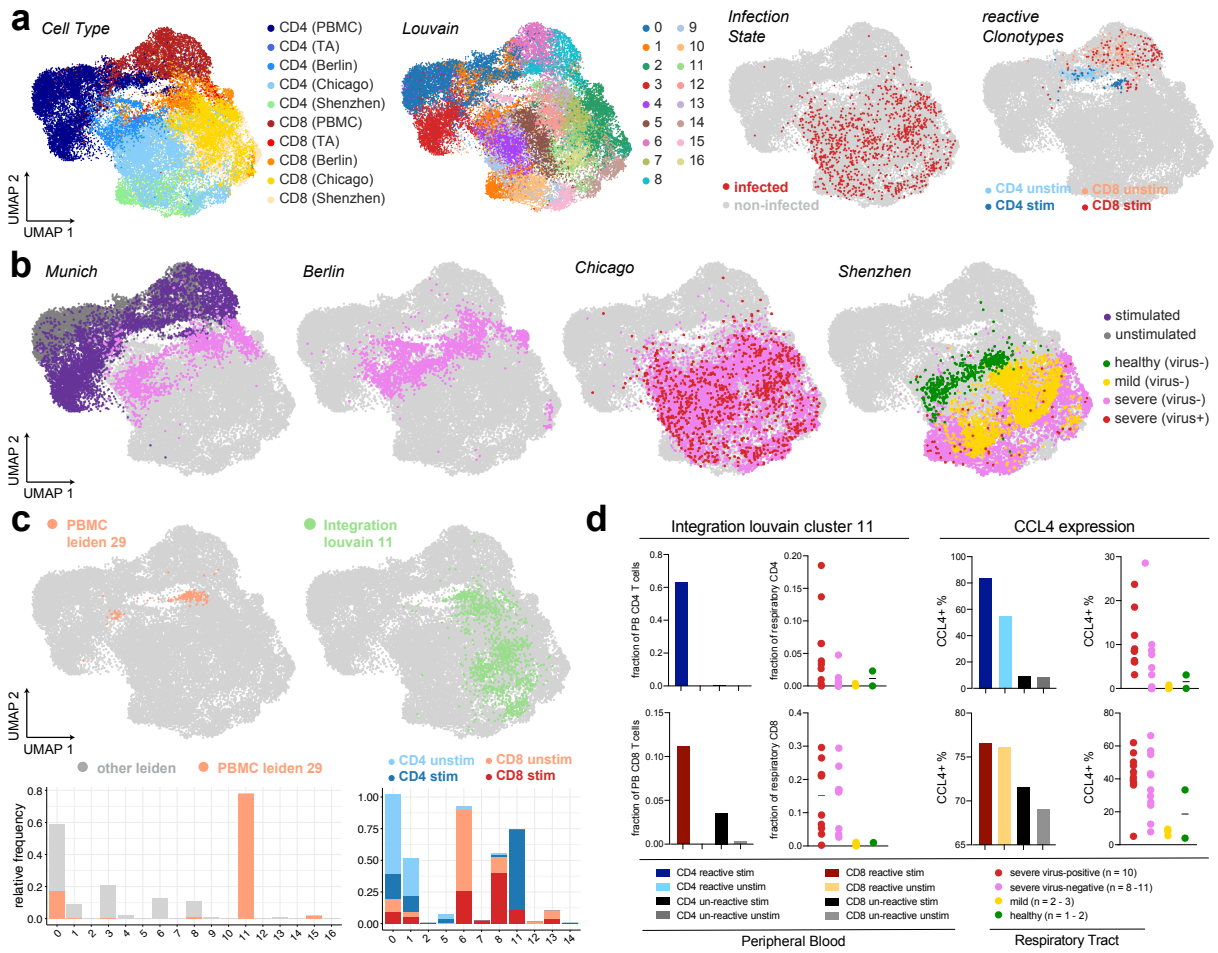


Figure 3.30: Phenotypic convergence of in vitro stimulated T cells from peripheral blood and T cells from the respiratory tract of severely diseased patients. **a** Integrated overview of previously analyzed T cells from PB and TA, and additional bronchoalveolar lavage fluid (BALF) samples from patients with mild disease, severe disease and healthy donors. UMAP colour coded by cell type, Louvain cluster, infection state and reactive clonotypes (n = 30,033 T cells from 28 patients). **b** Severity of patients/stimulation status of PB samples highlighted in the 4 different cohorts separately. **c** Location of IFNG-positive antigen reactive Leiden cluster 29 (see Fig. 3.27a) in integrated view. After re-clustering of integrated cells, highest fraction of this Leiden cluster 29 is found in the *integrated Louvain cluster 11*, which also contains large portions of reactive CD4/CD8 clonotypes. **d** Fractions of Louvain cluster 11 (left) or CCL4 expressing cells (right) among CD4/CD8 T cells from either PB or respiratory samples. For respiratory T cells, data is grouped by individual patients with indicated disease stages (n = 2 healthy, n = 3 mild, n = 11 severe virus$^-$, n = 10 severe virus$^+$).

These results reinforced the connections between disease stages of individual patients and phenotypic signatures of respiratory T cells, for which currently and previously activated PB T cell subsets provided a fitting framework. As final step in this TCR based analysis, the properties of antigen-reactive T cells were put into a bigger context by including other cell types at the affected site. Due to the extension to particularly immune cells, the intercellular communication between disease-relevant niches could be explored. As there were no viral transcripts detectable any longer in the patients from the Munich cohort, the analysis incorporated a recently published scRNA-seq data set.

3.4.3 Increased communication of T cells with virus$^+$ macrophages

The study from Grant and colleagues on collected BALF samples from 10 patients with severe COVID-19 within 48 hours of intubation (*Chicago cohort*). They reported an persistent enrichment of T cells and monocytes in the alveolar space, which suggested that infected alveolar macrophages attract T cells, which in turn produce IFNG to induce inflammatory cytokine release from alveolar macrophages and further promote T cell activation. This positive feedback loop is proposed to drive the persistent alveolar inflammation seen in severely affected patients.[301]

The majority of captured cell types were from the leukocyte compartment, with a smaller number of epithelial cells present as well (3.31a). Macrophages have been grouped into tissue-resident alveolar macrophages (TRAM) as well as monocyte-derived alveolar macrophages (MoAM). Owing to the short time period between intubation and sample collection, SARS-CoV-2 transcripts could still be detected in several patients, particularly in macrophages, either after direct infection or after phagocytosis of infected cells. Strikingly, certain subsets were characterized by higher infection fractions and formed separate clusters (TRAM2, MoAM2) (Fig. 3.31b). This additional grouping enabled the investigation of cross-talk between infected macrophages and T cells that bear the established antigen-reactive signatures, i.e. belong to the described *integration Louvain cluster 11*.

The `NicheNet` algorithm[232] was applied, which ranks ligands expressed by *sender* cells according to their ability to induce a set of target genes in the *receiver* population based on prior knowledge. As a first target list those genes were selected, which were up-regulated in virus$^+$ TRAM2 with respect to virus$^-$ TRAM1. The list included well-expected cytokines such as CCL2, CCL3, CCL4, CXCL9, CXCL10, CXCL11 ICAM1, STT1 (Fig. 3.31c). IFNG and TNF were predicted to be the most important ligands of T cells from the integrated data set. Interestingly, these were most dominantly expressed in *Louvain cluster 11*, which also highly expressed the predicted ligands CCL3 and CCL4. Looking closer into the monocyte derived subset, `NicheNet` predicted similar T cell ligands inducing the transcriptomic changes between virus$^+$ and virus$^-$ MoAM (immature MoAM1 vs. mature MoAM2/3, respectively). The prediction pattern of the top predicted ligands also coincided with other Louvain cluster regions which were enriched for severely diseased patients, particularly 2, 5, 7, 11, 12.
Conversely, to elucidate whether SARS-CoV-2 transcript carrying macrophages would likewise signal back to T cells with the antigen-reactive signature, the target gene list was set to the genes distinguishing reactive from unreactive T cells from the stimulated condition (Fig. 3.31d,e). These target gene sets, including GZMB, IFNG, TNF, CCL3 and CCL4 in stimulated reactive CD4 T cells, led to the prediction of macrophage derived co-stimulatory ligands such as IL-15, IL-18, CCL4, CCL8 or CXCL9, which have been described to be up-regulated in macrophages from the respiratory tract during COVID-19 already in the original studies from the Berlin and Shenzhen cohorts.[300,302] For the CD8 counterpart, macrophage derived co-stimulatory ligands CCL2 and SPP1 were predicted to drive CCL3 and CCL4 expression, while IL-15, IL-18, ICAM1, ADAM17, CD80 and CD86 might drive IFNG, TNF and GZMB expression. Intriguingly, most of these ligands were preferentially expressed by the very macrophage subtypes in which SARS-CoV-2 transcript was detectable (TRAM2 and MoAM1), indicating specific ligand-receptor cross-talk between respiratory T cells with antigen-reactive signatures and virus$^+$ macrophages.

In summary, unbiased prediction of ligands responsible for gene expression changes in the corresponding target cell types highlighted an increased level of interaction, especially between SARS-CoV-2 transcript-positive macrophages and T cells belonging to Louvain clusters that had higher proportion of cells from severely affected patients.

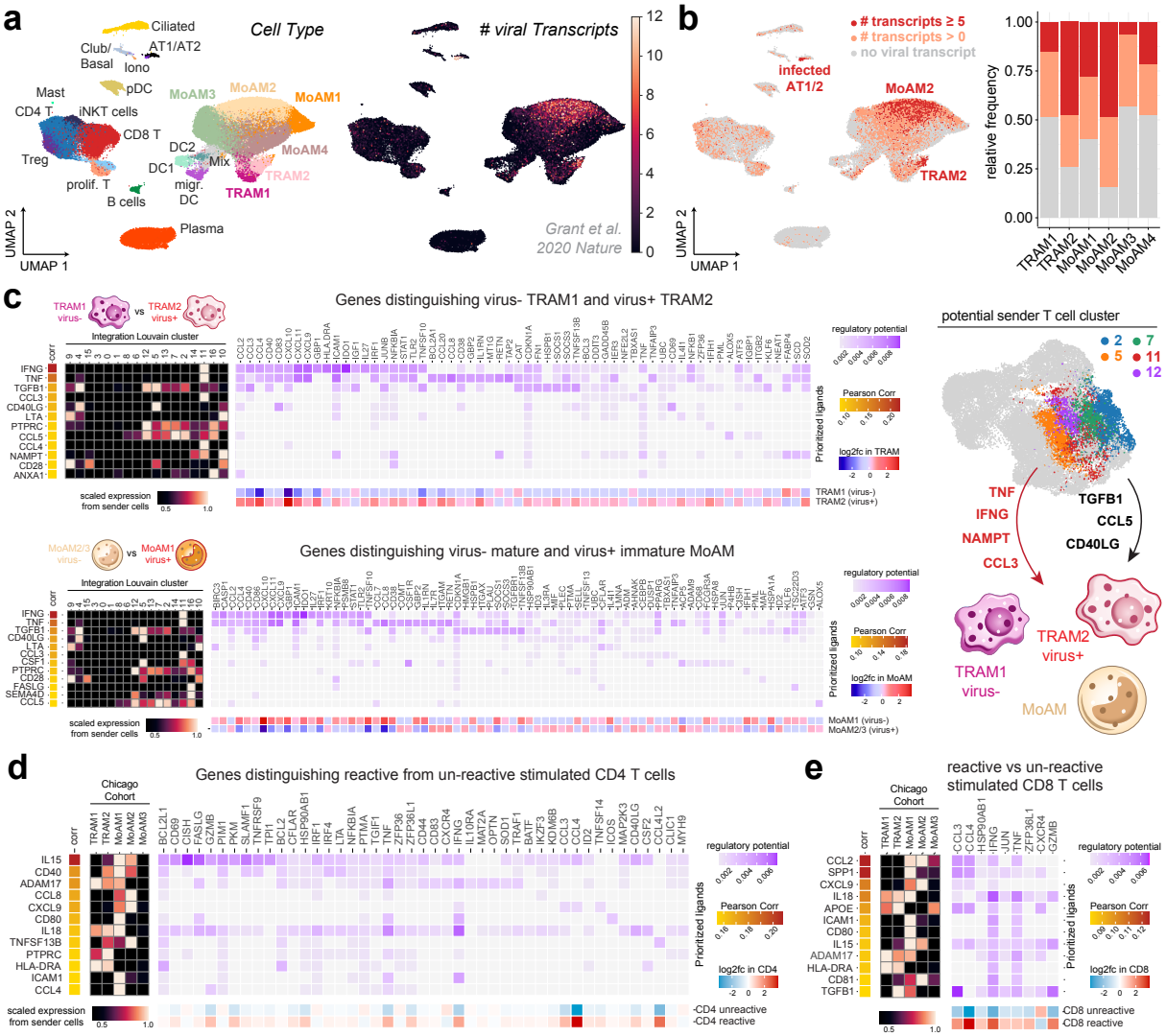


Figure 3.31: Cell-cell communication between established T cell subsets and selected cell types in the scRNA-seq reference cohort of COVID-19 patients with NicheNet. **a** UMAP plot based on Chicago data set of 77,146 cells from 10 patients with severe COVID-19 within 48h after intubation, coloured by cell type and number of viral transcripts detected per cell.[301] **b** Proportions of infected cells from tissue resident alveolar macrophages (TRAM) and monocyte derived alveolar macrophages (MoAM). For clarity, highly infected cells (5+ viral transcripts) are coloured red. **c** Potential ligands in T cell subsets from respiratory tract of patients with severe COVID-19 explaining up-regulated genes observed between virus$^-$ TRAM1 and virus$^+$ TRAM2 (upper), virus$^-$ MoAM2/3 and virus$^+$ MoAM1 (lower). The pearson correlation coefficient is used to rank the ligands' ability to predict the expression changes in the target gene set. Graphical summary of key ligands expressed by T cell subsets. **d, e** Likewise, top ranked ligands TRAM and MoAM, best predicting the signatures in the reactive clonotypes for CD4 (d) and CD8 T cells (e) respectively.

Finally, it is has not escaped our notice that the more we learn about the human genome, the more there is to explore.

The Human Genome Project[32]

Chapter 4

Discussion and Outlook

This thesis presents the innumerable directions for exploration of scRNA-seq with particular focus on data analysis and the vast complexity of the lung. After establishing a custom work flow and outlining relevant analysis tools for the uniform pre-processing of each data set, general descriptions and visualizations were generated. These overviews were convenient for guiding the attention to the most striking patterns, while accounting for the unique properties of each data set during analysis.

By balancing both wet lab techniques and bioinformatic approaches, the validity of the derived hypotheses could be assessed and allowed for biologically sound conclusions. In many cases, the plausibility was further tested by systematic comparisons against other published data sets. This extended the generalizability and mitigated interpretations driven by non-biological artefacts. It also serves as a case study to demonstrate how the emerging collaborative efforts in the single-cell field and human biology overall can be leveraged to enhance scientific progress. The resources predominantly contain healthy donor tissue, but will soon be covering samples from a variety of disease contexts. By now the construction of disease atlases has been set into motion, for instance the Idiopathic Pulmonary Fibrosis Cell Atlas[304] and COVID-19 Cell Atlas,[305] which are continuously being adapted as new data arrives within the multi-institutional collaboration.

At the start of this work, stem cells have been the center of attention, how they give rise to progenitors of the human lung and liver during the first weeks of development. The used protocol induces step-wise activation of developmental pathways such as Sonic Hedgehog or Wnt/β-catenin. The high temporal resolution and dynamic expression patterns rendered this data set as a suitable basis to get familiar with longitudinal and trajectory analysis. A number of these pathways were encountered in later analyses as well, which was expected as some are essential for proper lung physiology and are considered to contribute to age-associated chronic lung diseases.[269] Especially the impairment of alveolar stem cell function during tissue regeneration has potential to facilitate pathogenesis. Disease development is not only restricted to epithelial cell dysfunction, but occurs as a result of a complex network of shifted cellular communication, in which other nearby cells release mediators that prolong the inflammatory reaction and hamper regeneration in general. In the next pages, the main messages from the separate chapters are reinforced and put into a broader perspective by the use of recent literature in the field.

4.1 Reactivation of developmental pathways

The lung is capable of responding to acute damage when spatially restricted stem and progenitor cells re-enter the cell cycle and differentiate to promote repair. Tissue regeneration is a complex process orchestrated by the cross-talk between these cells and their respective niches. The underlying mechanisms are still far from resolved, especially how chronic inflammation or repetitive insults affect progressive tissue destruction. It has been suggested that the pathological milieu in addition to a susceptible environment like increased age, comorbidities or sustained particle exposure impair stem cell function and promote dysfunctional regenerative processes after injury.

The progenitor populations evidently use signalling pathways that play important roles during development of the lung.[306] This work underlines the general sentiment in the field, on how the reactivation of such developmental pathways during regeneration is altered in certain human lung diseases.
Basal cells for instance are the main stem cells of the large airways, and are known to differentiate towards secretory and ciliated cell fate via the involvement of Notch signalling during airway epithelial development.[307] Already in the early lung development trajectory from chapter 3.1 the induction of the Notch pathway was identifiable by the gradual up-regulation of its transcription factors HES1 and DLK1. The data shown in chapter 3.2 corresponds to this notion, as epithelial cells from lungs of ILD patients displayed higher expression of Notch receptors (NOTCH2, NOTCH3) compared to healthy controls, while a number of their ligands were up-regulated by other cell types as well.

Another key pathway contributing to progenitor stem cell function and repair is the Wnt pathway, which appeared in results throughout this thesis. Again, chapter 3.1 described how some of its components, e.g. WNT5A and DKK1, show a gradual increase during lung speciation as well. Previous gene expression profiling has already shown that human IPF lungs display higher expression of Wnt signalling players.[269] This increase was also present in the integrated data set, as many of these ligands were among the most up-regulated genes, not only in the epithelial compartment. Following communication analyses, many prominent edges featured interactions between receptors of the Wnt pathway and their respective ligands. Strikingly, many edges between disease-induced epithelial and stromal population featured these genes. For instance, WNT7A, WNT7B, WNT9A, WNT10A, SFRP2, DKK1 were predicted to influence a remarkable portion of the target genes in both directions, and further generated positive feedback loops by heavy autocrine signalling.
To switch to another chronic disease, the importance of developmental programs is evident in COPD as well. Cigarette smoke has been shown to reduce canonical Wnt signalling in human bronchial epithelial cells,[308] and components of this pathway, namely CTNNB1, CSK-3β and TCF4, have been reported to have significantly lower levels in peripheral lung tissue of COPD patients relative to smokers without chronic lung disease.[309] The NF-κB pathway as a master regulator of inflammation further functions as crucial coordinator of cell differentiation, proliferation, and survival. Findings in the field suggest that the two signalling pathways cross-regulate each of their activities, and influence the progression of inflammation. However, both positive and negative cross-regulation has been observed depending on the cellular or tissue context.[310]

Further investigation of the smoke exposed mice described in chapter 3.3 beyond the single-cell data established a link between the Lymphotoxin-β receptor and Wnt/β-catenin signalling. Primary AT2 cells from human and mouse cell lines exhibited a down-regulation of the Wnt target genes Axin2, Nkd1, Lgr5 and Tcf4 following supplementation with LTβR agonists. This phenomenon could be reversed by LTβR inhibition. To confirm that this is indeed a result of decreased Wnt/β-catenin signalling, a NIK kinase-specific inhibitor acting on the non-canonical NF-κB pathway was added, again showing decreased expression of the aforementioned Wnt ligands.[172] These results add further functional context to the observations presented in chapter 3.3, on how alveolar epithelial cells up-regulate genes associated with canonical NF-κB signalling upon prolonged smoke exposure. LTβR signalling decreased such Wnt/β-catenin.
Treatment with LTβR-Ig disrupted LTβR signalling and resulted in down-regulation of NF-κB mediators, reversing the effects and potentially paving the way for Wnt-induced regeneration. Ultimately, this suggested that blockade of LTβR signalling represents a viable therapeutic option that not only prevents iBALT formation but also kick-starts tissue-regenerative strategies.

As briefly introduced in chapter 3.1, the lung mesenchyme regulates the growth and branching of the endoderm during early stages of development. With this in mind, the corresponding differentiation protocol was supplemented by FGF10 and SHH, both transcription factors essential in foregut and lung development. The literature suggests that the mesothelium might be involved in lung repair by reactivating developmental programs as well.[275] The presented integrative approach in chapter 3.2 supported this role, as fibroblast growth factors and Wnt target genes have been significantly up-regulated in mesothelial cells from ILD patients. Moreover, receptor-ligand analysis ranked these genes frequently as top ligands and implicated a high potential to influence the gene expression shifts in a variety of other cell types.

The induction of genes from the listed developmental pathways reflects how the lung attempts regeneration, still the response seems to be defective and aberrant communications in diseased patients could hinder the tissue from reaching its baseline state. Consequently, it remains an important task to identify new targets that interact with pathways that are involved in the pathogenesis of chronic lung diseases in order to circumvent impaired mechanisms and to catalyze proper regeneration. It is key to delineate cell-specific communications, as the results presented in this work and many other studies show-case how various cellular compartments are involved with specific response programs, all interacting in a complex network. The knowledge on these receptor-ligand pairs has been expanding rapidly in recent years, providing a valuable resource that can be leveraged and holds promise for therapeutic approaches in the future.

4.2 Persistence of otherwise transient cell states

One of the main scientific achievements this work has contributed to is the description of a previously unknown transient alveolar differentiation intermediate and their transcriptomic programs during terminal differentiation towards AT1 fate. The frequency and level of Krt8 expression of this novel state was highest during the fibrotic phases following tissue injury, and both gradually went down en route to regeneration. The transient state was enriched for pathways essential during lung regeneration, such as MYC, TNF-α signalling via NF-κB, oxidative phosphorylation,[256,257,258] and also exhibited features of senescence.

Cellular senescence is a process that promotes the elimination of unwanted cells by the means of tissue remodelling. This can be divided into three steps: First, senescent cells arrest their own cell proliferation. Second, a secretory phenotype is induced and recruits immune and mesenchymal cells to clear damaged cells and modify the ECM. Finally, nearby progenitors are mobilized to repopulate the tissue. These programs are regulated by the interplay of signalling molecules, some that directly activate cell cycle inhibitors (p14, p15, p16, p17, p21, and p27), and some indirectly via TP53, which in turn inhibit cyclin-dependent protein kinases. Senescence plays an essential role during embryonic development in order to eliminate transient structures. Alternatively, it can be triggered upon cellular damage or stress and appears to be impaired in aged tissue or pathological contexts.[311] Aging in particular is marked by progressive deterioration of tissue function over time, resulting in an increased susceptibility to environmental challenges. The irreversibly arrested proliferation of aged or damaged cells caused by cellular senescence is known as one of the hallmarks of aging. Due to the decline in regenerative capacity it is not surprising that increased age is a prominent risk factor in a variety of diseases.[312]

Although senescence is primarily a state of cell cycle arrest after extensive proliferation, it does participate in a variety of pathologies. As seen in chapter 3.2, fibrotic diseases are part of them. For instance, to limit the fibrotic response, senescence is induced in the activated fibroblast populations during the regenerative phase. In diseased individuals however, the cells might fail to induce apoptosis in the otherwise transient populations and prevent their clearance. On the contrary, senescent cells can rather negatively impact the surrounding tissue by continuously secreting proinflammatory cytokines. The aberrant activity of p53, TGF-β and Wnt pathway genes has been reported in IPF patients,[313] and was captured in the presented data as well.
In the bleomycin-injured mouse lungs the Krt8$^+$ ADI cells displayed higher scores for pathways associated with regeneration. After the inflammatory phase, in which cytokines for phagocytic immune cells had been released, these cells significantly up-regulated mediators such as Cdkn1a, which inhibits CDK-cyclin complexes and results in proliferative arrest, or Ccn1, another apoptosis related gene.

