https://doi.org/10.15388/vu.thesis.937
https://orcid.org/0000-0001-8175-504X

VILNIUS UNIVERSITY

Justina Žvirblytė

Single-Cell Transcriptomic Analysis of Healthy and Diseased Human Tissues

DOCTORAL DISSERTATION

Natural Sciences,
Biochemistry (N 004)

VILNIUS 2025

The dissertation was prepared between 2021 and 2025 at Vilnius University Life Sciences Center, Institute of Biotechnology.

Academic Supervisor – Prof. Dr. Linas Mažutis (Vilnius University, Natural Sciences, Biochemistry – N 004).
Academic Consultant – Dr. Rapolas Žilionis (Vilnius University, Natural Sciences, Biochemistry – N 004).

This doctoral dissertation will be defended in a public meeting of the Dissertation Defence Panel:

Chairwoman – Dr. Daiva Dabkevičienė (Vilnius University, Natural Sciences, Biochemistry – N 004).

Members:
Assoc. Prof. Dr. Johan Henriksson (Umeå University, Sweden, Natural Sciences, Biochemistry – N 004),
Assoc. Prof. Dr. Monika Mozerė (Vilnius University, Natural Sciences, Biochemistry – N 004),
Prof. Dr. Saulius Serva (Vilnius University, Natural Sciences, Biochemistry – N 004),
Prof. Dr. Kęstutis Sužiedėlis (Vilnius University, Natural Sciences, Biochemistry – N 004).

The dissertation shall be defended at a public meeting of the Dissertation Defence Panel at 2 pm on June 19th 2026, in Room R401 of the Life Sciences Center (Vilnius University).
Address: Saulėtekio av. 7, R401, Vilnius, Lithuania, LT-10257

https://doi.org/10.15388/vu.thesis.937
https://orcid.org/0000-0001-8175-504X

VILNIAUS UNIVERSITETAS

Justina Žvirblytė

Sveikų ir patologinių žmogaus audinių pavienių ląstelių transkriptomikos tyrimai

DAKTARO DISERTACIJA

Gamtos mokslai,
Biochemija (N 004)

VILNIUS 2025

Disertacija rengta 2021–2025 metais Vilniaus universiteto Gyvybės mokslų centro Biotechnologijos institute.

Mokslinis vadovas – **prof. dr. Linas Mažutis** (Vilniaus universitetas, gamtos mokslai, biochemija – N 004).
Mokslinis konsultantas – **dr. Rapolas Žilionis** (Vilniaus universitetas, gamtos mokslai, biochemija – N 004).

Gynimo taryba:
Pirmininkė – **dr. Daiva Dabkevičienė** (Vilniaus universitetas, gamtos mokslai, biochemija – N 004).

Nariai:
doc. dr. Johan Henriksson (Umea universitetas, Švedija, gamtos mokslai, biochemija – N 004),
doc. dr. Monika Mozerė (Vilniaus universitetas, gamtos mokslai, biochemija – N 004),
prof. dr. Saulius Serva (Vilniaus universitetas, gamtos mokslai, biochemija – N 004),
prof. dr. Kęstutis Sužiedėlis (Vilniaus universitetas, gamtos mokslai, biochemija – N 004).

Disertacija ginama viešame Gynimo tarybos posėdyje 2026 m. birželio mėn. 19 d. 14 val. Gyvybės mokslų centre (Vilniaus universitetas), R401 auditorijoje. Adresas: Saulėtekio al. 7, R401, Gyvybės mokslų centras, LT-10257, Vilnius, Lietuva.

CONTENTS

ABBREVIATIONS

AF – amniotic fluid
AFSC – amniotic fluid stem cells
AI – artificial intelligence
APC – antigen precenting cells
AVR – ascending *vasa recta*
BHB – barcoding hydrogel beads
CAF – cancer-associated fibroblast
ccRCC – clear cell renal cell carcinoma
DA – differential abundance
DEG – differentially expressed genes
DGE – differential gene expression
DVR – descending *vasa recta*
ECM – extracellular matrix
EMT – epithelial-mesenchymal transition
HPC – high-performance computing
HVG – highly variable genes
ICB – immune checkpoint blockade
lncRNA – long non-coding RNA
MSC – mesenchymal stem cells
PBMC - peripheral blood mononuclear cells
PCA – principal component analysis
PE – phycoerythrin
PTFE – polytetrafluoroethylene
QC – quality control
RBC – red blood cells
RT – reverse transcription
scRNA-seq – single-cell RNA sequencing
TAM – tumor associated macrophage
TF – transcription factor
TKI – tyrosine kinase inhibitors
TLS – tertiary lymphoid structures
TME – tumor microenvironment
TS – template switching
TV – tumor vasculature
UMAP - Uniform Manifold Approximation and Projection

INTRODUCTION

Human tissues display remarkable organizational and functional complexity, comprising a wide spectrum of specialized cellular phenotypes that elicit functions supporting organism-level homeostasis. Although most somatic cells carry the same genome, their distinct identities and roles arise from the differential use of it – a finely tuned, context-dependent pattern of gene regulation that results in a particular transcriptional profile. In disease, this balance is perturbed and cellular diversity is further amplified. For example, in the context of cancer, phenotypic expansion is evident not only in the emergence of malignant cell states, but in the reshaping of the entire tumor microenvironment (TME), driving heterogeneous responses of neighboring non-transformed or infiltrating cells. Another biological context representing vast cellular diversity is the development of an organism, as it requires plastic transition through various progenitor and intermediate cell states until proper structural and functional organization is achieved.

Gene expression analysis, also known as transcriptomics, is a widely used approach to study the cellular phenotypes in health and disease. However, population level measurements provide an average gene expression value and mask the diversity of cells within the sample. In order to investigate the inherent or disease-induced phenotypic heterogeneity, technologies that enable genome-wide gene expression profiling of single cells are necessary. A notable advancement came in 2009, when sequencing of single-cell mRNA was demonstrated for the first time (1). Even though this achievement laid the foundation of single-cell RNA sequencing (scRNA-seq) field, the initial technology was extremely labor-intensive, expensive, and would only allow profiling of several hundred to a couple thousand cells. Naturally, investigation of complex tissues harboring rare phenotypes requires scale, precision and cost-effectiveness to implement, and that became the focus of further technology development. In 2015, a major breakthrough in the field of scRNA-seq occurred with the establishment of microfluidics-driven droplet-based cell isolation and barcoding techniques. Two independent groups, side-by-side, published seminal droplet-based high-throughput scRNA-seq methods inDrops (2) and Drop-seq (3), dramatically increasing the scale and availability of single-cell profiling experiments. These innovations paved the way for a quick and widespread adoption of scRNA-seq technology seen worldwide today.

Single-cell RNA sequencing is an outstanding tool to efficiently probe biological systems that may harbor constituents lacking established markers for cellular characterization with conventional approaches such as immunocytochemistry, cytometry or qPCR. The power of genome-wide single-cell transcriptomic profiling was swiftly demonstrated by the discovery of new, previously uncharacterized cell types in the human body (4,5). Furthermore, efforts cataloguing cellular diversity in various healthy and disease-affected tissues led to advances in understanding complex biological systems, especially cancers (6–8). Solitary efforts soon converged into an ongoing global research initiative the "Human Cell Atlas" that aims to comprehensively catalogue every cell type across human lifespan in health and disease (9). In general, scRNA-seq has already made major contributions to our understanding of fundamental biological processes, and challenged the prevailing understanding of the "cell type" itself, revealing a vast variety of context-dependent cellular states and transitory continua (10).

The work presented in this thesis is centered around the investigation of cellular diversity in health and disease, using high-throughput single-cell transcriptomics. First, an improved version of droplet-based scRNA-seq method, termed inDrops-2, is presented by multiregional profiling of lung carcinoma tissue. Next, the developed method is applied to comprehensively characterize the cell heterogeneity in clear cell renal cell carcinoma (ccRCC) and healthy kidney tissues. Finally, taking an advantage of inDrops-2, this work concludes by reporting the first-ever comprehensive transcriptomic atlas of uncultured human amniotic fluid (AF) cells.

Study goal

Characterization of the transcriptional profile of single cells in human tissues (lung, kidney and amniotic fluid) using single cell RNA sequencing

Objectives:

- Validate inDrops-2 suitability for clinical sample profiling by analyzing methanol-fixed and preserved lung carcinoma specimens
- Compare the cellular composition of healthy kidney and clear cell renal cell carcinoma specimens
- Delineate the heterogeneity of endothelial and immune cells in ccRCC samples
- Construct an atlas of human amniotic fluid cells
- Characterize the phenotypes found in human amniotic fluid

Scientific novelty

Nowadays, commercial scRNA-seq systems, due to their ease-of-use, reproducibility and data quality are a primary choice for increasingly relevant and popular single-cell transcriptomic studies. However, such platforms lack flexibility and can become financially unfeasible. Open-source methods can accommodate the diverse needs of researchers for large-scale cell profiling at a lower cost and provide room for customization of the assays. In this work, the advantages of droplet-based open-source scRNA-seq platform inDrops-2 is showcased by profiling preserved, chemically hashed and multiplexed lung carcinoma specimens. Not only did this work validate the performance of the improved method, but, unexpectedly, revealed multiple rare, potentially clinically relevant phenotypes in lung carcinoma samples, that were overlooked in previous studies.

Profiling clear cell renal cell carcinoma with inDrops-2 method resulted in the generation of high-resolution transcriptomic atlas, yielding novel insights into the phenotypes present in the tumor microenvironment and their potential roles. Notably, this work introduces a tumor-associated endothelial tip cell phenotype, previously uncharacterized in the context of ccRCC. Considering that advanced and metastatic ccRCC treatment most commonly targets angiogenesis, detailed characterization of the tumor-associated endothelial compartment presented in this work is of particular relevance. The results presented in this thesis support the notion that tumor endothelial cells favor tumor progression through the expression of metastasis promoting factors, specific extracellular matrix components and indirectly via targetable interactions with immune cells in the TME. The novelty and relevance of the results presented in this work are already being proven by the high interest of the scientific community, significantly surpassing the impact metrics of the journal, as evidenced by rapidly increasing citation metrics and the active incorporation of the analyzed data in high-impact research.

Finally, the amniotic fluid atlas construction exemplifies the use of scRNA-seq technology as a discovery tool to probe uncharted territories and characterize biological systems with unknown phenotypes. Even though AF has been used in research and utilized as a source of stem cells for decades, the characterization of its cellular compartment is mostly limited to cell culture and cytometry using a predetermined set of markers. Due to challenges associated with limited cell count and sample availability, this biological system was also overlooked by large-scale efforts cataloguing cell types across tissues and conditions. The work presented in this thesis sheds light on the native cellular composition of the AF at two developmental timepoints,

post-conception week 16 and 20, and, for the first time, delineates the heterogeneity of fetal tissue cells floating within the AF. Notably, our results demonstrate the presence of diverse immune and non-immune cells, ranging from progenitor to highly specialized differentiated states. The non-immune compartment mostly consists of epithelial cells, likely shed from the fetal skin, intestinal tract, kidney and lungs. Interestingly, it also contains intermediate epithelial-mesenchymal phenotypes that do not seem to have published analogues in other biological systems. Moreover, diverse immune cells of fetal origin are detected, including both lymphoid and myeloid lineages, varying in abundance between the two developmental timepoints assessed. The last part of this work highlights the broad utility of scRNA-seq technology for gaining novel insights into cell diversity, presented as a unique atlas of fetal-derived cells in AF.

Defending statements

- scRNA-seq based on inDrops-2-TS approach is well-suited for profiling rare phenotypes in clinical samples
- Clear cell renal cell carcinoma (ccRCC) is highly infiltrated by T cells and tumor-associated macrophages
- Tumor endothelial cells support angiogenesis in ccRCC and comprise several phenotypes, including a tip cell-like population
- Human amniotic fluid harbors fetal immune, lung, intestine and kidney cells

1. LITERATURE REVIEW

1.1. The rise of single-cell transcriptomics

Biological systems are inherently heterogeneous and highly complex. Human tissues, in particular, harbor immense structural and cellular diversity, reflecting the functional specialization needed to support the physiological demands of the organism. As most somatic cells within an organism carry the same genome, such diversity stems from the differential use of it – a finely tuned regulation of gene expression that defines cell identity and function. In the context of various cellular pathologies, such as cancer, the homeostasis is disrupted and the cellular diversity is further amplified, as oncogenic transformation not only introduces novel phenotypes, but alters the environment, eliciting diverse responses of the non-transformed cells (6).

Gene expression analysis – transcriptomics – is a widely used approach to study the cellular phenotypes in health and disease. However, in order to decipher the inherent and pathology-imposed heterogeneity, single-cell profiling technologies are needed, as bulk population-level measurements provide limited information and mask the diversity. Moreover, to uncover the subtleties in gene expression programs, transcriptome-wide profiling is necessary. The transition from bulk, sample level analysis to high-resolution transcriptomics is not straightforward, as technical challenges are faced in throughput, labor, cost and the limiting amount of starting material. A single cell contains an estimated 1-50 pg of total RNA, and mRNA represents only 1-5% of this amount (11), which is not sufficient for sequencing library preparation. Thus, a single-cell transcriptomics method must include efficient single-cell compartmentalization, mRNA capturing, labeling and amplification strategies for sequencing library preparation, followed by bioinformatic analysis tailored for unbiased interpretation of highly dimensional data.

The first attempt to measure gene expression in single cells dates back to 1992, when Eberwine et al. measured the expression of several genes in live rat hippocampus neurons, using *in vivo* reverse transcription (RT) followed by *in vitro* transcription (IVT) and marker gene evaluation on a Southern blot (12). Later, single-cell mRNA or cDNA amplification was improved and combined with microarrays, substantially scaling up the number of cells and genes analyzed (13,14). Eventually, in 2009, Tang et al. paired these developments with sequencing, and demonstrated unbiased assessment of transcriptome-wide gene expression in a single cell for the first time (1). The first single cell RNA sequencing (scRNA-seq) experiments were done by

manually picking cells and placing them in separate tubes for RT followed by amplification, allowing to profile tens to hundreds of cells only. Stemming from the need for higher throughput and lower cost, tubes were replaced with microtiter plates and mRNA barcoding approach was developed (15). It greatly simplified the procedure: after RT, barcoded cDNA can be pooled and treated as a single sample for sequencing library preparation, and the transcripts are assigned to cells using barcode sequences during bioinformatic analysis. Further increase in experimental scale was achieved via automated cell isolation, utilizing integrated fluidic circuits (16), cell sorting and robotic liquid handling (17). Despite significant advancement, these systems were limited in the number of cells analyzed and high cost.

A major breakthrough in the field of scRNA-seq occurred in 2015, when two independent groups introduced droplet-based methods for cell isolation and barcoding: inDrops (2) and Drop-seq (3). Here, microfluidic devices are used to encapsulate individual cells together with barcode-carrying microspheres, cell lysis and RT reagents into nanoliter-scale aqueous droplets suspended in oil. Upon cell lysis within the droplets, mRNA is captured with poly-(dT) RT primers and barcoded during RT, enabling parallel processing of tens of thousands of cells. Subsequently, emulsion is broken and libraries are prepared on pooled material, simplifying the procedure and significantly reducing the cost. This innovation dramatically increased the throughput and laid the foundation for quick and widespread adoption of scRNA-seq technology. Today, commercialized by 10X Genomics, Chromium™ platform (18) (analogue to inDrops) represents the most popular choice for single-cell transcriptomics worldwide.

Droplet based scRNA-seq methods rely on Poisson statistics to co-encapsulate a single cell with a single barcoding microsphere, meaning that the barcode pool diversity needs to be substantially higher than the number of cells assayed, to avoid barcode duplication. To overcome this limitation and further increase the scale of cells assayed, combinatorial *in-situ* indexing strategy, initially developed for transposase-accessible chromatin sequencing (scATAC-seq) (19), was applied to scRNA-seq. Here, cells are not isolated individually, instead, sub-pools of several hundred fixed and permeabilized cells or nuclei are distributed across 96 or 384-well plates, harboring well-specific first barcode sequences. Then, after mRNA barcoding *in situ* during RT, the cells are pooled and randomly redistributed across a new multi-well plate, where a second barcode sequence is added to the cDNA via ligation (20) or PCR (21) reactions. Such *split-and-pool* procedure can be repeated until the barcode diversity is sufficient to avoid barcode collisions, making the method highly scalable – to the order of millions of cells. Additional advantage of this

approach is that no specialized microfluidics or cell sorting equipment is needed to achieve single-cell resolution. Commercialized by Parse Biosciences and Scale Biosciences, *split-and-pool* approach is gaining popularity, however, it has been shown to recover fewer input cells and lower number of genes and transcripts per cell, as compared to droplet-based methods (22).

The aforementioned technological advancements that enabled access to single-cell level information spurred the development of other single-cell -omic techniques. For instance, multiple methods have been published for single-cell genome sequencing (23,24), DNA methylation (25), histone modification (26), chromatin accessibility (19,27), chromatin conformation (28) or proteome profiling (29). Merging of these methodologies to extract a holistic understanding of the cell is an inevitable goal of single-cell analysis, thus, an area of active development is single-cell multi-omics. There are methods for simultaneous profiling of transcriptome and surface proteins (30,31), transcriptome and chromatin accessibility (32), transcriptome and DNA methylation (33), transcriptome and histone modifications (34), T and B cell receptor arrangements with transcriptome (35), among others (36). Moving forward, approaches to assess different combinations of more than two modalities are of high interest, such as simultaneous methylation, genome and transcriptome profiling (37–39) clonotype, surface protein, CRISPR perturbation and transcriptome assessment (40) and others (41). Evidently, single-cell transcriptomics remains at the cornerstone of multi-omic integration efforts.

The ability to profile hundreds of thousands of single cells in an unbiased fashion has immense potential not only to answer some of the most fundamental questions of cell, developmental or cancer biology, but also to generate hypotheses, fueling data-driven research. ScRNA-seq has already proven instrumental in the discovery of new cell types (4,5), as well as shaping the general understanding of cell phenotypes in the human body toward a plastic and intricate continuum, rather than a discrete set of stationary states (6,10). Empowered by single-cell transcriptomics, an ambitious global collaborative research initiative – the "Human Cell Atlas" – launched in 2016 (9,42). The goal of this consortium is to comprehensively catalogue all cell types across the human lifespan in health and disease, providing a foundation for understanding human biology and improving diagnosis, monitoring, and treatment of various pathologies. Today, more than 66 million cells from more than 10 thousand donors worldwide have been profiled as a part of this ongoing effort (43). Ultimately, upon multi-modal atlas integration into a foundation model, taking advantage of developments in machine learning and

artificial intelligence (AI), the “Human Cell Atlas” promises to transform translational research and impact clinical innovations (10,44).

1.1.1. Droplet-based scRNA-seq sample preparation

One of the key aspects of single-cell transcriptome profiling is the compartmentalization of the mRNA barcoding reaction. In the most commonly used droplet-based assays compartmentalization is achieved via encapsulation of single cells, barcoding entities and assay reagents into aqueous droplets. Thus, an important pre-requisite for any such study is efficient sample preparation yielding a suspension of high-quality single cells.

Sample preparation is one of the most important steps in single-cell transcriptomics that has a direct and immense effect on the quality of data obtained. Irrespective of tissue type, an optimal sample for droplet-based scRNA-seq should contain single cells of high viability (ideally >90%) without cellular debris or clumps, with as little handling-induced alterations in their gene expression profiles as possible (45,46). For cultured cells or liquid tissues, such as blood, cerebrospinal and amniotic fluids, the sample preparation is relatively straightforward, as the cells are already in suspension – a couple washes with media or resuspension buffer and optional red blood cell (RBC) lysis step is sufficient. However, for solid tissues, specific dissociation procedures must be employed. A typical tissue dissociation protocol involves mechanical disruption and enzymatic digestion to break down extracellular matrix (ECM) components, followed by washing and filtration steps to remove dead cells, debris, and cell clumps (45). Importantly, depending on tissue type, certain more fragile cell types (i.e. epithelial) might not survive dissociation procedure and there is a known bias toward immune cell capture in scRNA-seq protocols (47). Therefore, dissociation needs to be carefully optimized, including the selection of dissociation enzymes, times, temperatures, washing and centrifugation conditions, as well as buffers used (46,48). For example, it has been shown that the use of cold active proteases for dissociation can minimize the stress response signature expression (48,49). Furthermore, the presence of EDTA (above 0.1 mM) or divalent cations like Mg^{2+} and Ca^{2+} in resuspension buffers can hinder the efficiency of RT, leading to reduced cDNA output. To preserve RNA integrity during sample handling, it is essential to use RNase-free reagents and incorporate RNase inhibitors into the workflow (45). Additionally, handling times should be minimized to avoid transcriptional artefacts and cell death, as ambient RNA released from dying cells will get barcoded in each droplet, resulting in noisy expression profiles (discussed in section 1.2.2).

Depending on the study goals, it might be beneficial to enrich a certain subpopulation of cells in the dissociated samples. Fluorescence-activated cell sorting (FACS) and magnetic-activated cell sorting (MACS) are commonly employed for this task, for instance, to deplete dead cells or separately target immune (CD45+) or non-immune (CD45-) cell fractions. However, prolonged handling required by these methods might introduce stress artifacts (50) and result in cell loss, which is problematic when working with low numbers of cells.

The majority of scRNA-seq protocols require the processing of fresh, viable samples, which presents significant logistical and practical challenges, particularly in the context of clinical sample acquisition and handling. Additionally, certain sample types, such as brain, are virtually impossible to dissociate without damaging the membranes and releasing RNA. To overcome these issues, single-nucleus RNA-seq (snRNA-seq) can be employed, using nuclei extracted from fresh or frozen tissues as an input instead of cells. Profiling of the nuclear mRNA fraction was shown to faithfully recapitulate scRNA-seq data, proving it sufficient to distinguish cell types and delineate transcriptional programs (48,51). In certain tissues, nuclei profiling can even surpass scRNA-seq in detecting rare and fragile cell types (47,52). While nuclei profiling recovers similar cellular diversity, the cell type proportions are vastly different as compared to whole-cell transcriptomics (47,53). Moreover, snRNA-seq data is enriched in nascent, unspliced RNAs (generating intronic reads that are not always used in downstream analysis) and long non-coding RNA (lncRNA) (52,54). Therefore, while single-nucleus profiling can benefit analysis of fragile or biobanked tissues, limitations and biases should be taken into account.

Another attractive option, especially for clinical samples and longitudinal studies, is the use of long-term cell preservation techniques, such as cryopreservation or chemical fixation. Cryopreservation of cell suspensions in media supplemented with DMSO and high concentrations of FBS was shown to effectively preserve single-cell transcriptomes (55) and is compatible with droplet-based scRNA-seq (56). However, the freeze-thaw cycles may compromise cell viability, induce stress responses, as well as alter sample composition, leading to the loss of specific populations (i.e. epithelial cells) (48). Due to these issues, cryopreservation of samples for scRNA-seq is not widely adopted. Alternatively, several fixation techniques have been developed, with denaturing methanol fixation being the most widely adopted due to its compatibility with platforms like 10x Genomics and Drop-seq, its ability to preserve RNA integrity for extended periods, and its suitability for batching and transport (57,58). The protocol consists of resuspension of cells

in cold PBS followed by dropwise addition of ice-cold 100% methanol to a final concentration of 80%, and subsequent storage at –20°C or –80°C for up to several months, followed by a quick rehydration before encapsulation. Methanol fixation was shown to preserve cell population structure and yield good quality data for cell lines and mouse brain (57). However, successful fixation and, most importantly, rehydration are challenging for primary cells or cells with high protease and RNase content (i.e. immune cells). For these samples, to avoid RNA degradation, rehydration buffers must contain RNA protective agents. High concentration (3X-5X) saline sodium citrate (SSC) supplemented with RNase inhibitors and DTT was shown to be the buffer of choice for preserving RNA quality in primary cells, however, SSC must be diluted significantly (typically below 0.125X) upon encapsulation, as it is known to interfere with RT (58). Nonetheless, fixed cells might be prone to RNA leakage resulting in ambient RNA contamination (48), and cell handling time upon rehydration must be minimized.

Although methanol fixation has proven effective for cell suspensions, immediate tissue dissociation is not always feasible at the time of collection. Consequently, there is considerable interest in alternative fixation protocols that preserve intact tissues while remaining compatible with subsequent dissociation and single-cell analysis. Recently, a protocol using Lomant's Reagent (dithiobis(succinimidyl propionate) (DSP)), a reversible crosslinker fixative, was shown to preserve RNA integrity, library complexity, and cellular composition, while diminishing stress-related artifacts, in solid mouse and human tissues (59). DSP fixation is reversible with reducing agents such as DTT, and does not interfere with subsequent enzymatic digestion of tissues, making it an attractive option for longitudinal study designs. Nonetheless, this preservation technique, termed "FixNCut", is very recent and its broader applicability remains to be validated beyond the original study (59).

1.1.2. Droplet-based single-cell barcoding and sequencing

Droplet-based scRNA-seq workflows typically begin with encapsulation – cells in suspension are partitioned into aqueous water-in-oil droplets along with RT and lysis reagents and barcoding microspheres (beads). This step takes place within the microchannels of a microfluidic device operating under laminar flow conditions, which ensure that distinct aqueous streams, such as those carrying cells or lysis reagents remain separate until droplet formation occurs (**Figure 1.1, A**). Upon encapsulation, the droplets physically separate individual cells and act as independent reaction compartments for mRNA capture and/or barcoding during RT. The unparalleled ultra-high-throughput

of droplet-based systems (encapsulating up to 300 000 cells/hour (60)) ensures efficient capture of large cell populations and representation of rare cell types/states.

Two key technical aspects that support reliable single-cell analysis are droplet uniformity and stability. Consistent droplet size is achieved through microfluidic channel design and precision syringe pump-controlled flow rates. Droplet stability is ensured with the use of surfactants, typically introduced with the oil phase. These amphiphilic molecules orient themselves at the droplet interface – with hydrophilic heads facing the aqueous phase and hydrophobic or fluorophilic tails facing the oil – thereby reducing interfacial tension and preventing coalescence (61). Droplet volumes can be tuned depending on the assay performed, in a range from a few femtoliters to several nanoliters (61), with ~1 nl droplets most commonly used for single-cell transcriptomics (2,3).

The goal of cell partitioning is to co-encapsulate single cells with primer-carrying microspheres that enable capture and barcoding of individual transcriptomes. Cell loading is a random process and follows a Poisson distribution, where the probability of having x cells per droplet is expressed as $P(X=x)=e^{-\lambda}[\lambda^x/x!]$. In this equation, λ denotes the average number of cells per droplet. Cell loading concentration directly affects λ, allowing researchers to adjust cell occupancy by dilution. Typically, a λ of value ~0.1-0.3 is used, ensuring that only a fraction of droplets will contain a cell, resulting in low rates of cell doublets or multiplets (62). However, cell suspensions are prone to sedimentation, distorting the expected occupancy over time. Additionally, due to cell-to-cell variability in size, shape and density, cell capture efficiencies might vary. To circumvent this effect, the density or viscosity of cell suspension buffers is adjusted using compounds such as OptiPrep, dextran or xanthan gum (2,63). Equally critical to successful cell barcoding is the efficient encapsulation of the barcoding beads. These beads carry abundant copies (10^6-10^9) of RT oligonucleotides, either bound to hard microparticles (Drop-seq, (3)), or entangled in dissolvable (10X Genomics, (18)) or insoluble hydrogel beads (inDrops (2,62)). Depending on the design, beads can be introduced in a controlled fashion, reaching >80% occupancy (inDrops and 10X Genomics approach) or in a random manner, again following Poisson (Drop-seq approach). The latter results in barcoding of only a small fraction of cells, which is not desirable when aiming to capture rare cell types, or if the amount of starting material is low.

Upon encapsulation, cells are lysed and the released mRNA is captured by RT primers. Several targeting strategies are available, including transcriptome-wide capture using specific probes (64) or targeting the 3’ or 5’

ends of mRNA molecules. This section will focus on the 3' capture, as it is the most widely used approach. RT primer design details also differ depending on amplification strategy used (IVT or template-switching (TS) approach). However, the main parts, essential for single-transcriptome barcoding, are consistent and include: 1) poly-(dT) stretch for polyadenylated mRNA capture; 2) a bead-unique cell barcode sequence for identification of individual cells; 3) a random, primer-unique molecular identifier (UMI) sequence, used to quantify transcripts; and 4) a sequencing adapter (**Figure 1.1, B**). If IVT-based amplification is used, the primer also contains a T7 promoter region, while the TS-approach primers contain a PCR handle.

The TS-based protocols exploit intrinsic terminal transferase activity of reverse transcriptase enzyme (typically derived from Moloney murine leukemia virus (M-MLV)). Upon cDNA synthesis, the enzyme adds several non-templated nucleotides, primarily cytosines, to the end of the newly synthesized cDNA molecule, which serve as an annealing site for the template switching oligo (termed TSO). This primer then serves as a template for the reverse transcriptase to extend the cDNA in a process known as template switching. The reaction is performed in droplets after the release of RT primers, photocleaved from the barcoding beads in inDrops-2 (60) or released after bead dissolution in 10X Genomics approach. Then, since all cDNAs contain barcode information and TSO sequence, emulsions are broken and barcoded cDNA library is processed in bulk. Alternatively, in Drop-seq approach, RT is performed in bulk after recovering single-transcriptome bound microparticles. In any case, after RT, all cDNAs contain a universal sequence introduced via template switching, and this site is used for PCR primer binding and library amplification. Amplified cDNA is subjected to enzymatic fragmentation, adapter ligation and indexing PCR, which is used to uniquely label different libraries. The resulting final libraries are then pooled and sequenced.

In the IVT-based protocols, after RT and second strand synthesis reactions, the T7 promoter sequence present on the barcoding RT primer is used to transcribe the cDNA molecules into antisense RNA, thereby linearly amplifying the libraries (62). The resulting RNAs can then be fragmented not only enzymatically, as in TS-based approach, but also chemically (i.e. by using zinc-ion mediated hydrolysis) (62). The fragments are then prepared for sequencing by a second RT reaction, again converting RNA to DNA, and a few cycles of indexing PCR. The linear amplification strategy offers benefits over exponential (PCR), due to lower error rate and reduced bias toward abundant transcripts (65). However, IVT-based protocols are laborious and lengthy, thus, TS approach is dominating in the field (45).

There are two main approaches for scRNA-seq library sequencing: short-read sequencing of 3′ or 5′ end fragments, and long-read full-length transcript sequencing. Technically, once cDNA is generated and not yet fragmented, both options are available, and the choice of sequencing approach depends on particular study goals. For studies focusing on cell heterogeneity and differential gene expression analysis, gold-standard short-read Illumina sequencing is sufficient. It allows cost-effective high-throughput quantification of transcripts, however, the sequence information is mostly lost. Thus, its utility for studying transcript isoforms, gene fusions, alternative splicing events, or single nucleotide polymorphisms (SNPs) is limited. In contrast, long-read technologies (i.e. PacBio or Oxford Nanopore) sequence the entire transcript, allowing direct resolution of isoform diversity. This approach can benefit studies where transcript structure or sequence variation is biologically relevant, such as cancer, developmental biology and immunology. Nonetheless, it is currently limited by lower throughput, higher error rate and cost, as well as increased technical complexity compared to short-read methods (66,67).

An example of a final scRNA-seq library structure is depicted in **Figure 1.1**, panel **C**. It comprises universal P5 and P7 sequences, needed for DNA hybridization to the Illumina flow cell to enable sequencing-by-synthesis; Read 1 and Read 2 primer binding sites; cell barcode, UMI and cDNA insert sequences; i5 and i7 library indices. Standard paired-end Illumina sequencing reads are allocated in a way that Read 1 captures the cell barcode and UMI information, Read 2 – cDNA fragment sequence, and index reads are used to determine the library index, needed for demultiplexing of libraries. Sequencing depth choice depends on monetary resources available, the biological question raised and method used, and typically ranges in 20-100 thousand reads per cell (2,45,60).

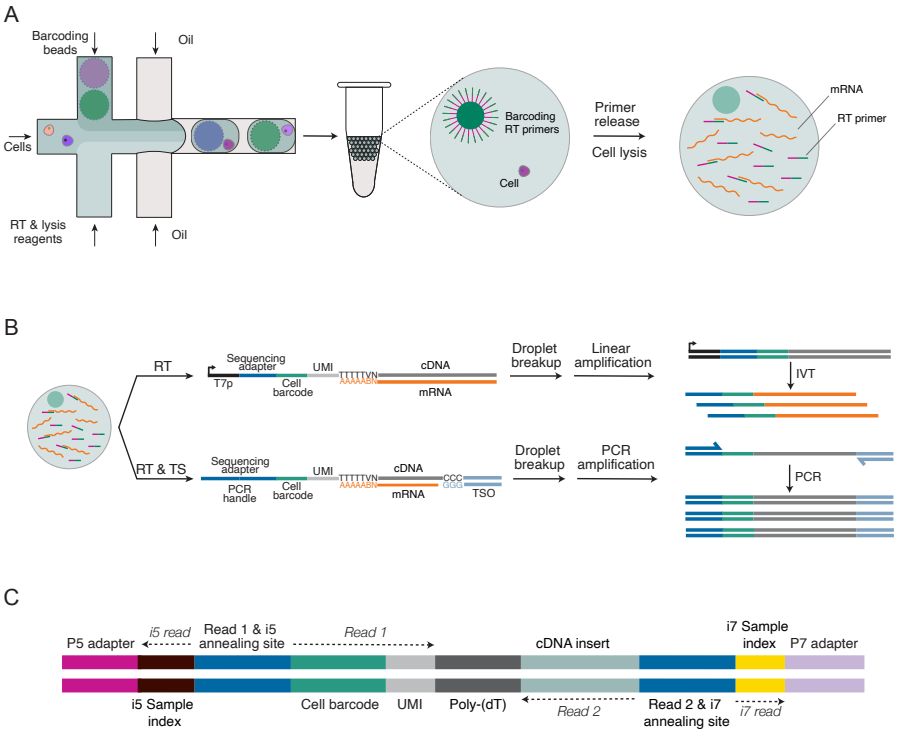


Figure 1.1. A – schematics of cell encapsulation in a microfluidics device and mRNA capture in droplets. **B** – schematics depicting RT primer design and cDNA amplification for both template-switching and *in vitro* transcription library preparation approaches. **C** – the final sequencing library structure.

1.1.3. Single-cell transcriptomics data analysis

The analysis of scRNA-seq data is a complex, iterative multi-step process. The goal of the initial processing of the sequencer output is to generate cell × gene matrices, where genome-wide expression values for each cell are stored. The matrices are then used to construct a low-dimensional data representation, and the process involves quality control, normalization and dimensionality reduction, followed by various downstream analyses, tailored to address the specific research questions raised (**Figure 1.2**) (68). Guidelines for data analysis are available (68,69), yet no "one button" solution exists. Every dataset presents unique challenges and questions, tackled utilizing various tools implemented across a range of programming languages. For convenience, these tools have now been collected into frameworks, the most popular ones being R-based Bioconductor (70) and Seurat (71), and Python-based scverse (72) ecosystem, with Scanpy package (73) at its core.

Simpler, graphical user interface platforms are available from 10X Genomics and Parse Biosciences, as well as non-commercial Granatum (74), ASAP (75), SCiAp (76) and others. However, the use of such platforms is not widespread due to limited control and flexibility.

Today, the analysis of scRNA-seq data has evolved into its own separate research field, and numerous tools and algorithms are continuously being developed, with over 1700 published to date (77). With such abundance, it is increasingly difficult to navigate which method performs the best for a particular task, thus, independent benchmarking efforts are necessary. However, the field lacks standardized evaluation frameworks and such lack of consistency often results in conflicting assessments of the same analytical methods. Custom benchmarks designed by tool developers frequently rely on selectively chosen datasets and performance metrics that emphasize the strengths of their own approaches, potentially compromising objectivity. Moreover, even in independent efforts, different datasets and metrics are used, resulting in <10% overlap between benchmarks, as reported recently (78). This highlights the unavoidably subjective and iterative nature of scRNA-seq analysis: comparisons and "sanity checks" have to be made numerous times, selecting methods and parameters most suitable for a particular dataset and research question. The following sections provide a brief overview of commonly used approaches and key considerations associated with each major step of the analysis.

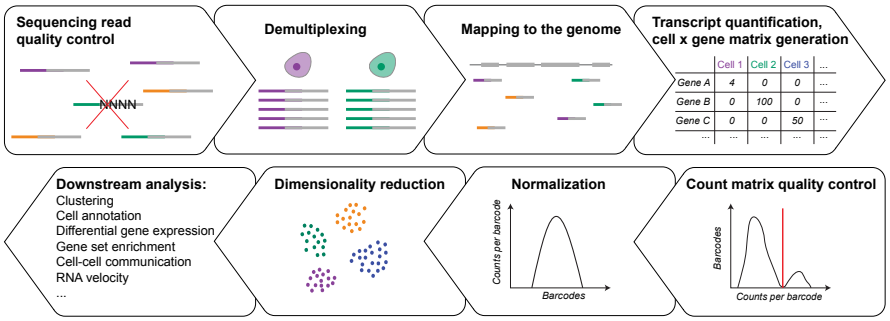


Figure 1.2. The data analysis workflow from quality control of sequencing data to low-dimensional representation, followed by various downstream analyses.

1.1.4. Single-cell transcriptomics data pre-processing

The first task in scRNA-seq analysis is to convert the output of the sequencing machines into a cell × gene (a.k.a barcode × feature, or count) matrix, which is the basis of downstream analyses that eventually yield

biological insight. The matrix generation is a mapping/quantification task, comprised of four main stages: 1) assigning reads to cells (cell barcode demultiplexing); 2) mapping of the sequenced cDNA reads to a reference; 3) assigning reads to genes; and 4) counting the number of unique initial RNA molecules (UMI deduplication). This procedure is very computationally demanding, as it involves processing hundreds of millions of reads, and requires the use of high-performance computing machines (HPC). Multiple parameters govern each step of the process and are barcoding method-specific. Usually, the entire pipeline is executed by a dedicated software such as STARsolo (79), CellRanger (18), Alevin (80), Kallisto-bustools (81), zUMI (82), SEQC (6) or other.

Starting with the sequencer output, upon base calling, FASTQ files are generated for each library – demultiplexed by the unique index sequences. Then, quality of reads is assessed and low-quality reads are removed, typically using FastQC (83). Next, adequate quality reads are demultiplexed (sorted) by barcode sequences, most often supplied as a whitelist (2,18) that also allows barcode correction. Once the reads are sorted, their origins are determined by mapping, which can be done on the organism's genome or transcriptome. It might appear counter-intuitive, but it is preferable to map reads to the genome, as scRNA-seq and especially snRNA-seq data contain reads spanning introns or intergenic regions (84). Additionally, mapping to the transcriptome has been shown to increase the amount of multi-mapping reads, which is deemed not desirable (85). Considering that a typical dataset has ~10-15% splice junction spanning sequences, splice-aware aligners have to be used, with the most popular one being STAR (86).

