Dissertation

submitted to the

Combined Faculty of Natural Sciences and Mathematics

of the Ruperto Carola University Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

presented by

M.Sc. Timo Benedikt Trefzer

born in: Lörrach, Germany

oral examination: 08.02.2022

Comparison of single cell transcriptomics technologies and their application to investigate cellular heterogeneity in healthy and diseased lung

Referees: Prof. Dr. Karsten Rippe
Prof. Dr. Christian Conrad

Acknowledgements

First, I would like to thank Prof. Dr. Christian Conrad and Prof. Dr. Roland Eils for supporting me throughout my doctoral work. In particular, I thank Prof. Dr. Christian Conrad for his supervision and for giving me the opportunity to develop as a scientist and freely explore the fields of bioinformatics and single cell genomics. Similarly, I thank Prof. Dr. Roland Eils for providing an interdisciplinary work environment and funding the projects presented in this thesis. I thank the team of the Thorax Klinik Heidelberg, namely Prof. Dr. Michael Meister, Dr. Marc Schneider and Liz Meister, for the possibility to work in the interesting field of lung cancer biology and for our close collaboration with fruitful discussions. In addition, I thank Prof. Dr. Henrik Kaessmann and his group, as well as Dr. Jan-Philipp Mallm and Katharina Bauer, for interesting and insightful meetings as we conducted the work on establishing single cell technologies. For their help with animal tissue procurement and sequencing experiments during my first two years, I thank Thoomke Brüning and Dr. Katharina Mößinger. I am also grateful to Dr. David Ibberson for support with sequencing, as well as Manuela Schäfer, Corinna Sprengart and Sandra Bodogh for help with navigating administrative challenges. Furthermore, I would like to thank Prof. Dr. Detlef Arendt and Prof. Dr. Christoph Plass for guidance and helpful advice during my thesis advisory committee meetings, as well as Prof. Dr. Karsten Rippe and Prof. Dr. Stefan Wiemann for examining my thesis.

Special thanks go to my current and former colleagues, in particular Katharina Jechow for constant support in the lab, and the whole eilslabs department for a nice working atmosphere as well as fun and productive meetings and retreats. I also thank Dr. Teresa Krieger and Dr. Luca Tosti for very helpful advice on my work and proofreading this thesis.

Finally, I would like to thank my family, especially my parents Irene and Dieter Trefzer as well as my grandparents, for constant support and positive influence throughout my life and Teresa who always encouraged me and supported me throughout the last years.

Contributions

If not stated otherwise, all data presented in this thesis were obtained and analysed by myself under supervision of Prof. Dr. Christian Conrad and Prof. Dr. Roland Eils. Animal material was collected in collaboration with the group of Prof. Dr. Henrik Kaessmann, with Thoomke Brüning and Dr. Katharina Mößinger. Patient material was provided by the Lung Biobank Heidelberg. Immunohistological staining and scoring was performed by Dr. Marc Schneider. Next generation sequencing was performed by the genomics core facility at EMBL Heidelberg.

I highly appreciate that my scientific work was part of joint projects, reflecting that research generally constitutes a collaborative endeavour within an overarching scientific environment. Therefore, I will use the term ‘we’ throughout this thesis.

Abstract

Multicellular organisms rely on the concerted interaction of a multitude of cells, which are often highly specialised and give rise to complex tissues. As vastly different cellular phenotypes emerge from the same genotype that is shared across all cells of an organism, the transcriptome represents a key mediator driving different cell types and cell states that give rise to functional tissues. These are also subject to environmental factors or intrinsic changes that may disrupt homeostasis and lead to disease. In the human lung, the effects of tobacco smoke exposure, still the greatest risk factor for lung cancer, have not been fully resolved at the cellular level. Moreover, cellular heterogeneity may be significant for the emergence of lung cancer in never smokers, a growing proportion of global cases. A focused investigation of cellular heterogeneity in the healthy lung and lung cancers is therefore highly warranted.

During the last decade, technological advancements have made it possible to interrogate the transcriptome of single cells by novel next generation sequencing approaches. While previous studies were limited to averaging transcriptome information over many cells, single cell RNA sequencing (scRNA-seq) technologies are now enabling the investigation of cellular phenotypes in healthy and diseased tissues at unprecedented resolution.

In this thesis, I adapt different scRNA-seq technologies to process fresh or biobanked samples from different tissues and species, thus enabling comparisons across diverse origins. We identify specific advantages, limitations and experimental challenges associated with each technology.

I then perform a comprehensive single-cell transcriptomics study of healthy lung and lung adenocarcinoma (LADC). Based on twelve healthy lung samples, we generate a reference cell atlas that provides a rich resource for investigating cellular diversity in the human alveolar lung. Its utility is demonstrated by probing the expression of genes that are implicated in host cell entry of SARS-CoV-2 virus, thereby contributing to our understanding of coronavirus infections. By comparing single cell profiles from smokers and never smokers, we also resolve the involvement of distinct cell types in the maintenance of an inflammatory state in smoker lungs, and we identify key mediators of inflammatory processes induced by tobacco smoke exposure in fibroblasts and endothelial cells.

To investigate cell type diversity and microenvironment interactions in LADC, I analyse 26 tumour tissue samples and resolve functional malignant cell subpopulations linked by a differentiation hierarchy in both smokers and never smokers. They comprise proliferating and intermediate undifferentiated cells as well as two differentiated tumour cell states implicated in cancer progression and invasiveness. Distinct macrophage and fibroblast subpopulations which contribute to a tumourigenic environment are also detected. A subset of proliferating tumour cells show differential immune modulating activity dependent on smoking status, with implications for future treatment approaches.

Taken together, these results provide a comparison of rapidly developing scRNA-seq technologies for use in further studies and demonstrate their utility to dissect cellular heterogeneity and identify transcriptional programmes in the healthy and diseased lung. By applying these technologies, I add to our understanding of SARS-CoV-2 entry into human lung cells, define the alveolar lung cell types affected by tobacco smoke exposure, and provide deeper insight into cellular heterogeneity of LADC and the tumour microenvironment. These findings represent a valuable reference for future translational studies.

Zusammenfassung

Mehrzellige Organismen sind auf das koordinierte Zusammenspiel einer Vielzahl von Zellen angewiesen, die oftmals hochspezialisiert sind und sich zu komplexen Geweben zusammenfügen. Das Transkriptom setzt dabei phänotypische Unterschiede zwischen den Zelltypen und Zuständen funktionaler Gewebe um, die alle aus demselben Genotypen hervorgehen. Diese Gewebe sind außerdem Umweltfaktoren oder intrinsischen Veränderungen ausgesetzt, die die Homöostase stören und zu Krankheiten führen können. In der menschlichen Lunge sind die Auswirkungen von Tabakrauch, der immer noch den größten Risikofaktor für Lungenkrebs darstellt, bisher nicht vollständig auf zellulärer Ebene aufgeklärt worden. Außerdem könnte zelluläre Heterogenität von Bedeutung für die Entstehung von Lungenkrebs bei Nierauchern sein, die einen zunehmenden Anteil der globalen Krankheitsfälle ausmachen. Eine zielgerichtete Untersuchung zellulärer Heterogenität in der gesunden Lunge und Lungenkrebs ist daher dringend erforderlich.

Technologische Fortschritte im Verlauf der letzten zehn Jahre ermöglichen es inzwischen, das Transkriptom einzelner Zellen mittels neuartiger Sequenzierungsmethoden zu analysieren. Während vorherige Studien darauf beschränkt waren, den Durchschnitt transkriptioneller Information über viele Zellen zu messen, kann man mittels Einzelzell-RNA-Sequenzierung nun zelluläre Phänotypen in gesunden und erkrankten Geweben in noch nie dagewesener Auflösung entschlüsseln.

In dieser Dissertation implementiere ich unterschiedliche Technologien der Einzelzell-RNA-Sequenzierung, um Gewebe aus frischen Biopsien oder einer Biobank zu prozessieren, sodass Proben unterschiedlichen Ursprungs verglichen werden können. Wir identifizieren dabei die spezifischen Vorteile, Einschränkungen und experimentellen Herausforderungen der jeweiligen Technologien.

Anschließend führe ich eine umfassende Einzelzell-Transkriptom-Studie von gesundem Lungengewebe und Adenokarzinom der Lunge (LADC) durch. Basierend auf zwölf Proben von gesundem Lungengewebe generieren wir einen Zellatlas, der als Referenz und Ressource dient, um zelluläre Diversität in der distalen menschlichen Lunge zu untersuchen. Seinen Nutzen demonstrieren wir anhand der Expression von Genen, die am Wirtszelleintritt des SARS-CoV-2-Virus beteiligt sind, und tragen damit zu unserem Verständnis von Coronavirus-Infektionen bei. Indem wir Einzelzell-Profile von Rauchern und Nierauchern vergleichen, bestimmen wir auch die Beteiligung verschiedener Zelltypen an der Aufrechterhaltung eines inflammatorischen Milieus in Raucherlungen und identifizieren Schlüsselfaktoren von Entzündungsprozessen, die durch Tabakrauch in Fibroblasten und Endothelzellen induziert werden.

Um die Diversität von Zelltypen in LADC und deren Interaktionen mit dem Mikromilieu zu beleuchten, analysiere ich auch 26 Tumorgewebeproben und ermittle funktionale maligne Zellpopulationen in Rauchern und Nierauchern, die durch eine Differenzierungshierarchie verbunden sind. Sie beinhalten sich teilende und differenzierende Zellen sowie zwei ausdifferenzierte Tumorzellstadien, die mit Tumorprogression und Invasivität assoziiert sind. Wir detektieren auch verschiedene Subpopulationen von Makrophagen und Fibroblasten, die zu einer tumorfördernden Umgebung beitragen. In Abhängigkeit vom Raucherstatus weist eine Untergruppe der sich teilenden Tumorzellen unterschiedliche Immunmodulationsaktivität auf, mit Relevanz für zukünftige Behandlungsansätze.

Insgesamt liefern diese Ergebnisse einen Vergleich neuartiger Technologien für die Einzelzell-RNA-Sequenzierung, der in zukünftigen Studien genutzt werden kann. Sie demonstrieren ihren Nutzen für die Analyse zellulärer Heterogenität und die Entschlüsselung von Transkriptionsprogrammen in gesundem und erkranktem Lungengewebe. Durch Anwendung dieser Technologien trage ich zu unserem Verständnis des Eintritts von SARS-CoV-2 in menschliche Lungenzellen bei, bestimme die Zelltypen der distalen Lunge, die durch Tabakrauch beeinträchtigt werden, und zeige die zelluläre Heterogenität in LADC und dessen Tumormikroumgebung auf. Diese Erkenntnisse stellen eine wertvolle Referenz für zukünftige translationale Studien dar.

Table of Contents

1 Introduction

1.1 Characterisation of cellular phenotypes

Cells are the basic unit of life on earth and every multicellular organism comprises a multitude of morphologically and functionally very distinct cells that rely on the same genome. Our understanding of cells as the fundamental building blocks of organisms arises from works such as the 'Cell Theory' by Rudolf Virchow, and others, formulated in the 19th century. These observations were based on two centuries of scientific discovery, including Robert Hooke's first description of a cell in 1665 which was made possible by Antoni van Leeuwenhoek's microscope [1].

Since then, researchers have tried to classify cells into different cell types that react to and interact with their environment in a characteristic way, and thereby contribute to the makeup of tissues and ultimately whole organisms. This classification was initially based on morphological differences between cells, but with increasing knowledge and understanding of the molecular processes underlying cellular phenotypes, descriptions were expanded to include their different constituent molecular layers, namely DNA, RNA and proteins.

A typical cell contains 6 pg of DNA, 50,000-300,000 mRNA molecules (5-30 pg) and millions of proteins (20-200 pg) [2-5]. For the longest time, classification of phenotypes at the level of single cells focused on the identification of proteins using microscopy technologies, like immunofluorescence light microscopy. Through progressive improvements in these technologies, imaging based classification of cells now encompasses the quantitative characterisation of molecular traits, for example of RNA molecules using fluorescence in-situ hybridisation (RNA-FISH) and structural characteristics of DNA using various fluorescent dyes [6-8]. However, all of these approaches require some prior knowledge about the target, such as protein structure for epitope detection by antibodies, or RNA molecules for hybridisation to a consensus sequence.

The first step towards an unbiased method to interrogate the molecular makeup of cells was taken with microarray technology, where isolated RNA was transcribed to cDNA, fluorescently labelled and hybridised to known complementary DNA sequences [9]. While this increased the number of genes that could be probed, the first truly untargeted approach only emerged with the ability to reconstruct the sequence of an unknown DNA molecule [10]. Today, a variety of what are now called next generation sequencing technologies (NGS) allow us to determine the sequence of any given DNA or RNA molecule at high precision [11, 12]. Fundamentally, these sequencing technologies rely on fragmenting DNA into smaller pieces (typically around 300 nucleotides) or transcribing RNA into cDNA with a reverse transcriptase. Nucleotide fragments are then fixated to a surface, and a DNA polymerase catalyses the sequential incorporation of fluorescently labelled nucleotides into antisense strands. By fluorescence imaging of this replication process, the order of nucleotides in the given fragment – the sequence – can be inferred. This sequence can now be compared to a reference genome of the appropriate species to determine its original genomic position, or the position of its template in the case of RNA.

The first human reference genome was made available by a huge international collaborative effort, the Human Genome Project, which started in 1990 and finished in 2003 [13]. With the availability of NGS, this reference has since been refined, and reference genomes for a multitude of other species have also been constructed. While the Human Genome Project cost as much as $3 billion, these new technologies enable the sequencing of a whole human genome in a matter of days at the cost of just a few thousand US dollars, making it feasible to address scientific questions rapidly and at scale [13].

However, until recently, our ever-improving capability to characterise the molecular phenotype of cells was not applicable at the level of single cells, because the amount of material required for NGS technologies far exceeded the typical quantity of DNA or RNA found in one cell and technologies to capture and amplify nucleic acids at single cell resolution were lacking. Studies of cellular genomes and transcriptomes were therefore limited to averaging over a large number of cells. When tissue samples were processed, this typically included multiple distinct cell types; and even in cases where individual cell types could be isolated beforehand by sorting techniques, bulk analyses were likely to overlook profound heterogeneity in gene expression between cells, given accumulating evidence for transcriptomic variability even within the same cell type [14, 15].

1.2 Single cell RNA sequencing

While DNA sequencing plays a vital role in the diagnosis of genetic diseases and the identification of mutations in cancers, RNA sequencing provides a snapshot of the transcriptional activity of a cell, its transcriptome, as a proxy for its functional state at a given timepoint. Sequencing the transcriptome of one single cell was pioneered in the early 1990s, when methods to create cDNA from the minute amounts of RNA present in a cell were developed [16, 17] and later applied to microarray-based identification of gene expression [18-21]. The first analysis of single cell transcriptomes based on NGS was published in 2009 on early embryonic development [22]. Since then, single cell RNA sequencing (scRNA-seq) technology has developed rapidly and is now applied across all areas of biology, sprouting large international efforts to study biological systems at unprecedented resolution. One such effort, the Human Cell Atlas Project, aims to chart the transcriptomes of all cells in the human body [23].

This rapid spread of a rather new methodology was made possible by the development of various technological approaches to address three major obstacles inherent in studying the molecular makeup of a single cell. Firstly, cells have to be isolated from their natural environment and compartmentalised in separate reaction volumes. Secondly, the minute amount of RNA in one cell requires novel experimental procedures for recovering and amplifying this material. Thirdly, the acquired sequence data poses novel challenges for analysis due to high technical variability and sparseness in single cell derived transcriptome datasets (Figure 1.1).

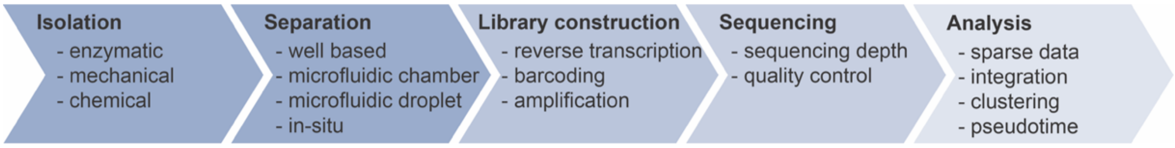


Figure 1.1 Challenges of single cell sequencing. There are three main challenges in scRNA-seq studies. (i) Isolation of cells from their natural environment and separation in distinct reaction volumes. (ii) Construction of sequencing libraries from minute amounts of material and appropriate sequencing. (iii) Computational analysis of extremely sparse data with high technical variability.

1.2.1 Isolation of single cells

To analyse the transcriptomes of single cells without prior knowledge, they first need to be isolated from the surrounding tissue and contacts with other cells or macromolecules such as the extracellular matrix (ECM) have to be broken enzymatically, chemically or mechanically. This process of extraction leads to the loss of spatial information about the cell and its natural microenvironment, and might also induce changes in the transcriptome which can affect downstream analysis [24-26].

Subsequent isolation of each cell's RNA into separate reaction compartments is the next critical step for sequencing individual transcriptomes. Initially, this was achieved by manual manipulation of single cells by pipetting [22] or limiting dilution [27]. However, the inefficiency of these technique made it unfeasible to process more than a few hundred cells. One of the earliest methods employed to increase the throughput of single cell RNA sequencing was flow cytometry, which allows for sorting of cells into separate wells of a reaction plate based on their physical properties or fluorescence signal (FACS) [28, 29]. When combined with fluorescence labelling of marker proteins, flow cytometry thereby also enables the analysis of specific subpopulations of cells [30].

A major breakthrough that finally enabled the sequencing of thousands or even millions of single cell transcriptomes in a single study was achieved by miniaturisation of the reaction volume to the nanolitre scale. This advancement not only reduced the cost of reagents, but also considerably improved the sensitivity of these assays, enabling scRNA-seq approaches to construct a more faithful representation of the actual transcriptome in a single cell [31]. Miniaturisation is here most commonly realised through microfluidic [32-34] (compare Figure 2.1) or plate-based technologies [35, 36]. Recent approaches have aimed to further reduce costs by utilising the cell itself as a reaction compartment, confining RNA molecules inside the cell membrane while rendering them accessible to the required reagents [37]. This method avoids the need for expensive microfluidic equipment, and it has been shown to identify cell types at similar efficiency compared to competing technologies while being much easier to scale at lower cost. Still in the developing stage, future investigations of this technology will provide more insights into its applicability for different tissues and experimental robustness [37].

1.2.2 Library construction for NGS

After the successful isolation of each cell's RNA, a sequencing library suited for NGS has to be constructed. This typically consists of three steps, including cell lysis, reverse transcription of RNA into cDNA, and cDNA amplification to provide enough input material for sequencing technologies.

Lysis is typically achieved using a hypotonic buffer and mild detergents, and first strand cDNA synthesis is performed with a poly(dT) primer and reverse transcriptase to select for poly(A) messenger RNA (mRNA). As current approaches result in reverse transcription of only 10-20% of RNA

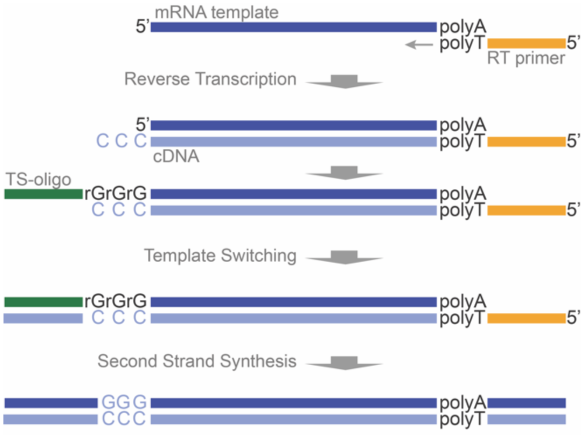


Figure 1.2 Template switching. By using a specific reverse transcriptase paired with a primer that anneals to introduced, protruding nucleotides at the 3' end of the nascent cDNA, template switching improves transcription efficiency and enables transcription of the full length of an mRNA molecule.

molecules, this step introduces another major bias into the scRNA-seq data [38, 39]. Different strategies for reverse transcription of mRNA exist, but the most common today are poly(A)-tailing and template switching [40], with the majority of all current high-throughput protocols relying on the latter.

Template switching not only provides superior efficiency for the minute RNA input amounts in scRNA-seq, it also ensures, unlike poly(A)-tailing, transcription of the full length of an mRNA molecule. This technology harnesses a specific property of the Moloneymurine leukemia virus (MMLV) reverse transcriptase, which is thought to add three protruding nucleotides (+CCC) to the 3' end of the nascent cDNA [41, 42] (Figure 1.2). A second primer (template switching oligo; TS-oligo) containing a matching sequence of three riboguanosins (rGrGrG) can then anneal to the nascent cDNA and allows the reverse transcriptase to switch templates; thus, the transcript can be elongated with a known nucleotide sequence and the entire mRNA sequence can be amplified.

Subsequently, the extremely low amount of transcribed cDNA needs to be further increased either by linear in-vitro amplification or exponential polymerase chain reaction (PCR). While linear amplification is less prone to introduce bias by preferential amplification of certain genes or different ratios of gene products, most protocols rely on exponential PCR as it is much less labour intensive [43, 44]. Therefore, the majority of scRNA-seq technologies are fundamentally based on the SMART method, which combines reverse transcription by template switching and PCR to generate a sequencing library suited for modern NGS from the extremely low amounts of RNA in one cell [40].

To date, single cell sequencing technologies have predominantly focused on examining the transcriptome of single cells, but sequencing-based methods have also been developed for other molecular properties. These include DNA [45, 46], epigenetic modifications and chromatin accessibility [47-49], and protein expression as determined by antibody sequencing [50]. Recent advancements even enable the integration of different approaches to simultaneously analyse multiple modalities from the same cell, termed multiomics [50-63].

1.2.3 Challenges in scRNA-seq data analysis

After using NGS to acquire sequence information for each library molecule, termed the sequencing reads, this data has to be pre-processed before downstream analysis. Usually, standard quality measures for RNA sequencing are applied, including a certainty score for each base of the sequence and overall base composition to discard low quality libraries [64]. In scRNA-seq approaches, reads also need to be assigned to the cell they derive from; this is achieved by a barcode sequence unique to each cell which is integrated into each read during reverse transcription of the cDNA or library construction. To translate sequence information into gene expression data, each read is aligned to a reference genome and a matrix of read counts per gene in each cell is constructed. However, due to the often-used exponential amplification of the sequencing library and missed molecules, these counts may not reflect the true molecule count of RNA across different cells (section 1.2.2). To circumvent this problem, bulk RNA sequencing and early scRNA-seq methods added a mix of known RNA molecules during library construction, reasoning that the read counts for this “spike-in” RNA would enable inference of any technical bias that should affect the cell’s native RNA in the same way [65]. The main problem with this method for single cell transcriptomics is that the amount of “spike-in” RNA needs to be delicately balanced. While too little might not result in useful information, too much would mask the signal from the cells’ RNA [65]. Given the high variability in RNA amount and composition for each individual cell, this method is consequently not very well suited for scRNA-seq. Therefore, most technologies now introduce short sequences called unique molecular identifiers (UMI) into each molecule during reverse transcription. As they are unique to each molecule, these identifiers can then be used to distinguish and remove PCR duplicates emerging during amplification, reducing potential biases introduced in the cDNA amplification step.

Another fundamental problem for the analysis of scRNA-seq data is the substantial sparseness of the data, meaning a high number of zero-counts, often referred to as “drop-outs” [66-68]. The latter term, however, is somewhat misleading, because zero-counts include values that occur due to technical noise as well as values that represent truly unexpressed genes [67, 68]. Zero values attributable to technical variation can be caused by different degradation susceptibility of RNA molecules, transcription as well as amplification differences or stochastic events during sequencing, especially in lowly expressed genes. Thus, beyond the expression level of each gene, data sparsity is affected by cell capture technology, library construction method and sequencing depth. Together with biological noise, e.g. oscillating expression of cell cycle genes, this hampers cell type and cell state identification and further downstream analyses [69, 70]. Suitable normalisation techniques for scRNA-seq data are therefore required to mitigate the bias introduced by zero-counts. Finding the optimal normalisation method remains an unsolved challenge although new techniques are constantly developed, which often model RNA counts using probabilistic approaches to estimate the true gene expression for each cell [66, 70-78].

The complex nature of many biological problems demands sampling not only once but e.g. at different time points, different locations or across different organisms. In scRNA-seq studies, methods to account for technical and biological noise are thus required when handling multiple samples. Accordingly, apart from normalization techniques, the development of computational approaches to correct for batch effects is also a rapidly evolving field. Depending on the specific experimental set-up, the most suitable method can be selected for batch correction [76, 79-94].

To derive biological insight from the data thus processed, a common approach is to define groups of cells with shared transcriptional characteristics. Most often this is achieved by reducing the dimensionality of the data by principal component analysis (PCA) [95] and finding communities of cells by constructing a shared nearest neighbour (SNN) graph [96]. The distinguishing features of cell populations are then characterised for example by differential gene expression analysis, often using the Wilcoxon rank sum test [97] or other methods specifically developed for single cell RNA analysis [98, 99].

Another popular method to distinguish groups of cells with similar expression profiles is non-negative matrix factorisation (NMF), which decomposes the gene-by-cell count matrix ***V*** into the product of two matrices, ***V*** = ***W*** x ***H*** (Figure 1.3). Gene sets that contribute to each factor may then be determined based on the gene-by-factor matrix ***W***, while groups of cells with similar expression profiles may be inferred from the factor-by-cell matrix ***H***. In practice, the dimensions of the factor matrices ***W*** and ***H*** are usually chosen to be much smaller than those of the original matrix [100] and the factorisation is approximated by ***V*** = ***W*** x ***H*** + ***U***, where the contribution of the residual matrix ***U*** is minimised. A soft clustering approach may then be used to identify populations as well as subpopulations of cells from the factor-by-cell matrix ***H*** without the need for iterative clustering.

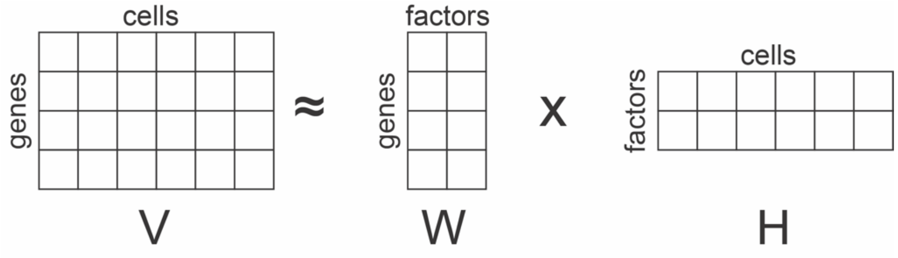


Figure 1.3 Illustration of approximate matrix decomposition. Matrix ***V*** is decomposed into the product of two matrices ***W*** and ***H***.

However, choosing a meaningful value for the number of factors presents a challenge and relies on biological insight. Ultimately, the cell populations identified by NMF with their transcriptional characterisations can give insight into the diversity of distinct cell types present in the tissue, as well as different functional states that may be occupied by cells of the same cell type.

In most tissues, cells exist not only in transcriptionally distinct populations of cell types but on a continuum of gradual changes, for example along a differentiation path such as the development from lymphoid progenitor cells to dendritic cells, B cells, T cells and natural killer cells, which may again comprise different cell states (e.g. activated and non-activated T cells). To analyse these trajectories, over 70 computational methods have already been developed [101]. They all aim to sort cells based on their transcriptional similarity along an axis of change most often termed pseudotime. The unknown topology of the underlying process represents a significant challenge here. Most methods therefore assume a linear or tree like model of cell state relation [102, 103], but new approaches also try to infer pseudotemporal ordering in complex graph based topologies [101, 104]. Once a trajectory is determined, a major hurdle in the further downstream analysis of transcriptome data is the identification of genes that change along this trajectory. Until now, most analytical approaches are limited to differential gene expression between branches of the trajectory or distinct subpopulations of cells along the pseudotime axis.

Despite the challenges and open problems associated with scRNA-seq data processing, single cell transcriptomics has already led to remarkable discoveries during the past decade, including the identification of a new cell type in the lung [105] and the application of single-cell technologies to guide therapeutic intervention [106]. As I will explore further in this work, it also offers a promising avenue to better understand tumour heterogeneity and evolution [107, 108].

1.3 Cell types of the human lung

As described above, cell type characterisation has long been driven by microscopic approaches. In the lung, they have led to the discovery of dozens of cell types [109, 110]. In recent years, microscopic imaging has been complemented with molecular analyses of marker protein expression that enable a more refined description of lung cell types [111, 112]. Single cell transcriptomics, in particular, has now been employed to further resolve the diversity of cell types in the lung, and has facilitated the detection of a novel rare cell type as well as the subdivision of known cell types into multiple classes [105, 113-115]. With increasing amounts of data resulting in more fine-grained classifications, the traditional distinction between defined cell types on the one hand and different states of the same cell type on the other hand is beginning to blur [116]. In this section, I will introduce the major cell types of the lung as they are canonically classified today (Figure 1.4).

As the lung constitutes a metabolically very active tissue and to permit gas exchange between air and blood, there is a high degree of vascularisation in the lung, facilitated by endothelial cells that line arteries, veins and lymphatic vessels [117, 118]. The basic structural and functional integrity of the lung is maintained by stromal cells, comprising smooth muscle cells, pericytes, mesothelial cells and various fibroblasts [119].

The main function of the lung is gas exchange, which requires a huge surface area. Starting from the trachea, human lung airways therefore progressively split from proximal to distal into ever smaller passages that total about 2^21 – 2^23 branches, culminating in highly vascularised spherical structures, the alveoli, which represent 99% of the total surface area in the lung [110, 120]. Lung airways are lined with a continuous epithelial layer predominantly comprised of basal, ciliated, club, goblet, mucous, serous and neuroendocrine cells. The composition of these epithelial cells changes from proximal to distal airways, with fewer mucosal, ciliated and basal cells and increased numbers of club and other secretory cells in more distal branches [121, 122]. In addition, the more specialised alveolar type 1 and type 2 cells (AT1/2) are found only in the alveoli [110, 121, 123]. Here, cell composition differs entirely from the proximal branches as alveoli contain a single squamous epithelial layer of thin, flat AT1 cells that are the main facilitators of gas exchange and in close contact with the vascular system [124, 125]. AT2 cells are also found in large numbers, mainly responsible for the maintenance of alveolar surface tension by surfactant production and regulation. They further play a role in immune regulation and interaction with the microenvironment via secretory factors [126, 127]. While basal cells predominantly effect tissue homeostasis and damage repair in the proximal lung [128, 129], AT2 cells in alveoli can also undergo proliferation and are capable of differentiation into AT1 cells, making them another prominent candidate for lung cancer cells of origin [127, 130, 131].

Apart from its role in gas exchange, the lung serves as a first line of defence against toxins and pathogens, due to its direct contact with the environment. Therefore, immune cell types are abundant in the lung, comprising the main cell types of the immune system, such as B and T lymphocytes, plasma, natural killer and basophil cells, as well as neutrophils, monocytes, dendritic cells and different macrophage populations, including specialised alveolar macrophages [119, 132-134].

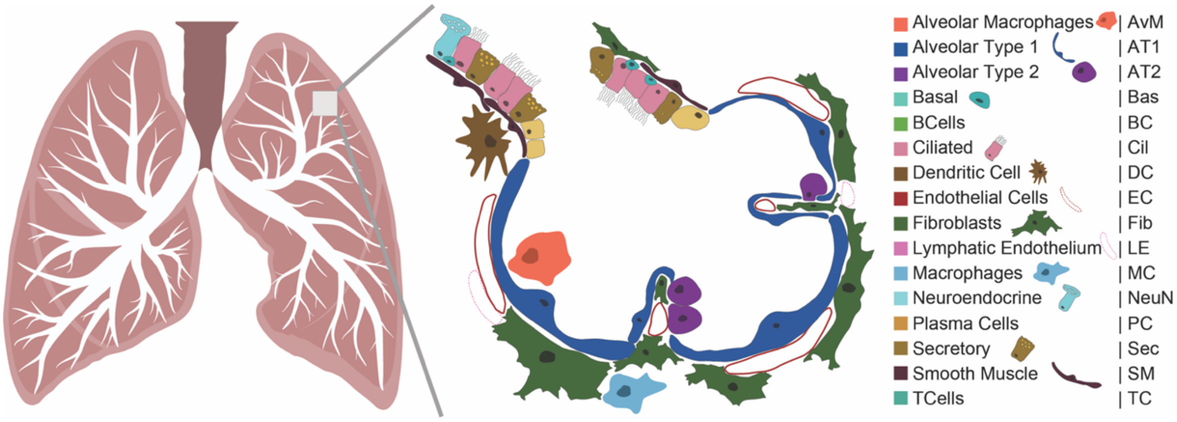


Figure 1.4 Cell types in human lung alveoli. The alveolar space of the lung comprises specialised cell types such as alveolar type 1 and 2 cells (AT1/2) and alveolar macrophages (AvM), as well as cells also found in more proximal parts of the lung such as basal (Bas), ciliated (Cil), secretory (Sec), neuroendocrine (NeuN), smooth muscle (SM), endothelial (EC), lymphatic endothelial cells (LE) and fibroblasts (Fib). Immune cells, including B cells (BC), plasma cells (PC), T cells (TC), macrophages (MC) and dendritic cells (DC), are also present in the lung.

1.4 SARS-CoV-2

In the work culminating in this thesis, I generated scRNA-seq data of samples from healthy lung and lung tumours with the goal to study lung cancer heterogeneity. In a first step, as detailed in section 2.2.1, this resulted in an atlas of healthy lung cell types that supports and refines the cell type classification described above. While this work was in progress, in late 2019, a novel coronavirus variant affecting the respiratory system emerged and rapidly spread across the globe [135-137]. Due to its similarity to other respiratory syndrome coronaviruses, such as severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV), it was named SARS-CoV-2, and the disease it causes became known as COVID-19 [138]. Since this virus predominantly provokes damage in the lung, we decided to share insights gained from our healthy lung samples with the scientific community to help address this severe global health threat [139].

Like other coronaviruses, SARS-CoV-2 has a major spike protein that initiates cell infection by binding to a receptor on the target cell membrane, leading to fusion of the viral lipid envelope with the cell's membrane. To be activated, after receptor binding but before membrane fusion, the spike protein needs to undergo proteolytic cleavage. Several proteases have been suggested as potential activators of the spike protein, including Furin, Cathepsin L, Transmembrane Serine Protease 2 (TMPRSS2), TMPRSS11A and TMPRSS11D [140-142]. Of these, Furin and TMPRSS2 are most often observed to play a role in activation of a broad range of virus proteins [143]. Studies conducted soon after the emergence of SARS-CoV-2 showed that infection was dependent on the virus binding to Angiotensin-Converting Enzyme 2 (ACE2), and entry of the virus was blocked by a TMPRSS2 inhibitor in a cell culture model [144-146]. These results suggested that ACE2 represents the receptor binding SARS-CoV-2 while TMPRSS2 activates the spike protein.

We therefore used our atlas of healthy lung cells to add to the knowledge of cell type specific expression of ACE2 and candidate proteases as described in section 2.2.2.

1.5 Lung cancer

While our characterisation of healthy lung cell types contributed to the understanding of SARS-CoV-2 entry into cells [139], the main focus of this thesis was the investigation of cellular heterogeneity and microenvironment interactions in healthy lung tissue and lung cancer in patients with or without a smoking history. In addition to the characterisation of healthy lung samples as described in section 2.2.1, scRNA-seq was therefore also performed on lung cancer samples, specifically lung adenocarcinoma (LADC).

Lung cancer is the leading cause of cancer deaths worldwide and has the third highest incidence among all cancer types, after breast and prostate cancer, with an age standardised rate (ASR) of 31.5 per 100,000 in men and 14.6 per 100,000 in women worldwide. Among all cancers, it has the highest mortality rate in men (25.9 ASR per 100,000) and the second highest in women (11.2 ASR per 100,000) after breast cancer [147].

The vast majority of lung cancer cases can be attributed to tobacco smoke, which accounts for about 90% of all cases [148]. Accordingly, the incidence of lung cancer has declined since the 1980s in the United States of America and other industrialised countries, after the health risks imposed by tobacco smoke were acknowledged and smoking prevention programmes established [147, 149, 150]. One of the most important demographic contributors to lung cancer cases, as with many other cancer types, is age [147, 151]. More than 90% of all cases occur in patients over 55 years of age, with the highest incidence in women between 75 and 79 (365.8 per 100,000) and men between 85 and 89 (585.9 per 100,000) years of age [150]. Other occupational and environmental risk factors include asbestos [152, 153], indoor smoke from cooking and heating with fire, air pollution and diesel exhaust [154-156], radon [150, 157], as well as radiation therapy [158, 159]. In addition, there are other lung diseases that have been associated with an increased risk for developing lung cancer, most likely due to inflammatory responses [160]. As an example, chronic obstructive pulmonary disease (COPD), a chronic inflammation of the lung, elevates the risk for lung cancer two to six times. While it is estimated that up to 50% of lifelong smokers develop COPD and the mortality rate of lung cancer correlates with the presence of this comorbidity, it also acts as a risk factor independent of age, smoking or sex [161, 162].

Clinical manifestations of lung cancer are diverse and classification has traditionally been based on pathological observations. Therefore, all lung cancers are broadly divided into two subclasses based on histopathology, non-small cell lung carcinoma (NSCLC) and small cell lung carcinoma (SCLC), with NSCLC according for about 80% of all cases [163, 164]. As indicated by the name, SCLC presents as very small cells with little cytoplasm; it is highly aggressive and, in contrast to NSCLC, most often not suited for surgical resection [165]. NSCLC is further divided into adenocarcinoma, squamous cell carcinoma and large cell carcinoma [164]. Technical developments and an improved understanding of biological mechanisms that drive cancer progression have enabled an even more refined stratification of these histological classes to provide guidance for more targeted therapeutic decisions [166]. While SCLC is very highly associated with smoking and ionising radiation [167], the most frequent type of NSCLC, lung adenocarcinoma, is most prevalent in non-smoking patients [168]. LADC presently accounts for about 40% of all lung cancer cases, and as tobacco consumption is declining in high income countries, it appears likely that its share will increase further in the future [169-171]. To address this challenge, novel mechanistic insights into LADC biology and the role of tobacco smoke in its aetiology are urgently needed. In this thesis, I will thus focus on LADC and its molecular composition at the single-cell level in smokers compared to never-smokers.

1.5.1 Lung adenocarcinoma

The histopathology of LADC shows an invasive epithelial neoplasm with high heterogeneity in growth patterns, which include lepidic, acinar, papillary, micropapillary and solid [163, 164, 172] (Figure 1.5). Lepidic growth, common in LADC, occurs along alveolar structures. Acinar patterns represent glandular structures in the stroma. Papillary and micropapillary tumours have a protruding papillary architecture, while the micropapillary type lacks vascularisation in its core. Solid growth is defined by the absence of the distinct features shown by the other patterns, and appears as homogenous sheets [163, 173]. Since most LADC cases present as a mixed phenotype, tumours are classified according to the most prevalent histopathological pattern, with notice of those co-occurring.

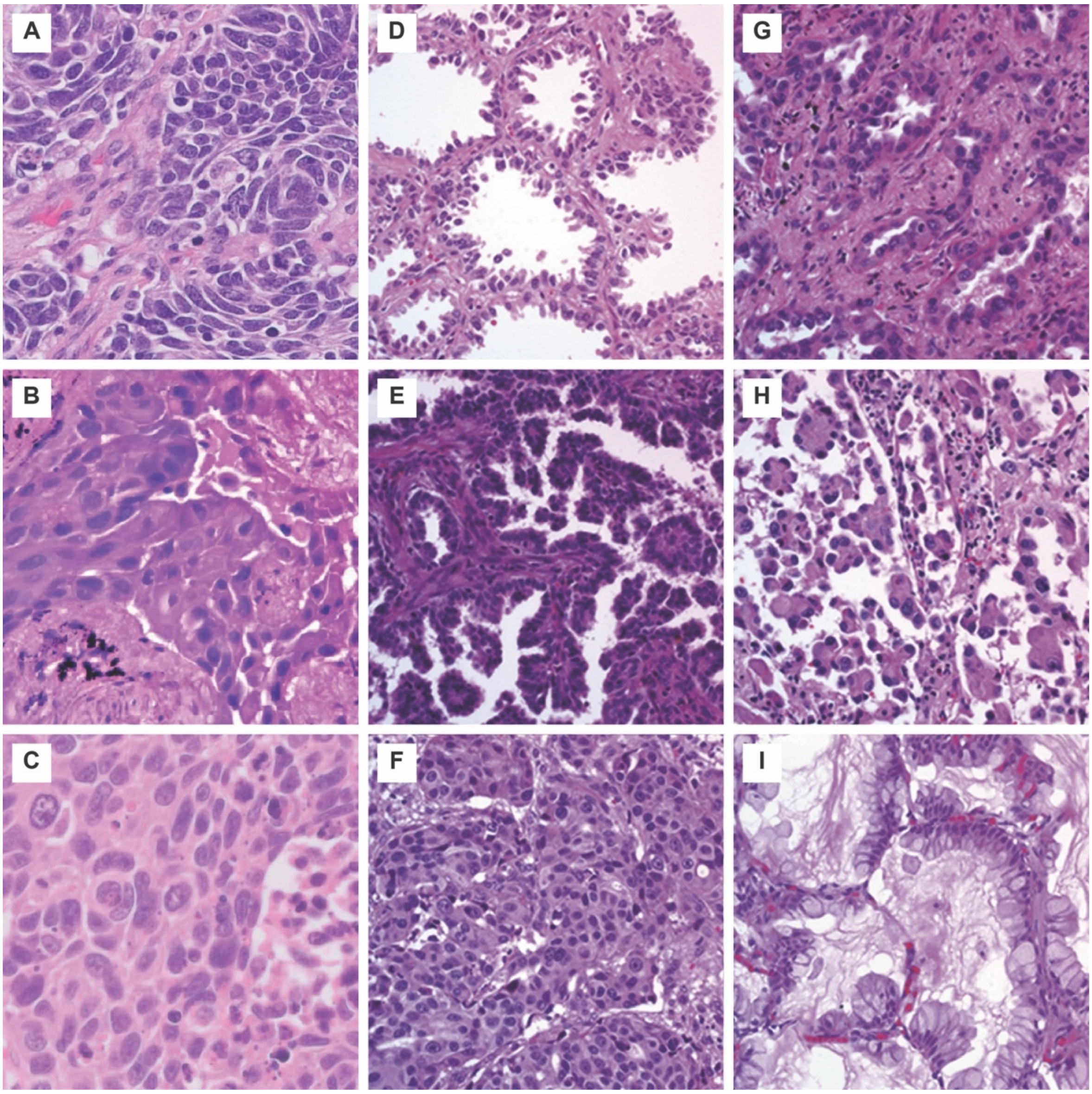


Figure 1.5 Subtypes of lung cancer based on histopathology. The histological subtypes of lung cancer include **(A)** small cell lung carcinoma and **(B-I)** non-small cell lung carcinoma. The latter is subdivided into **(B)** squamous and **(C)** large cell carcinoma and **(D-I)** various subtypes of lung adenocarcinoma. These comprise **(D)** lepidic, **(E)** papillary, **(F)** solid, **(G)** acinar, **(H)** micropapillary and **(I)** mucinous growth patterns. (adapted from [174-176]; original magnification: 40x for A,C and 400x for B,D-I).