During the time of analysis other research groups discovered intermediate states in alveolar regeneration which were highly similar to the described Krt8$^+$ ADI. Kobayashi et al. (2020)[84] titled the cell state pre-alveolar type-1 transitional cell state (PATS). The authors showed how these cells arise after alveolar injury by lipopolysaccharide (LPS) treatment of mouse lungs, but could also generate them in ex vivo alveolar organoid

cultures. Their listed key markers Cldn4, Krt19 and Sfn were highly expressed in the Krt8$^+$ ADI population as well. Elaborating on the functional interpretation, Kobayashi and colleagues further validated that PATS are vulnerable to mechanical stretch-induced DNA damage, which can easily occur during the intensive stretching these cells endure in differentiation. They furthermore stress the importance of p53, and show via chromatin immunoprecipitation (ChIP) that this senescence marker can directly bind to numerous PATS-associated gene loci and directly control their expression.[84]

Another independent study by Choi et al. (2020)[314] found an equal transient alveolar state, but approached it from a different angle. The authors performed scRNA-seq on lineage-labelled in vivo mouse AT2 cells and ex vivo AT2 cell-derived organoids to define a regeneration trajectory. The regeneration-specific cell population was titled damage-associated transient progenitors (DATP), marked by Ndrg1, Cldn4, and Krt8 expression. Intriguingly, Choi et al. could confirm that chronic inflammation, mediated by sustained Il-1β levels, prevented AT1 differentiation and eventually led to an aberrant accumulation of the transient cells. The signature of DATP was enriched in Il-1β-treated organoids relative to control organoids on day 6, corroborating the effect of Il-1β in a more targeted fashion. Quantitative PCR on isolated macrophage populations from uninjured mouse lungs revealed interstitial macrophages as specific source of Il-1β expression, which was further up-regulated upon bleomycin injury.[314]
These results highlight the involvement of immune cells, and the authors suggest that Il-1β skews differentiation of AT2 cells towards the AT1 fate. While this thesis did not delineate the specifics of AT2 and macrophage interaction in greater detail, a sharp increase in receptor-ligand numbers was indeed seen during early inflammation. The experiments from Choi et al. confirmed that inflammatory stimuli can direct the cell fate behaviour of AT2 stem cells during lung injury repair by the means of these example interactions.

The mentioned studies are a valuable validation for the presence and potential function of the ADI, as the conclusions across independent labs show high agreement and complement the results presented here. However, this work was able to dive deeper into transcriptomic shifts underlying the regeneration trajectory and potential cellular communication, owing to the dense temporal resolution and focus on computational exploration.

Linking to the more relevant human aspect as outlined in chapter 3.2, Haberman et al. (2020)[86] and Adams et al. (2020)[261] could describe a novel cell population called *aberrant basaloid cells* which were highly specific to human IPF lungs and were not found in healthy donors. During the integrative analysis their presence has been shown across multiple data sets and their co-expression of basal epithelial, mesenchymal and senescence markers has been confirmed during the independent annotation. Again, common senescence genes were induced in alveolar cells from ILD patients, such as the kinase inhibitors CDKN1A, CDKN2A, encoding for p21 and p16, as well as the Cyclins CCND1, CCND2. Creating a link to the non-human setting is crucial in order to facilitate study of the origin of this cell population. Currently it would not be possible to capture their appearance in human patients, as the presence of these aberrant cells is already established in the end stage disease lungs that are available for analysis. The novel cross-species comparison at the end of that chapter show-cased the transcriptomic similarity of the Krt8$^+$ ADI from injured mouse lungs to the basaloid population.

Of particular interest was the sustained expression of genes in human ILD lungs, which were down-regulated in mouse lungs after the fibrotic phase. Prominent examples would be the ADI markers KRT8, SPRR1A, LCN, and other transiently up-regulated genes TNC and SOX4. Likewise, the strong intercellular interaction between transient epithelial and stromal cell populations, which is progressively diminished during regeneration in the mouse model, appears to persist in the human disease condition. This ongoing signalling could be one of many factors that might explain their persistence.

Nonetheless, cell lines or mouse models introduce highly artificial conditions that do not always translate to human disease. As a glimpse into how this issue could be circumvented in the future, precision cut lung slices (PCLS) were used to model human fibrogenesis in a first pilot study and provided encouraging results.
The slices were treated with a profibrotic cocktail of cytokines to mirror pathological conditions, or PBS as control. Following standard processing, a first overview for day 1 and 7 is shown in Fig. 4.1. Within this framework, the response to perturbation in all major tissue resident cells can be studied, while they are still embedded in their natural niche environment. The drastic induced cell-state shifts corresponded well to the changes seen in human ILD, most strikingly the induction of the aberrant basaloid / Krt8$^+$ ADI cell state could be recapitulated exclusively after fibrotic cocktail treatment in this ex vivo tissue culture model. Motivated by this proof of concept it could be possible to use scRNA-seq coupled to PCLS as a powerful platform for drug discovery and analysis of cell plasticity mechanisms directly in human lung tissue.

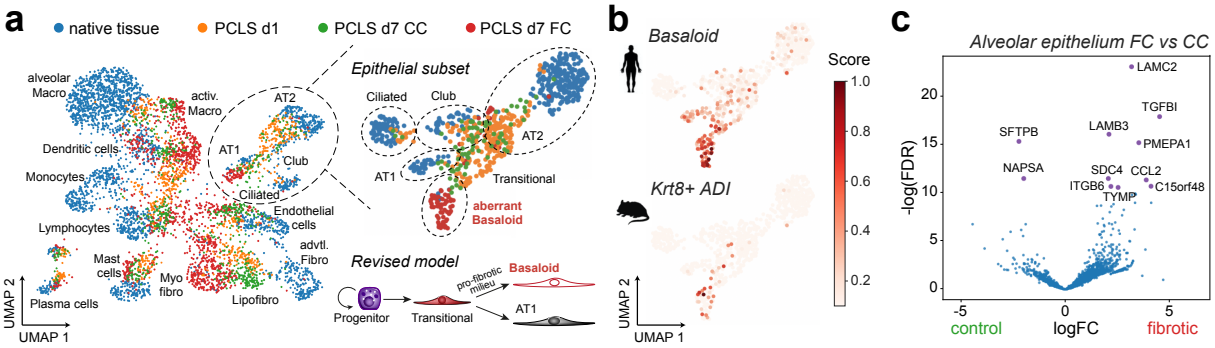


Figure 4.1: Outlook: Basaloid cells can be induced in human ex vivo tissue slice culture. **a** UMAP of cells from ex vivo tissue culture, either native tissue or after slicing (PCLS). Colour indicates time point and treatment condition; control (CC) or fibrotic cocktail (FC). Zoom into epithelial populations and scheme of proposed differentiation hierarchy following epithelial injury (left). **b** Epithelial UMAP overlaid by basaloid (human) and Krt8$^+$ ADI signature score (mouse). **c** Volcano plot showing preliminary results from differential gene expression between CC treated versus FC treated alveolar cells, including basaloid population.

Overall, the observations in both the mouse and human context lead to the hypothesis, that the non-permissive pathological milieu may prevent efficient removal of senescent cells and instead promote their accumulation, which ultimately contributes to the pathological manifestations seen in chronic lung diseases. These normally regenerative intermediate states displayed conspicuously high levels of cellular interaction, further encouraging recruitment, crowding, self-amplification and finally blocking lung regeneration by their abnormal and irreversible state.

4.3 Impaired AT1 cell regeneration in COVID-19

Adams et al. (2020)[261] defined the aberrant basaloid cells as highly specific for IPF lungs, as these could not be found in control patients. However, these cells were also present in some COPD lungs in their data set, albeit in strongly decreased numbers. They might reflect a common pathology that is shared in ILD and COPD lungs, as both are associated with increased age, smoking habits, accelerated cellular senescence and progressive loss of alveolar epithelium.[261]

The cytokine storm triggered during alveolar damage induces respiratory distress which may progress to ARDS, a disease whose features are commonly found in post-mortem histology of lung tissues of deceased COVID-19 patients. As outlined in chapter 3.4 the cross-talk between SARS-CoV-2 infected or bystander T cell and macrophages can cause a positive feedback loop resulting in sustained alveolar inflammation, eventually leading to alveolar damage. Specifically, lung parenchymal remodelling, characterized by fibroblast proliferation, alveolar obliteration, and micro-honeycombing could be observed in cryobiopsies of COVID-19 patients.[315]
Even after the virus has been eradicated in patients who have recovered from COVID-19, the removal of the cause of lung damage does not rewind the damage itself. A fraction of patients who have recovered from COVID-19 continue to battle longer lasting symptoms, varying from mild, in terms of fatigue and body aches, to severe forms requiring long term oxygen therapy and persisting lung fibrosis.[316] Still, it is not clear if the cause of these fibrotic features is the viral infection, the secondary cytokine cascade, or the ventilation.

Another remarkable study by Delorey et al. (2021)[317] set out to generate a single-cell atlas of lung, kidney, liver and heart based on autopsies from individuals with COVID-19, in which their lung data set spanned 16 donors and 106,792 cells.

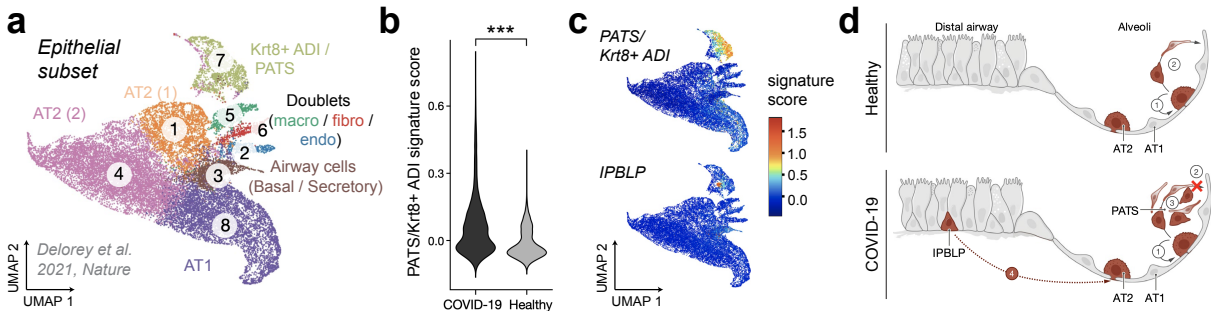


Figure 4.2: Relevance of failed alveolar regeneration in tissue samples from lung donors with COVID-19. **a** UMAP of 21,661 epithelial cells from 16 donors coloured by published annotations. **b** Distribution of PATS signature scores for 17,655 cells from COVID-19 and 24,000 cells from healthy lung pneumocytes (*** denotes p-value $< 2.2 \times 10^{-16}$, one-sided Mann–Whitney U test). **c** UMAP as in (a) coloured by signature score for the PATS/Krt8^{+} ADI or intrapulmonary basal-like progenitor (IPBLP) cell programs. **d** Proposed model of epithelial cell regeneration. In healthy alveoli, AT2 cells self-renew (1) and differentiate into AT1 cells (2). In COVID-19, AT2 cell self-renewal (1) and AT1 differentiation (2) are inhibited, resulting in PATS accumulation (3) and recruitment of airway-derived IPBLP cells to alveoli (4). Figures, legends taken and adapted from Delorey et al. (2021).[317]

The motivation for mentioning this publication at this point is that during the annotation of their epithelial compartment, the authors encountered a subset which corresponds to the previously described Krt8$^+$ ADI[171] / PATS[84] / DATP.[314] Following signature scoring of the epithelial cells, the similarity to PATS cells was confirmed and shown to be significantly higher in COVID-19 lungs compared to controls (Fig. 4.2).
Consistent with fibrosis in severe COVID-19, Delorey et at. report fibroblast expansion accompanied by loss of alveolar epithelial cells. The presence of Krt8$^+$ ADI-like cells hints at the invocation of regenerative cascades to re-establish the cells lost to infection.[317]

While the pathological manifestations of these diseases are quite different, a common niche of aberrant senescent cells that fails to trigger proper repair mechanisms could be a shared factor that participates in their onsets. Nonetheless, the existence of these disease-specific cells has been confirmed in a variety of studies in the last year - now additional studies are needed to assess the source of these cells and how to potentially manipulate them in order avoid pathogenesis.

4.4 Multi-omics and spatial data integration

The results of this thesis are strongly based on the analysis of transcriptomic data. Due to the explosion of single-cell genomics in recent years, scRNA-sequencing has been established as the most common technique - for now. As briefly described in the introduction, additional features can be consulted when trying to characterize cells. Since the initial report of single-cell transcriptomics in 2009[52] other single-cell omics technologies have evolved into central tools for biological research. These methods aim to measure for instance epigenetic modifications, non-genetic changes to the genome that have an regulating effect on gene expression, which can be via addition of chemical groups, e.g. methylation, or by modifications to chromatin, affecting which parts of the DNA are accessible during transcription.
Proteins on the other end reflect the organism's phenotype. It is now known that mRNA is not always translated into protein. Owing to these different control mechanisms at the transcriptomic and proteomic level, but also due to mRNA half-life and protein localization, there is no perfect correlation between mRNA readouts and protein profiles.[318] Thus accessing the proteomic layer provides a more relevant measure of the end product.

Nowadays, the single-cell layer can be harnessed by multiple data modalities besides transcriptomics. By measuring different types of molecules, such as DNA, RNA, protein, and chromatin at the highest possible resolution, it is feasible to accumulate more information on how the expression of genes is influenced. Although their development has been considerably lagging behind the scRNA-seq field, their applications are expanding rapidly. Many groups independently developed technologies for profiling single-cell genomes,[319,320] and different types of epigenomic modifications, such as DNA methylation,[321,322] histone modifications,[323] chromatin organization and accessibility.[324,325]
Instead of treating these layers as isolated portraits, researchers are increasingly combining them to achieve a more nuanced picture of cellular processes. Such integrative approaches of single-cell omics techniques explore the interplay and correlation of the different building blocks, ultimately enabling the construction of a multi-omics profile for the same cell.

Single-cell nucleosome, methylation and transcription sequencing (scNMT) for instance uses a methyltransferase to label open chromatin followed by bisulfite sequencing, adapted to determine the pattern of methylation, and finished with RNA-seq.[326] By combining gene-expression profiles, methylation and chromatin accessibility in the same cells, study of the interactions between the epigenome and the transcriptome became feasible. Another prominent combinatorial technique is CITE-seq, which integrates protein and transcriptome measurements into a single-cell readout by using oligonucleotide-labeled antibodies.[327] Although the number of proteins that can be measured is limited by the availability of antibodies, this approach is already widely applied due to its compatibility with existing single-cell sequencing approaches.

Heterogeneity analyses are gradually being impacted by the inclusion of the spatial component. Cellular interaction can only occur in cells that are in close proximity, thus the relative locations of cells within tissue is critical to understand their relationship and find spots of altered activity in disease pathology. However, due to the cell dissociation the information on location cannot be preserved. An increasing number of methods have been developed to supplement scRNA-seq data with spatial information. One of them would be SCRINSHOT, which directly hybridizes probes onto mRNA on fixated tissue sections, followed by amplification and sequential detection using fluorophore-labelled oligonucleotides. It relies on the specificity of the generated probes to ensure ligation activity to the correct sites and therefore requires prior knowledge on the cells of interest.[328] The great advantage however lies in the fact that the hybridization step is based on transcripts rather than proteins, and is thus not restricted by the availability of antibodies.
10X Genomics quickly realized the potential of spatial transcriptomics and established *Visium*, a commercial platform to capture gene expression profiles and their local information simultaneously. This is facilitated by slides containing barcoded spots, which are made up of millions of spatially barcoded capture oligos. This technique can be applied to either fresh frozen tissue, or FFPE tissues. The tissue is permeabilized in order to release mRNA from the cells for the frozen case, or ligated probe pairs from the cells for the latter. The released molecules then bind to the oligonucleotides on the capture area. Spatial barcodes are added via an extension reaction, such that the molecules can be traced back to their location after pooling them for library construction and sequencing.[329] The popularity and validity of this technique has already been demonstrated by a growing portfolio of peer-reviewed publications in the last years.

Such studies have proven that multimodal data analysis can achieve a more detailed characterization of cellular phenotypes than transcriptome measurements alone. Nonetheless, these modalities add further complexity to the already high dimensional data. Due to the recency of these integrative approaches, several challenges remain in defining suitable data structures, efficient computational methods and comprehensible visualizations.[330] More effort will be put into the advancement of a framework for the joint analysis. Still, the reward of this endeavour will be a bridge between the molecular omics components within their tissue context, streamlining the results to the physically possible and relevant cellular interactions. This is of great importance to compress and make sense of the wealth of data. Naturally, such biologically meaningful conclusions will expand our understanding of mechanisms in homeostasis, disease and further inform therapeutic methods to alleviate or even prevent pathological conditions in the distant future.

Chapter 5

Appendix

This segment includes additional data base information and panels that are of interest but were too lengthy to include in the main text. They are therefore constrained to this separate space for reference.

- List of the key abbreviations and acronyms used throughout the thesis.
- Name and description of hand-picked genes encountered during intercellular communication analysis in human ILD cohort data.
- Overview of quality control (sample-wise number of transcripts and genes, percentage of mitochondrial counts), filtering thresholds and pre-processing parameters (selection of highly variable genes, knn graph construction, batch correction, UMAPs) of all presented data sets.
- Compartment-wise subset UMAPs and matrixplots of selected cell type marker used during annotation process for the individual human patient cohorts.

Abbreviations

scRNA-seq	Single-cell RNA-sequencing	PCLS	Precision-cut lung slices
PCR	Polymerase Chain Reaction	WHO	World Health Organization
(c)DNA	(Complementary) Deoxyribonucleic acid	COPD	Chronic obstructive pulmonary disease
(m)RNA	(Messenger) Ribonucleic acid	ILD	Interstitial lung disease
FACS	Fluorescence Activated Cell Sorting	IPF	Interstitial pulmonary fibrosis
MACS	Magnetic Activated Cell Sorting	ARDS	Acute respiratory distress syndrome
FISH	Fluorescence In Situ Hybridization	COVID-19	Coronavirus disease 2019
Drop-seq	Droplet-sequencing	UMI	Unique molecular identifier

aCap	Aerocytes	IM	Interstitial macrophages
ADI	Alveolar differentiation intermediate	LEC	Lymphatic endothelial cells
AM	Alveolar macrophages	Meso	Mesothelium
AT1/2	Alveolar epithelial type 1/2 cells	NEC	Neuroendocrine cells
DC	Dendritic cells	VEC	Vascular endothelial cells
Fibro	Fibroblasts	nc/cl Mono	non-/classical Monocytes
gCap	General capillary cells	SMC	Smooth muscle cells
TRAM	Tissue-resident alveolar macrophages	MoAM	Monocyte-derived alveolar macrophages

iPS	Induced pluripotent stem cells	ECM	Extracellular matrix
FE	Foregut endoderm	FGF	Fibroblast Growth Factor
DE	Definite endoderm	SHH	Sonic Hedgehog Signalling Molecule
DE	Early lung progenitors	CXCL	CXC motif chemokin ligand
Bleo	Bleomycin	CCL	C-C motif chemokine ligand
PBS	Phosphate-buffered saline	TGF-β	Transforming Growth Factor β
iBALT	inducible bronchus-associated lymphoid	PBMC	Peripheral blood mononuclear cells
CS	Cigarette smoke	TA	Tracheal aspirates
FA	Filtered air	BALF	Bronchoalveolar lavage fluid
LTβR	Lymphotoxin β-receptor	TCR	T cell receptor
ROS	Reactive oxygen species	OTR	Orthotopic TCR replacement

PC	Principal component	DC	Diffusion component
t-SNE	T-distributed stochastic neighbour embedding	UMAP	Uniform manifold approximation and projection
DiffMap	Diffusion map	dpt	Diffusion pseudo time
BBKNN	Batch balanced k-nearest neighbours	PAGA	Partition-based graph abstraction
FDR	False discovery rate	GSEA	Gene set enrichment analysis
logFC	Log2 fold change	hvgs	Highly variable genes

gene	gene name	UNIPROT summary for the encoded protein
ACVR1, 2A	Activin A Receptor type 1, type 2A	Receptor that mediates the functions of activins, which are growth and differentiation factors which belong to the TGF-β superfamily.
ADAM9, 12, 17	Metallopeptidase Domain 9, 12, 17	May mediate cell-cell, cell-matrix interactions and regulate the motility of cells via interactions with integrins.
ANXA1	Annexin A1	Anti-inflammatory, promotes resolution of inflammation and wound healing.
APP	Amyloid Beta Precursor Protein	cell surface receptor that performs physiological functions on the surface of neurons relevant to neurite growth, neuronal adhesion and axonogenesis.
AREG	Amphiregulin	member of the epidermal growth factor family. Interacts with the EGF/TGF-α receptor to promote the growth of normal epithelial cells.
BMP2, 3, 4, 5, 6	Bone Morphogenetic Protein 2, 3, 4, 5, 6	Growth factor of the TGF-β superfamily that plays essential roles in many developmental processes. Initiates the canonical BMP signalling cascade by associating with the receptors BMPR1A and BMPR2
CADM1	Cell Adhesion Molecule 1	Mediates homophilic cell-cell adhesion. Interaction with CRTAM promotes NK cell cytotoxicity and IFN-γ secretion by $CD8^+$ cells.
CALR	Calreticulin	Resides primarily in the endoplasmic reticulum and is involved in cell adhesion.
CCL2	C-C Motif Chemokine Ligand 2	Involved in immunoregulatory and inflammatory processes. Chemotactic activity for monocytes and basophils but not for neutrophils or eosinophils.
CCL7	C-C Motif Chemokine Ligand 7	Chemokine which attracts macrophages during inflammation and metastasis. In vivo substrate of MMP2, an enzyme which degrades components of the ECM.
CCN1	Cellular Comm. Network Factor 1	Promotes the adhesion of endothelial cells and plays a role in cell proliferation, angiogenesis, apoptosis, and extracellular matrix formation
CCN2	Cellular Comm. Network Factor 2	Mitogen that is secreted by vascular endothelial cells, plays a role in chondrocyte proliferation and differentiation, cell adhesion and is related to PDGF.
CDH1, 2, 3, 4, 7, 11	Cadherin 1, 2, 3, 4, 7, 11	Cadherins are cell adhesion proteins and preferentially interact with themselves. Involved in regulation of cell-cell adhesions, mobility and epithelial proliferation.
CDKN1A	cyclin dependent kinase inhibitor 1A	Regulator of cell cycle progression at G1, tightly controlled by the tumor suppressor p53 in response to stress stimuli.
CDKN2B	cyclin dependent kinase inhibitor 2B	Lies adjacent to CDKN2A in a region that is frequently mutated and deleted in many tumors. Its expression was found to be dramatically induced by TGFβ.
CLDN2	Claudin 2	Claudins are major integral membrane proteins localized exclusively at tight junctions and regulate tissue-specific physiologic properties of tight junctions.
CLU	Clusterin	ECM chaperone that can be found in the cell cytosol under stress conditions. Prevents stress-induced aggregation of blood plasma proteins.
COL1A1, 2	Collagen Type I Alpha 1, 2, Chain	Type I is a fibril-forming collagen found in most connective tissues and is abundant in bone, cornea, dermis and tendon.
COL3A1	Collagen Type III Alpha 1 Chain	Collagen type III occurs in most soft connective tissues along with type I collagen.
COL4A1	Collagen Type IV Alpha 1 Chain	Type IV collagen proteins are integral components of basement membranes.
COL5A1, 2, 3	Collagen Type V Alpha 1, 2, 3 Chain	Type V collagen is a member of group I collagen (fibrillar forming collagen) and binds to DNA, heparan sulfate, thrombospondin, heparin, and insulin.
COL6A2, 3	Collagen Type VI Alpha 2, 3 Chain	Type VI collagen are major structural components of microfibrils and act as a cell-binding protein.
CTHRC1	Collagen Triple Helix Repeat 1	May play a role in the cellular response to arterial injury through involvement in vascular remodeling.
COMP	Cartilage Oligom. Matrix Protein	Non-collagenous ECM protein that may play a role in the structural integrity of cartilage via its interaction with collagens and fibronectin
CXCL1	C-X-C Chemokine Ligand 1	Plays a role in inflammation and as a chemoattractant for neutrophils. Aberrant expression is associated with the growth and progression of certain tumors.
CXCL12	C-X-C Chemokine Ligand 12	Plays a role in many cellular functions, including immune surveillance, inflammation response, tumor growth and metastasis.
CX3CL1	C-X3-C Chemokine Ligand 1	Ligand for CX3CR1 and integrins. Exerts immune response, inflammation, cell adhesion and chemotaxis.
DAG1	Dystroglycan 1	Central component of dystrophin-glycoprotein complex that links the ECM and the cytoskeleton in the skeletal muscle.
DKK1	Dickkopf Wnt Signalling Pathway Inhibitor 1	Antagonizes canonical Wnt signalling. Plays an important role in vertebrate development. by locally inhibiting Wnt regulated processes such as antero-posterior axial patterning and limb development.