An aspect that is often overlooked in scRNA-seq data pre-processing, but has immense impact downstream, is mapping and the genome annotation used. Even after more than 20 years of human genome assembly, the exact number of genes remains unknown – the gene and transcript annotations are continuously updated and expected to expand further (87). Two main catalogues used are GENCODE (88) and National Center for Biotechnology Information (NCBI) maintained RefSeq (89). Despite their widespread use, these catalogues report different numbers of protein-coding genes and other features. Additionally, even within a single annotation, ambiguities arise when read mapping is performed: a fraction of reads does not map confidently; some reads align to multiple locations, introns or intergenic regions; and some map to locations that have multiple overlapping annotations. Naturally, a question arises: which of these reads to include into the count matrix? In scRNA-seq studies the multi-mappers, irrespective of true or annotation-imposed multi-mapping, are usually discarded from analysis (84). Bulk RNA-seq data

analysis showed that such exclusion can distort the expression levels of certain genes and potentially affect downstream analyses (90). While still a topic of debate in the field, inclusion of multi-mappers was shown to benefit analysis for certain gene groups (i.e. paralogs) (79). Multi-mapped reads can be included in the count matrix by simple uniform distribution across the multiple mapped genes, or using sophisticated probabilistic methods, such as Maximum Likelihood Estimation and Expectation Maximization (MLE-EM) algorithms (79). Similarly, inclusion of intronic sequences can not only provide a unique analytical outlook (discussed in section 1.2.4), but also benefit gene detection sensitivity, especially in snRNA-seq datasets (79,91). Nonetheless, post-mapping inclusion of multi-mappers and intronic or intergenic sequences in the count matrix represents a trade-off between assignment confidence and data retention. Recently, a genome reference annotation optimization strategy was proposed, constituting 3' UTR extension, intronic read incorporation and resolution of gene overlaps (91). The use of such augmented genome annotation was shown to improve cell profiling resolution by recovering missing marker genes and even certain cell subtypes (91).

Irrespective of the feature inclusion strategy used, for quantification of gene expression levels, the reads have to be traced back to the original transcripts. During library preparation, cDNA is amplified numerous times and reads do not represent unique molecules anymore. For that reason, the UMI sequence, originating from the RT primer, is used for deduplication. In this process, each read that has a unique cell barcode, UMI sequence and maps to the same feature is treated as a single molecule. Due to amplification and sequencing errors, artificial UMI sequences might get generated, and current pre-processing pipelines have multiple sophisticated options for UMI correction (79,80). Finally, after UMI deduplication, the count matrix of dimensions barcodes × features is populated with numerical values – the expression levels of each gene for each barcode.

1.1.5. Count matrix quality control

Once the count matrices are generated, quality control (QC) measures need to be taken to ensure that barcode entries in it represent actual single cells. Not every barcode in the matrix corresponds to a viable single cell. At this stage, the quality of sample preparation becomes evident, underscoring the importance of quick and efficient sample handling (**Figure 1.3**). During cell encapsulation process, the barcoding bead occupancy is >80%, and the cell occupancy is typically ~10-20%. That means that any RNA released from

dying or damaged cells enters every droplet and gets barcoded. To filter out this noise and low-quality cells, three main metrics are considered: the total number of counts per barcode, the fraction of mitochondrial gene counts per barcode and the number of genes expressed per barcode (68,69,84). Depending on the biological system investigated, it might be beneficial to assess the fraction of ribosomal and hemoglobin genes as well (69), as it might indicate low-complexity transcriptomes or contamination. Some pre-processing pipelines, such as CellRanger and STARsolo, attempt automatic barcode filtering based on the count distribution, however, it is important to evaluate at least two metrics together. Most often, the thresholds are selected manually based on distributions observed. Barcodes with very low transcript and gene counts represent ambient RNA or damaged cells, while high fraction of mitochondrial genes is associated with cell death (69). Regarding ambient RNA contamination, simple removal of low count barcodes might not be enough, as the leaked RNA enters the droplets irrespective if it contains a cell or not. Numerous algorithms have been developed for quantification and removal of ambient RNA contamination, popular ones being SoupX (92), CellBender (93) and DecontX (94). Simply put, these tools estimate the background expression profile and attempt to remove it by correcting gene expression levels. An independent benchmark of these tools on mouse kidney dataset revealed CellBender as the method of choice regarding cleanup related to marker gene expression analysis (95). Another study demonstrated that previously published single-transcriptome profiling studies on mouse brain suffer from neuronal transcript contamination in glial cells, and background removal aids in recovery of other overlooked rare subpopulations (96). However, as these tools augment the count matrix, which is generally not desirable, it is necessary to carefully examine the data before and after cleanup. Whether or not to perform background correction depends on the extent of contamination and the overall goals of the study. Strong knowledge in biology of the system investigated is a must, and certain tasks, such as clustering and classification of cells, are robust to some background noise (95).

Additionally, since cell loading is a Poisson process, a small fraction of droplets contains two or more cells, whose transcriptomes will be represented by a single barcode. Such events are especially frequent if the cell suspension contained any cell aggregates. These hybrid transcriptomes need to be removed as they can compromise downstream analysis or be mistakenly identified as a novel cell phenotype. Several efficient doublet detection methods have seen widespread adoption, such as Scrublet (97) and DoubletFinder (98). These tools work by generating artificial doublets out of

any random barcode pairs in the dataset, constructing a *k*-nearest neighbor graph in the PCA space and assigning a doublet score to each barcode depending on the proportion of doublets among its neighbors. The user can then select which barcodes to exclude based on the distribution of doublet scores. Independent benchmark using simulated cell line and real peripheral blood mononuclear cell (PBMC) scRNA-seq datasets positioned R package DoubletFinder as top performer in terms of accuracy, but not computational efficiency (99). Overall, the QC of the count matrix is an iterative process, and the thresholds might need to be adjusted multiple times based on the results of downstream analysis.

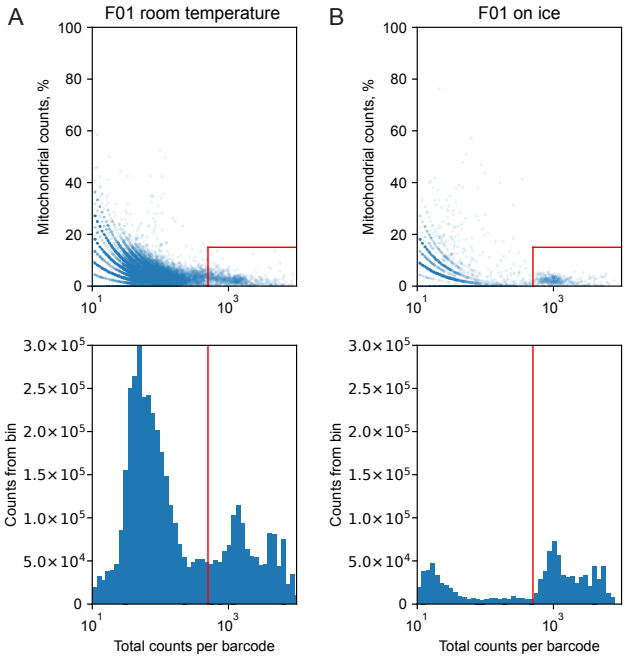


Figure 1.3. Assessing mitochondrial count fraction and total counts per barcode provides information on sample quality. High values of mitochondrial count fraction are associated with cell death, while low total counts per barcode indicate ambient RNA released from dying cells. **A** – amniotic fluid sample prepared for barcoding at room temperature, scatterplot and histogram indicates cell damage. **B** – the same sample prepared for barcoding on ice has higher quality for both metrics.

1.1.6. From count matrices to low-dimensional embeddings

The entries in the filtered cell × gene count matrix represent successful capture, reverse transcription, amplification and sequencing of a given transcript. Each of these steps is not ideally efficient and adds a degree of

variability. Additionally, lysis efficiency and sequencing depth differences contribute to the gene expression variation between even the most biologically identical cells. It is well known that scRNA-seq data suffers from dropout – up to >97% of entries in the matrix are 0 (100). These zeros can be biological or technical in origin. It is natural that cells express only a certain set of genes needed to support their function and metabolism. Additionally, transcript capture is a random process and RT reaction has limited efficiency, resulting in failure to detect some transcripts and variability of gene expression levels among cells. Thus, to account for sampling effects and enable statistical analysis, data needs to be normalized. It is widely recognized that normalization is the most critical step in scRNA-seq data processing, having the highest impact on downstream analysis (68,69,101). A standard approach inherited from bulk RNA-seq is to normalize the expression of all cells globally using a set scaling factor (median counts per cell or a certain number of total counts per cell) (68). However, it operates under the assumption that all cells in the dataset are of the same transcriptome size, which is almost always violated when analyzing heterogeneous cell populations (69,102). Thus, complex normalization algorithms suited for sparse scRNA-seq data have been developed. For instance, a popular tool Scran (103) pools cells with similar count depth and estimates pool-based size factors using linear regression. Another approach utilizes analytical approximation of Pearson residuals, fitting a linear model with sequencing depth as a covariate to obtain transformed matrices (104). Interestingly, a recent benchmark of 22 transformation methods showed that a simple shifted logarithm transformation $\log(y/s + y0)$ (where y is a particular gene's expression, s is a cell size factor and $y0 = 1$) performs just as well or better than sophisticated alternatives (105). These findings demonstrate that choosing a transformation based on conceptual reasons does not necessarily yield better downstream analysis results.

Another source of unwanted variation in the data is related to processing of cells in batches (i.e. libraries, samples, donors). Batch effects are typically observed in low-dimensional visualizations and should be removed to ensure they are not mistaken with biological signal (i.e. might appear similar to a patient-specific population, phenotypical shift as a response to treatment and other cases). There are numerous algorithms for this task which can be divided into two major groups: batch correction (tools that remove the batch effect augmenting the entire count matrix), or batch integration (tools that operate in lower-dimensional space and return augmented embeddings) (68). Methods initially developed for bulk RNA-seq or microarray data, ComBat (106) and limma (107) have been successfully applied to single-cell data for batch effect

removal. However, the augmented matrices should be used with caution due to over-correction issues, instead, adjusted lower-dimensional embeddings are preferred (108). Considering the non-linear nature of scRNA-seq data, methods utilizing nearest-neighbor information, such as MNN (109), BKNN (110) or Scanorama (111) have been developed. A benchmarking study evaluated 16 data integration tools, assessing their effectiveness in batch effect correction and preservation of biological variability (108). In complex settings, such as large-scale atlas integration, deep learning-based tools (i.e., scANVI (112) that includes cell type annotation information for alignment and scVI (113)), along with Scanorama, achieved top performance. However, for more straightforward integration tasks, linear embedding approaches like canonical correlation analysis (CCA) (114) and Harmony (115) performed well, with Harmony standing out as the preferred choice.

Upon log-normalization, the construction of a low-dimensional representation of the count matrix begins. In theory, single-cell transcriptomics datasets contain information on over 30 000 features in the human genome (89), in practice – around 15 000 genes are detected (68). A cell in such a dataset can be envisioned as a data point in the feature space, whose position is described by >15 000 values – dimensions. Drawing any conclusions or making biological interpretations out of this data is virtually impossible without some kind of simplification and interpretable representation. That becomes possible under the *manifold assumption* (116). A manifold is a mathematical construct that describes a lower-dimensional structure, on which the data points lie, positioned in a higher-dimensional space. This assumption is considered valid for scRNA-seq data, because gene expression is not random. For example, cells express defined transcriptional programs, governed by the activity of signaling and gene regulatory networks, and any transitions between cellular states tend to be gradual (116). Therefore, in a given single-cell transcriptomics dataset, the cells are ordered on a manifold that can be described by far fewer dimensions than the number of genes measured. Dimensionality reduction algorithms attempt to detangle the combination of linear and non-linear vectors needed to the describe this manifold (68,116). The first step in this process is informative feature selection. As mentioned, not all genes are important for the manifold structure. The first filtering applied is the removal of genes that are not abundantly expressed – i.e., in *n* cells by *m* counts. Permissive values for *n* and *m* should be chosen, aligned to the expected size of rare cell populations, as exclusion of their marker genes would result in failure to detect these phenotypes. Next, genes that describe the most variation in the data are selected (termed *Highly Variable Genes* or HVG), using metrics such as the Fano factor (mean-

variance ratio) (2). At this stage, the dimensionality is reduced to 1000-5000 HVGs; the exact number is arbitrary and was shown to be robust to downstream analysis (2,68). Further dimensionality reduction is performed using dedicated algorithms, such as principal component analysis (PCA). Even though the linear nature of this transformation is not sufficient to represent the data in 2D, it serves as a useful pre-processing step, because most of the variance is preserved and distances in the reduced dimensions have a consistent interpretation. The number of PCs to retain can be chosen manually using the elbow method (visual inspection of the cumulative variance plot) or automatically, by calculating the eigenvalues of shuffled data and retaining PCs that surpass this threshold (2,8). At this stage, the dimensionality is reduced to the range of tens to a couple hundred principal components. To preserve the distances between cells in a meaningful way, PCA space is utilized for *k*-Nearest Neighbor (*k*-NN) graph construction. A measure of distance (i.e. Euclidean) is used to connect *k* cells closest in the PCA space to a given cell, thereby recording the transcriptional similarity and preserving the local structure of the data. The graph serves as a foundational structure for many downstream analyses, including clustering, trajectory inference, and, critically, data visualization (68,69).

The neighborhood, or cell similarity concept is central to most scRNA-seq data visualization techniques, due to inherent non-linearity of the data. Thus, most tools for scRNA-seq data exploration and representation are non-linear manifold-learning approaches. A method that was considered the gold-standard in scRNA-seq visualization up until recently, t-distributed Stochastic Neighbor Embedding (t-SNE), records similarities of the data points in the high-dimensional space (i.e. PCA) and constructs an embedding by placing similar cells closer than dissimilar ones. It faithfully captures local relationships and groups similar cells, however, fails to record global structure, fragmenting any continuous progressions (i.e. differentiation) into clusters, which are positioned in an essentially meaningless way (116). Moreover, it lacks reproducibility due to stochastic nature and relies on super-parameters that are non-intuitive to adjust, but have a high impact on the overall representation (117). Thus, t-SNE plots should be interpreted with extra caution. The current gold-standard, Uniform Manifold Approximation and Projection (UMAP) (118) and force-directed layouts (FDL) such as SPRING (119) directly use the *k*-NN graph structure for building the embeddings represented in 2D or 3D. In these visuals, the general concept remains the same – cells with a similar transcriptional profile are portrayed in proximity. However, in contrast to t-SNE, SPRING is highly reproducible and exceptional at preserving continuous processes (119). Meanwhile, UMAP is

better at preserving the global structure of the manifold, including any transitions, and is very computationally efficient, making it superior for large-scale atlases (120). Besides these, other popular data representation algorithms, tailored for specific use cases exist. For instance, diffusion maps reflect differentiation trajectories and pseudotemporal ordering of cells, although the method returns more than two dimensions, complicating visualization, and does not scale well to large numbers of cells (121). Partition-based graph abstraction (PAGA) provides graph-like coarse-grained maps of cell groups, preserving the global data structure and reflecting both continuous and discrete phenotypes (122). Moreover, PAGA can be used to initialize other manifold learning algorithms, such as UMAP and FDL, improving their visualizations in interpretability and preservation of global topology (68,122). PHATE takes into account both local affinities and diffusion-based long-range relationships between data points, denoising the underlying structure and providing an information-rich low-dimensional output (123). Numerous other algorithms have been published, yet, not many have seen widespread adoption and, at least for now, UMAP remains the method of choice.

It is important to emphasize that while low-dimensional visualizations are vital for scRNA-seq data representation and exploration, no conclusions about the underlying biology should be drawn solely from the images obtained. Irrespective of the algorithm used, these visuals heavily depend on parameter tuning and can be manipulated easily, oftentimes inadvertently serving confirmation bias (124). Only downstream analysis (reviewed in the next section) can inform biological interpretations, and the images obtained should rather be used as illustrative summaries post-analysis. Nonetheless, the atlases can be very useful if assessed with caution: serving as a ground for initial data exploration, data-driven hypothesis formulation and technical artefact assessment, thereby iteratively informing QC process. Considering the data size and computational burden to generate embeddings, interactive exploration of scRNA-seq datasets is especially attractive and convenient. Therefore, multiple platforms have gained popularity, such as SPRING (119), cellxgene (125), ShinyCell (126) and open-access Broad Institute's Single Cell Portal (127). Importantly, these user-friendly tools not only support data exploration during the analysis phase, but also facilitate the dissemination and accessibility of results to non-specialists of scRNA-seq data, such as collaborators, clinicians and the public.

1.1.7. Downstream analysis

Downstream analysis of scRNA-seq data is an ever-expanding computational biology field with an extensive toolkit tailored to gain insight into the biological processes of complex multicellular systems. Owing to its unprecedented single-cell resolution, the most common and basic application of scRNA-seq technology is to investigate the phenotypic heterogeneity with the goal to characterize new cell types or states. Cell populations are identified by grouping cells with similar expression profiles into clusters. A classical algorithm for this task is *k*-means clustering (128), which iteratively identifies user determined *k* number of cluster centroids and assigns cells to the nearest one. However, the number of clusters *k* has to be supplied to the algorithm, although it is generally unknown *a priori*, and has to be determined heuristically; also, the algorithm is prone to generating clusters of equal size, thereby potentially masking rare populations (68). Due to these and other limitations, community detection methods that can use the underlying *k*-NN graph information are superior for scRNA-seq data clustering (69,129,130). Louvain clustering (131), first implemented for scRNA-seq data in PhenoGraph algorithm (132), determines the number of clusters without user input, and the granularity can be controlled adjusting the resolution parameter. It was shown to perform the best in independent benchmarks (133,134), and was therefore widely accepted by the research community as the default clustering approach (68). However, it also has flaws, such as poorly connected communities within clusters (135). Therefore, an improved version of this method was proposed, called Leiden clustering, which guarantees connected communities and is more computationally efficient (135). Use of this approach is now recommended as a part of scRNA-seq data analysis best practices for cell annotation tasks (69). It is important to emphasize that cluster number can be manipulated easily and does not necessarily correspond to the number of cell types or states in the dataset. Thus, clustering should be performed in an iterative fashion, evaluating the results obtained with multiple resolutions and algorithms (130).

Naturally, the next step is to assign a biological interpretation to the clusters obtained. This process is known as cell annotation. It can be performed automatically using dedicated tools such as classifier-based methods and reference mapping, or manually, examining cluster-specific marker genes (69). Ideally, both approaches should be combined, starting with automatic annotation to obtain general cluster labels, followed by manual expert curation (136). Initial automatic classification efforts included construction of simple Bayesian classifiers from publicly available bulk,

sorted cell expression profiles (8). Later, more sophisticated classifiers pre-trained on large, open-access scRNA-seq atlases were developed, CellTypist (137) being the most widely applied. CellTypist model can also be trained by the user on any annotated dataset, providing flexibility. The other group of tools – reference mapping – work by integrating query dataset with an existing annotated atlas, and perform label transfer on the resulting joint embedding. Dedicated algorithms exist for this task, but data integration methods, such as the aforementioned Harmony, can be readily used for label transfer as well (136). Nonetheless, automatic annotation results heavily depend on the quality of the reference or training datasets and preciseness of annotation, as well as quality and complexity of query data, and must be verified. Thus, despite being extremely labor intensive, manual annotation remains highly relevant. Typically, differential gene expression (DGE) analysis is performed, comparing the expression profile of a particular cluster with the rest of cells in the dataset. A simple statistical test (i.e. Wilcoxon rank-sum, with false-discovery rate (FDR) correction) is sufficient to determine the marker genes for each cluster (69). The resulting gene lists are then examined using literature, existing cell atlases and marker gene databases, to assign cell type identities.

Depending on the biological question investigated, partitioning of cells into discrete clusters might not be a suitable strategy altogether. Certain biological processes, such as transformation, differentiation or activation, are transitory in nature and require a different approach. Continuous processes are challenging to study with scRNA-seq, as the data provides a snapshot of the entire system, and the ability to reconstruct cellular trajectories depends on the abundance of cells along them. Additionally, the underlying trajectories can be diverse: linear, branching trees, cyclical or a mixture of these. Algorithms that order the cells based on their transcriptional similarity are called trajectory inference or pseudotime analysis methods. The pseudotemporal ordering was first introduced in Monocle (138), which uses a graph-learning technique called minimum spanning trees (MST) to extract trajectories from cell graphs embedded in low-dimensional space. However, the number of lineages and direction of the trajectory had to be supplied to the model. The latest version, Monocle 3, overcame these limitations and is widely used (139). Diffusion pseudotime (140) takes another approach and uses random walks on a weighted k-NN graph, yet the starting cell has to be determined by the user. Another popular tool, Slingshot (141) fits curves to cluster-based MST and does not require pre-determined start and end points. Palantir (142) uses diffusion maps and Markov chains to not only model trajectories, but estimate cell fate probabilities. Nonetheless, the performance of trajectory inference

methods heavily depends on the dataset complexity and the type of trajectory to uncover, and no single algorithm has clear superiority (69).

Trajectory inference by transcriptional similarity only might not have a biological meaning and requires validation with more sophisticated techniques (68,69). An innovative approach with a solid conceptual framework to infer dynamic, directed processes was introduced by La Manno et al., termed RNA velocity (143). The method first constructs the spliced and unspliced count matrices from the raw sequencing data using intron mapping reads, and models splicing kinetics by inspecting the ratio between the two forms of mRNA for a given gene. It operates under the assumption that unspliced mRNA precedes spliced, indicating activation of transcription. Consequently, a gradual increase in spliced counts indicates a progression in time, providing an estimate of the cell's future state (143,144). The method faithfully recovers expected trajectories of multiple types of differentiating cells and longitudinal sampling experiments. However, it was observed that in certain contexts that involve steady states or transcriptional bursts (i.e. hematopoiesis) the direction is inferred incorrectly (145). Thus, model assumptions and individual gene phase portraits have to be taken into account and assessed carefully before making any biological conclusions. Transitional process inference was recently advanced by CellRank, a tool combining RNA velocity dynamics with trajectory inference using the *k*-NN graph (146). Now improved and available as CellRank 2, this framework enables identification of initial and terminal states, estimation of cell fate probabilities and integration of other data modalities, advancing the analysis of dynamic cellular systems (147).

Transitional processes mentioned above involve orchestrated transcription of multiple genes or gene modules. Due to the inherent sparsity of scRNA-seq data and the dropout effect (discussed in section 1.2.3), studying gene co-expression patterns is challenging. Therefore, a group of algorithms that attempt to fill the missing values in a process known as data imputation were developed. They can be broadly divided into three groups (148). The first utilizes probabilistic models to infer which zeros in the count matrix are technical in nature and attempt to fill these values. For instance, scImpute borrows information from similar cells utilizing genes less affected by dropout (149), while SAVER fits a regression model imitating the transcript sampling procedure (150). The second group estimates a latent space representation of cells through matrix-based linear decomposition (i.e. singular vector decomposition in ALRA (151)) or deep learning methods (i.e. neural networks in scVI (113)). The third group adjusts all entries (both zero and non-zero) in the count matrix by smoothing or diffusing the expression values across

similar cells, i.e. neighbors in a graph as in MAGIC (152) or *k*-NN smoothing (153). This approach is particularly useful for gene-gene relationship reconstruction and allows transcription factor (TF) target prediction, as demonstrated by reconstruction of the transcriptional landscape of epithelial-mesenchymal transition (EMT) (152). An independent benchmark revealed that *k*-NN smoothing, MAGIC and SAVER outperform other methods in recovering missing gene expression values (148). However, imputation did not improve other downstream analysis such as clustering or trajectory inference. Overall, imputation can be beneficial as a denoising approach for gene co-regulation analysis, however, it should be used with caution or avoided altogether for other downstream analysis tasks.

Upon annotation of the cellular phenotypes, be it discrete or transitory, an atlas of the members present in the biological system investigated is obtained. However, these diverse cell types do not act in isolation – they are in constant contact with the surrounding matrix and each other, transmitting important signals that shape the behavior of the local microenvironment. Hence, various methods to systematically identify receptor-ligand interactions and intercellular signaling networks from scRNA-seq data have been developed. In the process known as cell-cell communication inference, these algorithms use manually curated repositories of known receptor-ligand and other interactions and model potential communication between phenotypes in the dataset. CellChat uses probabilistic models, network analysis and manifold learning that take into account multi-subunit expression of receptors, secreted and membrane-bound ligands, as well as co-factors (i.e. soluble agonists and antagonists), and outputs cell communication probabilities organized into signaling pathways (154). Meanwhile the continuously updated CellPhoneDB (155,156) uses permutations to identify significant interactions of ligand-receptor pairs, also taking into account subunit architecture of receptors and ligands. Newer versions of this tool support integration of spatial transcriptomics data to take into account cell co-localization for higher interaction confidence (157); modeling non-protein ligand-mediated signaling (i.e. hormone and small molecule) using expression of the last enzyme in the biosynthesis pathway as a proxy; as well as investigation of downstream TF activity, derived from scRNA-seq or scATAC-seq data (158). Nonetheless, transcriptional data cannot reliably portray actual activity of proteins, thus, cell-cell communication inference results should be assessed with caution. Moreover, independent benchmarking revealed that different databases provide uneven coverage of specific pathways and the choice of both method and database used heavily impacts the outcomes (159). Thus, the LIANA framework was developed, integrating various databases and methods into a

single platform, providing a consensus-based approach that improves robustness and reproducibility of predictions (159), and is now recommended as a starting point for cell-cell communication analysis (69).

Single-cell transcriptomics not only enables in-depth investigation of cellular heterogeneity, development and communication, but is also instrumental in disclosing compositional changes in tissues. The abundance of certain phenotypes (i.e. exhausted immune cells in cancer) can act as biomarkers for disease progression or response to treatment, highly relevant in the clinical context (160). Often, a general view of phenotype enrichment can be observed simply by clustering and examining abundance of cells across conditions in a low-dimensional embedding such as a UMAP. However, for reliable and statistically significant associations with experimental perturbations or clinical outcomes, differential abundance (DA) analyses are required. Several algorithms for DA testing have been published. For instance, Milo counts the number of cells of differing experimental conditions in local neighborhoods of the *k*-NN graph and applies negative binomial generalized linear models that take into account sample size differences (161). CNA utilizes random walks on the *k*-NN graph to obtain neighborhood abundance matrices, which are analyzed using PCA (162). Another method, DA-seq applies a logistic regression model and label permutation to identify DA cells between two conditions (163). While exact statistical approaches differ, these methods largely achieve the same results, and the selection of DA testing method mostly depends on the dataset analyzed. For instance, a recent benchmark found Milo to be least sensitive to technical and biological noise; Milo and DA-seq robust with respect to hyperparameter changes; while CNA was the most scalable method for DA analysis in large atlases (164). Overall, analyzing differences in cell abundance provides another layer of characterization for complex tissues, especially relevant in disease or treatment monitoring context.

Aside from aforementioned sophisticated algorithms, many research questions in transcriptomics benefit from relatively simple DGE analysis. For instance, the phenotypic differences between treated vs non-treated cells, patients, timepoints, conditions and other instances can be determined. However, care needs to be taken not to mistake technical artifacts with biological signal. It was recently shown that the performance of DGE workflows depends on batch effect intensity, sequencing depth, data sparsity and synergy of these aspects (165). Additionally, failure to acknowledge inter-individual differences was shown to generate false positives, thus, pseudobulk approaches that enable covariate modeling, (i.e. MAST (166), DESeq2 (167), limma (107)) should be preferred (168). A recent study published in Nature

presented single-cell memory-related gene expression changes (169), yet independent reanalysis with proper statistical assessment (pseudobulk with multiple hypothesis testing) revealed no significant differentially expressed genes (170). This highlights the importance to understand the statistical procedures needed for trustworthy results in scRNA-seq analysis.

The hundreds or thousands of genes obtained from DGE analysis are non-straightforward to interpret. Thus, various statistic approaches for gene set enrichment analysis (GSEA) (171) and over-representation analysis (ORA) are carried out to summarize results into interpretable terms, such as signaling pathways or biological processes. It was demonstrated that GSEA results are more sensitive to the choice of gene set database used rather than statistics applied (172). Commonly used databases include Gene Ontology, for broader biological or cellular process description (173); MSigDB, for diverse hallmark gene set activation evaluation (174); KEGG (175) and Reactome (176), for molecular signaling and metabolic pathway activity inference. A complimentary layer for understanding the molecular mechanisms behind the phenotypic diversity in scRNA-seq data is offered by gene regulatory network and transcription factor (TF) activity inference algorithms. Tools such as DoRothEA (177) leverage curated information from ChIP-seq datasets and perturbation experiments to infer TF activities based on the expression of targets. Unified frameworks that provide access to various databases and activity inference methods like decoupleR (178) and OmniPath (179), can aid the selection of suitable databases and tools. Overall, single-cell transcriptomics data is rich in information beyond the general description of transcriptional heterogeneity, and the research community is actively developing methods to address increasingly complex questions, from cellular dynamics and communication, to upcoming inevitable integration with other data modalities (69).

The recent rise of machine learning and artificial intelligence (AI) did not surpass the single-cell bioinformatics field. Rapidly evolving AI models offer promising tools to extract meaningful patterns from the highly-dimensional single-cell resolution data. Today, nearly every pre-processing and downstream analysis step detailed above has a corresponding deep-learning-based method available (113,180,181). Nonetheless, their adoption remains limited due to complexity, lack of interpretability and independent benchmarks, as well as substantial computational resources and extensive coding skills required for implementation (182).

Today, with developments in single-cell technologies and a vast amount of work already published, the primary challenge in single-cell genomics has shifted from generating high-quality data to interpreting it in a meaningful

way. Single-cell profiling, in combination with the fast-growing field of AI, holds potential to uncover fundamental biological mechanisms, drive therapeutic target and drug discovery, and fuel the precision medicine of the future. Reflecting this promise, emerging companies (i.e., Ensocell, Cellanome, Cellular Inteligence) are actively developing therapeutic solutions that involve the use of high-dimensional single-cell data. How these efforts will shape the next phase of biological discovery and translational research remains to be determined.

1.2. Healthy kidney and kidney cancer

1.2.1. Cellular basis of kidney physiology

Human kidneys are complex organs of the urinary system, primarily involved in metabolic waste filtration from the bloodstream to form urine. In addition to the excretory role, kidneys convert vitamin D to its active form and produce hormones such as erythropoietin and renin. Vitamin D is important for bone homeostasis, erythropoietin regulates erythropoiesis, while renin-angiotensin-aldosterone axis controls acid-base and electrolyte balance, plasma osmolarity and pressure of the blood and other bodily fluids (183). The main structural and functional unit of the kidney – the nephron – consists of a glomerulus, enclosed by the Bowman's capsule, and renal tubule, entangled in a peritubular capillary network (**Figure 1.4**). Glomerulus is a sieve-like ball of capillaries where blood filtration of metabolites and proteins smaller than albumin (<68 kDa) occur. The filtrate enters the Bowman's capsule and gets concentrated while passing along the tubular system, consisting of proximal tubule, Henle's loop and distal convoluted tubule, where water, small proteins, amino acids, carbohydrates and electrolytes are reabsorbed by the tubular epithelial cells. Non-absorbed compounds enter the collecting ducts to form urine, which is eventually excreted to the renal pelvis and the ureter (183). The vital physiological functions of the kidneys are executed by a multitude of diverse, spatially and phenotypically specialized cells, which will be briefly introduced in the following sections.

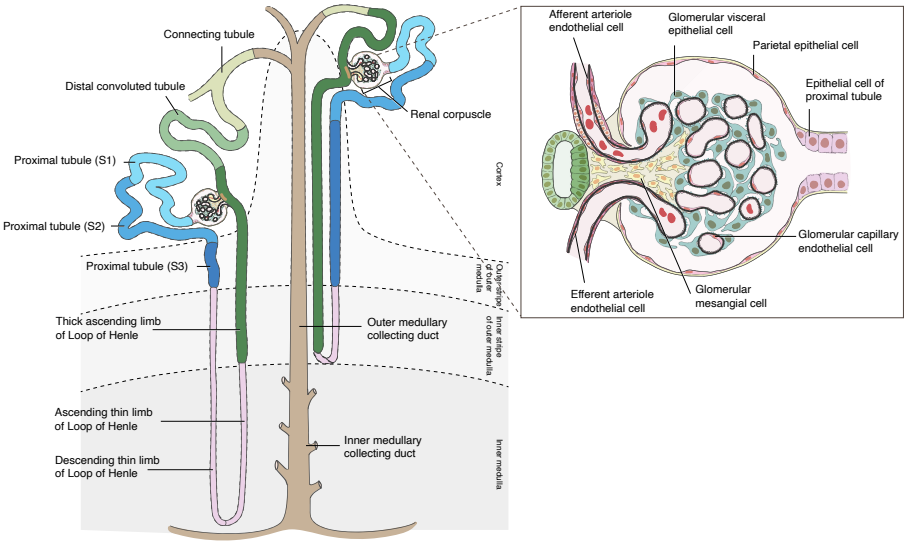


Figure 1.4. Schematics depicting the main structural unit of a kidney, the nephron and the different tubular segments. Additionally, a close up of the renal corpuscle – glomerulus enclosed by the Bowman's capsule – depicts the different cell types within this structure. Graphics were obtained from Human Reference Atlas, NIAID NIH BIOART Source, under accession numbers BIOART-000560 and BIOART-000562 with CC BY 4.0 license.

At least four distinct cell types reside in the Bowman's capsule – glomerular endothelium, parietal epithelial cells, mesangial cells and podocytes (183,184). Parietal epithelial cells line the Bowman's capsule, internalize albumin and, under inflammation, upon podocyte-induced activation, secrete basement membrane-like ECM (185–187). Some authors postulate that these cells can differentiate into podocytes (185). Podocytes are an integral component of the glomerular filtration barrier. These visceral epithelial cells have long foot processes that wrap glomerular capillaries, forming slit diaphragms, which maintain selective solute filtration based on size and charge (188,189). Additionally, podocytes are the main source of VEGF in the glomerulus, exerting proliferative and protective effects on endothelial cells (190). Mesangial cells are stromal cells very similar to pericytes and vascular smooth muscle cells (vSMCs), distinguished by their phagocytic capabilities. They support the structure of the glomerulus, secrete the glomerular basement membrane, regulate inflammation and recruit immune cells, as well as participate in injury repair (191). The glomerulus is challenging to dissociate, hence, recovery of glomerular cells in scRNA-seq studies is usually low (192). Studies employing enrichment strategies (FACS)

have shed light on the heterogeneity of mesangial cells and inter-species differences, providing guidance for translational studies (184).

The intricate tubular system of the nephron is comprised of many transcriptionally distinct segments, differing by location, functions and permeability to water and solutes (193). The proximal tubule can be divided into three segments (S1, S2, S3) on the basis of morphology and location, or into convoluted and straight proximal tubule (194). The main function of proximal tubule is reabsorption of water, glucose, amino acids, salts and bicarbonate (195), essential for maintaining organism-level acid-base balance. The segments have distinct transcriptional profiles, i.e., glucose co-transporter SGLT2 is expressed in S1 and S2 segments, while SGLT1 is enriched in S3 (196). The Henle's loop is composed of descending thin limb, ascending thin and thick limbs, further subdivided into cortical and medullary segments (194). The descending parts are permeable to water (highly express aquaporin AQP1), but not to solutes, while the ascending parts are impermeable to water, generating a high cortico-medullar osmotic gradient. The ascending thin part is responsible for urea collection from the interstitium, while the thick ascending segment reabsorbs essential divalent cations (Mg^{2+} and Ca^{2+}), as well as ammonia with sodium, potassium and chloride ions through NKCC2 transporter (encoded by *SLC12A1*). The distal convoluted tubule comprises two segments differing in sensitivity to adrenal hormone aldosterone, which stimulates sodium reabsorption and potassium secretion, regulating fluid volume and blood pressure. This part is also responsible for leftover chloride, calcium and magnesium reabsorption (197). The distal convoluted tubule extends into connecting tubule to the collecting duct, which also has three functionally distinct segments – cortical, outer medullary and inner medullary (194). The collecting tubule is mainly comprised of principal (PC) and intercalated cells (IC), the latter subdivided into type A, type B and noncanonical (non-A, non-B) types. The principal cells are dispersed along the entire collecting duct, while type A ICs dominate in the outer medullary part and type B ICs are found in the cortical part (194). The main function of PCs is vasopressin-regulated water uptake via aquaporins (uniquely expressed *AQP2*), as well as aldosterone-regulated sodium reabsorption and potassium secretion (198). Type A ICs are responsible for acidification of the urine, as they express H+ ATPase on the apical surface (towards tubular lumen), and chloride-bicarbonate transporter AE1 (encoded by *SLC4A1*) on the basolateral side. Meanwhile, type B ICs act in reverse – they express chloride-bicarbonate transporter pendrin (encoded by *SLC26A4*) on the apical side of the cell, and the H+ ATPase on the basolateral (199). In addition to acid-base regulation, intercalated cells have been reported to phagocytose uropathogenic bacteria

(200). Collecting duct cells are considered to be plastic and intermediate PC-IC and noncanonical phenotypes have been observed, but not yet investigated in detail (194,200).

Kidney vasculature comprises diverse and specialized endothelial cell types, uniquely adapted to the local environment (glomerulus, cortex and medulla) (201). Renal endothelia consist of various large and small vessels, including lymphatics, and organ-specific glomerular and *vasa recta* endothelium cells. Glomerular endothelium is highly fenestrated to allow passage of fluid, and, similarly to podocytes, is decorated in thick negatively charged glycocalyx, which acts as a barrier to protein passage and inhibits immune cell adhesion. It has adapted to withstand high blood pressure, and interacts tightly with podocytes ensuring efficient glomerular filtration (190). The *vasa recta*, composed of phenotypically distinct descending and ascending segments (DVR and AVR), sustain corticomedullary osmotic gradient necessary for urine concentration. Running in parallel to the loop of Henle, their countercurrent blood flow to the tubular filtrate facilitates exchange of water and solutes. The arterial-like descending *vasa recta*, surrounded by vSMCs, lose water and gain solutes. In contrast, the fenestrated venous-like ascending part reabsorbs water from the medullary interstitium (190,201,202). The lymphatic vessels are primarily found in the cortex, where they remove fluid and macromolecules from the interstitial space, but not in the medulla, as the AVR represents a hybrid phenotype with lymphatic-like features (203). Overall, the vasculature of the kidney displays remarkable heterogeneity and specialization, essential for organism-level homeostasis.

Aside from organ-specific function executing elements such as the vasculature and epithelia, kidney harbors a rich immune cell landscape, strategically distributed to provide immune surveillance. Single-cell transcriptomics has been instrumental in understanding the immune cell composition and dynamics in the developing and adult kidney (204). Briefly, both lymphoid and myeloid lineage cells are present in the kidney under homeostatic conditions, although myeloid cells are much more abundant (183,205). Lymphoid compartment consists of CD4 and CD8-positive T cells, NK, NKT and B cells (204), although it is argued if these cells reside in the organ or arrive from circulation (183). Myeloid lineage cells found in the kidney are neutrophils, mast cells and mononuclear phagocytes, the latter comprising classical and non-classical monocytes, dendritic cells and macrophages. These phagocytic cells are enriched in the medulla and pelvis, where they fight against any bacteria arising from the ureter (205). Macrophages found in the kidney are diverse, monocyte-derived inflammatory (M1 polarization) and anti-inflammatory (M2 polarization), or

tissue-resident, arising from the yolk sac progenitors during embryonic development (204). M1 macrophages are characterized by pro-inflammatory cytokine (IL-1β, 6, 8, 12) and tumor necrosis factor (TNF-α) production, important in host defense, whereas the M2 macrophages produce IL-10 and tumor growth factor (TGF-β), involved in tissue remodeling and inflammation suppression (206). According to this classical dichotomy, prenatally seeded progenitor-derived macrophages represent the M2 phenotype (204).

In scRNA-seq studies, there is little agreement on the abundance of the aforementioned different cell types in the healthy kidney, most probably due to dissociation issues. Generally, healthy kidney atlases are dominated by proximal tubule cells, which typically constitute from 40% to >75% of all epithelial cells (204,207–209). Recovery of other tubular segments varies (5-20%), and certain cell types, such as podocytes and mesangial cells, are critically underrepresented (<1-5%), likely due to their fragility and difficulties dissociating glomeruli (204,207,208,210). Nonetheless, single-cell studies have greatly deepened the understanding of development, function and phenotypic diversity of cells in the healthy kidney (192).