As our understanding of the molecular mechanisms involved in LADC development and the availability of targeted treatment options are improving, the histopathological classification of LADC is usually augmented using molecular features acquired by sequencing the tumour genome [164]. Most commonly in LADC, mutations in Kirsten rat sarcoma viral oncogene homolog (*KRAS*), epidermal growth factor receptor (*EGFR*) and proto-oncogene B-Raf (*BRAF*), as well as a *EML4-ALK* translocation are detected, with *EGFR* and *KRAS* found to be mutually exclusive

[166]. Mutations are also often detected in *PIK3CA*, *MET*, *HER2*, *MEK*, *NRAS*, *AKT* and *TP53* [166, 177-179]. Major advances over the past years have resulted in specific therapies targeting several of these aberrations, with one of the highest success rates in using tyrosine kinase inhibitors for tumours harbouring mutations in *EGFR*, *HER2* or *ALK* [166].

Clear relationships between these mutations and the aforementioned histological patterns would facilitate diagnostic procedures, but evidence for an association of histological phenotypes with one or the combination of several mutations remains contradictory [172, 180-184]. The co-occurrence of histological patterns in mixed tumours, impeding unambiguous classification, probably represents a major obstacle here.

While knowledge of the mutations present in the tumour genome can contribute to selecting the appropriate therapy for each patient, a better mechanistic understanding of phenotypic changes in cancer cells may be achieved through transcriptome studies. RNA sequencing of tumour samples has consequently provided additional insight into more complex mutational landscapes [185] and suggested biomarkers on the RNA level for stratification of NSCLC [186]. In this way, many genes involved in LADC and other lung cancers have been discussed, such as genes regulating oxidative phosphorylation, DNA replication and proliferation [187]. In addition, connections between gene expression and overall survival have been proposed, as e.g. *AGER* and *SPP1* have been associated with poor survival [188].

While these analyses have already had enormous impact on the treatment of patients, they lack the resolution to characterise the biology of single tumour cells and their interactions with the microenvironment. Both genetic and transcriptomic data are routinely generated by sequencing a piece of tissue acquired via biopsy, which comprises all of the different cell types found in a tumour, and thus represent an average across these cell types. In contrast, recent single cell sequencing technologies enable the refinement of mechanistic insights into cell types and cell states as discrete units, with immense potential to resolve intratumoural heterogeneity and tackle challenges for personalised therapy in diverse tumours.

1.5.2 Single cell RNA sequencing of LADC

During the tumourigenesis of LADC, several molecular processes contribute to increasing cellular diversity. Genomic instability occurs especially during the early stages of tumour development and leads to mutations, as detailed in section 1.5.1, as well as structural changes such as copy number variations [189-192]. Together with epigenetic alterations such as DNA methylation, histone modification and non-coding RNAs, this results in altered expression and regulation of oncogenes and tumour suppressor genes [193, 194]. By successive acquisition of these alterations, neoplastic and finally malignant cells emerge. This progression not only results in phenotypic differences between individual patient tumours, but also drives an evolutionary process that generates great heterogeneity within a single tumour [108, 195]. Intratumoural heterogeneity in LADC has up to now mainly been characterised by histology and immunofluorescent imaging, requiring a targeted investigation with a high degree of prior knowledge, or by multiregional sequencing [196-198]. Only within the last few years, it has become possible to investigate this heterogeneity with single cell sequencing approaches (section 1.2).

The first scRNA-seq studies examining transformed cells in LADC confirmed substantial intratumoural heterogeneity and found evidence for specific transcriptional signatures in subsets of cells. By conducting a longitudinal study of patients before, during and after targeted treatment, a signature resembling alveolar cells was found in tumours that persisted after treatment, while progressing tumours showed an increase in inflammatory signalling that declined in residual tumours [199]. Other studies identified deregulation of epithelial transcriptional programmes in tumour cells and possible differentiation paths to ciliated or alveolar phenotypes [200].

It is well acknowledged that cancer can only be fully understood by taking into account the tumour microenvironment (TME), which may hinder or contribute to tumour progression [189, 201]. Many studies have therefore focused on investigating this compartment. Different types of fibroblasts in the tumour microenvironment have been described, distinguished by the expression of genes such as different sets of collagens and endothelial cell expression signatures, that might contribute to angiogenesis and tissue remodelling [119, 200]. The immune compartment with its major role in cancer development has also received keen attention, especially due to its involvement in tumourigenic inflammatory processes and the need for tumour cells to evade destruction by the immune system [189].

Importantly, inflammation caused by extrinsic factors such as air pollution, tobacco smoke or virus infection has been linked to providing an immunosuppressive and tumourigenic environment [202]. Following the onset of neoplasia, the interplay of different immune cell types and other cells of the TME is very complex and contradicting observations have sometimes been reported, possibly caused by sampling at different disease stages or interpatient heterogeneity. Yet there is good evidence for changes in cell type composition and transcriptome profiles in the immune compartment that occur in the presence of LADC. These include a depletion of immune cells at the tumour site or compositional changes within the immune compartment that compromise anti-tumour immunity. For example, the number of cytolytic B cells is reduced while immunosuppressive, PPARG-expressing macrophages emerge in the presence of LADC. A multi-region single cell transcriptome study also found evidence for a loss of immune checkpoints in areas spatially closer to the tumour [203]. In addition, tumour-associated macrophages have also been observed to produce high levels of IL6 which may promote tumourigenesis, and tumours may themselves trigger inflammatory responses [204-206].

The diverse interactions between LADC and the immune compartment have profound implications for immunotherapy, which has proven very successful in only a subset of patients. As our understanding of this interplay is still nascent, it remains poorly understood how underlying interpatient differences in the immune compartment, but also heterogeneity within one tumour, might influence therapeutic outcomes [207-209].

1.5.3 Smoking and LADC

Many different diseases have been associated with tobacco smoke exposure, with chronic obstructive pulmonary disease and lung cancer having the highest mortality [210, 211]. The risk of developing lung cancer is increased 20-fold in smokers compared to never smokers [152].

The exact mechanisms by which tobacco smoke affects lung physiology are still not fully discerned, partly because tobacco smoke comprises a complex mixture of more than 5,000 chemicals [212]. It is, however, well established that smoking causes damage to the epithelium and alveoli of the lung [213, 214] and leads to impairment in ciliary function [215]. It also causes immigration of immune cells into the lung tissue [214] and invokes an inflammatory response [216]. In addition, many of the compounds found in tobacco smoke are known carcinogens that introduce genomic alterations, which have been shown to persist even in former smokers for many years [217, 218].

Among lung cancers, smoking is strongly associated with all histological subtypes (Figure 1.5). Smokers are especially likely to develop squamous cell carcinomas, while never smokers more often develop LADC [219-221] and small cell carcinomas only represent an estimated 1.5% of lung cancer cases among never smokers [222]. Tobacco smoke exposure also affects the genetic composition of lung cancers. Within NSCLC, the high mutational burden imposed by smoking is reflected in a higher overall mutation frequency in smokers [223], and there is evidence for a difference in driver mutations between smokers and never smokers. Specifically, *EGFR* mutations seem to be more common in never smokers, with one study finding *EGFR* mutations in 45% of all never smoking patients compared with 7% of smoking patients [224].

With the availability of whole transcriptome sequencing, it became feasible to assess smoking related changes in the lung on a more mechanistic level. Gene expression changes due to tobacco smoke exposure have thus been identified, some of which persisted for decades after smoking cessation [225-227]. The most significant of these transcriptional differences were related to immune processes, inflammation and cell death [228]. Smoking-induced inflammation has also been detected based on higher plasma levels of cytokines like IL6, CRP and fibrinogen in smokers [229], and an increase in alveolar macrophages as mediators of inflammation in the alveolar fluid [230-232] and histological sections [233].

As described in the previous section 1.5, an inflammatory environment is thought to contribute to cancer initiation and progression [160, 234, 235]. In smokers, persistent inflammation may prompt normally quiescent stem cells to proliferate, facilitating neoplastic transformation of lung epithelial cells [236, 237]. This transformation could be initiated or promoted by reactive oxygen species, released by immune cells like alveolar macrophages recruited to the site of inflammation, that might damage the DNA of surrounding cells and add to the already heavy mutational burden induced by the various DNA damaging agents contained in tobacco smoke [238].

Most of the work on the effects of smoking on the lung and its implication for lung cancer development has been conducted using lavage to obtain cells from the lung lumen or bulk biopsy material, therefore representing an average across diverse cell types. More recently, single cell RNA sequencing has been employed to interrogate the effects of tobacco smoke on individual cell types. This has uncovered dysregulation of peripheral blood monocytes in smokers [239], a shift in metabolic gene expression in bronchial cells [240], and transcriptomic changes which may impair tissue regeneration after injury in tracheal epithelial cells of smoker lungs [241]. While these studies revealed important mechanistic consequences of tobacco exposure for systemic immune cells and upper airway epithelial cells, little is known about alterations of single cell transcriptomes in the distal lung, with its distinct cellular composition, and how they might affect chronic lung disease initiation and progression.

1.5.4 LADC in women

In the year 2000, it was estimated that more than 50% of global lung cancer cases in women occurred in non-smokers, compared to only 15% in men [242]. While these numbers indicate a pronounced gender bias, there has been substantial controversy about whether female non-smokers are truly at a higher risk of developing lung cancer compared to their male counterparts, or whether the disparity is rather due to other factors like social behaviour or the longer life-span of women [243-245]. Besides this ongoing debate, there is good evidence for differences in disease mechanisms and therapeutic outcomes between women and men. In particular, younger patients, i.e. those under 50 years of age, are more often female, indicating a gender-associated risk for this patient group [246-251]. Smoking seems to have gender-specific effects on DNA integrity, as differences in smoking-induced DNA adducts have been identified in women and attributed to different metabolic processing of chemicals contained in tobacco smoke [252, 253]. Hormonal differences between men and women have been discussed as candidate drivers of this asymmetry. In LADC, estrogen receptor beta (ERβ) was found to be expressed more often in non-smoking patients, and in a higher proportion of women compared to men [254]. Expression of this receptor might directly impact therapeutic strategies, since ERβ has been shown to interact with EGFR, the LADC driver mutation most frequent in never-smokers (section 1.5.3), by facilitating the release of EGFR ligands from cells after ERβ stimulation [255, 256]. A lot of research has therefore focused on the role of ERβ in LADC development, but a clear mechanistic link to hormone levels or hormonal replacement therapy remains elusive as other confounding factors and the difficulty of defining discrete patient subgroups have presented major obstacles [257-262].

1.6 Aim of this study

Healthy tissues comprise a multitude of specialised cells that can each occupy various defined functional states to fulfil distinct tasks in a highly concerted manner.

The transcriptome of individual cells mediates the translation of genetic and epigenetic information into phenotypic traits and communication of cells with the environment. It therefore serves as a proxy from which cellular identity, as well as the susceptibility of cells to adverse environmental events such as viral infection or tobacco smoke exposure, may be inferred. Transcriptomic disruptions also accompany diseased cell states and neoplastic transformations. Moreover, in the same way that healthy tissues rely on the concerted interaction of different cell types, tumours rely on the interplay of heterogenous populations of neoplastic cells and their interactions with the environment.

During the last decade, a variety of new technologies have been developed to study cell identities and changes in the transcriptome at single cell resolution. Harnessing these technologies to answer biological questions first requires an evaluation of their suitability for different cases of application.

This thesis therefore aims to provide a comparison of four of the first commercially available single cell transcriptomics technologies and discusses their suitability for different biological applications. Using heterogeneous well-characterised tissues, mouse brain as well as testis, I demonstrate the successful implementation of each technology and highlight their advantages and disadvantages. I also compare single cell transcriptomes acquired from freshly dissected tissue to single nuclei transcriptomes obtained from fresh frozen tissue, further broadening the scope of applications to include biobanked samples.

In the second part of this thesis, I apply single cell transcriptomics to create a reference atlas of the human lung and study the susceptibility of different cell types to coronavirus infection based on the expression of mediators of viral infection. I further explore changes in cell type composition or the transcriptome after tobacco smoke exposure and in the presence of lung adenocarcinoma, as well as intratumoural heterogeneity in lung adenocarcinoma. As there is evidence for increased susceptibility to lung adenocarcinoma in young female never smokers compared to male never smokers, I particularly focus on this demographic. My results determine functional heterogeneity at the single cell level with translational relevance.

2 Results

The work presented in this thesis initially started with the aim of identifying and implementing appropriate single cell transcriptomics technologies to investigate cellular heterogeneity in a variety of mammalian species across a range of tissues, including brain and testis. To establish a suitable technology in the laboratory, I investigated approaches for tissue dissociation, and also obtained single nuclei from frozen samples to enable the handling of rare samples after storage or transport. Single cell technologies rapidly evolved and permitted to increase the numbers of analysed cells, resulting in a more faithful representation of cell type diversity in a given tissue. I therefore tested four different technologies as they became available (C1, iCell8, Dolomite-scRNASeq, 10x-Chromium; Figure 2.1). I successfully implemented each technology in our laboratory, enabling comparisons across tissues and species. Within the scope of my dissertation, I then applied this expertise to focus on single cell transcriptomics studies of the human lung.

Using surgical biopsy samples obtained from LADC patients, we explored cell type diversity in human lung tissue to create a cell atlas of the healthy human lung. This atlas was harnessed to identify potential mediators of SARS-CoV-2 infection, as well as changes in cell type composition and transcriptional profiles in response to tobacco smoke exposure. In addition to healthy lung tissue, we applied scRNA-seq to investigate LADC tumour samples, with a particular focus on comparing their developmental architecture and transcriptional profiles in smokers and never smokers.

In this chapter, I will first describe the implementation of the different experimental single cell RNA sequencing approaches and compare their outputs (section 2.1). I will then present a census of healthy lung cell types (section 2.2.1), and report on candidate effectors of SARS-CoV-2 infection (section 2.2.2) as well as smoking-induced inflammation (section 2.2.3). The chapter will conclude with a comprehensive investigation of malignant cell type heterogeneity and the tumour microenvironment in LADC (section 2.3). Figures and text in sections 2.2 and 2.3 are partly adapted from associated publications [139, 263]. I am joint first author of both publications and wrote the original manuscript for the latter.

2.1 Evaluation of different single cell transcriptomics technologies

Modern scRNA-seq technologies enable the simultaneous processing of thousands of cells from diverse tissues. When the work for this thesis started, the field of single cell genomics was still in its early stages, and studies at the time were limited to a few hundred cells at relatively high cost. In addition, protocols for the preparation and dissociation of tissues into healthy cells suitable for single cell applications had only been developed for a small number of tissues, including easily accessible and dissociable ones like the hematopoietic system or tissues with wide availability of reference data such as embryonic mouse brain [264, 265].

The first commercially available scRNA-seq solution was the C1-96 system (Fluidigm) [32], a circuit-based microfluidic system with a throughput of at most 96 cells at a time. Recent updates to the C1 system have increased the throughput to 10,000 cells, but were not available at the time this work was conducted. Its application is highly labour intensive and extremely expensive, at a cost of about 70 € per cell. As the microfluidic cell capture sites only allow for a narrow range of cell diameters to be processed, it also restricts experiments to the simultaneous analysis of cells that are similar in size. Using the C1 system, we processed live cells isolated from mouse embryonic forebrain tissue and identified characteristic gene expression profiles for different cell populations known to be present in this tissue (section 2.1.3). These successful results demonstrated the applicability of the C1 system; however, the emergence of novel (commercial) solutions enabling a higher throughput at a lower cost led us to apply new technologies to achieve a more comprehensive representation of single cells transcriptomic profiles in complex tissues.

These newer technologies included the iCell8 system (Takara Bio) [36], based on a flat chip containing 5,184 nanowells into which cells are distributed by limiting dilution. By keeping the reaction volume for reverse transcription of RNA in the nanolitre range, it ensures higher gene capture efficiency compared to other well based methods (see sections 1.2 and 4.2.3.2). The iCell8 system permits the simultaneous assessment of about 1,300 single cells and includes the possibility to image the chip before library construction, so that damaged cells or cell doublets can be excluded from further processing at an early experimental stage.

Droplet microfluidic systems offer an even higher throughput by encapsulating cells in oil immersed water droplets for cell lysis and mRNA capture. Among those, the scRNA-seq system from Dolomite-Bio as well as the Chromium Controller from 10x Genomics were tested in the context of this thesis. Similarly to the iCell8 system, they also limit the reaction volume to nanolitres (see sections 1.2, 4.2.3.3 and 4.2.3.4). However, reverse transcription in the Dolomite system takes place in millilitre reactions of pooled cells, whereas in the Chromium protocol RNA capture as well as reverse transcription happen inside the droplets, theoretically increasing sensitivity for lowly abundant genes. Unlike the previous systems, both the Dolomite-Bio and the Chromium system do not offer any imaging capability and therefore require rigorous quality control steps after sequencing to identify and exclude damaged cells and cell doublets.

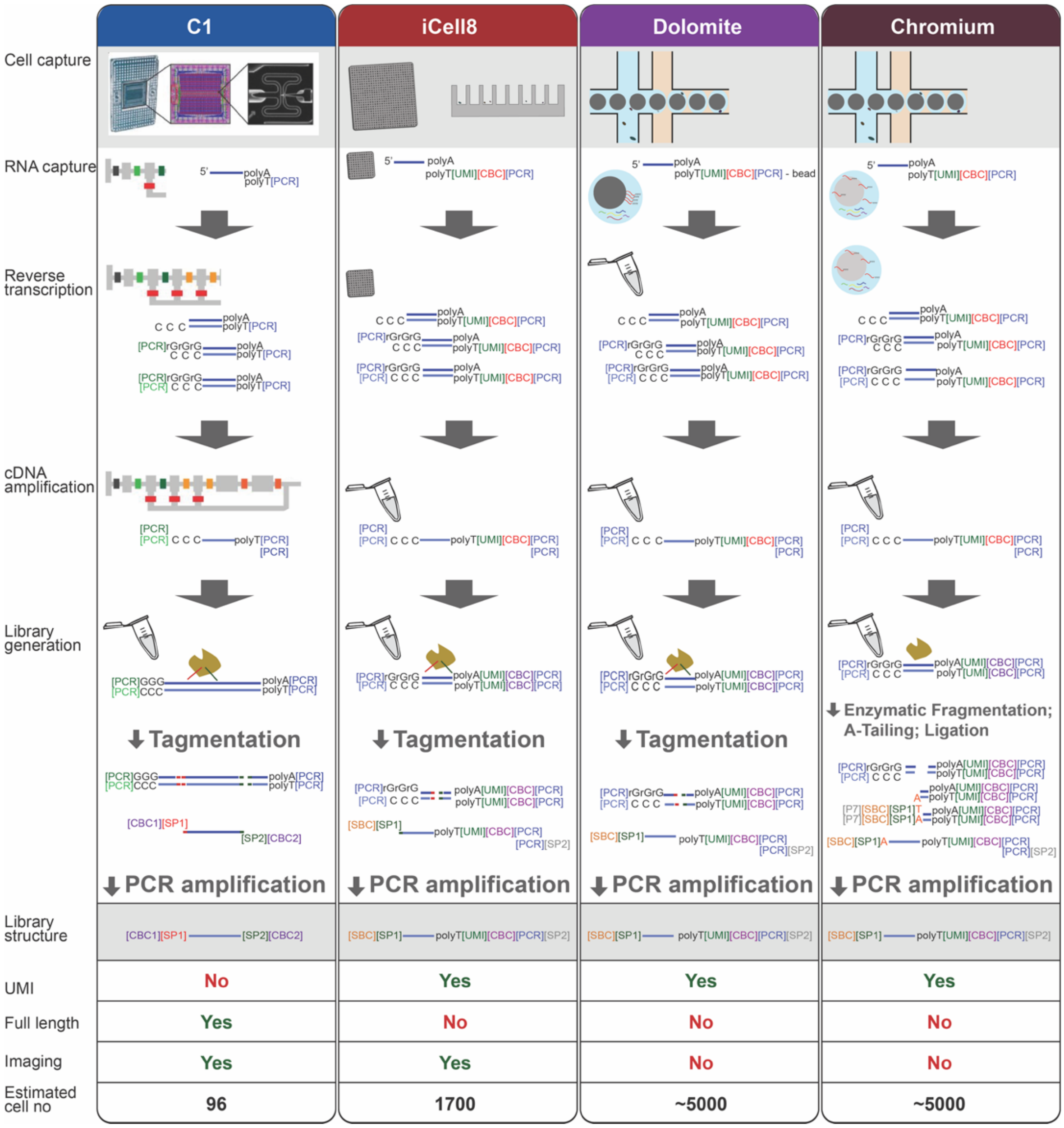


Figure 2.1 Technology overview. Overview of the four scRNA-seq technologies discussed in this thesis, including the respective cell and mRNA capturing technology as well as library generation methods. The C1 system uses consecutive microfluidic chambers to capture cells and conduct reverse transcription and cDNA amplification in nanolitre volumes separately for each cell. Libraries are generated in microliter reaction volumes for each cell using tagmentation, achieving full length transcript libraries. In the iCell8 system, limiting cell dilutions are dispensed across a chip containing nanolitre-sized wells, with single-cell capture achieved stochastically. Reverse transcription is conducted in nanolitre volumes for each cell separately inside the chip. In the Dolomite system, cell and mRNA capture are completed inside nanolitre droplets generated by microfluidics. In both the iCell8 and the Dolomite system, cDNA amplification and library generation are performed on pooled samples using one-sided tagmentation to enrich for 3' end fragments. The Chromium system uses nanolitre droplets for cell and mRNA capture as well as reverse transcription. cDNA amplification and library generation are performed on pooled samples by enzymatic fragmentation, A-tailing (addition of a non-template adenine), barcoded sequence ligation and 3' enriched PCR amplification. All techniques use PCR to amplify cDNA as well as the complete sequencing library to increase material available for NGS. A detailed description of the library generation process can be found in section 4.2.3. PCR: PCR-primer; SP: sequencing primer; CBC: cell barcode; SBC: cell barcode; UMI: unique molecular identifier. (images of C1 workflow are adapted from [32])

2.1.1 Sequencing of single cells and single nuclei

The first and one of the most crucial steps in scRNA-seq is isolating intact single cells. Cell connections and extracellular matrix components vary between different organs and species. Methods to isolate single cells from a tissue therefore need to be adapted to minimise any bias towards certain cell types or alterations to the transcriptome. I first developed and adapted methods to dissociate different tissues from several species, including *Mus musculus* (section 4.2.1) and *Anolis carolinensis* [266]. These dissociation methods, however, might not be most suitable for morphologically complex cell types such as neurons, since they require the disruption of all cell contacts which could damage the cell and significantly affect the transcriptome [24, 267]. Further, these methods are not applicable to frozen material, as the cytoplasmic membrane becomes porous when frozen and intact living cells are thus difficult to recover [267, 268].

A promising approach for these more challenging samples is single nucleus RNA sequencing (snRNA-seq), where intact nuclei are isolated and subsequently processed using analogous workflows to scRNA-seq. The nucleus harbours different RNA species compared to the cytoplasm, with a higher percentage of unspliced, early response and short turnover RNA [269-271]. Nevertheless, it has been shown to provide a faithful representation of the cell's transcriptome, enabling cell type identification and detection of rare subpopulations of cells [267, 268, 272-274]. To be able to process samples procured from human donors, for example precious biopsies stored in biobanks, I adapted snRNA-seq to study frozen samples. The most widely used method for this purpose today, using mild detergent and mechanical force, proved suitable for isolating single nuclei from healthy tissue (section 4.2.2). However, solid tumour tissue such as LADC required a harsher protocol employing a citric acid buffer, which enabled consistent isolation of single nuclei with fewer nuclei doublets or clumps compared to other protocols using detergents (Figure 2.2). Originally developed to prevent RNA degradation in ribonuclease-rich tissues such as the pancreas, it dramatically decreased undesirable extranuclear debris in tumour tissue samples (section 4.2.2).

Having established protocols to isolate intact single cells from fresh mouse forebrain and testis tissue as well as nuclei from fresh frozen samples of mouse forebrain, testis, human lung and lung cancer, I proceeded to compare different sequencing library generation protocols.

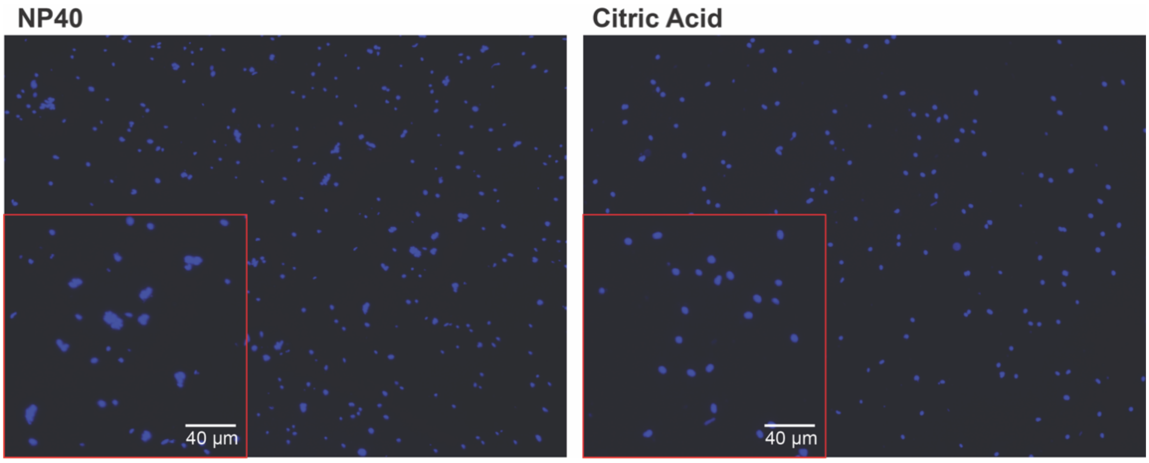


Figure 2.2 Comparison of nuclei isolation protocols. Intact nuclei were isolated by mechanical force either in the presence of the detergent NP40 (left) or citric acid (right). The latter protocol resulted in fewer nuclei doublets or clumps.

2.1.2 Library preparation, data pre-processing and quality control

The experimental implementation and comparison of scRNA-seq approaches focused on mouse testis and forebrain, representing two tissues with contrasting physiological roles and biological properties. All mouse samples were taken from SWISS mice (section 4.1.2) at embryonic (E13.5), juvenile (4 weeks) or adult (9 weeks) developmental stages. Testis or forebrain was dissected and immediately processed for single cell analysis or fresh frozen in liquid nitrogen. Single cells were isolated from fresh tissue by a combination of enzymatic digestion and mechanical dissociation, depending on the sample (section 4.2.1). From fresh frozen tissue, intact nuclei were isolated by chemical and mechanical disruption of the cytoplasmic membrane and washing of the extracted nuclei to reduce ambient cytoplasmic RNA.

Viability and integrity of cells and nuclei, respectively, was assessed by trypan blue staining.

Single cell transcriptomics libraries were created according to the technology specific protocols and sequenced on a MiSeq, NextSeq500 or HiSeq4000 next generation sequencer from Illumina (for more detail see section 4.2.3).

Single cell RNA sequencing data from each technology was first assessed for its quality by commonly used methods, namely cDNA fragment size analysis during library preparation as well as phred base quality score [275] and relative nucleotide contents after sequencing (Supplementary Figure 1 A-D). Data were discarded if predefined quality standards were not met (compare section 4.3.2). Reads were assigned to individual cells using the barcode added to each cell's scRNA-seq-library during preparation, and the resulting gene expression data was filtered to exclude cells with low read numbers identified as lying below the inflection point of cumulatively summed reads (e.g. Supplementary Figure 1 E). Following the alignment of the sequencing reads to the appropriate genome (section 4.3.1), libraries with low genome mapping ratio were also excluded. Successful alignment results in a count matrix of detected genes by cells. The count matrix was further filtered for cells with a minimum number of genes detected ($>$ 200), as entries below this threshold most likely represent partial cells or ambient RNA captured from the cell suspension. Likely cell doublets were excluded by discarding cells with an exceptionally high number of genes or RNA molecules detected, depending on the experiment (e.g. Supplementary Figure 2; Table 1; section 4.3.2).

After quality control, general metrics for each experiment (Table 2) including number of detected RNA molecules and genes per cell were assessed (Figure 2.3 A,B). The number of cells obtained from each experiment reflected the expected differences between technologies (Figure 2.1), with an average of 49 cells for C1, 237 cells for iCell8 (using a quarter of the chip capacity), 1,841 cells for Dolomite and 1,274 cells for Chromium experiments.

To compare the number of genes detected per cell across technologies, differences in read depth must be accounted for. Given the observed variety in read numbers, counts in all data sets were downsampled to equal numbers in each cell (section 4.3.3). This revealed that the adjusted average number of genes detected per cell is comparable for single cells and single nuclei of the forebrain using either the C1 or iCell8 system (Figure 2.3 C). On the contrary, in the case of testis, a significantly higher number of genes were detected from single cells compared to single nuclei, irrespective of the technology used (iCell8, Dolomite or Chromium), with the iCell8 system having a higher gene detection rate compared to the other technologies. These results indicate a strong difference in sensitivity depending on the adopted technology and the interrogated cell type, as we also found in a subsequent collaborative multi-centre study focused on benchmarking different scRNA-seq technologies to which I contributed the expertise gained from these explorative experiments [276].

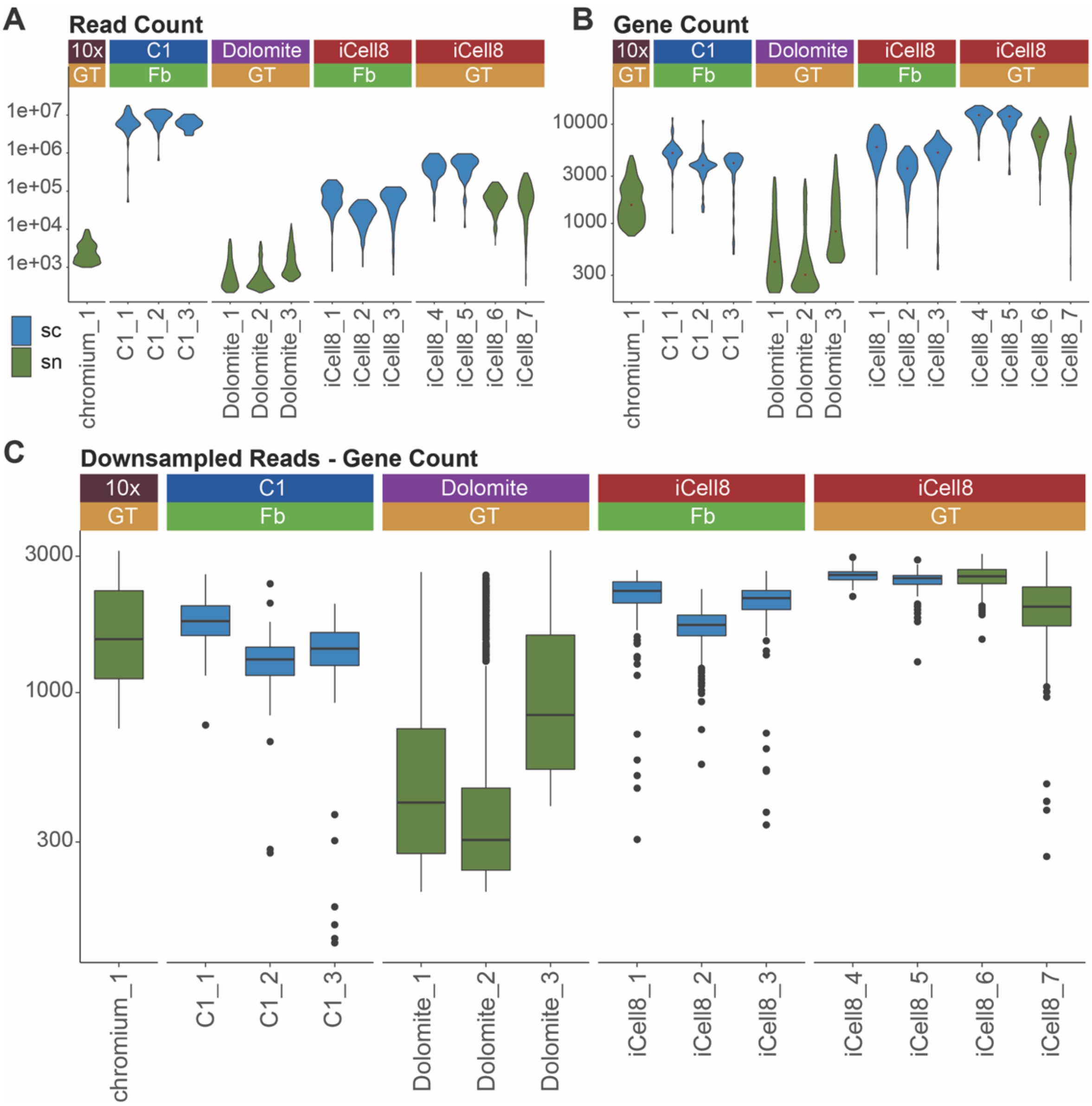


Figure 2.3 General scRNA-seq metrics. scRNA-seq libraries were generated using different technologies (Chromium 10x, Dolomite, iCell8, C1) from single cells (sc) of fresh tissue or single nuclei (sn) of fresh frozen tissue, either from mouse forebrain (Fb) or testis (GT). Each library was sequenced and aligned to the appropriate reference genome in order to relate reads to their corresponding genes. Violin plots show **(A)** read counts per cell and **(B)** the number of detected genes per cell. **(C)** To compare the number of detected genes across technologies, reads for each cell were downsampled to equal numbers for each cell. Box plots show the number of detected genes per cell after downsampling. An overview of all experiments can be found in Table 2.

2.1.3 Comparison of technologies using mouse forebrain single cells

As the first commercially available technologies for single cell transcriptomics, we compared the C1 and iCell8 systems by processing mouse embryonic forebrain, a tissue that had already been well characterised by scRNA-seq approaches in previous studies [265, 277]. Following dissection of the forebrain (section 4.1.2), single cells were isolated by enzymatic digestion (section 4.2.1) and subjected to the microfluidics-based workflow implemented in the C1 system or dispensed into nanowells using the iCell8 system (section 4.2.3). After sequencing and quality control, computational methods implemented in the R package ‘Seurat’ were employed to identify different cell types based on transcript count matrices (section 4.3.3). To this end, cells were clustered by first calculating the k-nearest neighbours (knn) and then constructing a shared nearest neighbour graph. Clusters were visualised using the Uniform Manifold Approximation and Projection (UMAP) algorithm for dimensionality reduction (Figure 2.4 A,C). Cell types were assigned using previously published reference data [277] by calculating a similarity score of single cell transcriptomes to the reference (compare section 4.3.7) and confirmed by gene set enrichment analysis of differentially expressed genes (Figure 2.4 B,D; section 4.3.9).

Using the iCell8 system, we could identify interneurons, neuronal progenitor, radial glia and endothelial cells as well as cells originating from the cortex or the choroid plexus (Figure 2.4 C). On the contrary, the cell clusters identified using the C1 system did not show unambiguous agreement with known cell types (Figure 2.4 A). Nevertheless, differential gene expression analyses among the different clusters revealed the specific expression of genes involved in cell proliferation, neuronal morphogenesis and oxidative processes, indicative of cellular functions expected in the developing mouse brain (Figure 2.4 B,D).

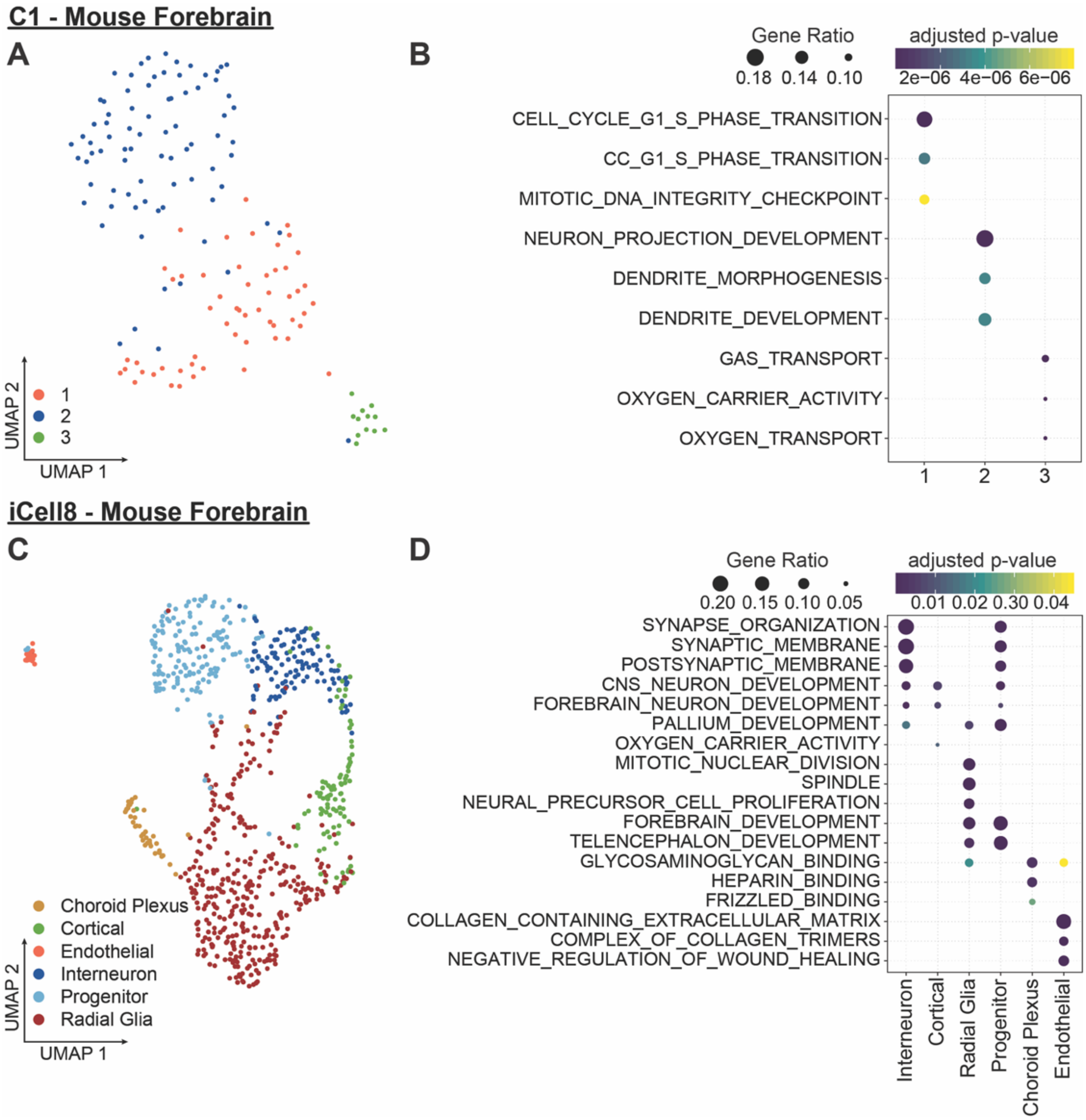


Figure 2.4 Cell identities in the mouse forebrain. Single cells from fresh mouse forebrain tissue (embryonic day E13.5) were used to generate scRNA-seq libraries. The resulting single cell transcriptomes were clustered by knn clustering. By comparison to reference data sets, cell identities were established for the iCell8 data, while the cell clusters identified in the C1 data did not show unambiguous agreement with known cell types. **(A,C)** UMAP visualisation of single cell transcriptomic data generated with **(A)** the C1 system or **(C)** the iCell8 system. Colours indicate inferred cell identities for the iCell8 data and cell clusters for the C1 data. **(B,D)** Gene set enrichment analysis of differentially expressed genes per cell identity or cluster corresponding to the UMAP visualisation shows gene ontology (GO) terms enriched in different cell populations. Dot sizes indicate the ratio of member genes present in the gene set that were detected in each cell population. Colours represent p-values (hypergeometric test after Benjamini-Hochberg correction).

These results demonstrate the applicability of both the C1 and iCell8 systems for explorative studies of tissue development, but also highlight the need for sufficient cell numbers to identify different cell types and states. While unambiguous cell type annotation could not be achieved based on the sparse data from the C1 system (147 cells; 3 experiments), cell numbers in the range of those obtained with the iCell8 system (768 cells; 3 experiments at quarter capacity) already enable significant observations in the mouse brain.

Therefore, it became apparent from an early time point that studying complex tissues would require transcriptomic data for several hundreds to thousands of cells. As the C1-96 system permits processing of only a very limited number of cells, the cost of using this system would thus have been prohibitively high [278], therefore we excluded this system from subsequent analyses. Instead, we focused on comparisons between the other three technologies (iCell8, Dolomite and Chromium).

2.1.4 Comparison of technologies using single nuclei from mouse testis

To evaluate the capability of the iCell8, Dolomite, and Chromium systems to generate data that allows the identification of distinct cell types, we compared their performance using mouse testis, which represents a very heterogenous tissue comprising cell types with a known developmental progression during spermatogenesis and other specialised cells [279]. To isolate single nuclei, samples from fresh frozen tissue were processed by mechanical dissociation in the presence of a detergent (section 4.2.2) and single cells from fresh tissue were isolated as detailed in section 4.2.1. After sequencing, we followed the same approach for data processing, dimensional reduction and visualisation as described above for mouse forebrain (section 2.1.3) to determine transcriptionally different cell clusters. Cell types were then assigned by differential expression of marker genes based on the literature (Figure 2.5).

In data obtained from the different scRNA-seq technologies, cell types were not represented in equal proportions (Figure 2.5). This discrepancy could be due to a divergent sampling bias between technologies or it might reflect the difference in the number of sampled cells (Figure 2.3), with lower cell numbers insufficient to capture the full complexity of the tissue. For example, the size of microfluidic channels might prevent the processing of cells with a certain size or shape, while different cell densities may lead to sedimentation bias during pipetting steps. While all of the major testis cell types were detected with the Chromium system, including the main stages of spermatogenesis, Sertoli cells were not detected in the Dolomite data and Leydig and spermatogonia in the iCell8 data (Figure 2.5).