gene	gene name	UNIPROT summary for the encoded protein
EDN1, 2	Endothelin 1, 2	Vasoconstrictors whose receptors are targets in the treatment of pulmonary arterial hypertension. Aberrant expression may promote tumorgenesis.
ENPP2	Ectonucleotide Pyrophosphatase 2	Stimulates the motility of tumor cells and has angiogenic properties, its expression is upregulated in several kinds of carcinomas.
DOCK3	Dedicator Of Cytokinesis 3	Suggested to affect the function of small GTPase involved in the regulation of actin cytoskeleton or cell adhesion receptors.
F2RL1	F2R like Trypsin Receptor 1	Generally promoting inflammation, regulates endothelial cell barrier integrity during neutrophil extravasation
FAP	Fibroblast Activation Protein Alpha	Involved in the control of fibroblast growth or epithelial-mesenchymal interactions during development, tissue repair, and epithelial carcinogenesis.
FAS	Fas Cell Surface Death Receptor	Member of the TNF-receptor superfamily. Central role in regulation of programmed cell death, and has been implicated in the pathogenesis various diseases of the immune system
FGF1, 2, 7	Fibroblast Growth Factor 1, 2, 7	FGF family members are involved in embryonic development, cell growth, morphogenesis, tissue repair, tumor growth and invasion.
FN1	Fibronectin 1	Involved in cell adhesion, migration processes, embryogenesis, wound healing, blood coagulation and maintenance of cell shape.
HBEGF	Heparin Binding EGF Like GF	Promotes SMC proliferation and may be involved in macrophage-mediated cell proliferation. Mitogenic for fibroblasts.
HMGB1	High Mobility Group Box 1	In the ECM compartment involved in regulation of the inflammatory response.
ICAM1	Intercellular Adhesion Molecule 1	Expressed on endothelial cells and cells of the immune system, involved in the binding of a cell to another cell or to the ECM.
IL24	Interleukin 24	Mostly synthesized by helper T cells, as well as through monocytes, macrophages, and endothelial cells. Important cytokines of the immune system.
IL32	Interleukin 32	Increased after the activation of T-cells. Induces the production of various cytokines such as TNF-α , IL8 and signal pathways of NFκB and p38 MAPK.
IL33	Interleukin 33	Involved in maturation of Th2 cells and the activation of mast cells, basophils, eosinophils and NK cells. Activates NFκB and MAPK signalling pathway.
INHBA, B	inhibin subunit beta A, B	Member of the TGF-β superfamily. Inhibins appear to oppose the functions of activins.
ITGA2, 3, 5, 8, 9	Integrin Subunit Alpha 2, 3, 5, 8, 9	Alpha subunit of a transmembrane receptor for collagens. Mediates the adhesion of platelets and other cell types to the ECM.
ITGB1, 4, 5, 6, 8	integrin subunit beta 1, 4, 5, 6, 8	Membrane receptors involved in cell adhesion and recognition in including embryogenesis, hemostasis, tissue repair, immune response and metastatis.
JAG1, 2	Jagged Canonical Notch Ligand 1, 2	Ligands for Notch receptors and involved in cell-fate decisions during hematopoiesis, enhances FGF-induced angiogenesis.
JAM2	Junctional Adhesion Molecule 2	Localized at the tight junctions of epithelial and endothelial cells. Acts as an adhesive ligand for interacting with a variety of immune cell types.
KDR	Kinase Insert Domain Receptor	Receptor of VEGF, essential role in the regulation of angiogenesis, vascular development, vascular permeability, and embryonic hematopoiesis. Promotes proliferation, survival, migration and differentiation of endothelial cells.
LAMB/C2	Laminin Subunit beta/gamma 2	Laminins are a family of ECM glycoproteins and the major noncollagenous constituent of basement membranes. Implicated in cell adhesion, differentiation, migration, signalling, neurite outgrowth and metastasis.
MDK	Midkine	Cytokine and GF that mediates inflammatory response, cell proliferation, cell adhesion, cell growth, cell survival, tissue regeneration and migration.
NAMPT	Nicotinamide Phosphoribosyl	Thought to be involved in metabolism, stress response and aging.
NCAM1	Neural Cell Adhesion Molecule	Involved in cell-to-cell/-matrix interactions during development/differentiation and in the expansion of T, B, NK cells.
NOTCH2	Notch Receptor 2, 3	The Notch pathway regulates interactions between adjacent cells. Receptor for membrane-bound ligands JAG1/2, DLL1 to regulate cell-fate determination.
NUPR1	Nuclear Protein 1	Transcription regulator that converts stress signals into a program that empowers cells with resistance to the stress induced environment. Participates in regulation of cell-cycle, apoptosis, autophagy and DNA repair responses.
THBS1, 2	Thrombospondin 1, 2	Adhesive glycoprotein that mediates cell-to-cell/-matrix interactions. Binds to fibrinogen, fibronectin, laminin, type V collagen and integrins.

gene	gene name	UNIPROT summary for the encoded protein
TGFA	Transforming Growth Factor α	Ligands for the EGF receptor, which activates a signalling pathway for cell proliferation, differentiation and development.
TGFB1, 2	Transforming Growth Factor β1, 2	Bind TGF-β receptors leading to activation of SMAD family TFs that regulate gene expression. Regulates cell proliferation, differentiation, and activation of other GFs including IFN-γ and TNF-α. Frequenctly up-regulated in tumor cells.
TNC	Tenascin C	ECM protein acting as ligand for several integrins. Stimulates angiogenesis by elongation, migration and sprouting of endothelial cells in tumors.
OCLN	Occludin	Membrane protein that is required for cytokine-induced regulation of the tight junction paracellular permeability barrier.
PAPPA	Pappalysin 1	Secreted metalloproteinase cleaving IGFBPs resulting in activation of the IGF pathway. Plays a role in bone formation, inflammation and wound healing.
PCDH7, 9	Protocadherin 7, 9	Belong to the protocadherin gene family, a subfamily of the cadherin superfamily.
PDFGB, C	Platelet Derived Growth Factor B, C	GF and potent mitogen for mesenchymal cells. Required for normal proliferation and recruitment of pericytes/vascular SMCs and for normal blood vessel development. Important in wound healing.
PECAM1	Platelet and Endot. Cell Adh. Molecule	Found on the surface of platelets, monocytes, neutrophils, some T-cells, and makes up a large portion of endothelial cell intercellular junctions.
PIK3CB	Phosphatidylinosit [...] Subunit β	Kinase PI3KB is part of the activation pathway in neutrophils which have bound immune complexes at sites of injury or infection.
PLXNA3, PLXNB2	Plexin A3	Semaphorin receptors that may be involved in cytoskeletal remodelling and apoptosis.
POSTN	Periostin	ECM protein that functions in tissue development and regeneration. Binds to integrins to support adhesion and migration of epithelial cells.
PRSS1, 2	Serine Protease 1, 2	Its upregulation is a characteristic feature of pancreatitis. Among its related pathways are Degradation of ECM.
RHOB	Ras Homolog Family Member B	Mediates apoptosis in neoplastically transformed cells after DNA damage and affects cell adhesion and GF signalling in transformed cells. Plays a negative role in tumorigenesis.
SDC3	Syndecan 3	May play a role in the organization of cell shape by affecting the actin cytoskeleton
SERPINE1	Serpin Family E Member 1	As PLAU inhibitor involved in the regulation of cell adhesion and spreading. Required for stimulation of keratinocyte migration during injury repair.
SFRP1, 2	secreted frizzled related protein 1, 2	Modulators of Wnt signalling through direct interaction with Wnts. Regulate cell growth and differentiation in specific cell types, antiproliferative effects on vascular cells.
SEMA3B, C	Semaphorin 3B, C	Semaphorin function in growth cone guidance during neuronal development and have been shown to act as a tumor suppressor by inducing apoptosis.
SOX4	SRY-Box TF 4	Involved in the regulation of embryonic development and determination of the cell fate. May function in the apoptosis pathway leading to cell death as well as to tumorigenesis.
VCAM1	Vascular Cell Adhesion molecule 1	Expressed by cytokine-activated endothelium, mediates leukocyte-endothelial cell adhesion and signal transduction.
VEGFA	Vascular Endothelial Growth Factor	Major GF active in angiogenesis, vasculogenesis and endothelial cell growth. Induces endothelial cell proliferation, promotes cell migration, inhibits apoptosis and induces permeabilization of blood vessels.
VIM	Vimentin	Responsible for maintaining cell shape and cytoskeletal interactions. Organizer of other critical proteins involved in cell attachment, migration, and signalling.
VWF	Von Willebrand Factor	Function in the adhesion of platelets to sites of vascular injury and the transport of various proteins in the blood.
WNT5, 7, 10 9A	Wnt Family Member 5, 7, 9, 10	Act as ligands to activate the different Wnt pathways. Implicated in stem cell control, as a proliferative and self-renewal signal, early development and later during the growth and maintenance of various tissues.

Table 5.1: Selection of genes encoding receptor and ligands of interest based on differential gene expression and encountered during `NicheNet` analysis of human ILD patients. Retrieved and shortened from UniProt.[249]

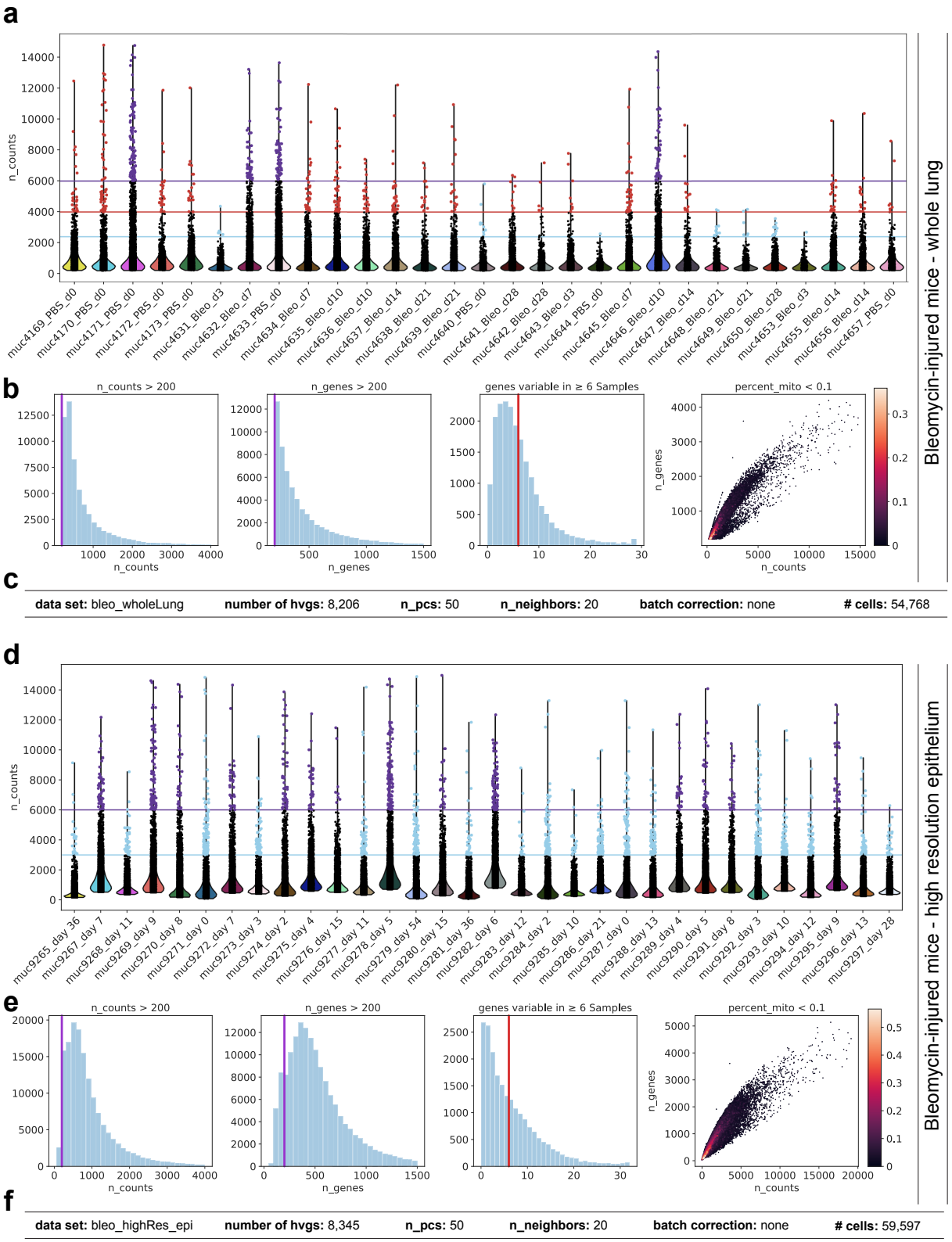


Figure 5.1: Overview of quality metrics and filter thresholds for data sets on bleomycin-treated mice whole lung (upper) and EpCam$^+$ enriched cells (lower). a, d Violin plots to display the distribution of number of transcripts. Coloured dots represent cells that exceeded upper threshold and were removed. **b, e** Histograms of count depth, number of genes per cell and number of genes that are among top 4000 hvgs in given number of samples, as well as scatter plot on these metrics. Threshold are indicated as vertical lines.

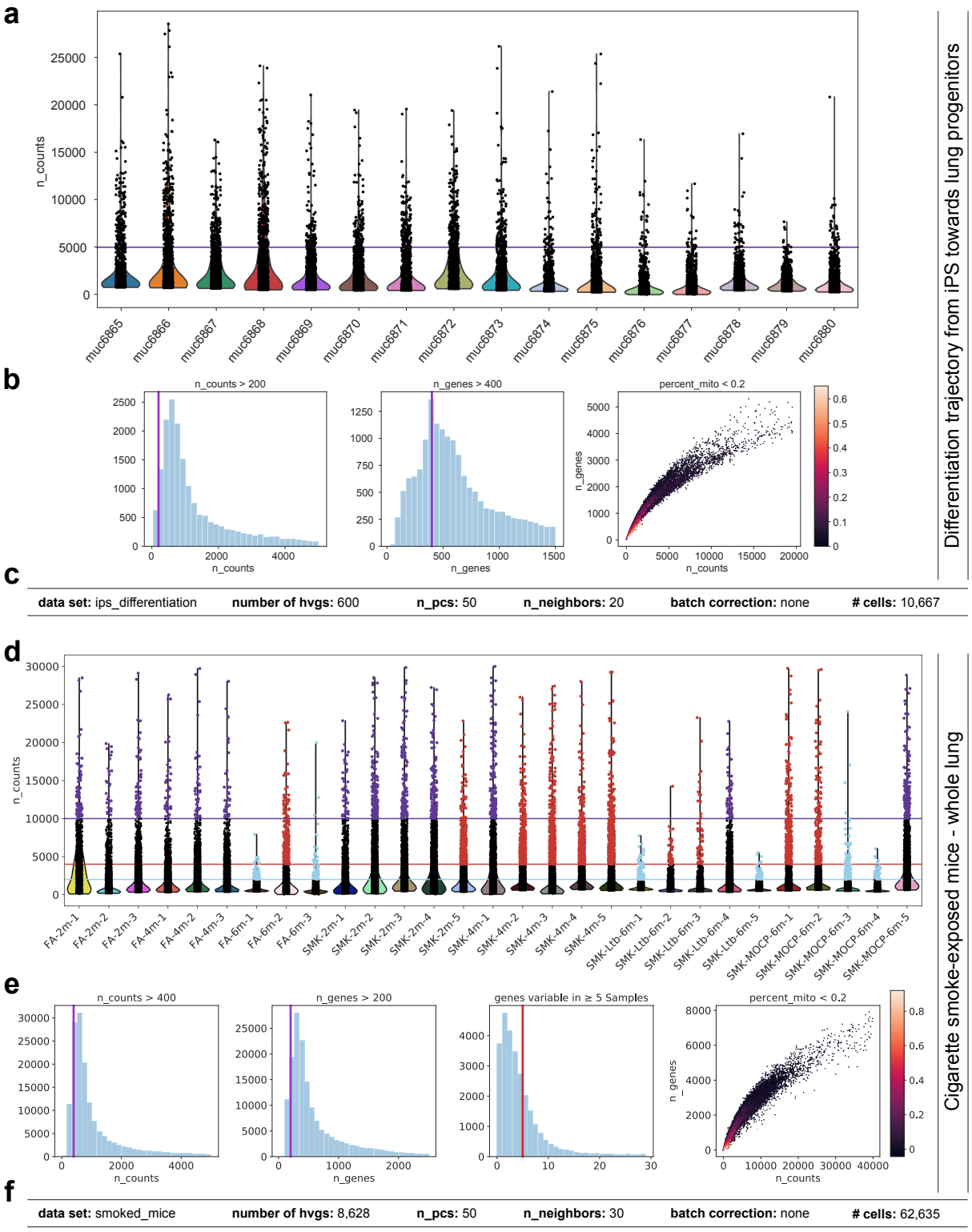


Figure 5.2: Overview of quality metrics and filter thresholds for data sets on iPS to lung differentiation (upper) and cigarette smoke exposed mice (lower). a, d Violin plots to display the distribution of number of transcripts. Coloured dots represent cells that exceeded upper threshold and were removed. **b, e** Histograms of count depth, number of genes per cell and number of genes that are among top 4000 hvgs in given number of samples, as well as scatter plot on these metrics. Threshold are indicated as vertical lines.

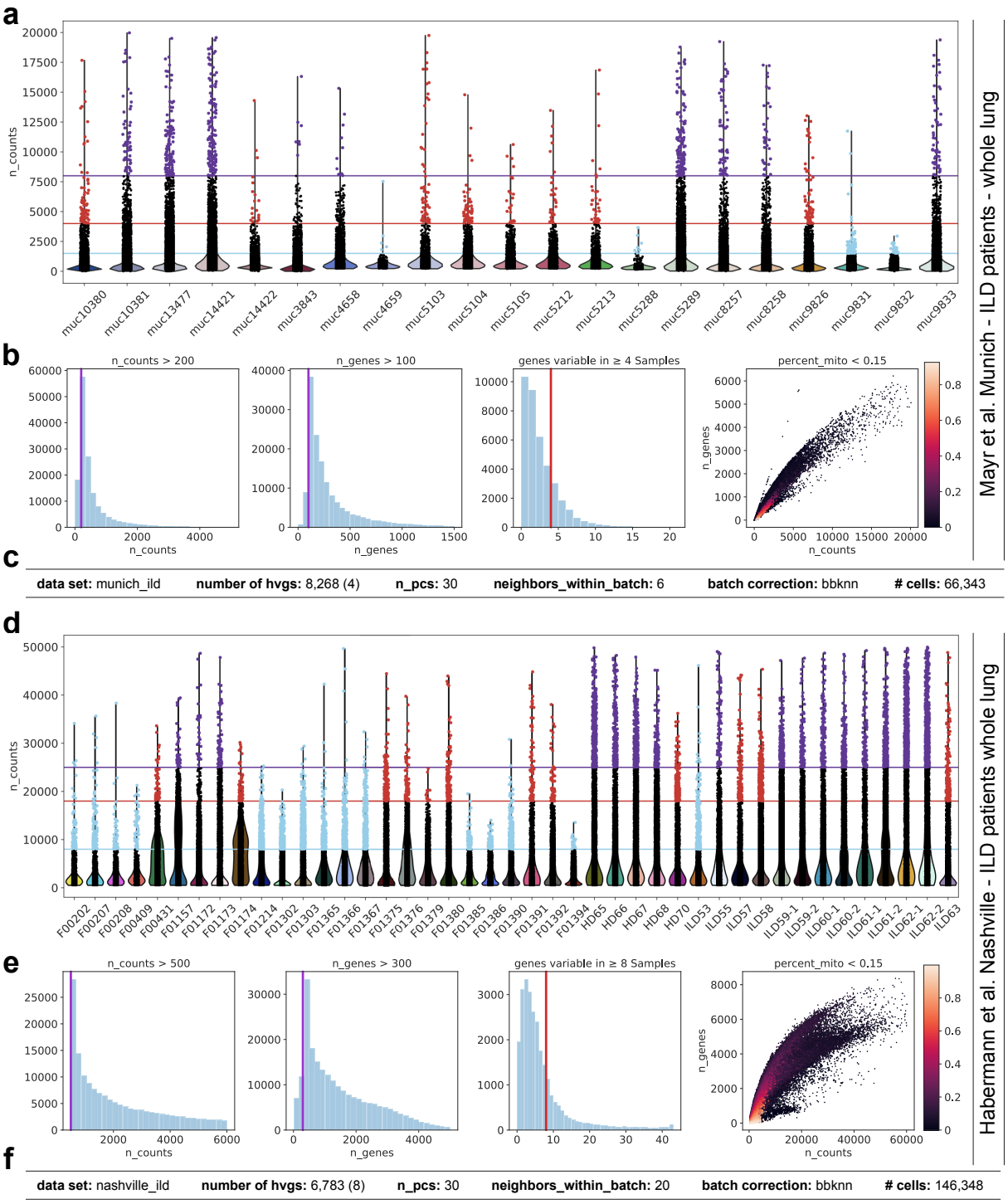


Figure 5.3: Overview of quality metrics and filter thresholds for human lung data sets from Munich (upper) and Nashville cohort (lower). a, d Violin plots to display the distribution of number of transcripts. Coloured dots represent cells that exceeded upper threshold and were removed. **b, e** Histograms of count depth, number of genes per cell and number of genes that are among top 4000 hvgs in given number of samples, as well as scatter plot on these metrics. Threshold are indicated as vertical lines.