1.2.2. Clear cell renal cell carcinoma

Kidney cancer is a group of malignancies of epithelial origin, with more than 10 subtypes characterized, the most prevalent ones being clear cell renal cell carcinoma (ccRCC) (more than 80% of all cases), papillary carcinoma (pRCC) (10-15% of cases) and chromophobe carcinoma (<5% of cases) (211). Kidney cancer incidence is particularly high in Northern and Central Europe, especially in Czech Republic and Lithuania (212). Risk factors for ccRCC are obesity, hypertension and tobacco, but they do not explain the geographic variation of disease incidence. A recent whole-genome sequencing study of ccRCC tumor specimens shed light on the mutational signatures from 11 countries and identified single base substitution (SBS) profiles associated with disease incidence (213). Interestingly, while a potential etiology (i.e., exposure to known lifestyle or environmental mutagens) has been suggested for a fraction of SBS signatures, the Lithuanian ccRCC mutational profiles are dominated by SBS40a and SBS40b, which do not yet have mutagenic exposure interpretation (213). Renal carcinoma incidence markedly increases with age and is about two times more prevalent in males than females. Due to asymptomatic nature, it is frequently detected accidentally during routine health checkups, and at a late stage. A review from 2017 reports 5-year survival rates for early 1st and 2nd stages at 95% and 88%, respectively, while prognosis worsens significantly for late 3rd and 4th stages, at 59% and 20%,

respectively (212). Standard treatment practice includes partial or radical nephrectomy, however, even with surgical intervention about a third of patients develop metastases, which are resistant to radiotherapy and chemotherapy (214). Classical options for advanced disease treatment include angiogenesis inhibitors targeting tyrosine kinases in the vascular endothelial growth factor (VEGF) signaling axis, as well as anti-proliferative mTOR inhibitors (211,214). The past decade has seen rapid progress in the development of targeted and immunotherapeutic strategies and their combinations for advanced and metastatic ccRCC treatment. Currently, best results are obtained with various combinations of immune checkpoint blockade (ICB) (anti-PD-1, anti-CTLA-4) and VEGF-axis targeted tyrosine kinase inhibitors (TKI) (215).

Response to treatment varies between patients, likely due to high heterogeneity of tumors, both at the level of genetic mutations in tumor cells and the constituents of the tumor microenvironment (discussed in section 1.3.3). The most frequent genomic alterations in clear cell renal cell carcinoma (ccRCC) include loss of chromosome 3p regions (in over 90% of cases) and mutations in the von Hippel–Lindau (*VHL*) gene (in more than 50% of cases) (211,216). These changes impair the degradation of hypoxia-inducible factors (HIFs), resulting in their abnormal accumulation even at sufficient oxygen levels, driving angiogenesis, proliferation and metabolic adaptation (217). Other frequent ccRCC mutations include members of epigenetic regulation machinery (*KDM5C, PBRM1, SETD2*) (218), as well as proliferation and migration pathway components (*PTEN, PIK3CA, MTOR*) (219). Tumor cells harboring various combinations of these and other mutations are spatially distributed in tumors, forming genetically distinct subpopulations, supporting branching tumor evolution model (220). That is especially relevant in the context of therapy, as treatment targeted to a particular signaling axis will only affect subpopulations harboring relevant mutations. A scRNA-seq study found metastatic ccRCC cell subpopulations with enhanced expression of either EGFR, Src or both signaling pathways, leading to development of drug combination regimen with increased efficacy over monotherapy in patient-derived xenograft *in vitro* and *in vivo* models (221). Thus, high resolution single-cell profiling technologies have tremendous potential to uncover renal tumor heterogeneity and advance therapy development.

To elucidate the molecular changes underlying renal carcinoma development, it is necessary to identify and characterize the cell population at the origin of tumorigenesis. Chromophobe renal carcinoma originates from oncogenic transformation of intercalated cells in the distal nephron, while the papillary and ccRCC arise from the proximal tubule epithelium (211,219).

Single-cell transcriptomic studies revealed that ccRCC tumor epithelium is highly transcriptionally similar to a subpopulation of rare *SLC17A3+, SLC7A13-, VCAM1+* proximal tubule cells (196,222). Interestingly, *VCAM1+* proximal tubule subpopulation with inflammatory gene signature expression has been observed in various kidney injuries, and is considered to represent a de-differentiated, progenitor-like phenotype (192). However, the tumor cells express significantly higher levels of other inflammation mediators, such as complement system molecules, which are associated with the degree of intratumoral macrophage infiltration (222). Thus, it is debated whether *VCAM1+* proximal tubule cells acquire further inflammatory characteristics giving rise to ccRCC cells, or if these changes appear independently (192). Additionally, the rare *VCAM1+* progenitor population has been reported to exist in homeostatic conditions and expand upon injury, and, utilizing fate tracing techniques, demonstrated to give rise to pRCC (223). These findings highlight the value of scRNA-seq for characterization of rare cell populations, and shed light on a potential cell of origin for ccRCC and pRCC.

1.2.3. The tumor microenvironment of ccRCC

The resolution offered by single-cell profiling techniques has influenced a conceptual shift in cancer research – the traditional tumor cell-focused view evolved into a systemic view of the tumor as an intricate, dynamic ecosystem, composed of diverse and interacting cellular players. Single-cell transcriptomics has been instrumental in revealing the phenotypic complexity and plasticity of the microenvironment of a multitude of tumors, including ccRCC (212). Aside from the intratumor mutational heterogeneity of cancer cells, ccRCC presents with highly immune and non-immune cell infiltrated tumor microenvironment (TME) that has an impact on disease progression, prognosis and treatment response (192). It has been established long ago that mononuclear cell infiltration alone is an independent predictor of patient survival in ccRCC (224), underscoring the importance of rigorous and high-resolution investigation of the TME, now widespread with the use of scRNA-seq.

Immune cells in the healthy kidney are responsible for pathogen surveillance, apoptotic cell clearance and overall maintenance of homeostasis (205). In cancer context, the immune cell phenotypes expand (6) and can undertake both tumor suppressing and tumor promoting roles. In ccRCC, the most abundant immune cells in the TME are T cells and macrophages (192,225).

Interestingly, in contrast to almost all other solid tumors, the degree of T cell infiltration in renal carcinoma correlates with poor, rather than positive prognosis (226). Multiple hypotheses have been proposed to explain this phenomenon, including ineffective antigen presentation, impaired effector functions, phenotypic heterogeneity, metabolic dysregulation and association with specific genomic alterations of tumor cells (227). Antigen presentation in tumors occurs in niches where antigen presenting cells (APC) reside, oftentimes similar to tertiary lymphoid structures (TLS). It has been suggested that TCF1+ stem-like T cell subpopulation in renal carcinoma gives rise to terminally differentiated T cells, and loss of APC-dense niches correlate with impaired T cell responses and disease progression (228). However, the origin of these stem-like T cells remains unclear and the proposed APC niche loss model has limitations (228). Evidently, mechanisms of impaired tumor immunity in ccRCC are multifaceted and highly complex. Tumor infiltrating T cells differ in clonality, cytotoxic potential and phenotype, hence, overall T cell infiltration degree is less informative than abundance of subtypes (i.e. the TCF1+ progenitors, regulatory T cells) (227,229). Therefore, single-cell profiling technologies are particularly relevant for ccRCC immune infiltrate investigation.

To this day, a multitude of CD4 and CD8 T cell phenotypes has been described in ccRCC TME, including proliferative, cytotoxic, helper, naïve, memory, regulatory and, most importantly, phenotypes at various stages of exhaustion (230–234). Using scRNA-seq and T cell receptor (TCR) profiling it has been shown that in ccRCC, CD8 T cells are more clonally expanded than CD4 T cells (235). Notably, many T cell clones found within the tumor were also detected in healthy adjacent tissues and blood circulation, suggesting that tumor-reactive T cells initially expand in the periphery before invasion and might hold disease monitoring potential (230,235). As mentioned, various subtypes of CD8 T cells in ccRCC have been described, but research mostly focused on exhausted phenotypes. T cell exhaustion is a common immune evasion mechanism exerted by cancer cells and other pro-tumorigenic cells in the TME (i.e. macrophages) via engagement of the inhibitory co-stimulatory receptors on T cells (227). Exhausted T cells are characterized by impaired proliferation and effector functions, and highly express inhibitory receptors encoded by *PDCD1*, *HAVCR2*, *LAG3* and *CTLA4* (230,231). Considering that exhaustion markers PD-1 and CTLA-4 are directly targeted by immunotherapy, high-resolution investigation of the exhausted phenotypes expressing these targets is especially clinically relevant, as it might provide clues into therapy success. A study utilizing mass cytometry on ccRCC samples described an overwhelming 22 T cell

phenotypes in the TME, and while most subpopulations were PD-1 positive, the expression of other inhibitory receptors (CTLA-4, TIM-3, 4-1BB) was limited to a few clusters only (225). This result highlights the need to assess the TME composition for personalized therapies, as in this case CTLA-4 targeting is unlikely to be effective. Additionally, the authors found correlation between the abundance of PD-1+ T cells, regulatory T cells and M2-like macrophages, providing insight into the coordinated fashion of immunosuppression in the ccRCC TME (225). Indeed, a study using scRNA-seq established that T cell exhaustion degree and abundance increases along advancing disease stage, contributing to progressive immune dysfunction (232). Moreover, inhibitory interactions between immunosuppressive macrophages and terminally exhausted CD8 T cells were associated with worse overall survival (232). Another single-cell profiling study revealed a presence of intriguing CD8 T cell subset co-expressing proliferative and exhausted transcriptional signatures, which was associated with non-responsiveness to anti-PD-1 immunotherapy, lower overall survival and higher histological tumor grade (230). Additionally, intratumorally expanded TCR clonotypes were shown to arise predominantly from the exhausted T cell populations (231), yet in another study they were seldom detected in peripheral blood (236), limiting disease monitoring potential. Overall, T cell dysfunction appears to plays a major role in immune evasion and progression of ccRCC tumors. Hence, T cell response to therapy, particularly ICB, is under active investigation. One study found a tissue-resident non-exhausted CD8 T cell population that expands upon ICB administration and correlates with better response to ICB combination regimens (231). Conversely, another study revealed that ICB exposure increased T cell checkpoint signature expression and anti-inflammatory signaling, postulated to hint toward a potential mechanism of adaptation underlying treatment resistance. Surprisingly, in the same study, terminally exhausted phenotypes were observed in responders (233). In summary, research suggests that T cell compartment has prognostic and predictive, as well as monitoring potential in ccRCC. Nonetheless, currently reported mechanisms of action and response are conflicting and remain to be delineated.

The myeloid compartment in ccRCC TME is composed of dendritic cells, both classical and non-classical monocytes, and heterogeneous populations of tumor-associated macrophages (TAMs), which are at the focus of current research efforts. Different phenotypes of TAMs are reported to play diverse roles in the TME, promoting angiogenesis, tumor growth, immune evasion, as well as exerting complete opposite – antitumor activities (227,231–233). Macrophage impact in ccRCC was appreciated even before single-cell

profiling era, as clinical studies have reported varying treatment outcomes with respect to macrophage infiltration of tumors (237,238). Particularly, macrophage signatures extracted from microarray data showed association with poor overall survival and decreased response to TKI treatment, and were enriched in high-risk patients (238). Thus, delineating phenotypic diversity and functions of macrophages in ccRCC TME is of clinical relevance.

Macrophages are tissue-resident or arise from circulating monocytes upon differentiation in response to local cues, such as chemokines and growth factors. These cells are conventionally divided into two polarization states – pro-inflammatory M1 phenotype, important in innate immune response and pathogen defense, and anti-inflammatory M2 phenotype, playing a role in tissue repair and immunomodulation (239). TAMs in various tumors, including ccRCC have been reported to challenge this dichotomy, as various intermediate or mixed phenotypes have been observed (6,192,225,230). A study utilizing mass cytometry found 17 TAM phenotypes, with diverse combinations of pro-tumor (CD163, CD204, CD206) and anti-tumor (CD169) marker levels (225). An interesting mixed phenotype was observed, characterized by expression of co-stimulatory ligands responsible for T cell exhaustion (CD273 and CD274), as well as chemokines CXCL10 and CCL8, involved in CD8+ and regulatory T cell chemotaxis. Counterintuitively, lower frequency of this phenotype was associated with shorter progression-free survival, while other, more M2-like macrophage populations were associated with worse outcomes (225). Another study reported a similar phenotype expressing CD273 and CD274, but with more pronounced M2-like features (*CD163*, *TREM2*, complement gene expression) (240). Importantly, this population associated with poor overall survival, whereas another, M1-like macrophage population was associated with better outcomes (240). While clear dichotomy of the M1 and M2 phenotypes most often cannot be established (230,231,233), it is generally accepted that TAMs are more M2-like as they express genes whose products are involved in immunomodulation (*VSIR*, *VSIG4*, *LGALS9*, *APOE*, complement genes). It has been reported that M2-like TAM abundance increases along advancing disease stage and these cells engage in inhibitory interactions with exhausted T cells, which associate with worse overall patient survival (232). Moreover, immunosuppressive C1Q+ TREM2+ APOE+ TAMs were shown to correlate with post-surgical disease recurrence (241), further highlighting the extent of involvement in disease progression. Of note, anti-inflammatory TAM subtypes, similarly to ccRCC tumor cells, express high levels of vascular endothelial growth factor (VEGF) (192,196,240). This may contribute to the differential responses to anti-angiogenic TKI therapies, which have been linked to macrophage

infiltration (238). Even though macrophage infiltration is considered to hinder TKI therapy response, it is likely that observed effects relate to the resolution of methods used in clinical studies – subpopulations hidden in bulk microarray data might drive opposing responses. Indeed, a scRNA-seq study revealed a mixed M1/M2 TAM subpopulation characterized by low class II major histocompatibility complex (MHC) receptor expression and high expression of angiogenesis and interferon signaling genes, which was associated with better progression-free survival after TKI therapy across multiple cohorts. These findings once again underscore the importance of high-resolution cell profiling technologies and provide insight into TAM phenotype-dependent response.

Macrophage involvement in more advanced, ICB-based therapies has also been investigated. One study reported that in a patient with complete ICB response, mixed phenotype TAMs, expressing *CD163,* had low infiltration across all tumor regions (231). Another single-cell study evaluated ccRCC TME composition and expression profiles in naïve and anti-PD-1 ICB and TKI receiving patients (233). The authors described various macrophage subpopulations with sparse expression of PD-1 ligands PD-L1 and PD-L2, but elevated expression of *LGALS9*, *VSIG4*, *VSIR*, which are also involved in T cell immune checkpoint and are associated with M2 polarization. Interestingly, in patients with partial response to ICB, macrophages exhibited a widespread shift towards a pro-inflammatory phenotype with elevated antigen presentation and proteosome function signatures. However, ICB-exposed versus naïve macrophages also exhibited elevated expression of the aforementioned M2-like genes. A puzzling upregulation of both pro- and anti-inflammatory properties in macrophages was postulated to potentially promote eventual resistance to ICB, often observed in ccRCC (233). Nonetheless, similarly to T cells, TAM involvement in ICB response is not yet well understood and requires further research.

The non-immune compartment of ccRCC TME, consisting of vasculature and stromal cells, remains much less characterized than the immune infiltrate, despite established involvement in disease progression and treatment response (242). Notably, the common *VHL* function loss in ccRCC results in excessive angiogenesis and highly vascularized tumor appearance, and angiogenesis-targeting TKI remains the main therapy option for advanced and metastatic disease treatment (211). Angiogenesis signatures extracted from microarray data effectively stratify patients, with high values associating with improved TKI response and better progression free and overall survival (238). Tumor vasculature in ccRCC is abnormal as compared to healthy kidney endothelium – lacking proper barrier integrity, involved in immune cell trafficking and

antitumor immunity suppression, as well as ECM remodeling (242). Single-cell studies have shed light on the heterogeneous nature of tumor endothelium in ccRCC (196,222,240,243,244). Most studies report two major phenotypes of tumor vasculature. Long et al., described *VCAM1*+ endothelium, expressing genes involved in immune cell trafficking and epithelial-mesenchymal transition (EMT), and *VCAM1*- population, associated with proliferation and vasculature development (243). Similarly, other groups reported presence of *ACKR1*+ (196) and *EDNRB*+ subpopulations (222). Additionally, in another study, the *ACKR1*+ endothelium signature was associated with worse overall and progression-free survival in the TCGA dataset (240). Of note, both *VCAM1*+ and *VCAM1*- endothelium had high numbers of inferred cell-cell interactions with tumor cells and TAMs, indicating involvement in TME modulation (243). In other reports, VEGF produced by cancer cells and macrophages was inferred to engage with endothelial receptors (*FLT1*, *KDR*, *NRP1*, *NRP2*) (196,240). Indeed, spatial transcriptomics revealed that collagen producing endothelial cells co-localize at the tumor-normal interface enriched in tumor cells and *IL1B*+ macrophages (234). A recent study on purified ccRCC tumor endothelial cells revealed notable differences to healthy kidney vasculature, including high expression of *IGFBP3*, which is involved in sprouting angiogenesis, and upregulation of ECM remodeling pathways (245). Interestingly, established primary cultures of tumor endothelium retained these properties, were resistant to cell death upon VEGF withdrawal and exhibited increased adherence of CD8+ T cells and monocytes as compared to normal counterparts (245). Together, these findings suggest an active tumor-promoting and immunoregulatory role of endothelial cells in the ccRCC TME and highlight the need of in-depth future investigation.

Stromal cell population in ccRCC is comprised of pericytes, vascular smooth muscle cells, fibroblasts and other mesenchymal stromal cells (240,244,246). These cells participate in ECM remodeling, cytokine secretion, angiogenesis, and may shape immune cell infiltration and function, as well as engage in direct communication with tumor and immune cells (247). Several single-cell profiling efforts have appreciated the roles and phenotypic heterogeneity of stromal cells in ccRCC TME. For instance, Alchahin *et al.* reported inflammatory capillary pericytes and fibroblasts, which expressed cancer-associated fibroblast (CAF) markers (*FAP*, *FN1*, *LRRC15*, *THY1*, and *TGFBI*) and associated with poor prognosis in two independent ccRCC cohorts (244). Another study detected three subpopulations of fibroblasts, marked by abnormal lipid metabolism in tumor specimens (240). Interestingly, one subpopulation had elevated activity of ZXDC, a TF regulating antigen presentation machinery expression, and associated with

poor progression-free and overall survival (240). In ccRCC bone metastasis samples, three pericyte and two mesenchymal stromal cell (expressing *NT5E*, *CXCL12*, *LEPR*) subpopulations were observed (246). Notably, a particular phenotype of the latter had high expression of collagen (types III, IV and VI) and was involved in RANK-RANKL signaling, essential for bone remodeling in metastasis. Moreover, this phenotype's signature correlated with poor progression-free and overall survival (246). Hence, stromal cells in ccRCC have prognostic value. Notably, increased abundance of CAF cells has been reported in recurrent RCC as compared to primary disease, and co-occurred with low CD8+ T cell infiltration (248). CAFs highly expressed LGALS1 (Gal1), which induced CD8+ T cell apoptosis in primary *in vitro* cultures. Moreover, *in vivo*, knockdown of Gal1 in CAFs suppressed tumor growth, reduced the proportion of apoptotic CD8+ T cells and enhanced infiltration, resulting in elevated efficacy of anti-PD-1 immunotherapy (248). Similarly, in another study, high expression of endosialin (CD248) by activated tumor-derived pericytes associated with low cytotoxic T cell infiltration (249). Inhibition of this glycoprotein increased T cell infiltration in ccRCC tumor-bearing mice and synergistically enhanced anti-PD-1 immunotherapy efficacy (249). Therefore, evidence suggest that stromal cells in the TME are actively contributing to tumor progression and impaired immunity, positioning them as an attractive target for combination therapies.

Given the evidence of tumor-promoting action by non-tumor cells in the TME, therapeutic strategies directed at various components of the TME are under active development. Having established that TAMs hinder T cell responses and promote immunosuppression favoring tumor growth, these cells are increasingly recognized as attractive targets for treatment. Strategies to deplete, reprogram or selectively kill tumor-promoting macrophages are actively being pursued (250). For instance, therapies in combination with ICB that activate CD40 on antigen presenting cells and inhibit TAM-expressed CSF1R have cleared early-phase clinical trials (251). Nonetheless, further research, especially functional studies involving *in vivo* models, is needed to elucidate exact action mechanisms and therapeutic intervention points of various TME components.

1.3. Human amniotic fluid (AF)

1.3.1. Biological and clinical significance of AF

Amniotic fluid (AF) is a clear, yellowish liquid that surrounds the fetus within the amniotic sac. The sac is composed of two layers: an inner layer of amnion epithelial cells and outer layer of vascularized chorion, both of fetal origin. AF is a highly complex and dynamic system that undergoes compositional and volume changes during pregnancy, tightly linked to the undergoing fetal development. AF maintains temperature and pressure homeostasis, provides mechanical cushioning, protecting the fetus from external forces, allows fetal movement and proper function of the umbilical cord. Aside from the physical protective role, it contains nutrients, growth factors, antimicrobial effectors and other biomolecules (252,253) (overviewed in section 1.4.2). The biological significance of AF in fetal development cannot be understated – as illustrated below, deviation in volume and composition might significantly negatively affect fetal development, while premature loss of contact to AF results in underdevelopment of lungs and gastrointestinal tract (254,255). Despite its immense importance, human amniotic fluid is understudied due to ethical constraints and technical challenges of sample acquisition, thus, most of current insights on the action of AF and its constituents are derived from animal models, harboring inherent translational limitations due to species-specific developmental differences. Nonetheless, progress in understanding AF physiology and composition has advanced prenatal diagnostic and therapeutic strategies, as well as opened new avenues for regenerative medicine and other clinical applications (255).

Amniotic fluid production starts as a filtration of maternal plasma through the fetal membranes at around the time of implantation. At 10 weeks of gestation, the volume of AF is ~25ml, while at week 20 it reaches ~400ml. Around week 8, the fetal kidneys start the production and excretion of urine. Shortly after, fetal swallowing begins – an important process required for proper gastrointestinal tract development and fetal nutrition (256). Until about week 20-22, before fetal skin keratinization begins, the AF and solutes in it can pass through the skin, fetal membranes and the placenta directly. Significant takeover of AF circulation by fetus becomes evident during the second half of the pregnancy (252). At week 27, the volume of AF reaches ~800ml and plateaus, before declining to ~400 ml before delivery. During the second and third trimester, after completion of fetal skin keratinization at around week 25, the main routes of abundant (>1L per day) AF circulation are

fetal swallowing and urination, intramembrane exchange, as well as excretion of oral, nasal, tracheal and pulmonary fluids (252,257).

The volume of AF itself is a clinical variable and a condition – accurate assessment of it is instrumental in evaluating fetal wellbeing and various pregnancy-related complications. For example, reduced amniotic fluid volume, termed oligohydramnios, may arise in case of maternal dehydration, preeclampsia, placental insufficiency, spontaneous rupture of membranes, as well as fetal renal and urinary tract anomalies, which substantially reduce urine output (255,257). This condition possesses serious risk to fetal development, as insufficient amount of fluid may result in umbilical cord compression, limiting nutrient uptake. Additionally, restriction to breathing and other movements might result in limb deformities and lung hypoplasia, likely due to intrapulmonary pressure loss and reduced entry of critical AF growth and signaling factors (255). Excess volume of AF, termed polyhydramnios, is likewise unfavorable, even though it is most often secondary to already existing fetal or maternal conditions. It can manifest due to maternal gestational diabetes, twin-twin transfusion syndrome, fetal anemia, and, most commonly, impaired fluid swallowing due to gastrointestinal obstructions (i.e. esophageal atresia) or neurological disorders, as well as the opposite – overproduction of fetal urine due to renal or cardiac anomalies (258). With recent advances in deep learning methods for medical image analysis, there is a rise in efforts to automate routine fetal biometrics and amniotic fluid volume measurements (259).

AF not only provides minimally invasive means to fetal health monitoring, but also serves as a valuable diagnostic medium. Amniotic fluid is obtained via transabdominal amniocentesis, it contains fetal cells, urine and lung secretions, which can be used for diagnostic purposes (255). Similarly to AF volume deviation, abnormal concentrations of various AF constituents can be used in the clinical setting and inform on underlying fetal pathologies. For example, fetal lung maturity can be assessed by determining the ratio of lecithin/sphingomyelin, surfactant/albumin and the presence of phosphatidyl glycerol in AF (254). Additionally, elevated AF acetylcholinesterase activity and α-fetoprotein levels are used as a means to evaluate potential neural tube defects (260) and Edwards or Down syndrome susceptibility. Fetal cells are mostly utilized for the detection of genetic abnormalities, such as chromosomal aneuploidies, deletions, duplications, as well as genetic variants related to various inherited conditions (255). Analysis of cell-free AF transcriptome to infer genetic, developmental, and environmental diseases is an area of active investigation and holds potential as well (261). However,

with widespread access to non-invasive prenatal testing (NIPT) in the developed world, frequency of amniocentesis procedures has seen a decline.

1.3.2. AF constituents and cellular composition

Human amniotic fluid is rich in bioactive molecules: carbohydrates, lipids, proteins and peptides, electrolytes, enzymes, lactate, pyruvate, as well as various hormones and growth factors (252,262). These AF constituents play important roles in fetal nutrition, as well as placental and fetal development. Fluid exchange is extremely important for lung development, while ingestion of AF promotes fetal growth, gastrointestinal tract development and intestinal epithelium differentiation (255). For instance, necrotizing enterocolitis (a life-threatening intestine inflammation and necrosis) primarily affects prematurely born infants, likely related to intestinal underdevelopment due to abruptly truncated exposure to AF. Moreover, the composition and concentration of various bioactive molecules in AF mirrors those in human breast milk, reflecting the nutritive and protective value of AF (253).

Solutes and amino acids present in the fluid diffuse from the placenta, and can cross the not-yet-keratinized fetal skin, as well as enter the digestive tract via fluid swallowing. The most important amino acids in AF are glutamine and arginine. The latter is a precursor for polyamines, which have a role in placental development and are associated with intestinal growth and function (254). Moreover, arginine stimulates nitric oxide production in the intestine, potentially contributing to vascularization and nutrient uptake (253). Glutamine, a key precursor for nucleic acid synthesis, is found at the highest concentration among other amino acids (263), and is particularly important for rapidly dividing intestinal mucosa cells. Interestingly, the concentration of both amino acids is lower in the serum of preterm infants who develop necrotizing colitis, even before the onset of symptoms (254), highlighting the importance of these AF constituents for gastrointestinal tract development and health.

Human AF also contains antioxidant vitamins, such as vitamin A and C, as well as folate and vitamin B12, which deficiencies have been linked to neural tube defects in fetuses (264). Especially important to fetal development and preparation for the postnatal life are the trophic growth factors present in AF. It is known to contain epidermal growth factor (EGF), transforming growth factor alpha (TGFα), transforming growth factor beta-1 (TGFβ1), insulin-like growth factor 1 (IGF1), fibroblast growth factor (FGF), hepatocyte growth factor (HGF), erythropoietin and granulocyte colony-stimulating factor (G-CSF) (254,255). The exact routes of action and significance of these

molecules in AF is difficult to investigate and remain speculative, mostly associated with general trophic effects (254,265). For instance, amniotic fluid was shown to promote growth of fetal intestinal cells *in vitro*, while addition of EGF, IGF, FGF, HGF and TGFα inhibitors partially reduced this effect, demonstrating the contribution of the aforementioned growth factors for proper intestinal development (266). Interestingly, while addition of recombinant growth factors promoted cell growth as well, the effect was less pronounced as compared to complete AF (266). These findings support the notion that AF is the primary source of trophic factors required for intestinal development and maturation. Importantly, the actions of certain factors appear to be coordinated with preparation for postnatal life. For instance, while EGF and other growth factor concentration correlates with gestational age, TGFβ1 is found only in late gestation and is thought to stimulate terminal intestinal cell differentiation and intestinal wound healing via promotion of cell migration (254). Taken together, while AF supports cell growth *in vitro* and harbors a variety of growth factors known to promote intestinal development, their action is likely synergistic and the relative contribution of each AF component *in situ* remains difficult to define.

AF also has an important defensive role as a part of innate immune system. It contains a vast array of antimicrobial, antifungal, anti-parasitic, anti-inflammatory and immunomodulatory compounds, as well as fetal immune cells (267). For example, active lactoferrin, lysozyme, calprotectin, psoriasin and alpha-defensins are found in vernix and AF (268), and maintain their activity against common pathogens *in vitro* even after removal of insoluble and cellular components (269). Concentration of these compounds increases significantly in preterm labor, premature rupture of membranes or infection. Generally, amniotic fluid is considered to be a sterile environment absent of any kind of microbiome or viriome. However, recent reports challenged this view with the discovery of intrauterine microbiome. Using 16S rRNA gene sequencing, microbial species were detected in amniotic fluid and meconium (the first stool of a newborn) (270), with meconium sharing features with amniotic fluid rather than maternal microbiome, indicating potential colonization *in utero* (271). Today, dozens of reports present contradictory findings with regards to the presence of intrauterine (fetal, placental, meconium, AF) microbiota and the topic is under active investigation. However, it was recently demonstrated that studies reporting the presence of bacterial sequences or even positive bacterial cultures from intrauterine samples suffer from various forms of unaccounted sample contamination (i.e. vaginal, fecal, skin), bioinformatic analysis ambiguities, as well as conceptual

discrepancies (272). Therefore, in absence of infection, the general consensus on AF sterility still stands.

The contribution of the cellular compartment in immunomodulation of the AF environment is less delineated, even though fetal immune cells are present in AF even in absence of infection. Fetal immune cell population in healthy pregnancy AF consists of mononuclear phagocytes (i.e. monocytes and macrophages), neutrophils, innate lymphoid cells (ILC), T cells and small numbers of NK and B cells (273). The composition of the immune cell compartment in AF changes along gestation. Flow cytometry analysis of AF leukocytes revealed that at week 15 to 20, the dominating immune cell populations are type 3 ILC and T cells; between week 20 to 30, B cells, neutrophils and monocytes emerge, T cells remain and ILC abundance starts to decline, with neutrophils dominating from week 30 to term. NK cells are more abundant between 15 to 30 weeks of gestation than at term, whereas monocytes and macrophages emerge around week 20 and remain constant until term (273). In the presence of intra-amniotic infection, all immune cell populations increase in abundance, except for ILC. These cells are the most abundant in the early weeks of the second trimester, express intraepithelial marker CD103 (gene *ITGAE*), and are phenotypically very similar to ILC found in fetal lung and intestines, likely reflecting their origins (274). Moreover, type 3 ILC from AF produce high levels of IL-17 and TNF upon stimulation *in vitro*, suggesting that they retain their activity upon escape from tissues to the fluid, and potentially participate in regulation of inflammation and intra-amniotic infection (274), even though they do not expand in these conditions (273). Fetal T cells in AF are also of mucosal origin, and constitute CD4 and CD8 positive populations, dominated by CD4 regulatory T cells that suppress T cell activity and responses against maternal antigens (275). Amniotic fluid B cells are a small, but constantly present population during the second and third trimester, they express CD5 and are considered to bridge the innate and adaptive immunity (273). While most immune cells in AF are of fetal origin, maternal immune cells can also be found in AF. For example, AF neutrophils can be predominantly either of fetal or maternal origin, or a mixture of both, especially during intraamniotic infection, when they increase in numbers significantly (273). During preterm gestation, neutrophils are primarily of fetal origin, whereas at term they are mostly maternal. Regardless of their origin, they actively participate in innate immunity host defense (276). Monocyte and macrophage origin, dynamics and activity in infection follow the same pattern as neutrophils. Potential source of fetal macrophages (Hofbauer cells) in AF is the placental villous tree (277). Interestingly,

phagocyte numbers are increased in AF of fetuses with neural tube defects (254).

Immune cells in AF likely participate in the events that lead to labor and delivery. Generally, the onset of labor is associated with increased concentrations of cytokines (i.e. IL-17, IL-8, IL-6), while spontaneous preterm labor and birth is often associated with intraamniotic inflammation (278). Immune cells are considered to be the source for these inflammatory components. For example, it was shown that fetal T cells, particularly CD4 positive population, increases in size, undergoes activation and secretes cytokines in the case of preterm birth in absence of inflammation (279). Therefore, immune cells are a dynamic and heterogeneous unit of the AF, actively participating in innate immunity and gestational processes.

Aside from immune cells, AF also harbors a heterogeneous population of cells that have shed away from the developing fetus. The origins of these cells were of researcher's interest as early as 1970-80s, as simple observation of AF cell morphology already indicated cellular heterogeneity. Majority of cells in AF are squamous epithelial cells, shed from the fetal skin, gastrointestinal tract, lungs, kidneys and bladder – organs that are in direct or indirect contact with AF (280–282). Colonic epithelial cells were also observed in AF (283). In certain conditions, the fluid can come in contact with remote tissues, i.e. neural tissue, and in that case glial cell presence in AF was also reported (282). The presence of amnion or trophoblast cells in AF is considered contemptuous (280). For AF cells that can establish colonies in culture, a term "amniocyte" was adopted and is still used in some scientific literature today. However, it was noticed early that the use of this term is misleading, as it suggests that a single cell type is established in culture, which is not the case (280). Additionally, AF supposedly harbors mesenchymal stem cells with broad differentiation potential (284–286), as well as lineage-restricted epithelial progenitors (287,288). The stem cells of AF are discussed in more detail in the next section.

1.3.3. AF as a stem cell niche

At the dawn of AF cell research in the late seventies, adherent AF cells were attempted to classify into three major groups, based on often arbitrary morphological criteria and growth characteristics. These groups are the flat E-type (epithelioid) cells, suggested to arise from fetal skin and urine; squamous-fibroblast intermediate AF-type (amniotic fluid specific) cells, supposedly originating from fetal membranes and trophoblast; and spindle-shaped F-type (fibroblastic) cells, hypothesized to originate from connective tissue and fetal

dermal fibroblasts (280,289). The proportion and growth dynamics of these subpopulations varies, with AF-type (~70% of all adherent cells) and E-type (~20%) co-existing at early culture stages and growing at slow or intermediate rate, and F-type (~10%), arising later with very high growth rate (280,289). However, this taxonomy is not necessarily consistent with regard to commonly accepted characteristics of cell types. For example, while in primary cultures both E and AF-type, but not F-type cells were positive for keratins, over 90% of further subcultured F-type fibroblastic morphology cells were also positive, which is inconsistent with the proposed mesenchymal phenotype (290).

Considering that AF cells readily grew in conventional culture conditions, there was a growing interest in the possibility that these might represent progenitor or stem cells, positioning AF as an ethical stem cell source for regenerative medicine applications. In 2003, several research groups reported the presence of mesenchymal stem cells in second and/or third trimester amniotic fluid. Prusa et al. demonstrated that ~0.1-0.5% of cells in AF express the pluripotency marker Oct-4 at the transcript and protein level, as well as stem cell factor CD117 (c-kit), mesenchymal marker vimentin and cyclin A (286). Intriguingly, Oct-4 positive cells were found in only about half of the samples analyzed. Others reported that cultured AF stem cells were positive for mesenchymal markers and class I MHC antigens, negative for class II MHC and hematopoietic markers, and readily differentiated toward fibroblasts, osteoblasts and adipocytes under defined culture conditions (284). Additionally, these cells were negative for endothelial markers (285). Roubelakis et al. further demonstrated that both round-shaped and spindle-shaped cells in primary AF cultures were negative for CD34, CD133, CD31, CD45, CD14 and HLA-DR, and positive for mesenchymal markers (CD73, CD105, CD166, CD29, CD44, CD49e) and class I MHC antigens. Interestingly, c-kit (CD117) was expressed at very low/undetectable levels in both populations, even though they could differentiate into multiple lineages and were positive for Sox2 and Oct-4 (291). Rahman et al. showed that AF mesenchymal stem cells express renal markers and are capable to endocytose exogenous albumin, suggesting that AFSC (amniotic fluid stem cells) are of renal origin (292). However, others postulate that these cells arise from the amnion membrane (293,294), thus, their under-investigated fetal tissue origins remain controversial.

Considering the aforementioned expression profile, adherence to culture dish plastic and robust differentiation toward multiple mesodermal lineages, AFSC are considered to be *bona fide* mesenchymal stem cells (MSC) (289). In 2007, de Coppi et al. isolated a rare c-kit (CD117) positive subpopulation

of cells from established murine and human AF cultures (~1% of cells), deemed pluripotent by authors (295). Unfortunately, in this work, comparison to previously reported multipotent mesenchymal AFSC was not made. Most current literature makes no distinction between c-kit+ AFSC and the stromal AFSC reported in 2003, assuming these cells are the same (296,297), while some authors consider them to be different based on growth characteristics and differentiation potential (289). Generally, c-kit+ AF cells share both the positive and negative marker profile with mesenchymal AFSC. It was demonstrated that c-kit+ AFSC have a very high proliferation rate, comparable to embryonic stem cells, and vastly broader differentiation potential than MSC (289,295). These AFSC were also positive for embryonic marker SSEA-4, retained long telomeres and stable karyotype for over 250 population doublings (295). Aside from Sox2, Oct-4 and SSEA-4, other embryonic stem cell markers were not detected.

The current consensus is that AFSC are not pluripotent, despite the expression of pluripotency markers, because they do not form teratomas *in vivo* (289). Yet, these cells can differentiate toward myogenic, adipogenic, osteogenic, endothelial, neural and hepatic lineages, showcasing that their differentiation potential is broader than bone marrow or other adult MSC (295,296). However, some of these results remain controversial, as the differentiation success is often measured assessing just a few markers and not every lineage specification (i.e. neural, hepatic) was confirmed at the functional *in vivo* level (289,298). Thus, AFSC are considered to reflect a broad multipotency state, which can be considered as an intermediate between pluripotency and multipotency (296).

Aside from mesenchymal AFSC, there have been efforts to characterize other progenitor cells in human AF. Da Sacco et al. proposed that AF cultures contain various organ (lung, liver, heart and kidney) progenitors, as some markers of these tissues were expressed in total AF cultures in bulk. The authors further detected metanephric mesenchyme cells, and from this cell line separated various kidney-specific cell type progenitors (i.e. podocyte, mesangial cell, tubule epithelial cell) (299). However, the characterization of these subpopulations' identity was limited to a handful of markers assessed via qPCR. Lesage et al. showed that some AF-MSC lines, when cultured on decellularized lung ECM, have lung characteristics (300). However, it remains elusive whether these cells exist in the native environment of AF, or if the observed phenotype represents a plastic cell response to specific culture conditions. Ditadi et al. reported the presence of c-kit+ hematopoietic progenitors in murine and human AF, and demonstrated the generation of erythroid, myeloid and lymphoid lineages *in vitro* and *in vivo* in

immunocompromised mice (294). However, while human hematopoietic progenitors gave rise to the three blood cell lineages *in vitro*, they failed to reconstitute the hematopoietic system *in vivo*. Intriguingly, both murine and human hematopoietic progenitors described in this study were negative for the canonical marker CD34, but expressed immune cell marker CD45 (294).

Generally, it appears that AFSC cell line derivation success, differentiation potential and marker expression profiles (both at the transcript and protein level) exhibit significant inter-sample or gestational age-attributed variability, further exacerbated by inconsistencies between individual reports (289,296,299). Despite some overlapping expression patterns of surface and pluripotency markers, it remains unproven whether all studies reporting on AFSC actually describe the same cells, and if the assigned characteristics are actually correct. Especially puzzling is the phenomenon of pluripotency gene expression without the capacity to form teratomas *in vivo* like embryonic or induced pluripotent stem cells do. In an excellent study, Ryan et al. critically examined methodologies on Oct-4 detection in various fetal MSC including AFSC, showing that most studies report Oct-4 messenger RNA or protein expression, but no study actually provides definitive evidence for the expression of functional Oct-4A, the true master regulator of pluripotency (301). *POU5F1* gene has multiple transcript isoforms, out of which only one generates the functional protein Oct-4A, moreover, a pseudogene on a different chromosome generates Oct4-B protein (gene *POU5F1B*) which is highly (96%) homologous and also localizes to the nucleus. Therefore, studies claiming Oct-4 expression at either or both transcript and protein levels do not actually target the "true" Oct-4 due to faulty primer design and non-specificity of antibodies used (301). Other pluripotency markers (i.e. *NANOG*) also have multiple pseudogenes, warranting similar re-analysis. Importantly, the authors further experimentally proved that using sound methodology no Oct-4 can be detected in fetal MSC (although AF as a source was not experimentally tested) (301). Indeed, a few years ago Vlahova et al. reported that AFSC, including the pure c-kit+ AFSC, do not express the "true" Oct-4 (302).