To further probe differences between scRNA-seq technologies in capturing transcriptional profiles of single cells, correlation coefficients of cell type specific gene expression patterns were calculated based on the average expression of highly expressed marker genes from the downsampled expression matrices (sections 4.3.4 and 4.3.5). While it was possible to assign cell types manually using known marker genes, the overall gene expression levels detected in these experiments were highly dependent on the technology used (Figure 2.6 A). To correct for this bias, we integrated data from the different systems using dimensionality reduction by canonical correlation (CCA) and mutual nearest neighbour (MNN) analysis (section 4.3.6). Following data integration, we observed a high degree of correlation between technologies for the gene expression profiles of the different cell types. This result demonstrates not only that the integration method employed here is capable of reducing technology-related noise which would otherwise obscure biological signals in the data, but also that representative transcriptional signatures for each cell type were extracted through our approach independent of library technology (Figure 2.6 B).

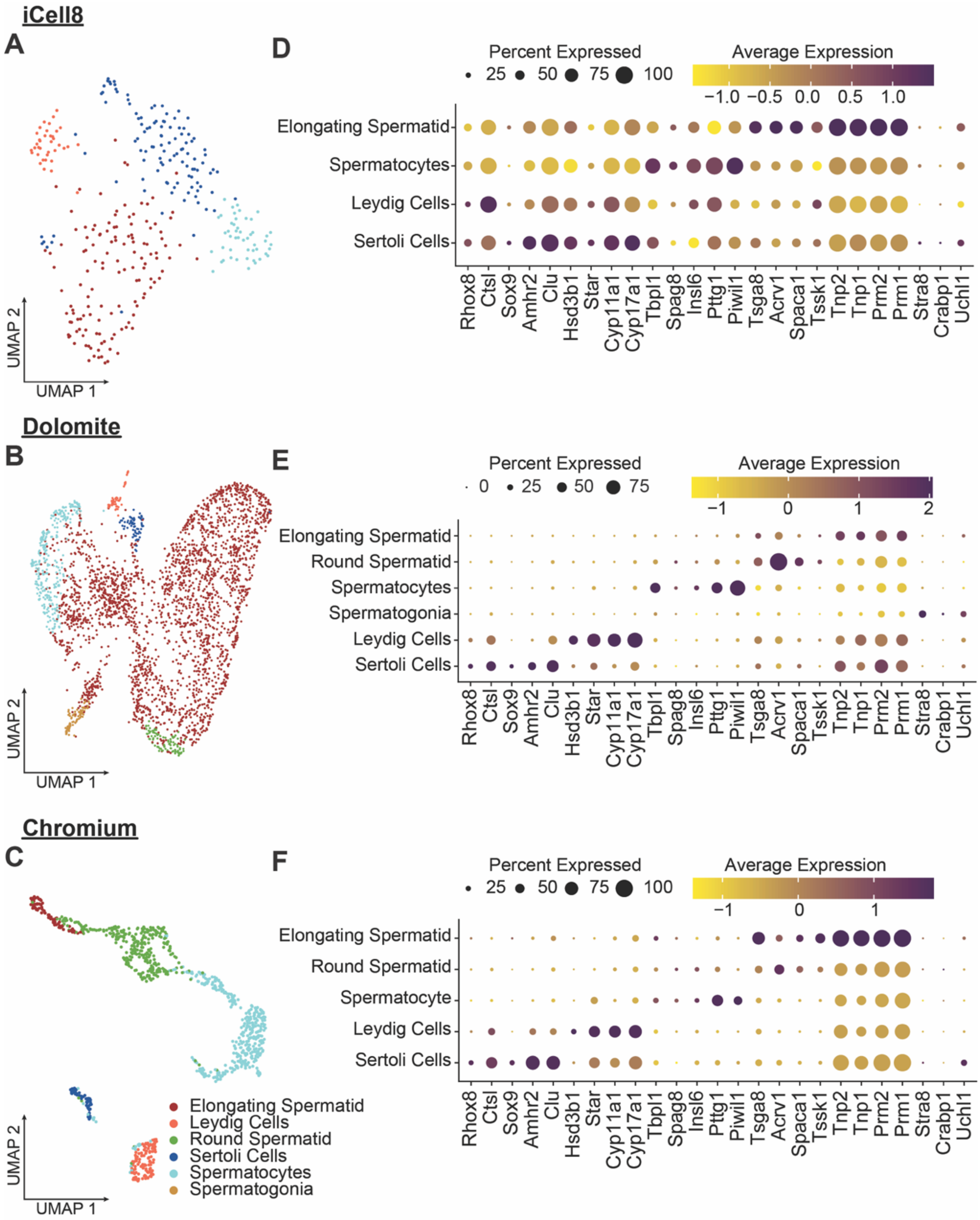


Figure 2.5 Cell types of mouse testis. Single cell data from one scRNA-seq experiment of fresh frozen mouse testis for each technology (Chromium, Dolomite, iCell8) was clustered by knn clustering and each cluster was assessed for expression of known cell type marker genes. **(A-C)** Inferred cellular identities for each cluster are shown on UMAP representations, indicated by colour. **(D-F)** The scaled average expression of known cell type marker genes is visualised across all clusters. Dot sizes represent the proportion of cells in each cluster where expression was detected.

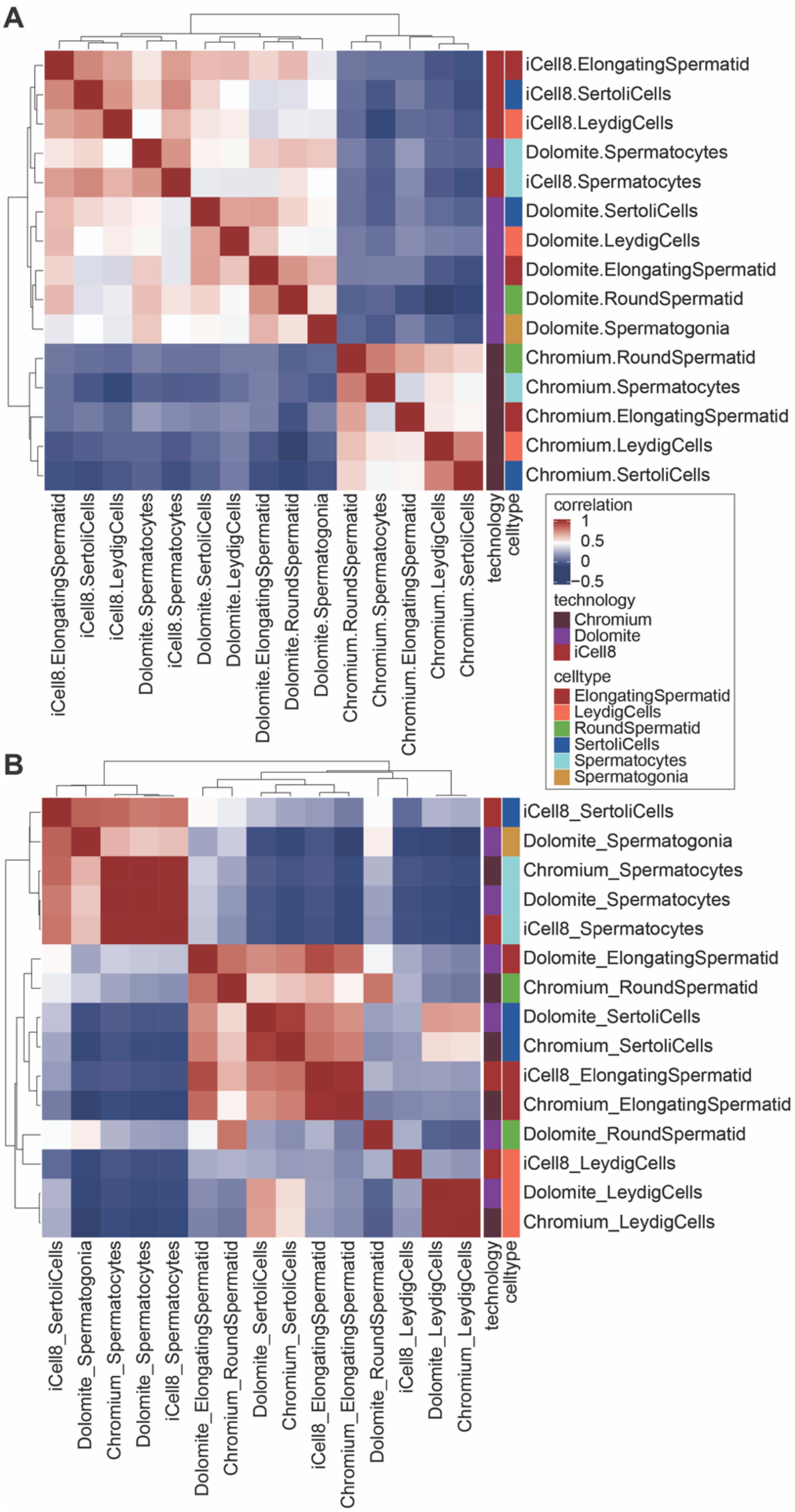


Figure 2.6 Correlation of cell type transcriptomes. Hierarchical clustering of Spearman correlation values representing the association between gene expression averaged across all cells of a given cell type processed with each technology. Correlation coefficients were calculated either based on **(A)** raw or **(B)** computationally integrated count matrices.

2.2 Cell type diversity in the healthy human lung

The comparison of scRNA-seq technologies described in the previous section showed that each technology was successful at resolving cell type diversity, while differences between technologies mostly concerned the number of cells that could be processed in one experiment, the associated cost, and the ease of implementation. Due to its superior throughput and robustness, the Chromium system has thus emerged as the technology of choice for many scRNA-seq applications, and I employed this system for the investigation of single cell heterogeneity in healthy lung and LADC presented below.

To study healthy human lung cells as well as LADC in patients with or without a smoking history, fresh frozen surgical lung tissue samples from patients with LADC were retrospectively obtained from the Lung Biobank Heidelberg and subjected to single nucleus RNA sequencing. In addition to tumour tissue, additional samples from normal lung tissue distant from the tumour had been obtained for a subset of patients during surgery. Samples originated from four patient groups: eight female smokers between 40 and 60 years of age, eight female never smokers and three male never smokers from the same age group, as well as seven elderly female never smokers between 75 and 90 years of age (Table 3 and Table 4). This diversity of sample origins allowed us to a analyse the composition and characteristic features of healthy lung and LADC taking into account age, gender and smoking history as potential determinants.

2.2.1 Identification of cell types in healthy human lung tissue

To create a reference map of gene expression profiles in untransformed human lung cells, I initially considered only the samples from healthy lung tissue, which comprised a total of 41,061 cells from three individuals for each of the patient groups after quality control as described earlier in section 2.1.2 (Supplementary Figure 2; Table 5; section 4.3.2). Their transcriptome profiles were first integrated using CCA and MNN analysis to eliminate technical variation, as described above (compare section 4.3.6), and clustered based on principal component analysis (PCA) (section 4.3.3 and Supplementary Figure 3 A,B). Cell clusters where then investigated for canonical marker gene expression. In addition, differential gene expression analysis using a Wilcoxon rank test to identify cluster specific genes and gene set enrichment analysis were performed to infer the cell type identity of each cluster (section 4.3.9). In this way, all of the major cell types that have been described for human lung alveoli to date could be identified in our data (Figure 2.7). They comprise the epithelial cells that line the alveoli and distal bronchi (basal cells, alveolar type 1 and 2 cells, ciliated cells, secretory cells, neuroendocrine cells), endothelial cells, smooth muscle cells, fibroblasts and different immune cells (B cells, T cells, dendritic cells, macrophages, alveolar macrophages).

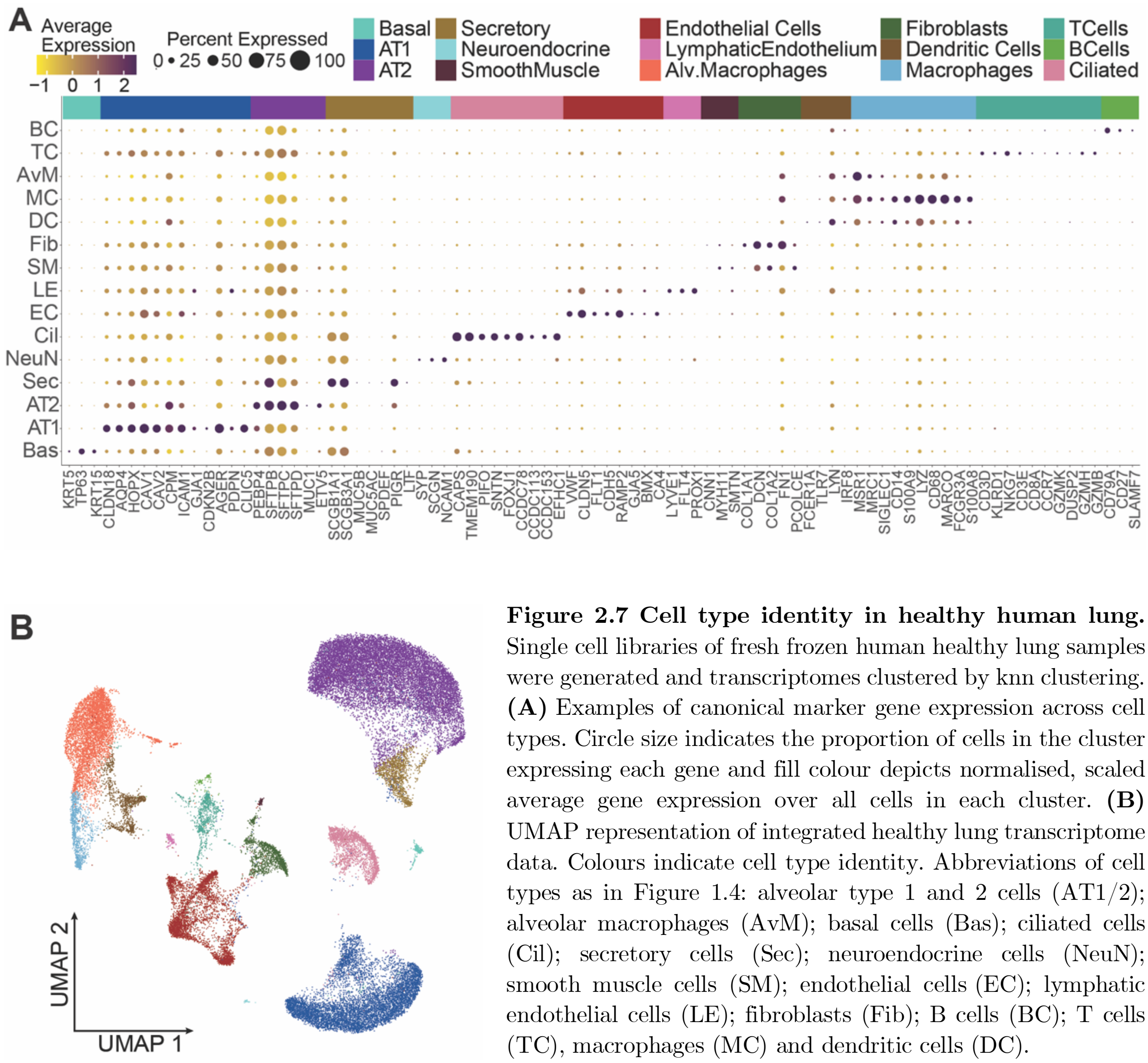


Figure 2.7 Cell type identity in healthy human lung. Single cell libraries of fresh frozen human healthy lung samples were generated and transcriptomes clustered by knn clustering. **(A)** Examples of canonical marker gene expression across cell types. Circle size indicates the proportion of cells in the cluster expressing each gene and fill colour depicts normalised, scaled average gene expression over all cells in each cluster. **(B)** UMAP representation of integrated healthy lung transcriptome data. Colours indicate cell type identity. Abbreviations of cell types as in Figure 1.4: alveolar type 1 and 2 cells (AT1/2); alveolar macrophages (AvM); basal cells (Bas); ciliated cells (Cil); secretory cells (Sec); neuroendocrine cells (NeuN); smooth muscle cells (SM); endothelial cells (EC); lymphatic endothelial cells (LE); fibroblasts (Fib); B cells (BC); T cells (TC), macrophages (MC) and dendritic cells (DC).

2.2.2 Susceptible cell types in the lung for SARS-CoV-2 infection

2.2.2.1 Expression of mediators of SARS-CoV-2 infection

Our single cell transcriptomics atlas of the diverse cell types in healthy lung alveoli provides a valuable resource for investigating changes in response to toxin exposure, neoplastic processes, infections and other challenges to tissue homeostasis. In the early stage of the SARS-CoV-2 pandemic, we harnessed this capability to investigate the expression of genes that are involved in SARS-CoV-2 infection. Specifically, we analysed the expression of the genes encoding ACE2, the receptor to which the virus binds, and TMPRSS2 as well as FURIN, two proteases that have been suggested to activate the viral spike protein and thereby enable virus entry into the cell (see section 1.4).

ACE2 expression was very low across cell types, with under five counts per million (CPM) (Figure 2.9 A; section 4.3.8). Therefore, we aggregated counts per sample and cell type for further analysis. As expected based on the literature, *ACE2* expression was highest in AT2 cells, and a slightly higher number of AT2 cells expressed *ACE2* compared to the other cell types (Figure 2.9 A,D). *TMPRSS2* was expressed at much higher levels overall, with above 150 CPM (Figure 2.9 C), and expression was also biased towards AT2 cells in accordance with previous studies [280]. *FURIN*, which has more recently been suggested as an alternative activator of the spike protein enabling virus entry into the host cell (see section 1.4), was expressed at high levels and with a bias towards AT2 cells as well (Figure 2.9 B). If FURIN, as well as TMPRSS2, can activate the spike protein, our data indicate that this would increase the number of cells susceptible to virus infection by 12%, although susceptible cells continue to represent only a small fraction (around 0.7%) of all alveolar cells (Figure 2.8). FURIN may also be present not only inside the cells expressing it but also in their local neighbourhood [281], which could further potentiate tissue susceptibility to SARS-CoV-2 infection, although intercellular activity of FURIN in lung tissue remains to be validated.

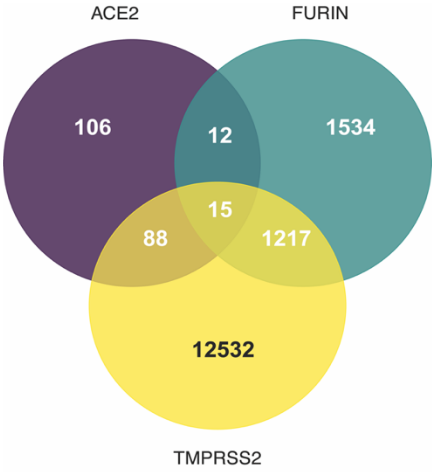


Figure 2.8 Number of cells expressing *ACE2*, *TMPRSS2* and *FURIN*. The Venn diagram indicates the number of cells expressing one or a combination of *ACE2*, *TMPRSS2* and *FURIN* in healthy lung tissue. In total, 25,557 cells were examined.

Overall, these findings suggest that only a small proportion of alveolar cells are susceptible to SARS-CoV-2 infection via expression of ACE2 and TMPRSS2, while FURIN acting as another activator of the viral spike protein after ACE2 receptor binding increases overall susceptibility to SARS-CoV-2 infection in lung alveoli by equipping more cells with proteolytic activity.

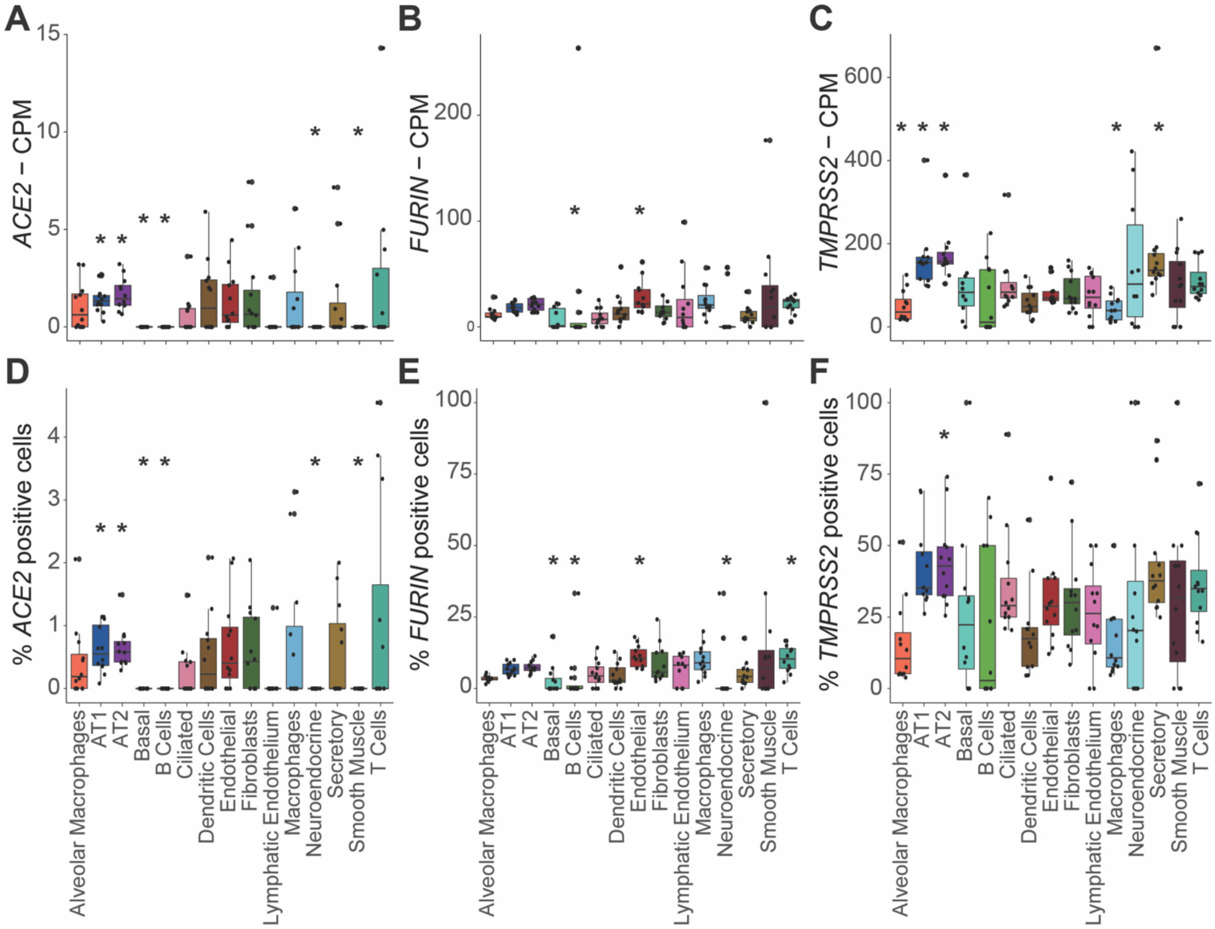


Figure 2.9 Expression levels of *ACE2*, *TMPRSS2* and *FURIN* in healthy lung. (A-C) Read counts were normalised per cell, aggregated for each cell type and CPM values calculated for *ACE2*, *TMPRSS2* and *FURIN* per cell type. **(D-F)** Percentage of cells of each cell type expressing *ACE2*, *TMPRSS2* or *FURIN*. * indicate significant differences between the CPM values or proportion of positive cells, respectively, of one cell type compared to all others (Mann-Whitney test; p-value $<$ 0.01).

2.2.2.2 Correlation of sex, age and smoking history with ACE2 expression

Initial studies of COVID-19 spreading suggested that susceptibility to SARS-CoV-2 infection correlates with age, sex and smoking status [282-285]. We therefore investigated these possible risk factors for COVID-19 in our data (Figure 2.10). We found no correlation of *ACE2* expression with regard to sex, age or smoking habit on the level of individual cell types (Figure 2.10 A-C).

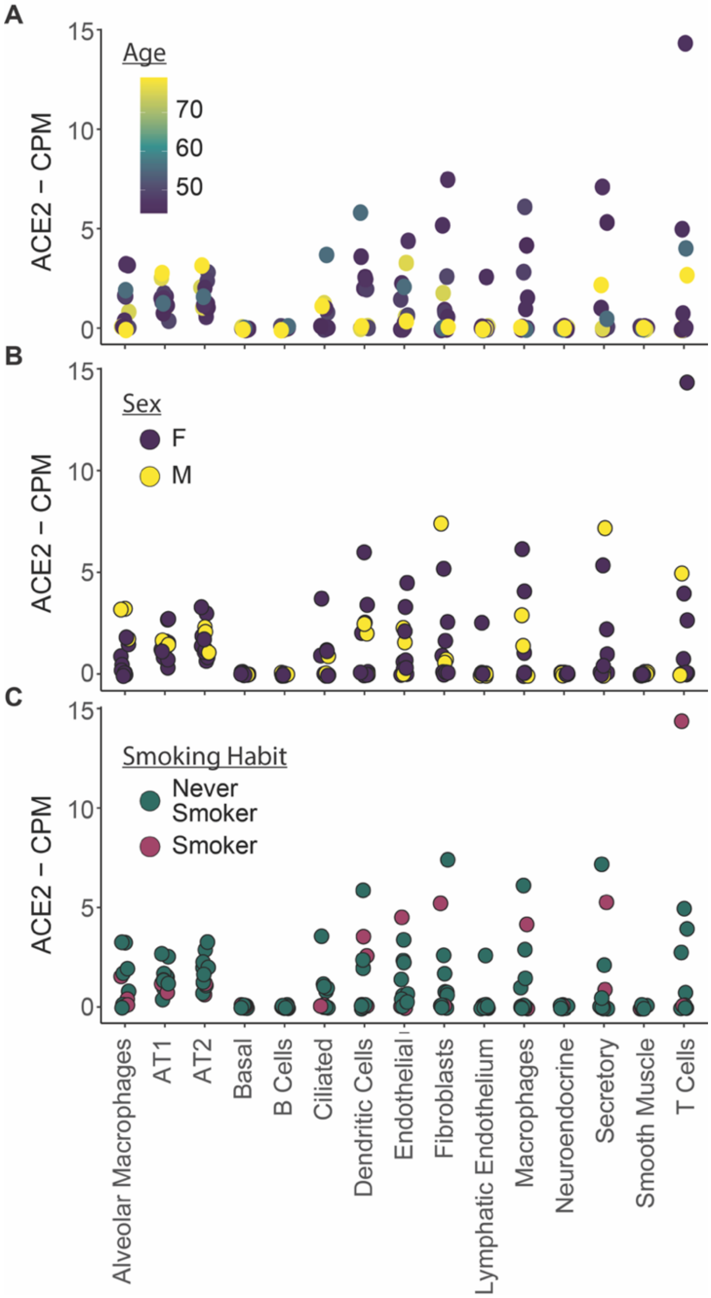


Figure 2.10 *ACE2* expression by age, sex and smoking habit. (A-C) *ACE2* transcript counts were normalised per cell, aggregated for each cell type and CPM values calculated. Shown are CPM values per cell type by **(A)** age, **(B)** sex and **(C)** smoking habit.

However, differences by age were observable when aggregating all reads from individual patients (Figure 2.11); this was likely due to the low overall expression levels of *ACE2* impeding detection in the unaggregated data. In samples from female patients, we identified a trend towards higher *ACE2* expression levels with older age (Figure 2.11 A; R^2 = 0.35; p = 0.09; ANOVA). We were not able to examine this relationship in men because our cohort only included male patients from the younger age group. When comparing three samples from male patients to five samples from female patients within the younger age group, we detected higher *ACE2* expression levels in males (Figure 2.11 B; p = 0.002; two-sided t-test).

While the small size of our cohort and the extremely low detection levels of *ACE2* imply clear limitations to our observations, these results contributed to rapidly increasing our knowledge of SARS-CoV-2 infection mechanisms and provided valuable initial insights for further studies.

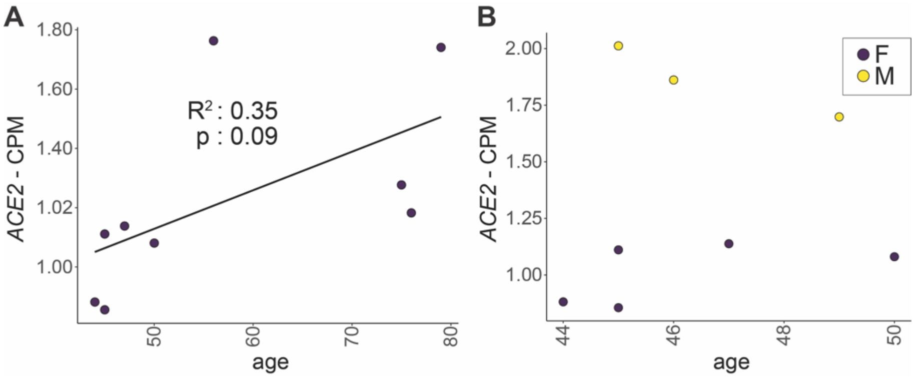


Figure 2.11 Correlation of *ACE2* expression with age and sex. *ACE2* read counts were normalised per cell, aggregated for each sample and CPM values calculated. **(A)** CPM values for *ACE2* as a function of patient age across female samples. A linear model was fitted to estimate R^2 and p-values. **(B)** CPM values for *ACE2* as a function of patient age for samples from young female and male patients (under the age of 55).

2.2.3 Effects of smoking on the healthy lung

2.2.3.1 Increase in alveolar macrophages and AT2 cells in smoker lungs

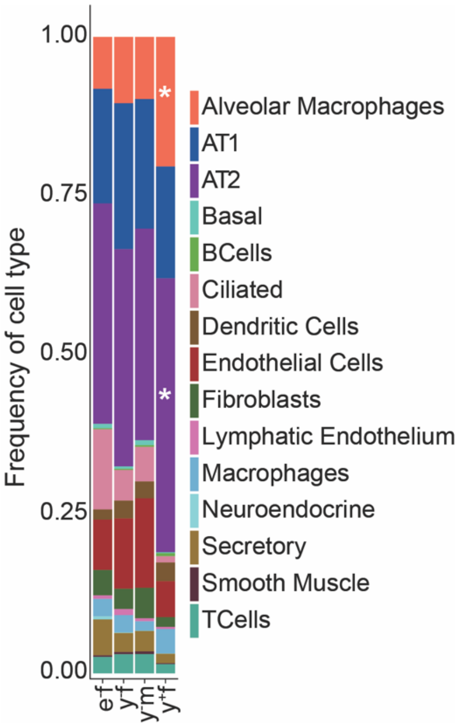


Figure 2.12 Cell type composition of healthy lung tissue in patient groups. Proportion of cells corresponding to each cell type in elderly female never smokers ($e^{-}f$), young female never smokers ($y^{-}f$), young male never smokers ($y^{-}m$) and young female smokers ($y^{+}f$). * denotes cell type proportions that differ significantly from those in young female never smokers based on a Bayesian model of compositional changes.

Exposure to toxins is capable of affecting cell type composition and the transcriptional landscape of lung alveoli. As our single cell atlas of healthy human lung transcriptomes was initially constructed to serve as a reference for the analysis of gene expression in LADC, I proceeded to investigate differences between samples from patients with or without a history of tobacco smoke exposure based on the cell type assignments determined above (section 2.2.1).

While cell type composition was broadly comparable between patient groups, we employed a Bayesian model of compositional changes ([286]; section 4.3.9) to evaluate statistically significant differences and identified fold changes (FC) of cell type frequencies between patient groups. Compared to young female never smokers, young female smoker lung samples showed increased numbers of alveolar macrophages (20.4% in young female smokers compared to 10.4% in never smokers; $\log_2(FC)=1.12$) and AT2 cells (43.1% in young female smokers compared to 34.1% in never smokers; $\log_2(FC)=0.43$) (Figure 2.12). These changes might reflect the adverse influence of smoking on lung cell type composition and increase of immune activity in the tissue.

2.2.3.2 Inflammation in smoking patients

Exposure to tobacco smoke causes damage to the lung and elevates cell death, leading to increased infiltration of leukocytes and activation of cellular repair mechanisms [213, 214]. This inflammatory environment is thought to promote lung cancer development and progression. To investigate the effect of smoking on gene expression for the diverse lung cell types, we focused on healthy tissue samples from female smokers and never smokers between 40 and 60 years of age, thus excluding age and gender as potential confounding factors.

Differential expression and gene ontology analysis revealed an enrichment of gene sets relating to inflammation and activation of immune response in smokers (Figure 2.13 A). Enrichment in gene ontology sets representing response to interleukin 1 was observed especially in alveolar type 2 cells, fibroblasts and endothelial cells. Fibroblasts and endothelial cells also exhibited a significant increase in expression of genes involved in cytokine mediated activity and myeloid leukocyte migration as well as T cell activation. Immune related cells, such as dendritic cells, macrophages and alveolar macrophages, showed higher expression of genes implicated in the response to interferon gamma. Genes involved in these pathways that were upregulated in smokers included *S100A9*, *SLC11A1* and *NFKB*, which contribute to leukocyte activation and migration, as well as *CCL2*, *CSF3* and *IL6*, as general mediators of inflammation (Figure 2.13 B).

To identify cell type interactions mediating this inflammatory response, I evaluated the expression levels of a curated set of ligand receptor pairs across immune cell types (dendritic cells, T cells, macrophages and alveolar macrophages) as well as fibroblasts, endothelial cells and smooth muscle cells of the alveolar region in female smokers and never smokers (section 4.3.13).

The overall number of putative ligand receptor interactions is equivalent in smoking and never-smoking patients (Figure 2.14 A). However, interactions of certain drivers of inflammation are increased in smoker lung samples as inferred from an increase in mean gene expression of ligand receptor pairs with significant specificity to given pairs of cell types (Figure 2.14 B).

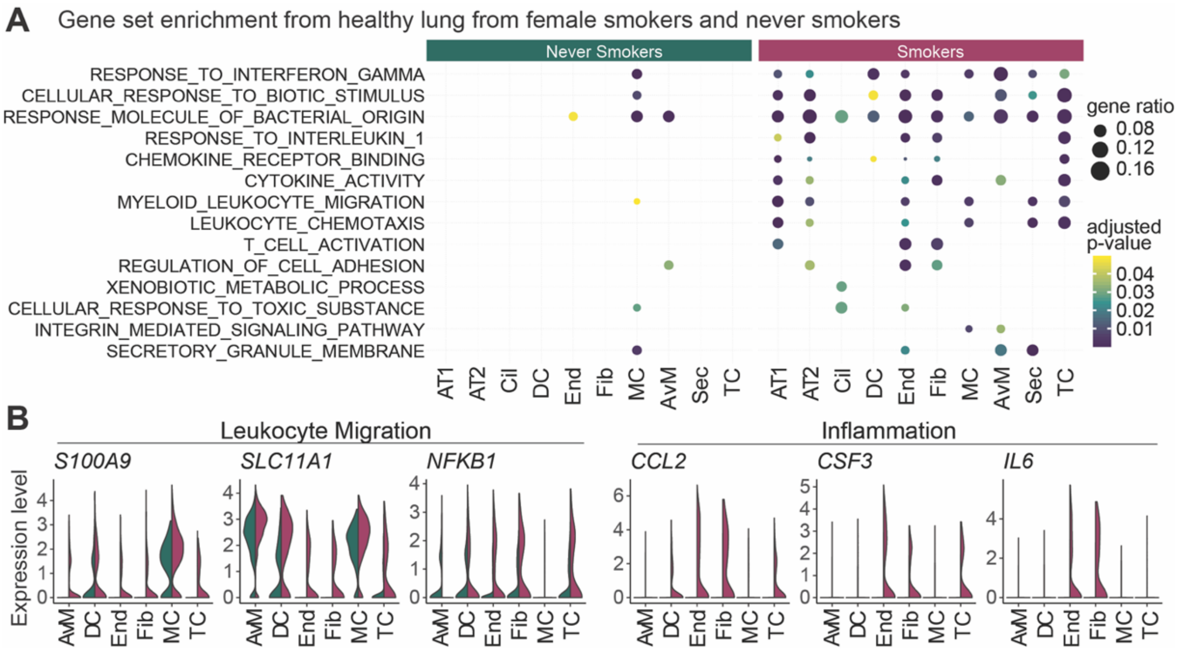


Figure 2.13 Inflammation in smoker lung. (A) Differential gene expression and gene set enrichment analysis between cells derived from never smokers and smokers by cell type. Dot sizes indicate the proportion of genes from each gene set enriched in a given cell type, while colours represent p-values (hypergeometric test after Benjamini-Hochberg correction). Cell types on the x axis as in Figure 1.4. **(B)** Violin plots of normalised single cell expression levels of inflammatory pathway genes across different cell types, split by smoking status (green: never smokers, red: smokers).

For example, inflammatory cytokines IL6 and CSF3 display an increased expression level in endothelial cells and fibroblasts, promoting activation of their corresponding ubiquitous receptors expressed from *IL6R* and *CSF1R* (Figure 2.14 C and Supplementary Figure 4 B). We also identified increased interactions of ICAM1 on endothelial cells, fibroblasts and muscle cells with various integrin complexes on cells of the immune system (Figure 2.14 B).

All patients included in this study were not under current antibiotic or anti-inflammatory medication, except for one young female smoker who received cortison treatment due to an underlying chronic condition. As cortison reduces inflammatory responses, we concluded that this single case was unlikely to bias our results, and at most would lead us to underestimate the increased inflammation seen in smokers.

Thus, gene expression changes in fibroblasts and endothelial cells contribute to an inflammatory environment in normal lung tissue in smoking patients, prompting the question whether these differences translate into different tumour phenotypes according to smoking status.

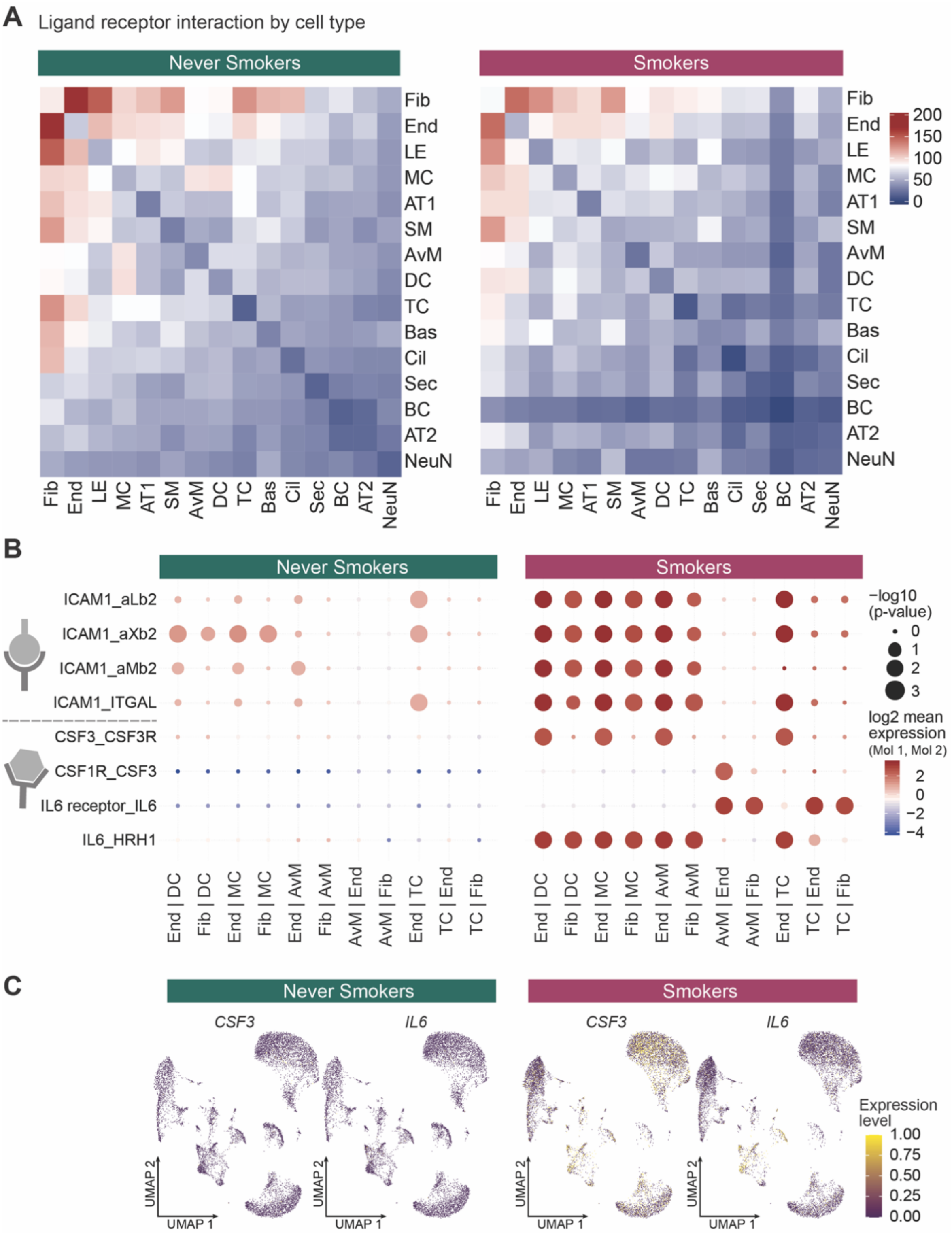


Figure 2.14 Cell interactions in smoker lung. (A) Overall number of inferred cell-to-cell interactions between the different annotated cell types, showing no significant difference by smoking habit. **(B)** Putative ligand-receptor interactions inferred from gene expression data, based on the mean expression of known ligand-receptor pairs in two given cell types. Cell type pairs on the x axis indicate the direction of the interaction (e.g. End | DC: ligand on endothelial cells and receptor on dendritic cells). Dot sizes represent the likelihood of cell-type specificity of a given receptor-ligand interaction, computed based on a random permutation of cell cluster labels as described in section 4.3.13. **(C)** Expression levels of inflammation mediating ligands *IL6* and *CSF3* in individual cells from never smokers and smokers. UMAP representation as in Figure 2.7 B.

2.3 Lung adenocarcinoma in female smokers and never smokers

Following the analysis of human lung cell type diversity and transcriptional changes in the context of SARS-CoV-2 infection and smoking, the following section will present my results on cellular heterogeneity and microenvironment interactions in LADC. Here, in addition to the twelve healthy lung tissue samples used to construct the single cell atlas described in the previous section, we also considered an additional 26 tumour tissue samples from the four patient groups, adding a total of 81,718 cells to the analysis. Tumour samples originated from eight female smokers between 40 and 60 years of age, eight female never smokers and three male never smokers from the same age group, as well as seven elderly female never smokers between 75 and 90 years of age (Table 3 and Table 4).

Importantly, samples obtained from the tumour bulk of lung adenocarcinomas during surgery consist not only of transformed cells, but also contain normal lung cells that shape the microenvironment around the tumour [189, 287]. Separating malignant from non-malignant cells is therefore a key task in single cell cancer analysis to enable the assessment of tumour heterogeneity and changes in non-transformed tissue cell types.

2.3.1 Identification of neoplastic and tumour microenvironment cells

Utilizing the established gene expression patterns of healthy lung tissue (section 2.2.1), we analysed cell type identities in all tumour samples. To this end, samples were integrated as described above (compare section 4.3.6 and Supplementary Figure 3 C-E); cells were then clustered and functionality within the 'Seurat' R package was used to calculate similarity scores that quantify the correspondence between each cell in the tumour samples and the cell types identified in healthy lung samples (section 4.3.7; Supplementary Figure 3 F). Additionally, clusters were manually probed for expression of cell type specific genes.