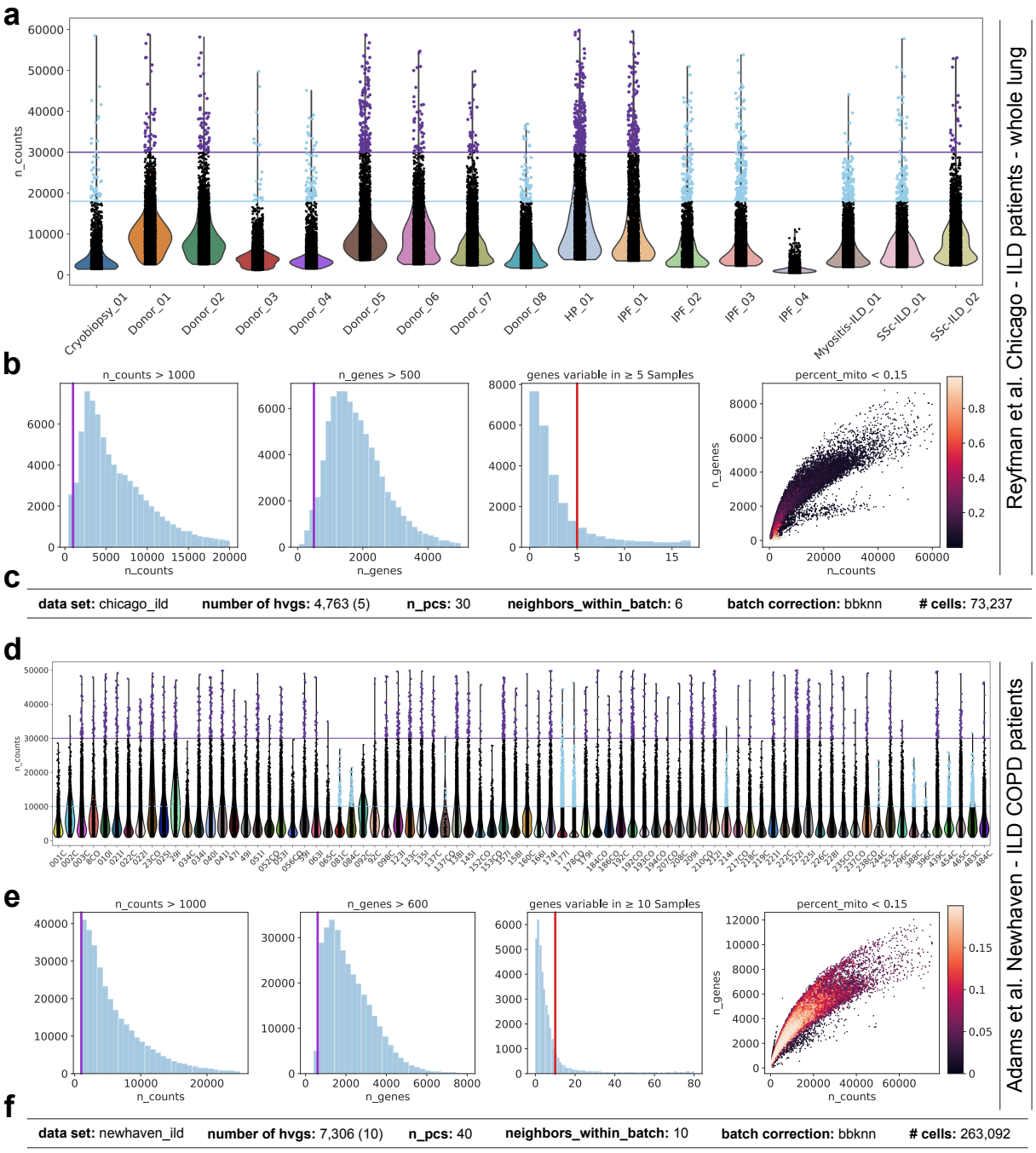


Figure 5.4: Overview of quality metrics and filter thresholds for human lung data sets from Chicago (upper) and Newhaven cohort (lower). a, d Violin plots to display the distribution of number of transcripts. Coloured dots represent cells that exceeded upper threshold and were removed. **b, e** Histograms of count depth, number of genes per cell and number of genes that are among top 4000 hvgs in given number of samples, as well as scatter plot on these metrics. Threshold are indicated as vertical lines.

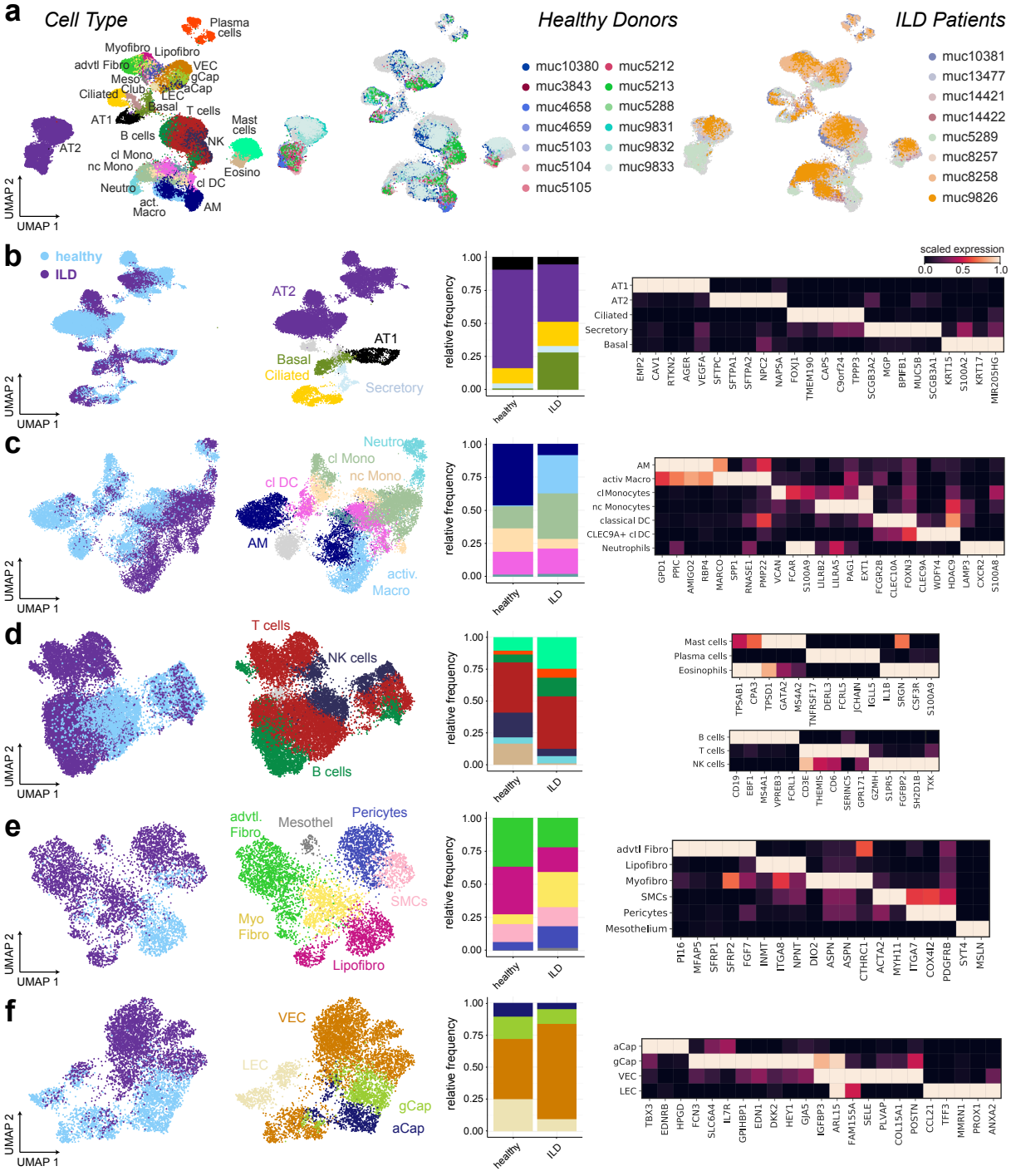


Figure 5.5: Compartment-wise annotation of cells from healthy donors and ILD patients of the Munich cohort. a UMAP coloured by cell type (left) and split view separated by healthy donors (middle) and patients diagnosed with (right). **b-f** Cell type annotation, relative composition within compartment and literature-derived marker genes of the epithelium (b), mononuclear phagocytes (c), lymphocytes and granulocytes (d), mesenchyme (e) and endothelium (f). Light gray cells from the cell type coloured UMAPs were estimated to be of low-quality and were excluded from further analysis.

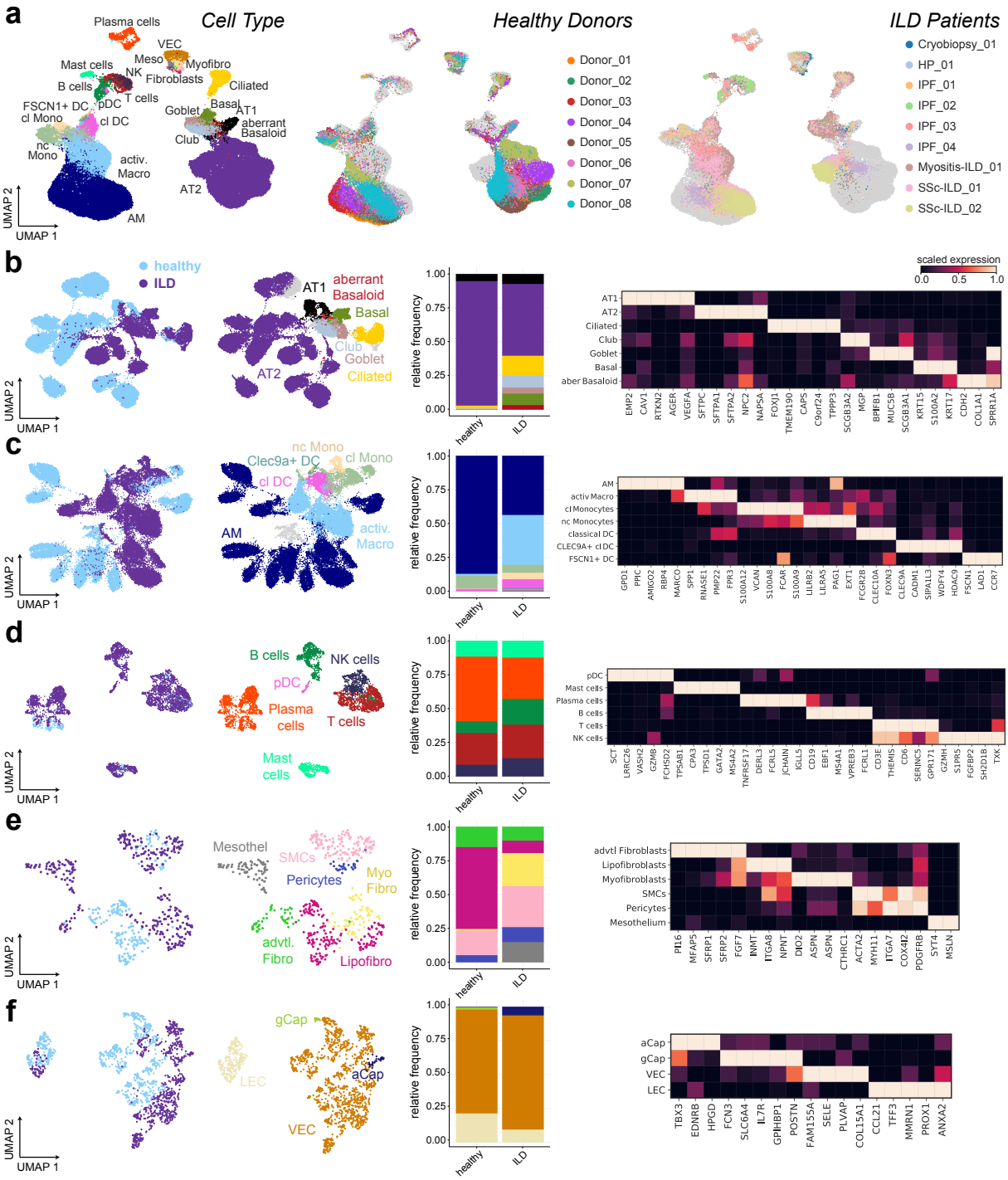


Figure 5.6: Compartment-wise annotation of cells from healthy donors and ILD patients of the Chicago cohort. **a** UMAP coloured by cell type (left) and split view separated by healthy donors (middle) and patients diagnosed with (right). **b-f** Cell type annotation, relative composition within compartment and literature-derived marker genes of the epithelium (b), mononuclear phagocytes (c), lymphocytes and granulocytes (d), mesenchyme (e) and endothelium (f). Light gray cells from the cell type coloured UMAPs were estimated to be of low-quality and were excluded from further analysis.

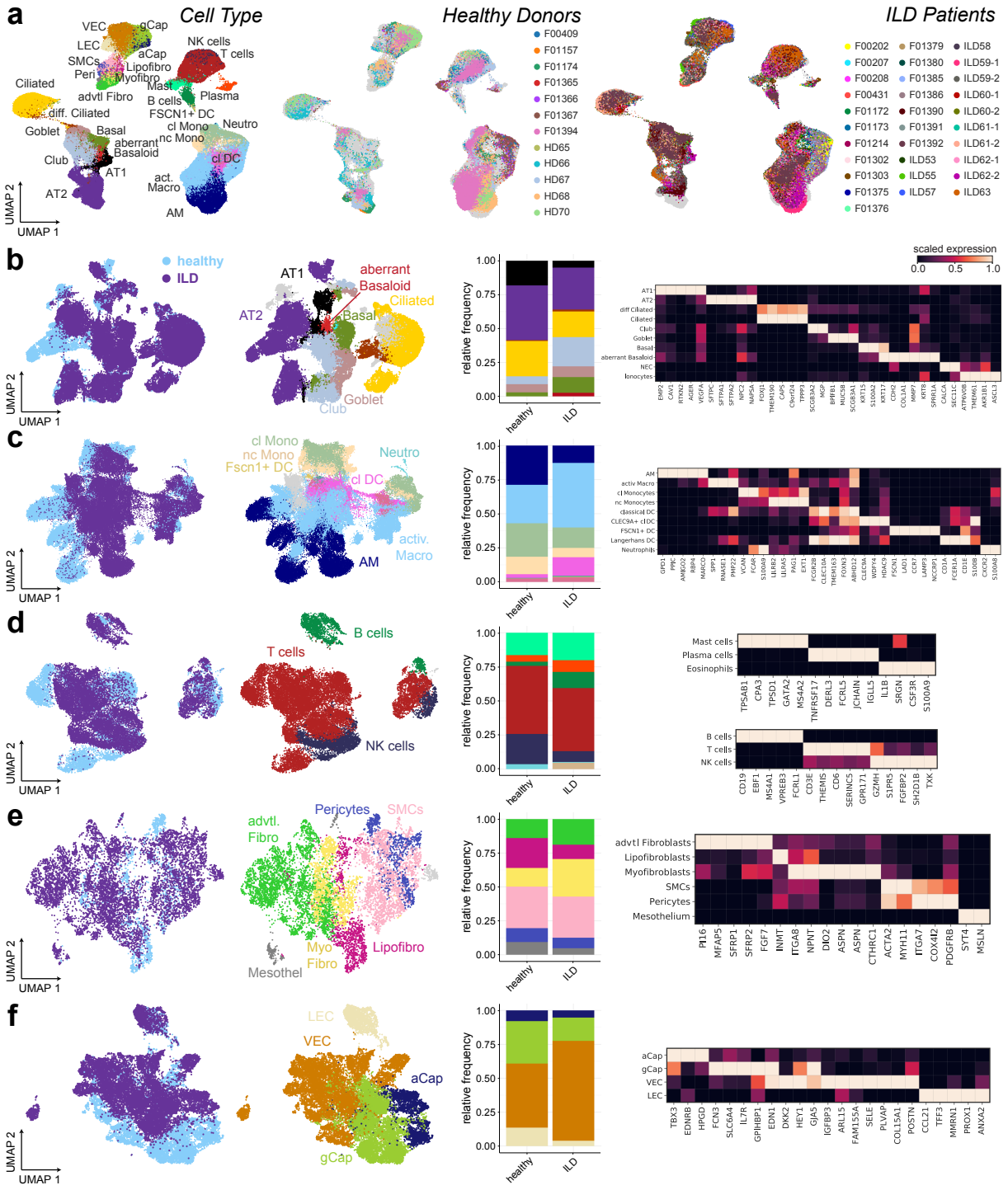


Figure 5.7: Compartment-wise annotation of cells from healthy donors and ILD patients of the Nashville cohort. **a** UMAP coloured by cell type (left) and split view separated by healthy donors (middle) and patients diagnosed with (right). **b-f** Cell type annotation, relative composition within compartment and literature-derived marker genes of the epithelium (b), mononuclear phagocytes (c), lymphocytes and granulocytes (d), mesenchyme (e) and endothelium (f). Light gray cells from the cell type coloured UMAPs were estimated to be of low-quality and were excluded from further analysis.

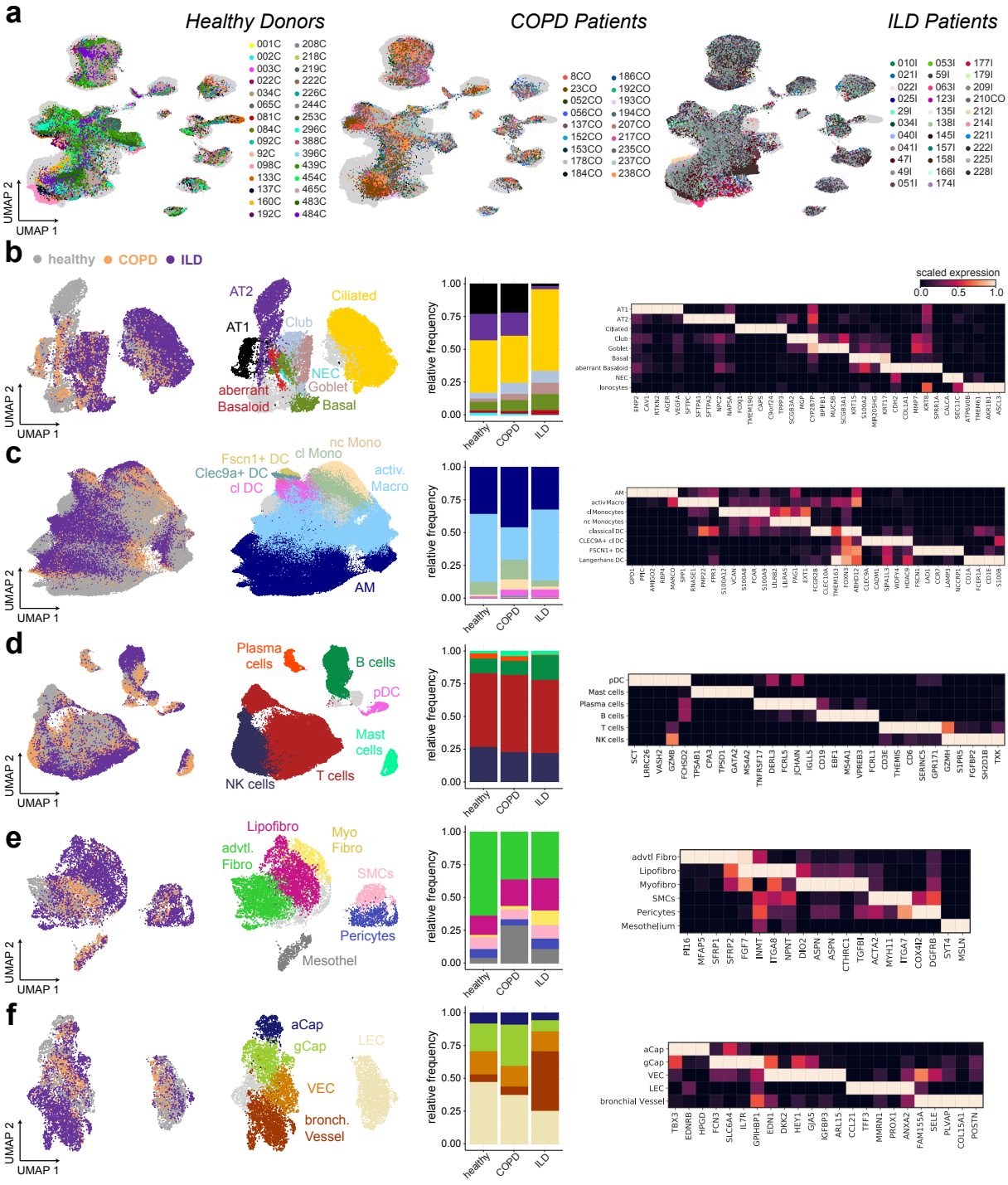


Figure 5.8: Compartment-wise annotation of cells from healthy donors and ILD patients of the Newhaven cohort. **a** Split UMAP separated by healthy donors (left), patients diagnosed with either COPD (middle) or ILD (right). **b-f** Cell type annotation, relative composition within compartment and literature-derived marker genes of the epithelium (b), mononuclear phagocytes (c), lymphocytes and granulocytes (d), mesenchyme (e) and endothelium (f). Light gray cells from the cell type coloured UMAPs were estimated to be of low-quality and were excluded from further analysis.

Bibliography

[1] Leigh Van Valen. *A new evolutionary law*. Evolutionary Theory, 1973, 1–30.

[2] Gibson LJ. "Cork: Structure, Properties, Applications." In: (2016).

[3] Gest H. "The discovery of microorganisms by Robert Hooke and Antoni Van Leeuwenhoek, fellows of the Royal Society." In: *Notes Rec R Soc Lond.* 58 (2004), pp. 187–201. DOI: 10.1098/rsnr.2004.0055.

[4] Tero AC. *Achiever's Biology*. Allied Publishers, 1990, p. 36.

[5] Mazzarello PA. "Unifying concept: the history of cell theory." In: *Nat Cell Biol* 1 (1999). DOI: https://doi.org/10.1038/8964.

[6] Bianconi E, Piovesan A, Facchin F, Beraudi A, Casadei R, Frabetti F, Vitale L, Pelleri MC, Tassani S, Piva F, Perez-Amodio S, Strippoli P, and Canaider S. "An estimation of the number of cells in the human body." In: *Ann Hum Biol.* 40 (2013), pp. 463–71. DOI: 10.3109/03014460.2013.807878.

[7] LC Junqueria, LCU Junqueira, and RO Kelley. *Basic histology*. McGraw-Hill/Appleton Lange, 1995.

[8] H Rudenberg and PG Rudenberg. *Origin and Background of the Invention of the Electron Microscope. Advances in Imaging and Electron Physics*. 2010, 207–286.

[9] Coons† AH, Creech HJ, and Jones RN. "Immunological Properties of an Antibody Containing a Fluorescent Group." In: *Exp. Biol. Med.* 42 (1941), pp. 200–202. DOI: https://doi.org/10.3181/00379727-47-13084P.

[10] Arthur G. "Albert Coons: harnessing the power of the antibody." In: *Lancet Respir Med.* 4 (2016), pp. 181–2. DOI: 10.1016/S2213-2600(16)00020-5.

[11] Fulwyler MJ. "Electronic separation of biological cells by volume." In: *Science* 150 (1965), pp. 910–1. DOI: 10.1126/science.150.3698.910.