It is widely accepted that AF and AF cultures contain heterogeneous cell populations, yet the in-bulk characterization methods used (qPCR, Western blot) severely limit the resolution of investigation, and the single-cell methods (i.e. cytometry) are limited by the number of targets and other experimental factors (i.e. antibody specificity). Additionally, the use of rodent animal models for AFSC investigation raises several questions as the amnion development differs vastly as compared to primates (297). For instance, in the seminal c-kit+ AFSC isolation study by de Coppi et al. AF collection procedure is poorly described, while Ditadi et al. study methodology includes

“the collection of amniotic fluid samples from mouse embryos between E9.5-E19.5, after removing the maternal uterine wall to expose the amniotic sacs”, which, in rodents, likely corresponds to exposure of visceral yolk sacs, resulting in the collection of exocoelomic, and not amniotic, fluid (297). Thus, it appears that the methodological limitations, inconsistencies and ambiguities in AFSC research field provides room for further investigations that would benefit from the use of present day, high-resolution techniques such as scRNA-seq.

Stem cells in AF are of great interest not only for regenerative medicine applications, but also as a means to study fetal development and congenital diseases. Recently, de Coppi group described derivation of epithelial organoids from AF of intestinal, kidney and lung identities, which were examined using scRNA-seq (287). Babosova et al. also demonstrated the ability to culture AF-derived epithelial fetal kidney and lung organoids, analyzed their single-cell transcriptomic profiles and showcased kidney organoid ability to integrate and contribute to the developing nephrons in kidney explant cultures (288). In both reports, the organoids spontaneously assembled in highly specific culture conditions without pre-selection of progenitor populations, and had characteristic transcriptomic profiles of multiple organ-restricted cell lineages. Both groups claimed that the organoids must originate from organ-specific progenitors present in AF, even though their presence in uncultured AF was not established. On a side note, the aforementioned de Coppi study included a scRNA-seq dataset of uncultured human AF cells. Intriguingly, AFSC, described in 2007, were not present in the dataset; likewise, there were no cells with pluripotency marker expression. Instead, the authors claimed that a minute fraction of squamous epithelial cells represented the kidney, lung and intestine epithelial progenitors, based on superficial assessment (i.e. lacking standard scRNA-seq analysis steps such as DGE analysis or cell annotation).

Overall, characteristics and origin of AF cells, including the stem and progenitor populations, are far from resolved and research in this topic remains highly relevant. Moreover, little is understood on the fundamental biological implications – aside from AF cell potential in regenerative and personalized medicine applications, currently there is no interpretation for their plausible role in the native environment.

2. MATERIALS AND METHODS

2.1. Ethics statement

Experiments using human specimens were performed in accordance with the ethical standards of Helsinki Declaration. All patients who donated specimens provided informed consent. For lung adenocarcinoma and squamous cell carcinoma tissues from patients undergoing a surgical resection at Memorial Sloan Kettering Cancer Center (MSKCC), the biospecimen collection and analysis followed the Institutional Review Board-approved protocol. For paired ccRCC and kidney sample collection at the National Cancer Institute (Vilnius, Lithuania) a Vilnius Regional Bioethics Committee approval No.2019/2-1074-586 was granted. For AF collection at the Vilnius University Hospital Santaros Klinikos, Center for Medical Genetics, a Vilnius Regional Bioethics Commitee approval No. 2022/4-1429-900 was granted.

2.2. Sample collection and clinical information

Lung tissues. Lung adenocarcinoma (n=2) and squamous cell carcinoma (n=1) specimens from treatment-naïve patients were collected during surgery at MSKCC. Available clinical information is provided in **Supplementary Table S1.**

Kidney and ccRCC. Fresh ccRCC tumor (n=8) and healthy-adjacent (n=9) paired kidney tissues were obtained from the National Cancer Institute (Vilnius, Lithuania). All patients were treatment-naïve. Samples were collected during an open or laparoscopic, partial or radical nephrectomy surgery, placed on ice and rapidly (<1 h) transferred to the laboratory for processing. Sample T1 (tumor from patient P1) was highly necrotic, thus excluded from analysis. Clinical characteristics of all samples are provided in **Supplementary Table S2.**

Amniotic fluid. Fresh amniotic fluid samples (n=26) were collected via amniocentesis at the Vilnius University Hospital Santaros Klinikos, Center for Medical Genetics. These samples were primarily collected for genetic testing of fetuses, and only the remaining 1,5-3 ml per sample were used for research. Clinical information related to samples profiled is provided in **Supplementary Table S3.**

2.3. Sample preparation for scRNA-seq

Lung carcinoma. Each specimen (n=3) was cut into three ~5–10 mm^3 sized pieces. Then, the tissues were minced with surgical blades and dissociated for

15 min at 37°C on the GentleMACS Octo Dissociator with Heaters (Miltenyi) using Human Tumor Dissociation Kit (Miltenyi Biotec, Cat. No. 130-095-929). After dissociation, the cell suspension was passed through 35 μm Cell Strainer Snap Cap (TFS, Cat. No. 08-771-23) and subjected to red blood cell lysis (ACK buffer, Lonza, Cat. No. BP10-548E) for 2 min at room temperature. One lung adenocarcinoma sample was resuspended in PBS (Gibco, cat. no. 20012027) supplemented with 0.04% (w/v) BSA (Roth, Cat. No. 8076.20) and used fresh for single-cell encapsulation. The rest of the cell suspensions were stained with live cell dye (Calcein AM, Invitrogen, Cat. No. C3009) and PE-conjugated anti-human CD45 antibody (BioLegend, Cat. No. 368510) mixture, and sorted into CD45 + and CD45- compartments with BD FACS Aria II instrument. Next, the sorted cells were spun down for 5 min at 300g at 4°C in a swinging bucket centrifuge and resuspended in 90% methanol. The methanol-fixed cells were then transferred to −80°C. After 30 days, the suspensions were retrieved and cells were prepared for encapsulation. Briefly, cells in methanol were placed on ice for 15 min and then centrifuged at 1000g for 10 min at 4°C in a swinging bucket centrifuge. Supernatant was removed, leaving ~50 μl on the cell pellet. Next, the cells were resuspended in 400 μl of ice-cold Rehydration Buffer 1 (3× SSC buffer (Invitrogen, Cat. No. 15557044), 80 mM dithiothreitol (DTT) (TFS, Cat. No. R0861), 0.2% BSA, 1 U/μl RiboLock RNase Inhibitor (TFS, Cat. No. EO0381)) and transferred onto a centrifugal tube filter (Millipore, Cat. No. UFC30DV25), pretreated with 1% BSA. The column was centrifuged at 50g for 45 s at 4°C. The flow through fraction was discarded. The cell suspension retained on top of the filter (~50 μl volume) was washed two more times with an ice-cold Rehydration Buffer 1 and once with an ice-cold Rehydration Buffer 2 (1× SSC, 40 mM DTT, 0.1% BSA, 1 U/μl RiboLock RNase Inhibitor). Afterwards, the cells were retrieved from the filter membrane, counted under hemocytometer and resuspended in 1X DPBS (Gibco, Cat. No. 14190144) supplemented with 0.04% (w/v) BSA and 16% OptiPrep (Sigma-Aldrich, Cat. No. D1556).

Kidney and ccRCC. Tumor tissues were minced and dissociated in an automated instrument gentleMACS Octo Dissociator with Heaters (Miltenyi Biotec) using Tumor Dissociation Kit (Miltenyi Biotec, Cat. No. 130-095-929) as per manufacturer's instructions. Healthy-adjacent tissues were dissociated using Tissue Dissociation Kit I (Miltenyi Biotec, Cat. No. 130-110-201). After dissociation, red blood cell removal was performed using RBC lysis reagent (Miltenyi Biotec, Cat. No.130-094-183). Then, cells were washed three times in ice-cold 1X DPBS at 500g for 5 min at 4°C. Cell viability and count were assessed on a hemocytometer using Trypan Blue dye

(Gibco, Cat. No. 15250061). No further enrichment or selection of cells was performed and the cells were resuspended in 1X DPBS supplemented with 0.04% (w/v) BSA and 15% OptiPrep.

Amniotic fluid. Upon receiving the fresh samples on ice, the fluid was centrifuged at 300g for 5 min at 4°C. Supernatant was then transferred to clean Protein LoBind tubes (Fisher scientific, Cat. No. 05414206) for storage at -80°C. The cells were washed 2 times by resuspending in 1 ml 1X DPBS with 0.04% BSA and centrifugation at 300g for 5 min at 4°C. Cell viability and count were assessed on a hemocytometer using Trypan Blue dye. For cell barcoding, cells were resuspended in 1X DPBS supplemented with 0.04% BSA and either 10% 500K MW Dextran (SERVA Feinbiochemica, Cat. No. 18695) (samples F1-4), 15% OptiPrep (samples F10-17) or 0.02% Xanthan Gum (Sigma Aldrich, Cat. No. G1253) (samples F29-50).

For all experiments, final cell concentration was adjusted to 400k/ml to reach encapsulation $\lambda \sim 0.2$.

2.4. Single-cell RNA sequencing

The workflow of inDrops single-cell sequencing is described in Zilionis et al. (62) and Juzenas et al. (60). Briefly, it consists of cell and reagent preparation, encapsulation into nanoliter droplets in a microfluidics chip, reverse transcription reaction in droplets and sequencing library preparation on post-RT pooled material.

2.4.1. Barcoding bead design and preparation

Synthesis and barcoding of hydrogel beads (BHB), carrying the RT primers is described in Zilionis et al. (62). Prior encapsulation, BHBs were washed 5 times in 1 ml bead washing buffer (1X Maxima H minus RT buffer (TFS, Cat. No. EP0751) with 1% Igepal CA-630 (Sigma-Aldrich, Cat. No. I8896-50ML)) using a tabletop centrifuge. Washed beads were packed removing as much supernatant as possible and loaded into a 0.56 mm inner diameter PTFE tubing (hereafter mentioned as tubing; Atrandi Biosciences, Cat. No. MAN-TUB2), which was connected to 1ml luer lock syringe (Fisher Scientific, Cat. No. 1482330) pre-filled with 500 µl HFE-7500 oil (hereafter mentioned as oil; 3M, Cat. No. 98-0212-2929-3). The tubing with BHBs was then protected from light by inserting it into a black opaque tubing. In this work, BHBs with several RT primer designs were used, provided in **Table 2.1.** The inDrops-2 TS-v2020 BHBs were purchased from Atrandi Biosciences, (Cat. No. DG-BHB-C).

Table 2.1. RT reaction primers: BHB-carried mRNA capture primer and template-switching primer sequences. The underlined nucleotides indicate T7 promoter; the numbers indicate the cell barcode sequence.

Name	Used for	Sequence (5'→3')
inDrops-2 IVT vs TS comparison	LUAD sample (IVT vs TS comparison)	CGATGACGTAATACGACTCACTATAGGG ATACCACCATGGCTTTCCCTACACGACGCTCTT CCGATCT[12345678901]GAGTGATTGCTTGTGAC GCCAA[12345678]NNNNNNNN TTTTTTTTTTTTTTTTTTV; Second step RT primer for IVT approach: GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC TNNNNNN
inDrops-2 TS_v1	Lung carcinoma	CTACACGACGCTCTTCCGATCT[12345678]CATG [12345678]NNNNNNNN TTTTTTTTTTTTTTTTTTT
inDrops-2 TS_v2020	ccRCC, kidney, amniotic fluid	TACGGCGACCACCGAGATCTACAC[12345678]A CACTCTTTCCCTACACG[12345678]NNNNNNTTT TTTTTTTTTTTTTTTTTTTTTTTTTTTVN
TSO	All samples	AAGCAGTGGTATCAACGCAGAGTACATrGrGrG

2.4.1. Reverse transcription mix preparation

For single-cell barcoding experiments, reverse transcription mix (at final concentration in droplets) consisted of 1X RT buffer, 25μM TSO (**Table 2.1**), 0.5 mM dNTP mix (Thermo Scientific, Cat. No. R0192), 10 U/μl Maxima H Minus reverse transcriptase (Thermo Scientific, Cat. No. EP0751), 1 U/μl RiboLock RNase Inhibitor (Thermo Scientific, Cat. No. EO0382) and 0.3% Igepal CA-630 (Sigma-Aldrich, Cat. No. I8896-50ML).

2.4.2. inDrops experiment

Having prepared the cells, BHBs and RT mix, cell encapsulation was performed in a microfluidics device (Atrandi Biosciences, Cat. No. MCN-C5), schematics depicted in **Figure 2.1**. Cell suspension and RT mix were loaded into pre-chilled 1ml syringes with 500 μl oil, mounted into precision pumps (Harvard Apparatus PHD 2000) and connected to the device via tubing. A 1 ml syringe was filled with Droplet Stabilization Oil (Atrandi Biosciences, Cat. No. MON-DSO2), mounted into a precision pump and connected to the device

via tubing. Once all reagents were primed and connected to the device, encapsulation was performed at these flow rates: cells and RT mix at 250 µl/hr, BHBs at 100-150 µl/hr, controlling the occupancy to reach 80-90%, and droplet stabilization oil at 700 µl/hr. RT mix and cell suspension were chilled during encapsulation using a glove filled with ice. The process was monitored under an inverse bright field microscope (Nikon Eclipse Ti) using a Phantom high-speed camera. Emulsion was collected into a 1.5 ml Eppendorf DNA LoBind tube (Fisher Scientific, Cat. No. 13698791) on ice.

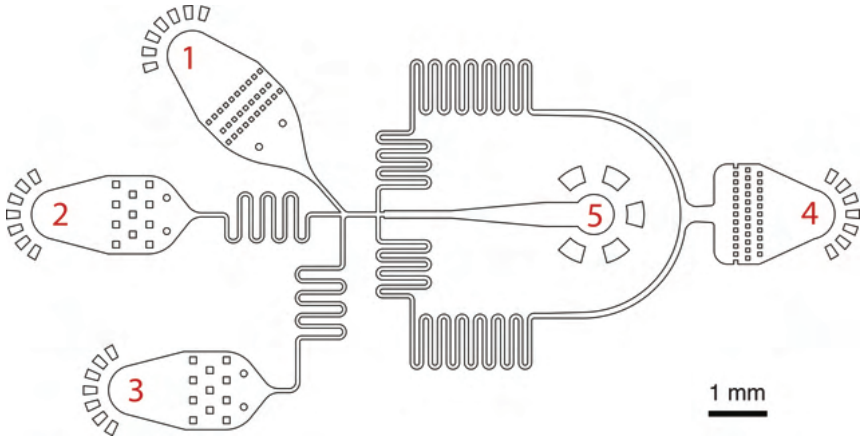


Figure 2.1. Design of a cell encapsulation and mRNA barcoding microfluidics device. The numbers indicate the barcoding hydrogel bead inlet (1), the cell suspension inlet (2), the RT and lysis mix inlet (3), the droplet stabilization oil inlet (4) and the collection outlet (5).

2.4.3. Reverse transcription

Following collection, the photocleavable RT primer release was performed by exposing the tube with an emulsion to a 350nm light either using LED device (Droplet Genomics, MHT-LAS1) for 20 s, or UV lamp (UVP, cat. no. 95–0127-01) for 5 min. Emulsions were aliquoted to contain 1000-5000 cells each. Then, the emulsions were transferred to a thermocycler to initiate cDNA synthesis: for IVT vs TS comparison, 42°C for 90 min; for all other samples 42°C for 60 min followed by 5 min at 85°C. Such cDNA-containing emulsion can be stored at -20°C prior sequencing library preparation.

2.4.4. Sequencing library preparation

For IVT vs TS experiment, detailed library preparation protocols for both approaches can be found in Juzenas et al. (60). For ccRCC, kidney and

amniotic fluid samples, library preparation process was conducted as follows. Primers used for library preparation are listed in **Table 2.2**. The emulsions were broken by adding up to 10% (v/v) emulsion breaker (Atrandi Biosciences, Cat. No. MON-EB1), and after a quick spin for 30 s at 300g, the supernatant was transferred onto a filter column (Zymo, Cat. No. C1004-250). The flow-through fraction containing barcoded cDNA was collected into a new 1.5 ml DNA LoBind tube by centrifugation for 1 min at 10 000g. The barcoded cDNA was purified twice with 0.8X AMPure XP magnetic beads (BeckMan Coulter, Cat. No. A63881) as per manufacturer's instructions. Next, cDNA was PCR amplified with KAPA HiFi Hot Start Ready Mix (Roche, Cat. No. KK2601) with 0.5µM of forward and reverse primers. PCR program is provided in **Table 2.3**. For DNA fragmentation and ligation, reagents and instructions from NEBNext® Ultra™ II FS DNA Library Prep Kit (NEB, Cat. No. E7805S) were used. Amplified DNA was fragmented using NEBNext Ultra II FS Reaction buffer and enzyme mix, for 8 min at 37°C, followed by 30 min incubation at 65°C. Then, fragmented DNA was purified by performing double size selection (0.6X-0.8X AMPure magnetic beads). Adapter ligation was performed using NEBNext Ultra II Ligation Master Mix and Enhancer, and 0.05 µM final concentration of ligation adapter, for 15 min at 20°C (**Table 2.2**). Next, ligated material was purified with 0.8X AMPure. Finally, the libraries were amplified by indexing PCR (**Table 2.4**) in a reaction mix comprising 1X KAPA HiFi Hot Start Ready Mix (Roche, Cat. No. KK2601) and 0.5 µM p5 and p7 indexes (**Table 2.2**). Once again, double size selection was performed using 0.6X-0.8X AMPure beads and final libraries were eluted in water. Library quality post 1st PCR and after final indexing was assessed using Bioanalyzer DNA High Sensitivity chip (Agilent, Cat. No. 50674626).

Table 2.2. Sequencing library preparation primers. *- indicates phosphorothioate bond. REV – reverse, FWD – forward.

Name	Sequence (5'→3')
cDNA amplification primers	
REV cDNA primer	AAGCAGTGGTATCAACGCAGAG
FWD cDNA primer	TACGGCGACCACCGAGATC
Ligation adapter	
Ligation adapter duplex	/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCA C/3ddC /5AmMC6/GCTCTTCCGATCT

Indexing PCR primers	
FWD PCR index primer p5	AATGATACGGCGACCACCGAGATCTACA*C
p7 index 1	CAAGCAGAAGACGGCATACGAGAT AACCTG GTGACTGGAGTTCAGACGTG*T
p7 index 2	CAAGCAGAAGACGGCATACGAGAT CCAAGT GTGACTGGAGTTCAGACGTG*T
p7 index 3	CAAGCAGAAGACGGCATACGAGAT GGTTCA GTGACTGGAGTTCAGACGTG*T
p7 index 4	CAAGCAGAAGACGGCATACGAGAT TTGGAC GTGACTGGAGTTCAGACGTG*T
p7 index 5	CAAGCAGAAGACGGCATACGAGAT ACCACT GTGACTGGAGTTCAGACGTG*T
p7 index 6	CAAGCAGAAGACGGCATACGAGAT CAGTGG GTGACTGGAGTTCAGACGTG*T
p7 index 7	CAAGCAGAAGACGGCATACGAGAT GTTGTC GTGACTGGAGTTCAGACGTG*T
p7 index 8	CAAGCAGAAGACGGCATACGAGAT TGACAA GTGACTGGAGTTCAGACGTG*T

Table 2.3. cDNA amplification PCR program.

Step	Temperature	Time
Initial denaturation	98°C	00:03:00
Denaturation	98°C	00:00:15
Annealing	67°C	00:00:20
Extension	72°C	00:01:00
Go to step 2, 15 cycles (16 in total)		
Final extension	72°C	00:01:00
Hold	4°C	Hold

Table 2.4. Indexing PCR program.

Step	Temperature	Time
Initial denaturation	98°C	00:00:45
Denaturation	98°C	00:00:20
Annealing	54°C	00:00:30
Extension	72°C	00:00:20
Go to step 2, 10 cycles (11 in total)		
Final extension	72°C	00:01:00
Hold	4°C	Hold

2.4.5. Sequencing

The final inDrops-2 (IVT) libraries were sequenced on the NextSeq550 and HiSeq2500 (Illumina) instrument (Read 1: 54 cycles; i7: 8 cycles, Read 2: 35 cycles or more). The inDrop-2 (TS) libraries were sequenced on the MiSeq, HiSeq2500, NextSeq550 and NovaSeq6000 (Illumina) platforms, without PhiX spike-in. The sequencing parameters were Read 1: 28 cycles; i7: 8 cycles, Read 2: between 35 and 92 cycles (depending on instrument and sequencing reagent kit).

The kidney and ccRCC libraries were sequenced on Illumina NextSeq 550 platform in multiple batches using either NextSeq 500/550 High Output Kit v2.5 (75 Cycles) (Illumina, Cat. No. 20024906) or NextSeq 500/550 High Output Kit v2.5 (150 Cycles) (Illumina, Cat. No. 20024907). Two batches of beads were used, and the sequencing parameters were Read 1: 51 cycles, Read 2: 35 cycles, i7: 6 cycles for the first batch. For the second it was Read 1: 16 cycles, Read 2: 62 cycles, i7: 6 cycles and i5: 8 cycles.

The amniotic fluid libraries were sequenced on Illumina NextSeq 2000 machine using NextSeq™ 1000/2000 P2 XLEAP-SBS™ Reagent Kit (100 Cycles) with the following settings: Read 1: 16 cycles, Read 2: 108 cycles, i7: 6 cycles, i5: 8 cycles.

2.5. Data analysis

2.5.1. Pre-processing

The goal of raw sequencing data pre-processing is to construct a cell x feature matrix. It consists of 1) sample and cell barcode demultiplexing (correction); 2) read mapping to the reference; 3) feature assignment; and 4) UMI deduplication. Throughout this work, solo-in-drops pipeline, (https://github.com/jsimonas/solo-in-drops), which is a wrapper around STARsolo (79) was used for that. MultiQC was used to assess the quality of sequencing library fastq files and STARsolo summary statistics.

For inDrops-2 IVT and TS libraries, STAR v2.7.10a was used to map reads to the human GRCh38 genome (GENCODE v41 annotation). STARsolo parameters were: --soloFeatures *GeneFull*; --soloType *CB_UMI_Complex*, --soloCBmatchWLtype *EditDist_2*, --soloUMIdedup *Exact.*

For methanol-fixed hashtag lung carcinoma libraries, the pre-processing was done using a combination of SEQC (6) and CITE-seq-Count (303) pipelines. The SEQC with default parameters was used to obtain cell x count

matrices, while CITE-seq-Count with the default parameters and --*no_umi_correction* was used to count hashtag sequences.

For kidney and ccRCC libraries, STAR v2.7.6a was used with the following parameters: --soloMultiMappers *Uniform*, --soloType *CB_UMI_Simple*, --soloUMIfiltering *MultiGeneUMI*, and --soloCBmatchWLtype *1MM*. Homo sapiens (human) genome assembly GRCh38 and Ensembl v93 annotations were used as the reference.

For amniotic fluid libraries, STAR v2.7.10a was used to map reads to the human GRCh38 genome (GENCODE v41 annotation). STARsolo parameters were: --soloMultiMappers *Uniform*, --soloType *CB_UMI_Complex*, --soloUMIfiltering *MultiGeneUMI*, --soloUMIdedup *Exact* and --soloCBmatchWLtype *EditDist_2*.

2.5.2. Quality control and doublet cleanup

Starting with cell x gene matrices, analysis was performed with Python using scanpy toolkit (73). Briefly, each library was filtered on total counts and percentage of mitochondrial counts, upon evaluation of total count distributions, as shown in **Figure 1.3**. For IVT vs TS comparison LUAD dataset, these values were set at 400 UMI per cell and 15% mitochondrial counts. For lung carcinoma dataset, cells were classified as singlets or doublets based on hash counts and assigned to corresponding samples with HashSolo (304), and already filtered matrix output was used. For kidney and ccRCC, all libraries were filtered at 20% mitochondrial counts and 400 UMI per cell, except for libraries T3.1, T9.1, N3.3, N4.3, N2.3, where 300 UMI threshold was used. For amniotic fluid libraries (61 sequenced in total), 15% mitochondrial count threshold and 400 UMI per cell was applied, except for some libraries (names are arbitrary as appear in lab notes): F10.1, F15.2x2 – 500 UMI, F34.3 – 800 UMI, and F34.2, F34, F47.1, F49.1, F49.2, F50.2, F13_1x3, F32, F29.2x2, F31.1_2x2, F46.1x2, F48.2x2 – 1000 UMI per cell.

Doublets were removed using Scrublet (v0.2.3) (97) algorithm in the same PCA space used for initial UMAP construction (general embedding construction is described in section 2.5.3). Scrublet was applied on each emulsion separately. Briefly, the procedure comprised of 1) calculating doublet scores for each cell in each emulsion using Scrublet; 2) very high-resolution graph-based clustering using scanpy's Louvain algorithm (resolution=60 for kidney and ccRCC, resolution=40 for AF); 3) evaluation of mean doublet score and fraction of predicted doublets per cluster; 4) manual inspection of doublet-rich clusters in the interactive SPRING application (119), 5) removal of clusters with high mean doublet score and doublet

fraction. For ccRCC and kidney dataset, the procedure was performed in two rounds and 913 cells were removed. Additionally, for this dataset, transcriptomes with >1% of total raw counts originating from hemoglobin genes (*HBB*, *HBA1*, *HBA2*, *HBD*) were considered as RBCs and removed from further analysis (47 cells). For AF samples, the doublet removal was performed once and 325 cells were removed. Additionally, for this dataset, a low-quality Louvain cluster was removed, comprising 843 cells. For IVT vs TS comparison, doublet removal was not performed. For lung carcinoma, doublet removal was based on hash counts and performed with HashSolo.

2.5.3. Embedding construction and clustering

Here, the general procedure for UMAP construction is described, and the exact parameters used for each embedding presented in this work are given in **Table 2.5**. This process is highly iterative, and the final parameter combinations used were selected upon inspection of many variations. Throughout this effort, an indispensable tool to assess the quality of the resulting embedding, explore data and clustering outcomes was used – an interactive, locally launched SPRING web interface (119).

Briefly, upon QC and doublet removal, final matrices were subjected to UMAP construction. The procedure consisted of 1) normalization to 10,000 total counts (CPTT), log-transformation and scaling; 2) selection of highly variable genes based on Fano factor as in Klein et al. (2); 3) PCA; 4) batch correction using Harmony (115); 5) neighborhood graph construction and 6) UMAP representation. After normalization, genes with at least *n_counts* (CPTT) in not less than *n_cells* were considered abundant and retained. Next, mitochondrial and ribosomal genes were excluded from highly variable gene candidates, and top *n_var* abundant and highly variable genes, based on Fano factor were used for PCA. PCA was performed in an iterative fashion on shuffled matrix to calculate the largest eigenvalue over 10 permutations, which was used as a threshold to retain eigenvalues above it, as described in (2), provided as *num_PCs* in **Table 2.5**. To remove batch effects, scanpy.external.pp.harmony_integrate() was used on *batch_variable*. Then, adjacency graph was constructed using sc.pp.neighbors() with parameter *n_neighbors* and UMAP representation was built using sc.tl.umap() parameter *min_dist*.

Clustering was performed either using graph-based spectral clustering with sklearn.cluster.SpectralClustering() function, scanpy Leiden implementation via sc.tl.leiden() function or PhenoGraph Leiden implementation via sc.external.tl.phenograph() function. Spectral clustering divides the graph into

a pre-determined number of clusters, and Leiden uses a resolution parameter, both provided as *resolution* in **Table 2.5**.

Table 2.5. Parameters used to construct UMAP embeddings, clustering approach used and resolution.

Name	*n_counts*	*n_cells*	*n_var*	*num_PCs*	*batch_variable*	*n_neighbors*	*min_dist*	*Clustering approach*	*resolution*
LUAD IVT vs TS **Figure 3.4**	10	10	2000	28	library	20	0.5	leiden	0.6
Lung carcinoma **Figure 3.5**	10	10	2000	55	-	30	0.5	PhenoGraph leiden	2
Lung non-immune **Figure 3.6**	5	10	2000	48	-	50	0.3	PhenoGraph leiden	0.6
Lung myeloid **Figure 3.7**	5	10	2000	36	-	50	0.3	PhenoGraph leiden	0.5
Lung lymphoid **Figure 3.8**	5	10	2000	21	-	50	0.3	PhenoGraph leiden	0.8
Kidney and ccRCC **Figure 3.10**	15	25	2000	71	beads	30	0.4	spectral	43
AF all cells **Figure 3.21**	10	20	2000	128	library	30	0.5	leiden	1
AF immune **Figure 3.22**	10	10	2000	43	library	20	0.6	spectral	16
AF non-immune **Figure 3.24**	10	10	2000	151	library	20	0.5	spectral	20

2.5.4. DGE analysis and cell annotation

To obtain cluster marker genes, differential gene expression analysis, comparing a given cluster to the rest of cells was performed (Mann Whitney U test with Bonferoni-Hochberg correction). Prior testing, genes were filtered based on abundance, using the same *n_counts* and *n_cells* values as in UMAP construction (**Table 2.5**). For AF samples, mitochondrial genes were also

excluded. Top 50 marker genes for each cluster (adjusted p value < 0.05) were used for in-depth literature analysis and manual cell type annotation.

2.5.5. CellTypist label transfer

For kidney and ccRCC dataset, CellTypist (137) label transfer was used to examine the similarity of assigned cell types to established phenotypes from publicly available datasets. For that, CellTypist models were trained in HPC cluster according to a tutorial available at https://www.celltypist.org/. For endothelial cells, Goveia et al. (305), endothelial cell scRNA-seq matrix and metadata was obtained from https://endotheliomics.shinyapps.io/lung_ectax/, the matrix was log-normalized, non-tumor endothelial and patient #5 specific cells were excluded. Then, the model was trained on the dataset without gene filtering and applied via celltypist.annotate() function to endothelial cell log-normalized matrix with parameter *majority_voting=True*. Similarly, a model was trained on Zhang et al. (222) dataset obtained from GEO (at GSE159115). The dataset was filtered to kidney and ccRCC epithelial cells without gene filtering and the model was applied for label transfer to our epithelial cell log-normalized matrix. To aid AF immune cell analysis, models were manually trained on published fetal and adult intestinal atlas (306), as well as fetal lung immune atlas leukocytes (307) and applied in the same manner.

2.5.6. Gene set over-representation analysis

For kidney and ccRCC, gene set over-representation analysis on top 100 DEGs was performed using gene sets obtained from the Hallmark Pathways of the MSigDB database v7.5.1. Marker genes were submitted to a hypergeometric test implemented in the enrichGO() function of the clusterProfiler R package, using genes that had nonzero UMI counts as a background reference. The pathways with Benjamini-Hochberg FDR values below 0.05 were considered as significantly over-represented.

For AF non-immune clusters, gene set over-representation analysis on top 200 DEGs was performed using Gene Ontology Biological Process 2023, MSigDB Hallmark 2020 and Reactome 2022 databases accessed via the gseapy package (308), using function gseapy.enrichr(). It also performs a hypergeometric test with Benjamini-Hochberg multiple hypothesis testing, using all annotated genes as a background reference. The terms filtered by adjusted p-value <0.05 were plotted using gseapy dotplot() function.

2.5.7. Sample heterogeneity quantification

For kidney and ccRCC dataset, Shannon entropy of samples was calculated for each broad cell category to assess heterogeneity as described in Chan et al.(309). Briefly, in each cell group (stromal, endothelial, tumor, lymphoid, myeloid, epithelial and cycling) entropy values were calculated for sample frequency. To account for differences in cell quantity per group, 100 cells were subsampled from each group 100 times with replacement, and Shannon entropy was calculated using function *scipy.stats.entropy()* from the scipy package. Cells from "Tumor cells 1" cluster were excluded, as they were sample specific.

2.5.8. Receptor-ligand interaction analysis

For kidney and ccRCC dataset, cell-cell interactions were inferred using CellphoneDB v.2.0.0 (156) with method "statistical_analysis" and default parameters. Log-normalized expression values for all cell types, excluding healthy epithelial cell populations and cycling cells were used. Significant (p value < 0.05) cell-cell interactions were explored and manually selected for plotting (**Figures 3.14, A**; **3.16, A**; **3.20, A** and **Supplementary Figure S3**). Cell-cell interaction signatures for survival analysis (as in **Figure 3.14, B**) were assembled by taking both the receptor and ligand genes.

2.5.9. Survival analysis

Survival analysis for selected gene signatures from kidney and ccRCC dataset was performed on TCGA KIRC cohort bulk RNA-seq (upper quartile FPKM normalized) and clinical data, downloaded from the NCI GDC Data Portal (310) via TCGAbiolinks R package. Cell type signature scoring was performed by calculating the mean value of the z-scored TCGA bulk RNA-seq expression values for all genes in a given signature. The association between signature score and overall survival time was assessed by Kaplan-Meier and multivariate Cox regression analyses. Log-rank tests and Wald tests, respectively, were used to evaluate statistical significance (at level of 0.05) of the performed survival analyses. For the Kaplan-Meier analysis, signature was stratified into high (greater or equal than the median signature score) and low (lower than the median). For the multivariate Cox regression analysis, the continuous signature score values were used with patient age and sex as covariates. The survival analyses were conducted using the survival and the survminer R packages.

2.5.10. Data and code availability

scRNA-seq data from inDrops-2 development (LUAD and lung carcinoma) (60) is deposited in in the European Nucleotide Archive (ENA) under accession number PRJEB71611. Raw and processed kidney and ccRCC scRNA-seq data from Zvirblyte et al. (311) is deposited in Gene Expression Omnibus (GEO) repository at GSE242299. Publicly available datasets used in kidney and ccRCC analysis in were downloaded from GEO (at GSE159115) and https://endotheliomics.shinyapps.io/lung_ectax/. Amniotic fluid dataset will be deposited into one of the aforementioned archives upon publication. The publicly available data used in AF analysis was downloaded from https://www.gutcellatlas.org/ and https://fetal-lung-immune.cellgeni.sanger.ac.uk/.

Jupyter notebooks for kidney and ccRCC analysis and visualization are provided at https://github.com/zvirblyte/2023_ccRCC. Most of the notebooks presented in this directory were taken directly and adapted from https://github.com/AllonMKlein/Pfirschke_et_al_2021 (8). They were re-used for LUAD, lung carcinoma and AF analysis as well. Any additional notebooks resulting from AF analysis will be made available upon publication.

3. RESULTS

The results of this thesis are structured into three major parts. The first part presents an improved single-cell RNA sequencing method inDrops-2 and its initial application for clinical sample profiling. The second part presents the use of this method for in-depth single-cell analysis of healthy kidney and kidney carcinoma clinical samples. In the third part, scRNA-seq is used to generate a human amniotic fluid transcriptional atlas.

3.1. inDrops-2: an improved single-cell RNA sequencing method

Upon original side-by-side publication of high-throughput droplet-based scRNA-seq methods inDrops (2) and Drop-seq (3) the technology was quickly commercialized (i.e., 10X Genomics Chromium platform) and gained popularity worldwide. To this day, single-cell analysis has greatly advanced our understanding on various complex biological systems, such as cancer (6,7) at unprecedented resolution and scale, and enabled outstanding discoveries of novel cell types in the human body (4,5). ScRNA-seq is the leading technology for large-scale international efforts cataloguing cellular diversity in various tissues and diseases (9,312), and is likely to remain highly relevant in the foreseeable future.

While commercial systems are attractive due to outstanding reproducibility and sensitivity, the open-source methods, such as inDrops, offer higher flexibility and lower cost. That is especially relevant to large-scale single-cell profiling efforts and research groups with limited financial resources. However, as compared to the commercial counterparts, open-source systems often suffer from lower transcript and gene capture rates (313). Moreover, it is well known that extended encapsulation times required result in undesirable transcriptional changes. Thus, in the development of inDrops-2, major attention was given to optimizations leading to substantial increase in method sensitivity, implementation of a more user-friendly template switching-based library preparation protocol, as well as establishment of compatible long-term cell preservation protocols. The utility of the updated method was demonstrated by single-cell analysis of preserved lung carcinoma clinical samples, revealing rare cell phenotypes. Overall, inDrops-2 (60) development was a collaborative effort involving various other technical optimizations and improvements that are out of scope of this thesis. In the following sections, results directly related to the remainder of this work are covered: the transition to a TS-based library preparation design (section 3.1.1)

that was used for the initial analysis of clinical lung tissues (section 3.1.2), as well as for profiling all the other samples in this thesis (sections 3.2 and 3.3).

3.1.1. Comparison of IVT and TS approach for primary cell profiling

The seminal inDrops method utilized IVT-based linear cDNA amplification approach (2). Here, upon mRNA barcoding RT reaction in droplets, the emulsion is broken and recovered cDNA is subjected to amplification via IVT, enabled by the presence of T7 RNA polymerase promoter sequence on the RT capture oligonucleotides. The following library preparation consists of chemical RNA fragmentation, a second RT reaction, and a few cycles of indexing PCR (62). In contrast, another seminal scRNA-seq method, Drop-seq (3), as well as the widely used commercial solutions, utilize a TS-based approach. Here, intrinsic terminal transferase activity of the reverse transcriptase results in non-templated nucleotide addition at the end of the synthesized cDNA molecule, serving as an annealing site for the TSO, enabling template switching. After this process, all cDNA molecules contain a universal sequence, exploited for exponential PCR amplification. The sequencing library preparation then involves enzymatic fragmentation of the cDNA, adapter ligation and indexing PCR. Both approaches have been successfully used for scRNA-seq library generation, however, the library preparation with TS is significantly shorter, more user-friendly, and widely used in the field (3,18,314), motivating the implementation into inDrops method as well.

For a direct comparison of TS and IVT-based library preparation, with a focus on clinical sample profiling, single primary lung adenocarcinoma (LUAD) cells were encapsulated into 1 nl droplets together with barcoding beads carrying T7p containing RT primers, lysis and RT reagent mix, and 25 μM TSO. The resulting emulsion, containing ~5000 cells, was subjected to RT reaction at 42°C for 90 min and then divided into two equal parts, processed in parallel following either the TS or IVT library preparation protocols, hereafter referred to as inDrops-2-TS and inDrops-2-IVT (**Figure 3.1**).

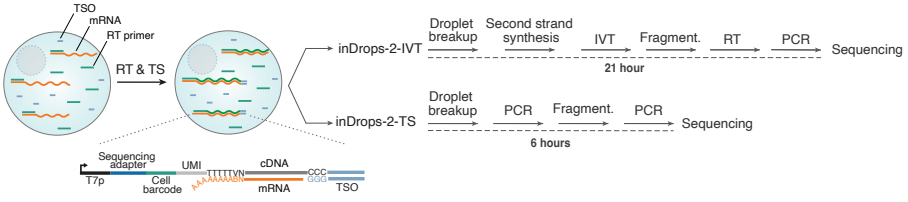


Figure 3.1. Schematics depicting the workflow for direct comparison of template-switching and *in vitro* transcription approach.

Key sensitivity metrics for scRNA-seq methods are transcript (UMI) and gene capture per cell. For faithful comparison, sequencing depth has to be taken into account, hence both libraries were downsampled to contain 20 000 raw reads per cell. The TS approach recovered a median of 2291,5 UMIs and 1081 genes, while the IVT-based library preparation resulted in 2197 and 1327 median UMIs and genes, respectively (**Figure 3.2, A**). Interestingly, inspection of the read mapping statistics with respect to different regions of a gene revealed a tendency of inDrops-2-IVT to recover more introns and 3'UTR exons, whereas the inDrops-2-TS captured more exons from protein coding sequences (**Figure 3.2, B**).