All cell types identified in healthy tissue were also detected in tumour samples, as verified by their expression of canonical marker genes (Figure 2.15). Cell type composition was comparable in tumours from all patient groups, and no statistically significant difference could be detected employing a Bayesian model of compositional changes (Figure 2.15 B; section 4.3.9). In addition, 37,596 cells derived from the tumour samples showed low similarity scores with known lung cell types (Supplementary Figure 3 F) and did not specifically express any of the used marker genes (Figure 2.15 C), suggesting that they represent neoplastic cells.

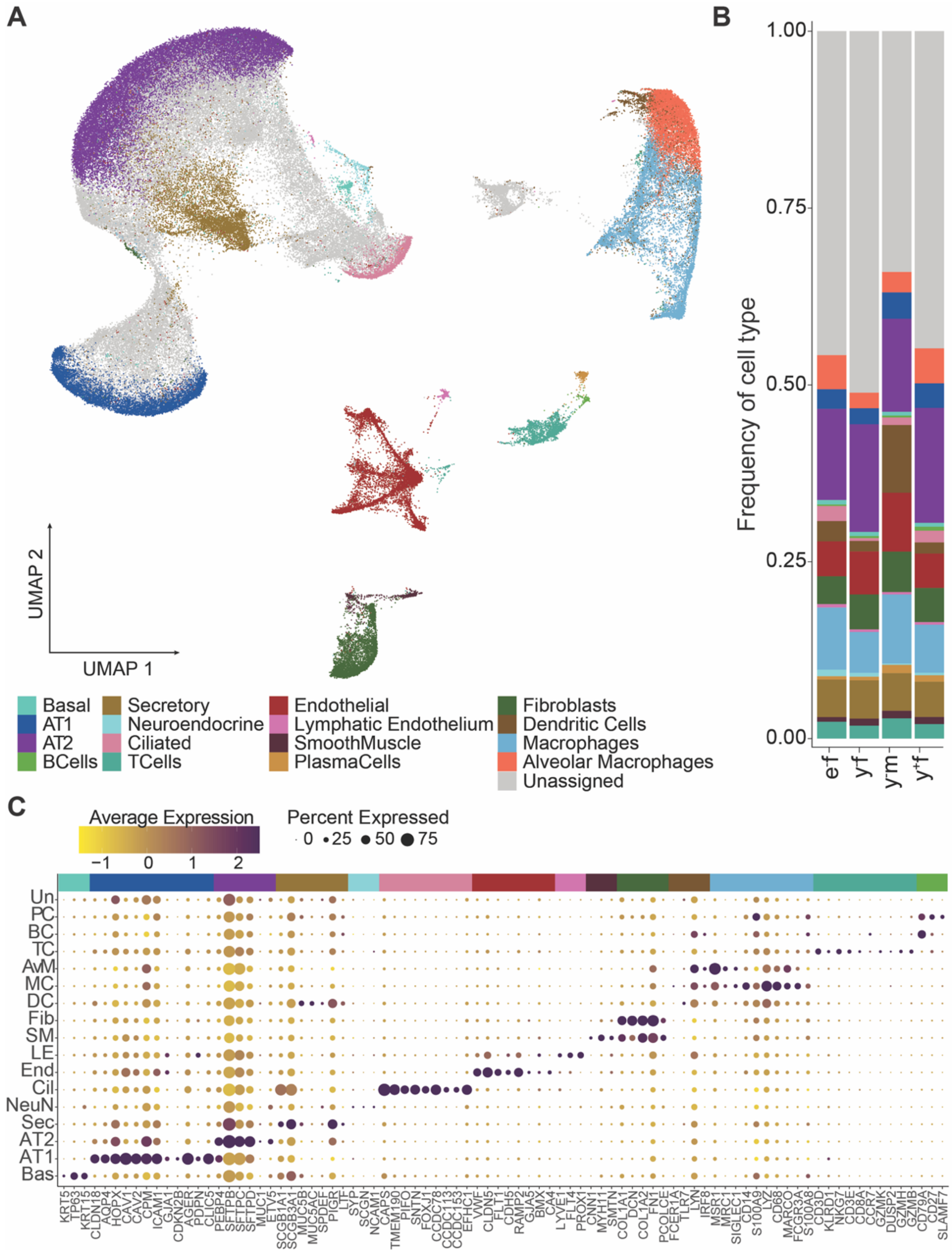


Figure 2.15 Cell type identity in lung tumour samples. Single cell libraries of fresh frozen LADC samples were generated and transcriptomes clustered by knn clustering. **(A)** UMAP representation of integrated transcriptome data from all lung samples, including healthy and tumour tissue. Grey cells represent cells that could not be assigned to any known lung cell type. **(B)** Cell type composition per patient group. Elderly female never smokers (e$^-$f), young female never smokers (y$^-$f), young male never smokers (y$^-$m) and young female smokers (y$^+$f). **(C)** Examples of canonical marker gene expression across the cell types identified. Circle size indicates the proportion of cells in the cluster expressing each gene and fill colour depicts scaled normalised average expression over all cells in each cluster. Colours indicate cell type identity in panels **(A)** and **(B)** and corresponding marker genes in **(C)**.

The deviation from endogenous gene expression signatures in tumour tissue is often caused by mutations or large-scale structural genomic aberrations, such as gains or deletions of chromosomal parts [288, 289]. To corroborate the malignant identity of unassigned cells, we inferred copy number variations (CNV) from transcriptomic data by comparing the average expression levels of genes in close proximity on the genome to a baseline derived from patient-matched normal lung samples (Figure 2.16). Clustering of cells according to their CNV profiles (Figure 2.16 A) revealed two clusters devoid of copy number variations, which included cells from all patients analysed (cluster 4 and 5). These two clusters mostly comprised cells which could be assigned a healthy lung cell type based on the similarity scores calculated previously, and were therefore deduced to represent cells belonging to the tumour microenvironment. In contrast, the remaining clusters harboured distinct losses or gains and were mostly patient-specific. These clusters were also enriched for cells that had not been assigned any healthy lung cell type based on transcriptional similarity scores (Figure 2.16 B).

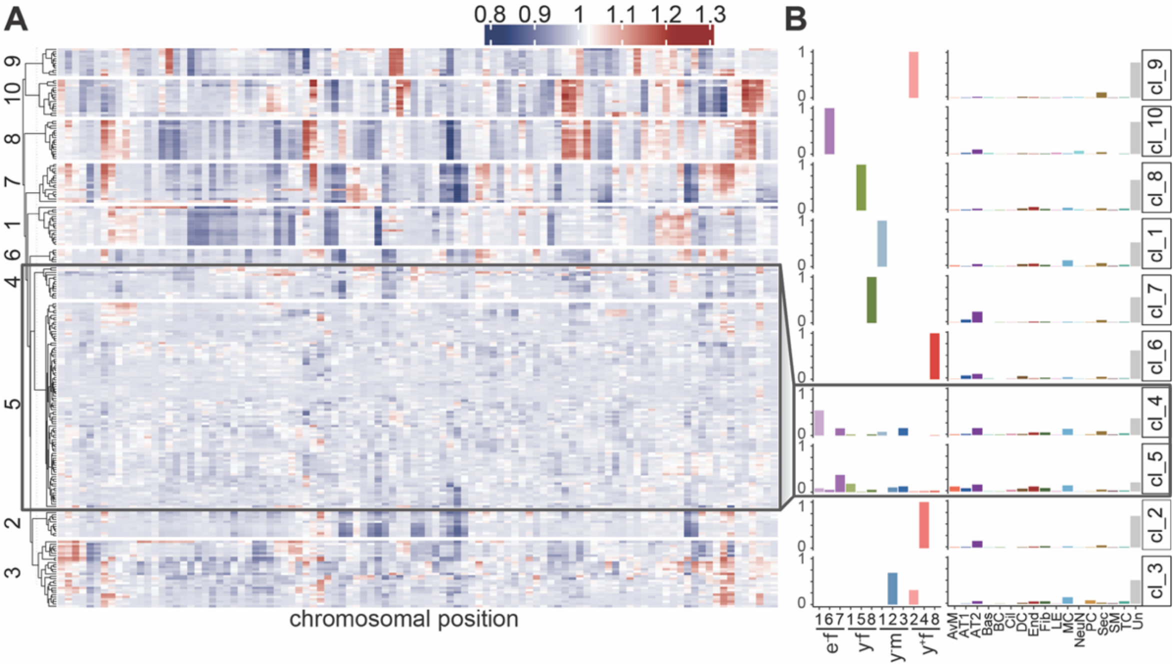


Figure 2.16 CNV profiles of tumour cells. **(A)** Heatmap of inferred CNV profiles from all cells of tumour samples with matched healthy tissue data. Colours indicate residual normalised expression levels (see section 4.3.11). Cell clusters with patient-specific gains (red) or losses (blue) were identified by hierarchical clustering. **(B)** Proportion of cells for each cluster in **(A)** by patient origin (left) or cell type identity (right). Patient groups comprised elderly never smokers female ($e^{-}f$), young never smokers female ($y^{-}f$), young smokers female ($y^{+}f$) and young never smokers male ($y^{-}m$). Cell type labels are defined in Figure 2.15.

We thus inferred that these previously unassigned cells correspond to the malignant tumour cell population. The distinct CNV profiles exhibited by individual patient tumours underline interpatient heterogeneity observable at the level of structural genomic aberrations (Figure 2.16 B).

2.3.2 High intratumoural heterogeneity with distinct cellular subtypes in young female patients

After harnessing our transcriptomics data from all patient groups for cell type assignment as described above, we focused our subsequent analysis on young female smokers and never smokers to compare single cell transcriptomes between patients with and without a smoking history, excluding sex and age as potential confounders. It has been widely demonstrated that solid tumours do not consist of one homogeneous malignant cell population but represent a heterogenous tissue of diverse cellular states [108]. A high degree of interpatient heterogeneity has also been observed in LADC [290, 291].

Clustering of the subset of 37,596 malignant cells from young female LADC patients based on their transcriptomes identified ten distinct cell clusters (Figure 2.17 A). All ten clusters comprised cells from both smokers and never smokers, with some heterogeneity between patients (Figure 2.17 B).

To elucidate the functional relevance of cell clusters, gene set enrichment analysis was performed based on genes differentially upregulated in each cluster. Enriched gene ontology (GO) terms were summarised into meta-signatures reflecting functional processes and used to guide naming of the clusters (Figure 2.18 A,B; Table 6).

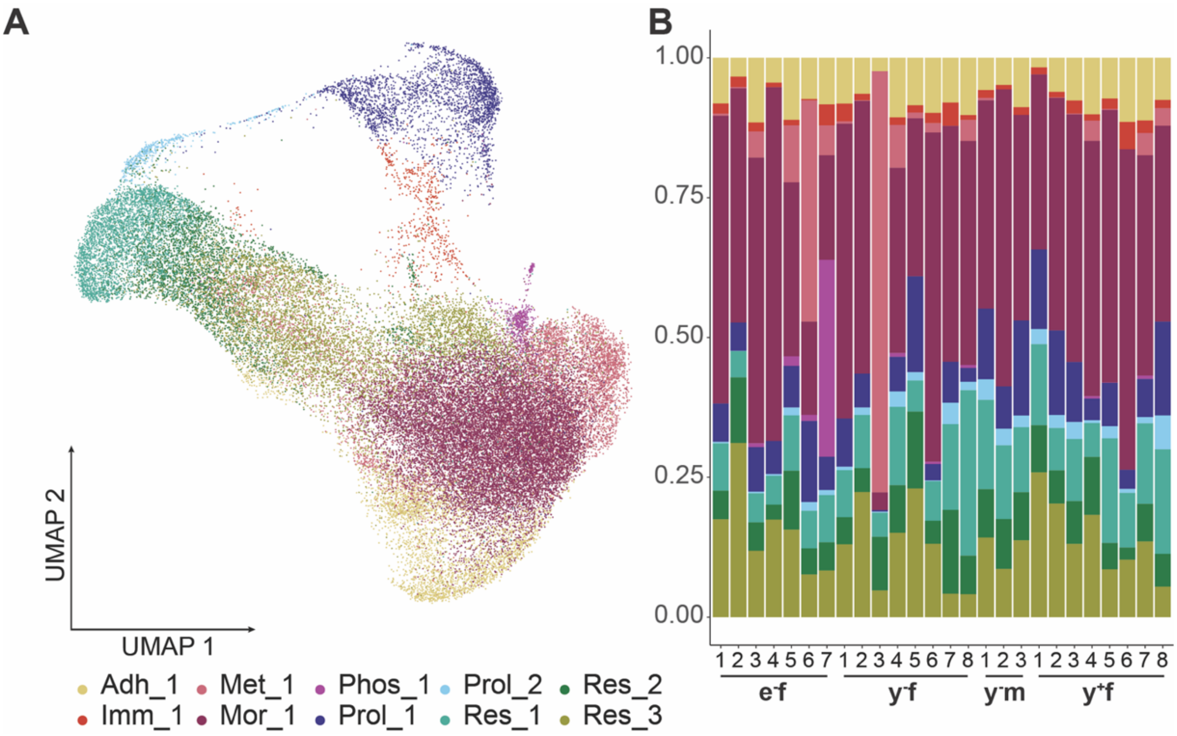


Figure 2.17 Malignant cell heterogeneity. (A) UMAP representation of cell clusters within the malignant cell compartment. Cluster names are defined in the text below. **(B)** Proportions of cells in malignant cell clusters by patient. Elderly female never smokers ($e^{-}f$), young female never smokers ($y^{-}f$), young female smokers ($y^{+}f$) and young male never smokers ($y^{-}m$).

Interestingly, we found that signatures overlap between clusters. While cells from the proliferative cluster labelled 'Prol_1' showed exclusive enrichment in mitosis related GO terms, e.g. MITOTIC_NUCLEAR_DIVISION and CELL_-CYCLE_G2_M_PHASE_TRANSITION, cells from cluster 'Prol_2' in addition expressed genes from GO terms related to cellular respiration, such as ATP_SYNTHESIS_COUPLED_ELECTRON_TRANSPORT and OXIDATIVE_PHOSPHORYLATION. Gene expression of mitotic marker genes, such as *AURKB* and *TOP2A*, was highly elevated in these two clusters compared to all others. Three further clusters showed elevated expression of respiration associated genes ('Res_1-3'). Two of these were also enriched for transcription related GO terms while one cluster, 'Res_3', was exclusively enriched for respiratory terms. While all cells in our data expressed genes related to respiration (e.g. *COX6B1* and *NDUFA4*) and transcription (e.g. *RPL38* and *SPCS1*) to varying degrees, these genes were most highly detected in clusters Prol_2 and Res_1-3 (Figure 2.18 B).

These overlapping signatures might represent different cell states or cell types differentiating from highly proliferating cells to more specialised functions, with the respiratory signature also indicating the high energy demand of tumour development.

Other tumour cells showed no specific expression of cell cycle associated genes, but had distinct signatures with GO terms involved in cell adhesion mediated by integrins ('Adh_1'; CELL_ADHESION_MEDIATED_BY_INTEGRIN) and cell morphology ('Mor_1'; e.g. REGULATION_OF_CELL_MORPHOGENESIS, ACTIN_BASED_CELL_PROJECTION). Associated genes showed distinct expression in the respective clusters, such as the integrin *ITGB8* in cluster 'Adh_1' and the fibroblast growth factor *FGF13* in cluster 'Mor_1'. These cells therefore likely contribute to the spatial architecture and integrity of the tumour tissue and are possibly involved in the initiation of metastatic processes, with cluster 'Mor_1' also being the largest cluster overall (15,884 cells).

One other cluster specifically expressed genes related to metabolism ('Met_1'), with distinct expression of e.g. the insulin mediator *IRS2* and the solute carrier *SLC16A14.* Another cluster was enriched for genes involved in phospholipid binding ('Phos_1'), such as *FCHSD2*, whose product promotes endocytosis of EGFR in cancer cells and thus reduces EGFR signalling [292], and the Growth Factor Receptor Bound Protein 2-Associated Protein 2 (*GAB2*), which also showed a low level of expression across the other clusters.

We further identified an immune related gene expression signature comprising GO terms such as LEUKOCYTE_PROLIFERATION, T_CELL_PROLIFERATION and MHC_CLASS_II_PROTEIN_COMPLEX. This signature was exclusively enriched in one cluster, 'Imm_1', which also showed enrichment of other functional terms including proliferation and morphology. The signature comprised genes such as *SLC11A1*, a divalent ion transporter, which contributes to natural resistance against infections with certain natural parasites and modulates macrophage mediated inflammation [293, 294]. In addition, 'Imm_1' displayed higher expression of *CD86* typically expressed by antigen presenting cells and *IRAK3*, which is part of the Toll-like receptor immune signal transduction pathway and is thought to promote tumour progression by contributing to an inflammatory environment [295]. Sets of genes are usually regulated by different transcription factors associated with distinct cellular functions. We analysed the regulation of gene sets through transcription factors employing a three-step algorithm, which first infers a list of putative target genes for each transcription factor based on coexpression. To avoid false positive target gene assignment, each gene list is filtered for the presence of associated transcription factor binding motifs in the genes' transcription start site. These genes are further ranked by importance based on expression level for each cell separately (section 4.3.12). In this way, we determined gene regulatory networks contributing to the functional heterogeneity observed before (Figure 2.18 C). Proliferating cells (Prol) highly express genes linked to networks regulated by *ATF4*, which is involved in stress responses and amino acid homeostasis [296, 297], and *POU5F1*, also known as *OCT4*, with a critical role in embryonic stem cell self-renewal [298, 299]. Cells enriched for the immune modulating signature (Imm) show additional expression of genes regulated by transcription factors *FOXN3* and *MEF2A*, which are known to be involved in cell cycle checkpoint control and contribute to epithelial-to-mesenchymal transition (EMT) [300, 301].

Together, these results identify eight functional subpopulations of malignant LADC cells in both smokers and never-smokers.

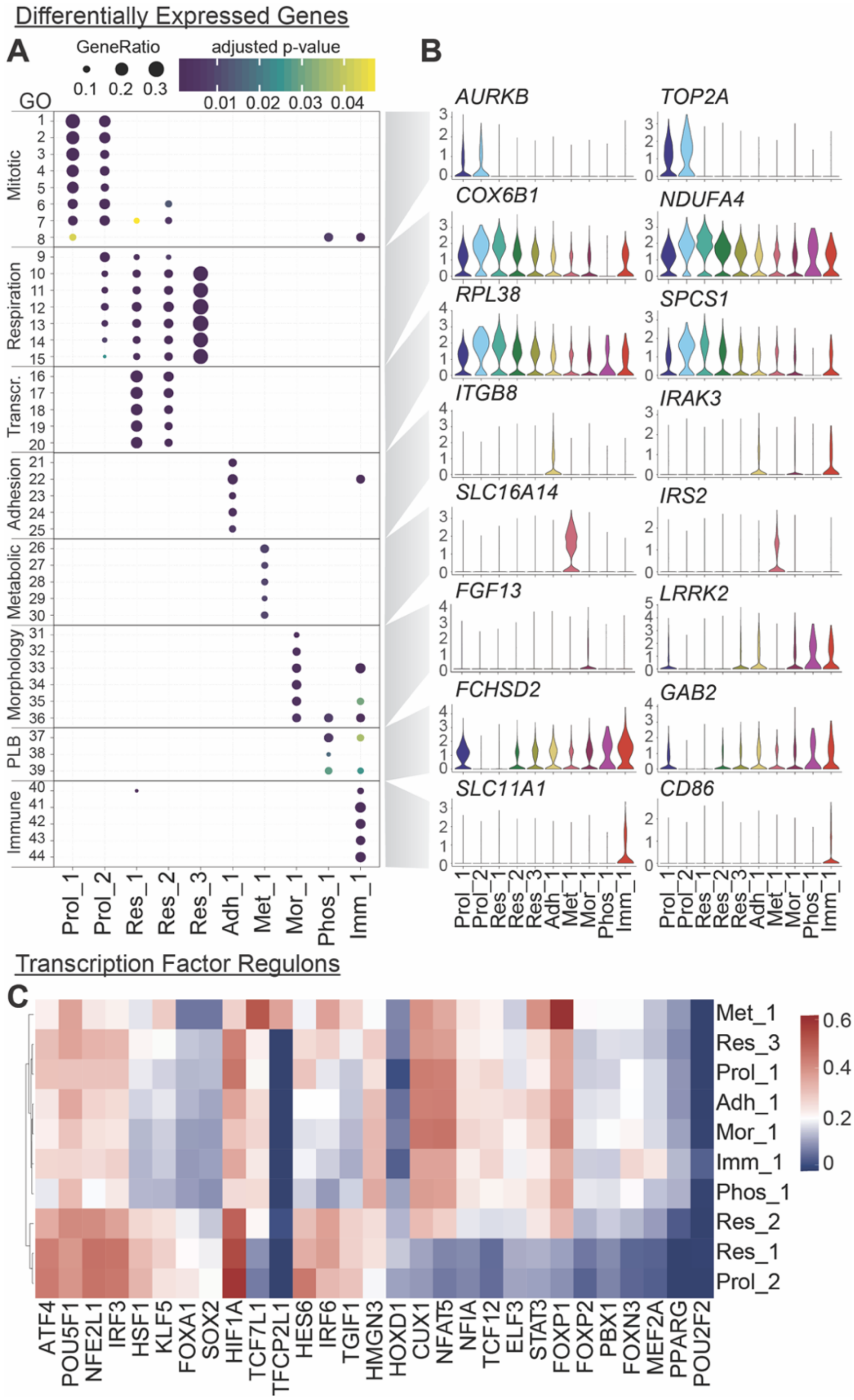


Figure 2.18 Functional heterogeneity of malignant LADC cells. (A) Gene set enrichment analysis of differentially expressed genes in malignant cells of female patients identified 44 cluster-specific GO terms (Table 6) that were combined into 8 functional signatures, named Proliferating (Prol_1/2), Respiration (Res_1/2), Adhesion (Adh_1), Metabolism (Met_1), Morphological (Mor_1), Phospholipid binding (Phos_1) and Immune modulation (Imm_1). Dot sizes indicate the ratio of member genes present in the gene set that were detected in each cell population. Colours represent p-values (hypergeometric test after Benjamini-Hochberg correction). **(B)** Normalised expression of representative genes for each functional signature across malignant cell clusters in female patients. **(C)** Transcription factor network analysis of all malignant cell clusters. Shown are enrichment scores for regulons, which consist of transcription factors and genes associated with a matching transcription factor binding site that are co-expressed in our data (see section 4.3.12).

2.3.3 Trajectory of differentiation and characterisation of malignant cells in the context of smoking history

As tumours are evolving and differentiating tissues [302], we applied a graph-based trajectory inference method (section 4.3.14) to malignant cell transcriptomes from young female never smokers and smokers to discern a differentiation trajectory linking the eight functional malignant cell subpopulations identified above. Pseudotemporal ordering assigned cells to four branches labelled S1-4, with the junction point S0 (Figure 2.19). One branch (S0-S1) consisted of mitotic cells (cluster 'Prol_1') and immune related signatures (cluster 'Imm_1') and was therefore selected as the trajectory origin, with pseudotime subsequently increasing through the junction point S0 towards the most distant points on each of the other branches (Figure 2.19 A,B). Differential expression and gene set enrichment analysis confirmed cell cycle related gene expression in cells on branch S0-S1, in line with previous findings. Branch S0-S4 comprised cells from all identified malignant clusters and was not significantly enriched for specific GO terms; as it was limited to cells at intermediate pseudotimes, some of which were cycling, this branch likely represents undifferentiated tumour cells. Cells on branch S0-S3 consistently expressed genes related to morphology (cluster 'Mor_1'), as well as cell adhesion, substrate binding and wound healing. Branch S0-S2 mainly harboured respiratory cells (clusters 'Res_1' and 'Res_2') with gene expression related to autophagy (Figure 2.19 C). Together with the respiratory signature of these clusters, this reflects the tight connection between oxidative phosphorylation and autophagic processes due to mitochondrial turnover or nutritional need in highly active tissues [303].

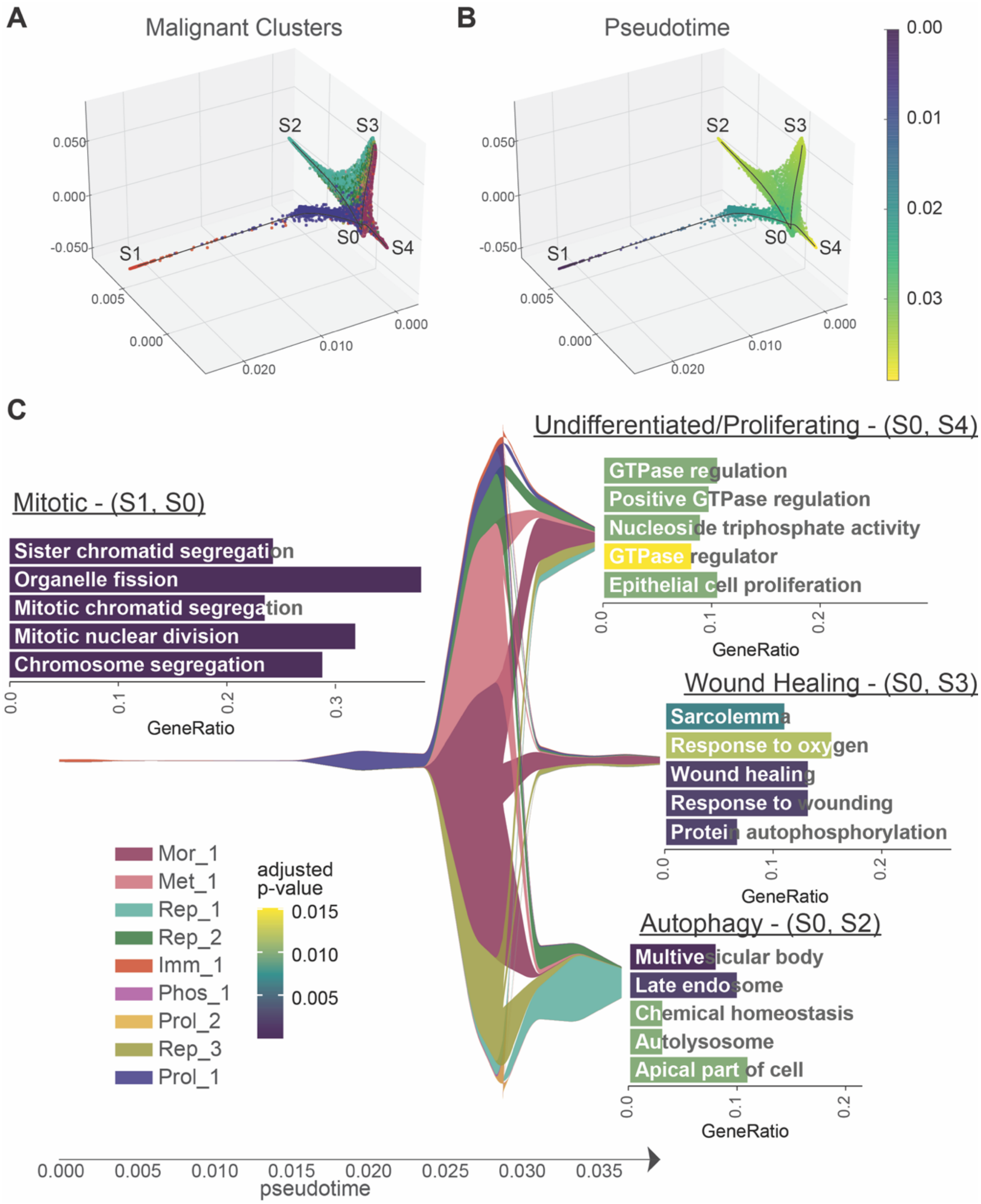


Figure 2.19 Malignant cell trajectory in female patients. (A) Three-dimensional projection of cellular gene expression profiles by Modified Locally Linear Embedding of all identified malignant cells from young female patients to infer a trajectory of differentiation with four branches (S1-4). **(B)** Same projection as in **(A)** with cells coloured by pseudotime. **(C)** Malignant cluster proportions along pseudotime are depicted. Differential expression and gene set enrichment analysis performed for each branch indicate enrichment of proliferative (S1-S0), undifferentiated (S4-S0), autophagy (S2-S0) or wound healing (S3-S0) signatures, as highlighted by the bar plots, with the x axis showing the proportion of gene set members enriched on each branch and the colour representing adjusted p-values (hypergeometric test after Benjamini-Hochberg correction). Full names of GO terms can be found in Table 7.

The developmental trajectory of LADC thus comprises proliferating and intermediate undifferentiated cells as well as two distinct differentiated tumour cell states. Importantly, equivalent trajectories were identified when this analysis was performed separately on malignant cells from young female smokers or never smokers (Figure 2.20), indicating shared functional tumour cell types and a conserved differentiation hierarchy regardless of smoking status.

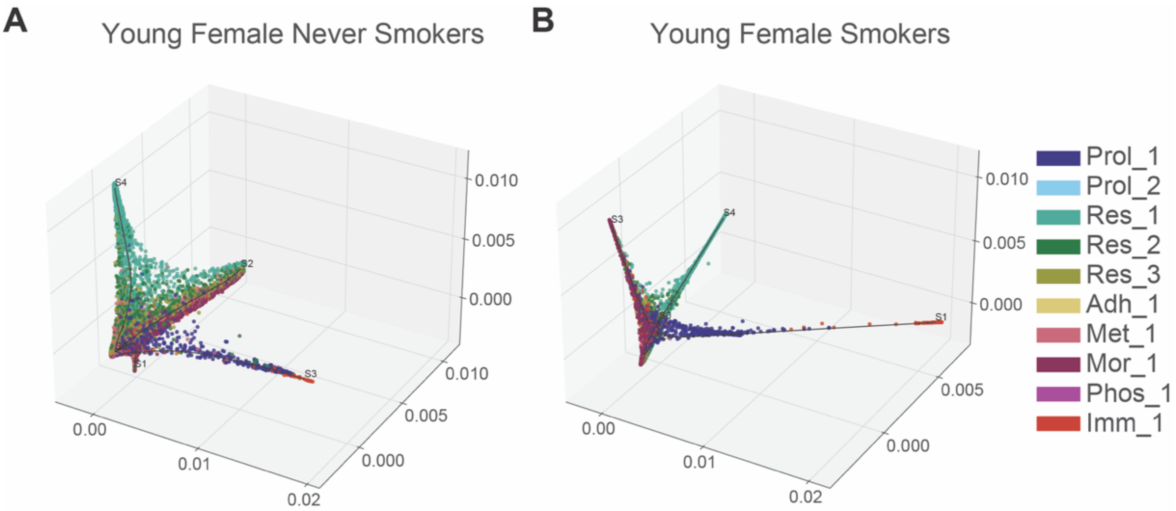


Figure 2.20 Malignant cell trajectory by smoking habit. Three-dimensional projection of cellular gene expression profiles by Modified Locally Linear Embedding of malignant cells from only **(A)** young female never smokers or **(B)** young female smokers.

While LADC from smokers and never smokers in our cohort share the same functional malignant cell types and differentiation trajectory (compare above and Supplementary Figure 5), tobacco smoke exposure might induce more subtle gene expression differences within malignant cell types. Comparing gene expression between smokers and never smokers for each malignant cell cluster separately, I observed that the majority of differentially expressed genes were unique to one or two patients, indicating substantial inter-patient transcriptional heterogeneity in agreement with the previous analyses (Figure 2.16 B and Supplementary Figure 3 C).

We therefore restricted our attention to genes that were differentially expressed in at least half of the female patients of the same smoking habit, and identified consistent gene expression changes across patients only for cluster ‘Imm_1’ (Figure

2.21). Here, gene set enrichment analysis uncovered a difference in immune modulating pathway gene expression, with genes including *ANXA1*, *C1QB* and *PAEP* upregulated in smokers, and genes such as *HLADQA2*, *HLA-DRB5* and *WFDC2* upregulated in never smokers. We also observed differential expression of genes involved in migration, EMT and metabolism, with *MSLN* and *FNDC3B* upregulated in smokers and *AGR3*, *CLDN10*, *IG2FR* and *PCDH7* upregulated in never smokers.

To validate our observations at the transcriptomic level, two exemplary candidate proteins involved in immune modulation pathways were stained in samples from both smokers and never smokers by immunohistochemistry (Figure 2.21 B,C). Representative stainings indicate an increased expression of ANXA1 and glycodelin (PAEP) at the protein level in the majority of female smokers. Quantification of staining intensity revealed a trend for upregulation of both proteins in young female smokers compared to never smokers. Moreover, staining intensity and average gene expression level based on scRNA-seq for each patient correlated for glycodelin, with a trend also observed for ANXA1 (Figure 2.21 D,E).

This divergence implies differential immune modulating capacity of proliferating tumour cells in female never smokers compared to smokers.

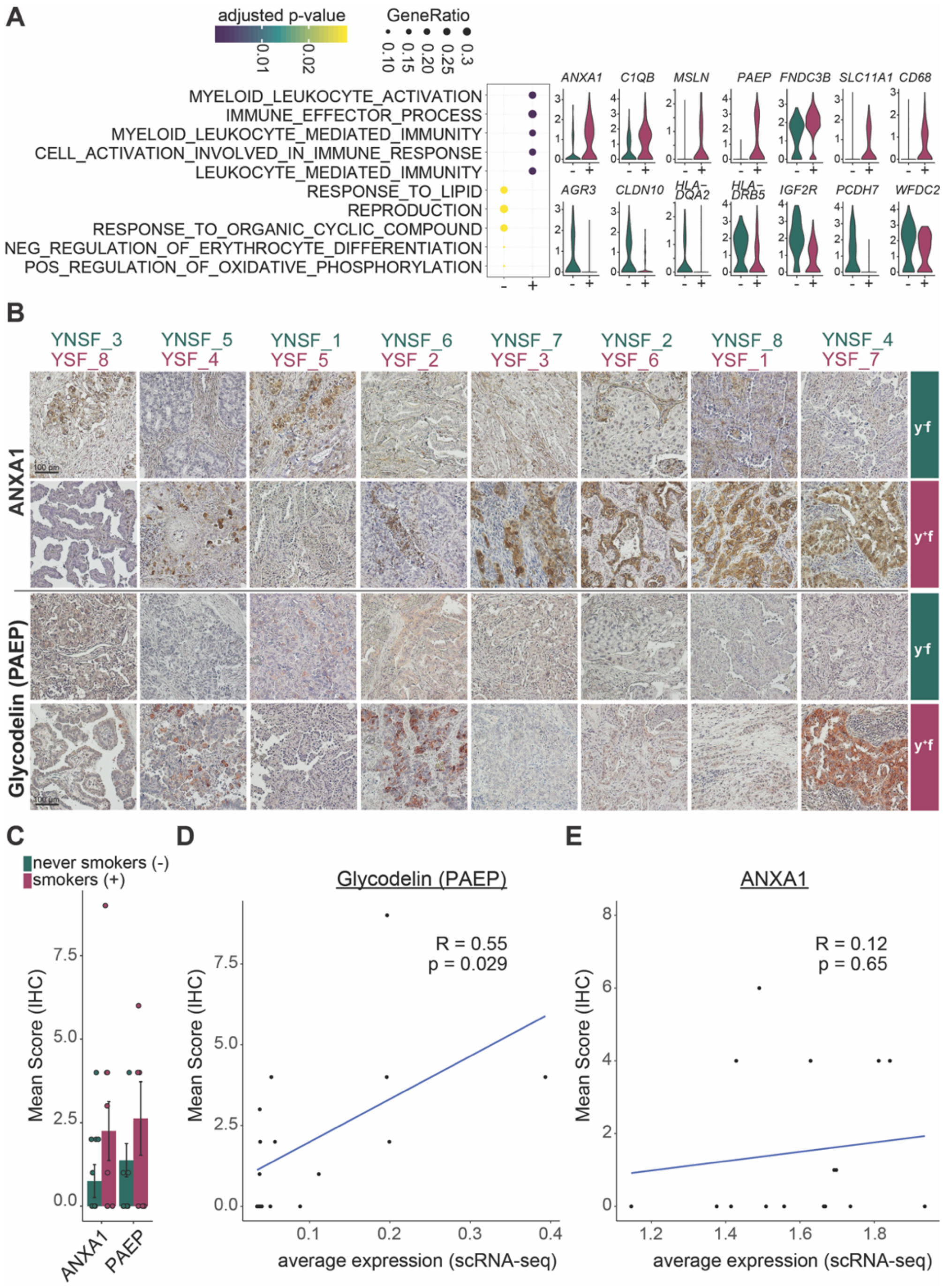


Figure 2.21 Immune modulating cell population in smokers and never smokers. Cells from young female patients in cluster Imm_1 were assessed for gene expression differences by smoking habit. **(A)** Dot plot indicates enriched GO terms in Imm_1 cells from smokers (+) and never smokers (-). Dot sizes indicate the ratio of member genes present in the gene set that were detected in each population. Colours represent p-values (hypergeometric test after Benjamini-Hochberg correction). Violin plots depict representative genes with significantly different expression levels between never smokers (green) and smokers (red). **(B)** Immunohistochemistry (IHC) staining of ANXA1 and glycodelin (PAEP) in tumour cryosections from samples of young female smokers and never smokers. For each patient, one representative staining is shown. **(C)** Quantification of IHC staining. Scoring was performed using five randomly selected tumour sections based on a combination of staining intensities and the number of positive cells; displayed are the mean ± s.e.m. **(D,E)** Correlation of protein expression determined by quantitative scoring of immunohistochemistry staining and average gene expression across all tumour cells for **(D)** Glycodelin (PAEP) and **(E)** ANXA1. Correlation visualised by a linear model and coefficient calculated using Pearson correlation. Immunohistological staining and scoring was performed by Marc A. Schneider.

2.3.4 Deregulation of the tumour microenvironment transcriptome in LADC

Transformed tumour cells are greatly dependent on interactions with their local surroundings, which might either hinder or promote tumour development [201]. Based on the gene expression signatures derived from healthy lung tissue, we already identified diverse cell types of the tumour microenvironment in the LADC tumour samples as described above (section 2.3.1). To delineate transcriptomic states within this compartment that may contribute to tumour progression, I used non-negative matrix factorisation (NMF) to decompose the gene expression matrix for all non-malignant cell types from both tumour and healthy lung tissue samples into the product of two matrices, with the first comprising signatures of co-expressed genes (factors) across all cells and the second capturing the contribution of all genes to these factors. This approach revealed factors that contribute to cell type identity, but also factors that separate cell types into distinct cell states (Figure 2.22).

Two of these factors (factor 5 and 6) represent two cell states within the macrophage population with decreased expression in tumour tissue compared to healthy lung (Figure 2.22 B,C), and contain genes involved in immune cell activation and inflammation (e.g. *PPARG*, *C1QA*, *MARCO*, *GRN* and *SLC11A1*, *MSR1*, *GPCPD1*, *CD68*). Specifically, factor 5 contains genes whose products play a role in macrophage activation such as *PLXDC1*, a receptor of ligand PEDF that enhances tumouricidal activity of macrophages [304, 305], and *SLC11A1*, a divalent transition metal transporter whose activity is associated with pro-inflammatory processes [306]. Downregulation of this signature indicates a reduced activation of macrophages in the presence of LADC. Factor 6 delineates another subpopulation of macrophages with lower expression of inflammatory genes in tumour tissue, including *PPARG* and *MARCO*. The latter has been suggested as a possible treatment target in NSCLC [307], since antibody targeting of MARCO expressing macrophages reduced tumour growth in a recent study [308]. Consistent with our observation that only a subset of macrophages downregulate MARCO, the same study found MARCO expression in only a subset of tumour-associated macrophages. Anti-MARCO antibody treatment was therefore most effective in combination with antibodies against other immune checkpoint markers [308].

Along with macrophages, fibroblasts can also support or hinder tumour development. NMF identified a population of fibroblasts with decreased expression of genes in the *SLIT/ROBO* pathway in tumour samples (factor 2). The *SLIT/ROBO* pathway has often been found to be differentially regulated in cancer, where its complex involvement in tumour progression may include beneficial as well as detrimental effects on tumour growth [309]. The decreased expression of *SLIT2* observed in tumour-associated fibroblasts here could facilitate tumour survival and progression [309], while *SLIT3* downregulation might enhance EMT [310]. Another population of fibroblasts showed increased expression of type I and type III collagens in neoplastic tissue (factor 10). As part of the tumour microenvironment, different extracellular matrix components provided by fibroblasts have been found to affect tumour behaviour [311]. Increased expression of type I and type III collagens, as observed here, is thought to promote invasion and metastasis in lung cancer [312-314].

Our results thus resolve different macrophage and fibroblast subpopulations in LADC, with distinct gene expression signatures contributing to a tumourigenic environment in both smokers and never-smokers.