[12] Bonner WA, Hulett HR, Sweet RG, and Herzenberg LA. "Fluorescence activated cell sorting." In: *Rev Sci Instrum.* 43 (1972), pp. 404–9. DOI: 10.1063/1.1685647.

[13] Picot J, Guerin CL, Le Van Kim C, and Boulanger CM. "Flow cytometry: retrospective, fundamentals and recent instrumentation." In: *Cytotechnology* 64 (2012), pp. 109–30. DOI: 10.1007/s10616-011-9415-0.

[14] Regev A et al. "Human Cell Atlas Meeting Participants." In: *Elife* 6 (2017). DOI: 10.7554/eLife.27041.

[15] Andrews TS and Hemberg M. "Identifying cell populations with scRNASeq." In: *Mol Aspects Med.* 59 (2018), pp. 114–122. DOI: 10.1016/j.mam.2017.07.002.

[16] Zamorano PL, Mahesh VB, and Brann DW. "Quantitative RT-PCR for neuroendocrine studies. A minireview." In: *Neuroendocrinology* 63 (1996), pp. 397–407. DOI: 10.1159/000127065.

[17] Alberts B, Johnson A, Lewis J, Raff M, and Roberts K. *Molecular biology of the cell.* Garland Science, 2008.

[18] Wang YC, Peterson SE, and Loring JF. "Protein post-translational modifications and regulation of pluripotency in human stem cells." In: *Cell Res.* 24 (2014), pp. 143–60. DOI: 10.1038/cr.2013.151.

[19] Alwine JC and Kemp DJ nd Stark GR. "Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes." In: *Proc Natl Acad Sci USA* 74 (1977), pp. 5350–4. DOI: 10.1073/pnas.74.12.5350.

[20] Trayhurn P. "Northern blotting." In: *Proc Nutr Soc.* 55 (1996), pp. 583–9. DOI: 10.1079/pns19960051.

[21] Mullis K, Faloona F, Scharf S, Saiki R, Horn G, and Erlich H. "pecific enzymatic amplification of DNA in vitro: the polymerase chain reaction." In: *S Cold Spring Harb Symp Quant Biol.* 51 (1986). DOI: 10.1101/sqb.1986.051.01.032.

[22] Medrano JF Wong ML. "Real-time PCR for mRNA quantitation." In: *Biotechniques* 39 (2005), pp. 75–85. DOI: 10.2144/05391RV01.

[23] Pardue ML and Gall JG. "Molecular hybridization of radioactive DNA to the DNA of cytological preparations." In: *Proc Natl Acad Sci USA* 64 (1969), pp. 4381–5. DOI: 10.1073/pnas.64.2.600.

[24] Rudkin G and Stollar B. "High resolution detection of DNA–RNA hybrids in situ by indirect immunofluorescence." In: *Nature* 265 (1977), 472–473. DOI: https://doi.org/10.1038/265472a0.

[25] O'Connor C. "Fluorescence in situ hybridization (FISH).". In: *Nature Education* 1 (2008), p. 171.

[26] Schena M, Shalon D, and Brown PO Davis RW. "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." In: *Science* 270 (1995), pp. 467–70. DOI: 10.1126/science.270.5235.467.

[27] Shalon D, Smith SJ, and Brown PO. "A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization." In: *Genome Res.* 6 (1996), pp. 639–45. DOI: 10.1101/gr.6.7.639.

[28] Lobenhofer EK, Bushel PR, Afshari CA, and Hamadeh HK. "Progress in the application of DNA microarrays." In: *Environ Health Perspect* 109 (2001), pp. 881–91. DOI: 10.1289/ehp.01109881.

[29] Wang C, Gong B, Bushel PR, Thierry-Mieg J, Thierry-Mieg D, Xu J, and Fang H. "The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance." In: *Nat Biotechnol.* 32 (2014), pp. 926–32. DOI: 10.1038/nbt.3001.

[30] Li J, Hou R, Niu X, Liu R, Wang Q, Wang C, Li X, Hao Z, Yin G, and Zhang K. "TComparison of microarray and RNA-Seq analysis of mRNA expression in dermal mesenchymal stem cells." In: *Biotechnol Lett.* 38 (2016), pp. 33–41. DOI: 10.1007/s10529-015-1963-5.

[31] Wang Z, Gerstein M, and Snyder M. "RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis." In: *Annual Review of Biomedical Data Science* 2 (2019), pp. 139–173. DOI: https://doi.org/10.1146/annurev-biodatasci-072018-021255.

[32] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, and Devon K. "International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome." In: *Nature* 409 (2001), pp. 860–921. DOI: 10.1038/3505706.

[33] Wang Y and Navin NE. "Advances and applications of single-cell sequencing technologies." In: *Mol Cell.* 58 (2015), pp. 598–609. DOI: 10.1016/j.molcel.2015.05.005.

[34] WLevsky JM and Singer RH. "Gene expression and the myth of the average cell." In: *Trends Cell Biol.* 13 (2003), pp. 4–6. DOI: 0.1016/s0962-8924(02)00002-8.

[35] Huang S. "Non-genetic heterogeneity of cells in development: more than just noise." In: *Development* 136 (2009), pp. 3853–62. DOI: 10.1242/dev.035139.

[36] Eldar A and Elowitz MB. "Functional roles for noise in genetic circuits." In: *Nature* 467 (2010), pp. 167–73. DOI: 10.1038/nature09326.

[37] 10x Genomics. *Single-Cell RNA-Seq: An Introductory Overview and Tools for Getting Started*. https://www.10xgenomics.com/blog/single-cell-rna-seq-an-introductory-overview-and-tools-for-getting-started. Accessed November 4, 2022.

[38] Trapnell C. "Defining cell types and states with single-cell genomics." In: *Genome Res.* 10 (2015), pp. 1491–8. DOI: 10.1101/gr.190595.115.

[39] Kapellos TS, Bonaguro L, Gemünd I, Reusch N, Saglam A, Hinkley ER, and Schultze JL. "Human Monocyte Subsets and Phenotypes in Major Chronic Inflammatory Diseases." In: *Front Immunol.* 10 (2019), p. 2035. DOI: 10.3389/fimmu.2019.02035.

[40] Waddington CH. *The Strategy of the Genes, a Discussion of Some Aspects of Theoretical Biology.* London: Allen Unwin, 1957.

[41] Darmanis S, Sloan SA, Croote D, Mignardi M, Chernikova S, Samghababi P, Zhang Y, Neff N, Kowarsky M, Caneda C, Li G, Chang SD, Connolly ID, Li Y, Barres BA, Gephart MH, and Quake SR. "Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma." In: *Cell Rep.* 21 (2017), pp. 1399–1410. DOI: 10.1016/j.celrep.2017.10.030.

[42] Liu SJ, Nowakowski TJ, Pollen AA, Lui JH, Horlbeck MA, Attenello FJ, He D, Weissman JS, Kriegstein AR, Diaz AA, and Lim DA. "Single-cell analysis of long non-coding RNAs in the developing human neocortex." In: *Genome Biol.* 17 (2016), p. 67. DOI: 10.1186/s13059-016-0932-1.

[43] Marques S, Zeisel A, Codeluppi S, van Bruggen D, Mendanha Falcão A, Xiao L, Li H, Häring M, Hochgerner H, Romanov RA, Gyllborg D, Muñoz Manchado A, La Manno G, Lönnerberg P, Floriddia EM, Rezayee F, Ernfors P, Arenas E, Hjerling-Leffler J, Harkany T, Richardson WD, Linnarsson S, and Castelo-Branco G. "Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system." In: *Science* 352 (2016), pp. 1326–1329. DOI: 10.1126/science.aaf6463.

[44] Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, and McCarroll SA. "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets." In: *Cell* 161 (2015), pp. 1202–1214. DOI: 10.1016/j.cell.2015.05.002.

[45] Belgrader P Ryvkin P Bent ZW Wilson R Ziraldo SB Wheeler TD McDermott GP Zhu J Gregory MT Shuga J Montesclaros L Underwood JG Masquelier DA Nishimura SY Schnall-Levin M Wyatt PW Hindson CM Bharadwaj R Wong A Ness KD Beppu LW Deeg HJ McFarland C Loeb KR Valente WJ Ericson NG Stevens EA Radich JP Mikkelsen TS Hindson BJ Bielas JH Zheng GX Terry JM. "Massively parallel digital transcriptional profiling of single cells." In: *Nat Commun.* 8 (2017), pp. 140–49. DOI: 10.1038/ncomms14049.

[46] Dharmat R, Kim S, Li Y, and Chen R. "Single-Cell Capture, RNA-seq, and Transcriptome Analysis from the Neural Retina." In: *Methods Mol Biol* 2092 (2020), pp. 159–186. DOI: 10.1007/978-1-0716-0175-4_12.

[47] Wang YJ, Schug J, Won KJ, Liu C, Naji A, Avrahami D, Golson ML, and Kaestner KH. "Single-Cell Transcriptomics of the Human Endocrine Pancreas." In: *Diabetes* 65 (2016), pp. 3028–38. DOI: 10.2337/db16-0405.

[48] Segerstolpe Å, Palasantza A, Eliasson P, Andersson EM, Andréasson AC, Sun X, Picelli S, Sabirsh A, Clausen M, Bjursell MK, Smith DM, Kasper M, Ämmälä C, and Sandberg R. "Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes." In: *Cell Metab.* 24 (2016), pp. 593–607. DOI: 10.1016/j.cmet.2016.08.020.

[49] Tritschler S, Theis FJ, and Böttcher A Lickert H. "Systematic single-cell analysis provides new insights into heterogeneity and plasticity of the pancreas." In: *Mol Metab.* 6 (2017), pp. 974–990. DOI: 10.1016/j.molmet.2017.06.021.

[50] Biase FH, Cao X, and Zhong S. "Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing." In: *Genome Res.* 24 (2014), pp. 974–990. DOI: 10.1101/gr.177725.114.

[51] Goolam M, Scialdone A, Graham SJL, Macaulay IC, Jedrusik A, Hupalowska A, Voet T, Marioni JC, and Zernicka-Goetz M. "Heterogeneity in Oct4 and Sox2 Targets Biases Cell Fate in 4-Cell Mouse Embryos." In: *Cell* 165 (2016), pp. 61–74. DOI: 10.1016/j.cell.2016.01.047.

[52] Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, and Surani MA. "mRNA-Seq whole-transcriptome analysis of a single cell." In: *Nat Methods.* 6 (2009), pp. 377–82. DOI: 10.1038/nmeth.1315.

[53] Eberwine J, Yeh H, Miyashiro K, Cao Y, Nair S, Finnell R, Zettel M, and Coleman P. "Analysis of gene expression in single live neurons." In: *Proc Natl Acad Sci USA* 89 (1992), pp. 3010–4. DOI: 10.1073/pnas.89.7.3010.

[54] Svensson V, Vento-Tormo R, and Teichmann SA. "Exponential scaling of single-cell RNA-seq in the past decade." In: *Nat Protoc.* 13 (2018), pp. 599–604. DOI: 10.1038/nprot.2017.149.

[55] Hwang B, Lee JH, and Bang D. "Single-cell RNA sequencing technologies and bioinformatics pipelines." In: *Exp Mol Med.* 50 (2018), p. 96. DOI: 10.1038/s12276-018-0071-8.

[56] Papalexi E and Satija R. "Single-cell RNA sequencing to explore immune cell heterogeneity." In: *Nat Rev Immunol.* 18 (2018), pp. 35–45. DOI: 10.1038/nri.2017.76.

[57] Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, and Kirschner MW. "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells." In: *Cell* 161 (2015), pp. 187–1201. DOI: 10.1016/j.cell.2015.04.044.

[58] Agresti JJ, Antipov E, Abate AR, Ahn K, Rowat AC, Baret JC, Marquez M, Klibanov AM, Griffiths AD, and Weitz DA. "Ultrahigh-throughput screening in drop-based microfluidics for directed evolution." In: *Proc Natl Acad Sci USA* 107 (2010), pp. 4004–9. DOI: 10.1073/pnas.0910781107.

[59] Zhang X, Li T, Liu F, Chen Y, Yao J, Li Z, Huang Y, and Wang J. "Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems." In: *Mol Cell* 73 (2019), pp. 130–142. DOI: 10.1016/j.molcel.2018.10.020.

[60] Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, Lönnerberg P, and Linnarsson S. "Quantitative single-cell RNA-seq with unique molecular identifiers". In: *Nat Methods* 11 (2014), pp. 163–6. DOI: 10.1038/nmeth.2772.

[61] Soumillon M, Cacchiarelli D, Semrau S, van Oudenaarden A, and Mikkelsen TS. "Characterization of directed differentiation by high-throughput single-cell RNA-Seq." In: *bioRxiv Preprint* (2014). DOI: 10.1101/003236.

[62] Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, Mildner A, Cohen N, Jung S, Tanay A, and Amit I. "Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types." In: *Science* 343 (2014), pp. 776–9. DOI: 10.1126/science.1247651.

[63] Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, and Sandberg R. "Smart-seq2 for sensitive full-length transcriptome profiling in single cells." In: *Nat Methods* 10 (2013), pp. 1096–8. DOI: 10.1038/nmeth.2639.

[64] Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, Leonhardt H, Heyn H, Hellmann I, and Enard W. "Comparative Analysis of Single-Cell RNA Sequencing Methods." In: *Mol Cell.* 65 (2017), pp. 631–643. DOI: 10.1016/j.molcel.2017.01.023.

[65] Hershel Raff and Michael G. Levitzky. *Medical Physiology: A Systems Approach*. Lange Medical Books, 2011.

[66] openstax. *Anatomy and Physiology - Organs and Structures of the Respiratory System*. https://assets.openstax.org/oscms-prodcms/media/documents/AnatomyandPhysiology-OP.pdf. Accessed November 4, 2022.

[67] Nikolić MZ, Sun D, and Rawlins EL. "Human lung development: recent progress and new challenges." In: *Development* 145 (2018). DOI: 10.1242/dev.163485.

[68] Travaglini KJ, Nabhan AN, Penland L, Sinha R, Gillich A, Sit RV, Chang S, Conley SD, Mori Y, Seita J, Berry GJ, Shrager JB, Metzger RJ, Kuo CS, Neff N, Weissman IL, Quake SR, and Krasnow MA. "A molecular cell atlas of the human lung from single-cell RNA sequencing." In: *Nature* 587 (2020), pp. 619–625. DOI: 10.1038/s41586-020-2922-4.

[69] Rock JR and Hogan BL. "Epithelial progenitor cells in lung development, maintenance, repair, and disease." In: *Annu Rev Cell Dev Biol.* 27 (2011), pp. 493–512. DOI: 10.1146/annurev-cellbio-100109-104040.

[70] Tata PR and Rajagopal J. "Plasticity in the lung: making and breaking cell identity." In: *Development* 144 (2017), pp. 755–766. DOI: 10.1242/dev.143784.

[71] Watson JK, Rulands S, Wilkinson AC, Wuidart A, Ousset M, Van Keymeulen A, Göttgens B, Blanpain C, Simons BD, and Rawlins EL. "Clonal Dynamics Reveal Two Distinct Populations of Basal Cells in Slow-Turnover Airway Epithelium." In: *Cell Rep.* 12 (2015), pp. 90–101. DOI: 10.1016/j.celrep.2015.06.011.

[72] Hogan BL, Barkauskas CE, Chapman HA, Epstein JA, Jain R, Hsia CC, Niklason L, Calle E, Le A, Randell SH, Rock J, Snitow M, Krummel M, Stripp BR, Vu T, White ES, Whitsett JA, and Morrisey EE. "Repair and regeneration of the respiratory system: complexity, plasticity, and mechanisms of lung stem cell function." In: *Cell Stem Cell* 15 (2014), pp. 123–38. DOI: 10.1016/j.stem.2014.07.012.

[73] Reynolds SD, Reynolds PR, Pryhuber GS, Finder JD, and Stripp BR. "Secretoglobins SCGB3A1 and SCGB3A2 define secretory cell subsets in mouse and human airways." In: *Am J Respir Crit Care Med.* 166 (2002), pp. 1498–509. DOI: 10.1164/rccm.200204-285OC.

[74] Adler KB, Tuvim MJ, and Dickey BF. "Regulated mucin secretion from airway epithelial cells." In: *Front Endocrinol.* 18 (2013), p. 129. DOI: 10.3389/fendo.2013.00129.

[75] Rawlins EL, Okubo T, Xue Y, Brass DM, Auten RL, Hasegawa H, Wang F, and Hogan BL. "The role of Scgb1a1+ Clara cells in the long-term maintenance and repair of lung airway,

but not alveolar, epithelium." In: *Cell Stem Cell* 4 (2009), pp. 525–34. DOI: 10.1016/j.stem.2009.04.002.

[76] Tata PR, Mou H, Pardo-Saganta A, Zhao R, Prabhu M, Law BM, Vinarsky V, Cho JL, Breton S, Sahay A, Medoff BD, and Rajagopal J. "Dedifferentiation of committed epithelial cells into stem cells in vivo." In: *Nature* 503 (2013), pp. 218–23. DOI: 10.1038/nature12777.

[77] Boucherat O, Boczkowski J, Jeannotte L, and Delacourt C. "Cellular and molecular mechanisms of goblet cell metaplasia in the respiratory airways." In: *Exp Lung Res.* 39 (2013), pp. 207–16. DOI: 10.3109/01902148.2013.791733.

[78] Schiller HB, Montoro DT, Simon LM, Rawlins EL, Meyer KB, Strunz M, Vieira Braga FA, Timens W, Koppelman GH, Budinger GRS, Burgess JK, Waghray A, van den Berge M, Theis FJ, Regev A, Kaminski N, Rajagopal J, Teichmann SA, Misharin AV, and Nawijn MC. "The Human Lung Cell Atlas: A High-Resolution Reference Map of the Human Lung in Health and Disease." In: *Am J Respir Cell Mol Biol.* 61 (2019). DOI: 10.1165/rcmb.2018-0416TR.

[79] Guha A, Vasconcelos M amd Cai Y, Yoneda M, Hinds A, Qian J, Li G, Dickel L, Johnson JE, Kimura S, Guo J, McMahon J, McMahon AP, and Cardoso WV. "Neuroepithelial body microenvironment is a niche for a distinct subset of Clara-like precursors in the developing airways." In: *Proc Natl Acad Sci USA* 109 (2012), pp. 12592–7. DOI: 10.1073/pnas.1204710109.

[80] Colby TV Mason R Williams MC Herzog EL Brody AR. "Knowns and unknowns of the alveolus." In: *Proc Am Thorac Soc.* 5 (2008), pp. 778–82. DOI: 10.1513/pats.200803-028HR.

[81] Evans MJ, Cabral LJ, Stephens RJ, and Freeman G. "Transformation of alveolar type 2 cells to type 1 cells following exposure to NO2." In: *Exp Mol Pathol.* 1 (2013), pp. 142–50. DOI: 10.1016/0014-4800(75)90059-3.

[82] Barkauskas CEa, Cronce MJ, Rackley CR, Bowie EJ, Keene DR, Stripp BR, Randell SH, Noble PW, and Hogan BL. "Type 2 alveolar cells are stem cells in adult lung." In: *J Clin Invest.* 123 (2013), pp. 3025–36. DOI: 10.1172/JCI68782.

[83] Desai TJ, Brownfield DG, and Krasnow MA. "Alveolar progenitor and stem cells in lung development, renewal and cancer." In: *Nature* 507 (2014), pp. 190–4. DOI: 10.1038/nature12930.

[84] Kobayashi Y, Tata A, Konkimalla A, Katsura H, Lee RF, Ou J, Banovich NE, Kropski JA, and Tata PR. "Persistence of a regeneration-associated, transitional alveolar epithelial cell state in pulmonary fibrosis." In: *Nat Cell Biol.* 22 (2020), pp. 934–946. DOI: 10.1038/s41556-020-0542-8.

[85] Clevers H, Loh KM, and Nusse R. "Stem cell signaling. An integral program for tissue renewal and regeneration: Wnt signaling and stem cell control." In: *Science* 346 (2014). DOI: 10.1126/science.1248012.

[86] Habermann AC, Gutierrez AJ, Bui LT, Yahn SL, Winters NI, Calvi CL, Peter L, Chung MI, Taylor CJ, Jetter C, Raju L, Roberson J, Ding G, Wood L, Sucre JMS, Richmond BW, Serezani AP, McDonnell WJ, Mallal SB, Bacchetta MJ, Loyd JE, Shaver CM, Ware LB, Bremner R, Walia R, Blackwell TS, Banovich NE, and Kropski JA. "Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis." In: *Sci Adv.* 6 (2020), p. 1972. DOI: 10.1126/sciadv.aba1972.

[87] Kameritsch P and Renkawitz J. "Principles of Leukocyte Migration Strategies." In: *Trends Cell Biol.* 30 (2020), pp. 818–832. DOI: 10.1016/j.tcb.2020.06.007.

[88] Maton D, Hopkins J, McLaughlin CW, Johnson S, Warner MQ, LaHart D, Wright JD, and Kulkarni DV. *Human Biology and Health.* Englewood Cliffs, 1997.

[89] Németh T, Sperandio M, and Mócsai A. "Neutrophils as emerging therapeutic targets." In: *Nat Rev Drug Discov.* 4 (2020), pp. 253–275. DOI: 10.1038/s41573-019-0054-z.

[90] Németh T and Mócsai A. "The role of neutrophils in autoimmune diseases." In: *Immunol Lett.* 143 (2012), pp. 9–19. DOI: 10.1016/j.imlet.2012.01.013.

[91] Coffelt SB, Wellenstein MD, and de Visser KE. "Neutrophils in cancer: neutral no more." In: *Nat Rev Cancer* 16 (2016), pp. 431–46. DOI: 10.1038/nrc.2016.52.

[92] Klion AD, Ackerman SJ, and Bochner BS. "Contributions of Eosinophils to Human Health and Disease." In: *Annu Rev Pathol.* 15 (2020), pp. 179–209. DOI: 10.1146/annurev-path/mechdis-012419-032756.

[93] Weller PF and Spencer LA. "Functions of tissue-resident eosinophils." In: *Nat Rev Immunol.* 17 (2017), pp. 746–760. DOI: 10.1038/nri.2017.95.

[94] Falcone FH, Haas H, and Gibbs BF. "The human basophil: a new appreciation of its role in immune responses." In: *Blood* 96 (2000), pp. 4028–38.

[95] Geissmann F, Manz MG, Jung S, Sieweke MH, Merad M, and Ley K. "Development of monocytes, macrophages, and dendritic cells." In: *Science* 327 (2010), pp. 656–61. DOI: 10.1126/science.1178331.

[96] Fogg DK, Sibon C, Miled C, Jung S, Aucouturier P, Littman DR, Cumano A, and Geissmann F. "A clonogenic bone marrow progenitor specific for macrophages and dendritic cells." In: *Science* 311 (2006), pp. 83–7. DOI: 10.1126/science.1117729.

[97] Pittet MJ, Nahrendorf M, and Swirski FK. "The journey from stem cell to macrophage." In: *Ann NY Acad Sci* 1319 (2014), pp. 1–18. DOI: 10.1111/nyas.12393.

[98] Auffray C, Sieweke MH, and Geissmann F. "Blood monocytes: development, heterogeneity, and relationship with dendritic cells." In: *Annu Rev Immunol.* 27 (2009), pp. 669–92. DOI: 10.1146/annurev.immunol.021908.132557.