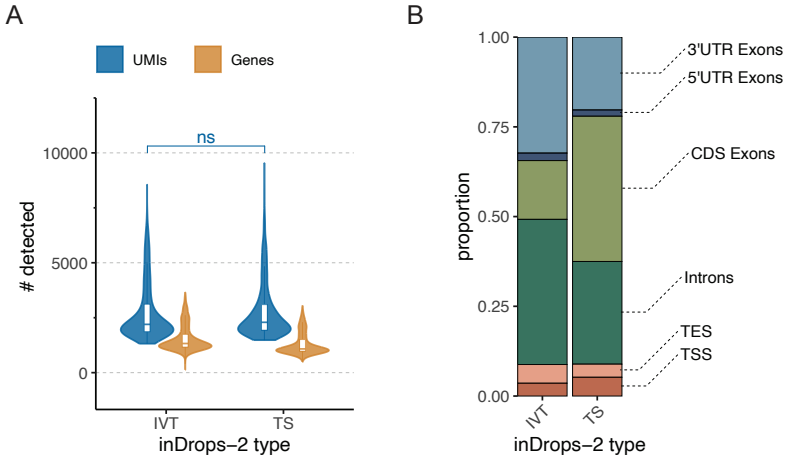


Figure 3.2. Comparison of IVT and TS inDrops-2 library preparation approaches. **A** – Number of UMIs and genes detected in primary lung adenocarcinoma cells in scRNA-seq libraries prepared by inDrops-2-IVT and inDrops-2-TS. Sequencing depth was normalized to 20 000 raw reads per cell. Boxplots display median (center point), first and third quartiles (lower/upper hinges), 1.5 × interquartile range (lower/upper whiskers). t-test used for comparison revealed no statistical significance ($p>0.05$). **B** – Fraction of reads mapping to different regions of a gene. CDS – coding sequence, TES – transcription end site, TSS – transcription start site, UTR – untranslated region.

Additionally, it was observed that inDrops-2-IVT has a tendency to recover more transcripts of longer genes, whereas inDrops-2-TS approach is biased toward shorter ones (**Figure 3.3, A**). Transcript biotype analysis revealed that inDrops-2-TS dataset had more ribosomal protein transcripts and a slightly higher fraction of lncRNA transcripts, as compared to the IVT approach (**Figure 3.3, B**).

In order to evaluate if the tendencies observed have any relevance to biological interpretation, differential gene expression analysis using MAST

was carried out, and the resulting gene lists were subjected to gene set enrichment analysis, using Gene Ontology (GO) terms for broad biological processes (BP) and cellular compartment (CC). The results revealed that indeed, genes associated with different biological processes and cellular compartments were significantly enriched between the two amplification approaches. Particularly, inDrops-2-TS was enriched for cellular compartment terms related to ribosomes, whereas inDrops-2-IVT – with nucleus (**Figure 3.3, C**). These results were corroborated by biological process enrichment results, with inDrops-2-TS having terms associated with biosynthetic processes and translation, and inDrops-2-IVT – with chromatin organization and histone modification (**Figure 3.3, C**). These results indicate that the choice of amplification strategy might have an impact on the biological interpretation in certain contexts, and should be taken into account. Some differences might be attributed to the efficiency of RT reaction and template switching. For example, the reverse transcriptase enzyme drop-off events might occur while synthesizing long transcripts, and while these fragments would be readily transcribed and amplified in the IVT approach, the template switching would not occur, especially considering that efficiency of this reaction is related to the presence of a G-cap (315). Considering that nuclear transcriptome profiling has been shown to be enriched in unspliced RNAs, generating intronic reads (52,54), it is not surprising that inDrops-2-IVT dataset, recovering more such reads, associates with nuclear biological processes and compartment.

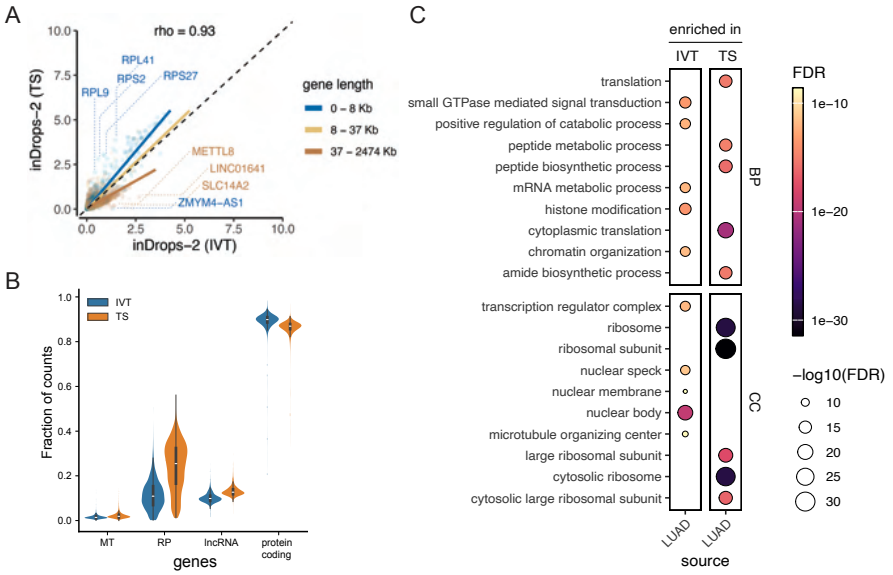


Figure 3.3. Comparison of IVT and TS inDrops-2 library preparation approaches with a focus on biological interpretation. **A** – Correlation between inDrops-2-IVT and inDrops-2-TS protocols in lung adenocarcinoma sample. Spearman's coefficient (rho) is depicted at the top of the scatter plot. Each dot represents normalized expression levels of detected genes. Dashed line (diagonal) divides the panel into two equal parts, whereas blue, yellow and brown lines display linear regression curves corresponding to short (0–8 kb), medium (8–37 kb) and long gene length (37–2474 kb) categories, respectively. **B** – fraction of total counts by gene biotype in filtered (threshold 400 UMI, 15% mitochondrial fraction) lung adenocarcinoma datasets for both IVT and TS approaches. **C** – Gene set enrichment analysis results. GSEA performed on ordered gene list by the level of differential expression (log2 fold change) between IVT and TS lung adenocarcinoma libraries. GO terms BP and CC refer to biological process and cellular compartment, respectively.

Next, to evaluate potential impact on the cellular phenotype recovery, the data generated using both amplification approaches were used for the construction of a transcriptional atlas, visualized via UMAP. Cells from inDrops-2-TS (n=2133 cells) and inDrops-2-IVT (n=2173 cells) libraries overlapped well and no sample-specific groups were observed (**Figure 3.4, A**). Nine major cell types were detected (**Figure 3.4, B**). The epithelium consisted of alveolar epithelial cells (*SFTPA1*, *SFTPB*) and a small group of club cells (*SCGB1A1*, *SCGB3A2*). The immune compartment was comprised of a large population of T cells (*CD3D*, *IL7R*), NK cells (*NKG7*, *PRF1*), B cells (*CD79A*, *MS4A1*), plasma cells (*IGKC*, *JCHAIN*), mast cells (*TPSB2*, *TPSAB1*), myeloid cells (*LYZ*, *C1QB*) and a tiny cluster of RBCs (*HBB*,

HBA1). Dataset composition analysis revealed that both inDrops-2-TS and inDrops-2-IVT libraries had virtually identical cellular composition, and no bias in cell phenotype capture was observed (**Figure 3.4, C**).

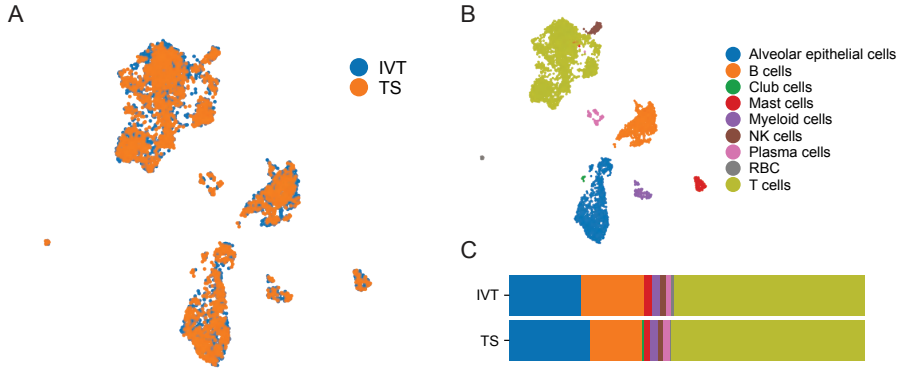


Figure 3.4. Comparison of library preparation approaches with a focus on transcriptional landscape reconstitution. **A** – UMAP representation of LUAD cells from IVT-based and TS-based libraries show complete overlap. Each dot represents a cell. n=4,302. **B** – UMAP of lung adenocarcinoma cells from both libraries, colored by major cell type. **C** – composition of TS- and IVT-based scRNA-seq datasets, colored by cell type. Both libraries have similar proportions of major cell types.

Taken together, TS-based amplification protocol was successfully implemented into inDrops-2 method for single-cell analysis of primary cells. The major advantage offered by this approach is rapid construction of sequencing libraries that does not require advanced molecular biology skillset as compared to the IVT-based counterpart. The transcript and gene capture rates were high and comparable to that of inDrops-2-IVT approach. Moreover, despite observed bias toward shorter and ribosomal protein transcripts, the method recovered the transcriptomic landscape of LUAD, proving sufficient for cellular composition profiling and atlas construction of primary cells derived from clinical samples.

3.1.2. inDrops-2-TS enables rare phenotype detection in lung carcinoma samples

Single-cell profiling of clinical samples presents logistic challenges, as the majority of scRNA-seq protocols require immediate handling and barcoding of fresh specimens. Moreover, lengthy manipulations during sample preparation induce transcriptional changes and can compromise data quality. Thus, inDrops-2 effort included the development of droplet

scRNA-seq compatible methanol-based primary cell preservation protocol, effectively safeguarding the cellular transcriptomes upon dissociation and preventing any further changes or RNA degradation. Taking it one step further, multiplexing of fixed samples, termed hashing, was introduced, taking advantage of methyltetrazine-modified DNA oligonucleotides, also known as 'ClickTags'. These labels do not rely on specific epitopes and are chemically attached to proteins via Diels–Alder reaction. That way, multiple specimens can be labeled, combined and barcoded in the same experiment, increasing scale, alleviating batch effects and aiding data cleanup via unbiased detection of true doublets. The capacity of inDrops-2 for clinical sample profiling in combination with this strategy was showcased by multiregional profiling of lung carcinoma specimens.

Briefly, primary lung carcinoma tissues (n=3) collected during surgery were cut into three regions each, dissociated and FACS-sorted into methanol to enrich and simultaneously fix CD45 positive and negative fractions. Following sorting, the methanol-fixed cell suspensions (n=18) were transferred to -80℃ for long-term storage. After 30 days, the samples were hashed with ClickTags while in methanol, pooled, rehydrated and barcoded using the inDrops-2-TS approach (**Figure 3.5, A**).

After sequencing, pre-processing and quality control, 32,937 high-quality cells were obtained, with high average UMI and gene count (6959 and 1966, respectively), similar to a previously published lung cancer single-cell dataset (309). Data analysis revealed a complex tumor microenvironment, composed of epithelial, mesothelial, endothelial and stromal cells, with abundant immune infiltration of both myeloid and lymphoid lineages (**Figure 3.5, B**). The CD45 positive and negative fractions separated well in concordance to the broad cell type annotations assigned (**Figure 3.5, C**). Samples from different patients mixed well in the immune cell clusters, suggesting shared infiltrating phenotypes, yet in the non-immune compartment patient-specific populations were observed (**Figure 3.5, D**), highlighting inter-patient variability. For a more granular characterization of the tumor microenvironment, further analysis was performed on the non-immune, myeloid and lymphoid fractions separately.

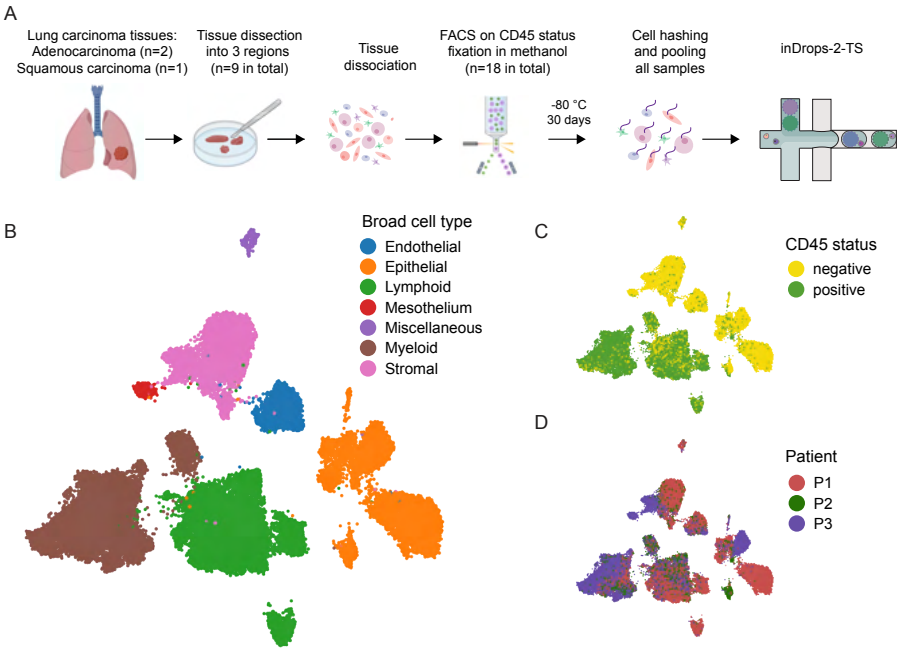


Figure 3.5. Profiling fixed, preserved and hashed primary lung carcinoma specimens. **A** – Schematics of lung carcinoma multi-regional tissue profiling experimental design. **B** – a UMAP of lung carcinoma samples, colored by broad cell type assignment shows capture of diverse phenotypes. n=32,937. **C, D** – lung carcinoma UMAPs colored by CD45 status and patient identity, respectively.

To gain insight into the non-immune cell diversity (n=12,521) and variability across samples and tumor regions, a separate atlas was built, clustered and annotated manually upon extensive literature review (**Figure 3.6, A**). Detailed analysis uncovered lung-specialized epithelial cells such as alveolar epithelial cells (AEC, markers *SFTPA1, SFTPC*), club (*SCGB1A1, SCGB3A2*), ciliated (*CAPS, PIFO*), neuroendocrine (*CALCA, UCHL1*) and basal (*KRT17, KRT15*) cells, as well as patient 3-enriched *SPINK1*-high club cell population and *MMP7*-high alveolar epithelial phenotype (**Figure 3.6, B**). *SPINK1*-high club cells featured expression of canonical club cell markers but, unexpectedly, also expressed distal lung marker *NAPSA* (also expressed by AEC) and *CEACAM6* (**Figure 3.6, B**), known to be involved in disease progression (316). Interestingly, among alveolar epithelial cells, two distinct states were observed – both subpopulations expressed canonical AEC markers (i.e. *SFTPA1*, *SFTPA2*, *SFTPC*), but one was marked by high *MMP7* and *PRSS2* expression (**Figure 3.6, B**). Expression of *PRSS2* is associated with invasive features (317), while *MMP7* is a widely used biomarker for

pulmonary fibrosis. Non-epithelial phenotypes included lymphatic (*CCL21*, *NR2F2*) and tumor endothelial cells (*CLDN5*, *ACKR1*), smooth muscle cells (*ACTA2*, *TAGLN*), mesothelium (*MSLN*, *UPK3B*) and fibroblasts (*COL1A2*, *FN1*) (**Figure 3.6 A, B**). The fibroblasts separated into two transcriptionally distinct groups, potentially involved in inflammation. For instance, complement-high fibroblasts had high expression of complement system constituents (i.e. *C7*, *CFD*) suggesting participation in complement-mediated inflammatory processes in the tumor microenvironment (**Figure 3.6, B**). The other fibroblast population was enriched for *HAS1* (**Figure 3.6, B**), suggesting similarity to an invasive *HAS1* expressing fibroblast population recently discovered in fibrotic lungs (318). Moreover, these cells expressed cytokines *CXCL1*, *CXCL2* and *IL6*, as well as a potent chemokine for monocytes *CCL2* (**Figure 3.6, B**). Thus, inDrops-2-TS approach captured interesting, previously undetected rare populations of epithelial and stromal cells, of potential interest for future investigations.

Several populations detected were not only patient-enriched (**Figure 3.6, C**), but also varied between tumor regions profiled. Importantly, inter-patient variability could not be attributed solely to tumor histology – lung adenocarcinomas (patient 1 and 3 samples) were vastly different in fibroblast and neuroendocrine cell fraction (**Figure 3.6, C, D**), while the squamous carcinoma (patient 2 sample) had similar fibroblast involvement as adenocarcinoma from patient 1. With regards to regional differences, patient 2 regions 1 and 3 lacked club cells, while these cells constituted a relatively large fraction of region 2, indicating presence of bronchioles in that region (**Figure 3.6, D**). Additionally, the aforementioned inflammatory fibroblast populations were largely constricted to patient 3, with HAS1-high group detected in two out of three regions (**Figure 3.6, D**). These findings highlight the complexity of tumor architecture and the need to recognize study limitations when aiming to build generalized cancer single-cell atlases of a single tissue biopsy.

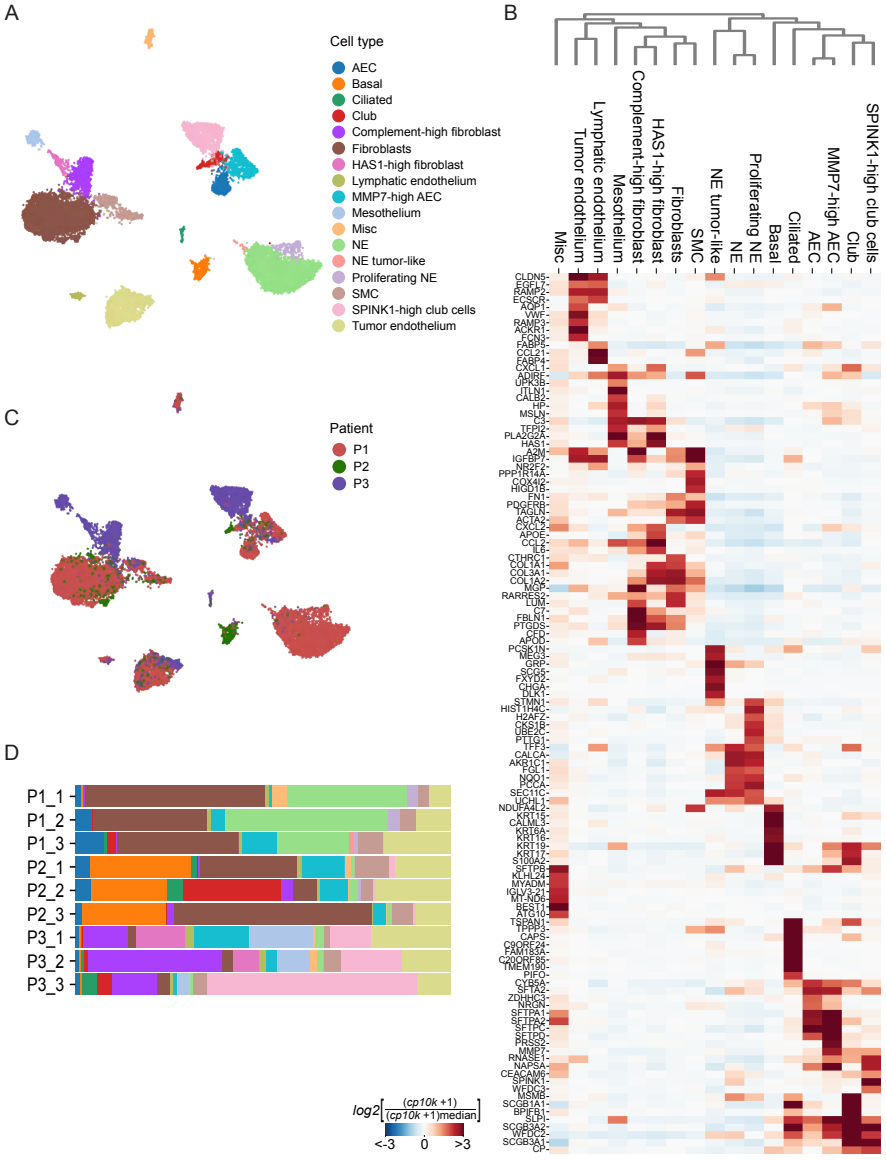


Figure 3.6. Assessing the heterogeneity of non-immune cell populations in lung carcinoma. **A** – a UMAP of non-immune cells of the lung carcinoma TME. Diverse lung-specific and stromal phenotypes are observed. n=12,521. **B** – a heatmap and hierarchical clustering dendrogram of top differentially expressed genes (ranked by fold-change) between the non-immune cell phenotypes, Mann-Whitney U test, Benjamini-Hochberg adjusted p-value <0.05. Markers *PIFO, UCHL1, CEACAM6, NR2F2, ACTA2, UPK3B, FN1, CXCL1* appear in the top 50 significant DEG lists, but were added manually to this plot. **C** – UMAP colored by patient ID highlights multiple patient-specific populations. **D** – sample composition, colors correspond to the legend

in **A**. Labels correspond to patient ID followed by an underscore and tumor region profiled from the same patient, i.e. P1_1, P1_2 etc.

Abundant immune infiltration was observed, with more than half of all profiled cells (>60%) being immune (**Figure 3.5, C**). Even though this result can be partially attributed to dissociation bias, it appears to be consistent across published lung carcinoma atlases (7,319,320). To assess immune cell phenotypic heterogeneity in detail, myeloid and lymphoid cells were separately extracted from the dataset for construction of a low-dimensional representation, clustered and annotated manually.

Myeloid cell analysis (n=9,921) revealed the presence of mast cells (*TPSPB2*, *TPSAB1*), monocytes (*S100A9*, *FCN1*), monocyte-like dendritic cells (*CCL17*, *CLEC10A*, as described in (8)), conventional type 1 dendritic cells (*CLEC9A*, *CST3*), activated dendritic cells (*CCR7*, *CCL22*), alveolar macrophages (*MARCO*, *FABP4*) and tumor-associated macrophages (TAM) (**Figure 3.7, A, B**). The TAM group displayed diverse M1- and M2-polarization associated profiles. M1-like macrophages expressed canonical inflammatory markers (i.e. *IL1B*, *CCL3*, *TNF*) (**Figure 3.7, B**). M2-like TAM 1 subpopulation expressed *APOE*, *APOC1*, *TREM2*, consistent with an immunosuppressive phenotype, while M2-like TAM 2 population had elevated levels of canonical M2 polarization markers *CD163* and *MRC1* (**Figure 3.7, B**). Interestingly, certain myeloid populations were highly patient-specific (**Figure 3.7, C**). For instance, all regions from patient 2 sample lacked conventional dendritic cells, potentially related to its distinct histology (squamous carcinoma) (**Figure 3.7, D**). Moreover, the M2-like TAM 2 population was present in all tumor regions from patients 1 and 2, but not patient 3. Meanwhile, M1-like TAMs were detected only in patient 3, and varied in abundance between regions profiled, likely reflecting a balance between inflammatory and immunosuppressive components in the TME (**Figure 3.7, D**). Interestingly, as mentioned previously, samples from this patient were enriched in complement-high fibroblasts (**Figure 3.6, D**), suggesting potential maintenance of inflammation, although the causal relationship between the presence of these cells and M1-like TAMs remains unresolved. Together, these results indicate myeloid cell diversity in lung carcinoma and patient-specific infiltration patterns, likely arising from differential immunomodulatory signals of the diverse non-immune cell types.

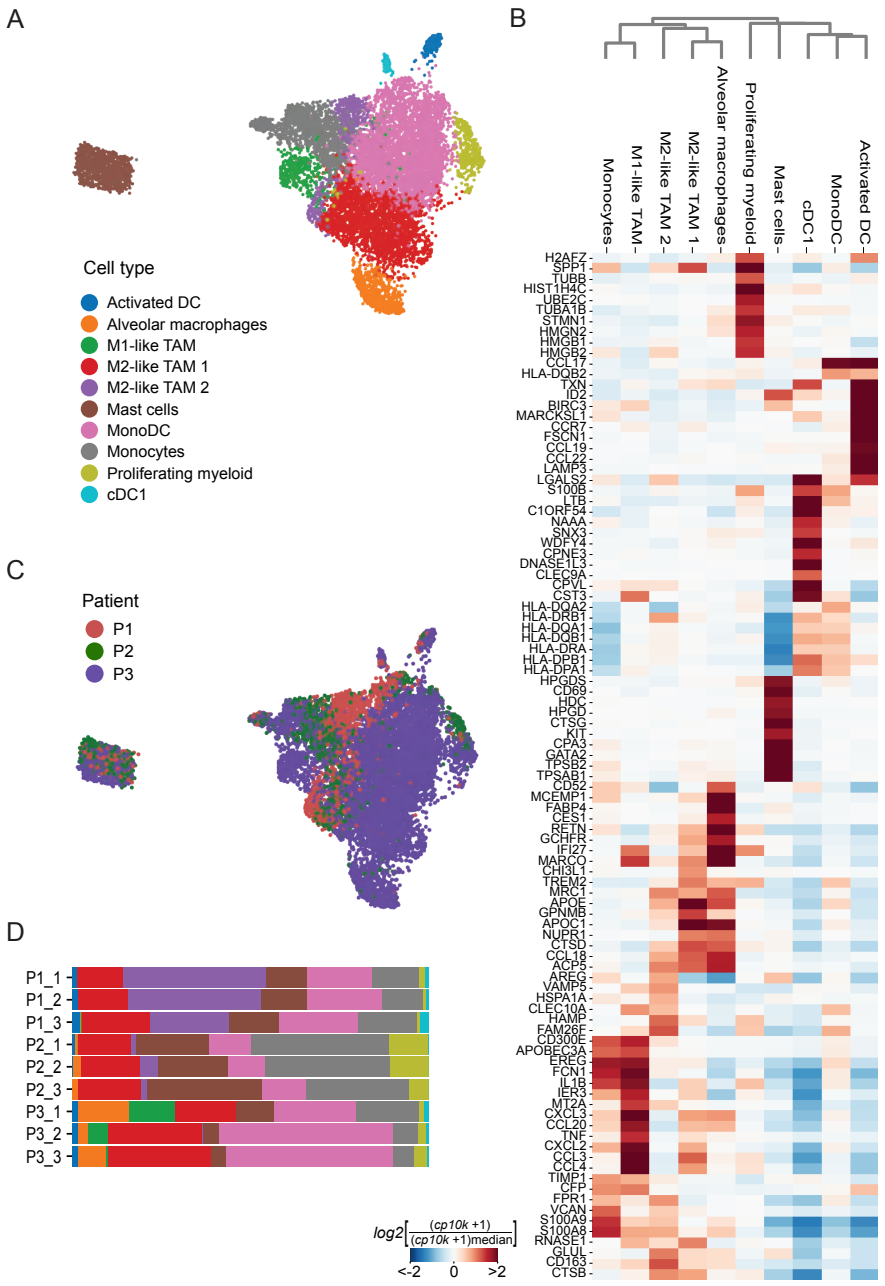


Figure 3.7. Analysis of myeloid cell populations in lung carcinoma samples. **A** – a UMAP of myeloid cells of the lung carcinoma TME. Several subpopulations of TAMs are detected. n=9,921. **B** – a heatmap and hierarchical clustering dendrogram of top differentially expressed genes (ranked by fold-change) between the myeloid cell phenotypes, Mann-Whitney U test, Benjamini-Hochberg adjusted p-value <0.05. Markers *CLEC10A, CLEC9A, TREM2, MRC1* appear in the top 50 significant DEG lists, but were

added manually to this plot. **C** – UMAP colored by patient ID. **D** – sample composition, colors correspond to the legend in **A**. Labels correspond to patient ID followed by an underscore and tumor region profiled from the same patient, i.e. P1_1, P1_2 etc.

In-depth assessment of lymphoid cells (n=10,495) revealed the presence of innate lymphoid cells (*CD3D*), plasmacytoid dendritic cells (*LILRA4*, *CLIC3*), B cells (*CD79A*, *MS4A1*), plasma cells (*IGHG4*, *JCHAIN*), NK cells (*NKG7*, *GZMB*) and a large fraction of various CD4 and CD8 T cell phenotypes, including a proliferating population (**Figure 3.8, A**). Particularly, in CD4 T cells, regulatory (*FOXP3*, *CTLA4*), naïve (*IL7R*, *CCR7*) and *CXCL13*-high phenotypes were observed (**Figure 3.8, B**). Meanwhile, CD8 T cell group was composed of effector memory (*CD52*, *CREM*) and cytotoxic (*GZMA*, *CCL4*) subtypes (**Figure 3.8, B**). Inter-patient (**Figure 3.8, C**) and inter-regional (**Figure 3.8, D**) differences in the lymphoid cell compartment appeared to be less pronounced as compared to non-immune or myeloid cells. Nonetheless, effector memory T cells were enriched in patient 3 tumor, whereas the *CXCL13*-high T cells were mostly present in patient 1 (**Figure 3.8, D**). This population appeared to coincide with higher fraction of B cells in patient 1 samples (**Figure 3.8, D**), supporting the selective chemoattractant role of CXCL13 in B cell recruitment.

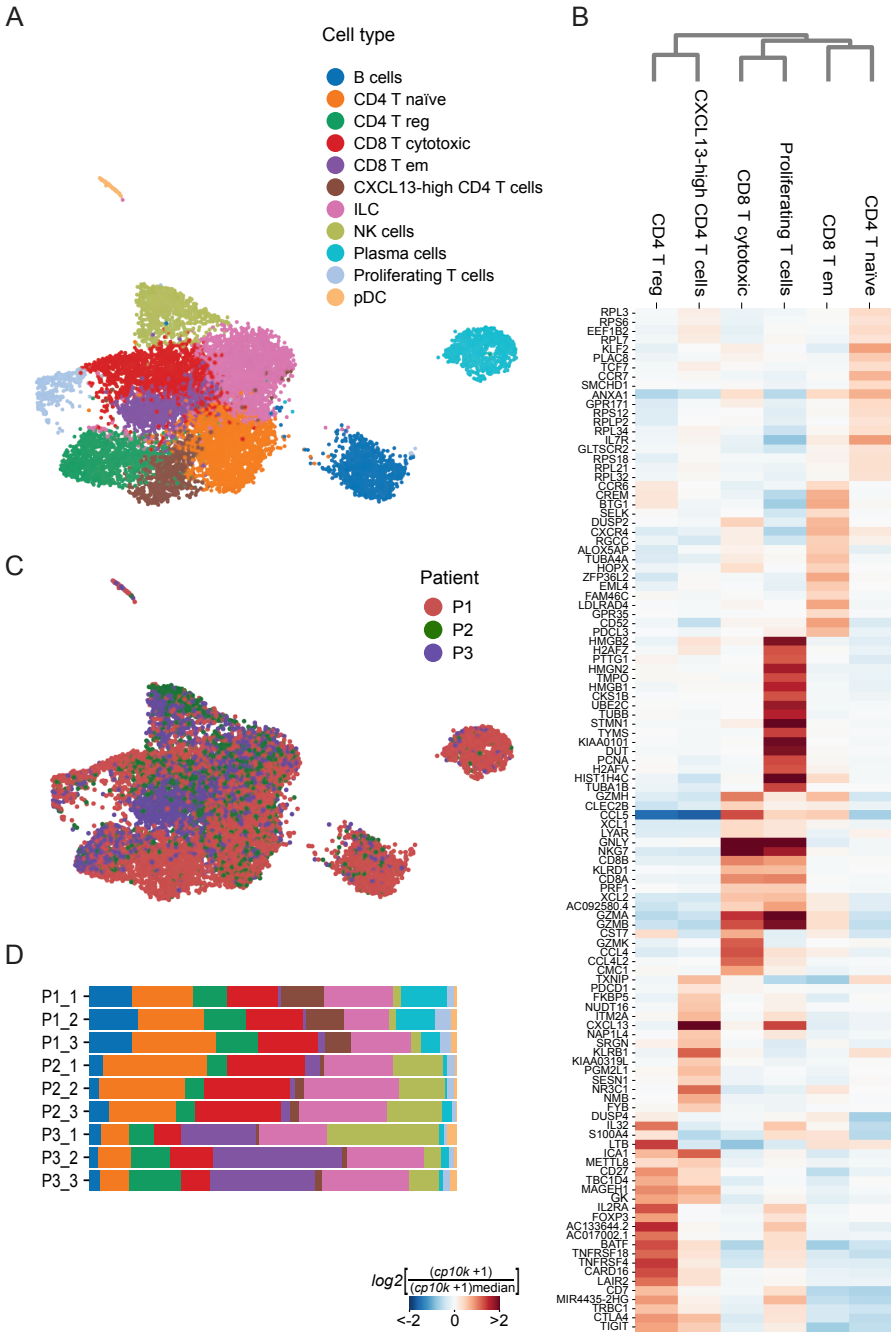


Figure 3.8. Analysis of lymphoid cell populations in lung carcinoma samples. **A** – a UMAP of lymphoid cells of the lung carcinoma TME. n=10,495. **B** – a heatmap and hierarchical clustering dendrogram of top differentially expressed genes (ranked by fold-change) between the T cell populations, Mann-Whitney U test, Benjamini-Hochberg adjusted p-value

<0.05. **C** – UMAP colored by patient ID. **D** – sample composition, colors correspond to the legend in **A**. Labels correspond to patient ID followed by an underscore and tumor region profiled from the same patient, i.e. P1_1, P1_2 etc.

Overall, these results illustrate the suitability of inDrops-2-TS approach for recovery of transcriptomic profiles of tens of thousands of cells in clinical tissue samples that have been preserved, stored long-term and multiplexed via hashing. Furthermore, inDrops-2-TS enabled the recovery of several rare, potentially clinically relevant phenotypes, such as the inflammatory fibroblast populations and CD4 *CXCL13*-high T cells, that were not described previously in the context of lung carcinoma.

3.2. Single-cell profiling of healthy kidneys and clear cell renal cell carcinoma

The most common kidney cancer, clear cell renal cell carcinoma (ccRCC) (>80% of cases) is often diagnosed late, frequently metastasizes and exhibits high genetic and phenotypic heterogeneity that influences treatment efficacy. Hallmark alterations include loss of 3p chromosome regions and *VHL* gene mutations, leading to pseudo-hypoxic conditions and highly vascularized tumor appearance (212). Single-cell profiling efforts, in the context of ccRCC, have provided valuable insight into the malignancy-related transcriptional programs (243), cell of origin of ccRCC (196), immune infiltrate (230,234) as well as its phenotypical changes along advancing disease stage (232) and in response to treatment (233). These studies mostly focused on tumor and immune cells, however, stromal and endothelial cells, which are particularly relevant in ccRCC, have often been overlooked and under-characterized. The work presented in this chapter aims to fill this gap in ccRCC TME characterization by single-cell profiling of ccRCC with an emphasis on vascular and stromal cells and their involvement in the TME. Additionally, a detailed atlas of adjacent healthy kidney tissue underscores the suitability of inDrops-2 for capture of rare and sensitive phenotypes and further highlights the vast structural and phenotypic disturbances imposed by ccRCC disease.

3.2.1. Global atlas of ccRCC and adjacent kidney reveals inter-patient variability and progenitor-like epithelial phenotype

To investigate the phenotypes within the ccRCC TME and healthy kidney tissue, we profiled fresh tumor (n=8) and paired adjacent healthy tissue (n=9) specimens using the droplet-based scRNA-seq inDrops-2-TS platform. With

the goal to capture diverse cell types and shorten cell handling time, the experimental strategy employed did not include cell enrichment of any kind (**Figure 3.9, A**). Following quality control, batch correction and doublet removal, a total of 50,236 single-cell transcriptomes were obtained and clustered using graph-based spectral clustering algorithm. Cell type annotation involved an extensive literature review on differentially expressed marker genes (adjusted p value < 0.05; cluster vs the rest of cells, Mann-Whitney U test with Benjamini-Hochberg correction) (**Figure 3.9, B**).

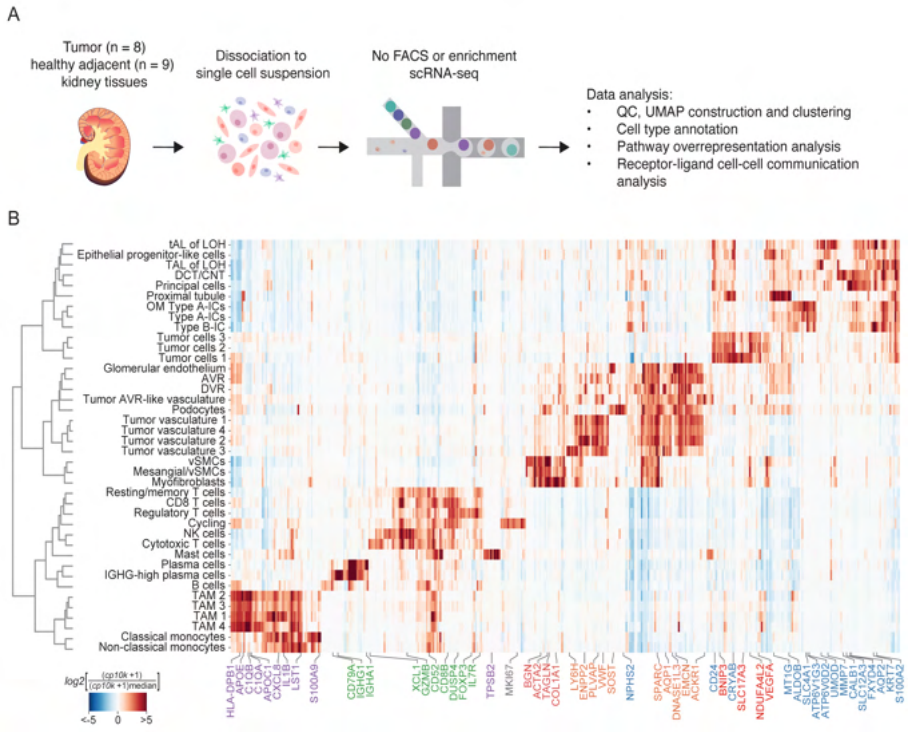


Figure 3.9. Building a single-cell atlas of kidney and ccRCC specimens. **A** – experimental design of kidney and ccRCC sample profiling. **B** – a global heatmap depicting population-specific markers and their expression levels, hierarchical clustering provided as a dendrogram. Only significant genes are shown (Benjamini-Hochberg adjusted p value < 0.05). Color of the gene name indicates major cell type (blue – epithelial, red – tumor, orange – endothelium, pink – stromal, green – lymphoid, purple – myeloid, grey – cycling). AVR – ascending vasa recta, DVR – descending vasa recta, vSMCs – vascular smooth muscle cells, LOH – loop of Henle, tAL – thin ascending limb, TAL – thick ascending limb, DCT/CNT – distal convoluted/connecting tubule, ICs – intercalated cells, OM – outer medullary, TAM – tumor associated macrophages.