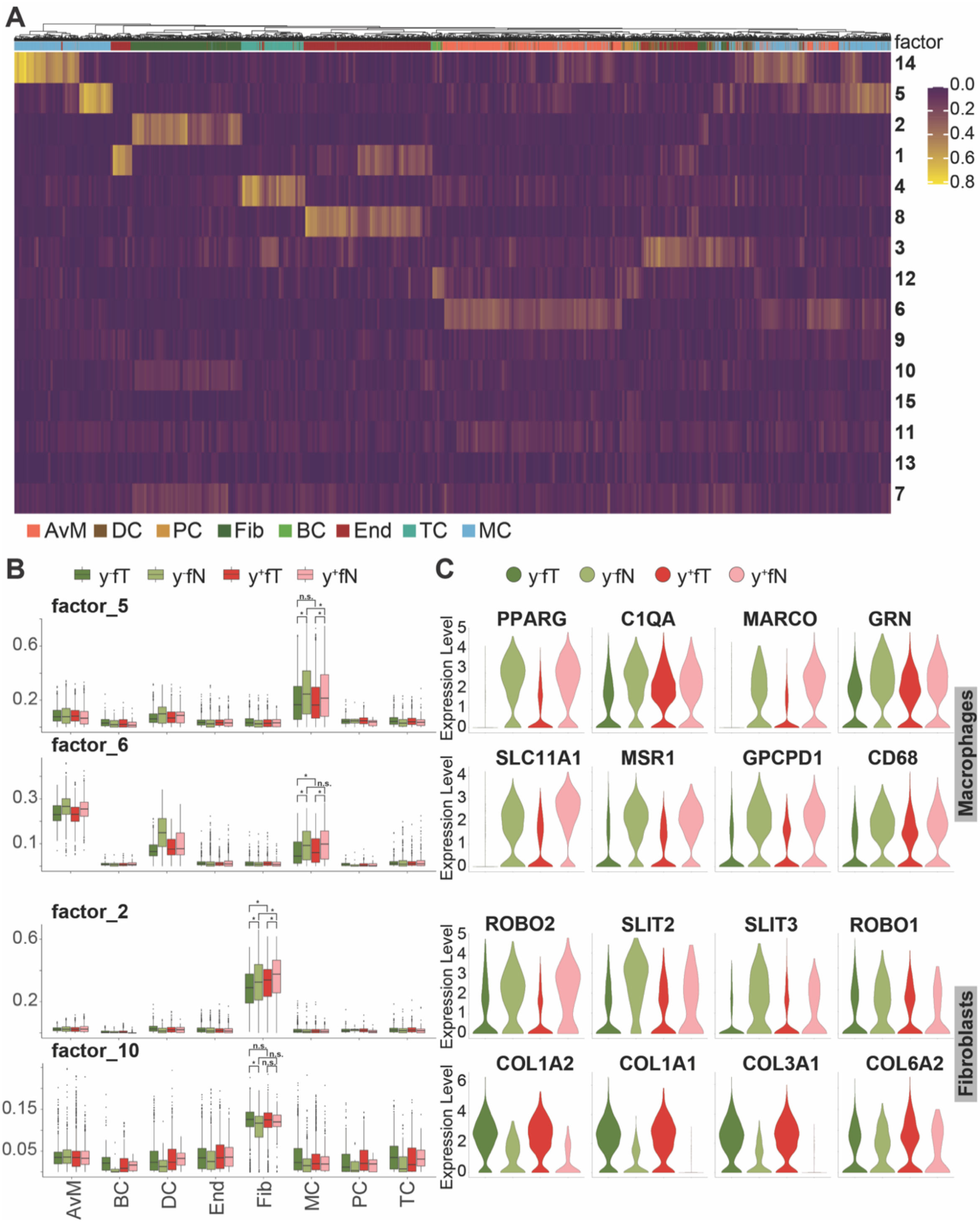


Figure 2.22 Microenvironment deregulation in the presence of LADC. (A) Using all non-malignant cells from young female patients in tumour and normal lung samples, gene expression signatures delineating cell types and states were identified by NMF. Colour scale indicates factor representation in each cell. **(B)** Contribution of selected factors to observed gene expression in different cell types is depicted separately for tumour and non-tumour tissue samples from never smokers (●/●) and smokers (●/●). * depict p-values < 0.001 calculated by two-sided ANOVA with post-hoc test using the Tukey's 'Honest Significant Difference' method. Statistical analysis is only shown for cell types of interest. **(C)** Expression levels of significant genes represented in the factors shown in **(B)**, across tumour and non-tumour tissue samples from never smokers (●/●) and smokers (●/●).

3 Discussion

Rapid technological developments over the past decade have enabled the interrogation of molecular features at the level of single cells using next generation sequencing. Through the miniaturisation of experimental procedures, lower sequencing costs and improved computational capabilities, it is now possible to investigate gene expression in large datasets comprising thousands or even millions of cells.

While previous studies at the bulk level largely provided insights into transcriptional features averaged over different cell types, single-cell profiling advances now enable the testing and re-evaluation of hypotheses at the level of individual cells. Distinct cell types, including those that occur at low frequency, and relationships between them may thus be identified. Moreover, differences in cellular composition or function between sample groups can now be interrogated at the single cell level, both in healthy tissues and under pathological conditions.

As a result, we have witnessed an explosion of new findings across all fields of the life sciences during the past few years [4, 315]. For example, single cell technology has facilitated charting previously inaccessible branches of the tree of life by investigating hard to culture microorganisms [316]. It has allowed invaluable insight into early embryonic development [317-320], interactions between fetal and maternal tissue [321] and spermatogenesis [322, 323]. Rare cell types with distinct roles in physiological or pathophysiological processes have been discovered in several tissues, including a new cell type of the mammalian lung involved in fluid regulation [105]. In addition, single cell approaches have made it possible to chart the immune system with its plethora of versatile cell types and functional states at unprecedented resolution [324], which led to the discovery of organ specific natural killer cell populations [325] as well as new subtypes of dendritic cells and monocytes [326].

As these examples indicate, the benefits of single cell technologies become especially apparent when investigating rapidly developing, evolving and heterogenous tissues. Single cell profiling is therefore also particularly suited to studying cancers, which often exhibit a high degree of cellular heterogeneity. Within a tumour cell population, rapid replication cycles and genomic instability drive evolutionary processes that determine tumour progression [195, 327]. These processes might also lead to adaption or resistance to therapeutic interventions [108], as has been demonstrated for NSCLC under tyrosine kinase inhibitor treatment [328]. To better understand the complex cellular architectures and behaviours underlying cancer disease trajectories, single cell sequencing technologies have therefore been employed to assess heterogeneity within and between patients in a great number of cancers [329-331], with important implications for patient stratification and treatment [332, 333].

3.1 Comparison of scRNA-seq technologies

To study single cell transcriptomes, a huge variety of different technologies have been developed over the past decade (compare section 1.2). With the initial aim to investigate cellular composition and single cell transcriptomes of different tissues across species, I compared four commercially available approaches in terms of their performance and suitability. These included the C1 (Fluidigm), the ICell8 (Wafergen) and the Single Cell RNA Seq-System (Dolomite-Bio), as well as the Chromium Controller (10x Genomics).

I successfully implemented all four of these technologies in our lab and used them to generate single cell transcriptome profiles from mouse forebrain and testis cells as well as nuclei, demonstrating their applicability for assessing cells of diverse morphology and their compatibility with different tissue dissociation protocols. As a result of system design, the number of single cell transcriptomes acquired per experiment varied greatly between technologies. We obtained data for an average of

49 cells in the C1, 237 cells in the iCell8 (using a quarter of the chip capacity), 1,841 cells in the Dolomite and 1,274 cells in the Chromium system.

The average number of genes detected per cell was also highly variable when accounting for read depth (Figure 2.3), with the highest number of genes per cell detected with the iCell8 system and comparable results in the C1 and Chromium systems. The lowest number of genes was detected using the Dolomite system, which might be due to its early developmental stage leaving room for optimisation. Detection of genes was comparable in nuclei and whole cells, despite the lower amount of input material.

Interestingly, recent studies have shown that increasing the number of cells is not strictly necessary for adding statistical power to single cell sequencing experiments, given sufficient read depth [39, 334]. To achieve an 80% true positive detection rate of genes across a simulated range of fold differences between two populations of cells, it has been estimated that 99 cells are required at a read depth of 1 million reads per cell. With a read depth of 500,00 reads per cell, the requisite cell number increased to 135, and the most cost effective was using 254 cells at a read depth of 250,000 per cell [334]. When designing single cell sequencing experiments, the trade-off between cell number and read depth for the desired sensitivity thus needs to be taken into account.

Apart from the sensitivity required to detect even lowly expressed genes, additional factors might also need to be considered during the design of a single cell transcriptomics study. At present, the C1 and iCell8 systems are the only available fully automated platforms that can generate full-length transcriptomic sequence information. When the experiments presented here were conducted, this was only true for the C1 system. The other single-cell profiling systems, in contrast, only provide data that is enriched for sequences from the 3' or 5' end of mRNA molecules. While this information may be sufficient for many scientific questions, as highlighted above (section 1.2), additional full-length information can be utilised to study otherwise undetectable transcriptional isoforms [22, 335-337], gene expression dynamics and splice variants [338], which are still only rarely considered in single cell sequencing studies.

Due to their geometry, the C1 and Chromium systems place considerable restrictions on the range of cell sizes that can be processed. The flexible microfluidic chip design of the Dolomite system is more permissive. However, only the iCell8 system is capable of examining very large cells such as multinucleated cardiomyocytes [339] or even multicellular structures like small cell spheroids and organoids [340]. To facilitate the processing of challenging samples like microorganisms or plant cells, the iCell8 and the Dolomite system (as well as its successor, the Nadia system) also allow for the flexible exchange of reagents and other technical parameters, extending the range of potential applications.

Finally, another aspect that might determine experimental design is the imaging capability built into the C1 and iCell8 systems. While the quality of the acquired images is limited, it enables the exclusion of cell doublets already at the stage of conducting the experiment, and offers the possibility to link morphological or other phenotypic information to single cell transcriptome data.

Comparison of single-cell transcriptome data for mouse testis generated using the Dolomite, iCell8 and Chromium technologies showed that all three enabled the identification of distinct cell types. Based on canonical marker gene expression, the Chromium system made it possible to identify all major cell types of mouse testis and sperm development and testis tissue, comprising Leydig and Sertoli cells, spermatogonia, spermatocytes, round and elongating spermatids. In contrast, Sertoli cells were not detected when using the Dolomite system, while the iCell8 system did not identify spermatogonia (section 2.1.4). For the iCell8 system, this might be primarily due to the low number of cells obtained, whereas the lower number of genes detected using the Dolomite system might hamper distinction of cell types. Consistently, cluster separation was clearest in the Chromium data, which might also be due to ambient RNA present in the single-nucleus solutions or lower sensitivity for the other systems.

When comparing single-cell transcriptomes across technologies, we also observed a higher correlation between different cell types processed with the same method, than between the same cell types using different methods. This indicates a significant degree of technical bias that is introduced by each method, affecting the observed transcriptome data. However, I demonstrated that this bias could be successfully corrected for using computational data integration methods (section 2.1.4), revealing the underlying cell type identity across technologies. Appropriate computational approaches thus enable the identification of common transcriptional features when comparing data from different single cell sequencing technologies, even in the presence of technical variation.

The data presented in this dissertation on the comparison of single cell sequencing technologies has some limitations. There is natural variation in the sampling process of tissue which might affect cell type composition. Moreover, the read depth per cell was not nearly exhausted and varies between experiments, which can only partly be overcome by subsampling of reads. To facilitate a more comprehensive comparison of different technologies, we therefore used the expertise acquired here in the context of an international multi-centre study that assessed differences between technologies and their applicability for creating cell atlases in a highly controlled manner [276]. As part of this study, we processed cells from a standardised mixture of cells from different tissues and species, distributed across all sites, using the Dolomite system. As in the above comparison of mouse testis tissue, the study revealed pronounced differences between technologies in the number of genes detected per cell and cell type composition, reinforcing the need for computational approaches to integrate data from different technologies when assembling large cell atlases. This study as well as our previous evaluation of the different technologies let us conclude that, given the distinct advantages of each technology, the most suitable choice at present for investigating heterogeneity in diverse tissues with a large number of cells would be the Chromium system as it provides high throughput, experimental robustness and relatively low cost.

3.2 Characterisation of healthy human lung tissue

Having implemented these scRNA-seq technologies in our lab, we sought to investigate tumour heterogeneity in healthy human lung tissue as well as lung adenocarcinoma of smokers and never smokers. As discussed above, the Chromium system was judged the most suitable for this task, with the most robust experimental workflow being imperative for the processing of rare clinical samples. The majority of single cell transcriptomics studies to date have relied on the acquisition of fresh material from surgical samples to isolate living cells, which poses significant logistical challenges especially when diseases occur relatively rarely, as is the case for lung adenocarcinoma in never smokers. Here, we present a retrospective approach performing single nuclei RNA sequencing from biobanked material. Consistent with our findings, recent studies have demonstrated that single-nucleus RNA sequencing provides adequate sensitivity and classification of cell types compared to using whole cells [341].

To create a cell atlas of the human alveolar lung that could serve as a reference for the comparison with LADC tumour tissue, we used patient matched samples of healthy tissue obtained at a distance from the tumour during surgery. These samples comprised a total of 41,061 cells from three individuals for each of four patient groups (young female smokers and never smokers, elderly female never smokers and young male never smokers). Based on this data, we established a reference map for all major cell types of the human lung, including epithelial and endothelial cell types, smooth muscle cells, fibroblasts and different immune cells (Figure 2.7).

3.2.1 SARS-CoV-2 infection of human lung cells

While our study of smoking effects on healthy lung tissue and LADC at the transcriptional level was underway, in 2019, the COVID-19 pandemic arose and many patients presented with severe lung disease following SARS-CoV-2 infection.

As this fast emerging global public health threat called for a concerted scientific effort to understand the pathophysiology of this new disease, we used our single cell transcriptome data to probe the expression of putative mediators of SARS-CoV-2 infection in healthy lung tissue [139]. Through rapid publication of our results, we contributed our reference cell atlas of the healthy human alveolar lung as a resource for the community and provided valuable information about human host factors for SARS-CoV-2 infection. Our findings thus supported the ongoing research on SARS-CoV-2 and might in the future be of further use to better understand the transcriptome of the lung in health and disease.

As the human lung single-cell atlas we generated was not initially intended for the study of viral infection, its use for this purpose certainly comes with limitations, including the relatively small cohort size. Nonetheless, our data include both smokers and never smokers from two adult age groups (40 to 56 years and 75 to 79 years) and thus comprise patient groups at a high risk for severe COVID-19, providing meaningful and immediately relevant insights. As the small sample number still limits the scope of my data for understanding the pathogenicity of SARS-CoV-2 in the context of different confounding factors such as age, gender and smoking history, investigation of gene and protein expression levels in a larger patient cohort will be required to test any hypothesis derived from this data.

Bearing in mind these limitations, we harnessed this lung single-cell atlas to assess the expression levels of *ACE2* across all human lung cell types. As ACE2 is currently the only receptor known to mediate SARS-CoV-2 binding to the host cell membrane, its expression is thought to render a cell susceptible to SARS-CoV-2 infection. We found *ACE2* expressed at extremely low levels overall, but with a significant enrichment in AT2 cells. Consistently, results from other groups quantifying *ACE2* expression in various tissues, including cell types of the respiratory tract such as nasal epithelia and lung, have corroborated a very low expression level in the lung and enrichment in AT2 cells [342-344]. While we did not observe an association of *ACE2* expression with age, sex or smoking status at the level of individual cell types, we observed a trend for age dependency of *ACE2* expression aggregated over all cell types. This finding was subsequently confirmed in a large meta study including the data presented here [345]. By making use of a greater number of cells, the meta analysis was also able to resolve cell type specific changes. It found that *ACE2* expression increased with age in AT2 cells and was elevated in men compared to women in secretory cells, AT1 and AT2 cells. Basal and secretory cells of past or current smokers displayed higher, and AT2 cells lower *ACE2* expression.

While these findings are suggestive of differential susceptibility to SARS-CoV-2 infection, it should be noted that socioeconomic factors may obstruct the assessment of age, sex and smoking as independent risk factors. A clear association of higher age with infection risk therefore remains controversial. The observed increased expression of *ACE2* in men compared to women supports observations of male individuals having a higher infection risk for SARS-CoV-2 [346, 347]. However, an increased infection risk has not been observed for smokers so far, although smoking has been associated with more severe symptoms [348].

An emergent question in COVID-19 research is why the viral load and possibly duration of infectiousness is much higher for SARS-CoV-2 compared to other coronaviruses, such as SARS-CoV or MERS-CoV [349]. Potential explanations comprise enhanced cleavage of the SARS-CoV-2 spike protein, resulting in higher infection rate, or an increased number of susceptible cell types. Employing our reference map of the human alveolar lung, we investigated additional host factors that might be involved in SARS-CoV-2 cell entry.

Coronaviruses are known to be able to enter into host cells via different endocytic pathways in the presence of proteases [146, 350-352]. SARS-CoV and SARS-CoV-2 both bind to the cell surface receptor ACE2, while TMPRSS2 has been identified as the main protease facilitating host cell entry [144, 145, 353, 354]. However, other proteases were also previously shown to enable coronavirus infection [343]. Of note, SARS-CoV-2 has a cleaving site for the protease FURIN that is absent in SARS-CoV [355, 356]. Recent studies suggest an increased binding affinity for ACE2 upon virus spike protein cleavage by FURIN [355], hypothesised to be caused by structural rearrangements of the cleaved spike protein as shown for other coronavirus spike proteins [357-359], which has now been confirmed by various in-vitro and animal model experiments [360-365].

We therefore investigated co-expression of *ACE2*, *TMPRSS2* and *FURIN* in the healthy human lung and found that the proposed contribution of FURIN to SARS-CoV-2 host cell entry would increase the number of cells susceptible to virus infection by about 12%, although susceptible cells continue to represent only a small fraction (0.7%) of all alveolar cells. Interestingly, FURIN might not only be active in its membrane bound form, but a secreted form has also been identified [281], which could further potentiate tissue susceptibility to SARS-CoV-2 infection in the neighbourhood of FURIN expressing cells. This intercellular activity of FURIN, however, still requires further exploration by future experimental approaches.

Taken together, these findings indicate that only a small proportion of alveolar cells are directly vulnerable to SARS-CoV-2 infection via ACE2 and TMPRSS2 expression, while FURIN might increase susceptibility by functioning as an additional protease able to cleave the virus spike protein and possibly even extending this effect to surrounding cells. Despite its limitations due to small sample numbers and the difficulty of assessing lowly expressed genes in scRNA-seq experiments, our healthy lung data provides a rich resource aiding further research into SARS-CoV-2 infection and acts as a reference for studies including primary samples of COVID-19 patients.

3.2.2 Smoking effects on non-tumour lung tissue

While the investigation of possible mediators of coronavirus infection using our healthy lung transcriptome data demonstrates its broader utility to the research community, the data was initially intended to serve as a reference for investigating transcriptional changes in response to tobacco smoke exposure and LADC. Analysing normal lung tissue samples in this context, we identified an increase in inflammation and immune activation induced by tobacco smoke exposure, with inflammatory signalling molecules such as CSF3, ICAM1 and IL6 mediating communication between immune cells as well as fibroblasts and endothelial cells (Figure 2.14).

Cell type composition was comparable across patient groups overall. However, an increased proportion of alveolar macrophages and AT2 cells in smoking patients was observable (Figure 2.12). These compositional changes might be a consequence of adverse effects of tobacco smoke. Furthermore, AT2 cells serve as alveolar stem cells and are capable of transdifferentiating into AT1 cells upon injury of the alveolar compartment [130]. Our findings are therefore suggestive of tissue damage and an overall inflammatory response with accompanying macrophage invasion into the tissue, consistent with higher alveolar macrophage numbers in smoker lungs also found in other studies [230, 233]. The increased proportion of AT2 cells that we observed might also reflect a higher proliferative activity of AT2 cells in smokers. Through increased cell division rates, AT2 cells could constitute a potential cell type of origin of LADC, in line with previous studies [366].

Similar consequences of tobacco smoke exposure have been proposed based on histology, lavage, elevated inflammatory molecules in peripheral blood or bulk transcriptome samples [213-216, 367]. While more recent single cell transcriptomic studies investigated the effects of tobacco smoke in systemic immune cells and upper airway epithelial cells [239-241], cell types in the alveolar region and their interplay had not been addressed at this resolution. My results therefore contribute to an improved understanding of smoking effects in the alveolar lung and might aid in the identification of therapeutic agents that could counteract the known tumourigenic effects of inflammation, a challenge that remains unsolved [368, 369].

3.3 LADC in female smokers and never smokers

The cell atlas of the human alveolar lung described in the previous section was established based on samples of healthy lung tissue obtained during surgery from patients with LADC. The second pillar of my work were samples obtained from the tumours themselves, which were also fresh frozen and subsequently processed for single nucleus RNA sequencing using the Chromium system. In total, 81,718 cells from tumour samples were included in this study after quality control, comprising 26 samples from four patient groups (young female smokers and never smokers, elderly female never smokers and young male never smokers). Patient-matched healthy lung samples, as already characterised (section 2.2.1), were available for three patients from each group. These data allowed me to explore cellular heterogeneity and interactions within LADC and the tumour microenvironment in patients with or without a smoking history. As lung cancer cases in never smokers exhibit a pronounced bias towards women [242], my study focused particularly on female smokers and never smokers.

A significant obstacle in the analysis of tumour tissues within this study was the substantial inter-patient heterogeneity, which precluded the direct inference of gene expression differences between male and female never smokers. Due to differences in genetic background, epigenetic modifications, patient history and comorbidities, this can only partly be overcome by larger sample sizes and molecular patient stratification. Computational methods that exclude patient specific features without losing biologically relevant signals will be necessary to further refine analyses of malignant cell populations across patients at the single cell level. In addition, our results based on single cell transcriptomics could be tested in larger patient collectives using bulk omics approaches. As smoking prevalence decreases, future studies should also address other environmental and intrinsic factors contributing to inflammation. These include inflammatory diseases such as chronic obstructive pulmonary disorder (COPD), which increases the risk of lung cancer independent of age, sex and smoking status [162, 370].

3.3.1 Identification of neoplastic and tumour microenvironment cells

Our atlas of single-cell transcriptomes from healthy lung tissue was used as a reference to annotate tumour sample cells, and cells that could not be assigned to any endogenous lung cell type were hypothesised to be the transformed cells of the tumour. The deviation from endogenous gene expression signatures in tumour tissue is often caused by mutations or large-scale structural genomic aberrations, such as gains or deletions of chromosomal parts [288, 289]. To corroborate the malignant identity of unassigned cells, we deduced copy number variations (CNV) from transcriptomic data by comparing the average expression levels of genes in close proximity on the genome to a baseline derived from patient-matched normal lung samples (section 2.3.1). Clustering of cells according to their inferred CNV profiles revealed cells devoid of copy number variations, which included cells from all patients analysed, as well as cells harbouring distinct losses or gains that were mostly patient-specific. Clusters containing CNVs were enriched for cells not representative of any healthy lung cell type, confirming that these previously unassigned cells are of malignant origin, while the remaining cells with low CNV prevalence were correctly annotated as cells belonging to the tumour microenvironment.

With this analysis, we were able to distinguish neoplastic cells and cells from the tumour microenvironment, which is still a challenging task in single cell transcriptomic experiments of cancer [371]. Here, we could make use of patient matched healthy lung samples as an appropriate reference for the comparison of tumour cell transcriptional profiles, thus overcoming a main obstacle for this inference analysis. Inferred CNV profiles could in principle be further analysed to investigate patient group specific genomic changes, although the present data did not establish any correlation with known genomic alterations in LADC (compare section 1.5.1) or identify novel variations with sufficient certainty. Single cell approaches assessing the genome or even new approaches to obtain DNA and RNA profiles from the same cell [55, 372] would greatly enhance the accuracy of distinguishing neoplastic cells and enable future investigations into genomic alterations at the single cell level as well as their relationship with transcriptional signatures.

3.3.2 High intratumoural heterogeneity with distinct cellular subtypes in young female patients

Tumour cell heterogeneity is increasingly recognised to play a crucial role in tumour progression, with implications for tumour evolution and efficacy of treatments [108, 199]. Previous single cell transcriptomic studies of LADC were often focused on resolving heterogeneity in the tumour microenvironment, rather than within malignant cells. Initial transcriptomic investigations of the neoplastic compartment have confirmed intratumoural cellular heterogeneity within single tumour sites [203], often driven by mutational differences [373]. However, an extensive characterisation of different subtypes of malignant cells is hampered by high interpatient heterogeneity and the difficulty of separating transformed cells from the TME. Having identified the malignant cells in tumour samples from our cohort, we therefore proceeded to investigate possible tumour heterogeneity within LADC.

The analysis of malignant cell transcriptomes revealed ten distinct cell populations with characteristic gene expression. Enrichment analysis of genes specific for each population identified eight expression signatures comprising proliferation, transcription, cellular respiration, cell adhesion, metabolism, morphological changes, phospholipid binding and immune related profiles. By linking these signatures to transcription factor networks, we found an association of proliferating malignant cells with stress response and amino acid homeostasis mediated by *ATF4* [296, 297], and with stem cell renewal mediated by *POU4F1*, also known as *OCT4* [298, 299]. *ATF4* inhibition is currently under debate as a possible drug target for cancer therapy. While only tested in biological in-vitro model systems so far, several strategies have been proposed for targeting *ATF4*, including upstream suppression of *ATF4* translation by inhibiting eukaryotic translation initiation factor 2 (*EIF2A*) phosphorylation, downstream suppression of *ATF4* targets, or inhibition of transcriptional activity [296]. Furthermore, the cell population enriched for an immune modulating signature showed additional expression of genes regulated by transcription factors *FOXN3* and *MEF2A*. Both are involved in cell cycle checkpoint control and contribute to EMT [300, 301].

Together, these results identified eight functional subpopulations of malignant LADC cells in both smokers and never smokers and provided mechanistical insight into underlying regulatory pathways that might aid in understanding and targeting LADC tumour heterogeneity in the clinic.

3.3.3 Trajectory of differentiation and characterisation of malignant cells in the context of smoking history

Despite the decline of tobacco smoking in industrialised countries, lung cancer remains the cancer with the highest mortality worldwide, and an increasing percentage of lung cancer patients present without a smoking history (section 1.5.3). Among never smokers that develop LADC, there also exists a bias towards women. I therefore used the single cell transcriptomic data to examine differences in the identified malignant cell clusters between young female smokers and never smokers. To probe the data for potential differences in the cellular differentiation hierarchy of the tumour, we employed pseudotemporal ordering and graph based trajectory inference to derive a differentiation trajectory. At its apex, this trajectory included the immune modulating cell population ('Imm_1') which comprised proliferating cells. We identified a differentiation path from proliferating through undifferentiated cells towards two differentiation states representing signatures of either autophagy or wound healing processes. Wound healing mechanisms have long been suggested to be involved in cancer progression, invasion and metastasis by creating a niche that fosters proliferation and tissue remodelling [374-378]. The role of autophagy, on the other hand, remains subject to debate; this process has been implicated in mitochondrial turnover in metabolically highly active cells as well as nutrient deficiency in poorly vascularised tissue [303, 379, 380]. The detected autophagy signature may thus reflect high metabolic activity or damage of cells along the trajectory.

Notably, no distinct cell populations unique to female smokers or never smokers were detected in either normal or malignant tissue samples, and the differentiation hierarchy was comparable across patient groups (section 2.2.1 and 2.3.1; Figure 2.20). However, we resolved distinct transcriptional properties of the cluster of malignant immune modulating cells ('Imm_1') according to smoking history, with increased expression of immune-related genes such as *ANXA1*, *C1QB*, *SLC11A1*, *CD68*, *PAEP* in smoker cells and *HLA-DQA2*, *HLA-DRB5*, *WFDC2* in female never smokers. In the same cell cluster, we also identified genes involved in migration and development that were specifically expressed in smokers (*MSLN*, *FNDC3B*) or never smokers (*AGR3*, *CLDN10*, *IGF2R*, *PCDH7*). These cells might therefore differentially modulate the immune microenvironment according to patient background and smoking status, with potential significance for immunotherapies.

A previous study applied trajectory inference approaches to LADC and normal lung cell transcriptomes simultaneously to investigate tumour progression, demonstrating that LADC comprise both transformed cells with high transcriptional similarity to normal epithelial cells and a subset of distinct tumour cells with increased expression of genes related to proliferation and migration [200]. Here, we focused exclusively on the tumour cells and determined a differentiation trajectory within the malignant cell population that consistently includes cycling as well as differentiating tumour cell states, in both smokers and never smokers.

3.3.4 Tumour microenvironment transcriptome is highly deregulated in LADC

Transformed tumour cells rely to a large extent on interactions with their surroundings, which might either hinder tumour development or work to its benefit [201]. We detected two distinct cell states within the macrophage population of tumour tissue with decreased expression of defined gene signatures compared to healthy lung (Figure 2.22). These signatures comprise genes involved in immune cell activation and inflammation, with one signature containing genes with a role in macrophage activation. Their downregulation in the tumour vicinity therefore indicates a more permissive environment for tumour growth. My results are consistent with findings from previous studies showing that a reduction in immune cell activation and a bias towards less tumouricidal immune cells in the tumour environment and might contribute to tumour progression [119, 381]. In particular, a subpopulation of macrophages expressing *MARCO* and *PPARG* was also identified in a recent single cell transcriptomics study of myeloid cells in NSCLC and associated with less favourable outcome [381].

Other cells of the tumour environment along with macrophages can also support or hinder tumour growth. We identified two populations of fibroblasts, one contributing to changes in extracellular matrix composition that are beneficial for tumour invasion and metastasis and another promoting EMT. Consistently, a recent single cell transcriptomic study of the tumour microenvironment in NSCLC found distinct fibroblast subpopulations with differential expression of EMT related genes [119]. Fibroblasts derived from tumour samples in this study also showed increased expression levels of type I and type III collagens compared to fibroblasts from healthy lung tissue, in agreement with my findings.

These results resolve how different macrophage and fibroblast subpopulations in LADC contribute to a tumourigenic environment, with implications for the design of therapeutic strategies targeting the TME.

3.4 Conclusion

Single cell sequencing technologies have transformed our ability to investigate cellular properties in health and disease, as well as across species, at unprecedented resolution. Ambitious efforts are underway, often as part of large international consortia, to chart the diverse cell types and states that make up all human tissues. By sequencing single cells under disease conditions, comparative information is additionally obtained that promises to enhance our understanding of pathophysiological processes.

Such ventures are driven by the explosion of technological approaches in the field of single cell sequencing over the past decade. To ensure their validity, it is necessary to carefully define experimental conditions and workflows that are appropriate for generating reproducible single cell data across laboratories.

The research presented in this thesis therefore initially set out to compare emerging technologies for single cell RNA sequencing and assessed their applicability for comparison between species, as well as between healthy and tumour tissue. Valuable insights into the specific advantages, limitations and experimental challenges of four popular technologies were obtained. Moreover, my results also contributed to a larger multi-centre study assessing the suitability of these technologies for collaborative cell atlas projects.

The ability to profile individual cells is particularly beneficial for investigating cell types and states as well as their interactions in complex tissues. Single cell sequencing research has already begun to transform our understanding of tumour heterogeneity and its interplay with the TME in a variety of cancers. Here, I applied the experimental insights gained from the comparison of different technologies to investigate healthy lung tissue, as well as tumour heterogeneity and the TME in LADC of smokers and never smokers, as discussed in the second part of this thesis. To address the challenge of limited sample availability, I adapted a protocol for single nucleus sequencing of biobanked fresh frozen material. My results demonstrate the feasibility of using this valuable material for retrospective studies, which greatly facilitates the investigation of rare diseases at single cell resolution and circumvents logistic obstacles for multi-institutional studies.

The single cell transcriptomics data I generated from healthy lung tissue provides a rich resource for investigating cellular diversity in the alveolar part of the human lung. Following the emergence of the SARS-CoV-2 virus, which severely affects this anatomical region in many patients, I harnessed my data to probe the expression of genes that are implicated in host cell entry and was thus able to make a timely contribution to our understanding of cellular susceptibility to coronavirus infection. Having demonstrated the utility of the human lung cell atlas for investigating pathophysiological mechanisms, I proceeded to address cellular heterogeneity in lung tissue and LADC in smokers and never smokers. Through my analysis of the single cell transcriptomics data I generated from healthy lung tissue, human lung cell types and mediators of inflammatory processes induced by tobacco smoke exposure in the alveolar part of the lung were resolved for the first time at single cell resolution. As female never smokers are particularly susceptible to LADC compared to their male counterparts, I investigated LADC samples with a focus on female patients. My results provide a refined description of cellular heterogeneity within LADC tumours and their microenvironment, defining transcriptional signatures for distinct transformed cell states. While the cell type composition and differentiation hierarchy of LADC were comparable in female smokers and never smokers, I identified a subset of cells with differential immune modulating activity dependent on smoking status. These findings will aid in the selection and development of treatments that take into account the complex interplay of disease aetiology, intratumoural heterogeneity and interactions with the tumour microenvironment.

4 Material and Methods

4.1 Sample procurement and Ethics agreement

4.1.1 LADC patients

Cryopreserved surgical lung tissue from patients with lung adenocarcinoma was provided by the Lung Biobank Heidelberg. All subjects gave their informed consent for inclusion before participation in the study.

This study was conducted in accordance with the Declaration of Helsinki and the Department of Health and Human Services Belmont Report. The use of biomaterial for this study was approved by the local ethics committee of the Medical Faculty Heidelberg (S-270/2001 (biobank vote) and S-056/2021 (study vote)).

Tumour tissue and an additional representative part of normal lung tissue distant from the tumour ($>$ 5 cm) was collected during routine surgical intervention. Pieces of 0.5-1 cm^3 were cut immediately after resection snap-frozen in liquid nitrogen within 30 min after resection, with no direct contact of samples and nitrogen. After snap-freezing, the vials were stored at -80°C and monitored regarding temperature until use.

4.1.2 Mice

All tissues from mice were obtained from outbred strain RjOrl:SWISS (Janvier Labs). From sacrificed mice of developmental age E13.5 the forebrain and for mice 4 weeks or 9 weeks after birth testis tissue was either snap frozen in liquid nitrogen for nuclei extraction (section 4.2.2) or immediately processed to isolate single cells (section 4.2.1). All procedures were approved by the Interfaculty Biomedical Research Institute of the University of Heidelberg, Germany, in accordance with federal guidelines.

In detail, mouse embryonic forebrain was dissected by retrieving the embryo from the freshly dissected uterus, put immediately in ice-cold phosphate buffered saline pH 7.4 (PBS). The whole procedure was performed within 30 min under a stereo microscope in ice-cold PBS. Decidua and Placenta were separated from the embryo by cutting the visceral yolk sac around these tissues. The visceral yolk sac was then cut from the embryo, and the brain freed from the embryos soft skin and skull. Forebrain was then cut before the midbrain, containing both cerebral hemispheres. The olfactory lobe was cut off. Testis tissue was dissected without the epididymis.

4.2 Experimental methods

Detailed information about reagents can be found in Table 9 and 10.

4.2.1 Isolation of cells from fresh tissue

Intact, living cells were isolated from fresh tissue immediately after dissection.

Mouse forebrain tissue was processed using the papain dissociation system (Worthington Biochemical) as to the manufacturer's instructions. All steps were conducted under sterile conditions. In detail the tissue was incubated for 30 min at 37°C under constant rotation, in prewarmed Earle's Balanced Salt Solution (EBSS) containing 20 U/ml papain, 1 mM L-Cystein, 1 mM EDTA and 0.005% DNase. Before use EBSS was saturated with O2 by vigorous shaking. To break up the tissue, it was pipetted two times with a 10 ml blow out pipette.

After incubation the mixture was triturated 10 times with a 10 ml blow out pipette and larger pieces were allowed to settle to the bottom. The cell suspension was carefully removed and laced in a 15 ml screw capped tube (Thermo Fisher Scientific) and centrifuged at room temperature and 300 g for 5 min. The supernatant was discarded and cells resuspended in 3 ml EBSS containing 1 mg/ml albumin, 1 mg/ml ovomucuoid inhibitor and 0.005 % DNase to stop the protease activity of papain and remove extracellular DNA. Cells were further cleared of any debris using a discontinues density gradient. The cell suspension was carefully layerd ontop of 5 ml EBSS containing 10 mg/ml albumin and 10 mg/ml ovomucuoid inhibitor in a 15 ml screw capped tube and centrifuged at 70 g for 6 min at 4°C. The cell pellet at the bottom was resuspended in ice cold Hanks Balanced Salt Solution (HBSS) (Sigma-Aldrich) containing 0.3% glucose (Sigma-Aldrich), filtered through a 20 µm cell strainer (pluriSelect Life Science) and kept at 4°C until further processing.

Mouse testis were processed with the mouse tumour dissociation kit (Milteny Biotech) according to the manufacturers protocol. Briefly whole testis tissue was placed together with 2.5 ml of the prewarmed vendors enzyme solution into a gentleMACS tube (Milteny Biotech), processed with a gentleMACS Dissociater (Milteny Biotech) running m_impTumor_02 and incubated at 37°C under constant rotation for 40 min. Afterwards the gentleMACS Dissociater program m_impTumor_03 was run and the mixture strained with a 70 µm strainer (pluriSelect Life Science) into a 15 ml screw cap tube. The strainer was washed with 10 ml RPMI-1640 (Gibco) and centrifuged for 7 min at 300 g. The supernatant was discarded and cells resuspended in 1 ml HBSS and kept at room temperature until further processing.

Cells were counted on a LUNA cell counter (Logos Biosystems).

4.2.2 Nuclei isolation from frozen tissue

For nuclei isolation an adaptation of a previously published protocol [268] was used. All solutions and material were precooled to 4°C and kept on ice for the whole procedure. Frozen tissue pieces, cut to cubes of about 5 mm, were placed in 1 ml of ice-cold homogenisation buffer (250 mM sucrose, 25 mM KCl, 5 mM MgCl2, 10 mM Tris buffer ph 7.5, 1 µM DTT, 0.4 U/µl RNAseIn, 0.2 U/µl SuperasIn, 0.1% NP40, 1 µg/ml Hoechst 33342) in a 1 ml Dounce Homogenizer (Wheaton) and crushed using five strokes of the lose pestle. After incubation at 4°C for 5 min the tissue was further homogenised by ten strokes of the tight pestle and strained through a 35 µm sized cell strainer into a 5 ml round bottom test tube (Corning). The strainer was washed with 500 µl homogenisation buffer and the mixture then centrifuged at 4°C and 500 g for 5 min. the cell pellet was resuspended carefully in 200 µl homogenisation buffer without NP40 using a 200 µl pipette and titrated ten times. Another 800 µl of detergent free homogenisation buffer was added and the solution centrifuged at 4°C, 500 g for 5 min. After discarding the supernatant, the pellet was resuspended in 200 µl detergent free homogenisation buffer using a 200 µl pipette and filtered again through a 35 µm cell strainer. The strainer was washed using 600 µl of ice-cold PBS and the solution mixed three times with a 1000 µl pipette. Isolated nuclei were now kept on ice until further processing.

Intact nuclei from snap frozen tissue of lung adenocarcinoma patients were isolated adopting the protocol described in [274]. Still frozen tissue was cut into cubic pieces of approximately 5 mm. They were transferred to a chilled 1 ml Dounce Homogenizer, filled with 1 ml of ice-cold homogenisation buffer (0.25 M Sucrose, 25 mM Citric Acid, 1 µg/ml Hoechst 33342). The tissue was broken by one stroke of the loose pestle and incubated for 5 min at 4°C and afterwards further broken down by five additional strokes of the loose pestle After another 5 min incubation at 4°C, the tissue was homogenised with ten strokes of the tight pestle and filtered through a 35 µm sized cell strainer. The filtrate was centrifuged at 500 g, 4°C for 5 min and the supernatant discarded. Nuclei in the pellet were resuspended in 700 µl homogenisation buffer, transferred into a new 1.5 ml Eppendorf Tube and again centrifuged at 500 g, 4°C for 5 min. After discarding the supernatant nuclei were resuspended in 100 µl, ice cold resuspension buffer (25 mM KCl, 3 mM MgCl2, 50 mM Tris-buffer pH 7.5, 0.4 U/µl RNaseIn, 0.4 U/µl SuperasIn, 1 µg/ml Hoechst 33342) and kept on ice until further processing.

Nuclei were counted on a Countess II FL Automated Cell Counter (Thermo Fisher Scientific) and diluted in resuspension buffer to 1 Mio nuclei/µl or less.

4.2.3 Single cell RNA sequencing library construction

4.2.3.1 Library construction C1

Cells were processed using the C1 Single-Cell Auto Prep System (Fluidigm) with the C1 Reagent Kit for mRNA Seq (Fluidigm) and SMARTer Ultra Low RNA Kit (Clontech). Cells were counted using Trypan blue solution (Gibco) with the LUNA automated cell counter and used in a concentration of 700 cells/µL. For loading of an IFC microfluidic chip (Fluidigm) designed for 5 to 10 µm cells, 6 µL cell suspension and 4 µL C1 suspension reagent were mixed and loaded with the mRNA Seq: Cell Load script. Cells are then captured at specific sites on the translucent IFC chip. Using a bright field microscope, the chip inspected for capture sites containing a single, viable cell. Doublets and dead cells were in this way excluded from further processing. Lysis, reverse transcription, and cDNA amplification were performed in the C1 system according to the manufacturer's instructions.

Subsequent library preparation was done using the Nextera XT DNA Library Preparation Kit (Illumina) as to the manufacturer's instructions.

Molar concentration of the sequencing libraries was quantified using the Qubit Fluorometer (Thermo Fisher Scientific), and fragment length was assessed using a Bioanalyzer 2100 (Agilent Technologies).

All libraries constructed with this protocol were sequenced on a MiSeq sequencer (Illumina) using paired-end protocol (75 bp).

4.2.3.2 Library construction iCell8

Living cells and nuclei were processed using the iCell8 single-Cell System (WaferGen) and iCell8 chip and reagent kit (WaferGen). Cell suspensions were stained with ReadyProbes Cell Viability Imaging Kit (Thermo Fisher Scientific), containing Hoechst 33342 and propidium iodide for 20 min at RT. Cell suspension was then diluted to 20 cell/ul and distributed with the liquid handler of the iCell8 System, into 5,184 the nanowell chip (WaferGen) and centrifuged at 300 g for 5 min at RT. After imaging all wells with the fluorescent microscope of the iCell8 System, cells were frozen in the chip at -80°C over night. Well images were inspected for single intact cells. Empty as well as wells containing dead or multiple cells were excluded from further processing. The nanowell chip was then thawed for 10 min at RT to lyse the cells and centrifuged at 3,220 g for 3 min at 4°C. Free polyA mRNA was annealed to pre-printed primers in the chip by heating the chip for 3 min at 73°C in a modified PCR cycler (Bio-Rad T100). The chip was subsequently centrifuged at 3,220 g for 3 min at 4 C and reverse transcription reagents applied to selected wells using the iCell8 System liquid handler. Reverse transcription was carried out with template switch extension at 42°C for 90 min and cDNA collected by centrifugation using the iCell8 Collection Kit (WaferGen) in a 1.5 ml microcentrifuge tube.

cDNA was then concentrated with the DNA Clean & ConcentratorTM-5 kit (Zymo Research) and single stranded DNA eliminated by exonuclease treatment (Exonuclease I; 37°C for 30 min and 80°C for 20 min). The product was further amplified by PCR (90°C for 1 min; 18 cycles of 95°C for 15 s and 65°C for 30 s; 68°C for 6 min; 72°C for 10 min). Library construction was then performed using the Nextera XT DNA Library Preparation Kit as by the manufacturer's instructions, which includes tagmentation to introduce amplification and sequencing primer, but amplification only using sequences included in the Nextera Transposase Sequence and the pre-printed polyT oligonucleotides. This enriches for 3' cDNA fragments of about 300 nt length.

Concentration of the sequencing libraries were quantified using the Qubit Fluorometer, and fragment length was assessed using an Agilent Bioanalyzer 2100. All libraries constructed with this protocol were sequenced on a HiSeq500 (Illumina) sequencer in high-output mode, paired-end 26 x 49 bp.

4.2.3.3 Library construction RNA-Seq System Dolomite

Single cell RNA libraries were constructed using the Dolomite µEncapsulator system (Dolomite Bio) as to the manufacturer's instructions and published protocols [34].