[99] Menezes S, Melandri D, Anselmi G, Perchet T, Loschko J, Dubrot J, Patel R, Gautier EL, Hugues S, Longhi MP, Henry JY, Quezada SA, Lauvau G, Lennon-Duménil AM, Gutiérrez-Martínez E, Bessis A, Gomez-Perdiguero E, Jacome-Galarza CE, Garner H, Geissmann F, Golub R, Nussenzweig MC, and Guermonprez P. "The Heterogeneity of Ly6Chi Monocytes Controls Their Differentiation into iNOS+ Macrophages or Monocyte-Derived Dendritic Cells." In: *Immunity* 45 (2016), pp. 1205–1218. DOI: 10.1016/j.immuni.2016.12.001.

[100] Epelman S, Lavine KJ, Beaudin AE, Sojka DK, Carrero JA, Calderon B, Brija T, Gautier EL, Ivanov S, Satpathy AT, Schilling JD, Schwendener R, Sergin I, Razani B, Forsberg EC, Yokoyama WM, Unanue ER, Colonna M, Randolph GJ, and Mann DL. "Embryonic and adult-derived resident cardiac macrophages are maintained through distinct mechanisms at steady state and during inflammation." In: *Immunity* 40 (2014), pp. 91–104. DOI: 10.1016/j.immuni.2013.11.019.

[101] Ginhoux F, Greter M, Leboeuf M, Nandi S, See P, Gokhan S, Mehler MF, Conway SJ, Ng LG, Stanley ER, Samokhvalov IM, and Merad M. "Fate mapping analysis reveals that adult microglia derive from primitive macrophages." In: *Science* 330 (2010), pp. 841–5. DOI: 10.1126/science.1194637.

[102] Yona S, Kim KW, Wolf Y, Mildner A, Varol D, Breker M, Strauss-Ayali D, Viukov S, Guilliams M, Misharin A, Hume DA, Perlman H, Malissen B, Zelzer E, and Jung S. "Fate mapping reveals origins and dynamics of monocytes and tissue macrophages under homeostasis." In: *Immunity* 38 (2013), pp. 79–91. DOI: 10.1016/j.immuni.2012.12.001.

[103] Byrne AJ, Maher TM, and Lloyd CM. "Pulmonary Macrophages: A New Therapeutic Pathway in Fibrosing Lung Disease." In: *Trends Mol Med.* 22 (2016), pp. 303–316. DOI: 10.1016/j.molmed.2016.02.004.

[104] Tan SY and Krasnow MA. "Developmental origin of lung macrophage diversity." In: *Development* 143 (2016), pp. 1318–27. DOI: 10.1242/dev.129122.

[105] Chakarov S et al. "Two distinct interstitial macrophage populations coexist across tissues in specific subtissular niches." In: *Science* 363 (2019). DOI: 10.1126/science.aau0964.

[106] Gibbings SL, Goyal R, Desch AN, Leach SM, Prabagar M, Atif SM, Bratton DL, Janssen W, and Jakubzick CV. "Transcriptome analysis highlights the conserved difference between embryonic and postnatal-derived alveolar macrophages." In: *Blood* 126 (2015), pp. 1357–66. DOI: 10.1182/blood-2015-01-624809.

[107] Misharin AV et al. "Monocyte-derived alveolar macrophages drive lung fibrosis and persist in the lung over the life span." In: *J Exp Med.* 214 (2017), pp. 2387–2404. DOI: 10.1084/jem.20162152.

[108] Keohane E, Otto C, and Walenga J. *Rodak's Hematology: Clinical Principles and Applications.* Elsevier, 2020, pp. 6–10.

[109] Pennock ND, White JT, Cross EW, Cheney EE, Tamburini BA, and Kedl RM. "T cell responses: naive to memory and everything in between." In: *Adv Physiol Educ.* 37 (2013), pp. 273–83. DOI: 10.1152/advan.00066.2013.

[110] Akkaya M, Kwak K, and Pierce SK. "B cell memory: building two walls of protection against pathogens." In: *Nat Rev Immunol.* 20 (2020), pp. 229–238. DOI: 10.1038/s41577-019-0244-2.

[111] Goldenberg NM and Kuebler WM. "Endothelial cell regulation of pulmonary vascular tone, inflammation, and coagulation." In: *Compr Physiol.* 5 (2015), pp. 531–59. DOI: 10.1002/cphy.c140024.

[112] Simionescu N and Simionescu M. "Endothelial Cell Dysfunctions." In: *New York: Plenum Press* (1992), p. 565.

[113] Gillich A, Zhang F, Farmer CG, Travaglini KJ, Tan SY, Gu M, Zhou B, Feinstein JA, Krasnow MA, and Metzger RJ. "Capillary cell-type specialization in the alveolus." In: *Nature* 586 (2020), pp. 785–789. DOI: 10.1038/s41586-020-2822-7.

[114] Kendall RT and Feghali-Bostwick CA. "Fibroblasts in fibrosis: novel roles and mediators." In: *Front Pharmacol.* 5 (2014), p. 123. DOI: 10.3389/fphar.2014.00123.

[115] Rock JR, Barkauskas CE, Cronce MJ, Xue Y, Harris JR, Liang J, Noble PW, and Hogan BL. "Multiple stromal populations contribute to pulmonary fibrosis without evidence for epithelial to mesenchymal transition." In: *Proc Natl Acad Sci USA.* 108 (2011), pp. 475–83. DOI: 10.1073/pnas.1117988108.

[116] Tsukui T, Sun KH, Wetter JB, Wilson-Kanamori JR, Hazelwood LA, Henderson NC, Adams TS, Schupp JC, Poli SD, Rosas IO, Kaminski N, Matthay MA, Wolters PJ, and Sheppard D. "Collagen-producing lung cell atlas identifies multiple subsets with distinct localization and relevance to fibrosis." In: *Nat Commun.* 11 (2020), p. 1920. DOI: 10.1038/s41467-020-15647-5.

[117] Sun KH, Chang Y, and Sheppard D Reed NI. "Alpha-Smooth muscle actin is an inconsistent marker of fibroblasts responsible for force-dependent TGFb activation or collagen production across multiple models of organ fibrosis." In: *Am J Physiol Lung Cell Mol Physiol.* 310 (2016), pp. 824–36. DOI: 10.1152/ajplung.00350.2015.

[118] Abe R, Donnelly SC, Peng T, Bucala R, and Metz CN. "Peripheral blood fibrocytes: differentiation pathway and migration to wound sites." In: *J Immunol.* 166 (2001), pp. 7556–62. DOI: 10.4049/jimmunol.166.12.7556.

[119] Corvol H, Flamein F, Epaud R, Clement A, and Guillot L. "Lung alveolar epithelium and interstitial lung disease." In: *Int J Biochem Cell Biol.* 41 (2009), pp. 1643–51. DOI: 10.1016/j.biocel.2009.02.009.

[120] Degryse AL, Tanjore H, Xu XC, Polosukhin VV, Jones BR, McMahon FB, Gleaves LA, Blackwell TS, and Lawson WE. "Repetitive intratracheal bleomycin models several features of idiopathic pulmonary fibrosis." In: *Am J Physiol Lung Cell Mol Physiol.* 299 (2010), pp. 442–52. DOI: 10.1152/ajplung.00026.2010.

[121] Brunekreef B and Holgate ST. "Air pollution and health." In: *Lancet* 360 (2002), pp. 1233–42. DOI: 10.1016/S0140-6736(02)11274-8.

[122] Olivieri D and Scoditti E. "Impact of environmental factors on lung defences." In: *European Respiratory Review* 14 (2005), pp. 51–56. DOI: 10.1183/09059180.05.00009502.

[123] World Health Organization. *Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019.* "https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death. Accessed December 30, 2020.

[124] World Health Organization. *The top 10 causes of death*. https://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death. Accessed November 4, 2022.

[125] Cersosimo RJ. "Lung cancer: a review." In: *Am J Health Syst Pharm.* 59 (2002), pp. 611–42. DOI: 10.1093/ajhp/59.7.611.

[126] West John B and Luks Andrew M. *Pulmonary Pathophysiology*. Wolters Kluwer, 2017.

[127] Singh D, Agusti A, Anzueto A, Barnes PJ, Bourbeau J, Celli BR, Criner GJ, Frith P, Halpin DMG, Han M, López Varela MV, Martinez F, Montes de Oca M, Papi A, Pavord ID, Roche N, Sin DD, Stockley R, Vestbo J, Wedzicha JA, and Vogelmeier C. "Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease: the GOLD science committee report 2019." In: *Eur Respir J.* 53 (2019). DOI: 10.1183/13993003.00164-2019.

[128] Allinson JP and Wedzicha JA. "Update in Chronic Obstructive Pulmonary Disease 2016." In: *Am J Respir Crit Care Med.* 196 (2017), pp. 414–424. DOI: 10.1164/rccm.201703-0588UP.

[129] Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, AlMazroa MA, and Memish ZA. "Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010." In: *Lancet* 380 (2012), pp. 2095–128. DOI: 10.1016/S0140-6736(12)61728-0.

[130] Barnes PJ, Burney PG, Silverman EK, Celli BR, Vestbo J, Wedzicha JA, and Wouters EF. "Chronic obstructive pulmonary disease." In: *Nat Rev Dis Primers* 1 (2015), p. 15076. DOI: 10.1038/nrdp.2015.76.

[131] SMarc Decramer, Wim Janssens, and Marc Miravitlles. "Chronic obstructive pulmonary disease., Management, and Prevention of Chronic Obstructive Lung Disease: the GOLD science committee report 2019." In: *Lancet* 379 (2012), 1341–51. DOI: 10.1016/S0140-6736(11)60968-9.

[132] Fletcher C and Peto R. "The natural history of chronic airflow obstruction." In: *Br Med J.* 1 (1977), pp. 1645–8. DOI: 10.1136/bmj.1.6077.1645.

[133] Barnes PJ. "Cellular and molecular mechanisms of chronic obstructive pulmonary disease." In: *Clin Chest Med.* 35 (2014), pp. 71–86. DOI: 10.1016/j.ccm.2013.10.004.

[134] Hogg JC and Timens W. "The pathology of chronic obstructive pulmonary disease." In: *Annu Rev Pathol.* 4 (2009), pp. 435–59. DOI: 10.1146/annurev.pathol.4.110807.092145.

[135] Barnes PJ. "Alveolar macrophages in chronic obstructive pulmonary disease (COPD).". In: *Cell Mol Biol.* 50 (2004). DOI: 10.1016/j.ccm.2013.10.004.

[136] Taylor AE, Finney-Hayward TK, Quint JK, Thomas CM, Tudhope SJ, Wedzicha JA, Barnes PJ, and Donnelly LE. "Defective macrophage phagocytosis of bacteria in COPD.". In: *Eur Respir J.* 35 (2010), pp. 1039–47. DOI: 10.1183/09031936.0003670.

[137] Hodge S, Hodge G, Scicchitano R, Reynolds PN, and Holmes M. "Alveolar macrophages from subjects with chronic obstructive pulmonary disease are deficient in their ability to phagocytose apoptotic airway epithelial cells." In: *Immunol Cell Biol.* 81 (2003), pp. 289–96. DOI: 10.1046/j.1440-1711.2003.t01-1-01170.x.

[138] Hogg JC, Chu F, Utokaparch S, Woods R, Elliott WM, Buzatu L, Cherniack RM, Rogers RM, Sciurba FC, Coxson HO, and Paré PD. "The nature of small-airway obstruction in chronic obstructive pulmonary disease". In: *N Engl J Med.* 350 (2004), pp. 2645–53. DOI: 10.1056/NEJMoa032158.

[139] Kuhn C and Senior RM. "The role of elastases in the development of emphysema." In: *Lung* 155 (1978), pp. 185–97.

[140] Edwards R. "The problem of tobacco smoking." In: *BMJ* 328 (2004), pp. 217–9. DOI: 10.1136/bmj.328.7433.217.

[141] Raghu G et al. "An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management." In: *Am J Respir Crit Care Med.* 183 (2011), pp. 788–824. DOI: 10.1164/rccm.2009-040GL.

[142] Barkauskas CE and Noble PW. "Cellular mechanisms of tissue fibrosis - New insights into the cellular mechanisms of pulmonary fibrosis." In: *Am J Physiol Cell Physiol.* 306 (2014), pp. 987–96. DOI: 10.1152/ajpcell.00321.2013.

[143] Raghu G, Chen SY, Yeh WS, Maroni B, Li Q, Lee YC, and Collard HR. "Idiopathic pulmonary fibrosis in US Medicare beneficiaries aged 65 years and older: incidence, prevalence, and survival, 2001-11." In: *Lancet Respir Med.* 2 (2014), pp. 566–72. DOI: 10.1016/S2213-2600(14)70101-8.

[144] King TE, Pardo A, and Selman M. "Idiopathic pulmonary fibrosis." In: *Lancet* 378 (2011), pp. 1949–61. DOI: 10.1016/S0140-6736(11)60052-4.

[145] Martinez FJ, Collard HR, Pardo A, Raghu G, Richeldi L, Selman M, Swigris JJ, Taniguchi H, and Wells AU. "Idiopathic pulmonary fibrosis." In: *Nat Rev Dis Primers* 3 (2017). DOI: 10.1038/nrdp.2017.74.

[146] Wynn TA. "Integrating mechanisms of pulmonary fibrosis." In: *J Exp Med.* 208 (2011), pp. 1339–50. DOI: 10.1084/jem.20110551.

[147] Byrne AJ, Maher TM, and Lloyd CM. "Pulmonary Macrophages: A New Therapeutic Pathway in Fibrosing Lung Disease." In: *Trends Mol Med.* 208 (2016), pp. 303–316. DOI: 10.1016/j.molmed.2016.02.004.

[148] Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, Chak S, Naikawadi RP, Wolters PJ, Abate AR, Butte AJ, and Bhattacharya M. "Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage." In: *Nat Immunol.* 20 (2019), pp. 163–172. DOI: 10.1038/s41590-018-0276-y.

[149] Desai O, Winkler J, Minasyan M, and Herzog EL. "The Role of Immune and Inflammatory Cells in Idiopathic Pulmonary Fibrosis." In: *Front Med. (Lausanne)* 5 (2018), p. 43. DOI: 10.3389/fmed.2018.00043.

[150] Helene M, Lake-Bullock V, Zhu J, Hao H, Cohen DA, and Kaplan AM. "T cell independence of bleomycin-induced pulmonary fibrosis." In: *J Leukoc Biol.* 65 (1999), pp. 187–95. DOI: 10.1002/jlb.65.2.187.

[151] Selman M and Pardo A. "Revealing the pathogenic and aging-related mechanisms of the enigmatic idiopathic pulmonary fibrosis." In: *Am J Respir Crit Care Med.* 189 (2014), pp. 1161–72. DOI: 10.1164/rccm.201312-2221PP.

[152] Bernard GR, Artigas A, Brigham KL, Carlet J, Falke K, Hudson L, Lamy M, Legall JR, Morris A, and Spragg R. "The American-European Consensus Conference on ARDS. Definitions, mechanisms, relevant outcomes, and clinical trial coordination." In: *Am J Respir Crit Care Med.* 149 (1994), pp. 818–24. DOI: 10.1164/ajrccm.149.3.7509706.

[153] Ranieri VM, Rubenfeld GD, Thompson BT, Ferguson ND, Caldwell E, Fan E, Camporota L, and Slutsky AS. "Acute respiratory distress syndrome: the Berlin Definition." In: *JAMA* 307 (2012), pp. 2526–33. DOI: 10.1001/jama.2012.5669.

[154] Matthay MA, Zemans RL, Zimmerman GA, Arabi YM, Beitler JR, Mercat A, Herridge M, Randolph AG, and Calfee CS. "Acute respiratory distress syndrome." In: *Nat Rev Dis Primers* 5 (2019), p. 18. DOI: 10.1038/s41572-019-0069-0.

[155] Matthay MA, Ware LB, and Zimmerman GA. "The acute respiratory distress syndrome." In: *J Clin Invest.* 8 (2012), pp. 2731–40. DOI: 10.1172/JCI60331.

[156] Ting C, Aspal M, Vaishampayan N, Huang SK, Wang F, Farver C, and Zemans RL. "Ineffectual AEC1 differentiation from KRT8hi transitional state without fibrosis is associated with fatal COVID-19 ARDS.". In: *bioRxiv* (2021). DOI: 10.1101/2021.01.12.426404.

[157] Hu B, Guo H, Zhou P, and Shi ZL. "Characteristics of SARS-CoV-2 and COVID-19." In: *Nat Rev Microbiol.* 19 (2021), pp. 141–154. DOI: 10.1038/s41579-020-00459-7.

[158] Eurosurveillance editorial team. "World Health Organization declares novel coronavirus (2019-nCoV) sixth public health emergency of international concern." In: *Euro Surveill.* 25 (2020). DOI: 10.2807/1560-7917.ES.2020.25.5.200131e.

[159] World Health Organization. *WHO Coronavirus Disease (COVID-19) Dashboard*. https://covid19.who.int/info. Accessed December 30, 2020.

[160] Guan WJ et al. "China Medical Treatment Expert Group for Covid-19. Clinical Characteristics of Coronavirus Disease 2019 in China." In: *N Engl J Med.* 382 (2020). DOI: 10.1056/NEJMoa2002032.

[161] Li W, Moore MJ, Vasilieva N, Sui J, Wong SK, Berne MA, Somasundaran M, Sullivan JL, Luzuriaga K, Greenough TC, Choe H, and Farzan M. "Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus." In: *Nature* 426 (2003), pp. 450–4. DOI: 10.1038/nature02145.

[162] Glowacka I, Bertram S, Müller MA, Allen P, Soilleux E, Pfefferle S, Steffen I, Tsegaye TS, He Y, Gnirss K, Niemeyer D, Schneider H, Drosten C, and Pöhlmann S. "Evidence that TMPRSS2 activates the severe acute respiratory syndrome coronavirus spike protein for membrane fusion and reduces viral control by the humoral immune response." In: *J Virol.* 85 (2011), pp. 4122–34. DOI: 10.1128/JVI.02232-10.

[163] Shang J, Wan Y, Luo C, Ye G, Geng Q, Auerbach A, and Li F. "Cell entry mechanisms of SARS-CoV-2." In: *Proc Natl Acad Sci USA* 117 (2020), pp. 11727–11734. DOI: 10.1073/pnas.2003138117.

[164] Sungnak W, Huang N, Bécavin C, Berg M, Queen R, Litvinukova M, Talavera-López C, Maatz H, Reichart D, Sampaziotis F, Worlock KB, Yoshida M, Barnes JL, and HCA Lung Biological Network. "SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes." In: *Nat Med.* 26 (2020), pp. 681–687. DOI: 10.1038/s41591-020-0868-6.

[165] Ziegler CGK et al. "SARS-CoV-2 Receptor ACE2 Is an Interferon-Stimulated Gene in Human Airway Epithelial Cells and Is Detected in Specific Cell Subsets across Tissues." In: *Cell* 28 (2020), pp. 1016–1035. DOI: 10.1016/j.cell.2020.04.035.

[166] Cevik M, Kuppalli K, Kindrachuk J, and Peiris M. "Virology, transmission, and pathogenesis of SARS-CoV-2." In: *BMJ.* 371 (2020). DOI: 10.1136/bmj.m3862.

[167] Muus C et al. "Single-cell meta-analysis of SARS-CoV-2 entry genes across tissues and demographics." In: *Nat Med.* (2021). DOI: 10.1038/s41591-020-01227-z.

[168] Liao M, Liu Y, Yuan J, Wen Y, Xu G, Zhao J, Cheng L, Li J, Wang X, Wang F, Liu L, Amit I, Zhang S, and Zhang Z. "Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19." In: *Nat Med.* 26 (2020), pp. 842–844. DOI: 10.1038/s41591-020-0901-9.

[169] Olmer R, Dahlmann J, Merkert S, Baus S, Göhring G, and Martin U. "Generation of a NKX2.1 knock-in reporter cell line from human induced pluripotent stem cells." In: *Stem Cell Res.* 39 (2019), p. 101492. DOI: 10.1016/j.scr.2019.101492.

[170] Ori O, Ansari M, Angelidis I, Theis FJ, Schiller HB, and Drukker M. "Single cell trajectory analysis of human pluripotent stem cells differentiating towards lung and hepatocyte progenitors." In: *bioRxiv Preprint* (2021). DOI: 10.1101/2021.02.23.432413.

[171] Strunz M, Simon LM, Ansari M, Kathiriya JJ, Angelidis I, Mayr CH, Tsidiridis G, Lange M, Mattner LF, Yee M, Ogar P, Sengupta A, Kukhtevich I, Schneider R, Zhao Z, Voss C, Stoeger T, Neumann JHL, Hilgendorff A, Behr J, O'Reilly M, Lehmann M, Burgstaller G, Königshoff M, Chapman HA, Theis FJ, and Schiller HB. "Alveolar regeneration through a Krt8+ transitional stem cell state that persists in human lung fibrosis." In: *Nat Commun.* 11 (2020). DOI: 10.1038/s41467-020-17358-3.

[172] Conlon TM et al. "Inhibition of LTβR signalling activates WNT-induced regeneration in lung." In: *Nature* 588 (2020), pp. 151–156. DOI: 10.1038/s41586-020-2882-8.

[173] Mayr CH, Simon LM, Leuschner G, Ansari M, Schniering J, Geyer PE, Angelidis I, Strunz M, Singh P, Kneidinger N, Reichenberger F, Silbernagel E, Böhm S, Adler H, Lindner M, Maurer B, Hilgendorff A, Prasse A, Behr J, Mann M, Eickelberg O, Theis FJ, and Schiller HB. "Integrative analysis of cell state changes in lung fibrosis with peripheral protein biomarkers." In: *EMBO Mol Med.* (2021). DOI: 10.15252/emmm.202012871.

[174] Fischer DS, Ansari M, Wagner KI, Jarosch S, Huang Y, Mayr CH, Strunz M, Lang NJ, D'Ippolito E, Hammel M, Mateyka L, Weber S, Wolff LS, Witter K, Fernandez IE, Leuschner G, Milger K, Frankenberger M, Nowak L, Heinig-Menhard K, Koch I, Stoleriu MG, Hilgendorff A, Behr J, Pichlmair A, Schubert B, Theis FJ, Busch DH, Schiller HB, and Schober K. "Single-cell RNA sequencing reveals ex vivo signatures of SARS-CoV-2-reactive T cells through 'reverse phenotyping'." In: *Nat Commun.* 12 (2021). DOI: 10.1038/s41467-021-/24730-4.

[175] Angelidis I, Simon LM, Fernandez IE, Strunz M, Greiffo FR Mayr CH, Tsitsiridis G, Ansari M, Graf E, Strom TM, Nagendran M, Desai T, Eickelberg O, Mann M, Theis FJ, and Schiller HB. "An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics." In: *Nat Commun.* 10 (2019), p. 963. DOI: 10.1038/s41467-019-08831-9.

[176] Hughes TK, Wadsworth MH, Gierahn TM, Do T, Weiss D, Andrade PR, Ma F, de Andrade Silva BJ, Shao S, Tsoi LC, Ordovas-Montanes J, Gudjonsson JE, Modlin RL, Love CJ, and Shalek AK. "Highly Efficient, Massively-Parallel Single-Cell RNA-Seq Reveals Cellular States and Molecular Features of Human Skin Pathology." In: *bioRxiv* 48 (2019), pp. 531–9. DOI: 10.1101/689273.

[177] Timmerman L. *DNA sequencing market will exceed 20 billion dollar, says Illumina CEO Jay Flatley. Forbes*. http://www.forbes.com/sites/luketimmerman/2015/04/29/qa-with-jay-flatley-ceo-of-illumina-the-genomics-company-pursuing-a-20b-market/#4dbd19943bf5. Accessed November 4, 2022.