The global atlas of all profiled cells revealed high diversity of phenotypes both in tumor specimens and healthy kidney (**Figure 3.10, A**). The ccRCC samples, spanning localized and locally advanced pT1a and pT3a pathological stages, exhibited high immune cell infiltration, consisting of several phenotypes of tumor-associated macrophages and T cells (**Figure 3.10 A, B**). The stromal cell compartment was composed of myofibroblast (type I, IV and VI collagens, *FN1, TIMP2, ACTA2*), vascular smooth muscle cell (vSMC; *TAGLN, ACTA2, SNCG*) and mesangial/vSMC (*BGN, PDGFRB, TAGLN*) populations. Strikingly, tumor endothelium completely separated from healthy-adjacent tissue endothelial populations and exhibited distinct expression pattern (**Figure 3.9, B**). This compartment was composed of ascending *vasa recta*-like endothelial cells (*ACKR1, DNASE1L3*) as well as heterogeneous subpopulations expressing tumor-associated endothelium markers *PLVAP, VWF, SPARC, INSR, ANGPT2* among others (**Figure 3.9, B**). While four out of five tumor endothelium subpopulations identified in our dataset were described previously (196,222,243), one small (n=151 cells) tumor vasculature subpopulation appeared to have not been described in the context of ccRCC (described in detail in section 3.2.4).

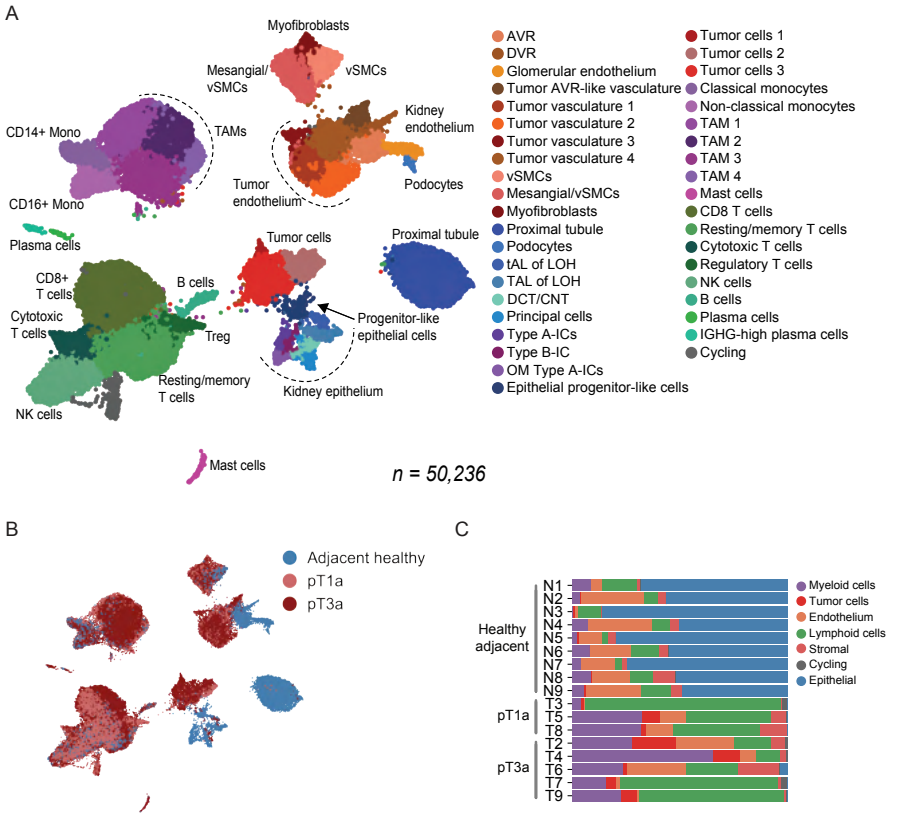


Figure 3.10. Transcriptional atlas of kidney and ccRCC. **A** – a global single-cell transcriptional map (UMAP) of kidney and ccRCC specimens. **B** – a UMAP of cells annotated by disease stage (adjacent healthy, pT1a and pT3a). **C** - sample composition by major cell type, depicting high immune infiltration and loss of epithelial cells in tumor specimens.

The tumor cells were marked by the expression of canonical markers such as *CA9*, *NDUFA4L2*, *VEGFA* and separated into three subpopulations, out of which one (Tumor cells 1) was patient-specific (**Supplementary Figure S1, A**). Notably, this population had elevated expression of progenitor-like markers *VCAM1* and *SLC17A3*, and appeared distinct from other tumor cells based on unsupervised hierarchical clustering (**Figure 3.9, B, Supplementary Figure S1, B**). These cells also expressed pan-cancer marker *MDK33*, genes *IFI27* and *SOD2* involved in interferon response (233) as well as *HLA-G*, involved in immunosuppressive interactions and *FABP7*, important for lipid uptake and storage in hypoxic conditions when *de novo* lipid synthesis is repressed (321). Consistently, gene set over-representation analysis revealed that this tumor cell population was not enriched for hypoxia, but instead enriched for oxidative phosphorylation and adipogenesis, in contrast to the

other tumor cells in our dataset (**Figure 3.18, A**). These results suggest that the patient-specific Tumor cells 1 population could represent an intermediate progenitor-tumor cell phenotype.

As expected, the cellular composition of tumor tissues displayed variability across the patients as compared to the matched healthy-adjacent control samples. Tumor samples exhibited almost complete loss of specialized kidney-specific epithelial and endothelial cells and were highly infiltrated by immune cells (**Figure 3.10, C**). Except for Tumor cells 1 population, no other phenotype was patient-specific and cell population composition analysis by patient ID confirmed adequate representation of all samples profiled (**Supplementary Figure S1, A**). In order to quantitatively assess sample heterogeneity, Shannon entropy values for each broad (i.e., epithelial, myeloid, stromal) cell category were calculated (309). Low entropy values for a cell group indicate that the phenotype is rarely shared between the samples – in other words, it indicates that the level of diversity within samples is high. The heterogeneity was highest for tumor sample stromal, endothelial and tumor cell groups, whereas the healthy adjacent tissues had relatively lower heterogeneity (**Supplementary Figure S1, C, D**). The lack of patient-specific populations (except for Tumor cells 1) and the diverse TME composition among patients in our and other ccRCC single-cell datasets (222,322) suggest that patient stratification may rely on the abundance of certain phenotypes, rather than unique, patient-specific ones. These results highlight the importance of revisiting strategies for biomarker selection in the context of personalized ccRCC treatment.

Healthy adjacent samples contained all major epithelial and endothelial cell populations characteristic of a healthy kidney (**Figure 3.10, A**) (51,204,323). Our experimental strategy excluding the cell enrichment procedures enabled the capture of cell types that are known to be highly sensitive to extended workflows (324). For instance, we were able to capture both ascending (*DNASE1L3*) and descending (*AQP1*, *SLC14A1*) parts of the vasa recta, as well as glomerular endothelium (*IGFBP5*, *SOST*) (**Figure 3.9, B**). The dominant broad cell type was epithelium (**Figure 3.10, C**), encompassing cells representing various specialized nephron segments (**Figure 3.10, A**). Additionally, we were able to recover rare and sensitive cell phenotypes, such as intercalating cells of the collecting duct of type A and B (expressing markers *ATP6V1G3* and *SLC26A4*, respectively), as well as podocytes (*NPHS2*, *PODXL*). Interestingly, in contrast to tumor samples, all healthy tissues comprised a small population (n=321) of epithelial progenitor-like cells (**Supplementary Figure S1, A**). These cells were similar to a rare proximal tubule cell, proposed ccRCC "cell of origin" (*VCAM1*+,

SLC17A3+, *SLC7A13*-) described by Young et al. (196), as they were positive for *VCAM1* and negative for *SLC7A13*, but in our case only a few cells expressed *SLC17A3* (**Figure 3.11, A**). However, this population also expressed genes associated with de-differentiated injured kidney epithelium, such as *PROM1* and *ITGB8* (192), as well as *CD24* and *SOX4*, which are implicated in kidney development and mark proximal tubule and distal nephron response to acute kidney injury (325) (**Figure 3.11, A**). To investigate the epithelial cell identities further and examine the similarity to the established ccRCC "cell of origin" proximal tubule (PT-B) phenotype, we trained a CellTypist model on publicly available ccRCC scRNA-seq dataset by Zhang et al. (222). CellTypist label transfer revealed complete agreement of cell type annotations for specialized nephron tubule populations and confirmed proximal tubule identity for tumor cells. Importantly, it supported the notation that epithelial progenitor-like phenotype is not a misannotated healthy nephron epithelium population, but indeed most similar to the PT-B phenotype (**Figure 3.11, B**). Therefore, the epithelial progenitor-like population in healthy-adjacent tissues most likely represents a de-differentiated phenotype sharing transcriptomic features with the cell of origin of ccRCC.

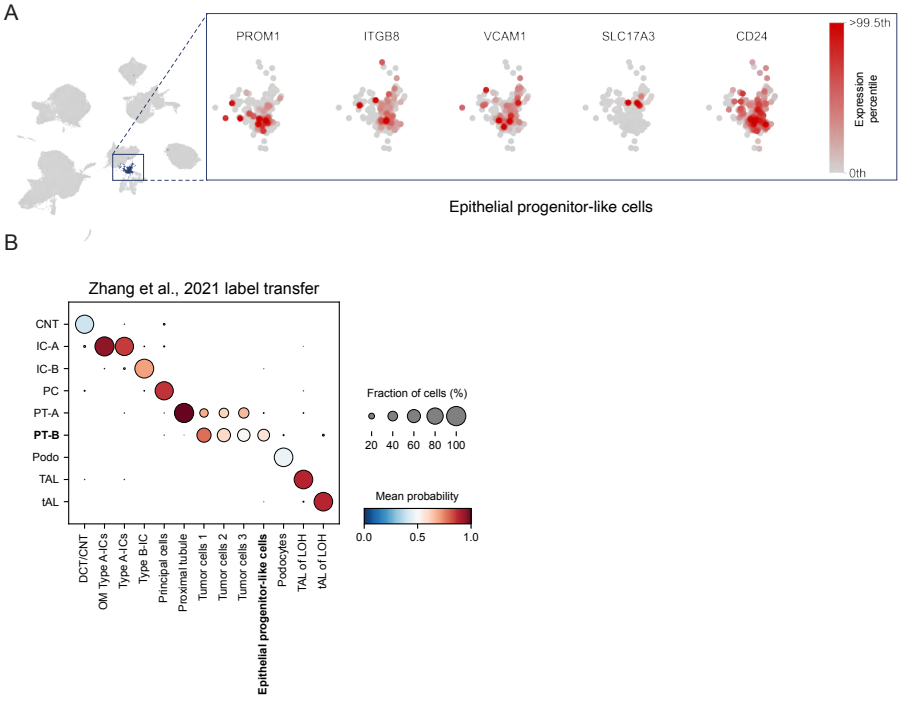


Figure 3.11. Analysis of epithelial progenitor-like cell population primarily found in healthy adjacent kidney tissues. **A** – expression of ccRCC cell of origin markers in epithelial progenitor-like cell population, normalized

to 99.5th percentile of a given gene's expression level. **B** – CellTypist model label transfer from Zhang et al. (222) dataset. X-axis labels are cell populations from this study and y-axis labels are from Zhang et al. The annotation of cell types in both studies is in agreement and epithelial progenitor-like cell population is the most similar to the cell-of-origin 'PT-B' phenotype from Zhang et al.

3.2.2. The ccRCC TME is highly infiltrated by TAMs exhibiting immunosuppressive interactions

Among solid tumors, ccRCC is known to have a highly immune infiltrated, dynamic TME. Recently, the immune compartment compositional changes along tumor stage progression (232) and response to immunotherapy treatment (233) were described and shown to have an impact on patient survival. Considering the ICB immunotherapy use for advanced and metastatic disease treatment (215), insight into the phenotypic states of immune cells is particularly relevant.

In concordance with previous ccRCC atlases (230–232), we identified all major lymphoid and myeloid cell populations within the immune compartment: plasma cells (*IGKC*, *IGHG1*), B cells (*CD79A*, *MS4A1*), natural killer (NK) cells (*GZMB*, *NKG7*), classical (*CD14*) and non-classical (CD16, gene *FCGR3A*) monocytes, mast cells (*TPSB2*) and two large groups of T cells and macrophages (**Figure 3.10, A**).

Tumor-associated macrophages enriched in tumor tissues clustered into four transcriptionally distinct subgroups (TAM 1-4) (**Figure 3.12, A**). Investigation of known macrophage polarization state marker genes revealed that TAM 1 and TAM 2 populations correspond to M1 and M2 polarization states, respectively. However, TAM 3 and TAM 4 cells did not appear to follow the traditional polarization dichotomy, although some of the alternative activation markers were elevated (**Figure 3.12, B**). For instance, while the immunosuppressive and tumor-associated marker *MARCO* was diminished in TAM 3/4 subpopulations, other immunosuppressive genes, such as *VSIG4* and *VSIR* were highly expressed in TAM 4 cells. Additionally, TAM 4 exhibited highest expression of complement component C1q genes (**Figure 3.12, B**), whose products are known to promote tumor progression in ccRCC by engaging with complement system molecules (i.e. C1s, C1r, C3) produced by the tumor cells (326). Interestingly, in our dataset the source of some of the interacting complement system components was not only the tumor cells, but also the stromal compartment, indicating potential stromal cell involvement (**Supplementary Figure S2**). These findings support the notion that ccRCC

TME is enriched for suppressive macrophages that respond to local cues, potentially impacting disease progression.

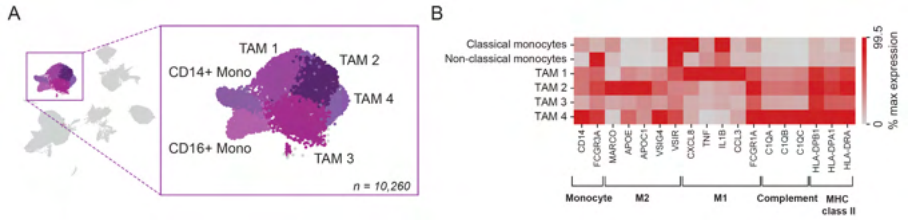


Figure 3.12. Assessing the heterogeneity of tumor-associated macrophage populations. **A** – a close-up of the myeloid cell compartment in the global UMAP. **B** – expression of polarization-associated markers in myeloid cell populations. Color intensity denotes cptt-normalized expression saturating at 99.5th percentile of a given gene's expression level.

The lymphoid compartment consisted of resting/memory T cells (*IL7R*, *CD52*), cytotoxic T cells (*XCL1*, *KLRB1*) CD8 T cells (*CD8B*, *DUSP4*), regulatory T cells (*FOXP3*, *TNFRSF4*), and natural killer cells (*GZMB*, *NKG7*) (**Figure 3.13, A**). These populations were heterogeneous in the expression of exhaustion markers (**Figure 3.13, B**) with classic immune checkpoint molecule *PDCD1* abundantly expressed in CD8 T cell cluster and *CTLA4* enriched in regulatory T cells. The cytotoxic T cell population and NK cells shared the exhaustion pattern of elevated *CD38*, *CD160*, *EOMES*, and *CD69* expression. Resting/memory T cells were the least exhausted compared to other lymphoid cell populations, as expected (**Figure 3.13, B**). Considering that the exhaustion profile of T cells and immunosuppressive features of TAMs are well established in ccRCC (231,232,327), we set to investigate their crosstalk with tumor cells.

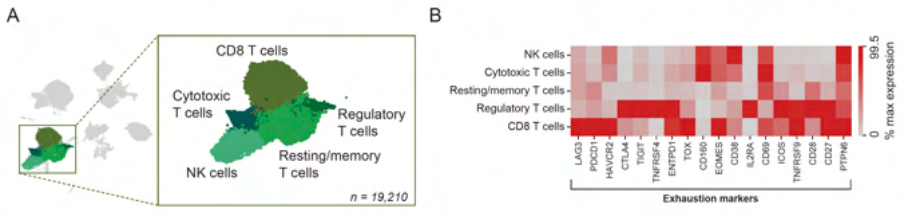


Figure 3.13. Analysis of ccRCC infiltrating lymphoid cells. **A** – a close-up of the lymphoid cell compartment in the global UMAP. **B** – expression of T cell exhaustion-associated markers in lymphoid cell populations. Color intensity denotes cptt-normalized expression saturating at 99.5th percentile of a given gene's expression level.

Cell-cell communication analysis with CellPhoneDB revealed interactions between immune and tumor cells related to immune suppression,

chemokine processing and sustained survival of tumor cells (**Figure 3.14, A**). For instance, tumor cells and monocytes/TAMs were predicted to exploit the immune checkpoint *HLA-G - LILRB1/2* axis. This interaction is known to promote the immunosuppressive M2 polarization and tumor immune escape (328). Additionally, tumor cells utilized *SPP1 - PTGER4* interaction to target both the pro-inflammatory (M1) and anti-inflammatory (M2) TAMs. In hepatocellular carcinoma, this interaction was shown to promote macrophage polarization towards tumor-supporting phenotype (329). Another noteworthy predicted interaction was the T cell co-stimulatory *CD27 - CD70* axis, involving CD8 T cells and CD4 regulatory T cells. Recently, this cell-cell interaction was linked to a pro-tumoral effect, driven by enhanced survival of regulatory T cells, T cell exhaustion induced by chronic stimulation, and recruitment of TAMs (330). Next, we evaluated the association of predicted interactions to clinical outcomes, using the publicly available The Cancer Genome Atlas (TCGA) ccRCC (KIRC cohort) bulk RNA-seq dataset. Interestingly, the expression of tumor-immune cell interaction signature (gene set of both receptors and ligands) was associated with significantly worse overall patient survival (**Figure 3.14, B**) and increased steadily along the progression of the disease (**Figure 3.14, C**). Together, our analysis of the ccRCC TME suggests an extensive immune-cancer cell interaction network related to immune-suppressive TME establishment, beneficial for tumor growth and survival.

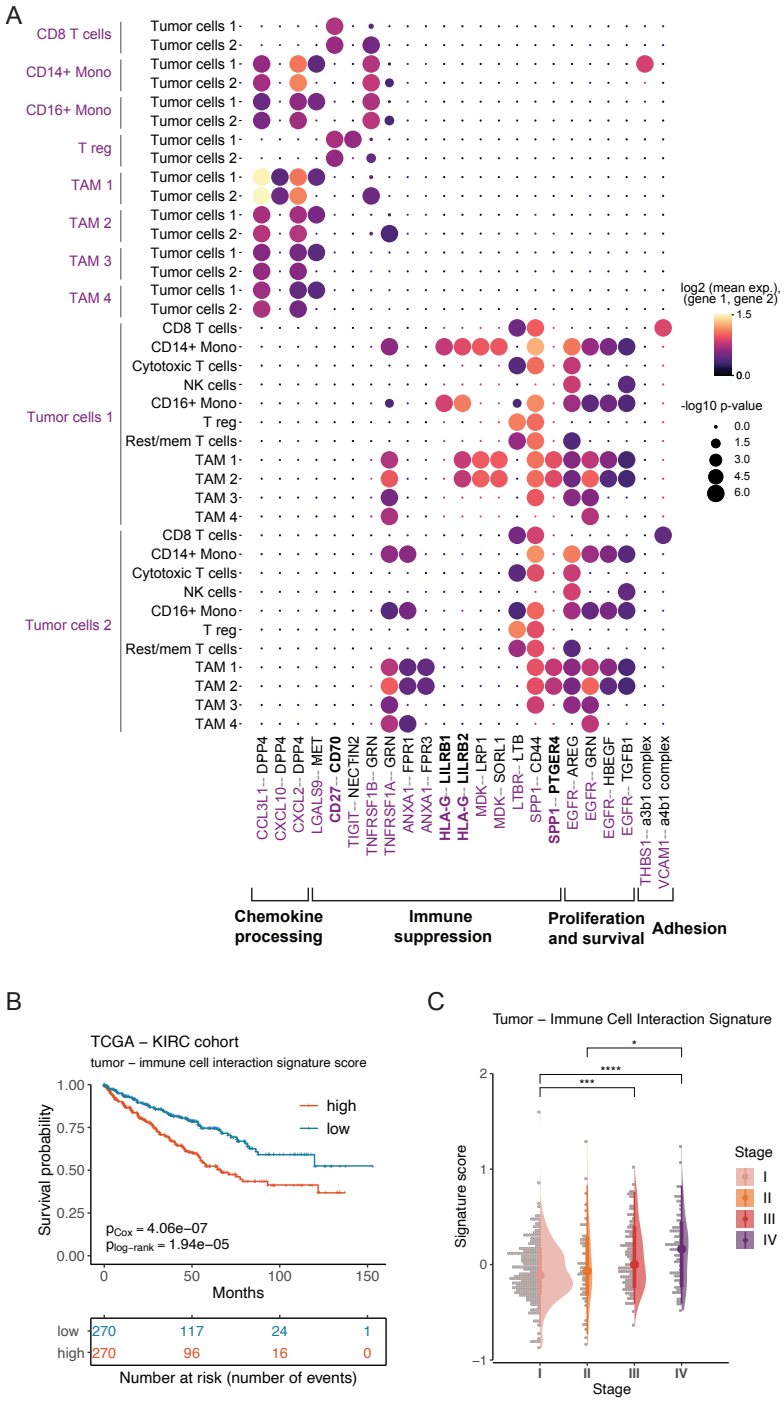


Figure 3.14. Investigation on tumor-immune interactions in the TME and their potential clinical significance. **A** – selected immunosuppressive interactions revealed by cell-cell communication analysis between immune and tumor cells using CellPhoneDB. **B** – Tumor - immune cell interaction

signature expression in TCGA KIRC cohort was associated with worse overall survival. C – Tumor and immune cell interaction signature increased along the progression of the ccRCC disease stages in the TCGA KIRC cohort. *p value < 0.05, ***p value < 0.001, ****p value < 0.0001 (Wilcoxon rank-sum test, Benjamini-Hochberg correction).

3.2.3. Tumor endothelial cells are diverse and distinct from healthy kidney endothelium

Most frequent genomic alterations in ccRCC result in pseudo-hypoxic conditions and production of angiogenic factors (211,216) that lead to highly vascularized tumor appearance, hence angiogenesis-related pathways remain the most common therapeutic target. To this day, the heterogeneity and possible regulatory role of tumor vasculature in ccRCC remains poorly described. Therefore, having identified an adequate population of tumor endothelium cells in our dataset, we set to investigate their heterogeneity and potential participation in TME intercellular communication.

Five tumor vasculature subpopulations were identified in our dataset (**Figure 3.15, A**) including a previously described ascending *vasa recta*-like population (222) and a novel, uncharacterized Tumor vasculature 3 (TV 3) population. Overall, tumor vasculature was markedly distinct from healthy kidney endothelium (**Figure 3.15, B**). Differential gene expression analysis (tumor vs healthy endothelium) revealed high tumor vasculature expression of *PLVAP*, *VWF*, *ANGPT2*, *SPARC*, *HSPG2*, *IGFBP7*, *INSR*, type IV collagen and others (**Figure 3.15, B**). Many of these genes have important implications in tumor vascularization and disease progression. For example, the fenestration maker *PLVAP* was recognized as a therapeutic target in hepatocellular carcinoma, where *PLVAP* blockade resulted in suppressed tumor growth (331); *ANGPT2* stimulates angiogenesis in autocrine manner and is involved in recruitment of immunosuppressive TAMs (332); *IGFBP7* is clinically used acute kidney injury urinary biomarker (333). Moreover, insulin receptor *INSR* that stimulates endothelial cell migration was overexpressed in tumor endothelium. This receptor was shown to be associated with poor overall survival in bladder cancer, which, similarly to ccRCC, is often resistant to VEGF pathway targeted therapy (334). These results highlight the fenestrated, abnormal nature of tumor vasculature cells and may aid in tumor-specific ccRCC vasculature identification in future studies.

Hierarchical clustering of healthy and tumor endothelium revealed, as expected, that AVR-like tumor vasculature was transcriptionally closer to the healthy endothelium populations than to tumor vasculature, while TV 3

population appeared the most distinct from the rest of tumor vasculature cells (**Figure 3.15, C**). This population was marked by expression of tip cell markers *LOX*, *PXDN*, *LY6H* and *PGF* (305,335) (**Figure 3.15, C**), suggesting an invasive phenotype. Moreover, TV 3, along with TV 1 and TV 4 displayed expression of extracellular matrix constituents, including pro-angiogenic collagen type IV and perlecan (*HSPG2*). Another population, TV 2, was marked by expression of genes involved in tumor progression. For instance, it expressed VEGF receptor *FLT1*, *ANGPT2*, *KCNE3*, *ESM1*, coagulation factor VIII (*F8*), involved in tumor-associated angiogenesis (201,336). Additionally, TV 2 expressed autotaxin (*ENPP2*), a potent stimulator of tumor development and invasion, which was shown to be involved in acquired resistance to antiangiogenic drug sunitinib in ccRCC (337) (**Figure 3.15, C**). Taken together, the ccRCC tumor endothelial cells appear to be heterogeneous and express tumor-promoting and angiogenesis-related factors.

Next, we sought to investigate the communication of the tumor vasculature with immune cells in the TME. Cell-cell communication prediction using CellPhoneDB revealed interactions involved in angiogenesis, immune suppression and adhesion (**Figure 3.16, A**). Interestingly, tumor vasculature delivered immunosuppressive signals previously thought to be confined to the tumor cells, such as the interactions between *TIGIT* and *NECTIN2* (**Supplementary Figure S3**), *HLA-F* and *LILRB1/2* (**Figure 3.16, A**). Also, several interactions with tumor endothelium mediated by myeloid cell produced *TNF* were observed, such as *TNF-NOTCH* (**Supplementary Figure S3**). This interaction is involved in *JAG1* expression induction, enhancing migration and proliferation of endothelial cells upon subsequent VEGF exposure (338). Moreover, tumor vasculature and immune cell communication signature gene set expression associated with significantly lower overall survival in TCGA KIRC cohort (**Figure 3.16, B**).

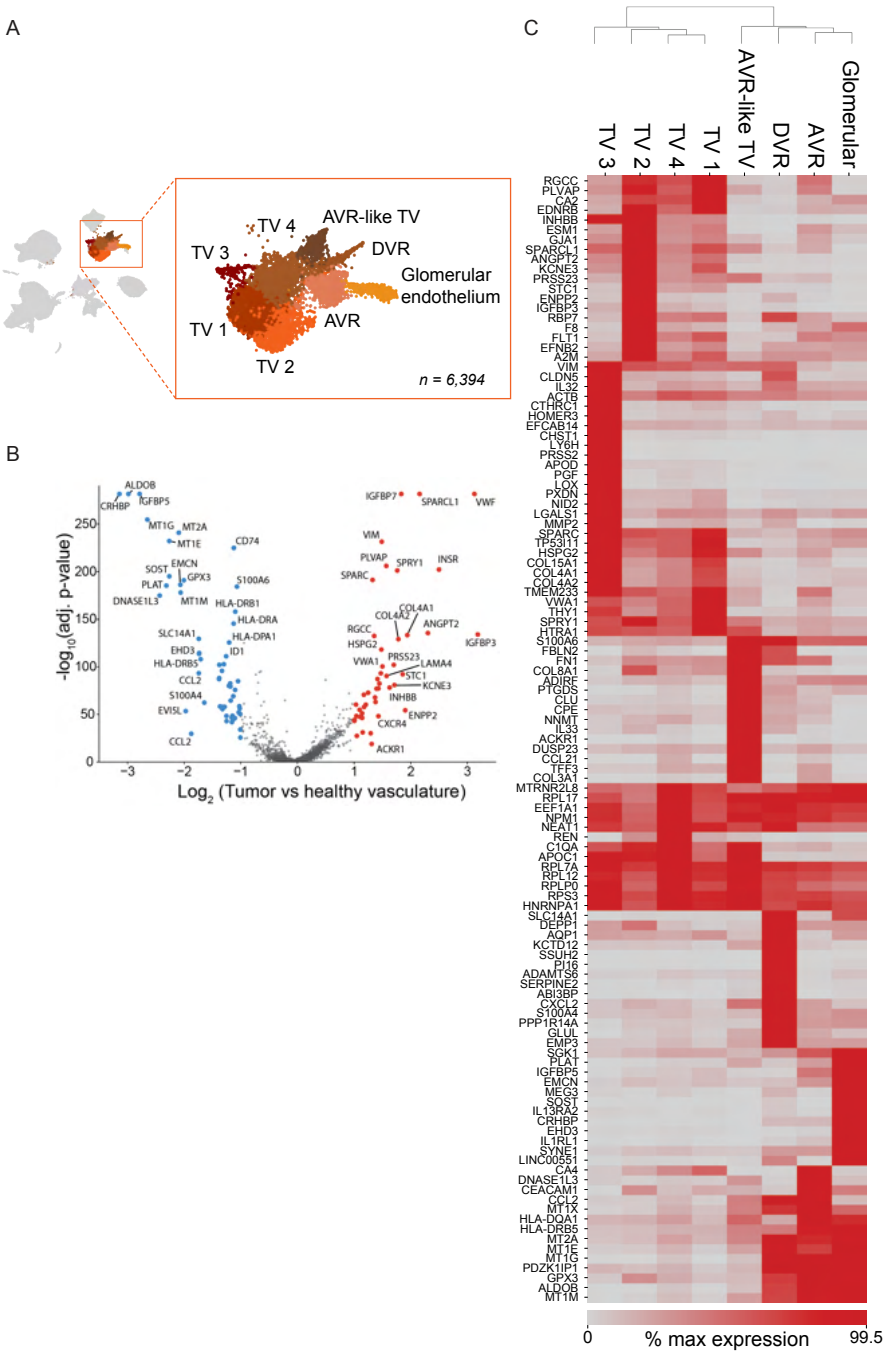


Figure 3.15. Tumor vasculature cell characterization. **A** – a close-up of the endothelial cell populations in the global UMAP. **B** – a volcano plot of tumor vs healthy endothelial cell differential gene expression results. Genes with fold-change of 2 and adjusted p-value <0.05 are highlighted and considered

significant. **C** - Differential gene expression between vasculature subpopulations. Only genes with Benjamini-Hochberg adjusted p value < 0.05 are shown. Color intensity denotes cptt-normalized expression saturating at 99.5th percentile of a given gene's expression level. AVR – ascending vasa recta, DVR – descending vasa recta, TV – tumor vasculature.

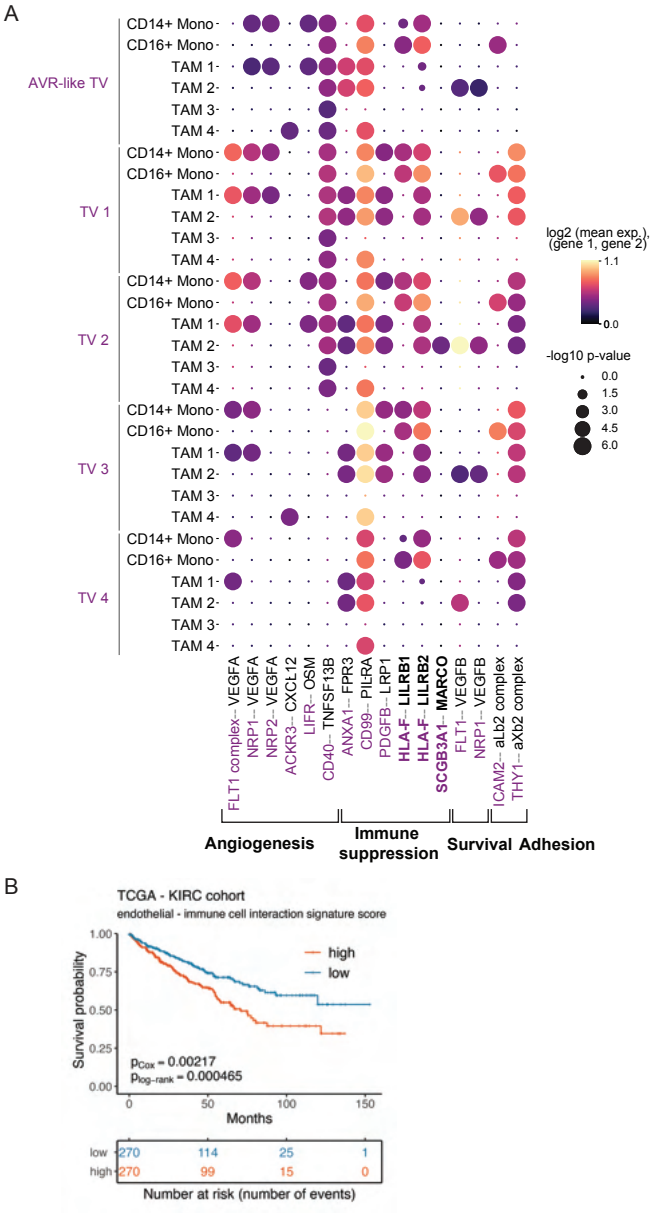


Figure 3.16. Tumor vasculature and immune cell communication assessment and potential clinical significance of the predicted interactions. **A** – selected interactions between tumor vasculature and myeloid cells, revealed

by cell-cell communication analysis using CellPhoneDB. **B** – endothelial - immune cell interaction signature expression in TCGA KIRC cohort was associated with worse overall survival.

Overall, these findings indicate potential tumor vasculature involvement in tumor progression and shaping of the TME niche via immunosuppressive communication with immune cells, as well as expression of tumor-promoting extracellular matrix components and angiogenesis-related genes.

3.2.4. Tumor endothelium subpopulation expresses genes involved in EMT, associated with worse patient survival

The tip cell-like tumor vasculature population (TV 3) expressed genes *LOX*, *PXDN*, *LY6H* and *PGF* (**Figure 3.17, A**), which are not only denoted as tip cell markers, but also have been described to exhibit tumor growth promoting functions. For example, VEGF family member placental growth factor (encoded by *PGF*), was shown to directly interact with VEGF receptors, increasing vascular permeability and plays a role in promotion of immunosuppressive macrophage polarization (339). Additionally, it was demonstrated that PGF-deficient mice tumors promote pro-inflammatory macrophage polarization and vasculature normalization (340). Lysil oxidase *LOX* and peroxidase *PXDN* are involved in growth factor induced endothelial cell proliferation and survival via cross-linking of the collagen type IV rich extracellular matrix and basement membrane (341). Furthermore, inhibition of ECM cross-linking through *LOX* knock-down was shown to impair vessel sprouting (305). Since tumor endothelial tip cell phenotype was not previously described in ccRCC, we assessed whether the TV 3 population is similar to an angiogenic tip cell in lung cancer described extensively by Goveia et al. (305). CellTypist label transfer revealed that transcriptionally, TV 3 population in our dataset corresponds to the tip cell phenotype (**Figure 3.17, B**). Therefore, the tumor vasculature 3 population likely represents the leading tip cell in angiogenic sprouting and, considering the similarity to the tip cells in lung cancer, could be similarly involved in promoting tumor progression.

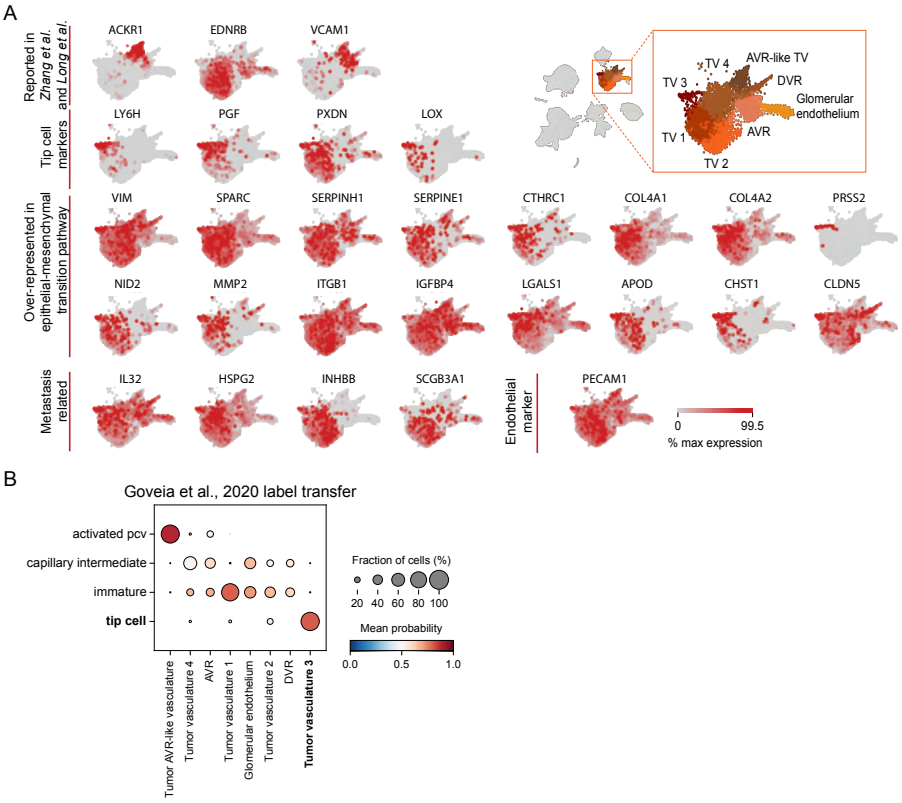


Figure 3.17. Discovering the tip cell-like endothelial cell phenotype in ccRCC. **A** – expression of metastasis associated genes, tip-cell markers and genes overlapping with epithelial-mesenchymal transition pathway in tumor vasculature clusters. Color intensity denotes cptt-normalized expression saturating at 99.5th percentile of a given gene's expression level. **B** – CellTypist model label transfer from Goveia et al. (305) dataset. X-axis labels are cell populations from this study and y-axis labels are from Goveia et al. Tumor vasculature populations in our study are similar to immature vessels and Tumor vasculature 3 population represents the tip cell phenotype. pcv – post-capillary vein.

Gene set over-representation analysis for tumor, tumor vasculature and stromal cell populations (top 100 marker genes) using the Molecular Signatures Database Hallmark gene sets revealed enrichment of hypoxia and glycolysis terms in tumor cells, as expected. Interestingly, significant enrichment for epithelial-mesenchymal transition term was observed for all tumor vasculature and stromal cell populations (**Figure 3.18, A**). Notably, specific genes overlapping with the EMT pathway differed between these subpopulations. We next assessed whether the overlapping gene sets have any effect in stratifying patient survival in the TCGA KIRC cohort. This analysis

revealed that EMT pathway overlapping genes for AVR-like tumor vasculature and the TV 3 phenotype were associated with worse overall survival (**Figure 3.18, B, C**). However, even though other tumor vasculature and stromal cells had significant EMT pathway overlap, no effect on patient survival in the TCGA KIRC cohort was observed (**Supplementary Figure S4**). Overall, our results highlight the presence of a previously uncharacterized tip cell-like tumor endothelium subpopulation in ccRCC, associated with invasive features potentially influencing clinical outcomes.