In detail, 600,000 barcoded beads (ChemGenes) were resuspended in 1,000 µl lysis buffer (500 µl nuclease free water; 300 µl of 20 % Ficoll PM-400; 10 µl of 20 % Sarkosyl; 30 µl of 0.5 M EDTA; 100 µl of 2 M Tris pH 7.5; 50 µl of 1 M DTT). All tubing was primed using either the same buffer as used for the cell suspension or QX200^{TM} Droplet Generation Oil for EvaGreen (Bio-Rad). The single cell suspension was placed in a 1.5 ml Eppendorf tube inside a magnetic stirrer to prevent settlement of cells. Beads were resuspended and injected into the 10 m sample loop using a 1 ml luer lock syringe. Connect the microfluidic chip to the appropriate tubing and start the oil, cell and bead pump with following setting (oil: 200 µl/min; cells: 60 µl/min; beads: 60 µl/min). Continuous flow of liquids in all tubing and homogenous formation of water droplets in oil was observed using a high speed camera (Dolomite Bio). After stable formation of droplets was established, droplets are collected in a 50 ml falcon tube.

After collection of about 1 ml of droplet emulsion, the oil phase (clear, lower phase) is removed using a 1000 µl pipette and droplets are broken by adding 30 ml 6x SSC buffer (Sigma-Aldrich) and 1 ml of perfluoroctanol (Sigm-Aldrich) under a fume hood, followed by vigorous shaking for three times. The solution is then centrifuged at 1,000 g for 1 min to create two separate phases with beads accumulating at the interface. The upper phase is removed until only a few millilitres remain above the interface and 30 ml 6x SCC buffer is added. After a few minutes two phases have again separated and the upper phase, containing the beads is transferred to a new 50 ml falcon tube. This phase is centrifuged at 1,000 g for 1 min and all supernatant but 1 ml are carefully removed. The remaining liquid is mixed and transferred to a 1.5 ml Eppendorf tube and centrifuged at 1,000 g for 1 min. The supernatant is discarded and beads are washed two times using 1 ml of 6x SCC buffer and one time with 300 µl of 5x RT buffer for Maxima H minus reverse transcriptase (Thermo Fisher).

The supernatant is removed and beads are resuspended in 200 µl of reverse transcription mix (75 µl nuclease free water; 40 µl Maxima 5x RT buffer; 40 µl of 20 % Ficoll PM-400; 20 µl of 10 mM dNTPs; 5 µl of RNAse inhibitor (40 U/µl);

10 µl of 50 µM template switch oligo; 10 µl Maxima H minus reverse transcriptase). The suspension was then incubated at RT for 30 min, followed by 90 min at 42°C with rotation at 1,000 rpm.

Beads are then washed with 1 ml of 10 mM Tris pH 8.0, by centrifugation at 1,000 g for 1 min and resuspended in 200 µl exonuclease mix (170 µl nuclease free water; 20 µl of 10x Exo I buffer; 10 µl Exo I). After incubation at 37°C for 45 min with rotation at 1,000 rpm, beads were washed as described above using one time 1 ml TE-SDS (10 mMTris pH 8.0; 1 mM EDTA; 0.5 % SDS), two times 1 ml TE-TW(10 mMTris pH 8.0; 1 mM EDTA; 0.01 % Tween-20) and one time 1 ml nuclease free water.

Amplification of cDNA by PCR is done on batches of 2,000 beads in 50 µl of PCR-mix (24.6 µl nuclease free water; 0.4 µl of 100 µM SMART PCR Primer; 25.0 µl of 2x Kapa HiFi Hotstart buffer) in a thermal cycler (95°C for 3 min; 4 cycles of: (98°C for 20 s; 65°C for 45 s; 72°C for 3 min); 9 cycles of: (98°C for 20 s; 67°C for 20 s; 72°C for 3 min); 72°C for 5 min). For cDNA prepared from nuclei 4 +11 cycles were used.

Amplification product was purified adding 30 µl AMPure XP beads (Beckman Coulter) and incubation for 5 min at RT, by subsequent pelletising and washing of magnetic beads on a magnetic stand twice with 200 µl of 80 % ethanol (Thermo Fisher Scientific). The pellet was air dried for 2 min and DNA eluted in 10 µl nuclease free water by incubation for 5 min at RT.

Yield and expected size of 1,300-2,000 nt) was determined using a BioAnalyzer High Sensitivity DNA Chip (Agilent Technologies).

600 pg of purified DNA in a total volume of 5 µl nuclease free water was transferred to a new PCR-tube and 10 µl of Nextera TD buffer (Illumina) as well as 5 µl Amplicon Tagment enzyme (Illumina) were added and mixed by pipetting on ice. After incubation at 55°C for 5 min in a thermal cycler 5 µl Neutralisation buffer was added, mixed by pipetting and incubated for 5 min at RT. Afterwards 15 µl of Nextera PCR mix, 8 µl nuclease free water, 1 µl of 10 µM DropSeq-P5 SMART PCR primer and 1 µl of appropriate 10 µM Nextera N70x oligo was added in this order.

PCR was run in a thermal cycler with following settings: (95°C for 30 s; 12 cycles of: (95°C for 10 s; 55°C for 30 s; 72°C for 30 s); 72°C for 5 min).

The Final library was purified as described above using 30 µl AMPure XP beads and elution in 10 µl nuclease free water.

Fragment size and concentration was analysed using a BioAnalyzer High Sensitivity DNA Chip.

All libraries constructed with this protocol were sequenced on a HiSeq500 sequencer (Illumina) in high-output mode, paired-end 26 x 49 bp.

4.2.3.4 Library construction Chromium^{TM}

Single cell RNA libraries were constructed with the Single Cell 3' Reagents Kit v2 (120237; 120236; 120262; 10x Genomics) and the Chromium^{TM} Controller (10x Genomics) according to the manufacturer's instructions using 16,000 nuclei as input. Briefly, 33.8 µl of cell suspension containing 16,000 nuclei were mixed with RT-master mix and carefully pipetted 5x on ice. The chip was loaded with 90 µl of cell/master mix suspension. After vortexing for 30 s, 40 µl RNA capture gel beads were loaded on the chip as well as 270 µl of partitioning oil according to the manufacturer's instructions. Subsequently the prepared chip was placed in the chromium controller and Single Cell A program run.

After completion, 100°µl of gel beads-in-emulsion (GEM) was transferred to a 8-tube strip (Thermo Fisher Scientific) and remaining emulsion was checked with the Countess II FL Automated Cell Counter for successful encapsulation of nuclei and beads.

Reverse transcription was carried out in a thermal cycler (Bio-Rad) with following specifications. 53°C for 45 min; 85°C for 5 min. cDNA was then cleaned using DynaBeads MyOne Silane (Thermo Fisher Scientific). After addition of 125 µl Recovery Agent and incubation at RT for 60 s, the lower phase containing recovery agent and partitioning oil was discarded. 200 µl of DynaBead Cleanup Mix was added (9 µl nuclease free water; 182 µl buffer sample clean up 1; 4 µl DynaBeads MyOne Silane; 5 µl Additive A), mixed by pipetting and incubated at RT for 10 min.

Solution was cleared using a magnet stand (10x Genomics) and the supernatant removed. 300 µl of 80 % ethanol were added and removed after 30 s. This wash was repeated one time, the tube strip briefly centrifuged and remaining ethanol removed. After air drying the pellet for 1 min it was resuspended in 35.5 µl Elution Solution (98 µl Buffer EB: 1 µl of 10 % Tween 20; 1 µl Additive A), mixed and incubated for 1 min. Solution was cleared using the magnet and 35 ml of the supernatant transferred to a new 8-tube-strip.

Cleaned cDNA was amplified by addition of 65 µl of Amplification Reaction Mix (8 µl nuclease free water, 50 µl amplification master mix; 5 µl cDNA Additive; 2 µl cDNA Primer Mix) and incubated in a thermal cycler using following PCR settings: 98°C for 3 min; 10 cycles of: (98°C for 15 s; 67°C for 20 s; 72°C for 1 min); 72°C for 1 min. After amplification 60 µl of SPRIselect Reagent Mix (Beckman Coulter) was added and magnetic beads washed with ethanol as described above. DNA was eluted using 40.5 µl Buffer EB and after incubation at RT for 2 min, 40 µl were transferred to a new 8-tube-strip. Successful reverse transcription and cDNA amplification was assessed using an Agilent TapeStation High Sensitivity D1000 ScreenTape (Agilent Technologies).

DNA was then fragmented and A-tailing achieved by adding of 15 µl Fragmentation Enzyme blend (2:1 in fragmentation buffer) on ice and incubating in a precooled thermal cycler (4°C) at 32°C for 5 min and 65°C for 30 min. Fragmented and A-tailed DNA was recovered using 30 µl of SPRIselect Reagent Mix. After incubation for 5 min and clearance with a magnet, the supernatant was transferred to a new tube strip. DNA in the supernatant was then cleaned using 10 µl of SPRIselect Reagent Mix as described above using two ethanol washes. DNA was eluted in 50.5 µl Buffer EB, incubated for 2 min, cleared with a magnet and 50 µl transferred to a new tube strip.

Sequencing libraries were constructed by ligation of adaptors, adding 50 µl of Adaptor Ligation Mix on ice (17.5 µl nuclease free water; 20 µl ligation buffer; 10 µL DNA ligase; 2.5 µl adaptor mix) and incubating at 20°C for 15 min. Ligation product was cleaned using 80 µl of SPRIselect Reagent Mix as described above, discarding the supernatant. DNA was eluted in 30.5 µl Buffer EB, incubated for 2 min, cleared with a magnet and 30 µl transferred to a new tube strip.

Individual indexing and amplification of each library was done by addition of 60 µl sample Index PCR Mix (8 µl nuclease free water; 50 µl amplification master mix; 2 µl SI-PCR primer and 10 µl of individual chromium i7 sample index) on ice. PCR was executed using following settings: 98°C for 45 s; 15 cycles of: (98°C for 20 s; 54°C for 30 s; 72°C for 20 s); 72°C for 1 min. The finished library was recovered using 60 µl of SPRIselect Reagent Mix. After incubation for 5 min and clearance with a magnet, the supernatant was transferred to a new tube strip. DNA in the supernatant was then cleaned using 20 µl of SPRIselect Reagent Mix as described above using two ethanol washes. DNA was eluted in 35.5 µl Buffer EB, incubated for 2 min, cleared with a magnet and 35 µl transferred to a new tube strip.

Quantity and quality of each library were assessed using an Agilent TapeStation High Sensitivity D1000 ScreenTape.

All libraries constructed with this protocol were sequenced on HiSeq4000 (Illumina), paired-end 26x74 bp, one sample per lane.

4.2.4 Immunohistochemical staining

Paraffin-embedded tissue sections were deparaffinized and peroxidases were blocked for 10 min at room temperature (RT) using 3 % H2O2 (Applichem, Darmstadt, Germany). Antigen retrieval was performed in a steamer with sodium-citrate-buffer (10 mM sodium citrate, 0.05% Tween 20, pH 6.0) for 15 min. The staining procedure for the polyclonal anti-glycodelin antibody (sc-12289, Santa Cruz Biotechnology, Heidelberg, Germany) was performed with DAKO EnVision+ System-HRP (AEC) for rabbit primary antibodies (Dako, Hamburg, Germany). The tissue slides were incubated overnight at 4°C with an anti-glycodelin antibody at a concentration of 2.5 µg/ml. A linker (rabbit anti-goat IgG, A27001, Thermo Scientific) antibody was used for 30 min at room temperature before tissue sections were incubated with secondary antibody for another 30 min at room temperature. Visualization of glycodelin was performed with AEC+ Substrate-Chromogen (Dako). For ANXA1 staining, the staining procedure was performed with SignalStain® DAB Substrate Kit (#8059, Cell Signaling) according to manufacturer's instructions. Cell nuclei were stained using Mayer's Hematoxylin Solution (Sigma-Aldrich, Munich, Germany). Staining was observed with an Olympus IX-71 inverted microscope. Pictures were taken with an Olympus Color View II digital camera and Olympus Cell-F software (cellSense dimension, V1.11, Olympus, Hamburg, Germany). Tiffs were assembled into figures using Photoshop CS6 (Adobe, San José, CA, USA). Only changes in brightness and contrast were applied. Scoring was performed by multiplication of staining intensity (0-3) with the proportion of positive cells (0-4). For each patient, five randomly selected pictures were analysed and median was calculated.

4.3 Computational analysis of single cell transcriptomic data

All computational analysis was performed on an HPC system running CentOS Linux v7.6.1810 using R v3.6, python v3.7.6, command line tools and software packages with versions indicated in Table 8. All statistical analysis was conducted using R with the indicated methods or as described in individual software packages. Visualisation was done employing the ggplot2 [382] or ComplexHeatmaps [383] packages in R.

4.3.1 Sequence alignment

For all sequence alignments the human GRCh37 or mouse GRCm38 genome assembly of the Genome Reference Consortium was used.

4.3.1.1 zUMIs

For alignment of scRNA-seq data generated with the C1-, iCell8- or Dolomite system the zUMI pipeline v2.9.4 has been employed [384]. Briefly this pipeline filters reads by user defined quality settings and spurious barcodes under a set threshold. Here phred score [275] of 20 and a minimum number of reads per barcode of 100 was used. Then reads are mapped using the STAR aligner v2.7.3a, which is aware of splicing sites and can therefore determine intronic and exonic reads [385]. Reads are then assigned to genes using Rsubread v2.0.0 featureCounts [386] and count matrices are generated using R, considering information from UMIs to correct for amplification biases during library construction.

4.3.1.2 CellRanger

scRNA-seq data generated with the Chromium system (section 4.2.3.4) were aligned using an implementation of the STAR alignment method [385] made available as CellRanger v2.1.1. This workflow considers UMIs and therefore corrects for amplification biases during library construction. The output is a matrix of cell barcodes and counted observations of genes.

4.3.2 Quality control

As a measure of data quality, several parameters commonly used in next generation sequencing analysis, have been considered.
First library quality has been determined by fragment size and libraries deviating from a theoretical optimal size and distribution were not processed further. cDNA was expected to be around 2,000 nt to ensure full length reverse transcription, no

RNA degradation and no contamination with genomic DNA for human and mouse samples.

Sequence data was further assessed for Phred base quality score [275], which should have a uniform high value above 25 for at least 30 bases and a base content, which is expected to be uniform across the length of the read, with equal amounts for T and A or C and G bases. For this the implementation of FastQC v0.11.9 was used. Sequence data not fulfilling these criteria were not further processed (e.g. Supplementary Figure 1 A-D).

In addition, data with a mapping efficiency of less than 80 % total reads mapped to the reference genome, as given by the respective alignment algorithm was discarded. All these measures were taken to avoid analysing data were library construction or sequencing failed.

Barcodes belonging to intact and not partial single cells or nuclei or empty reaction volumes with ambient RNA where filtered by building the cumulative sum of all reads over barcodes, starting from the barcode with the highest number of associated reads. Only barcodes up to the first infliction point of the curve have been considered valid barcodes (e.g. Supplementary Figure 1 E).

In single cell or nuclei sequencing there is also a likelihood of doublets or multiplets, RNA of more than one cells with the same barcode, which would hinder meaningful analysis. To avoid this, the count matrices for each sample after sequence alignment was filtered for cells having an extreme number of genes or UMIs detected, which indicates RNA from multiple cells. Further, cells with extremely low number of detected genes and UMIs have been discarded to avoid analysing cell fragments or empty reaction compartments harbouring only ambient RNA. Detailed information of threshold applied to each experiment can be found in Table 1 and Table 5.

4.3.3 Data processing, clustering and visualisation

First assessment, clustering and visualisation of individual data has been done using functionalities implemented in the Seurat software package v3.1.3 [88].

Briefly, to account for differences in total amount of RNA found in individual cells read count data was normalised by dividing counts for each cell by the total number of counts in this cell, multiplied by 10,000 and natural-log transformed.

Data was then centred so that each gene has a mean of 0 across the whole data set and scaled by the standard deviation of each gene. To reduce outlier effects scaled data higher than 10 was set to 10.

Dimensionality of the data is reduced using PCA, which is used as an input to construct a Shared Nearest Neighbour (SNN) Graph, by first determining the k-nearest neighbours of each cell and then calculating the neighbourhood overlap between each cell and a set number of its nearest neighbours.

Clusters of cells are then determined by an algorithm using the SNN graph and louvain clustering [96, 387].

The data is then visualised with the Uniform Manifold Approximation and Projection for Dimension Reduction algorithm [388].

4.3.4 Downsampling of read counts

Counts from each cell were downsampled to 20,000 reads per cell using the SampleUMI function implemented in the Seurat software. Cells that were below this threshold have been discarded.

4.3.5 Correlation of average gene expression in cell types

Average expression per cell type was calculated on downsampled data and Spearman correlation calculated between cell types determined by the same and all used scRNA-seq technologies as implemented in the R stats package.

4.3.6 Sample integration

To correct for technical biases between different single cell sequencing library data and preserve biological differences data from different samples were integrated as implemented in the Seurat analysis R package [88].

This approach uses canonical correlation analysis to perform dimension reduction on all data sets that will be aligned. To do this, canonical correlation vectors are identified that describe a shared gene correlation structure between two data sets. In contrast to e.g. multiple regression analysis, correlations of genes inside and between data sets can be taken into account. Canonical correlation vectors are further L2-normalised, that is the sum of the squares of all elements will be up to 1, to mitigate global effects such as sampling differences. In this dimensionally reduced space, k-nearest neighbours for all cells inside a data set are calculated for a given k. Afterwards, for each cell in one data set the k-nearest neighbours in the other data set is calculated. If two cells are both found to be in the set of cells defined as nearest neighbours of each other, they are considered to be mutual nearest neighbours and considered to serve as anchors between the data sets, since they are likely belonging to the same cell type and state.

The algorithm further applies a two-step method to avoid incorrectly identified anchors. First, for every pair of mutual nearest neighbours identified in the dimensionally reduced space, k-nearest neighbours of the second cell of the pair are identified in the original high-dimensional data that included the first cell. If the first cell of the pair does not appear within the first 200 nearest neighbours, the pair will be removed from the list of anchors, thereby filtering out false anchor pairs. This search is performed using the top 200 genes that have the highest contribution to the previously identified canonical correlation vectors. Second, for each cell of an anchor pair, k-nearest neighbours in its own and the paired data set are calculated with k equal to 30. This results in four matrices that are combined to calculate a neighbourhood graph. On this the shared neighbour overlap for each anchor pair is calculated. Together with the distance of each cell from the first data set to the anchor, a weight matrix is calculated. Using the identified anchors, a matrix is calculated where each column is the difference between the two vectors defined by each of the anchor cells. This matrix, together with the weight matrix, is then used to calculate a transformation matrix, that is applied to the original data and results in an expression matrix of the same dimensions as the original data [88].

4.3.7 Cell type inference from reference data

Cell type prediction from scRNA-seq reference data is implemented in the ‚TransferData' function of the Seurat R package [88]. Briefly, anchor cells in the reference data that define a mapping between the query and the reference data are determined as described above (compare section 4.3.6). The resulting weight matrix is multiplied by a binary anchor identity matrix of cell type labels times anchor cells, which results in a prediction score for each query cell and possible cell type label.

4.3.8 Calculation of CPM values

Read count data was normalised by dividing counts for each cell by the total number of counts in this cell, multiplied by 10,000 followed by natural-log transformation. CPM values were then calculated by aggregating reads for each cell type separate for each sample and dividing aggregated reads through the sum of all reads in this cell type followed by multiplication with 1,000,000.

4.3.9 Cell type compositional changes

The statistical analysis of cell type compositional changes has to overcome many obstacles inherent to single cell experiments, including low number of replicates and sample size as well as high technical variability. It further has to take into account proportional changes, so that a reduction in one cell type is not falsely interpreted as an increase in other cell types. We therefore employed here a Bayesian model approach developed and published as ‘sccoda’ [286], which is based on a hierarchical Dirichlet-Multinomial distribution.

4.3.10 Differential expression and gene set enrichment analysis

Differential expression analysis to find which genes in one population of cells is higher expressed than in another population of cells, is performed using the Mann-Whitney U test as implemented in Seurat’s FindMarker (as test.use = “wilcox”) [97], if not stated otherwise. This non-parametric test finds genes for which the probability to have a higher expression value in one population of cells (Y) compared to another (X) is different than the probability of a higher expression in X compared to Y.

To decrease the influence of interpatient heterogeneity in the analysis of differences between patients with different smoking habit, the before mentioned differential expression analysis was performed as follows for two populations of cells (X and Y), where cells from one patient could only belong to one population of cells at a time. All cells from one patient were compared to all cells from all patients belonging to the respective other population of cells. Then differentially expressed genes for each patient were sorted by average logarithmic fold change and p-value. Only genes with a p-value less than 0.01 have been considered. Of this sorted list, the first 150 genes were taken. Afterwards each differentially expressed gene was counted for its occurrence in their group of patients belonging to the same population. From this list, only genes were considered to be differentially expressed between X and Y, that were found in at least 30 % of patients from this group.

Genes that were found to be differentially expressed were further analysed for their enrichment in gene sets from The Molecular Signatures Database (MSigDB) [389], specifically C5 (Gene Ontology) v7.2. For this enrichment analysis as implemented in clusterProfiler v3.14.0 [390] has been employed. Here a hypergeometric test assessed the significance of defined differentially expressed genes being overrepresented in given sets of genes compared to random sampling.

4.3.11 Copy number variation inference from mRNA expression data

Cancer genomes are inherently instable [288, 289] and may therefore have a divergent copy number for parts of the genome, from the usual two copies in the human genome. This is termed copy number variation (CNV).

To assess possible differences in CNVs a method included in the inferCNV package v1.2.0 [391] was used. This method compares RNA expression changes in genes on proximal genomic position to a reference. As a reference single cell transcriptome data from patient matched normal samples was used to infer these changes in tumour samples. In detail, the average expression of genes in a moving window of 100 genes from a list, ordered by genomic position was used and compared to the average expression of the same genomic location in the reference. Higher average expression in the tumour sample was then considered to potentially harbour a gain in genome copies in this region, while lower expression potentially harbours a loss of genome copies.

4.3.12 Transcription factor network analysis

Transcription factors are a major regulator for expression of genes that together promote a phenotypic trait in a cell. Therefore, a cell state should be better defined by this regulatory network rather than the expression of a single gene.

We investigated these networks employing the pySCENIC software package v0.10.3 [392], that infers gene regulatory networks for a known set of transcription factors in single cell transcriptomic data, by a three-step algorithm.

First, a ranked list of pairwise comparisons between target genes and regulators (i.e. transcription factors) is produced. For this, an algorithm using a Random Forest Model to predict the strength of putative links between target gene and transcription factor as described in Huynh-Thu and implemented in GRNBoost2 [392] was used.

Sets of genes were derived from this list of genes, that have been assigned a potential regulator and a relevance score, by discarding gene – regulator pairs with less than

0.001 relevance score and deriving sets that have either a higher than 0.001 or 0.005 relevance score. From these sets the 50 genes for each transcription factor were taken and only the top 5, 10 and 50 transcription factors for each gene kept. For these gene sets the spearman correlation between transcription factor and gene are calculated and the sets are split in positive and negative correlation. Further only gene sets with more than 20 members were further analysed.

This list is only based on co-expression and might therefore harbour many false positive results. To filter these potential false positives each set of genes was therefore analysed in a next step for enrichment of transcription factor binding motifs, by searching for motifs that are enriched in the transcription start site of the gene and for each motif putative target genes in the gene set are predicted. For further analysis, only genes that have a positive correlation with the transcription factor binding motif are kept. This is implemented in RcisTarget [392] and references based on the GRCh37 genome as provided by [393]was used.

Finally, in a third step each gene set is scored for their importance in each cell using the AUCell method described in [392]. This method ranks all genes for one cell and all genes in the gene set by expression in this cell. Subsequently the area under the curve is calculated for a curve constructed on a cartesian system with x values representing the ranked genes from high to low expression and y values being the position of the particular gene in the ordered gene set list (e.g. 1 for gene at position 1, 2 for gene at position 2). Therefore, an enrichment of highly expressed genes in the gene set results in a high area under the curve value.

The final result is now a matrix of cells and gene sets, that are defined by a transcription factor (transcription factor modules), with entries being the area under the curve scores. For comparison with gene expression in identified cell types, the median for AUC scores by cell type was calculated and the most informative modules determined by hierarchical clustering and PCA.

4.3.13 Receptor – Ligand interactions

Cell to cell interactions via ligand and receptor proteins mediate diverse biological processes and shape a cell's and therefore tissues phenotype. Based on the RNA expression levels for these proteins we inferred likely interactions in the single cell data. For this a curated database of ligand-receptor pairs, that also considers the subunit composition for the receptor, and statistical framework as implemented in CellPhoneDB v 2.1.4 [394] was used. The analysis first calculates the mean expression for each gene in the database, pooled by cell cluster annotation and the percentage of cells in this cluster expressing the gene. Through iterative random shuffling of cell labels a null distribution for each gene pair is then derived (1,000 iterations), taking into account the expression levels. This is compared to the observed mean of ligand and receptor in two clusters of cells and a p-value for their expression specifically in this pair of clusters derived from the null distribution. Ligand-receptor pairs are then ranked by p-value and significant interactions determined. For this analysis integrated expression matrices, as described in section 4.3.6 were used.

4.3.14 Trajectory inference

Inference of a possible developmental trajectory of cells was realised using a framework employing principal graph inference, published as STREAM v1.0 [102]. Integrated expression matrices (section 4.3.6) were used to first define variable genes using non-parametric local regression, which were then used to reduce the dimensionality of the data employing modified locally linear embedding [102]. This method provides a continuous embedding, by considering local similarity to its neighbours. In this space cells are clustered using the affinity propagation method [395]. The result is then used to construct a minimum spanning tree to use as an initial tree structure for the construction of an elastic principal graph [396].

4.3.15 Non-negative matrix factorisation

To assess sets of genes expressed in a similar way across genes an implementation of non-negative matrix factorisation, NNLM v0.4.3 [100] with initialisation adopted from [397].

Briefly integrated count data was used for this analysis (section 4.3.6) and all negative values set to zero. Genes that were not expressed in the data have been discarded for further analysis. The data is then natural log transformed and the expression matrix decomposed into two matrices with a chosen number of factors. One gene by factor matrix and a second factor by cell matrix. The influence of each factor on a given cell is then assessed by dividing each factor value of a cell through the sum of factor values in a cell from the second matrix. Accordingly, the influence of each gene on a given factor was calculated on the second matrix.

5 References

1. Harris, H., *The birth of the cell.* 2000: Yale University Press.
2. Marinov, G.K., et al., *From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing.* Genome Res, 2014. **24**(3): p. 496-510.
3. Ho, B., A. Baryshnikova, and G.W. Brown, *Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome.* Cell Syst, 2018. **6**(2): p. 192-205 e3.
4. Wang, Y. and N.E. Navin, *Advances and applications of single-cell sequencing technologies.* Mol Cell, 2015. **58**(4): p. 598-609.
5. Cheung, M.C., et al., *Intracellular protein and nucleic acid measured in eight cell types using deep-ultraviolet mass mapping.* Cytometry A, 2013. **83**(6): p. 540-51.
6. Mannack, L.V., S. Eising, and A. Rentmeister, *Current techniques for visualizing RNA in cells.* F1000Res, 2016. **5**.
7. Weil, T.T., R.M. Parton, and I. Davis, *Making the message clear: visualizing mRNA localization.* Trends Cell Biol, 2010. **20**(7): p. 380-90.
8. Ploeger, L.S., et al., *Fluorescent stains for quantification of DNA by confocal laser scanning microscopy in 3-D.* Biotech Histochem, 2008. **83**(2): p. 63-9.
9. Hoheisel, J.D., *Microarray technology: beyond transcript profiling and genotype analysis.* Nat Rev Genet, 2006. **7**(3): p. 200-10.
10. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.
11. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
12. Goodwin, S., J.D. McPherson, and W.R. McCombie, *Coming of age: ten years of next-generation sequencing technologies.* Nat Rev Genet, 2016. **17**(6): p. 333-51.
13. Collins, F.S., M. Morgan, and A. Patrinos, *The Human Genome Project: lessons from large-scale biology.* Science, 2003. **300**(5617): p. 286-90.
14. Li, L. and H. Clevers, *Coexistence of quiescent and active adult stem cells in mammals.* Science, 2010. **327**(5965): p. 542-5.
15. Huang, S., *Non-genetic heterogeneity of cells in development: more than just noise.* Development, 2009. **136**(23): p. 3853-62.
16. Brady, G., M. Barbara, and N.N. Iscove, *Representative in vitro cDNA amplification from individual hemopoietic cells and colonies.* Methods Mol Cell Biol, 1990. **2**(1): p. 17-25.
17. Eberwine, J., et al., *Analysis of gene expression in single live neurons.* Proc Natl Acad Sci U S A, 1992. **89**(7): p. 3010-4.
18. Klein, C.A., et al., *Combined transcriptome and genome analysis of single micrometastatic cells.* Nat Biotechnol, 2002. **20**(4): p. 387-92.

19. Kurimoto, K., et al., *An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis.* Nucleic Acids Res, 2006. **34**(5): p. e42.

20. Xie, D., et al., *Rewirable gene regulatory networks in the preimplantation embryonic development of three mammalian species.* Genome Res, 2010. **20**(6): p. 804-15.

21. Tietjen, I., et al., *Single-cell transcriptional analysis of neuronal progenitors.* Neuron, 2003. **38**(2): p. 161-75.

22. Tang, F., et al., *mRNA-Seq whole-transcriptome analysis of a single cell.* Nat Methods, 2009. **6**(5): p. 377-82.

23. Regev, A., et al., *The Human Cell Atlas.* Elife, 2017. **6**.

24. Muller, C.M., A. Vlachos, and T. Deller, *Calcium homeostasis of acutely denervated and lesioned dentate gyrus in organotypic entorhino-hippocampal co-cultures.* Cell Calcium, 2010. **47**(3): p. 242-52.

25. Mattei, D., et al., *Enzymatic Dissociation Induces Transcriptional and Proteotype Bias in Brain Cell Populations.* Int J Mol Sci, 2020. **21**(21).

26. O'Flanagan, C.H., et al., *Dissociation of solid tumor tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses.* Genome Biol, 2019. **20**(1): p. 210.

27. Gross, A., et al., *Technologies for Single-Cell Isolation.* Int J Mol Sci, 2015. **16**(8): p. 16897-919.

28. Fulwyler, M.J., *Electronic separation of biological cells by volume.* Science, 1965. **150**(3698): p. 910-1.

29. Dittrich, W. and W. Göhde, *Flow-through chamber for photometers to measure and count particles in a dispersion medium.* 1973.

30. Chao, M.P., J. Seita, and I.L. Weissman, *Establishment of a normal hematopoietic and leukemia stem cell hierarchy.* Cold Spring Harb Symp Quant Biol, 2008. **73**: p. 439-49.

31. Wu, A.R., et al., *Quantitative assessment of single-cell RNA-sequencing methods.* Nat Methods, 2014. **11**(1): p. 41-6.

32. Pollen, A.A., et al., *Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex.* Nat Biotechnol, 2014. **32**(10): p. 1053-8.

33. Klein, A.M., et al., *Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells.* Cell, 2015. **161**(5): p. 1187-1201.

34. Macosko, E.Z., et al., *Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets.* Cell, 2015. **161**(5): p. 1202-1214.

35. Gierahn, T.M., et al., *Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput.* Nat Methods, 2017. **14**(4): p. 395-398.

36. Goldstein, L.D., et al., *Massively parallel nanowell-based single-cell gene expression profiling.* BMC Genomics, 2017. **18**(1): p. 519.

37. Rosenberg, A.B., et al., *Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding.* Science, 2018. **360**(6385): p. 176-182.

38. Islam, S., et al., *Quantitative single-cell RNA-seq with unique molecular identifiers.* Nat Methods, 2014. **11**(2): p. 163-6.
39. Svensson, V., et al., *Power analysis of single-cell RNA-sequencing experiments.* Nat Methods, 2017. **14**(4): p. 381-387.
40. Zhu, Y.Y., et al., *Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction.* Biotechniques, 2001. **30**(4): p. 892-7.
41. Chenchik, A., *Generation and use of high-quality cDNA from small amounts of total RNA by SMART PCR.* Gene cloning and analysis by RT-PCR, 1998.
42. Kulpa, D., R. Topping, and A. Telesnitsky, *Determination of the site of first strand transfer during Moloney murine leukemia virus reverse transcription and identification of strand transfer-associated reverse transcriptase errors.* EMBO J, 1997. **16**(4): p. 856-65.
43. Eberwine, J. and T. Bartfai, *Single cell transcriptomics of hypothalamic warm sensitive neurons that control core body temperature and fever response Signaling asymmetry and an extension of chemical neuroanatomy.* Pharmacol Ther, 2011. **129**(3): p. 241-59.
44. Hashimshony, T., et al., *CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification.* Cell Rep, 2012. **2**(3): p. 666-73.
45. Navin, N., et al., *Tumour evolution inferred by single-cell sequencing.* Nature, 2011. **472**(7341): p. 90-4.
46. Zong, C., et al., *Genome-wide detection of single-nucleotide and copy-number variations of a single human cell.* Science, 2012. **338**(6114): p. 1622-6.
47. Greenleaf, W.J., *Assaying the epigenome in limited numbers of cells.* Methods, 2015. **72**: p. 51-6.
48. Cusanovich, D.A., et al., *Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing.* Science, 2015. **348**(6237): p. 910-4.
49. Buenrostro, J.D., et al., *Single-cell chromatin accessibility reveals principles of regulatory variation.* Nature, 2015. **523**(7561): p. 486-90.
50. Fan, H.C., G.K. Fu, and S.P. Fodor, *Combinatorial labeling of single cells for gene expression cytometry.* Science, 2015. **347**(6222): p. 1258367.
51. Macaulay, I.C., et al., *G&T-seq: parallel sequencing of single-cell genomes and transcriptomes.* Nat Methods, 2015. **12**(6): p. 519-22.
52. Golomb, S.M., et al., *Multi-modal Single-Cell Analysis Reveals Brain Immune Landscape Plasticity during Aging and Gut Microbiota Dysbiosis.* Cell Rep, 2020. **33**(9): p. 108438.
53. Fiskin, E., et al., *Single-cell multimodal profiling of proteins and chromatin accessibility using PHAGE-ATAC.* Nat Biotechnol, 2021.
54. Xing, Q.R., et al., *Parallel bimodal single-cell sequencing of transcriptome and chromatin accessibility.* Genome Res, 2020. **30**(7): p. 1027-1039.
55. Dey, S.S., et al., *Integrated genome and transcriptome sequencing of the same cell.* Nat Biotechnol, 2015. **33**(3): p. 285-289.

56. Angermueller, C., et al., *Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity.* Nat Methods, 2016. **13**(3): p. 229-232.

57. Cheow, L.F., et al., *Single-cell multimodal profiling reveals cellular epigenetic heterogeneity.* Nat Methods, 2016. **13**(10): p. 833-6.

58. Cao, J., et al., *Joint profiling of chromatin accessibility and gene expression in thousands of single cells.* Science, 2018. **361**(6409): p. 1380-1385.

59. Clark, S.J., et al., *scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells.* Nat Commun, 2018. **9**(1): p. 781.

60. Chen, S., B.B. Lake, and K. Zhang, *High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell.* Nat Biotechnol, 2019. **37**(12): p. 1452-1457.

61. Liu, L., et al., *Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity.* Nat Commun, 2019. **10**(1): p. 470.

62. Zhu, C., et al., *An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome.* Nat Struct Mol Biol, 2019. **26**(11): p. 1063-1070.

63. Andor, N., et al., *Joint single cell DNA-Seq and RNA-Seq of cancer reveals subclonal signatures of genomic instability and gene expression.* NAR Genom Bioinform, 2020. **2**(2): p. lqaa016.

64. Ewing, B., et al., *Base-calling of automated sequencer traces using phred. I. Accuracy assessment.* Genome Res, 1998. **8**(3): p. 175-85.

65. Jiang, L., et al., *Synthetic spike-in standards for RNA-seq experiments.* Genome Res, 2011. **21**(9): p. 1543-51.

66. Grun, D., L. Kester, and A. van Oudenaarden, *Validation of noise models for single-cell transcriptomics.* Nat Methods, 2014. **11**(6): p. 637-40.

67. Hicks, S.C., et al., *Missing data and technical variability in single-cell RNA-sequencing experiments.* Biostatistics, 2018. **19**(4): p. 562-578.

68. Bacher, R. and C. Kendziorski, *Design and computational analysis of single-cell RNA-sequencing experiments.* Genome Biol, 2016. **17**: p. 63.

69. Buettner, F., et al., *Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells.* Nat Biotechnol, 2015. **33**(2): p. 155-60.

70. Kharchenko, P.V., L. Silberstein, and D.T. Scadden, *Bayesian approach to single-cell differential expression analysis.* Nat Methods, 2014. **11**(7): p. 740-2.

71. Lytal, N., D. Ran, and L. An, *Normalization Methods on Single-Cell RNA-seq Data: An Empirical Survey.* Front Genet, 2020. **11**: p. 41.

72. Hafemeister, C. and R. Satija, *Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression.* Genome Biol, 2019. **20**(1): p. 296.

73. Bacher, R., et al., *SCnorm: robust normalization of single-cell RNA-seq data.* Nat Methods, 2017. **14**(6): p. 584-586.

74. Vallejos, C.A., J.C. Marioni, and S. Richardson, *BASiCS: Bayesian Analysis of Single-Cell Sequencing Data.* PLoS Comput Biol, 2015. **11**(6): p. e1004333.
75. Lun, A.T., K. Bach, and J.C. Marioni, *Pooling across cells to normalize single-cell RNA sequencing data with many zero counts.* Genome Biol, 2016. **17**: p. 75.
76. Risso, D., et al., *A general and flexible method for signal extraction from single-cell RNA-seq data.* Nat Commun, 2018. **9**(1): p. 284.
77. Lopez, R., et al., *Deep generative modeling for single-cell transcriptomics.* Nat Methods, 2018. **15**(12): p. 1053-1058.
78. Qiu, X., et al., *Single-cell mRNA quantification and differential analysis with Census.* Nat Methods, 2017. **14**(3): p. 309-315.
79. Zhang, H., et al., *A multitask clustering approach for single-cell RNA-seq analysis in Recessive Dystrophic Epidermolysis Bullosa.* PLoS Comput Biol, 2018. **14**(4): p. e1006053.
80. Polanski, K., et al., *BBKNN: fast batch alignment of single cell transcriptomes.* Bioinformatics, 2020. **36**(3): p. 964-965.
81. Lotfollahi, M., F.A. Wolf, and F.J. Theis, *scGen predicts single-cell perturbation responses.* Nat Methods, 2019. **16**(8): p. 715-721.
82. Welch, J.D., et al., *Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity.* Cell, 2019. **177**(7): p. 1873-1887 e17.
83. Barkas, N., et al., *Wiring together large single-cell RNA-seq sample collections.* bioRxiv, 2018: p. 460246.
84. Johnson, T.S., et al., *LAmbDA: label ambiguous domain adaptation dataset integration reduces batch effects and improves subtype detection.* Bioinformatics, 2019. **35**(22): p. 4696-4706.
85. Gao, X., et al., *ClusterMap: compare multiple single cell RNA-Seq datasets across different experimental conditions.* Bioinformatics, 2019. **35**(17): p. 3038-3045.
86. Johansen, N. and G. Quon, *scAlign: a tool for alignment, integration, and rare cell identification from scRNA-seq data.* Genome Biol, 2019. **20**(1): p. 166.
87. Haghverdi, L., et al., *Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors.* Nat Biotechnol, 2018. **36**(5): p. 421-427.
88. Stuart, T., et al., *Comprehensive Integration of Single-Cell Data.* Cell, 2019. **177**(7): p. 1888-1902 e21.
89. Butler, A., et al., *Integrating single-cell transcriptomic data across different conditions, technologies, and species.* Nat Biotechnol, 2018. **36**(5): p. 411-420.
90. Kiselev, V.Y., A. Yiu, and M. Hemberg, *scmap: projection of single-cell RNA-seq data across data sets.* Nat Methods, 2018. **15**(5): p. 359-362.
91. Shaham, U., et al., *Removal of batch effects using distribution-matching residual networks.* Bioinformatics, 2017. **33**(16): p. 2539-2546.

92. Korsunsky, I., et al., *Fast, sensitive and accurate integration of single-cell data with Harmony.* Nat Methods, 2019. **16**(12): p. 1289-1296.

93. Hie, B., B. Bryson, and B. Berger, *Efficient integration of heterogeneous single-cell transcriptomes using Scanorama.* Nat Biotechnol, 2019. **37**(6): p. 685-691.

94. Lin, Y., et al., *scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets.* Proc Natl Acad Sci U S A, 2019. **116**(20): p. 9775-9784.

95. Jolliffe, I.T., *Principal component analysis.* 2nd ed. Springer series in statistics. 2002, New York: Springer.

96. Waltman, L. and N.J. van Eck, *A smart local moving algorithm for large-scale modularity-based community detection.* The European Physical Journal B, 2013. **86**(11): p. 471.

97. Wilcoxon, F., *Individual Comparisons by Ranking Methods.* Biometrics Bulletin, 1945. **1**(6): p. 80-83.

98. McDavid, A., et al., *Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments.* Bioinformatics, 2013. **29**(4): p. 461-7.

99. Finak, G., et al., *MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data.* Genome Biol, 2015. **16**: p. 278.

100. Lin, X. and P.C. Boutros, *Optimization and expansion of non-negative matrix factorization.* BMC Bioinformatics, 2020. **21**(1): p. 7.

101. Saelens, W., et al., *A comparison of single-cell trajectory inference methods.* Nat Biotechnol, 2019. **37**(5): p. 547-554.

102. Chen, H., et al., *Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM.* Nat Commun, 2019. **10**(1): p. 1903.

103. Trapnell, C., et al., *The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells.* Nat Biotechnol, 2014. **32**(4): p. 381-386.

104. Wolf, F.A., et al., *PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells.* Genome Biol, 2019. **20**(1): p. 59.

105. Montoro, D.T., et al., *A revised airway epithelial hierarchy includes CFTR-expressing ionocytes.* Nature, 2018. **560**(7718): p. 319-324.

106. Kim, D., et al., *Targeted therapy guided by single-cell transcriptomic analysis in drug-induced hypersensitivity syndrome: a case report.* Nat Med, 2020. **26**(2): p. 236-243.

107. Sun, G., et al., *Single-cell RNA sequencing in cancer: Applications, advances, and emerging challenges.* Mol Ther Oncolytics, 2021. **21**: p. 183-206.

108. Bedard, P.L., et al., *Tumour heterogeneity in the clinic.* Nature, 2013. **501**(7467): p. 355-64.

109. Young, J., *Malpighi's "De Pulmonibus".* Proc R Soc Med, 1929. **23**(1): p. 1-11.

110. Gehr, P., M. Bachofen, and E.R. Weibel, *The normal human lung: ultrastructure and morphometric estimation of diffusion capacity.* Respir Physiol, 1978. **32**(2): p. 121-40.

111. Hermans, C. and A. Bernard, *Lung epithelium-specific proteins: characteristics and potential applications as markers.* Am J Respir Crit Care Med, 1999. **159**(2): p. 646-78.

112. Franks, T.J., et al., *Resident cellular components of the human lung: current knowledge and goals for research on cell phenotyping and function.* Proc Am Thorac Soc, 2008. **5**(7): p. 763-6.