[178] Goodwin S, McPherson, J, and W McCombie. "Coming of age: ten years of next-generation sequencing technologies." In: *Nat Rev Genet.* 17 (2016), 333–351. DOI: https://doi.org/10.1038/nrg.2016.49.

[179] Illumina. *An introduction to Next-Generation Sequencing Technology*. https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf. Accessed November 4, 2022.

[180] Broad Institute. *Drop-seq - Java tools for analyzing Drop-seq data*. https://github.com/broadinstitute/Drop-seq. Accessed November 4, 2022.

[181] 10X Genomics. *Building Cell Ranger 3.0.2*. https://github.com/10XGenomics/cellranger. Accessed November 4, 2022.

[182] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, and Gingeras TR. "STAR: ultrafast universal RNA-seq aligner." In: *Bioinformatics.* 29 (2013), pp. 15–21. DOI: 10.1093/bioinformatics/bts635.

[183] Zappia L, Phipson B, and Oshlack A. "Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database." In: *PLoS Comput Biol.* 14 (2018). DOI: 0.1371/journal.pcbi.1006245.

[184] Butler A, Hoffman P, Smibert P, Papalexi E, and Satija R. "Integrating single-cell transcriptomic data across different conditions, technologies, and species." In: *Nat Biotechnol.* 36 (2018), pp. 411–420. DOI: 10.1038/nbt.4096.

[185] Wolf FA, Angerer P, and Theis FJ. "SCANPY: large-scale single-cell gene expression data analysis." In: *Genome Biol.* 6 (2018), p. 15. DOI: 10.1186/s13059-017-1382-0.

[186] Kiselev V, Andrews T, Westoby J, Davis McCarthy D, Büttner M, Jimmy Lee J, Polanski K, Müller SY, Madissoon E, Ballereau S, Do Nascimento Lopes Primo M, Nunez RM, and Hemberg M. *Analysis of single cell RNA-seq data.* https://scrnaseq-course.cog.sanger.ac.uk/website/introduction-to-single-cell-rna-seq.html. Accessed November 4, 2022.

[187] Satija Lab. *Seurat - Guided Clustering Tutorial.* https://satijalab.org/seurat/v3.2/pbmc3k_tutorial.html. Accessed November 4, 2022.

[188] Luecken MD and Theis FJ. "Current best practices in single-cell RNA-seq analysis: a tutorial." In: *Mol Syst Biol.* 15 (2019). DOI: 10.15252/msb.20188746.

[189] Griffiths JA, Scialdone A, and Marioni JC. "Using single-cell genomics to understand developmental processes and cell fate decisions." In: *Mol Syst Biol.* 14 (2018). DOI: 10.15252/msb.20178046.

[190] Wolock SL, Lopez R, and Klein AM. "Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data." In: *Cell Syst.* 24 (2019). DOI: 10.1016/j.cels.2018.11.005.

[191] Lun AT, Bach K, and Marioni JC. "Pooling across cells to normalize single-cell RNA sequencing data with many zero counts." In: *Genome Biol.* 17 (2016). DOI: 10.1186/s13059-016-0947-7.

[192] Vallejos CA, Risso D, Scialdone A, Dudoit S, and Marioni JC. "Normalizing single-cell RNA sequencing data: challenges and opportunities." In: *Nat Methods* 14 (2017), pp. 565–71. DOI: 10.1038/nmeth.4292.

[193] Anders S and Huber W. "Differential expression analysis for sequence count data." In: *Genome Biol.* 11 (2010). DOI: 10.1186/gb-2010-11-10-r106.

[194] Townes FW, Hicks SC, Aryee MJ, and Irizarry RA. "Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model." In: *Genome Biol.* 20 (2019), p. 295. DOI: 10.1186/s13059-019-1861-6.

[195] Satija R, Farrell JA, Gennert D, Schier AF, and Regev A. "Spatial reconstruction of single-cell gene expression data." In: *Nat Biotechnol.* 33 (2015), pp. 495–502. DOI: 10.1038/nbt.3192.

[196] Traag VA. "Faster unfolding of communities: speeding up the Louvain algorithm." In: *Phys Rev E Stat Nonlin Soft Matter Phys.* 92 (2015). DOI: 10.1103/PhysRevE.92.032801.

[197] De La Porte J, Herbst BM, Hereman W, and Van Der Walt SJ. "An Introduction to Diffusion Maps". In: *Department of Mathematical Sciences, University of Stellenbosch* (2008). DOI: 10.1.1.309.674.

[198] Trapnell C. "Defining cell types and states with single-cell genomics." In: *Genome Res.* 25 (2015), pp. 1491–8. DOI: 10.1101/gr.190595.115.

[199] Heimberg G, Bhatnagar R, El-Samad H, and Thomson M. "Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing." In: *Cell Syst.* 2 (2016), pp. 239–250. DOI: doi:10.1016/j.cels.2016.04.001.

[200] Pearson K. "On Lines and Planes of Closest Fit to Systems of Points in Space". In: *Philosophical Magazine* 11 (1901), pp. 559–572. DOI: 10.1080/14786440109462720.

[201] Hotelling H. "Analysis of a complex of statistical variables into principal components". In: *Journal of Educational Psychology* 24 (1933), pp. 417–441.

[202] Jolliffe IT. *Principal Component Analysis*. Springer, 2002.

[203] Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, Ginhoux F, and Newell EW. "Dimensionality reduction for visualizing single-cell data using UMAP.". In: *Nat Biotechnol.* (2018). DOI: 10.1038/nbt.4314.

[204] Diaz-Papkovich A, Anderson-Trocmé L, and Gravel S. "A review of UMAP in population genetics." In: *J Hum Genet.* 66 (2021), pp. 85–91. DOI: 10.1038/s10038-020-00851-4.

[205] van der Maaten L. "Visualizing Data using t-SNE". in: *Journal of Machine Learning Research* 9 (Nov. 2008), pp. 2579–2605. DOI: https://doi.org/10.1101/174029.

[206] *How to Use t-SNE Effectively*. URL: https://distill.pub/2016/misread-tsne/.

[207] Kobak D and Berens P. "The art of using t-SNE for single-cell transcriptomics." In: *Nat Commun.* 10 (2019), pp. 494–498. DOI: 10.1038/s41467-019-13056-x.

[208] McInnes L, Healy J, and Melville J. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." In: *ArXiv e-prints* (2018). DOI: arXiv:1802.03426.

[209] Belkin M and Niyogi P. "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering." In: *Advances in Neural Information Processing Systems* 14 (2001), pp. 586–691.

[210] Coifman RR, Lafon S, Lee AB, Maggioni M, Nadler B, Warner F, and Zucker SW. "Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps." In: *Proc Natl Acad Sci USA* 24 (2005), 7426–31. DOI: 10.1073/pnas.0500334102.

[211] Haghverdi L, Büttner M, Wolf FA, Buettner F, and Theis FJ. "Diffusion pseudotime robustly reconstructs lineage branching." In: *Nat Methods.* 13 (2016), pp. 845–8. DOI: 10.1038/nmeth.3971.

[212] Coifman RR and Lafon S. "Diffusion Maps." In: *Applied and Computational Harmonic Analysis* 21 (June 2006), pp. 5–30.

[213] Lancichinetti A and Fortunato S. "Community detection algorithms: a comparative analysis." In: *Phys Rev E Stat Nonlin Soft Matter Phys.* 80 (2009). DOI: 10.1103/PhysRevE.80.056117.

[214] Wolf FA, Hamey FK, Plass M, Solana J, Dahlin JS, Göttgens B, Rajewsky N, Simon L, and Theis FJ. "PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells." In: *Genome Biol.* 20 (2019), p. 59. DOI: 10.1186/s13059-019-1663-x.

[215] Saelens W, Cannoodt R, Todorov H, and Saeys Y. "A comparison of single-cell trajectory inference methods." In: *Nat Biotechnol.* 37 (2019), pp. 547–554. DOI: 10.1038/s41587-019-0071-9.

[216] Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, Purdom E, and Dudoit S. "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics." In: *BMC Genomics.* 19 (2018), p. 477. DOI: 10.1186/s12864-018-4772-0.

[217] Robrecht Cannoodt R, Saelens W, Sichien D, Tavernier S, Janssens S, Guilliams M Lambrecht B, De Preter K, and Yvan Saeys Y. "SCORPIUS improves trajectory inference and identifies novel modules in dendritic cell development." In: *bioRxiv* (2016). DOI: doi:https://doi.org/10.1101/079509.

[218] Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, Strobl DC, Zappia L, Dugas M, Colomé-Tatché M, and Theis FJ. "Benchmarking atlas-level data integration in single-cell genomics." In: *Nat Methods* 19 (2022), pp. 41–50. DOI: 10.1038/s41592-021-01336-8.

[219] Polanski K, Young MD, Miao Z, Meyer KB, Teichmann SA, and Park JE. "BBKNN: fast batch alignment of single cell transcriptomes." In: *Bioinformatics* 36 (2020), pp. 964–965. DOI: 10.1093/bioinformatics/btz625.

[220] Hie B, Bryson B, and Berger B. "Efficient integration of heterogeneous single-cell transcriptomes using Scanorama." In: *Nat Biotechnol.* 37 (2019), pp. 685–691. DOI: 10.1038/s41587-019-0113-3.

[221] Lopez R, Regier J, Cole MB, Jordan MI, and Yosef N. "Deep generative modeling for single-cell transcriptomics." In: *Nat Methods* 15 (2018), pp. 1053–1058. DOI: 10.1038/s41592-018-0229-2.

[222] Young MD and Behjati S. "SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data." In: *Gigascience* 9 (2020). DOI: 10.1093/gigascience/giaa151.

[223] Wang T, Li B, Nelson CE, and Nabavi S. "Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data." In: *BMC Bioinformatics* 20 (2019), p. 40. DOI: 10.1186/s12859-019-2599-.

[224] Kharchenko PV, Silberstein L, and Scadden DT. "Bayesian approach to single-cell differential expression analysis." In: *Nat Methods* 11 (2014), pp. 740–2. DOI: 10.1038/nmeth.2967.

[225] Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, Prlic M, Linsley PS, and Gottardo R. "MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data." In: *Genome Biol.* 16 (2015), p. 278. DOI: 10.1186/s13059-015-0844-5.

[226] Qiu X, Hill A, Packer J, Lin D, Ma YA, and Trapnell C. "Single-cell mRNA quantification and differential analysis with Census." In: *Nat Methods* 14 (2017), pp. 309–315. DOI: 10.1038/nmeth.4150.

[227] Fischer DS and Hölzlwimmer F. *diffxpy: Fast and scalable differential expression analysis on single-cell RNA-seq data.* https://github.com/theislab/diffxpy. Accessed November 4, 2022.

[228] Bar-Joseph Z, Gerber G, Simon I, Gifford DK, and Jaakkola TS. "Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes." In: *Proc Natl Acad Sci USA* 100 (2003), pp. 10146–51. DOI: 10.1073/pnas.1732547100.

[229] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, and Sherlock G. "Gene ontology: tool for the unification of biology." In: *Nat Genet.* 25 (2000), pp. 25–9. DOI: 10.1038/75556.

[230] Klopfenstein DV, Zhang L, Pedersen BS, Ramírez F, Warwick Vesztrocy A, Naldi A, Mungall CJ, Yunes JM, Botvinnik O, Weigel M, Dampier W, Dessimoz C, Flick P, and Tang H. "GOATOOLS: A Python library for Gene Ontology analyses." In: *Sci Rep.* 8 (2018), p. 10872. DOI: 10.1038/s41598-018-28948-z.

[231] Efremova M, Vento-Tormo M, Teichmann SA, and Vento-Tormo R. "CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes." In: *Nat Protoc.* 15 (2020), pp. 1484–1506. DOI: 10.1038/s41596-020-0292-x.

[232] Browaeys R, Saelens W, and Saeys Y. "NicheNet: modeling intercellular communication by linking ligands to target genes." In: *Nat Methods.* 17 (2020), pp. 159–162. DOI: 10.1038/s41592-019-0667-5.

[233] Green MD, Chen A, Nostro MC, d'Souza SL, Schaniel C, Lemischka IR, Gouon-Evans V, Keller G, and Snoeck HW. "Generation of anterior foregut endoderm from human embryonic and induced pluripotent stem cells." In: *Nat Biotechnol.* 29 (2011), pp. 267–72. DOI: 10.1038/nbt.1788.

[234] Takahashi K and Yamanaka S. "Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors." In: *Cell* 126 (2006), pp. 663–76. DOI: 10.1016/j.cell.2006.07.024.

[235] Hawkins F, Kramer P, Jacob A, Driver I, Thomas DC, McCauley KB, Skvir N, Crane AM, Kurmann AA, Hollenberg AN, Nguyen S, Wong BG, Khalil AS, Huang SX, Guttentag S, Rock JR, Shannon JM, Davis BR, and Kotton DN. "Prospective isolation of NKX2-1-expressing human lung progenitors derived from pluripotent stem cells." In: *J Clin Invest.* 127 (2017), pp. 2277–2294. DOI: 10.1172/JCI89950.

[236] Li Y, Eggermont K, Vanslembrouck V, and Verfaillie CM. "NKX2-1 activation by SMAD2 signaling after definitive endoderm differentiation in human embryonic stem cell." In: *Stem Cells Dev* 22 (2013), pp. 1433–42. DOI: 10.1089/scd.2012.0620.

[237] Litingtung Y, Lei L, Westphal H, and Chiang C. "Sonic hedgehog is essential to foregut development." In: *Nat Genet.* 20 (1998), pp. 58–61. DOI: 10.1038/1717.

[238] Bellusci S, Grindley J, Emoto H, Itoh N, and Hogan BL. "Fibroblast growth factor 10 (FGF10) and branching morphogenesis in the embryonic mouse lung." In: *Development* 124 (1998), pp. 4867–78.

[239] Gonzales LW, Guttentag SH, Wade KC, Postle AD, and Ballard PL. "Differentiation of human pulmonary type II cells in vitro by glucocorticoid plus cAMP.". In: *Am J Physiol Lung Cell Mol Physiol.* 283 (2002), pp. 940–51. DOI: 10.1152/ajplung.00127.2002.

[240] Shu W, Lu MM, Zhang Y, Tucker PW, Zhou D, and Morrisey EE. "Foxp2 and Foxp1 cooperatively regulate lung and esophagus development." In: *Development* 134 (2007), pp. 1991–2000. DOI: 10.1242/dev.02846.

[241] Carpentier A, Tesfaye A, Chu V, Nimgaonkar I, Zhang F, Lee SB, Thorgeirsson SS, Feinstone SM, and Liang TJ. "Engrafted human stem cell-derived hepatocytes establish an infectious HCV murine model." In: *J Clin Invest.* 124 (2014), pp. 4953–64. DOI: 10.1172/JCI75456.

[242] Kim E, Jiang M, Huang H, Zhang Y, Tjota N, Gao X, Robert J, Gilmore N, Gan L, and Que J. "Isl1 Regulation of Nkx2.1 in the Early Foregut Epithelium Is Required for Trachea-Esophageal Separation and Lung Lobation." In: *Dev Cell* 51 (2019), pp. 675–683. DOI: 10.1016/j.devcel.2019.11.002.

[243] van Tuyl M, Liu J, Groenman F, Ridsdale R, Han RN, Venkatesh V, Tibboel D, and Post M. "Iroquois genes influence proximo-distal morphogenesis during rat lung development." In: *Am J Physiol Lung Cell Mol Physiol.* 290 (2006), pp. 777–789. DOI: 10.1152/ajplung.00293.2005.

[244] Lin CR, Kioussi C, O'Connell S, Briata P, Szeto D, Liu F, Izpisua-Belmonte JC, and Rosenfeld MG. "Pitx2 regulates lung asymmetry, cardiac positioning and pituitary and tooth morphogenesis." In: *J Clin Invest.* 401 (1999), pp. 279–82. DOI: 10.1038/45803.

[245] Villavicencio EH, Walterhouse DO, and Iannaccone PM. "The sonic hedgehog-patched-gli pathway in human development and disease." In: *Am J Hum Genet.* 67 (2000), pp. 1047–54. DOI: 10.1016/S0002-9297(07)62934-6.

[246] Turner DA, Hayward PC, Baillie-Johnson P, Rue P, Broome R, Faunes F, and Martinez Arias A. "Wnt/β-catenin and FGF signalling direct the specification and maintenance of a neuromesodermal axial progenitor in ensembles of mouse embryonic stem cells." In: *Development* 141 (2014), pp. 4243–53. DOI: 10.1242/dev.112979.

[247] Falix FA, Aronson DC, Lamers WH, and Gaemers IC. "Possible roles of DLK1 in the Notch pathway during development and disease." In: *Biochim Biophys Acta.* 1822 (2012), pp. 988–95. DOI: 10.1016/j.bbadis.2012.02.003.

[248] Moeller A, Ask K, Warburton D, Gauldie J, and Kolb M. "The bleomycin animal model: a useful tool to investigate treatment options for idiopathic pulmonary fibrosis?" In: *Int J Biochem Cell Biol.* 40 (2008), pp. 362–82. DOI: 10.1016/j.biocel.2007.08.011.

[249] UniProt Consortium. "UniProt: the universal protein knowledgebase in 2021." In: *Nucleic Acids Res.* 49 (2021), pp. 480–489. DOI: 10.1093/nar/gkaa1100.

[250] El Agha E, Moiseenko A, Kheirollahi V, De Langhe S, Crnkovic S, Kwapiszewska G, Szibor M, Kosanovic D, Schwind F, Schermuly RT, Henneke I, MacKenzie B, Quantius J, Herold S, Ntokou A, Ahlbrecht K, Braun T, Morty RE, Günther A, Seeger W, and Bellusci S. "Two-Way Conversion between Lipogenic and Myogenic Fibroblastic Phenotypes Marks the Progression and Resolution of Lung Fibrosis." In: *Cell Stem Cell* 20 (2017), pp. 261–273. DOI: 10.1016/j.stem.2016.10.004.

[251] Marvin KW, George MD, Fujimoto W, Saunders NA, Bernacki SH, and Jetten AM. "Cornifin, a cross-linked envelope precursor in keratinocytes that is down-regulated by retinoids." In: *Proc Natl Acad Sci USA* 89 (1992), pp. 11026–30. DOI: 10.1073/pnas.89.22.11026.

[252] Dumic J, Dabelic S, and Flögel M. "Galectin-3: an open-ended story." In: *Biochim Biophys Acta.* 1760 (2006), pp. 616–35. DOI: 10.1016/j.bbagen.2005.12.020.

[253] LaBaer J, Garrett MD, Stevenson LF, Slingerland JM, Sandhu C, Chou HS, Fattaey A, and Harlow E. "New functional activities for the p21 family of CDK inhibitors." In: *Genes Dev.* 11 (1997), pp. 847–62. DOI: 10.1101/gad.11.7.847.

[254] Lund SA, Giachelli CM, and Scatena M. "The role of osteopontin in inflammatory processes." In: *J Cell Commun Signal.* 3 (2009), pp. 311–22. DOI: 10.1007/s12079-009-0068-0.

[255] Tsai CC, Wu SB, Kau HC, and Wei YH. "Essential role of connective tissue growth factor (CTGF) in transforming growth factor- TGF-β1-induced myofibroblast transdifferentiation from Graves' orbital fibroblasts." In: *Sci Rep.* 8 (2018), p. 7276. DOI: 10.1038/s41598-018-25370-3.

[256] Riemondy KA, Jansing NL, Jiang P, Redente EF, Gillen AE, Fu R, Miller AJ, Spence JR, Gerber AN, Hesselberth JR, and Zemans RL. "Single cell RNA sequencing identifies TGF-βas a key regenerative cue following LPS-induced lung injury." In: *JCI Insight* 5 (2019). DOI: 10.1172/jci.insight.123637.

[257] Cheng DS, Han W, Chen SM, Sherrill TP, Chont M, Park GY, Sheller JR, Polosukhin VV, Christman JW, Yull FE, and Blackwell TS. "Airway epithelium controls lung inflammation and injury through the NF-kappa B pathway." In: *J Immunol.* 178 (2007), pp. 6504–13. DOI: 10.4049/jimmunol.178.10.6504.

[258] McConnell AM, Yao C, Yeckes AR, Wang Y, Selvaggio AS, Tang J, Kirsch DG, and Stripp BR. "p53 Regulates Progenitor Cell Quiescence and Differentiation in the Airway." In: *Cell Rep.* 17 (2016), pp. 2173–2182. DOI: 10.1016/j.celrep.2016.11.007.

[259] Pan X, Zhao J, Zhang WN, Li HY, Mu R, Zhou T, Zhang HY, Gong WL, Yu M, Man JH, Zhang PJ, Li AL, and Zhang XM. "Induction of SOX4 by DNA damage is critical for p53 stabilization and function." In: *Proc Natl Acad Sci USA* 106 (2009), pp. 3788–93. DOI: 10.1073/pnas.0810147106.

[260] Gulati S and Thannickal VJ. "The Aging Lung and Idiopathic Pulmonary Fibrosis." In: *Am J Med Sci.* 357 (2019), pp. 384–389. DOI: 10.1016/j.amjms.2019.02.008.

[261] Adams TS, Schupp JC, Poli S, Ayaub EA, Neumark N, Ahangari F, Chu SG, Raby BA, Deluliis G, Januszyk M, Duan Q, Arnett HA, Siddiqui A, Washko GR, Homer R, Yan X, Rosas IO, and Kaminski N. "Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis." In: *Sci Adv.* 6 (2020), p. 1972. DOI: 10.1126/sciadv.aba1983.

[262] Reyfman PA et al. "Single-Cell Transcriptomic Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis." In: *Am J Respir Crit Care Med.* 199 (2019), pp. 1517–1536. DOI: 10.1164/rccm.201712-2410OC.

[263] AMcDonough JE, Ahangari F, Li Q, Jain S, Verleden SE, Herazo-Maya J, Vukmirovic M, Deluliis G, Tzouvelekis A, Tanabe N, Chu F, Yan X, Verschakelen J, Homer RJ, Manatakis DV, Zhang J, Ding J, Maes K, De Sadeleer L, Vos R, Neyrinck A, Benos PV, Bar-Joseph Z, Tantin D, Hogg JC, Vanaudenaerde BM, Wuyts WA, and Kaminski N. "Transcriptional regulatory model of fibrosis progression in the human lung." In: *JCI Insight* 4 (2019). DOI: 10.1172/jci.insight.131597.

[264] Ebina M, Shimizukawa M, Shibata N, Kimura Y, Suzuki T, Endo M, Sasano H, Kondo T, and Nukiwa T. "Heterogeneous increase in CD34-positive alveolar capillaries in idiopathic pulmonary fibrosis." In: *Am J Respir Crit Care Med.* 169 (2004), pp. 1203–8. DOI: 10.1164/rccm.200308-1111OC.

[265] Zhao MJ, Chen SY, Qu XY, Abdul-Fattah B, Lai T, Xie M, Wu SD andZhou YW, and Huang CZ. "Increased Cthrc1 Activates Normal Fibroblasts and Suppresses Keloid Fibroblasts by Inhibiting TGF-β/Smad Signal Pathway and Modulating YAP Subcellular Location." In: *Curr Med Sci.* 316 (2018), pp. 894–902. DOI: 10.1007/s11596-018-1959-1.