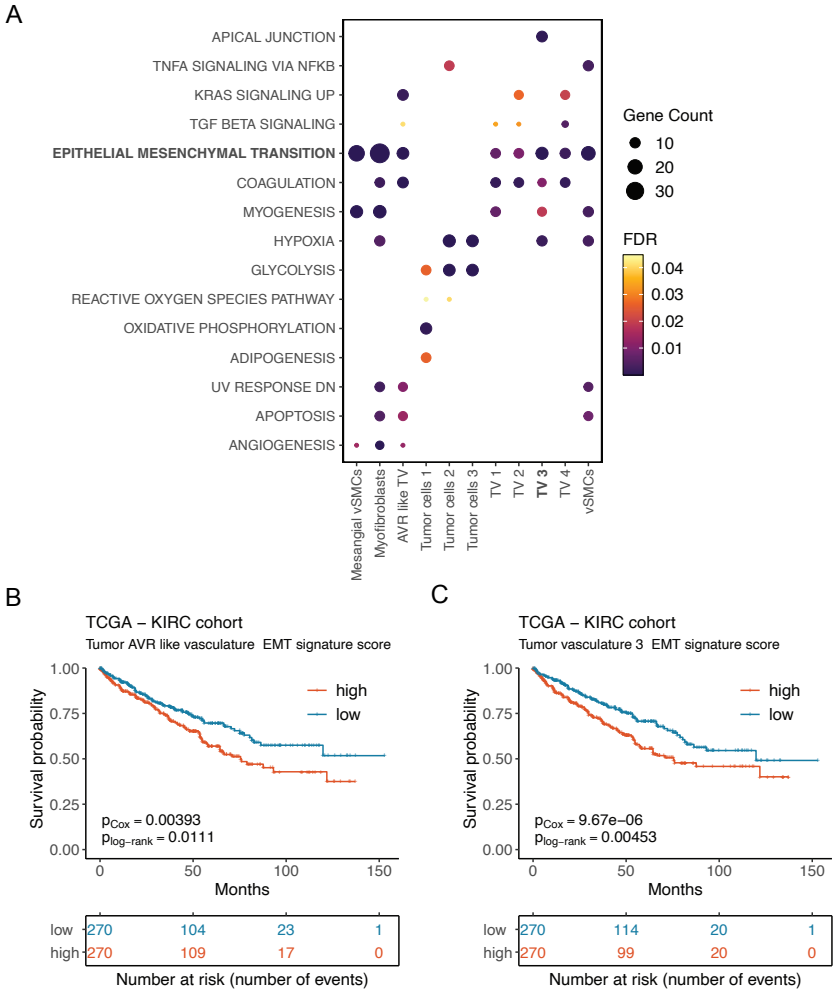


Figure 3.18. MSigDB Hallmark pathway over-representation analysis in tumor, stromal and tumor endothelial cells, and the potential clinical significance of overlapping signatures. **A** – gene set over-representation analysis revealed abundant enrichment of epithelial-mesenchymal transition pathway. **B** – tumor AVR-like vasculature and **C** – tip-like TV 3 signature

genes overlapping with EMT pathway associated with worse overall survival in the TCGA KIRC cohort.

3.2.5. Stromal cells remodel the ECM and communicate with TAMs

Stromal cells have been recognized as important components of the ccRCC TME (322), yet they have received much less attention as compared to tumor or immune cells in previous ccRCC atlases. We detected three populations within stromal cells: vascular smooth muscle cells (vSMCs), myofibroblasts and mesangial/vSMCs (**Figure 3.19, A**). The vSMCs were marked by expression of *TAGLN*, *ACTA2*, *MYH11*, while myofibroblasts were enriched for markers *ACTA2* and *TIMP1*, as well as ECM constituents (collagen types I, III, IV, VI and fibronectin) (**Figure 3.19, B**). Precise annotation of the third stromal cell population was challenging due to simultaneous expression of vSMC genes and mesangial marker *PDGFRB*. Interestingly, for this population, DGE analysis revealed substantial transcriptional differences between tumor and healthy-adjacent tissue. Mesangial/vSMC population found in tumor samples exhibited higher expression of tumor cell marker *NDUFA4L2* as well as stress-related genes, such as *CD36*, which is upregulated in chronic kidney disease and associated with poor prognosis in ccRCC (342,343) and renin (*REN*), which is expressed by mesangial cells under disturbed homeostasis (344) (**Figure 3.19, C**). The transcriptome changes of this stromal cell population might reflect a reaction to the disruptive environmental changes exerted by the tumor.

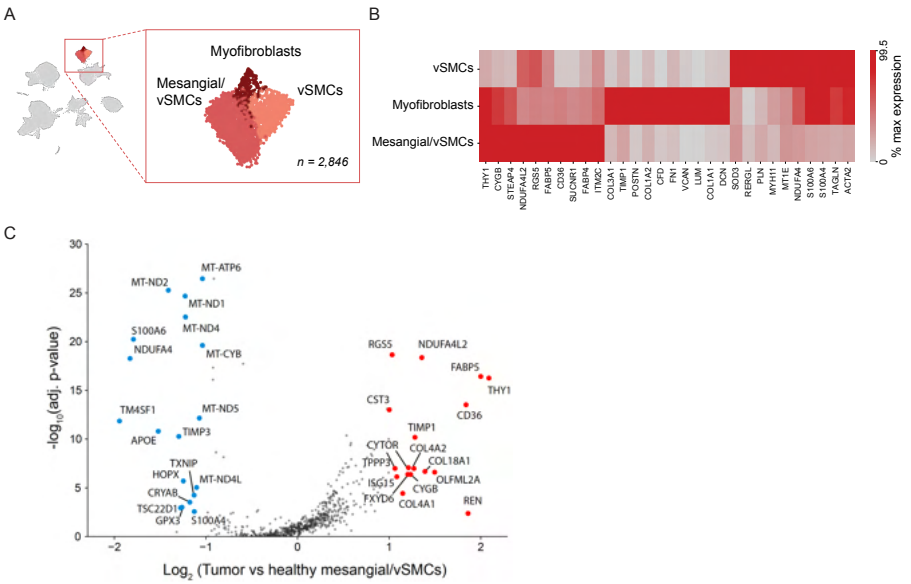


Figure 3.19. Assessing the heterogeneity of stromal cells in the ccRCC TME. **A** – a close-up of the global UMAP stromal cell populations. **B** – Differential gene expression results between stromal cell subpopulations. Only genes with Benjamini-Hochberg adjusted p value < 0.05 are shown. Color intensity denotes cptt-normalized expression saturating at 99.5th percentile of a given gene's expression level. **C** – a volcano plot of differentially expressed genes between mesangial/vSMC cells from tumor vs healthy-adjacent samples. Genes with fold-change of 2 and adjusted p-value <0.05 are highlighted and considered significant. The asymmetry of the central position of the volcano plot is due to difference in cell size with tumor cells having higher fraction of non-zero genes. This effect was not corrected to maintain consistency in differential gene expression analysis performed.

Intracellular communication analysis between stromal and immune cells predicted interactions related to stromal cell survival and proliferation, as well as immune cell adhesion and suppression (**Figure 3.20, A**). Interestingly, stromal cells mostly engaged with TAM 1 and TAM 2 subpopulations in a suppressive manner. For example, *ANXA1-FPR1* interaction was predicted, known to be involved in anti-inflammatory macrophage polarization and tumor progression in various cancers (345,346). We also detected myofibroblast and mesangial/vSMC interaction with cytotoxic T cells via *HLA-E-KLRC1*. This interaction was recently proposed as a targetable mode of T cell exhaustion in bladder cancer (347), moreover, treatment of *HLA-E* positive tumors with anti-NKG2A (encoded by *KLRC1*) antibodies showed high therapeutic potential due to observed restoration of the anti-tumor immunity (348). As with endothelial-immune cell communication, stromal-

immune interaction signature (receptor and ligand gene set) expression in the TCGA KIRC cohort associated with worse overall survival (**Figure 3.20, B**) and increased with advancing disease stage (**Figure 3.20, C**). Collectively, our results indicate that stromal cells could be actively involved in ccRCC TME modulation through therapeutically relevant paths.

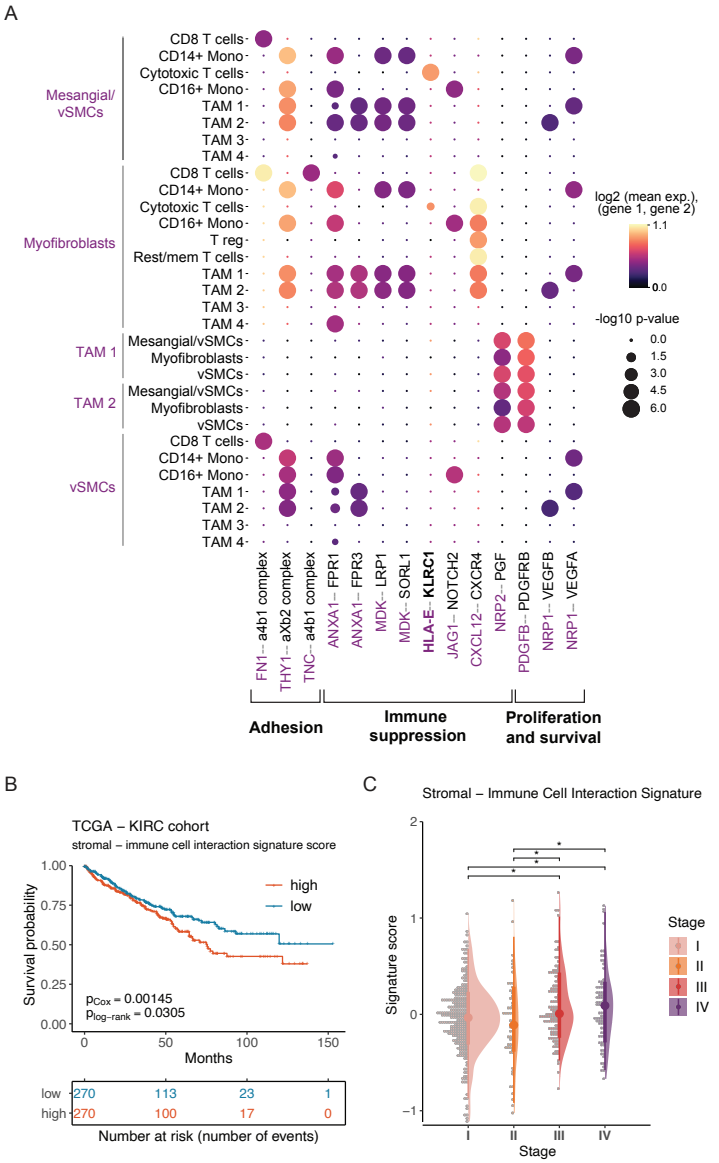


Figure 3.20. Stromal and immune cell communication assessment and potential clinical significance of the predicted interactions. **A** – selected interactions between stromal and myeloid cells, revealed by cell-cell communication analysis using CellPhoneDB. **B** – stromal - immune cell

interaction signature expression in TCGA KIRC cohort was associated with worse overall survival and **C** – increased steadily along the progression of the disease stage.

3.3 An atlas of uncultured human amniotic fluid cells

Amniotic fluid (AF) is a dynamic environment providing mechanical cushioning to the developing fetus and aiding the development of the intestinal tract and the respiratory system. The AF contains a heterogeneous population of cells shed from the fetus that are routinely used in prenatal diagnostics, and is considered to maintain a small fraction (<1%) of multipotent stem cells, characterized (and oftentimes selected) by the expression of c-kit (295), mesenchymal markers (CD44, CD90 and CD105) and pluripotency markers (*SOX2*, *ZFP42* (Rex1), *POU5F1* (Oct-4) and *NANOG*) (349,350). However, the absolute majority of studies to date assess the properties of *in vitro* cultured cells, while characterization of uncultured AF cells is critically lacking. The AF remains overlooked even by the Human Cell Atlas initiative, which harbors a vast collection of single-cell atlases of various tissues and biological systems – there is no comprehensive scRNA-seq investigation of AF cellular contents published to date. Therefore, in the final part of this thesis, single-cell RNA sequencing is applied to investigate the uncultured cells of human AF, providing the first detailed single-cell transcriptomic atlas of this biological niche.

3.3.1. Profiling uncultured human amniotic fluid cells

To investigate the cellular composition of human AF, we profiled fresh AF samples obtained via amniocentesis from consenting donors undergoing genetic testing due to various medical indications. For most patients, no fetal genetic anomaly was detected (**Supplementary Table S3**). Following the same practice as with kidney profiling, we coordinated swift sample delivery (<2hr on ice) and minimized sample handling time excluding cell enrichment of any kind. The collected samples span two developmental timepoints: post-conception week (PCW) 16 (n=19) and PCW 20 (n=7). The cells were washed and subjected to barcoding on the inDrops2-TS platform. After cleanup and quality control, the atlas contained 50,157 cells in total (42,472 cells PCW 16; 7,685 cells PCW 20) which were separated into erythroid, immune and non-immune compartments (**Figure 3.21, A**) based on the expression of canonical markers and the clustering pattern. These fractions, except for the erythroid

cells, were further analyzed in detail (i.e. visualized, clustered, annotated) separately.

Having constructed a broad representation of all cells in AF samples, we asked whether they are of fetal or maternal origin. Demultiplexing was performed using Freemuxlet algorithm, which evaluates single-nucleotide polymorphism (SNP) trends in the raw sequencing data. Results revealed that the absolute majority of cells are likely of fetal origin (94.6% of all cells) (**Figure 3.21, B**). A small fraction (<1%) of cells were assigned maternal origins. Since these cells correspond to immune and erythroid populations, they likely represent blood contamination arising from the sampling procedure. Some, mostly erythroid cells (4,4%) could not be assigned confidently to either origin, likely due to lower library complexity of these cells, however, due to fetal hemoglobin expression (genes *HBG1*, *HBG2*) they are most likely of fetal origin. Considering the minute size of likely maternal cell contamination and generally challenging SNP inference from short-read sequencing data, further analysis was performed on all cells disregarding the inferred origin status. Sample composition analysis by broad annotation revealed a tendency that at earlier gestational age (PCW 16), for most samples, majority of cells profiled correspond to the immune compartment, whereas at a later developmental timepoint (PCW 20) non-immune cells start to dominate (**Figure 3.21, C**), likely reflecting the accumulation of cells shed from various fetal surfaces. Interestingly, in PCW 16 samples affected by chromosome 21 trisomy, non-erythroid immune cells appeared to be noticeably diminished, although the current sample size is insufficient for any further insight.

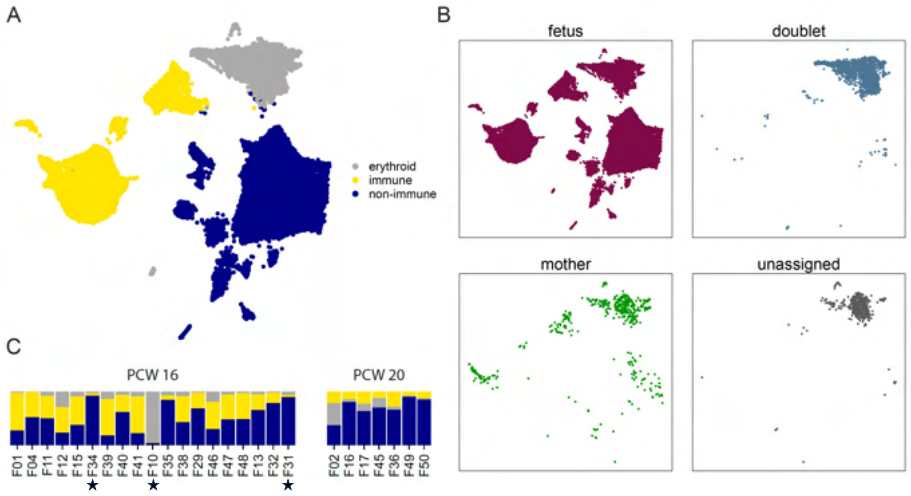


Figure 3.21. Human amniotic fluid single-cell atlas. **A** – a UMAP representation of cells in AF, annotated by broad cell group. **B** – cell origin inference with Freemuxlet predicted fetal origin for the absolute majority of cells. **C** – sample composition by broad cell group, PCW 16 samples had a tendency to comprise more immune cells, while in PCW 20 samples, non-immune cells were the dominant group. Stars denote samples from fetuses affected by chromosome 21 trisomy. PCW – post-conception week.

The number of cells passing quality control differed between samples, ranging from a few hundred to over 4 thousand cells (**Supplementary Figure S5, A**), yet there were no single sample-specific phenotypes observed (**Supplementary Figure S5, B**).

3.3.2. Macrophages and innate lymphoid cells dominate the immune cell population in human AF

Previous investigations into immune cell compartment of human AF established the presence of heterogeneous, mostly innate-immunity related cell populations even in the absence of infection, yet the techniques utilized were limited to the detection of several known marker-defined groups. Motivated by this lack of information, we set out to characterize the AF immune cells at higher resolution by their transcriptomic profiles. To achieve this, a separate embedding was constructed for the immune cell compartment. After cleanup of minor RBC and epithelial cell contamination (~4% cells previously annotated as immune), the dataset was clustered using spectral clustering, and DGE analysis was conducted to obtain marker genes, which were then assessed in the literature to assign cell annotations. Additionally, considering that likely tissue origin for some immune cells is the developing

intestine and lung (274), automatic annotation using CellTypist models, manually trained on published fetal and adult intestinal atlas (306), as well as fetal lung immune atlas leukocytes (307) was performed to aid annotation.

The immune compartment (n=16,942 cells) of AF consisted of both lymphoid and myeloid lineage cells. The most prominent cell type was type 3 innate lymphoid cells (ILC3, **Figure 3.22, A**), expressing markers *RORC* and *KIT* (**Figure 3.22, B** heatmap), which separated into two subpopulations (**Figure 3.22, A**, ILC3 1 and ILC3 2). Additionally, we detected ILC progenitors (*HPN*, *SCN1B*) and proliferating ILC (*UBE2C*, *TUBB*), as well as ILC population specific to samples F01, F02 and F04 (**Figure 3.22, A, B**). This population highly expressed *GCDH*, *CAGE1*, *RGS7*, *CLEC6A* among others (**Figure 3.22, B**), however, these genes do not have major implications in innate lymphoid cell biology and were expressed by non-immune cells in these samples as well (data not shown). Natural killer cells (*GNLY*, *NKG7*), central memory CD8 T cells (*CD27*, *CCR7*, *CD8B*) and a tiny population of mature B cells (*CD79A*, *IGKC*) were also detected (**Figure 3.22, A, B**). In the myeloid cell compartment, we found a small population of basophils (*CLC*, *HDC*), mast cells (*TPSB2*, *TPSAB1*) and monocytes (*S100A9*, *FCN1*), as well as larger group of antigen-presenting cells, such as Langerhans/DC (expressing MHC class II machinery, *CD207*) and macrophages (*CD68*, *CD14*, complement system). The latter group was diverse, and separated into three subpopulations: likely monocyte-derived macrophages, marked by monocytic gene *FTL*, *AIF1*, *S100A9* expression and two groups of M2-polarized (CD206 positive, gene *MRC1*), anti-inflammatory macrophages (**Figure 3.22, A**). Of the latter, one population was marked by *LYVE1*, as well as *RNASE1*, *SPP1* and *VSIG4* expression, similar to a population recently described in the fetal intestine (306). Interestingly, while these cells were positive for M2 polarization markers, such as *MRC1*, *CD163*, *DAB2*, they also expressed chemokine *CCL2* associated with pro-inflammatory functions (**Figure 3.22, B**). *LYVE1* macrophages represent a tissue-resident population that was reported to be closely associated with blood vessels in various tissues (i.e., heart, lung) (351). The other M2-polarized macrophage subpopulation expressed matrix metalloproteinases (i.e. *MMP9*, *MMP12*, *MMP14*, *MMP19*), and while *MMP9* is considered to mark pro-inflammatory macrophages, expression of *APOE*, *APOC1*, *TREM2*, *CTSB*, *CTSD* hints toward an alternatively-activated, tissue remodeling phenotype (352). Additionally, we observed a tendency that the immune cell composition of AF changes with gestational age – in PCW 16 samples, lymphoid lineage cells dominated, whereas PCW 20 specimens were enriched in myeloid cells (**Figure 3.22, C**).

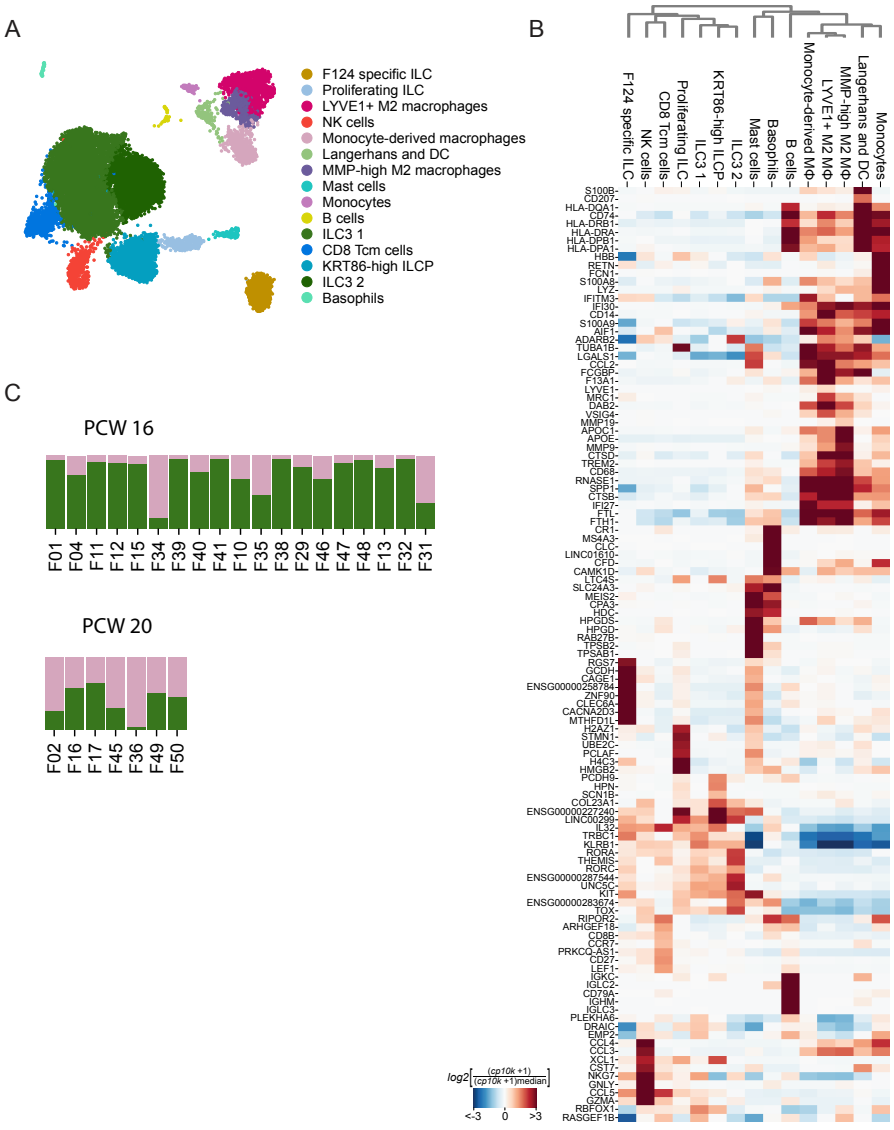


Figure 3.22. Analysis of immune cells in human AF. **A** – a UMAP of immune cells in AF, annotated by cell type. n=16,942. **B** – a heatmap and hierarchical clustering dendrogram of top differentially expressed genes (ranked by fold-change) between the immune cell phenotypes, Mann-Whitney U test, Benjamini-Hochberg adjusted p-value <0.05. Markers *SCN1B*, *CCR7*, *HDC*, *CD68*, *MRC1*, *MMP9*, *CD8B*, *RGS7*, *CD14*, *TREM2* appear in the top 50 significant DEG lists, but were added manually to this plot. Genes *KIT*, *LYVE1*, *VSIG4*, *MMP19*, *RORC*, *CD207* were added manually. **C** – sample immune cell compartment composition by broad lineage (green – lymphoid, pink – myeloid). Lymphoid cells tended to dominate at PCW 16, with myeloid lineage emerging at PCW 20.

Automatic annotation with CellTypist models manually trained on previously published fetal tissue atlases provided further support for the assigned cell identities (**Figure 3.23, A, B**). Both lung and intestine models' predictions coincided with most manually assigned cell types, however, monocyte-derived macrophages were assigned erythroid (or RBC) identity in both cases, likely reflecting model sensitivity to minor hemoglobin ambient RNA contamination present in this cluster that comes from a few samples (**Figure 3.22, B,** gene *HBB*). Interestingly, the gastrointestinal atlas model assigned Natural Cytotoxicity Receptor (NCR) negative lymphoid tissue inducer phenotype for both ILC3 populations, and NCR positive phenotype for a fraction of ILCP and cycling ILC cells. Moreover, both models predicted a fraction of ILC3 1 population a T cell identity – follicular helper and NK T cells by the intestine model (**Figure 3.23, A**, pale purple), and type 3 innate T cells from the lung model (**Figure 3.23, B**, pale pink). Inspection of curated ILCP, ILC, T and NK cell-related genes highlighted differences between ILC3 1 and ILC3 2 populations and revealed heterogeneity within the ILC compartment (**Figure 3.23, C**). Unexpectedly, ILC progenitors were positive for *NCR1* and *NCR2*, and expressed chemoattractant *XCL1* and *XCL2,* normally associated with activated T cells and NK cells. Moreover, hair keratins *KRT86* and *KRT81* appeared as pronounced markers for ILCP, and while ILC-specific expression of these genes can be detected in published fetal tissue atlases, in the literature their marker status and role remain elusive. Proliferating ILC were also positive for NCRs and ILC markers, as well as T and NK cell related genes, likely reflecting a mix of cycling ILC phenotypes present in this cluster. Sample F01, F02 and F04-specific ILC presented the highest T cell and cytotoxicity signature expression (*CD8B*, *TRBC1*, *TRBC2*, *NKG7* etc.), yet were positive for ILC markers (especially *KIT*, *CA2*) (**Figure 3.23, C**). Both ILC3 groups were indeed negative for *NCR2* and expressed ILC markers, yet ILC3 2 population had a more pronounced ILC signature (genes *RUNX3*, *RORA*, *RORC*, *CCR6*). Interestingly, ILC3 1 population had higher expression of T cell related (*KLRB1*, *TRBC1*, *TRBC2*) and some NK cell, cytotoxicity related genes (*NKG7*, *GZMA*). Thus, it appears that ILC3 1 population has a more T cell-like expression profile, while ILC3 2 cells likely represent a purer ILC phenotype (**Figure 3.23, C**). On a side note, in ILC3 1 population, the T cell-related gene expression was not limited to the cells with a CellTypist predicted T cell identity (**Supplementary Figure S6**). Moreover, it was observed that in different samples, either ILC3 1 or ILC3 2 phenotype is preferentially abundant (**Figure 3.23, D**), and the abundance of ILC3 1 has a tendency to coincide with a higher fraction of CD8 T cells. Naturally, that could suggest that the T cell-related gene expression in ILC3 1 population is

ambient RNA contamination of the co-present CD8 T cells, however, in such case all other cells in these samples (i.e. macrophages) would have the same level of T cell gene expression, yet that was not observed (**Supplementary Figure S6**). Thus, T cell related gene expression in ILC3 1 population is not an artifact, but true biological signal, emphasizing the complexity and continuity of the phenotypical space of the lymphoid cell lineage, and the potential T cell exerted impact on the ILCs in AF. On a side note, ILC3 2 group appeared to be preferentially abundant in samples from fetuses affected with genetic abnormalities (one triploidy (F46) and one XXY sample (F15)).

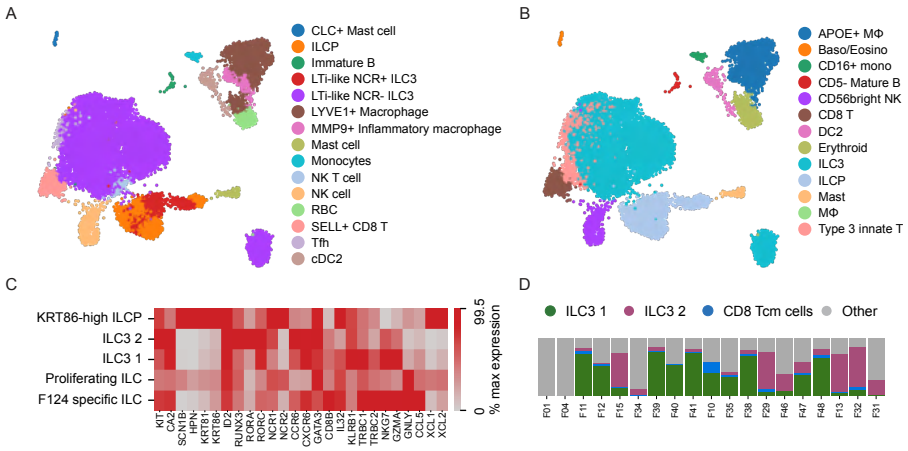


Figure 3.23. Analysis of immune cells in human AF using models built from publicly available data. **A** – a UMAP of immune cells in AF, annotated using CellTypist model built from the gastrointestinal atlas (306) and **B** – fetal lung leukocyte atlas (307). **C** – a heatmap depicting selected ILCP, ILC, T and NK cell-related cptt-normalized gene expression in AF ILC populations, color saturating at 99.5th percentile of a given gene's expression level. ILC3 population 1 had higher expression of T cell-related genes. **D** – PCW 16 sample immune cell composition, with ILC3 1 and 2 populations highlighted in green and purple, respectively. Samples seemed to contain preferentially one of these subpopulations.

While the presence of ILC and macrophages in AF is established in the literature, we report the unprecedented scope of immune cell diversity (i.e. presence of mast, dendritic cells, basophils, NK cells, cycling ILC population etc.) and delineate phenotypic heterogeneity of macrophages and innate lymphoid cells in human AF for the first time.

3.3.3. AF contains highly specialized tissue shed cell populations

The AF is known to contain fetal cells from various surfaces in contact with the fluid, shed naturally during development. Absolute majority of them (95-99%) are considered to be dead, except for a tiny fraction of mesenchymal stem or, as recently proposed instead, epithelial progenitor cells (287). Despite interest in the potential applications of the stem cells, the non-manipulated (i.e. sorted by known markers, cultured) non-immune AF cell compartment remains severely under-investigated. Following the same approach as with immune cells (i.e. embedding construction, clustering, DGE), non-immune cells were characterized. Additionally, the assigned identities were assessed further with marker gene set over-representation analysis using Gene Ontology Biological Process 2023, MSigDB Hallmark 2020 and Reactome 2022 databases.

Even though sample preparation did not involve any kind of live cell enrichment, inspection of raw counts in unfiltered libraries indicated low ambient RNA profiles and mitochondrial gene count fractions (**Supplementary Figure S7, A**). Additionally, non-immune cells (n=26,160) in our dataset had adequate QC metrics typically assessed in scRNA-seq data analysis when excluding dead or dying cells (**Supplementary Figure S7, B**). Thus, it appears that we were able to capture predominantly live cells.

The most abundant cell type in the non-immune fraction was squamous epithelial cells, as expected due to fetal surface shedding (**Figure 3.24, A**). This large group consisted of squamous epithelium, enriched in metallothionein expression (*MT1X*, *MT2A*); two subpopulations of keratinocyte-like squamous epithelial cells, marked by *KRT6A* and cornified envelope marker *SPRR2A* expression; esophagus-like squamous epithelium (*S100A14*, *CEACAM6, CAPN14*); lncRNA and NOMO-enriched (*NOMO1, NOMO2, NOMO3*) subpopulations (**Figure 3.24, A, B**). Interestingly, among the squamous epithelia we detected a relatively small (n=346 cells) population highly expressing chemokines *CCL20*, *CXCL2*, *CXCL3*, *CXCL8,* involved in response to tissue damage and inflammation (**Figure 3.24, B**). Additionally, we detected cells likely comprising hair follicles, as this population had elevated expression of genes highly expressed in inner root sheath and companion layer cells (i.e. *KRTDAP*, *LY6G6C*, **Figure 3.24, B**), as determined by interactive exploration of skin development atlas (353). We also detected an epithelial (*CDH1*+) population marked by plasticity-associated *SOX4* and *PLCG2* (309), as well as lung progenitor marker *ID2* expression (**Figure 3.24, B**), hence annotated as lung progenitors. Several populations of mature, highly specialized fetal tissue cell types were observed. These included umbrella

cells of the urothelium (enriched in uroplakins, i.e. *UPK1A*, *UPK2*); podocytes of the kidney (*PODXL*, *NPHS1*); ciliated cells of the lung (*CAPS*, *TMEM190*); a mix of goblet and enteroendocrine cells (*FCGBP*, *ZG16*, *CHGA*); enterocytes (*APOA1*, *SELENOP*, *ANPEP*) (**Figure 3.24, B**). Intriguingly, a few cell populations were challenging to annotate due to simultaneous expression of epithelial and mesenchymal markers (**Figure 3.25**). Namely, a population annotated as epithelial-mesenchymal expressed mesenchymal marker *VIM* together with epithelial keratins (*KRT8*, *KRT18*). Another group, *MMP7*-high fibroblast-epithelial cells had high expression of *MMP7* (**Figure 3.24, B**), mesenchymal markers *VIM*, *CD44* and transgelins, simultaneously with canonical epithelial markers, such as *EPCAM*, *CDH1*, *KRT7* and others (**Figure 3.25**). Finally, a *KRT5*-high fibroblast population was observed, with high fibronectin and other mesenchymal marker expression, yet enriched for basal stratified epithelium marker *KRT5*, as well as other keratins (**Figure 3.25**).

Gene set over-representation analysis of the top 200 differentially expressed genes using multiple public databases corroborated the assigned annotations. For instance, cilliated cells showed significant enrichment of cillium assembly and movement related pathways; enterocytes displayed enrichment of chylomicron assembly; endosomal/vacuolar pathway was activated in goblet and enteroendocrine cells (**Figure 3.26**). Accordingly, hair follicle cells upregulated genes related to estrogen response, which is crucial for hair follicle growth phase regulation; myofibroblasts had enrichment of collagen synthesis and ECM organization, while podocyte DEGs had highest overlap with podocyte development and differentiation pathways (**Figure 3.26**). Analysis of squamous epithelium subpopulations revealed enrichment of keratinization, keratinocyte and epidermal cell differentiation in both keratinocyte-like clusters, with keratinocyte-like 2 cluster presenting with a higher degree of overlap, perhaps reflecting a more differentiated nature than cluster 1 (**Figure 3.27**). NOMO-high and generic squamous epithelium ORA results reflected high expression of metallothioneins and their associated metal sequestration functions, while inflammatory epithelium DEGs had significant overlap with TNF-α signaling via NFκ-B pathway (**Figure 3.27**).

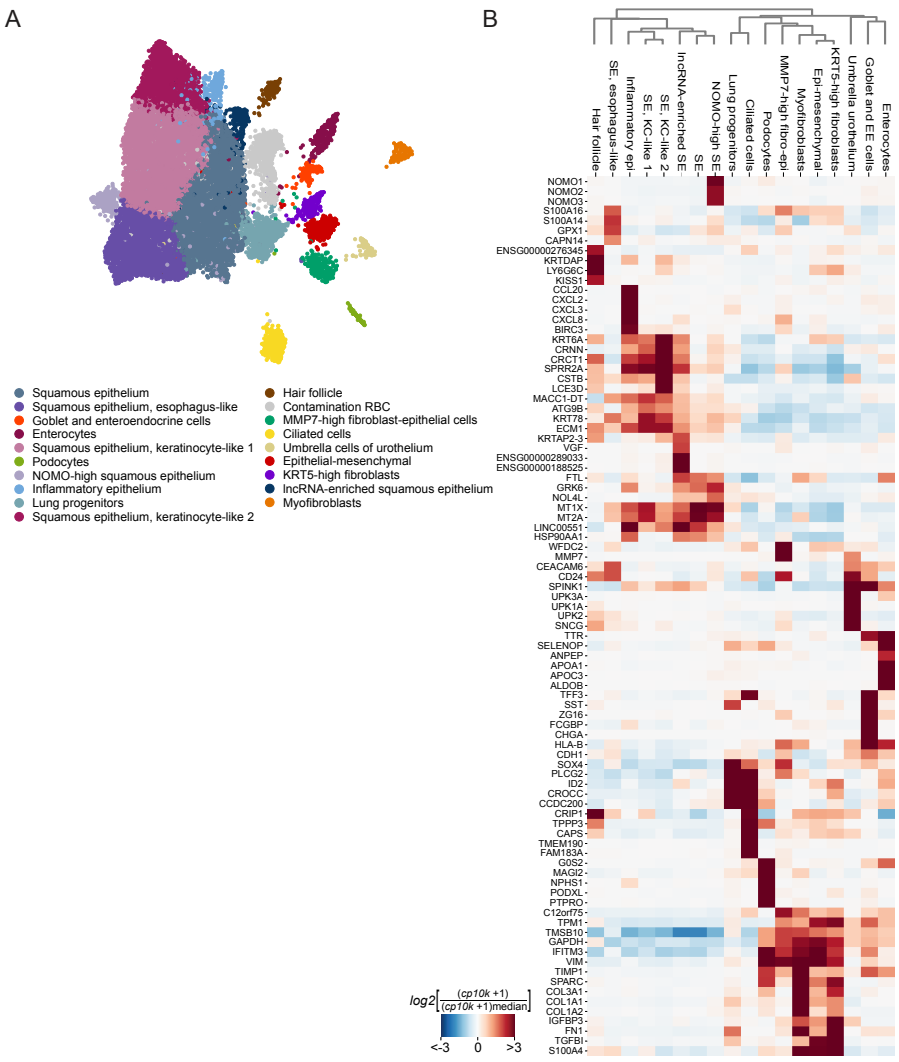


Figure 3.24. Analysis of non-immune cells in human AF. **A** – a UMAP of non-immune cells in AF, annotated by cell type. n=26,160. **B** – a heatmap and hierarchical clustering dendrogram of top differentially expressed genes (ranked by fold-change) between the non-immune cell phenotypes, Mann-Whitney U test, Benjamini-Hochberg adjusted p-value <0.05. Markers *CAPN14, CHGA, ANPEP* appear in the top 50 significant DEG lists, but were added manually to this plot. Gene *CDH1* was added manually. EE – enteroendocrine cells, Epi – epithelial, KC – keratinocyte, SE – squamous epithelium.

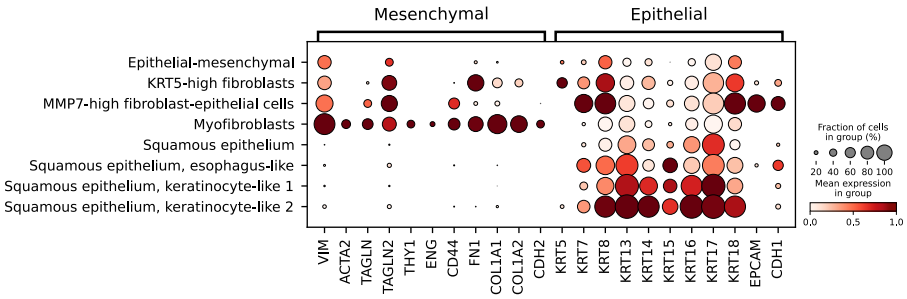


Figure 3.25. Analysis of mesenchymal and epithelial marker expression in squamous epithelium, myofibroblasts and intermediate epithelial-mesenchymal phenotypes. The latter presented heterogeneous expression of markers from both groups. The dot size depicts the fraction of cells expressing a given gene, while color indicates the log-cptt-normalized expression level, scaled per variable.

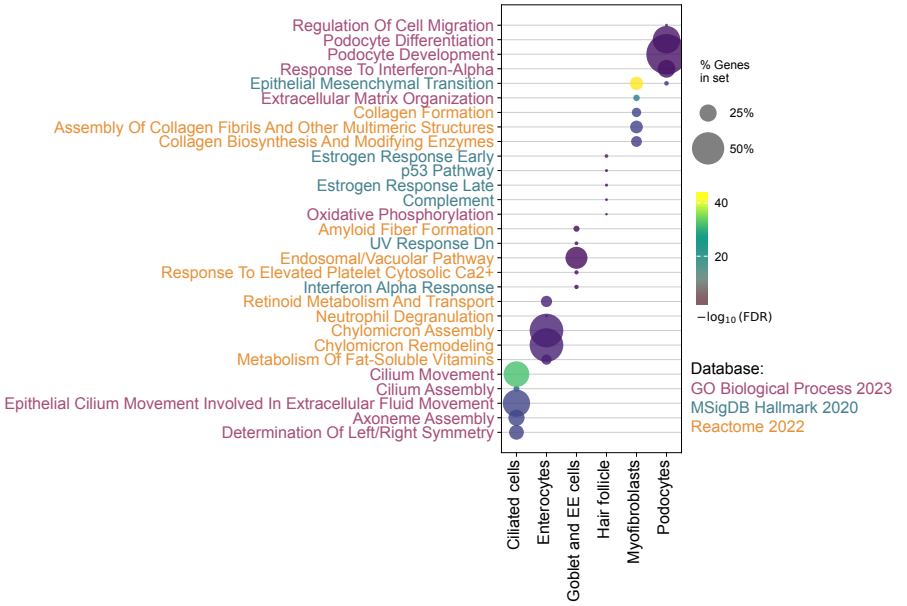


Figure 3.26. Gene set over-representation analysis of the top 200 differentially expressed genes for each specialized cell population. Overlapping terms are color-coded by the database used. Results support highly specialized cell functions in concordance to annotations assigned.