113. Reyfman, P.A., et al., *Single-Cell Transcriptomic Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis.* Am J Respir Crit Care Med, 2019. **199**(12): p. 1517-1536.

114. Vieira Braga, F.A., et al., *A cellular census of human lungs identifies novel cell states in health and in asthma.* Nat Med, 2019. **25**(7): p. 1153-1163.

115. Treutlein, B., et al., *Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq.* Nature, 2014. **509**(7500): p. 371-5.

116. Travaglini, K.J., et al., *A molecular cell atlas of the human lung from single-cell RNA sequencing.* Nature, 2020. **587**(7835): p. 619-625.

117. Weibel, E.R., *Morphological basis of alveolar-capillary gas exchange.* Physiol Rev, 1973. **53**(2): p. 419-95.

118. Weber, E., et al., *Pulmonary lymphatic vessel morphology: a review.* Ann Anat, 2018. **218**: p. 110-117.

119. Lambrechts, D., et al., *Phenotype molding of stromal cells in the lung tumor microenvironment.* Nat Med, 2018. **24**(8): p. 1277-1289.

120. Weibel, E.R., *Lung morphometry: the link between structure and function.* Cell Tissue Res, 2017. **367**(3): p. 413-426.

121. Mercer, R.R., et al., *Cell number and distribution in human and rat airways.* Am J Respir Cell Mol Biol, 1994. **10**(6): p. 613-24.

122. Crystal, R.G., et al., *Airway epithelial cells: current concepts and challenges.* Proc Am Thorac Soc, 2008. **5**(7): p. 772-7.

123. Dean, C.H. and R.J. Snelgrove, *New Rules for Club Development: New Insights into Human Small Airway Epithelial Club Cell Ontogeny and Function.* Am J Respir Crit Care Med, 2018. **198**(11): p. 1355-1356.

124. Herzog, E.L., et al., *Knowns and unknowns of the alveolus.* Proc Am Thorac Soc, 2008. **5**(7): p. 778-82.

125. Williams, M.C., *Alveolar type I cells: molecular phenotype and development.* Annu Rev Physiol, 2003. **65**: p. 669-95.

126. Sarode, P., et al., *Epithelial cell plasticity defines heterogeneity in lung cancer.* Cell Signal, 2020. **65**: p. 109463.

127. Fehrenbach, H., *Alveolar epithelial type II cell: defender of the alveolus revisited.* Respir Res, 2001. **2**(1): p. 33-46.

128. Rock, J.R., S.H. Randell, and B.L. Hogan, *Airway basal stem cells: a perspective on their roles in epithelial homeostasis and remodeling.* Dis Model Mech, 2010. **3**(9-10): p. 545-56.

129. Morrisey, E.E., *Basal Cells in Lung Development and Repair.* Dev Cell, 2018. **44**(6): p. 653-654.

130. Barkauskas, C.E., et al., *Type 2 alveolar cells are stem cells in adult lung.* J Clin Invest, 2013. **123**(7): p. 3025-36.

131. Chen, Z., et al., *Non-small-cell lung cancers: a heterogeneous set of diseases.* Nat Rev Cancer, 2014. **14**(8): p. 535-46.

132. Agostini, C., et al., *Pulmonary immune cells in health and disease: lymphocytes.* Eur Respir J, 1993. **6**(9): p. 1378-401.

133. Hewitt, R.J. and C.M. Lloyd, *Regulation of immune responses by the airway epithelial cell landscape.* Nat Rev Immunol, 2021. **21**(6): p. 347-362.

134. Guo, X., et al., *Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing.* Nat Med, 2018. **24**(7): p. 978-985.

135. Wu, F., et al., *A new coronavirus associated with human respiratory disease in China.* Nature, 2020. **579**(7798): p. 265-269.

136. Zhu, N., et al., *A Novel Coronavirus from Patients with Pneumonia in China, 2019.* N Engl J Med, 2020. **382**(8): p. 727-733.

137. Wang, C., et al., *A novel coronavirus outbreak of global health concern.* Lancet, 2020. **395**(10223): p. 470-473.

138. Wu, Y., et al., *SARS-CoV-2 is an appropriate name for the new coronavirus.* Lancet, 2020. **395**(10228): p. 949-950.

139. Lukassen, S., et al., *SARS-CoV-2 receptor ACE2 and TMPRSS2 are primarily expressed in bronchial transient secretory cells.* EMBO J, 2020. **39**(10): p. e105114.

140. Heald-Sargent, T. and T. Gallagher, *Ready, set, fuse! The coronavirus spike protein and acquisition of fusion competence.* Viruses, 2012. **4**(4): p. 557-80.

141. Millet, J.K. and G.R. Whittaker, *Host cell proteases: Critical determinants of coronavirus tropism and pathogenesis.* Virus Res, 2015. **202**: p. 120-34.

142. Hoffmann, M., H. Hofmann-Winkler, and S. Pöhlmann, *Priming Time: How Cellular Proteases Arm Coronavirus Spike Proteins.* Activation of Viruses by Host Proteases, 2018: p. 71-98.

143. Klenk, H.D. and W. Garten, *Host cell proteases controlling virus pathogenicity.* Trends Microbiol, 1994. **2**(2): p. 39-43.

144. Hoffmann, M., et al., *SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor.* Cell, 2020. **181**(2): p. 271-280 e8.

145. Walls, A.C., et al., *Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein.* Cell, 2020. **181**(2): p. 281-292 e6.

146. Wang, H., et al., *SARS coronavirus entry into host cells through a novel clathrin- and caveolae-independent endocytic pathway.* Cell Res, 2008. **18**(2): p. 290-301.

147. Sung, H., et al., *Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA Cancer J Clin, 2021. **71**(3): p. 209-249.

148. Islami, F., et al., *Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the United States.* CA Cancer J Clin, 2018. **68**(1): p. 31-54.
149. Youlden, D.R., S.M. Cramb, and P.D. Baade, *The International Epidemiology of Lung Cancer: geographical distribution and secular trends.* J Thorac Oncol, 2008. **3**(8): p. 819-31.
150. Torre, L.A., R.L. Siegel, and A. Jemal, *Lung Cancer Statistics.* Adv Exp Med Biol, 2016. **893**: p. 1-19.
151. Campisi, J., *Aging, cellular senescence, and cancer.* Annu Rev Physiol, 2013. **75**: p. 685-705.
152. Alberg, A.J., et al., *Epidemiology of lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines.* Chest, 2013. **143**(5 Suppl): p. e1S-e29S.
153. Alberg, A.J., et al., *Respiratory cancer and exposure to arsenic, chromium, nickel, and polycyclic aromatic hydrocarbons.* Clinics in Occupational and Environmental Medicine, 2002. **2**(4): p. 779-801.
154. Turner, M.C., et al., *Long-term ambient fine particulate matter air pollution and lung cancer in a large cohort of never-smokers.* Am J Respir Crit Care Med, 2011. **184**(12): p. 1374-81.
155. Ris, C., *U.S. EPA health assessment for diesel engine exhaust: a review.* Inhal Toxicol, 2007. **19 Suppl 1**: p. 229-39.
156. Olsson, A.C., et al., *Exposure to diesel motor exhaust and lung cancer risk in a pooled analysis from case-control studies in Europe and Canada.* Am J Respir Crit Care Med, 2011. **183**(7): p. 941-8.
157. Samet, J.M., et al., *Lung cancer mortality and exposure to radon progeny in a cohort of New Mexico underground uranium miners.* Health Phys, 1991. **61**(6): p. 745-52.
158. Rubino, C., et al., *Radiation dose, chemotherapy and risk of lung cancer after breast cancer treatment.* Breast Cancer Res Treat, 2002. **75**(1): p. 15-24.
159. Inskip, P.D., M. Stovall, and J.T. Flannery, *Lung cancer risk and radiation dose among women treated for breast cancer.* J Natl Cancer Inst, 1994. **86**(13): p. 983-8.
160. Gomes, M., et al., *The role of inflammation in lung cancer.* Adv Exp Med Biol, 2014. **816**: p. 1-23.
161. Lundback, B., et al., *Not 15 but 50% of smokers develop COPD?--Report from the Obstructive Lung Disease in Northern Sweden Studies.* Respir Med, 2003. **97**(2): p. 115-22.
162. Young, R.P., et al., *COPD prevalence is increased in lung cancer, independent of age, sex and smoking history.* Eur Respir J, 2009. **34**(2): p. 380-6.
163. Zheng, M., *Classification and Pathology of Lung Cancer.* Surg Oncol Clin N Am, 2016. **25**(3): p. 447-68.

164. Travis, W.D., et al., *The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification.* J Thorac Oncol, 2015. **10**(9): p. 1243-1260.

165. Gazdar, A.F., P.A. Bunn, and J.D. Minna, *Small-cell lung cancer: what we know, what we need to know and the path forward.* Nat Rev Cancer, 2017. **17**(12): p. 725-737.

166. Herbst, R.S., D. Morgensztern, and C. Boshoff, *The biology and management of non-small cell lung cancer.* Nature, 2018. **553**(7689): p. 446-454.

167. Cook, R.M., Y.E. Miller, and P.A. Bunn, Jr., *Small cell lung cancer: etiology, biology, clinical features, staging, and treatment.* Curr Probl Cancer, 1993. **17**(2): p. 69-141.

168. Wakelee, H.A., et al., *Lung cancer incidence in never smokers.* J Clin Oncol, 2007. **25**(5): p. 472-8.

169. Organization, W.H., *WHO report on the global tobacco epidemic 2019: offer help to quit tobacco use.* 2019: World Health Organization.

170. Bilano, V., et al., *Global trends and projections for tobacco use, 1990-2025: an analysis of smoking indicators from the WHO Comprehensive Information Systems for Tobacco Control.* Lancet, 2015. **385**(9972): p. 966-76.

171. Forey, B. and P.N. Lee, *New edition of International Smoking Statistics.* Int J Epidemiol, 2007. **36**(2): p. 471-2.

172. Travis, W.D., et al., *International association for the study of lung cancer/american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma.* J Thorac Oncol, 2011. **6**(2): p. 244-85.

173. Lu, F., et al., *Identification of lung adenocarcinoma mutation status based on histologic subtype: Retrospective analysis of 269 patients.* Thorac Cancer, 2016. **7**(1): p. 17-23.

174. Bhatti, V., et al., *Histopathological Spectrum and Immunohistochemical Profile of Lung Carcinomas: A 9-Year Study from a Tertiary Hospital in North India.* Int J Appl Basic Med Res, 2019. **9**(3): p. 169-175.

175. Yoshimura, M., et al., *Molecular Pathology of Pulmonary Large Cell Neuroendocrine Carcinoma: Novel Concepts and Treatments.* Front Oncol, 2021. **11**: p. 671799.

176. Cha, M.J., et al., *Micropapillary and solid subtypes of invasive lung adenocarcinoma: clinical predictors of histopathology and outcome.* J Thorac Cardiovasc Surg, 2014. **147**(3): p. 921-928 e2.

177. Heist, R.S. and J.A. Engelman, *SnapShot: non-small cell lung cancer.* Cancer Cell, 2012. **21**(3): p. 448 e2.

178. Calvayrac, O., et al., *Molecular biomarkers for lung adenocarcinoma.* Eur Respir J, 2017. **49**(4).

179. Kan, Z., et al., *Diverse somatic mutation patterns and pathway alterations in human cancers.* Nature, 2010. **466**(7308): p. 869-73.

180. Motoi, N., et al., *Lung adenocarcinoma: modification of the 2004 WHO mixed subtype to include the major histologic subtype suggests correlations between*

papillary and micropapillary adenocarcinoma subtypes, EGFR mutations and gene expression analysis. Am J Surg Pathol, 2008. **32**(6): p. 810-27.

181. Yoshizawa, A., et al., *Validation of the IASLC/ATS/ERS lung adenocarcinoma classification for prognosis and association with EGFR and KRAS gene mutations: analysis of 440 Japanese patients.* J Thorac Oncol, 2013. **8**(1): p. 52-61.
182. Yoshida, A., et al., *Comprehensive histologic analysis of ALK-rearranged lung carcinomas.* Am J Surg Pathol, 2011. **35**(8): p. 1226-34.
183. Jokoji, R., et al., *Combination of morphological feature analysis and immunohistochemistry is useful for screening of EML4-ALK-positive lung adenocarcinoma.* J Clin Pathol, 2010. **63**(12): p. 1066-70.
184. Kadota, K., et al., *Associations between mutations and histologic patterns of mucin in lung adenocarcinoma: invasive mucinous pattern and extracellular mucin are associated with KRAS mutation.* Am J Surg Pathol, 2014. **38**(8): p. 1118-27.
185. Seo, J.S., et al., *The transcriptional landscape and mutational profile of lung adenocarcinoma.* Genome Res, 2012. **22**(11): p. 2109-19.
186. Tang, H., et al., *Comprehensive evaluation of published gene expression prognostic signatures for biomarker-based lung cancer clinical studies.* Ann Oncol, 2017. **28**(4): p. 733-740.
187. Xu, H., et al., *Gene expression profiling analysis of lung adenocarcinoma.* Braz J Med Biol Res, 2016. **49**(3).
188. Bang, M.S., et al., *Transcriptome analysis of non-small cell lung cancer and genetically matched adjacent normal tissues identifies novel prognostic marker genes.* Genes & Genomics, 2017. **39**(3): p. 277-284.
189. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.
190. Markovic, J., et al., *Genomic instability in patients with non-small cell lung cancer assessed by the arbitrarily primed polymerase chain reaction.* Cancer Invest, 2008. **26**(3): p. 262-8.
191. Qiu, Z.W., et al., *Genome-wide copy number variation pattern analysis and a classification signature for non-small cell lung cancer.* Genes Chromosomes Cancer, 2017. **56**(7): p. 559-569.
192. Deng, Z.M., et al., *Analysis of genomic variation in lung adenocarcinoma patients revealed the critical role of PI3K complex.* PeerJ, 2017. **5**: p. e3216.
193. Langevin, S.M., R.A. Kratzke, and K.T. Kelsey, *Epigenetics of lung cancer.* Transl Res, 2015. **165**(1): p. 74-90.
194. Hawes, S.E., et al., *DNA hypermethylation of tumors from non-small cell lung cancer (NSCLC) patients is associated with gender and histologic type.* Lung Cancer, 2010. **69**(2): p. 172-9.
195. Merlo, L.M., et al., *Cancer as an evolutionary and ecological process.* Nat Rev Cancer, 2006. **6**(12): p. 924-35.
196. de Bruin, E.C., et al., *Spatial and temporal diversity in genomic instability processes defines lung cancer evolution.* Science, 2014. **346**(6206): p. 251-6.

197. Zhang, J., et al., *Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing.* Science, 2014. **346**(6206): p. 256-9.

198. Jamal-Hanjani, M., et al., *Tracking the Evolution of Non-Small-Cell Lung Cancer.* N Engl J Med, 2017. **376**(22): p. 2109-2121.

199. Maynard, A., et al., *Therapy-Induced Evolution of Human Lung Cancer Revealed by Single-Cell RNA Sequencing.* Cell, 2020. **182**(5): p. 1232-1251 e22.

200. Kim, N., et al., *Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma.* Nat Commun, 2020. **11**(1): p. 2285.

201. Maman, S. and I.P. Witz, *A history of exploring cancer in context.* Nat Rev Cancer, 2018. **18**(6): p. 359-376.

202. Singh, N., et al., *Inflammation and cancer.* Ann Afr Med, 2019. **18**(3): p. 121-126.

203. Sinjab, A., et al., *Resolving the spatial and cellular architecture of lung adenocarcinoma by multiregion single-cell sequencing.* Cancer Discov, 2021. **11**(10): p. 2506-2523.

204. Lavin, Y., et al., *Innate Immune Landscape in Early Lung Adenocarcinoma by Paired Single-Cell Analyses.* Cell, 2017. **169**(4): p. 750-765 e17.

205. Chen, J., et al., *Single-cell transcriptome and antigen-immunoglobin analysis reveals the diversity of B cells in non-small cell lung cancer.* Genome Biol, 2020. **21**(1): p. 152.

206. Ma, K.Y., et al., *Single-cell RNA sequencing of lung adenocarcinoma reveals heterogeneity of immune response-related genes.* JCI Insight, 2019. **4**(4): p. e121387.

207. Waldman, A.D., J.M. Fritz, and M.J. Lenardo, *A guide to cancer immunotherapy: from T cell basic science to clinical practice.* Nat Rev Immunol, 2020. **20**(11): p. 651-668.

208. Li, B., et al., *Comprehensive analyses of tumor immunity: implications for cancer immunotherapy.* Genome Biol, 2016. **17**(1): p. 174.

209. Lim, S.M., M.H. Hong, and H.R. Kim, *Immunotherapy for Non-small Cell Lung Cancer: Current Landscape and Future Perspectives.* Immune Netw, 2020. **20**(1): p. e10.

210. Sousa, C., et al., *Diffuse smoking-related lung diseases: insights from a radiologic-pathologic correlation.* Insights Imaging, 2019. **10**(1): p. 73.

211. Ezzati, M. and A.D. Lopez, *Estimates of global mortality attributable to smoking in 2000.* Lancet, 2003. **362**(9387): p. 847-52.

212. Talhout, R., et al., *Hazardous compounds in tobacco smoke.* Int J Environ Res Public Health, 2011. **8**(2): p. 613-28.

213. Todisco, T., et al., *Normal reference values for regional pulmonary peripheral airspace epithelial permeability. Influence of pneumonectomy and the smoking habit.* Respiration, 1989. **55**(2): p. 84-93.

214. Wright, J.L., et al., *Airway inflammation and peribronchiolar attachments in the lungs of nonsmokers, current and ex-smokers.* Lung, 1988. **166**(5): p. 277-86.

215. Wanner, A., *A review of the effects of cigarette smoke on airway mucosal function.* Eur J Respir Dis Suppl, 1985. **139**: p. 49-53.

216. Lee, J., V. Taneja, and R. Vassallo, *Cigarette smoking and inflammation: cellular and molecular mechanisms.* J Dent Res, 2012. **91**(2): p. 142-9.

217. Mao, L., et al., *Clonal genetic alterations in the lungs of current and former smokers.* J Natl Cancer Inst, 1997. **89**(12): p. 857-62.

218. Wistuba, II, et al., *Molecular damage in the bronchial epithelium of current and former smokers.* J Natl Cancer Inst, 1997. **89**(18): p. 1366-73.

219. Khuder, S.A., *Effect of cigarette smoking on major histological types of lung cancer: a meta-analysis.* Lung Cancer, 2001. **31**(2-3): p. 139-48.

220. Samet, J.M., et al., *Lung cancer in never smokers: clinical epidemiology and environmental risk factors.* Clin Cancer Res, 2009. **15**(18): p. 5626-45.

221. Toh, C.-K. and W.-T. Lim, *Lung cancer in never-smokers.* Journal of Clinical Pathology, 2007. **60**(4): p. 337-340.

222. Kurahara, Y., et al., *Small-cell lung cancer in never-smokers: a case series with information on family history of cancer and environmental tobacco smoke.* Clin Lung Cancer, 2012. **13**(1): p. 75-9.

223. Govindan, R., et al., *Genomic landscape of non-small cell lung cancer in smokers and never-smokers.* Cell, 2012. **150**(6): p. 1121-34.

224. Rudin, C.M., et al., *Lung cancer in never smokers: molecular profiles and therapeutic implications.* Clin Cancer Res, 2009. **15**(18): p. 5646-61.

225. Bosse, Y., et al., *Molecular signature of smoking in human lung tissues.* Cancer Res, 2012. **72**(15): p. 3753-63.

226. Zhang, L., et al., *Impact of smoking cessation on global gene expression in the bronchial epithelium of chronic smokers.* Cancer Prev Res (Phila), 2008. **1**(2): p. 112-8.

227. Beane, J., et al., *Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression.* Genome Biol, 2007. **8**(9): p. R201.

228. Pintarelli, G., et al., *Cigarette smoke alters the transcriptome of non-involved lung tissue in lung adenocarcinoma patients.* Sci Rep, 2019. **9**(1): p. 13039.

229. Elisia, I., et al., *The effect of smoking on chronic inflammation, immune function and blood cell composition.* Sci Rep, 2020. **10**(1): p. 19480.

230. Harris, J.O., E.W. Swenson, and J.E. Johnson, 3rd, *Human alveolar macrophages: comparison of phagocytic ability, glucose utilization, and ultrastructure in smokers and nonsmokers.* J Clin Invest, 1970. **49**(11): p. 2086-96.

231. Pacht, E.R., et al., *Deficiency of vitamin E in the alveolar fluid of cigarette smokers. Influence on alveolar macrophage cytotoxicity.* J Clin Invest, 1986. **77**(3): p. 789-96.

232. Pons, A.R., et al., *Decreased macrophage release of TGF-beta and TIMP-1 in chronic obstructive pulmonary disease.* Eur Respir J, 2005. **26**(1): p. 60-6.

233. Wallace, W.A., M. Gillooly, and D. Lamb, *Intra-alveolar macrophage numbers in current smokers and non-smokers: a morphometric study of tissue sections.* Thorax, 1992. **47**(6): p. 437-40.

234. Cho, W.C.S., et al., *The role of inflammation in the pathogenesis of lung cancer.* Expert Opinion on Therapeutic Targets, 2011. **15**(9): p. 1127-1137.

235. Mantovani, A., et al., *Cancer-related inflammation.* Nature, 2008. **454**(7203): p. 436-44.

236. Houghton, A.M., M. Mouded, and S.D. Shapiro, *Common origins of lung cancer and COPD.* Nat Med, 2008. **14**(10): p. 1023-4.

237. Walser, T., et al., *Smoking and lung cancer: the role of inflammation.* Proc Am Thorac Soc, 2008. **5**(8): p. 811-5.

238. Weitzman, S.A. and L.I. Gordon, *Inflammation and cancer: role of phagocyte-generated oxidants in carcinogenesis.* Blood, 1990. **76**(4): p. 655-63.

239. Martos, S.N., et al., *Single-cell analyses identify dysfunctional CD16(+) CD8 T cells in smokers.* Cell Rep Med, 2020. **1**(4): p. 100054.

240. Duclos, G.E., et al., *Characterizing smoking-induced transcriptional heterogeneity in the human bronchial epithelium at single-cell resolution.* Sci Adv, 2019. **5**(12): p. eaaw3413.

241. Goldfarbmuren, K.C., et al., *Dissecting the cellular specificity of smoking effects and reconstructing lineages in the human airway epithelium.* Nat Commun, 2020. **11**(1): p. 2485.

242. Parkin, D.M., et al., *Global cancer statistics, 2002.* CA Cancer J Clin, 2005. **55**(2): p. 74-108.

243. Zang, E.A. and E.L. Wynder, *Differences in lung cancer risk between men and women: examination of the evidence.* J Natl Cancer Inst, 1996. **88**(3-4): p. 183-92.

244. Bain, C., et al., *Lung cancer rates in men and women with comparable histories of smoking.* J Natl Cancer Inst, 2004. **96**(11): p. 826-34.

245. Henschke, C.I. and O.S. Miettinen, *Women's susceptibility to tobacco carcinogens.* Lung Cancer, 2004. **43**(1): p. 1-5.

246. Skarin, A.T., et al., *Lung cancer in patients under age 40.* Lung Cancer, 2001. **32**(3): p. 255-64.

247. Subramanian, J., et al., *Distinctive characteristics of non-small cell lung cancer (NSCLC) in the young: a surveillance, epidemiology, and end results (SEER) analysis.* J Thorac Oncol, 2010. **5**(1): p. 23-8.

248. Blanco, M., et al., *Bronchogenic carcinoma in patients under 50 years old.* Clin Transl Oncol, 2009. **11**(5): p. 322-5.

249. Kuo, C.W., et al., *Non-small cell lung cancer in very young and very old patients.* Chest, 2000. **117**(2): p. 354-7.

250. Liu, N.S., et al., *Adenocarcinoma of the lung in young patients: the M. D. Anderson experience.* Cancer, 2000. **88**(8): p. 1837-41.

251. Green, L.S., et al., *Bronchogenic cancer in patients under 40 years old. The experience of a Latin American country.* Chest, 1993. **104**(5): p. 1477-81.

252. Ryberg, D., et al., *Different susceptibility to smoking-induced DNA damage among male and female lung cancer patients.* Cancer Res, 1994. **54**(22): p. 5801-3.
253. Mollerup, S., et al., *Sex differences in lung CYP1A1 expression and DNA adduct levels among lung cancer patients.* Cancer Res, 1999. **59**(14): p. 3317-20.
254. Wu, C.T., et al., *The significance of estrogen receptor beta in 301 surgically treated non-small cell lung cancers.* J Thorac Cardiovasc Surg, 2005. **130**(4): p. 979-86.
255. Dubey, S., J.M. Siegfried, and A.M. Traynor, *Non-small-cell lung cancer and breast carcinoma: chemotherapy and beyond.* Lancet Oncol, 2006. **7**(5): p. 416-24.
256. Stabile, L.P., et al., *Combined targeting of the estrogen receptor and the epidermal growth factor receptor in non-small cell lung cancer shows enhanced antiproliferative effects.* Cancer Res, 2005. **65**(4): p. 1459-70.
257. Mollerup, S., et al., *Expression of estrogen receptors alpha and beta in human lung tissue and cell lines.* Lung Cancer, 2002. **37**(2): p. 153-9.
258. Stabile, L.P., et al., *Human non-small cell lung tumors and cells derived from normal lung express both estrogen receptor alpha and beta and show biological responses to estrogen.* Cancer Res, 2002. **62**(7): p. 2141-50.
259. Paulus, J.K., et al., *Haplotypes of estrogen receptor-beta and risk of non-small cell lung cancer in women.* Lung Cancer, 2011. **71**(3): p. 258-63.
260. Liu, Y., et al., *Reproductive factors, hormone use and the risk of lung cancer among middle-aged never-smoking Japanese women: a large-scale population-based cohort study.* Int J Cancer, 2005. **117**(4): p. 662-6.
261. Bae, J.M. and E.H. Kim, *Hormonal Replacement Therapy and the Risk of Lung Cancer in Women: An Adaptive Meta-analysis of Cohort Studies.* J Prev Med Public Health, 2015. **48**(6): p. 280-6.
262. Schwartz, A.G., et al., *Reproductive factors, hormone use, estrogen receptor expression and risk of non small-cell lung cancer in women.* J Clin Oncol, 2007. **25**(36): p. 5785-92.
263. Trefzer, T., et al., *Intratumoural heterogeneity and immune modulation in lung adenocarcinoma of female smokers and never smokers.* bioRxiv, 2021: p. 2021.05.18.444603.
264. Velten, L., et al., *Human haematopoietic stem cell lineage commitment is a continuous process.* Nat Cell Biol, 2017. **19**(4): p. 271-281.
265. Zeisel, A., et al., *Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq.* Science, 2015. **347**(6226): p. 1138-42.
266. Marin, R., et al., *Convergent origination of a Drosophila-like dosage compensation mechanism in a reptile lineage.* Genome Res, 2017. **27**(12): p. 1974-1987.
267. Bakken, T.E., et al., *Single-nucleus and single-cell transcriptomes compared in matched cortical cell types.* PLoS One, 2018. **13**(12): p. e0209648.

268. Krishnaswami, S.R., et al., *Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons.* Nat Protoc, 2016. **11**(3): p. 499-524.

269. Chen, L., *A global comparison between nuclear and cytosolic transcriptomes reveals differential compartmentalization of alternative transcript isoforms.* Nucleic Acids Research, 2009. **38**(4): p. 1086-1097.

270. Orphanides, G. and D. Reinberg, *A unified theory of gene expression.* Cell, 2002. **108**(4): p. 439-51.

271. Abdelmoez, M.N., et al., *SINC-seq: correlation of transient gene expressions between nucleus and cytoplasm reflects single-cell physiology.* Genome Biol, 2018. **19**(1): p. 66.

272. Lake, B.B., et al., *A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA.* Sci Rep, 2017. **7**(1): p. 6031.

273. Habib, N., et al., *Massively parallel single-nucleus RNA-seq with DroNc-seq.* Nat Methods, 2017. **14**(10): p. 955-958.

274. Tosti, L., et al., *Single-Nucleus and In Situ RNA-Sequencing Reveal Cell Topographies in the Human Pancreas.* Gastroenterology, 2021. **160**(4): p. 1330-1344 e11.

275. Ewing, B. and P. Green, *Base-calling of automated sequencer traces using phred. II. Error probabilities.* Genome Res, 1998. **8**(3): p. 186-94.

276. Mereu, E., et al., *Benchmarking single-cell RNA-sequencing protocols for cell atlas projects.* Nat Biotechnol, 2020. **38**(6): p. 747-755.

277. Loo, L., et al., *Single-cell transcriptomic analysis of mouse neocortical development.* Nat Commun, 2019. **10**(1): p. 134.

278. Haque, A., et al., *A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications.* Genome Med, 2017. **9**(1): p. 75.

279. Cheng, C.Y. and D.D. Mruk, *The biology of spermatogenesis: the past, present and future.* Philos Trans R Soc Lond B Biol Sci, 2010. **365**(1546): p. 1459-63.

280. Sungnak, W., et al., *SARS-CoV-2 Entry Genes Are Most Highly Expressed in Nasal Goblet and Ciliated Cells within Human Airways.* ArXiv, 2020.

281. Vidricaire, G., J.B. Denault, and R. Leduc, *Characterization of a secreted form of human furin endoprotease.* Biochem Biophys Res Commun, 1993. **195**(2): p. 1011-8.

282. Brussow, H., *The Novel Coronavirus - A Snapshot of Current Knowledge.* Microb Biotechnol, 2020. **13**(3): p. 607-612.

283. Chen, N., et al., *Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study.* Lancet, 2020. **395**(10223): p. 507-513.

284. Huang, C., et al., *Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China.* Lancet, 2020. **395**(10223): p. 497-506.

285. Zhang, J.J., et al., *Clinical characteristics of 140 patients infected with SARS-CoV-2 in Wuhan, China.* Allergy, 2020. **75**(7): p. 1730-1741.

286. Büttner, M., et al., *scCODA: A Bayesian model for compositional single-cell data analysis.* bioRxiv, 2020: p. 2020.12.14.422688.

287. Altorki, N.K., et al., *The lung microenvironment: an important regulator of tumour growth and metastasis.* Nat Rev Cancer, 2019. **19**(1): p. 9-31.

288. Negrini, S., V.G. Gorgoulis, and T.D. Halazonetis, *Genomic instability--an evolving hallmark of cancer.* Nat Rev Mol Cell Biol, 2010. **11**(3): p. 220-8.

289. Stranger, B.E., et al., *Relative impact of nucleotide and copy number variation on gene expression phenotypes.* Science, 2007. **315**(5813): p. 848-53.

290. Kazdal, D., et al., *Prevalence of somatic mitochondrial mutations and spatial distribution of mitochondria in non-small cell lung cancer.* Br J Cancer, 2017. **117**(2): p. 220-226.

291. Kazdal, D., et al., *Spatial and Temporal Heterogeneity of Panel-Based Tumor Mutational Burden in Pulmonary Adenocarcinoma: Separating Biology From Technical Artifacts.* J Thorac Oncol, 2019. **14**(11): p. 1935-1947.

292. Xiao, G.Y., A. Mohanakrishnan, and S.L. Schmid, *Role for ERK1/2-dependent activation of FCHSD2 in cancer cell-selective regulation of clathrin-mediated endocytosis.* Proc Natl Acad Sci U S A, 2018. **115**(41): p. E9570-E9579.

293. Cellier, M., et al., *Human natural resistance-associated macrophage protein: cDNA cloning, chromosomal mapping, genomic organization, and tissue-specific expression.* J Exp Med, 1994. **180**(5): p. 1741-52.

294. Correa, M.A., et al., *Slc11a1 (Nramp-1) gene modulates immune-inflammation genes in macrophages during pristane-induced arthritis in mice.* Inflamm Res, 2017. **66**(11): p. 969-980.

295. Jain, A., S. Kaczanowska, and E. Davila, *IL-1 Receptor-Associated Kinase Signaling and Its Role in Inflammation, Cancer Progression, and Therapy Resistance.* Front Immunol, 2014. **5**: p. 553.

296. Singleton, D.C. and A.L. Harris, *Targeting the ATF4 pathway in cancer therapy.* Expert Opin Ther Targets, 2012. **16**(12): p. 1189-202.

297. Harding, H.P., et al., *An integrated stress response regulates amino acid metabolism and resistance to oxidative stress.* Mol Cell, 2003. **11**(3): p. 619-33.

298. Takeda, J., S. Seino, and G.I. Bell, *Human Oct3 gene family: cDNA sequences, alternative splicing, gene organization, chromosomal location, and expression at low levels in adult tissues.* Nucleic Acids Res, 1992. **20**(17): p. 4613-20.

299. Niwa, H., J. Miyazaki, and A.G. Smith, *Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells.* Nat Genet, 2000. **24**(4): p. 372-6.

300. Yu, W., et al., *MEF2 transcription factors promotes EMT and invasiveness of hepatocellular carcinoma through TGF-beta1 autoregulation circuitry.* Tumour Biol, 2014. **35**(11): p. 10943-51.

301. Busygina, V., et al., *Multiple endocrine neoplasia type 1 interacts with forkhead transcription factor CHES1 in DNA damage response.* Cancer Res, 2006. **66**(17): p. 8397-403.

302. Burrell, R.A., et al., *The causes and consequences of genetic heterogeneity in cancer evolution.* Nature, 2013. **501**(7467): p. 338-45.

303. Lee, J., S. Giordano, and J. Zhang, *Autophagy, mitochondria and oxidative stress: cross-talk and redox signalling.* Biochem J, 2012. **441**(2): p. 523-40.

304. Cheng, G., et al., *Identification of PLXDC1 and PLXDC2 as the transmembrane receptors for the multifunctional factor PEDF.* Elife, 2014. **3**: p. e05401.

305. Martinez-Marin, D., et al., *PEDF increases the tumoricidal activity of macrophages towards prostate cancer cells in vitro.* PLoS One, 2017. **12**(4): p. e0174968.

306. Atkinson, P.G. and C.H. Barton, *Ectopic expression of Nramp1 in COS-1 cells modulates iron accumulation.* FEBS Lett, 1998. **425**(2): p. 239-42.

307. La Fleur, L., et al., *Expression of scavenger receptor MARCO defines a targetable tumor-associated macrophage subset in non-small cell lung cancer.* Int J Cancer, 2018. **143**(7): p. 1741-1752.

308. Georgoudaki, A.M., et al., *Reprogramming Tumor-Associated Macrophages by Antibody Targeting Inhibits Cancer Progression and Metastasis.* Cell Rep, 2016. **15**(9): p. 2000-11.

309. Jiang, Z., et al., *Targeting the SLIT/ROBO pathway in tumor progression: molecular mechanisms and therapeutic perspectives.* Ther Adv Med Oncol, 2019. **11**: p. 1758835919855238.

310. Zhang, C., et al., *Effects of Slit3 silencing on the invasive ability of lung carcinoma A549 cells.* Oncol Rep, 2015. **34**(2): p. 952-60.

311. Nissen, N.I., M. Karsdal, and N. Willumsen, *Collagens and Cancer associated fibroblasts in the reactive stroma and its relation to Cancer biology.* J Exp Clin Cancer Res, 2019. **38**(1): p. 115.

312. Zou, X., et al., *Up-regulation of type I collagen during tumorigenesis of colorectal cancer revealed by quantitative proteomic analysis.* J Proteomics, 2013. **94**: p. 473-85.

313. Willumsen, N., et al., *Serum biomarkers reflecting specific tumor tissue remodeling processes are valuable diagnostic tools for lung cancer.* Cancer Med, 2014. **3**(5): p. 1136-45.

314. Hirai, K., et al., *The spread of human lung cancer cells on collagens and its inhibition by type III collagen.* Clin Exp Metastasis, 1991. **9**(6): p. 517-27.

315. Tang, X., et al., *The single-cell sequencing: new developments and medical applications.* Cell Biosci, 2019. **9**: p. 53.

316. Rinke, C., et al., *Insights into the phylogeny and coding potential of microbial dark matter.* Nature, 2013. **499**(7459): p. 431-7.

317. Tang, F., et al., *Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis.* Cell Stem Cell, 2010. **6**(5): p. 468-78.

318. Xue, Z., et al., *Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing.* Nature, 2013. **500**(7464): p. 593-7.
319. Gao, S., et al., *Tracing the temporal-spatial transcriptome landscapes of the human fetal digestive tract using single-cell RNA-sequencing.* Nat Cell Biol, 2018. **20**(6): p. 721-734.
320. Wagner, D.E., et al., *Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo.* Science, 2018. **360**(6392): p. 981-987.
321. Vento-Tormo, R., et al., *Single-cell reconstruction of the early maternal-fetal interface in humans.* Nature, 2018. **563**(7731): p. 347-353.
322. Wang, M., et al., *Single-Cell RNA Sequencing Analysis Reveals Sequential Cell Fate Transition during Human Spermatogenesis.* Cell Stem Cell, 2018. **23**(4): p. 599-614 e4.
323. Chen, Y., et al., *Single-cell RNA-seq uncovers dynamic processes and critical regulators in mouse spermatogenesis.* Cell Res, 2018. **28**(9): p. 879-896.
324. Papalexi, E. and R. Satija, *Single-cell RNA sequencing to explore immune cell heterogeneity.* Nat Rev Immunol, 2018. **18**(1): p. 35-45.
325. Crinier, A., et al., *High-Dimensional Single-Cell Analysis Identifies Organ-Specific Signatures and Conserved NK Cell Subsets in Humans and Mice.* Immunity, 2018. **49**(5): p. 971-986 e5.
326. Villani, A.C., et al., *Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors.* Science, 2017. **356**(6335).
327. Cahill, D.P., et al., *Genetic instability and darwinian selection in tumours.* Trends Cell Biol, 1999. **9**(12): p. M57-60.
328. Turke, A.B., et al., *Preexistence and clonal selection of MET amplification in EGFR mutant NSCLC.* Cancer Cell, 2010. **17**(1): p. 77-88.
329. Lawson, D.A., et al., *Tumour heterogeneity and metastasis at single-cell resolution.* Nat Cell Biol, 2018. **20**(12): p. 1349-1360.
330. Darmanis, S., et al., *Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma.* Cell Rep, 2017. **21**(5): p. 1399-1410.
331. Young, M.D., et al., *Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors.* Science, 2018. **361**(6402): p. 594-599.
332. Caswell, D.R. and C. Swanton, *The role of tumour heterogeneity and clonal cooperativity in metastasis, immune evasion and clinical outcome.* BMC Med, 2017. **15**(1): p. 133.
333. Lim, B., Y. Lin, and N. Navin, *Advancing Cancer Research and Medicine with Single-Cell Genomics.* Cancer Cell, 2020. **37**(4): p. 456-470.
334. Ziegenhain, C., et al., *Comparative Analysis of Single-Cell RNA Sequencing Methods.* Mol Cell, 2017. **65**(4): p. 631-643 e4.
335. Karlsson, K. and S. Linnarsson, *Single-cell mRNA isoform diversity in the mouse brain.* BMC Genomics, 2017. **18**(1): p. 126.
336. Shalek, A.K., et al., *Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells.* Nature, 2013. **498**(7453): p. 236-40.

337. Tian, L., et al., *Comprehensive characterization of single cell full-length isoforms in human and mouse with long-read sequencing.* bioRxiv, 2020: p. 2020.08.10.243543.

338. Hayashi, T., et al., *Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs.* Nat Commun, 2018. **9**(1): p. 619.

339. Yekelchyk, M., et al., *Mono- and multi-nucleated ventricular cardiomyocytes constitute a transcriptionally homogenous cell population.* Basic Res Cardiol, 2019. **114**(5): p. 36.

340. Tirier, S.M., et al., *Pheno-seq - linking visual features and gene expression in 3D cell culture systems.* Sci Rep, 2019. **9**(1): p. 12367.

341. Ding, J., et al., *Systematic comparison of single-cell and single-nucleus RNA-sequencing methods.* Nat Biotechnol, 2020. **38**(6): p. 737-746.

342. Sungnak, W., et al., *SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes.* Nat Med, 2020. **26**(5): p. 681-687.

343. Hamming, I., et al., *Tissue distribution of ACE2 protein, the functional receptor for SARS coronavirus. A first step in understanding SARS pathogenesis.* J Pathol, 2004. **203**(2): p. 631-7.

344. Lee, I.T., et al., *ACE2 localizes to the respiratory cilia and is not increased by ACE inhibitors or ARBs.* Nat Commun, 2020. **11**(1): p. 5453.

345. Muus, C., et al., *Single-cell meta-analysis of SARS-CoV-2 entry genes across tissues and demographics.* Nat Med, 2021. **27**(3): p. 546-559.

346. Borges do Nascimento, I.J., et al., *Novel Coronavirus Infection (COVID-19) in Humans: A Scoping Review and Meta-Analysis.* J Clin Med, 2020. **9**(4): p. 941.

347. Liu, Y., et al., *Clinical and biochemical indexes from 2019-nCoV infected patients linked to viral loads and lung injury.* Sci China Life Sci, 2020. **63**(3): p. 364-374.

348. Hopkinson, N.S., et al., *Current smoking and COVID-19 risk: results from a population symptom app in over 2.4 million people.* Thorax, 2021. **76**(7): p. 714-722.

349. Petersen, E., et al., *Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics.* Lancet Infect Dis, 2020. **20**(9): p. e238-e244.

350. Simmons, G., et al., *Characterization of severe acute respiratory syndrome-associated coronavirus (SARS-CoV) spike glycoprotein-mediated viral entry.* Proc Natl Acad Sci U S A, 2004. **101**(12): p. 4240-5.

351. Matsuyama, S., et al., *Protease-mediated enhancement of severe acute respiratory syndrome coronavirus infection.* Proc Natl Acad Sci U S A, 2005. **102**(35): p. 12543-7.

352. White, J.M. and G.R. Whittaker, *Fusion of Enveloped Viruses in Endosomes.* Traffic, 2016. **17**(6): p. 593-614.

353. Yan, R., et al., *Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2.* Science, 2020. **367**(6485): p. 1444-1448.