[266] Yu G, Tzouvelekis A, Wang R, Herazo-Maya JD, Ibarra GH, Srivastava A, de Castro JPW, DeIuliis G, Ahangari F, Woolard T, Aurelien N, Arrojo E Drigo R, Gan Y, Graham M, Liu X, Homer RJ, Scanlan TS, Mannam P, Lee PJ, Herzog EL, Bianco AC, and Kaminski N. "Thyroid hormone inhibits lung fibrosis in mice by improving epithelial mitochondrial function." In: *Nat Med.* 24 (2018), pp. 39–49. DOI: 10.1038/nm.4447.

[267] Warrier S, Balu SK, Kumar AP, Millward M, and Dharmarajan A. "Wnt antagonist, secreted frizzled-related protein 4 (sFRP4), increases chemotherapeutic response of glioma stem-like cells." In: *Oncol Res.* 21 (2013), pp. 93–102. DOI: 10.3727/096504013X13786659070154.

[268] Mereu E, Iacono G, Guillaumet-Adkins A, Moutinho C, Lunazzi G, Santos CP, Miguel-Escalada I, Ferrer J, Real FX, Gut I, and Heyn H. "matchSCore: Matching Single-Cell Phenotypes Across Tools and Experiments." In: *bioRxiv* (2018). DOI: https://doi.org/10.1101/314831.

[269] Baarsma HA and Königshoff M. "'WNT-er is coming': WNT signalling in chronic lung diseases." In: *Thorax.* 72 (2017), pp. 746–759. DOI: 10.1136/thoraxjnl-2016-209753.

[270] Pfaff EM, Becker S, Günther A, and Königshoff M. "Dickkopf proteins influence lung epithelial cell proliferation in idiopathic pulmonary fibrosis." In: *Eur Respir J.* 37 (2011), pp. 79–87. DOI: 10.1183/09031936.00142409.

[271] Diaz-Horta O, Abad C, Sennaroglu L, Foster J 2nd, DeSmidt A, Bademci G, Tokgoz-Yilmaz S, Duman D, Cengiz FB, Grati M, Fitoz S, Liu XZ, Farooq A, Imtiaz F, Currall BB, Morton CC, Nishita M, Minami Y, Lu Z, Walz K, and Tekin M. "ROR1 is essential for proper innervation of auditory hair cells and hearing in humans and mice." In: *Proc Natl Acad Sci USA* 113 (2016), pp. 5993–8. DOI: 10.1073/pnas.1522512113.

[272] Kim S. "Interleukin-32 in inflammatory autoimmune diseases." In: *Immune Netw.* 14 (2014), pp. 123–7. DOI: 10.4110/in.2014.14.3.123.

[273] Zolak JS, Jagirdar R, Surolia R, Karki S, Oliva O, Hock T, Guroji P, Ding Q, Liu RM, Bolisetty S, Agarwal A, Thannickal VJ, and Antony VB. "Pleural mesothelial cell differentiation and invasion in fibrogenic lung injury." In: *Am J Pathol.* 182 (2013), pp. 1239–47. DOI: 10.1016/j.ajpath.2012.12.030.

[274] Mubarak KK, Montes-Worboys A, Regev D, Nasreen N, Mohammed KA, Faruqi I, Hensel E, Baz MA, Akindipe OA, Fernandez-Bussy S, Nathan SD, and Antony VB. "Parenchymal trafficking of pleural mesothelial cells in idiopathic pulmonary fibrosis." In: *Eur Respir J.* 39 (2012), pp. 133–40. DOI: 10.1183/09031936.00141010.

[275] Batra H and Antony VB. "The pleural mesothelium in development and disease." In: *Front Physiol.* 39 (2014), p. 284. DOI: 10.3389/fphys.2014.00284.

[276] Polverino F, Cosio BG, Pons J, Laucho-Contreras M, Tejera P, Iglesias A, Rios A, Jahn A, Sauleda J, Divo M, Pinto-Plata V, Sholl L, Rosas IO, Agustí A, Celli BR, and Owen CA. "B Cell-Activating Factor. An Orchestrator of Lymphoid Follicles in Severe Chronic Obstructive Pulmonary Disease." In: *Am J Respir Crit Care Med.* 192 (2015), pp. 695–705. DOI: 10.1164/rccm.201501-0107OC.

[277] Bracke KR, Verhamme FM, Seys LJ, Bantsimba-Malanda C, Cunoosamy DM, Herbst R, Hammad H, Lambrecht BN, Joos GF, and Brusselle GG. "Role of CXCL13 in cigarette smoke-induced lymphoid follicle formation and chronic obstructive pulmonary disease." In: *Am J Respir Crit Care Med.* 188 (2013), pp. 343–55. DOI: 10.1164/rccm.201211-20550C.

[278] Aloisi F and Pujol-Borrell R. "Lymphoid neogenesis in chronic inflammatory diseases." In: *Nat Rev Immunol.* 6 (2006), pp. 205–17. DOI: 10.1038/nri1786.

[279] Dejardin E, Droin NM, Delhase M, Haas E, Cao Y, Makris C, Li ZW, Karin M, Ware CF, and Green DR. "The lymphotoxin-beta receptor induces different patterns of gene expression via two NF-kappaB pathways." In: *Immunity.* 17 (2002), pp. 525–35. DOI: 10.1016/s1074-7613(02)00423-5.

[280] Kratz A, Campos-Neto A, Hanson MS, and Ruddle NH. "Chronic inflammation caused by lymphotoxin is lymphoid neogenesis." In: *J Exp Med.* 183 (1996), pp. 1461–72. DOI: 10.1084/jem.183.4.1461.

[281] John G, Kohse K, Orasche J, Reda A, Schnelle-Kreis J, Zimmermann R, Schmid O, Eickelberg O, and Yildirim AÖ. "The composition of cigarette smoke determines inflammatory cell recruitment to the lung in COPD mouse models." In: *Clin Sci (Lond).* 126 (2014), pp. 207–21. DOI: 10.1042/CS20130117.

[282] Fetterman JL, Sammy MJ, and Ballinger SW. "Mitochondrial toxicity of tobacco smoke and air pollution." In: *Toxicology* 391 (2017), pp. 18–33. DOI: 10.1016/j.tox.2017.08.002.

[283] Hosgood HD 3rd, Liu CS, Rothman N, Weinstein SJ, Bonner MR, Shen M, Lim U, Virtamo J, Cheng WL, Albanes D, and Lan Q. "Mitochondrial DNA copy number and lung cancer risk in a prospective cohort study." In: *Carcinogenesis.* 31 (2010), pp. 847–9. DOI: 10.1093/carcin/bgq045.

[284] Grieshaber-Bouyer R, Radtke FA, Cunin P, Stifano G, Levescot A, Vijaykumar B, Nelson-Maney N, Blaustein RB, Monach PA, and Nigrovic PA. "The neutrotime transcriptional signature defines a single continuum of neutrophils across biological compartments." In: *Nat Commun.* 12 (2021), p. 2856. DOI: 10.1038/s41467-021-22973-9.

[285] Tilley AE, Walters MS, Shaykhiev R, and Crystal RG. "Cilia dysfunction in lung disease." In: *Annu Rev Physiol.* 77 (2015), pp. 379–406. DOI: 10.1146/annurev-physiol-021014-071931.

[286] Hoenderdos K and Condliffe A. "The neutrophil in chronic obstructive pulmonary disease." In: *Am J Respir Cell Mol Biol.* 48 (2013), pp. 531–9. DOI: 10.1165/rcmb.2012-0492TR.

[287] Schlenner S, Pasciuto E, Lagou V, Burton O, Prezzemolo T, Junius S, Roca CP, Seillet C, Louis C, Dooley J, Luong K, Van Nieuwenhove E, Wicks IP, Belz G, Humblet-Baron S, Wouters C, and Liston A. "NFIL3 mutations alter immune homeostasis and sensitise for arthritis pathology." In: *Ann Rheum Dis.* 78 (2019), pp. 342–349. DOI: 10.1136/annrheumdis-2018-213764.

[288] Pesci A, Majori M, Cuomo A, Borciani N, Bertacco S, Cacciani G, and Gabrielli M. "Neutrophils infiltrating bronchial epithelium in chronic obstructive pulmonary disease." In: *Respir Med.* 92 (1998), pp. 863–70. DOI: 10.1016/s0954-6111(98)90389-4.

[289] Karlsson AC, Humbert M, and Buggert M. "The known unknowns of T cell immunity to COVID-19." In: *Sci Immunol.* 5 (2020). DOI: 10.1126/sciimmunol.abe8063.

[290] Le Bert N, Tan AT, Kunasegaran K, Tham CYL, Hafezi M, Chia A, Chng MHY, Lin M, Tan N, Linster M, Chia WN, Chen MI, Wang LF, Ooi EE, Kalimuddin S, Tambyah PA, Low JG, Tan YJ, and Bertoletti A. "SARS-CoV-2-specific T cell immunity in cases of COVID-19 and SARS, and uninfected controls." In: *Nature* 584 (2020), pp. 457–462. DOI: 10.1038/s41586-020-2550-z.

[291] Fuchs YF, Sharma V, Eugster A, Kraus G, Morgenstern R, Dahl A, Reinhardt S, Petzold A, Lindner A, Löbel D, and Bonifacio E. "Gene Expression-Based Identification of Antigen-Responsive CD8+ T Cells on a Single-Cell Level." In: *Front Immunol.* 10 (2019), p. 2568. DOI: 10.3389/fimmu.2019.02568.

[292] Szabo PA, Levitin HM, Miron M, Snyder ME, Senda T, Yuan J, Cheng YL, Bush EC, Dogra P, Thapa P, Farber DL, and Sims PA. "Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease." In: *Nat Commun.* 10 (2019), p. 4706. DOI: 10.1038/s41467-019-12464-3.

[293] Jorgensen JL, Esser U, Fazekas de St Groth B, Reay PA, and Davis MM. "Mapping T-cell receptor-peptide contacts by variant peptide immunization of single-chain transgenics." In: *Nature* 355 (1992), pp. 224–30. DOI: 10.1038/355224a0.

[294] Yokota S Takiguchi M Kobayashi N Takata H. "Down-regulation of CXCR4 expression on human CD8+ T cells during peripheral differentiation." In: *Eur J Immunol.* 34 (2004), pp. 3370–8. DOI: 10.1002/eji.200425587.

[295] Schober K, Müller TR, and Busch DH. "Orthotopic T-Cell Receptor Replacement - An Enabler for TCR-Based Therapies." In: *Cells* 9 (2020), p. 1367. DOI: 10.3390/cells9061367.

[296] Bacher P, Rosati E, Esser D, Martini GR, Saggau C, Schiminsky E, Dargvainiene J, Schröder I, Wieters I, Khodamoradi Y, Eberhardt F, Vehreschild MJGT, Neb H, Sonntagbauer M, Conrad C, Tran F, Rosenstiel P, Markewitz R, Wandinger KP, Augustin M, Rybniker J, Kochanek M, Leypoldt F, Cornely OA, Koehler P, Franke A, and Scheffold A. "Low-Avidity CD4+ T Cell Responses to SARS-CoV-2 in Unexposed Individuals and Humans with Severe COVID-19." In: *Immunity* 53 (2020), pp. 1258–1271. DOI: 10.1016/j.immuni.2020.11.016.

[297] Wang M, Windgassen D, and Papoutsakis ET. "Comparative analysis of transcriptional profiling of CD3+, CD4+ and CD8+ T cells identifies novel immune response players in T-cell activation." In: *BMC Genomics* 9 (2008), pp. 1258–1271. DOI: 10.1186/1471-2164-9-225.

[298] Parga-Vidal L, Behr FM, Kragten NAM, Nota B, Wesselink TH, Kavazovic I, Covill LE, Schuller MBP, Bryceson YT, Wensveen FM, van Lier R, van Dam TJP, Stark R, and van Gisbergen KPJM. "Hobit identifies tissue-resident memory T cell precursors that are regulated by Eomes." In: *Sci Immunol.* 6 (2021), eabg3533. DOI: 10.1126/sciimmunol.abg3533.

[299] Cheuk S, Schlums H, Gallais Sérézal I, Martini E, Chiang SC, Marquardt N, Gibbs A, Detlofsson E, Introini A, Forkel M, Höög C, Tjernlund A, Michaelsson J, Folkersen L, Mjösberg J, Blomqvist L, Ehrström M, Stahle M, Bryceson YT, and Eidsmo L. "CD49a Expression Defines Tissue-Resident CD8+ T Cells Poised for Cytotoxic Function in Human Skin." In: *Immunity* 46 (2017), pp. 287–300. DOI: 10.1016/j.immuni.2017.01.009.

[300] Liao M, Liu Y, Yuan J, Wen Y, Xu G, Zhao J, Cheng L, Li J, Wang X, Wang F, Liu L, Amit I, Zhang S, and Zhang Z. "Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19." In: *Nat Med.* 26 (2020), pp. 842–844. DOI: 10.1038/s41591-020-0901-9.

[301] Grant RA et al. "Circuits between infected macrophages and T cells in SARS-CoV-2 pneumonia." In: *Nature* 590 (2021), pp. 635–641. DOI: 10.1038/s41586-020-03148-w.

[302] Chua RL et al. "COVID-19 severity correlates with airway epithelium-immune cell interactions identified by single-cell analysis." In: *Nat Biotechnol.* 38 (2020), pp. 970–979. DOI: 10.1038/s41587-020-0602-4.

[303] Bystry RS, Aluvihare V, Welch KA, Kallikourdis M, and Betz AG. "B cells and professional APCs recruit regulatory T cells via CCL4." In: *Nat Immunol.* 2 (2001), pp. 1126–32. DOI: 10.1038/ni735.

[304] Human Cell Atlas. *Idiopathic Pulmonary Fibrosis Cell Atlas.* http://www.ipfcellatlas.com/. Accessed November 4, 2022.

[305] Human Cell Atlas. *COVID-19 Cell Atlas.* https://www.covid19cellatlas.org/. Accessed November 4, 2022.

[306] Stabler CT and Morrisey EE. "Developmental pathways in lung regeneration." In: *Cell Tissue Res.* 367 (2017), pp. 677–685. DOI: 10.1007/s00441-016-2537-0.

[307] Rock JR, Onaitis MW, Rawlins EL, Lu Y, Clark CP, Xue Y, Randell SH, and Hogan BL. "Basal cells as stem cells of the mouse trachea and human airway epithelium." In: *Proc Natl Acad Sci USA* 106 (2009), pp. 12771–5. DOI: 10.1073/pnas.0906850106.

[308] Wang R, Ahmed J, Wang G, Hassan I, Strulovici-Barel Y, Hackett NR, and Crystal RG. "Down-regulation of the canonical Wnt β-catenin pathway in the airway epithelium of healthy smokers and smokers with COPD.". In: *PLoS One* 6 (2011), p. 14793. DOI: 10.1371/journal.pone.0014793.

[309] Guo L, Wang T, Wu Y, Yuan Z, Dong J, Li X, An J, Liao Z, Zhang X, Xu D, and Wen FQ. "WNT/β-catenin signaling regulates cigarette smoke-induced airway inflammation via the PPARδ/p38 pathway." In: *Lab Invest.* 96 (2015), pp. 218–29. DOI: 10.1038/labinvest.2015.101.

[310] Ma B and Hottiger MO. "Crosstalk between Wnt/β-Catenin and NF-κB Signaling Pathway during Inflammation." In: *Front Immunol.* 7 (2016), p. 378. DOI: 10.3389/fimmu.2016.00378.

[311] Munoz-Espin D and Serrano M. "Cellular senescence: from physiology to pathology." In: *Nat Rev Mol Cell Biol.* 15 (2014), pp. 482–96. DOI: 10.1038/nrm3823.

[312] Parimon T, Hohmann MS, and Yao C. "Cellular Senescence: Pathogenic Mechanisms in Lung Fibrosis." In: *Int J Mol Sci.* 22 (2021), p. 6214. DOI: 10.3390/ijms22126214.

[313] Xu Y, Mizuno T, Sridharan A, Du Y, Guo M, Tang J, Wikenheiser-Brokamp KA, Perl AT, Funari VA, Gokey JJ, Stripp BR, and Whitsett JA. "Single-cell RNA sequencing identifies diverse roles of epithelial cells in idiopathic pulmonary fibrosis." In: *JCI Insight.* 1 (2016), e90558. DOI: 10.1172/jci.insight.90558.

[314] Choi J, Park JE, Tsagkogeorga G, Yanagita M, Koo BK, Han N, and Lee JH. "Inflammatory Signals Induce AT2 Cell-Derived Damage-Associated Transient Progenitors that Mediate Alveolar Regeneration." In: *Cell Stem Cell* 27 (2020), pp. 366–382. DOI: 10.1016/j.stem.2020.06.020.

[315] Carsana L, Sonzogni A, Nasr A, Rossi RS, Pellegrinelli A, Zerbi P, Rech R, Colombo R, Antinori S, Corbellino M, Galli M, Catena E, Tosoni A, Gianatti A, and Nebuloni M. "Pulmonary post-mortem findings in a series of COVID-19 cases from northern Italy." In: *Lancet Infect Dis.* 20 (2020), pp. 1135–1140. DOI: 10.1016/S1473-3099(20)30434-5.

[316] Rai DK, Sharma P, and Kumar R. "Post covid 19 pulmonary fibrosis.." In: *Indian J Tuberc.* 68 (2021), pp. 330–333. DOI: 10.1016/j.ijtb.2020.11.003.

[317] Delorey TM et al. "COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets." In: *Nature* 595 (2021), pp. 107–113. DOI: 10.1038/s41586-021-03570-8.

[318] Rogers S, Girolami M, Kolch W, Waters KM, Liu T, Thrall B, and Wiley HS. "Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models." In: *Bioinformatics* 24 (2008), pp. 2894–900. DOI: 10.1093/bioinformatics/btn553.

[319] Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, Muthuswamy L, Krasnitz A, McCombie WR, Hicks J, and Wigler M. "Tumour evolution inferred by single-cell sequencinge." In: *Nature* 472 (2011), pp. 90–4. DOI: 10.1038/nature09807.

[320] Zong C, Lu S, Chapman AR, and Xie XS. "Genome-wide detection of single-nucleotide and copy-number variations of a single human cell." In: *Science* 338 (2012), pp. 1622–6. DOI: 10.1126/science.1229164.

[321] Smith ZD Gu H Clement K Pop R Akopian V Klages S Santos DP Tsankov AM Timmermann B Ziller MJ Kiskinis E Gnirke A Meissner A Charlton J Downing TL. "Global delay in nascent strand DNA methylation." In: *Nat Struct Mol Biol.* 25 (2018), pp. 327–332. DOI: 10.1038/s41594-018-0046-4.

[322] Gu H, Raman AT, Wang X, Gaiti F, Chaligne R, Mohammad AW, Arczewska A, Smith ZD, Landau DA, Aryee MJ, Meissner A, and Gnirke A. "Smart-RRBS for single-cell methylome and transcriptome analysis." In: *Nat Protoc.* 16 (2021), pp. 4004–4030. DOI: 10.1038/s41596-021-00571-9.

[323] Rotem A, Ram O, Shoresh N, Sperling RA, Goren A, Weitz DA, and Bernstein BE. "Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state." In: *Nat Biotechnol.* 33 (2015), pp. 1165–72. DOI: 10.1038/nbt.3383.

[324] Buenrostro JD, Wu B, Chang HY, and Greenleaf WJ. "ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide." In: *Curr Protoc Mol Biol.* 109 (2015), pp. 1–9. DOI: 10.1002/0471142727.mb2129s109.

[325] Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, Steemers FJ, Trapnell C, and Shendure J. "Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing." In: *Science* 348 (2015), pp. 910–4. DOI: 10.1126/science.aab1601.

[326] Clark SJ, Argelaguet R, Kapourani CA, Stubbs TM, Lee HJ, Alda-Catalinas C, Krueger F, Sanguinetti G, Kelsey G, Marioni JC, Stegle O, and Reik W. "scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells." In: *Nat Commun* 9 (2018), p. 781. DOI: 10.1038/s41467-018-03149-4.

[327] Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, and Smibert P. "Simultaneous epitope and transcriptome measurement in single cells." In: *Nat Methods* 14 (2017), pp. 865–868. DOI: 10.1038/nmeth.4380.

[328] Sountoulidis A, Liontos A, Nguyen HP, Firsova AB, Fysikopoulos A, Qian X, Seeger W, Sundström E, Nilsson M, and Samakovlis C. "SCRINSHOT enables spatial mapping of cell states in tissue sections with single-cell resolution." In: *PLoS Biol.* 18 (2020). DOI: 10.1371/journal.pbio.3000675.

[329] 10X Genomics. *Spatially resolved biology.* https://www.10xgenomics.com/spatial-transcriptomics. Accessed November 4, 2022.

[330] Palla G, Spitzer H, Klein M, Fischer DS, Schaar AC, Kuemmerle LB, Rybakov S, Ibarra IL, Holmberg O, Virshup I, Lotfollahi M, Richter S, and Theis FJ. "Squidpy: a scalable framework for spatial single cell analysis." In: *bioRxiv* (2021). DOI: https://doi.org/10.1101/2021.02.19.431994.

Acknowledgements

The completion and success of this journey wouldn't have been possible without the support I have received throughout the years, which I want to acknowledge in these last pages. First and foremost, I am incredibly grateful to have been in this interface position, with a first row view on how experimental data is generated, and access to cutting-edge algorithms to explore it - accompanied by first class feedback from experts in the field.

Fabian Theis, for his continued supervision and advise throughout my PhD and opening the door to the Institute of Computational Biology - already way back during the start of my master's thesis. Always impressed by his ability to make time for all of his PhDs and his "big-picture view" to connect people to advance their projects and science in general.
All member of the ICB, and of the Machine Learning Group. In particular Lukas Simon, Malte Lücken, David Fischer and Benjamin Schubert for supervision, fruitful discussions and explanations to make sense of the wealth of data.

Herbert Schiller, for his guidance and for pushing me to newer heights I never deemed possible. Also for the spontaneous outburst of ideas lead to amazing results more often than not, this working experience will likely echo throughout my future endeavors.
All members of the Schiller Lab, especially the senior members Max, Ilias, Gabi, Christoph, Laura and Janine, which I could pester with organizational questions, but more importantly, for providing a welcoming work environment that made the PhD experience fun.
The *juniors* Niklas and Lukas for alleviating my computational burden towards the end.

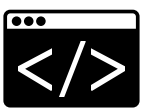
The CPC Research School for their groundwork and introduction to lung biology and translational methods to get me settled before attempting any data exploration. Thank you Claudia, Silke, Karin, Mareike and everyone involved!
The many excellent collaborations - in fact too many to list here - that were established across the institute. I will always be thankful for being part of these fascinating projects and am retrospectively glad I never turned down any of them.

Babs and Sandy for providing motivation and distraction especially during the restricted times in lockdown. And of course for putting up with my moods.
Michelle, for sharing struggles and achievements throughout the last years, both academically and privately. Always important to let off steam to get ready for the next success.
The Balli Balli Tangaballi Gang, for endless laughter and frequent breaks from everyday life, which were essential to not lose touch with the real world.
Fatima, for always having an open ear and keeping the focus on the things that matter.

Finally my family, for their never-ending support, interest in my work, and loving environment. Helma, Hans and my parents, for encouraging me to set out on this journey, and my brothers who never missed an opportunity to keep me grounded.

You all have my everlasting gratitude.

آپ سب کا میرا لازوال شکر ہے۔


Breathe, Code, Coffee in Mug
Then gasp – a novel Breakthrough?
... Nope, just another Bug.

"Work related Haiku" - First Theis Lab Group Day 2018