354. Zhou, P., et al., *A pneumonia outbreak associated with a new coronavirus of probable bat origin.* Nature, 2020. **579**(7798): p. 270-273.
355. Wu, C., et al., *Furin: A Potential Therapeutic Target for COVID-19.* iScience, 2020. **23**(10): p. 101642.
356. Coutard, B., et al., *The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade.* Antiviral Res, 2020. **176**: p. 104742.
357. Kirchdoerfer, R.N., et al., *Pre-fusion structure of a human coronavirus spike protein.* Nature, 2016. **531**(7592): p. 118-21.
358. Walls, A.C., et al., *Cryo-electron microscopy structure of a coronavirus spike glycoprotein trimer.* Nature, 2016. **531**(7592): p. 114-117.
359. Wrapp, D., et al., *Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation.* Science, 2020. **367**(6483): p. 1260-1263.
360. Johnson, B.A., et al., *Loss of furin cleavage site attenuates SARS-CoV-2 pathogenesis.* Nature, 2021. **591**(7849): p. 293-299.
361. Wrobel, A.G., et al., *SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects.* Nature Structural & Molecular Biology, 2020. **27**(8): p. 763-767.
362. Hoffmann, M., H. Kleine-Weber, and S. Pohlmann, *A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells.* Mol Cell, 2020. **78**(4): p. 779-784 e5.
363. Bestle, D., et al., *TMPRSS2 and furin are both essential for proteolytic activation of SARS-CoV-2 in human airway cells.* Life Sci Alliance, 2020. **3**(9): p. e202000786.
364. Lau, S.Y., et al., *Attenuated SARS-CoV-2 variants with deletions at the S1/S2 junction.* Emerg Microbes Infect, 2020. **9**(1): p. 837-842.
365. Klimstra, W.B., et al., *SARS-CoV-2 growth, furin-cleavage-site adaptation and neutralization using serum from acutely infected hospitalized COVID-19 patients.* J Gen Virol, 2020. **101**(11): p. 1156-1169.
366. Rowbotham, S.P. and C.F. Kim, *Diverse cells at the origin of lung adenocarcinoma.* Proc Natl Acad Sci U S A, 2014. **111**(13): p. 4745-6.
367. Beane, J., et al., *Characterizing the impact of smoking and lung cancer on the airway transcriptome using RNA-Seq.* Cancer Prev Res (Phila), 2011. **4**(6): p. 803-17.
368. Brasky, T.M., et al., *Non-steroidal anti-inflammatory drugs and small cell lung cancer risk in the VITAL study.* Lung Cancer, 2012. **77**(2): p. 260-4.
369. Zappavigna, S., et al., *Anti-Inflammatory Drugs as Anticancer Agents.* Int J Mol Sci, 2020. **21**(7): p. 2605.
370. Durham, A.L. and I.M. Adcock, *The relationship between COPD and lung cancer.* Lung Cancer, 2015. **90**(2): p. 121-7.
371. Fan, J., K. Slowikowski, and F. Zhang, *Single-cell transcriptomics in cancer: computational challenges and opportunities.* Exp Mol Med, 2020. **52**(9): p. 1452-1465.

372. Andor, N., et al., *Joint single cell DNA-seq and RNA-seq of gastric cancer cell lines reveals rules of in vitro evolution.* NAR Genom Bioinform, 2020. **2**(2): p. lqaa016.

373. Wu, F., et al., *Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer.* Nat Commun, 2021. **12**(1): p. 2540.

374. Dvorak, H.F., *Tumors: wounds that do not heal-redux.* Cancer Immunol Res, 2015. **3**(1): p. 1-11.

375. Schafer, M. and S. Werner, *Cancer as an overhealing wound: an old hypothesis revisited.* Nat Rev Mol Cell Biol, 2008. **9**(8): p. 628-38.

376. Dvorak, H.F., *Tumors: wounds that do not heal. Similarities between tumor stroma generation and wound healing.* N Engl J Med, 1986. **315**(26): p. 1650-9.

377. Virchow, R., *Aetiologie der neoplastischen Geschwulste/Pathogenie der neoplastischen Geschwulste.* 1863.

378. Rybinski, B., J. Franco-Barraza, and E. Cukierman, *The wound healing, chronic fibrosis, and cancer progression triad.* Physiol Genomics, 2014. **46**(7): p. 223-44.

379. Rosenfeldt, M.T. and K.M. Ryan, *The multiple roles of autophagy in cancer.* Carcinogenesis, 2011. **32**(7): p. 955-63.

380. Yun, C.W. and S.H. Lee, *The Roles of Autophagy in Cancer.* Int J Mol Sci, 2018. **19**(11).

381. Zilionis, R., et al., *Single-Cell Transcriptomics of Human and Mouse Lung Cancers Reveals Conserved Myeloid Populations across Individuals and Species.* Immunity, 2019. **50**(5): p. 1317-1334 e10.

382. Wickham, H., *ggplot2: Elegant Graphics for Data Analysis.* 2016: Springer-Verlag New York.

383. Gu, Z., R. Eils, and M. Schlesner, *Complex heatmaps reveal patterns and correlations in multidimensional genomic data.* Bioinformatics, 2016. **32**(18): p. 2847-9.

384. Parekh, S., et al., *zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs.* Gigascience, 2018. **7**(6): p. giy059.

385. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.

386. Liao, Y., G.K. Smyth, and W. Shi, *featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.* Bioinformatics, 2014. **30**(7): p. 923-30.

387. Blondel, V.D., et al., *Fast unfolding of communities in large networks.* Journal of Statistical Mechanics: Theory and Experiment, 2008. **2008**(10): p. P10008.

388. McInnes, L., J. Healy, and J. Melville *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.* 2018. arXiv:1802.03426.

389. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.

390. Yu, G., et al., *clusterProfiler: an R package for comparing biological themes among gene clusters.* OMICS, 2012. **16**(5): p. 284-7.

391. Patel, A.P., et al., *Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma.* Science, 2014. **344**(6190): p. 1396-401.

392. Aibar, S., et al., *SCENIC: single-cell regulatory network inference and clustering.* Nat Methods, 2017. **14**(11): p. 1083-1086.

393. Aerts lab. *Cis-target databases.* 2018 July 22, 2020]; Available from: https://resources.aertslab.org/cistarget/databases/homo_sapiens/hg19/refseq_r45/mc9nr/gene_based/.

394. Efremova, M., et al., *CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes.* Nat Protoc, 2020. **15**(4): p. 1484-1506.

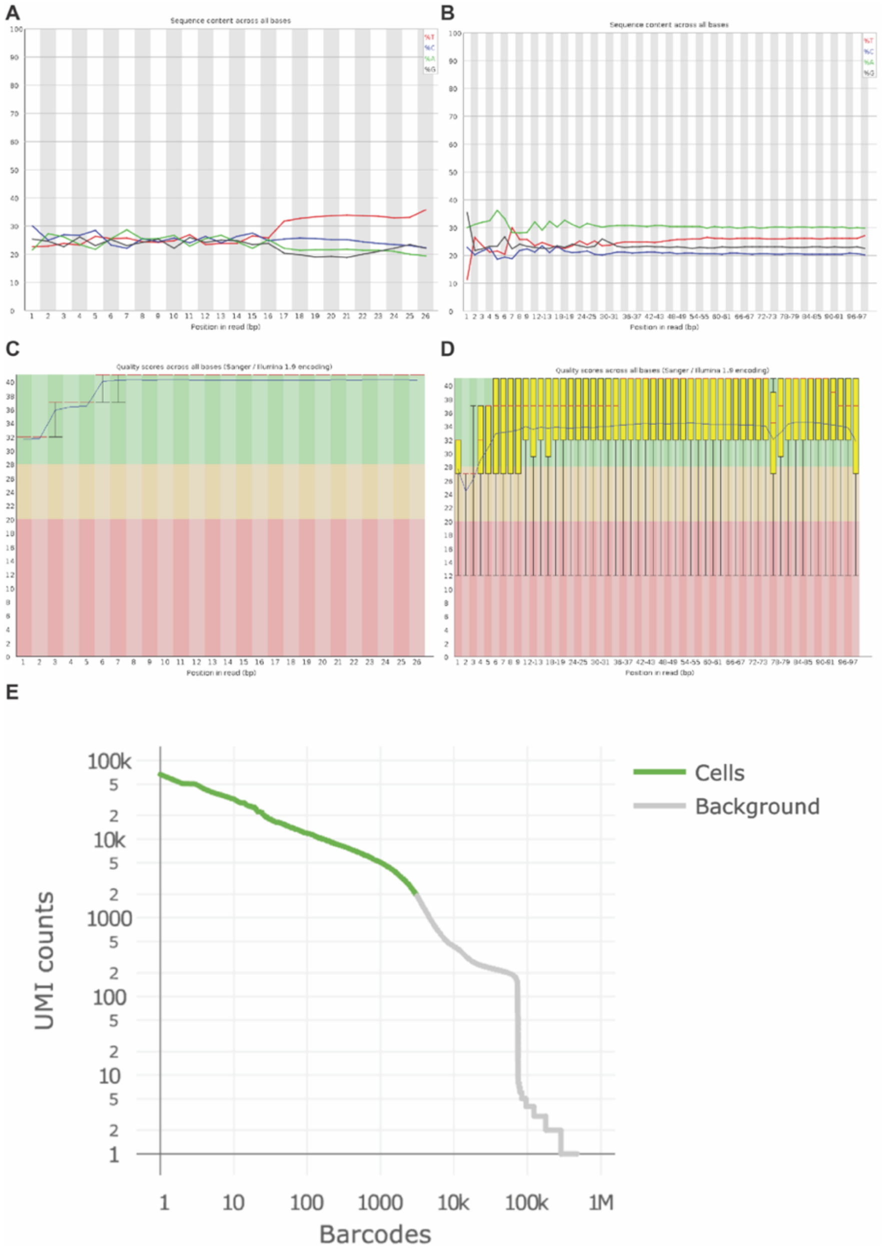
395. Frey, B.J. and D. Dueck, *Clustering by Passing Messages Between Data Points.* Science, 2007. **315**(5814): p. 972-976.

396. Albergante, L., et al., *Robust and Scalable Learning of Complex Intrinsic Dataset Geometry via ElPiGraph.* Entropy (Basel), 2020. **22**(3): p. 296.

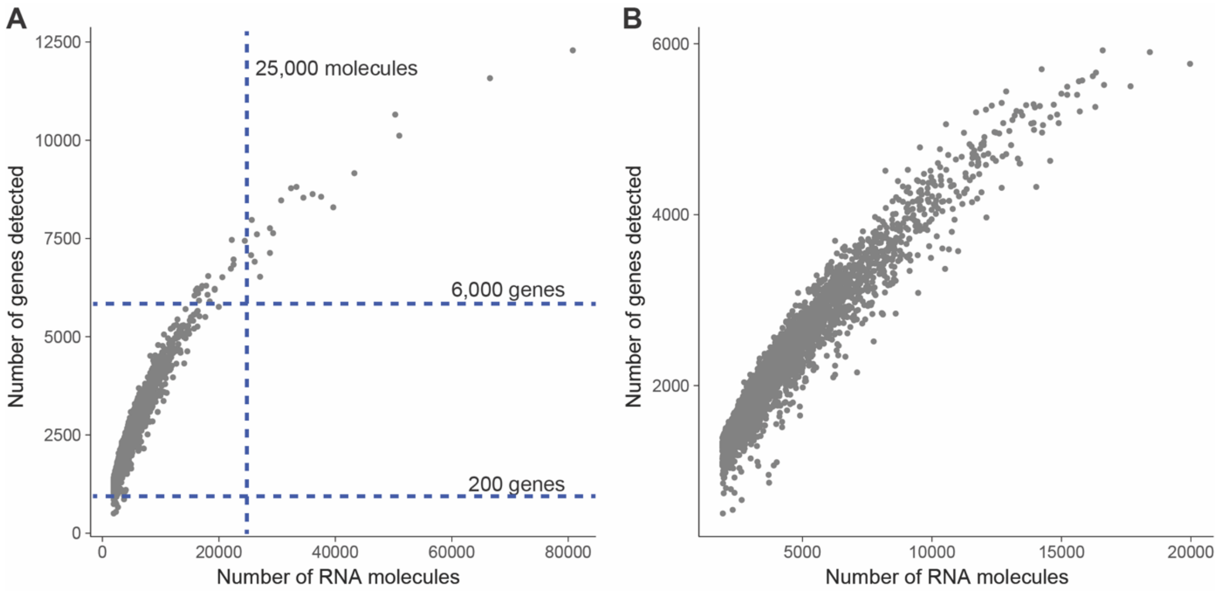
397. Wu, Y., P. Tamayo, and K. Zhang, *Visualizing and Interpreting Single-Cell Gene Expression Datasets with Similarity Weighted Nonnegative Embedding.* Cell Syst, 2018. **7**(6): p. 656-666 e4.

6 Appendix

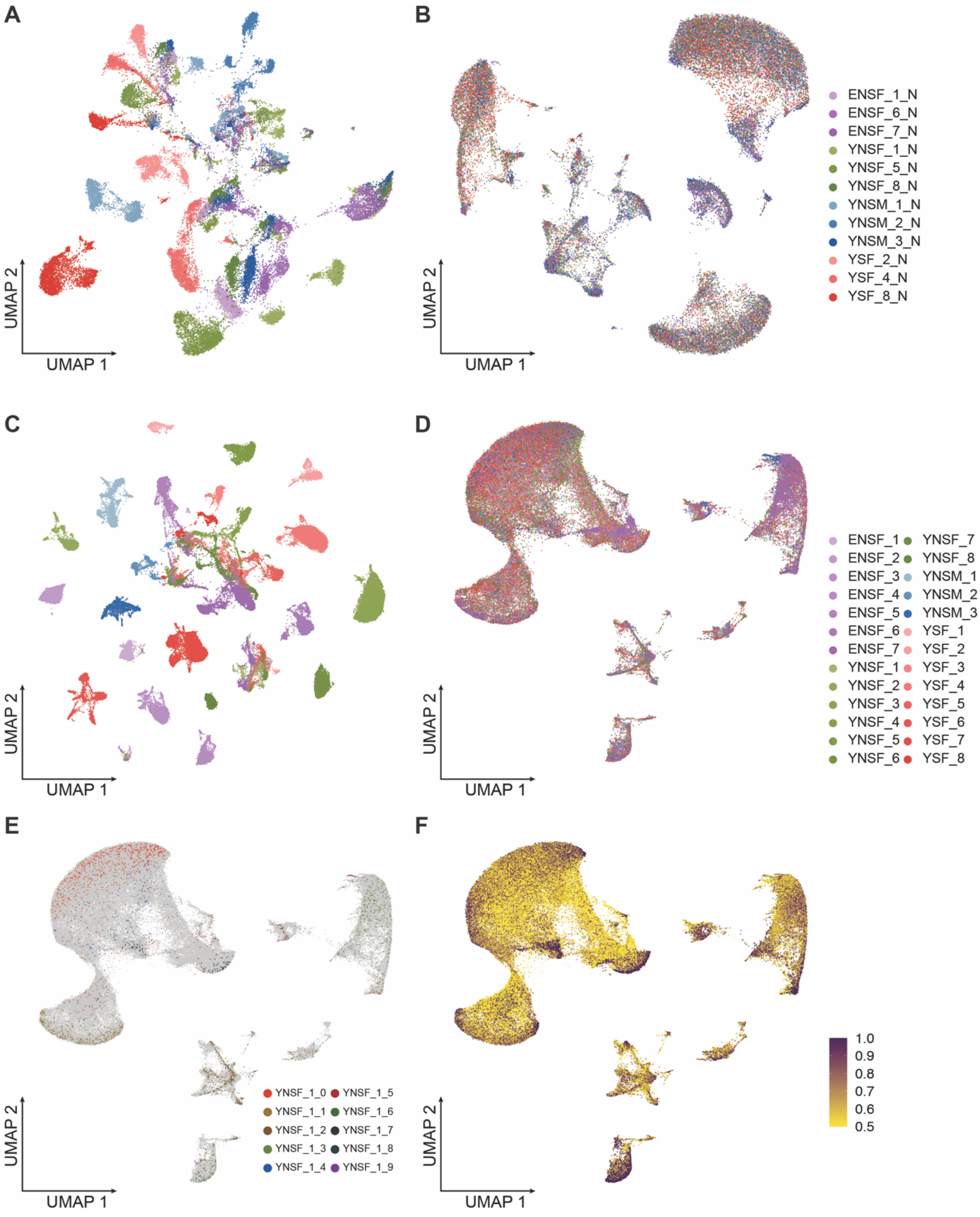
6.1 Supplementary Figures



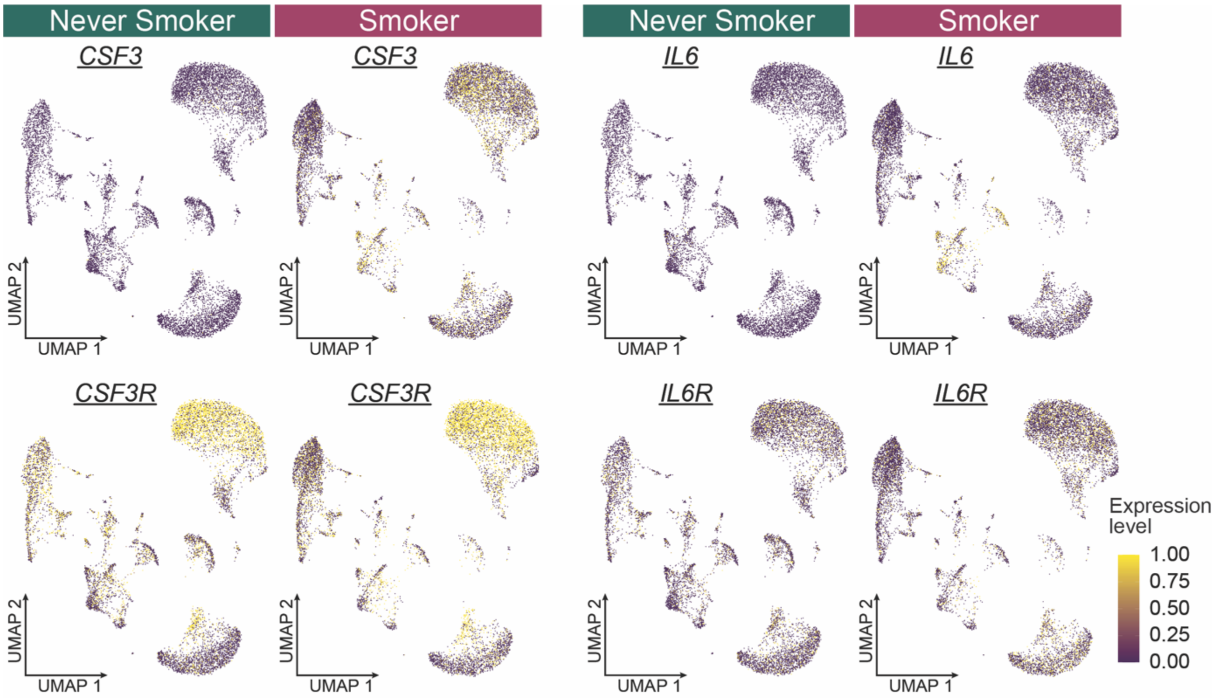
Supplementary Figure 1 Quality control for sequencing and alignment. After sequencing, read files are assessed for base content per position in **(A)** read 1, containing the cell barcode and UMI and **(B)** read 2, containing RNA sequence information. The base content should be comparable across the whole read length assuming an equal distribution of bases. Reads were also assessed using the phred quality score for identification of the base (compare section 4.3.2) for **(C)** read 1 and **(D)** read 2. **(E)** After alignment and generation of gene counts per cell, those data that correspond to intact cells are identified by the first infliction point of the ordered UMI count distribution (green), while the remainder are discarded as background (grey). Example data of sample ENSF_1 is shown.



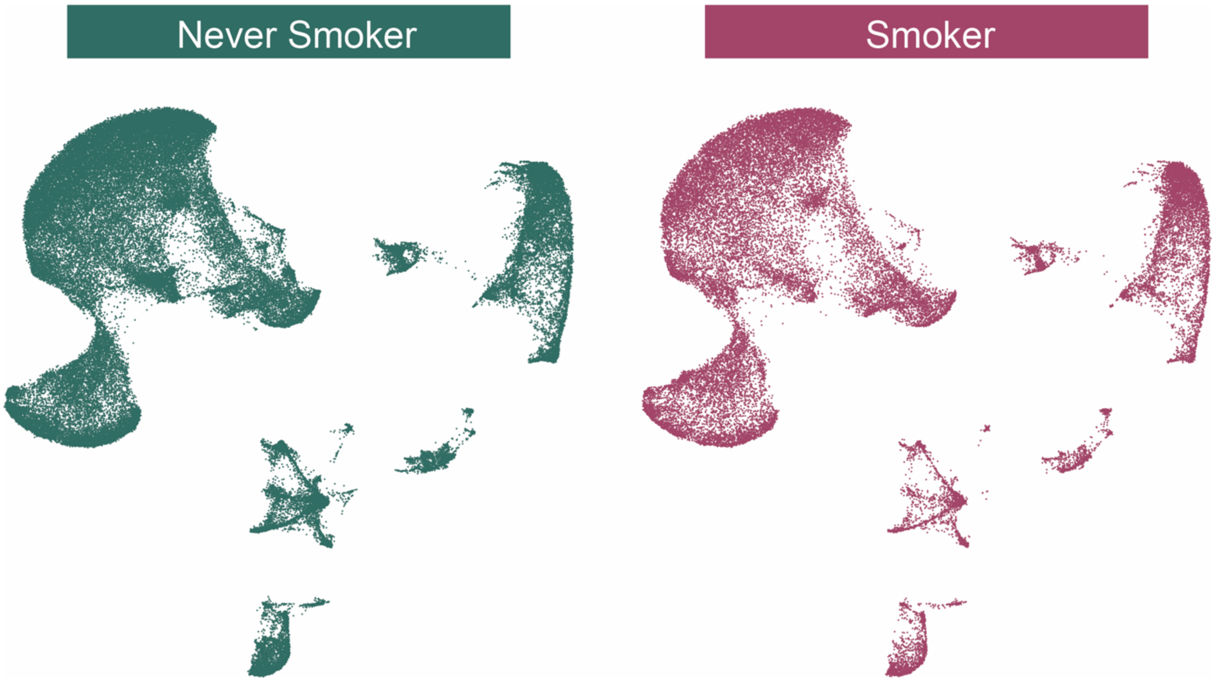
Supplementary Figure 2 Cell filtering based on read and gene count. To avoid analysis of partial cells, all cells with less than 200 detected genes are filtered out of the data set. To minimise the amount of multiplets in the data set, a maximum cut-off of detected genes (here 6,000) and detected molecules (here 25,000) was chosen for each experiment individually (compare section 4.3.2). Depicted are detected genes by number of RNA molecules per cell **(A)** before and **(B)** after filtering. Example data of sample ENSF_1 is shown.



Supplementary Figure 3 Integration of LADC data sets. scRNA-seq count data from healthy tissue samples of LADC patients **(A)** before and **(B)** after integration using MNN and CCA (section 4.3.6) and visualised by UMAP. scRNA-seq count data from tumour samples of LADC patients **(C)** before and **(D)** after integration using MNN and CCA (section 4.3.6) and visualised by UMAP. **(E)** Projection of clusters identified for one individual patient (YNSF_1) through separate analysis onto the integrated UMAP representation of all tumour samples, showing that sample integration does not affect cluster identity. **(F)** For cells originating from tumour tissue samples, cell types were assigned by comparing gene expression signatures to the healthy lung reference data. Depicted are confidence scores for the assignment (see section 4.3.7).



Supplementary Figure 4 Ligand and receptor expression. Ligand and receptor gene expression of selected chemokines across all non-malignant cells. UMAP representation is the same as in Figure **2.7**.



Supplementary Figure 5 UMAP by smoking habit. UMAP representation of cells from all smokers and never smokers separated by smoking habit.

6.2 Supplementary Tables

Table 1 QC filter parameters for technology evaluation samples

Pseudonym	minimal gene count	maximal gene count	maximal percent mitochondrial	maximal molecule count
Default	200	10000	0.1	10000
iCell8_1	200	10000	0.1	200000
iCell8_4	200	16000	0.1	1000000
iCell8_2	200	8000	0.1	60000
iCell8_6	200	12000	0.1	180000
iCell8_3	200	10000	0.1	130000
iCell8_5	200	16000	0.1	1000000
iCell8_7	200	14000	0.1	300000
Dolomite_1	200	3000	0.1	7500
Dolomite_2	200	3000	0.1	5000
Dolomite_3	400	5000	0.1	20000
Dolomite_4	200	800	0.1	1400
C1_1	200	100000	0.1	3000000
C1_2	200	100000	0.1	3000000
C1_3	200	100000	0.1	3000000

Table 2 Overview of experiments for technology evaluation

Experiment ID	scRNA-seq technology	Species	Isolation	Tissue	Cell Number
C1_1	C1	mouse	single cells	Forebrain	51
C1_2	C1	mouse	single cells	Forebrain	54
C1_3	C1	mouse	single cells	Forebrain	42
Chromium_1	10x	mouse	single nuclei	Testis	965
Dolomite_1	Dolomite	mouse	single nuclei	Testis	786
Dolomite_2	Dolomite	mouse	single nuclei	Testis	1351
Dolomite_3	Dolomite	mouse	single nuclei	Testis	3081
iCell8_1	iCell8	mouse	single cells	Forebrain	206
iCell8_2	iCell8	mouse	single cells	Forebrain	443
iCell8_3	iCell8	mouse	single cells	Forebrain	119
iCell8_4	iCell8	mouse	single cells	Testis	80
iCell8_5	iCell8	mouse	single cells	Testis	64
iCell8_6	iCell8	mouse	single nuclei	Testis	329
iCell8_7	iCell8	mouse	single nuclei	Testis	417

Table 3 Cohort description of LADC samples

Cohort Description		
Parameter	**n**	**(%)**
Median Age	52 (40–88)	
Total	**26**	**100**
Male	3	12
Female	23	88
Histology		
Adeno	26	100
Therapy		
OP	15	58
OP/RT	1	4
OP/ChT	9	34
OP/RT/ChT	1	4
Smoking status		
Never Smokers	18	69
Smokers	8	31
Pathological Stage (7th TNM edition)		
IA	2	8
IB	9	35
IIA	4	15
IIB	2	8
IIIA	7	27
IIIB	2	8
ECOG		
0	26	100

Table 4 LADC patient sample information

Pseudonym	Age	Sex	Packyears	Smoking Habit
ENSF_1	75	F	0	never smoker
ENSF_2	78	F	0	never smoker
ENSF_3	79	F	0	never smoker
ENSF_4	85	F	0	never smoker
ENSF_5	88	F	0	never smoker
ENSF_6	76	F	0	never smoker
ENSF_7	79	F	0	never smoker
YNSF_1	50	F	0	never smoker
YNSF_2	55	F	0	never smoker
YNSF_3	40	F	0	never smoker
YNSF_4	57	F	0	never smoker
YNSF_5	45	F	0	never smoker
YNSF_6	51	F	0	never smoker
YNSF_7	54	F	0	never smoker
YNSF_8	56	F	0	never smoker
YSF_1	52	F	40	smoker
YSF_2	47	F	45	smoker
YSF_3	51	F	30	smoker
YSF_4	45	F	30	smoker
YSF_5	46	F	35	smoker
YSF_6	52	F	80	smoker
YSF_7	53	F	100	smoker
YSF_8	44	F	40	smoker
YNSM_1	49	M	0	never smoker
YNSM_2	45	M	0	never smoker
YNSM_3	46	M	0	never smoker

Table 5 QC filter parameters for LADC samples

Pseudonym	minimal gene count	maximal gene count	maximal molecule count
ENSF_1	200	6000	25000
ENSF_1_N	200	6000	50000
ENSF_2	200	9000	30000
ENSF_3	200	10000	50000
ENSF_4	200	9000	40000
ENSF_5	200	7000	25000
ENSF_6	200	6000	20000
ENSF_6_N	200	7000	30000
ENSF_7	200	4000	12000
ENSF_7_N	200	6000	40000
YNSF_1	200	8000	30000
YNSF_1_N	200	6000	30000
YNSF_2	200	9000	40000
YNSF_3	200	10000	30000
YNSF_4	200	6000	30000
YNSF_5	200	8000	20000
YNSF_5_N	200	6000	30000
YNSF_6	200	6000	12000
YNSF_7	200	3000	7000
YNSF_8	200	7500	15000
YNSF_8_N	200	8000	40000
YNSM_1	200	8000	40000
YNSM_1_N	200	6000	30000
YNSM_2	200	6000	20000
YNSM_2_N	200	9000	70000
YNSM_3	200	9000	50000
YNSM_3_N	200	5000	15000
YSF_1	200	9000	90000
YSF_2	200	7500	30000
YSF_2_N	200	6500	50000
YSF_3	200	7500	30000
YSF_4	200	7000	25000
YSF_4_N	200	6000	20000
YSF_5	200	7500	25000
YSF_6	200	10000	20000
YSF_7	200	7000	30000
YSF_8	200	8000	30000
YSF_8_N	200	4000	10000

Table 6 GO signatures for malignant cell clusters. GO annotations as in Figure 2.18 with corresponding signature annotation.

IDs	GO_TERMS	Signature
1	ORGANELLE_FISSION	Mitotic
2	MITOTIC_NUCLEAR_DIVISION	Mitotic
3	CHROMOSOMAL_REGION	Mitotic
4	CONDENSED_CHROMOSOME	Mitotic
5	CHROMOSOME_CENTROMERIC_REGION	Mitotic
6	CELL_CYCLE_G2_M_PHASE_TRANSITION	Mitotic
7	REGULATION_OF_CELL_CYCLE_G2_M_PHASE_TRANSITION	Mitotic
8	SMALL_GTPASE_BINDING	Mitotic
9	ANAPHASE_PROMOTING_COMPLEX_DEPENDENT_CATABOLIC_PROCESS	Respiration
10	ATP_SYNTHESIS_COUPLED_ELECTRON_TRANSPORT	Respiration
11	RESPIRATORY_ELECTRON_TRANSPORT_CHAIN	Respiration
12	ATP_METABOLIC_PROCESS	Respiration
13	OXIDATIVE_PHOSPHORYLATION	Respiration
14	RESPIRASOME	Respiration
15	RESPIRATORY_CHAIN_COMPLEX	Respiration
16	PROTEIN_LOCALIZATION_TO_ENDOPLASMIC_RETICULUM	Transcription
17	ESTABLISHMENT_OF_PROTEIN_LOCALIZATION_TO_ENDOPLASMIC_RETICULUM	Transcription
18	COTRANSLATIONAL_PROTEIN_TARGETING_TO_MEMBRANE	Transcription
19	CYTOSOLIC_RIBOSOME	Transcription
20	NUCLEAR_TRANSCRIBED_MRNA_CATABOLIC_PROCESS_NONSENSE_MEDIATED_DECAY	Transcription
21	CELL_ADHESION_MEDIATED_BY_INTEGRIN	Adhesion
22	RESPONSE_TO_MOLECULE_OF_BACTERIAL_ORIGIN	Adhesion
23	PROTEIN_COMPLEX_INVOLVED_IN_CELL_ADHESION	Adhesion
24	INTEGRIN_BINDING	Adhesion
25	EXOGENOUS_PROTEIN_BINDING	Adhesion
26	ORGANIC_ACID_TRANSPORT	Metabolism
27	UNSATURATED_FATTY_ACID_METABOLIC_PROCESS	Metabolism
28	ICOSANOID_METABOLIC_PROCESS	Metabolism
29	PROSTANOID_METABOLIC_PROCESS	Metabolism
30	MONOCARBOXYLIC_ACID_TRANSPORT	Metabolism
31	NEURON_PROJECTION_ARBORIZATION	Morphology
32	ACTIN_BASED_CELL_PROJECTION	Morphology
33	CELL_LEADING_EDGE	Morphology
34	REGULATION_OF_CELL_MORPHOGENESIS	Morphology
35	SYNAPSE_ORGANIZATION	Morphology
36	ACTIN_BINDING	Morphology
37	PHOSPHOLIPID_BINDING	Phospholipid Binding
38	RESPONSE_TO_IMMOBILIZATION_STRESS	Phospholipid Binding
39	PHOSPHATIDYLINOSITOL_BINDING	Phospholipid Binding
40	MHC_CLASS_II_PROTEIN_COMPLEX	Immune Related
41	LEUKOCYTE_PROLIFERATION	Immune Related
42	REGULATION_OF_LEUKOCYTE_PROLIFERATION	Immune Related
43	T_CELL_PROLIFERATION	Immune Related
44	PHAGOCYTOSIS	Immune Related

Table 7 GO signatures for trajectory identity. GO annotations depicted in Figure 2.19.

Branch	ID	GO term
(S1,S0)	Sister chromatid segregation	GO_SISTER_CHROMATID_SEGREGATION
	Organelle fission	GO_ORGANELLE_FISSION
	Mitotic chromatid segregation	GO_MITOTIC_CHROMATID_SEGREGATION
	Mitotic nuclear division	GO_MITOTIC_NUCLEAR_DIVISION
	Chromosome segregation	GO_CHROMOSOME_SEGREGATION
(S0,S4)	GTPase regulation	GO_REGULATION_OF_GTPASE_ACTIVITY
	Positive GTPase regulation	GO_POSITIVE_REGULATION_OF_GTPASE_ACTIVITY
	Nucleoside triphosphate activity	GO_NUCLEOSIDE_ TRIPHOSPHATASE_REGULATORY_ACTIVITY
	GTPase regulator	GO_GTPASE_REGULATOR_ACTIVITY
	Epithelial cell proliferation	GO_EPITHELIAL_CELL_PROLIFERATION
(S0,S3)	Sarcolemma	GO_SARCOLEMMA
	Response to oxygen	GO_RESPONSE_TO_OXYGEN_LEVELS
	Wound healing	GO_REGULATION_OF_WOUND_HEALING
	Response to wounding	GO_REGULATION_OF_RESPONSE_TO_WOUNDING
	Protein autophosphorylation	GO_NEG_REGULATION_OF_PROTEIN_AUTOPHOSPHORYLATION
(S0,S2)	Multivesicular body	GO_MULTIVESICULAR_BODY
	Late endosome	GO_LATE_ENDOSOME
	Chemical homeostasis	GO_CHEMICAL_HOMEOSTASIS_WITHIN_A_TISSUE
	Autolysosome	GO_AUTOLYSOSOME
	Apical part of cell	GO_APICAL_PART_OF_CELL

Table 8 Software versions

Software	Version
CellPhoneDB	2.1.4
CellRanger	2.1.1
clusterProfiler	3.14.0
FastQC	0.11.9
inferCNV	1.2.0
NNLM	0.4.3
pySCENIC	0.10.3
Python	3.7.6
R	3.6.0
Rsubread	2.0.0
Seurat	3.1.3
STAR	2.7.3a
STREAM	1.0.0
zUMIs	2.9.4

Table 9 Reagents and materials

Name	Vendor	Catalogue Number
Agilent High Sensitivity D1000	Agilent Technologies	5067-4626
AMPure beads	Beckman Coulter	A63880
Beads, barcoded	Chemgenes	MACOSKO-2011-10
BSA	Sigma Aldrich	A8806
C-tubes	Milteny Biotech	130-093-237
C1 Reagent Kit for mRNA Seq	Fluidigm	100-6201
C1 Single-Cell Auto Prep IFC for mRNA Seq (5-10 µm)	Fluidigm	100-5759
Cell Strainer, 20 µm	pluriSelect Life Science	43-50020
Cell Strainer, 70 µm	pluriSelect Life Science	43-50070
Chromium i7 Multiplex Kit	10x Genomics	PN-120262
Chromium Single Cell 3’ Library & Gel Bead Kit v2	10x Genomics	PN-120237
Chromium Single Cell A Chip Kit	10x Genomics	PN-120236
Citric Acid Monohydrate	Sigma Aldrich	C1909-500G
Dissociation Kit, mouse tumor	Milteny Biotech	130-096-73
DNA Clean & Concentrator-5	Zymo Research	D4013
dNTPs 10 mM	Takara	4025
Dounce Homogeniser	Wheaton	T7482-1
Droplet Oil, EvaGreen	Bio-Rad	186-4006
DTT (1M)	AppliChem	A3668
Dyna Beads MyOne silane	Thermo Fisher Scientific	37002D
Earle’s Balanced Salt Solution	Thermo Fisher Scientific	10010
EDTA 0.5 M	Thermo Fisher Scientific	15575-20
Ethanol	Thermo Fisher Scientific	E/0600DF/C17
Exonuclease I enzyme and buffer	New England Biolab	M0293L
Ficoll PM-400 20%	Sigma-Aldrich	F5415
Glucose	Sigma-Aldrich	G7021-100G
Hanks Balanced Salt Solution	Thermo Fisher Scientific	14175095
Hoechst 33258	Thermo Fisher Scientific	H3569
iCell8 chip and reagent kit	WaferGen	430-000233
Isopropanol	Thermo Fisher Scientific	P/7500/PC17
Kapa HiFi Hotstart Readymix	Sigma-Aldrich	KK3605
Magnesium chloride, Rnase free (1 M MgCl2)	Thermo Fisher Scientific	AM9530
Maxima H- Rtase and 5x RT Buffer	Thermo Fisher Scientific	EP0753
N70X oligo	Illumina	FC-131-2001

Nextera XT DNA Library Preparation Kit	Illumina	FC-131-1024
NP-40 Surfact-Amps Detergent Solution	Thermo Fisher Scientific	28324
Nuclease free water	Thermo Fisher Scientific	AM9937
Papain Dissociation System	Worthington	LK003150
Perfluorooctanol	Sigma-Aldrich	370533
Phosphate Buffered Saline	Thermo Fisher Scientific	AM9625
Potassium Chloride, Rnase free (2 M KCl)	Thermo Fisher Scientific	AM9640G
ReadyProbes Cell Viability Imaging Kit	Thermo Fisher Scientific	R37610
RNase Inhibitor (RNAseIn) 40 U/µL-2,500 units	Thermo Fisher Scientific	AM2682
Round Bottom Polystyrene Test Tube 5 mL , with Cell Strainer Snap Cap	Falcon	352235
RPMI-1640	Gibco	21875034
Sarkosyl 20 %	Sigma-Aldrich	L7414
SDS (20% in H20)	Sigma-Aldrich	5030
SMARTer Ultra Low RNA Kit	Takara	634833
SMARTer® Ultra® Low RNA Kit for the Fluidigm® C1™ System	Clontech	634833
SPRIselect	Beckman Coulter	B23317
SSC, 20x	Thermo Fisher Scientific	15557036
Sucrose	Sigma-Aldrich	S9378-1KG
SUPERase•In™ Nase Inhibitor (20 U/µL) 10,000 units	Life Technologies	AM2696
Tris buffer, pH7.5; Rnase free (2M)	Sigma-Aldrich	T2944
Tris buffer, pH8; Rnase free (1M)	Thermo Fisher Scientific	AM9856
Trypan Blue Solution	Gibco	15250061
TWEEN-20	Sigma-Aldrich	P7949

Table 10 Primer sequences

Name	Sequence	Vendor
Primer TSO	AAGCAGTGGTATCAACGCAGAGTGAATrGrGrG	Exiqon
Primer SMART PCR	AAGCAGTGGTATCAACGCAGAGT	Exiqon
New-P5-SMART PCR oligo	AATGATACGGCGACCACCGAGATCTACACGC CTGTCCGCGGAAGCAGTGGTATCAACGCAGAGT*A*C (*locked nucleic acid)	Exiqon
Custom Read 1 primer	GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTAC	Exiqon

6.3 List of Figures

6.4 List of Tables

6.5 Abbreviations

AT1	alveolar type 1 cells
AT2	alveolar type 2 cells
AvM	alveolar macrophages
Bas	basal cells
BC	B cells
CCA	canonical correlation analysis
Cil	ciliated cells
CNV	copy number variation
COPD	chronical obstructive pulmonary disease
CPM	counts per million
DC	dendritic cells
DNA	deoxyribonucleic acid
EC	endothelial cells
ECM	extracellular matrix
EMT	epithelial to mesenchymal transition
FACS	fluorescence activated cell sorting
FC	fold change
Fib	fibroblasts
FISH	fluorescence in situ hybridisation
IHC	immunohistochemistry
knn	k-nearest neighbours
LADC	lung adenocarcioma
LE	lymphatic endothelial cells
MC	macrophages
MMLV	moloneymurine leukemia virus
MNN	mutual nearest neighbour
mRNA	messenger ribonucleic acid
MSigDB	molecular signature database
NeuN	neuroendocrine cells
NGS	next generation sequencing
NMF	non-negative matrix factorisation
NSCLC	non-small cell lung carcinoma
PC	plasma cells
PCA	principal component analysis
PCR	polymerase chain reaction
RNA	ribonucleic acid
SARS-CoV	severe acute respiratory syndrome corona virus
SCLC	small cell lung carcinoma
scRNA-seq	single cell RNA sequencing
Sec	secretory cells
SM	smooth muscle cells
SMART	Switching Mechanism at the 5′ end of RNA Template
snRNA-seq	single nucleus RNA sequencing
TC	T cells
TME	tumour microenvironment
UMAP	uniform manifold approximation and projection
UMI	unique molecular identifier

Units

°C	degree Celsius
ASR	age standardised rate
cm	centimeter
g	gravitational force equivalent
h	hours
M	molar
mg	milligram
min	minutes
ml	milliliter
mm	millimeter
mM	millimolar
ng	nanogram
nl	nanoliter
nm	nanometer
nt	nucleotide
pg	picogram
rpm	revolutions per minute
RT	room temperature
s	seconds
µg	microgram
µl	microliter
µm	micrometer

Gene names

ACE2	angiotensin-converting enzyme 2
AGR3	anterior gradient 3
ALK	anaplastic lymphoma receptor tyrosine kinase
ANXA1	annexin A1
ATF4	activating transcription factor 4
AURKB	aurora kinase B
BRAF	proto-oncogene B-Raf
C1QB	complement component 1, Q subcomponent, B chain
CD68	cluster of differentiation 68
CD86	cluster of differentiation 86
CLDN10	claudin 10
COX6B1	cytochrome C oxidase subunit 6B1
EGFR	epidermal growth factor receptor
eIF2α	eukaryotic translation initiation factor 2 alpha
EML4	echinoderm microtubule associated protein like 4
ERbeta	estrogen receptor beta
FCHSD2	FCH and double SH3 domains 2
FGF13	fibroblast growth factor 13
FNDC3B	fibronectin type III domain containing 3B
FOXN3	forkhead box N3
GAB2	growth factor receptor bound protein 2-associated protein 2
HLA-DQA2	human leukocyte antigen class II histocompatibility antigen, DQ Alpha 2
HLA-DRB5	human leukocyte antigen class II histocompatibility antigen, DR-5 beta
IGF2R	insulin like growth factor 2 receptor
IRAK3	interleukin 1 receptor associated kinase 3
IRS2	insulin receptor substrate 2
ITGB8	integrin subunit beta 8
KRAS	kirsten rat sarcoma viral oncogene homolog
MARCO	macrophage receptor with collagenous structure
MEF2A	myocyte enhancer factor 2A
MSLN	mesothelin
NDUFA4	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 4
PAEP	progestagen associated endometrial protein
PCDH7	protocadherin-7
POU4F1	POU class 4 homeobox 1
POU5F1	POU class 5 homeobox 1
PPARG	peroxisome proliferator-activated receptor gamma
RPL38	ribosomal protein L38
SLC11A1	solute carrier family 11 member 1
SLC16A14	solute carrier family 16 member 14
SPCS1	signal peptidase complex subunit 1
TMPRSS2	transmembrane serine protease 2
TOP2A	DNA topoisomerase II alpha
WFDC2	whey-acidic protein type disulfide core domain 2