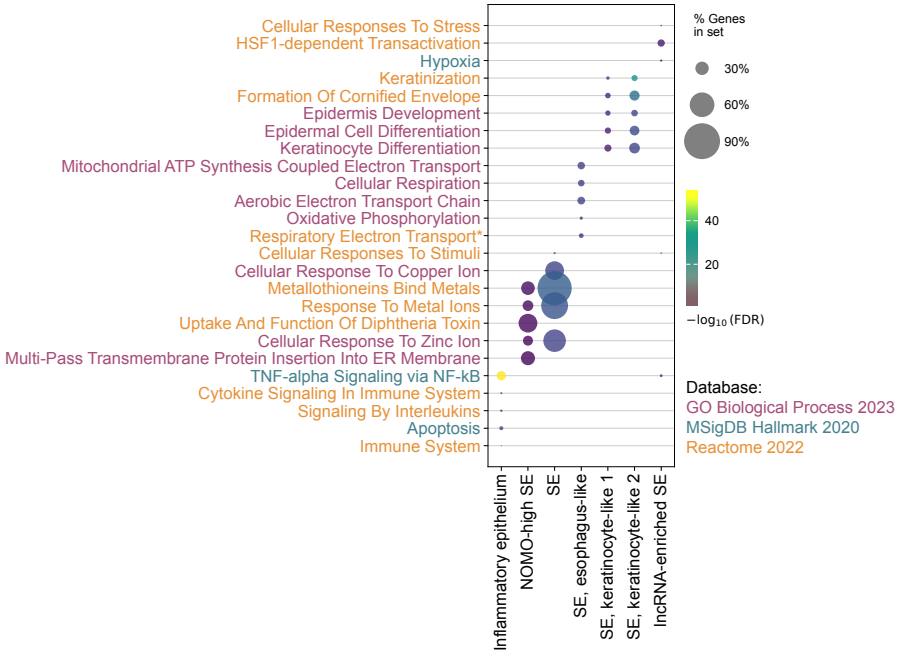


Figure 3.27. Gene set over-representation analysis of the top 200 differentially expressed genes for each epithelial cell population. Overlapping terms are color-coded by the database used. Results support annotations assigned. A * indicates a term shortened for visualization purposes, full term was "Respiratory Electron Transport, ATP Synthesis By Chemiosmotic Coupling, Heat Production By Uncoupling Proteins".

With regards to intermediate phenotypes observed, all three clusters presenting with both epithelial and mesenchymal marker gene expression showcased significant overlap with epithelial-mesenchymal transition pathway (**Figure 3.28**). Interestingly, epithelial-mesenchymal population was also enriched in hypoxia, angiogenesis and glycolysis-related processes, while *KRT5*-high fibroblasts expectedly had overlap with collagen formation and ECM organization pathways (**Figure 3.28**). Both *KRT5*-high fibroblasts and *MMP7*-high fibroblast-epithelial cells showed significant enrichment of plateled degranulation pathway, likely reflecting the secretion of various ECM constituents and growth factors, as the pathway contains not solely platelet markers. Additionally, these populations had overlap with terms related to apoptosis and estrogen response, which might indicate active cellular processes related to survival or growth factor response (**Figure 3.28**).

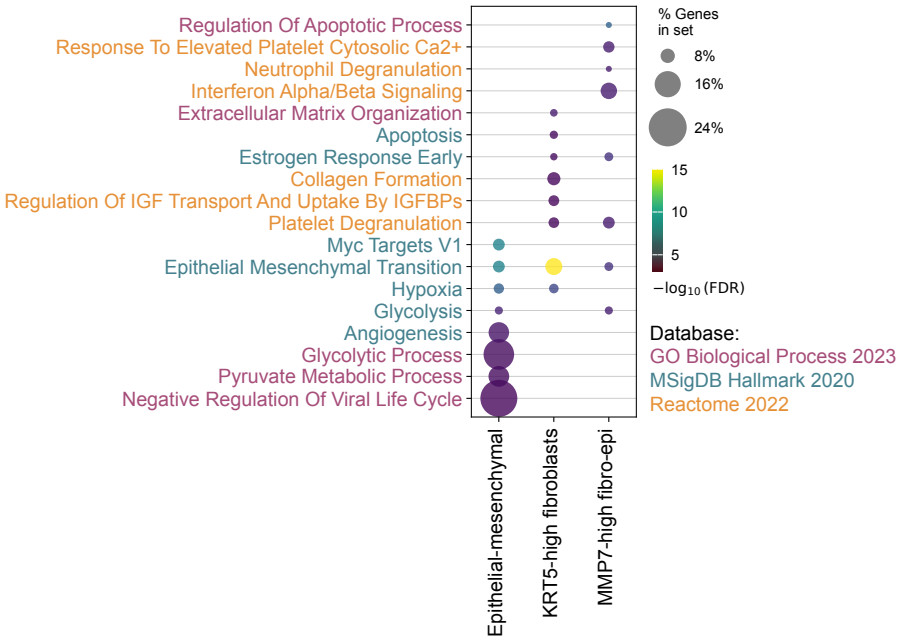


Figure 3.28. Gene set over-representation analysis of the top 200 differentially expressed genes for each intermediate phenotype population. Overlapping terms are color-coded by the database used. Notable term is "Epithelial Mesenchymal Transition", which significantly overlapped for all three populations.

Overall, these results highlight the heterogeneity of specialized tissue cells shed into AF, as well as the presence of intriguing intermediate phenotypes, to our knowledge, not described elsewhere.

3.3.4. Uncultured cells in AF do not express pluripotency-related genes

Considering the potential applications and interest in AF-derived stem cells, and the suitability of droplet-based single-cell transcriptomics technology to uncover such rare (estimated <1%) populations, we sought to investigate the presence of mesenchymal AFSCs in native, uncultured fluid. Additionally, the presence of tissue-specific (kidney, lung, intestinal) epithelial progenitors was examined. For that, expression of AFSC (286,295,350,354,355) and epithelial progenitor (287,288) marker genes was assessed in the non-immune cell populations.

Interestingly, co-expression of pluripotency markers, such as Oct-4 (gene *POU5F1*), *SOX2*, *NANOG*, Zfp42 (gene *REX1*) was not detected in any of the populations, likewise for AFSC-associated marker c-kit (also known as

CD117, gene *KIT*) and HLA-DR genes, reported to be positive in a fraction of AFSCs (**Figure 3.29**). These results are unexpected, considering that AFSCs were reported to be isolated solely based on the presence of c-kit. Important consideration is that the presence of a transcript does not indicate the presence of a protein and vice versa. However, even aside c-kit, it is evident that the expected transcriptional profile was not present in any of the non-immune populations. Expression of other, mesenchymal markers, such as *CD44*, CD90 (gene *THY1*), CD73 (gene *NT5E*), CD105 (gene *ENG*) and CD29 (gene *ITGB1*) was limited to myofibroblast cells (**Figure 3.29**). Additionally, *MMP7*-high fibroblast-epithelial cells expressed *CD44*, while integrin β1 (*ITGB1*) was expressed to a certain extent in most populations, consistently with its general involvement in cell adhesion.

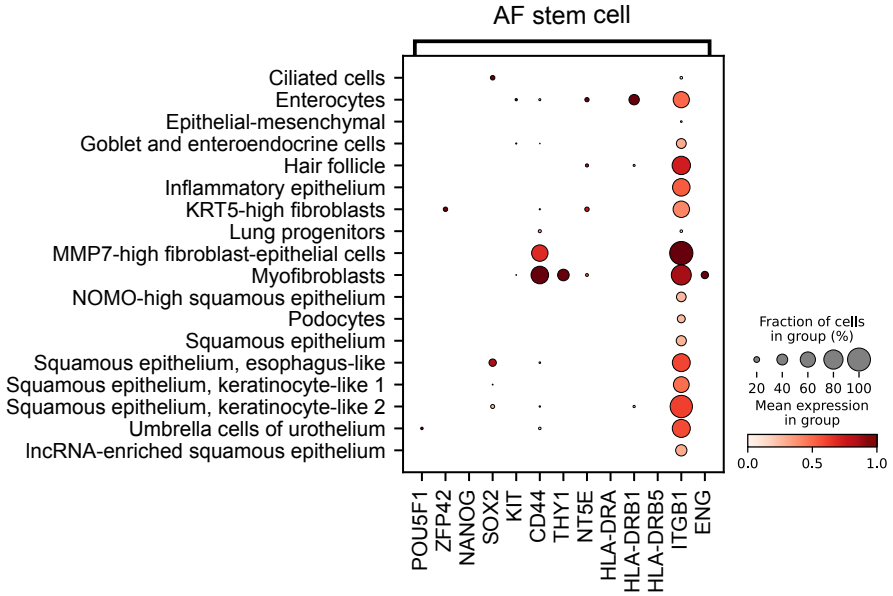


Figure 3.29. Analysis of amniotic fluid mesenchymal stem cell-associated marker gene expression in all non-immune cell populations. No cell type had consistent co-expression of the genes assessed. The dot size depicts the fraction of cells expressing a given gene, while color indicates the log-cptt-normalized expression level, scaled per variable.

Similarly to the pluripotency profile, no population had coherent co-expression of any of the tissue-specific progenitor marker gene sets, compiled from Gerli et al. (287) and Babosova et al. (288) (**Figure 3.30**). Lung progenitor population expressed *ID2*, *PLCG2* and *SOX4*, but not other markers in the set, indicating either non-canonical phenotype or not precise assignment of identity, although the other lung-derived cell population –

ciliated cells – expressed the same genes (**Figure 3.30**). Interestingly, a fraction of *MMP7*-high epithelial-mesenchymal cells expressed a notable portion of genes from the kidney progenitor signature: embryonic development-related transcription factor-encoding *PAX2*, *PAX8*, *EMX2*, *LHX1*, *GATA3* and *POU3F3* (**Figure 3.30**). Thus, it is likely that this intermediate phenotype does indeed originate from the fetal kidneys and might have stem or progenitor properties, although not pluripotency.

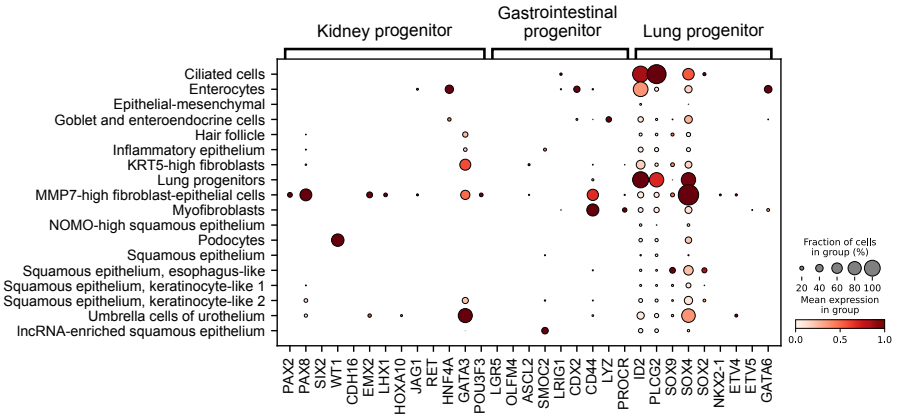


Figure 3.30. Analysis of tissue-specific epithelial progenitor-associated marker gene expression in all non-immune cell populations. Gene sets were compiled from Gerli et al. (287) and Babosova et al. (288). Some MMP7-high fibroblast-epithelial cells had expression of several markers from the kidney progenitor signature. The dot size depicts the fraction of cells expressing a given gene, while color indicates the log-cptt-normalized expression level, scaled per variable.

Together, these results indicate the absence of conventional, widely cultured and studied mesenchymal and c-kit positive AFSCs in the uncultured or otherwise manipulated AF samples. Likewise, the presence of gastrointestinal progenitors could not be established, while the population annotated as lung progenitors does not completely align with their proposed signature and will require further assessment. Nonetheless, we discover that *MMP7*-high phenotype, co-expressing some mesenchymal and epithelial features, likely corresponds to a progenitor population derived from fetal kidneys and is present in AF.

4. DISCUSSION

The first part of results presents an improved droplet-based scRNA-seq method inDrops-2, with higher sensitivity and user-friendly TS-based library preparation protocol. Using the updated method, multi-regional profiling of lung carcinoma tissues (n=18) was performed, resulting in characterization of heterogeneous phenotypes present in the tumor microenvironment. This analysis led to recovery of not only all major specialized lung epithelial, stromal and infiltrating immune populations (7,319,320), but also resulted in characterization of several cell phenotypes with potential clinical significance that could spark interest for future studies.

One of the interesting phenotypes observed is *CXCL13*-high CD4 T cells. CXCL13 expressing CD4 T cells have been demonstrated to be involved in tertiary lymphoid structure formation, as CXCL13 is a potent attractant for B and other immune cells (356). Indeed, in our dataset, *CXCL13*-high CD4 T cell abundance coincided with higher B cell numbers. Interestingly, it was demonstrated that abundance of *PDCD1*-high CXCL13 producing CD8 T cells predict response to PD-1 blockade therapy and correlate with increased overall survival in non-small cell lung cancer (357). In our dataset, *CXCL13*-high CD4 T cells were also positive for *PDCD1*, likely reflecting a similar, potentially clinically relevant phenotype. Other noteworthy populations captured are patient-specific *HAS1*-high fibroblasts and *SPINK1*-high club cells. *HAS1*-high fibroblasts were recently discovered in idiopathic pulmonary fibrosis samples as invasive contributors to pathologic ECM formation (318), yet their presence in lung carcinoma had not been described previously. Thus, we hypothesize that this population could be of interest for future studies, especially considering the growing interest on stromal cell involvement in cancer (358). *SPINK1*-high club cells featured expression of canonical club cell markers (*SCGB3A1*, *SCGB3A2*), as well as distal lung marker *NAPSA* and *CEACAM6*. *SPINK1* is upregulated and its expression correlates with adverse outcomes in a multitude of cancers (359), moreover, it was shown to induce proliferation, migration and invasion of lung adenocarcinoma cancer cells in vitro (360). Meanwhile, *CEACAM6* has been implicated in lung cancer progression and associates with poor clinical outcomes (316). Such expression pattern in *SPINK1*-high club cells is a novel finding and hints towards an altered, potentially pro-tumor phenotype, however, it remains unclear whether these changes are imposed by the tumor microenvironment signaling or malignant transformation of these cells. Another interesting finding was two phenotypes of alveolar epithelial cells – both of them expressed canonical AEC markers (i.e. *SFTPA1*, *SFTPA2*,

SFTPC), but one population had elevated *MMP7* and *PRSS2* expression. *MMP7* is a widely used biomarker for pulmonary fibrosis, while *PRSS2* expression is associated with invasive and metastasis promoting features (317). Recently it was demonstrated that *PRSS2* and *MMP7* are co-expressed in pre-alveolar transitional state epithelial cells in idiopathic pulmonary fibrosis (361). Thus, the *MMP7*-high alveolar epithelial cell population might be involved in disease progression and plasticity.

It is important to acknowledge that the analysis of lung carcinoma samples described in this work is limited to transcriptomic characterization and insights into the roles of the aforementioned novel phenotypes remain speculative. Nonetheless, in the context of inDrops-2 development, multi-regional lung carcinoma sample analysis highlighted the broad potential of the method for the recovery of rare, interesting phenotypes in clinical samples that have been preserved, stored long-term and multiplexed. Considering the limited availability and logistic challenges with fresh tissue samples, this work represents a major advancement, enabling longitudinal studies and biobanked specimen investigation. Importantly, the high-throughput profiling of single cells with inDrops-2 is highly customizable, scalable, and inexpensive, democratizing single-cell technologies.

The second part of this thesis presents an in-depth investigation of the cellular heterogeneity within the ccRCC tumor microenvironment and healthy kidney tissue. To achieve this, single-cell transcriptomes of human ccRCC tumor samples along with healthy adjacent tissues were profiled using the inDrops-2-TS platform. In contrast to previous studies, we did not employ cell enrichment methods, and rapid isolation of cells resulted in the ability to capture rare and sensitive phenotypes, significantly depleted or absent in previous reports. Given that the immune compartment in our dataset largely recapitulated previous findings (230–234,244), we mainly focused on the phenotypic heterogeneity and cellular interactions of the often overlooked and underappreciated endothelial and stromal cell populations.

Endothelial cells are very important in ccRCC development and to this day remain the main therapeutic targets in advanced and metastatic disease (211). The highlight of our analysis in the endothelial compartment was the discovery of tip-like cells, not detected previously in the context of ccRCC. Moreover, these cells were enriched for EMT pathway genes which associated with poor overall patient survival. Previous single-cell investigations of the ccRCC TME also captured endothelial cells, however, they were most often represented by two major phenotypic groups. For example, Long et el. reported *VCAM1*+ and *VCAM1*- vasculature populations, while Zhang et al.

presented *ACKR1*+ and *EDNRB*+ endothelium groups (222,362). Consistently, in our dataset we identified a population, namely AVR-like vasculature, co-expressing the *vasa recta* marker *ACKR1* and *VCAM1*, however, *EDNRB* was detected in tumor vasculature 1, 2 and 4, but not tumor vasculature 3. This provides further support that the endothelial (*PECAM1*+) TV 3 phenotype has not been characterized in ccRCC (**Figure 3.17, A**).

While not previously found in ccRCC, the tip cell population (TV 3) in our dataset corresponds to a tip cell phenotype observed in lung cancer (*LOX*, *PXDN*, *PGF*, *LXN*, type IV collagen enriched) (**Figure 3.17**) that correlated with worse patient survival (305). Interestingly, the authors found this phenotype to be the most congruent across species and tumor types, including kidney cancer, as determined using bulk proteomics data. This raises a question why previous single-cell atlases of ccRCC did not capture this rare phenotype. Furthermore, Goveia et al. demonstrated that *LOX* knock-down impairs vessel sprouting, suggesting that the TV 3 population might likewise be of interest for future research as a potential therapeutic target.

Congruent with our results, Long et al. showed that the *VCAM1*+ (AVR-like TV in our dataset) is enriched for the EMT signature (362). However, in our dataset, not only AVR-like vasculature, but all tumor vasculature and stromal cell populations had significant overlap with the EMT pathway. On the other hand, the overlapping gene set relation to patient survival was only pronounced for the AVR-like and TV 3 tumor vasculature, further emphasizing the diversity of tumor endothelial cells. In another study, endothelial and stromal cell association with EMT was also showed, but no distinction of tumor and healthy endothelium was made, furthermore, the authors reported lower endothelial cell abundance in tumor samples as compared to healthy tissues (244). Such discrepancies between different studies could be related to technical aspects of sample preparation and data analysis, and further underscore the need and importance for accurate phenotypic characterization of tumor endothelial cells in ccRCC.

Our findings suggest two major modes of action of the tumor vasculature cells in the ccRCC TME. First, remodeling of the ECM by deposition of various ECM constituents and expression of their modifying agents, and second, participation in cellular communication, involved in immune suppression and angiogenesis promotion. Using spatial transcriptomic profiling of ccRCC specimens, it was shown that collagen-producing endothelial cells localize at the tumor-normal interface, rich in IL1B+ macrophages and EMT-enriched tumor cells (234). Our results also suggest that tumor vasculature cells might contribute to EMT and interact with TAMs. Interestingly, the intercellular communication analysis uncovered various

interactions between tumor endothelial and stromal cells with immune cells that are of clinical relevance. For example, a phase I-II clinical trial for LILRB1 and LILRB2 inhibitor as a monotherapy or in combination with anti PD-1 ICB began in 2021, for treatment of advanced or metastatic solid tumors, including ccRCC (ID NCT04913337). LILRB2 inhibition reprograms myeloid cells to a pro-inflammatory state, while LILRB1 blockade stimulates the reprogramming of both myeloid and lymphoid cells. Our results suggest that *LILRB1/2*+ immune cells interact not only with tumor cells, but also with tumor endothelial cells. Likewise, regulatory T cell-expressed TIGIT was predicted to engage with NECTIN2, expressed by endothelial cells. This interaction has increasingly gained attention over the last few years and is currently being investigated in a multitude of clinical trials (363). Another noteworthy predicted interaction was between TV 2-expressed *SCGB3A1* and TAM 2-expressed *MARCO*. The gene *SCGB3A1* encodes a secretoglobin family member produced by endothelial cells, and it was recently demonstrated that it is a critical component of a pro-metastatic niche, inducing stem cell properties in cancer cells, while macrophages are required to maintain the pro-tumor niche (364). To our knowledge, this interaction was not previously observed in the context of ccRCC.

Stromal cells in our dataset exhibited immunosuppressive interactions with immune cells, suggesting their participation in tumor-promoting niche maintenance. Differential gene expression pattern observed in mesangial/vSMC population between tumor and healthy adjacent tissues further supports this notion. Additionally, the communication signature expression correlated with worse overall patient survival and higher disease stage in the TCGA KIRC cohort. On a side note, recent studies have shown stromal cell expansion in recurrent RCC compared to primary disease, furthermore, inhibition of stromal cell-derived Galectin-1 resulted in reduced tumor mass and improved anti-PD-1 ICB efficacy in murine models (365). Another area of active research is combined targeting of PDGFR+ stromal and VEGFR+ endothelial cells, as it was shown to delay tumor vascularization and provide clinical benefit in pancreatic neuroendocrine cancer (332). These observations, together with our limited results, highlight the need for in-depth characterization and further functional validation of tumor-promoting features of stromal cells in ccRCC. Understanding the role of stromal cells in ccRCC TME could be beneficial for novel, targeted therapy development.

The profiling of ccRCC TME presented in this thesis is not without limitations. Single-cell RNA sequencing studies, in general, suffer from data sparsity and tissue dissociation bias. For example, adhesive cells, such as epithelial or tumor, are more challenging to dissociate into a single cell

suspension needed for encapsulation, as opposed to infiltrating immune cells (324). The high degree of immune infiltration is a commonly accepted characteristic of ccRCC, however, the exact cellular composition of tumor specimens in our and other single-cell profiling efforts (231,244,362) is likely to be affected by various experimental variables, including dissociation and/or enrichment protocol used. Hence, the degree of immune infiltration is likely to be inflated. We aimed to minimize these biases by reducing the sample handling time and avoiding cell selection (i.e. FACS), as it is known to damage fragile cells. Finally, the sparsity of obtained data did not permit the use of sophisticated algorithms such as pseudotime or RNA velocity, which could provide further insights into the ccRCC tumor microenvironment. Nonetheless, despite the existing limitations, our study heavily complements ccRCC characterization at the single cell level. For instance, focusing on under-characterized tumor vasculature and stromal cell populations, we introduce and describe a tumor-associated endothelial tip cell phenotype, not previously detected in the context of ccRCC. Additionally, summarizing various bioinformatic analyses performed, we propose that tumor endothelial cells favor tumor progression via expression of metastasis promoting factors, specific ECM constituents and targetable interactions with immune cells in the TME. Undoubtedly, future functional studies will be needed to validate our findings and elucidate the exact roles of the described diverse tumor vasculature and stromal cell phenotypes. Other genomic, epigenomic, *in vitro* and *in vivo* assays (i.e. animal models) would greatly clarify the role of aforementioned populations, especially the tip cells, in disease progression, response to treatment or as potential therapeutic targets in ccRCC.

The patient-enriched populations observed in both lung carcinoma and ccRCC datasets highlight the immense potential of single-cell technologies for search of patient-stratifying biomarkers, fueling the development of personalized or precision medicine applications. To this day, information on the genetic landscape of tumors has been extensively used in therapeutics, with a multitude of treatments developed targeting pathways affected by particular tumor mutations. This approach has seen great success and had a substantial impact on cancer management worldwide. Taking it one step further, the knowledge accumulated by the swift adoption of single-cell profiling technologies in cancer research invites to revisit the very definition of a biomarker. Accumulating research highlights the impact of non-tumor cells, such as immune, endothelial, and stromal cell populations residing in the TME in modulating disease progression and response to therapy. Therefore, specific tumor-associated phenotypes and their abundance are increasingly being recognized as potential targets of therapy or a means to

evaluate disease progression, prognosis or therapy response, underlining the utility of scRNA-seq investigations.

The final part of this thesis described the generation of the first comprehensive transcriptional atlas of uncultured human amniotic fluid cells. For that, inDrops2-TS platform was applied to profile fresh AF amniocentesis samples from consenting donors, excluding cell enrichment of any kind. Currently, there is a single, recent report by Gerli et al. (287) where scRNA-seq was applied to profile uncultured AF cells from 12 amniocentesis samples of 15-34 week gestational age, yet the main focus of the study was derivation and characterization of epithelial organoids, formed by subjection of AF cells to sophisticated lineage-specific culture conditions. In this study, the atlas contained mostly epithelial cells, mirroring our findings, and several immune cell populations, such as macrophages, neutrophils, monocytes, B cells, T cells, NK cells and erythroblasts. Interestingly, the authors did not detect innate lymphoid cells, known to dominate the lymphoid compartment of AF (273). Moreover, a rather large population of neutrophils was found, even though it is widely accepted that recovery of notoriously fragile granulocytes is severely limited in droplet-based scRNA-seq protocols due to the presence of intrinsic nucleases and low amount of RNA (366). Furthermore, the authors omitted standard scRNA-seq analysis steps such as clustering, DGE and annotation, therefore the cell type labels seem to be simply superimposed onto the plot and are absent in the processed data deposited in NCBI GEO repository (GSE220994), complicating any meaningful comparison to our results. Hence, our atlas, albeit being not the first effort to profile uncultured AF cells at the single cell level, represents the first comprehensive description of the cell populations within human AF.

For a more granular view of the phenotypes in the AF, immune and non-immune cell fractions were analyzed separately. The immune compartment was dominated by type 3 innate lymphoid cells, as previously reported utilizing flow cytometry (273). Nonetheless, we uncovered transcriptional diversity of ILCs, reporting progenitor, cycling and two populations differing in the expression of T cell-related genes (**Figure 3.23, C**). Unexpectedly, ILC progenitors (*SCN1B*, *HPN*, *KIT*) expressed both natural cytotoxicity receptors, even though NCR2 expression was reported to be absent in ILCPs from the fetal intestine (306), fetal liver, adult blood and tonsils (367). Another interesting finding was the highly specific expression of hair keratins *KRT86* and *KRT81* in the ILCP cluster. Upon interactive exploration of published fetal tissue atlases, *KRT86* and *KRT81* expression was observed in ILCs from the fetal skin (353), lung (307), and intestine (306),

yet their role in ILC or ILCP function or development remains elusive. The presence of mucosal origin fetal T cells in human AF has been reported before, with the population consisting mostly of CD4 regulatory T cells that suppress T cell activity and responses against maternal antigens (275). In our dataset, however, regulatory (*FOXP3+)* T cells were not present at all – instead, we detected a population of likely central memory (*CCL5*, *CD27*, *CCR7*, *CD8B*) CD8 T cells, contradicting previous reports. With regards to myeloid cells, we uncovered several macrophage phenotypes, namely monocytic gene-enriched, *LYVE1* positive and MMP-high (also *TREM2* positive) M2-polarized macrophages. *LYVE1* macrophages are a known tissue-resident population that was reported to be closely associated with blood vessels in various tissues (i.e., heart, lung) (351). They have also been described in the fetal intestine (306). Interestingly, a recent study utilizing single-cell profiling of fetal skin revealed that both *LYVE1* and *TREM2* macrophages support skin angiogenesis, and the findings were further validated by introducing autologous macrophages to a skin organoid system, recapitulating key skin development aspects (353). Even though these macrophages were reported to reside in dermis, considering the abundant skin shedding into the AF, it is possible to envision that the likely tissue of origin for the observed macrophages could be both the skin and the intestine. Overall, the analysis of the immune cells in AF established the presence of vast, previously unknown phenotypic diversity. Surely, our analysis is limited to transcriptomic profiling and functional assessment is lacking. For instance, it remains unknown if the cells escaped to AF can migrate back toward the fetus and exert their functions. Nonetheless, it is astounding (personally) that sampling of the AF can provide a non-invasive means to glimpse into the immune system development.

With regards to non-immune cells in AF, it was unexpected to find that even without live cell enrichment (in contrast to the sole uncultured AF atlas published to date (287)), we seem to have been able to capture predominantly live cells, at least according to what can be inferred from their transcriptomic features. Generally, adherent cells detached from tissues undergo anoikis – a form of programmed cell death, initiated due to loss of contact to neighboring cells and ECM. When profiling single cells from fresh tissue samples, commonly used QC criteria for dead cell exclusion is low UMI count and high mitochondrial gene count percentage, indicating leaked or degraded RNA, as well as low number of genes expressed. That was not the case in our dataset, as QC criteria were sufficient – cells appeared to be predominantly alive at the time of encapsulation (**Supplementary Figure S7**). One plausible scenario could be that the detached cells or cellular debris floating in AF for a

prolonged amount of time have completely lost their mRNA, as the ambient RNA profile in our samples appeared to be low.

In this work, the presence of highly specialized tissue cells in AF was established, including podocytes, enterocytes, goblet and enteroendocrine cells, ciliated cells and umbrella cells of the urothelium (**Figure 3.24**). The tissues of their origin (gastrointestinal tract, lungs and kidneys) are in direct contact with the fluid, and the cells likely exit via the major AF circulation routes. Studies from the 80's already investigated the presence of such cells in AF (280–282), however, it was limited to microscopy-based identification, and their transcriptomic analysis presented here is, to our knowledge, unprecedented.

Regarding AFSCs, their presence could not be established in the uncultured fluid, as evaluated by conventional stem and mesenchymal marker gene co-expression. A valid concern is the reported rarity of this broadly multipotent phenotype. Prusa et al. demonstrated that only ~0.1-0.5% of cells in amniocentesis samples express the pluripotency marker Oct-4, stem cell factor CD117 (c-kit) and mesenchymal marker vimentin (286). Moreover, such cells were detected not in all samples analyzed. In our atlas, the smallest observed populations were only 38 cells in size (basophils and B cells), comprising only 0.076% of all cells profiled. Additionally, the sample size in our dataset is 26 donors. Thus, it is unlikely, even considering the rarity, that such population would not be captured solely due to random sampling. Additionally, Ryan et al. showed that most studies on fetal MSCs, including AFSCs, report Oct-4 mRNA or protein expression due to faulty primer design and non-specificity of antibodies used, as *POU5F1* gene has multiple transcript isoforms, only one being the functional protein Oct-4A (301). Moreover, a pseudogene generates a highly (96%) homologous, nucleus-localizing protein that is indistinguishable by commercial antibodies (301). Furthermore, Vlahova et al. reported that AFSCs, including the purified c-kit+ AFSCs, do not express the "true" Oct-4 (302). Additionally, the uncultured fluid atlas (287) with de Coppi as a senior author, who reported the c-kit+ AFSCs back in 2007 (295), omits any comment on the absence of AFSCs in their atlas. Therefore, the methodology issues discussed above, together with our results, indicate that the native AF environment does not harbor the broadly multipotent mesenchymal phenotype seen in culture.

The absence of AFSCs does not preclude the possibility of other stem or progenitor cells residing in the AF niche. Recently, Babosova et al. (288) and Gerli et al. (287) reported formation of tissue-specific epithelial organoids from AF samples. In the latter study, as mentioned above, the presence of epithelial progenitors was first attempted to be established in the uncultured

fluid single-cell atlas. Among the epithelium cluster, shown to be enriched for 'Cornified Envelope' Gene Ontology term, the authors found scattered and minute populations of gastrointestinal, kidney and lung progenitors. However, these identities were assigned on the basis of non-zero *average* expression of progenitor marker gene sets (used in **Figure 3.30**), meaning that co-expression of these markers was not established and the obtained average values could be driven solely by a few genes in the list. Meanwhile, Babosova et al. did not investigate uncultured cells. Thus, the presence of epithelial progenitors remains to be confirmed by further studies. In our atlas, a population expressing plasticity-associated *SOX4* and *PLCG2* (309), as well as a distal lung tip progenitor cell marker *ID2* (368) was annotated as lung progenitors, even though the other lung progenitor markers (i.e. *ETV4*, *ETV5*, *NKX2-1*) were not expressed. Therefore, either these cells exhibit a non-canonical progenitor transcriptional state, or are not truly lung progenitors. Further analysis will be needed to refine their identity.

An especially puzzling finding was the three intermediate phenotypes observed, namely, *KRT5*-high fibroblasts, epithelial-mesenchymal and *MMP7*-high fibroblast-epithelial cells. It is important to emphasize that since no fetal atlases that we assessed seem to contain analogous populations, the assigned labels are preliminary and will require further refinement. One of these puzzling cell clusters, *MMP7*-high fibroblast-epithelial cells, showcased high expression of *MMP7*, mesenchymal markers *VIM*, *CD44* and transgelins, simultaneously with canonical epithelial markers, such as *EPCAM*, *CDH1*, *KRT7* and others (**Figure 3.25**). At the same time, a fraction of cells in this population were positive for the development-related transcription factor expression, namely, *PAX2*, *PAX8*, *EMX2*, *LHX1*, *GATA3* and *POU3F3* from the kidney progenitor signature (**Figure 3.30**). There is evidence in the literature that vimentin and CD44 is expressed in adult injured, de-differentiated kidney epithelial cells (369), additionally, these markers are generally associated with EMT. Therefore, it is likely that the MMP7-high population is actually of epithelial phenotype, with mesenchymal marker expression designating a particular de-differentiated, or progenitor, state. Additionally, considering kidney-associated marker expression, a likely source for this cell phenotype in AF is the developing fetal kidneys.

Taken together, these findings shed light on the cellular composition and transcriptional landscape of human AF. Nonetheless, the atlas is far from complete – at the moment, we are collecting more samples, with a focus on trisomy 21 affected fetuses. The bioinformatic analysis presented in this work is descriptive and limited in depth, and other techniques, such as pseudotime, RNA velocity and cell-cell interaction inference will be performed once the

dataset is complete. Additionally, to disentangle the AFSC identity, we are performing AF culture and single-cell profiling experiments, with the hopes to identify the native cell-of-origin for the cultures, and characterize AFSCs further. It is an exciting prospect, with the potential to yield insight into fundamental cellular properties, such as plasticity, extending beyond the specific biological niche studied.

CONCLUSIONS

1. inDrops-2-TS and inDrops-2-IVT methods have comparable gene and transcript capture characteristics
2. inDrops-2-TS method enables the recovery of rare phenotypes
3. Single-cell profiling of kidney samples by inDrops-2 revealed all major structural cell types, including distinct epithelial cells of nephron segments, whereas ccRCC was enriched in tumor-infiltrating immune cells and tumor-associated endothelium
4. In ccRCC TME, T cells present exhausted phenotypes, tumor-associated macrophages display pro-inflammatory and immunosuppressive phenotypes, whereas endothelial cells include a novel tip cell-like population. Together, these cells likely contribute to maintenance of tumor-supportive TME
5. Immune cells in amniotic fluid (AF) mostly consist of innate lymphoid cells and macrophages
6. Non-immune cells in AF include highly specialized organ-specific cells shed from the fetal skin, kidney, intestine and lung, and transitory epithelial-mesenchymal phenotypes
7. Cells harboring cultured amniotic fluid stem cell transcriptional profile are not detected in uncultured AF

REFERENCES

1. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009 May;6(5):377–82. doi:10.1038/nmeth.1315

2. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Cell. 2015 May 21;161(5):1187–201. doi:10.1016/j.cell.2015.04.044 PubMed PMID: 26000487.

3. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015 May 21;161(5):1202–14. doi:10.1016/j.cell.2015.05.002 PubMed PMID: 26000488.

4. Plasschaert LW, Žilionis R, Choo-Wing R, Savova V, Knehr J, Roma G, et al. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. Nature. 2018 Aug;560(7718):377–81. doi:10.1038/s41586-018-0394-6

5. Villani AC, Satija R, Reynolds G, Sarkizova S, Shekhar K, Fletcher J, et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science. 2017 Apr 21;356(6335):eaah4573. doi:10.1126/science.aah4573

6. Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, Prabhakaran S, et al. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. Cell. 2018 Aug;174(5):1293-1308.e36. doi:10.1016/j.cell.2018.05.060

7. Laughney AM, Hu J, Campbell NR, Bakhoum SF, Setty M, Lavallée VP, et al. Regenerative lineages and immune-mediated pruning in lung cancer metastasis. Nat Med. 2020 Feb;26(2):2. doi:10.1038/s41591-019-0750-6

8. Zilionis R, Engblom C, Pfirschke C, Savova V, Zemmour D, Saatcioglu HD, et al. Single-Cell Transcriptomics of Human and Mouse Lung Cancers Reveals Conserved Myeloid Populations across Individuals and Species. Immunity. 2019 May 21;50(5):1317-1334.e10. doi:10.1016/j.immuni.2019.03.009 PubMed PMID: 30979687.

9. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The Human Cell Atlas. eLife. 2017 Dec 5;6:e27041. doi:10.7554/eLife.27041

10. Rood JE, Wynne S, Robson L, Hupalowska A, Randell J, Teichmann SA, et al. The Human Cell Atlas from a cell census to a unified foundation model. Nature. 2025 Jan;637(8048):1065–71. doi:10.1038/s41586-024-08338-4

11. Boon WC, Petkovic-Duran K, Zhu Y, Manasseh R, Horne MK, Aumann TD. Increasing cDNA yields from single-cell quantities of mRNA in standard laboratory reverse transcriptase reactions using acoustic microstreaming. J Vis Exp JoVE. 2011 Jul 11;(53):e3144. doi:10.3791/3144 PubMed PMID: 21775961; PubMed Central PMCID: PMC3346307.

12. Eberwine J, Yeh H, Miyashiro K, Cao Y, Nair S, Finnell R, et al. Analysis of gene expression in single live neurons. Proc Natl Acad Sci. 1992 Apr;89(7):3010–4. doi:10.1073/pnas.89.7.3010

13. Tietjen I, Rihel JM, Cao Y, Koentges G, Zakhary L, Dulac C. Single-Cell Transcriptional Analysis of Neuronal Progenitors. Neuron. 2003 Apr 24;38(2):161–75. doi:10.1016/S0896-6273(03)00229-0 PubMed PMID: 12718852.

14. Kurimoto K, Yabuta Y, Ohinata Y, Ono Y, Uno KD, Yamada RG, et al. An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. Nucleic Acids Res. 2006 Mar 1;34(5):e42. doi:10.1093/nar/gkl050

15. Islam S, Kjällquist U, Moliner A, Zajac P, Fan JB, Lönnerberg P, et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. 2011 Jul 1;21(7):1160–7. doi:10.1101/gr.110882.110

16. Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods. 2013 Nov;10(11):1093–5. doi:10.1038/nmeth.2645

17. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, et al. Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. Science. 2014 Feb 14;343(6172):776–9. doi:10.1126/science.1247651

18. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017 Jan 16;8(1):14049. doi:10.1038/ncomms14049

19. Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, et al. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. Science. 2015 May 22;348(6237):910–4. doi:10.1126/science.aab1601

20. Rosenberg AB, Roco CM, Muscat RA, Kuchina A, Sample P, Yao Z, et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science. 2018 Apr 13;360(6385):176–82. doi:10.1126/science.aam8999

21. Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science. 2017 Aug 18;357(6352):661–7. doi:10.1126/science.aam8940

22. Elz AE, Gratz D, Long A, Sowerby D, Hadadianpour A, Newell EW. Evaluating the practical aspects and performance of commercial single-cell RNA sequencing technologies. bioRxiv. 2025 May 24;2025.05.19.654974. doi:10.1101/2025.05.19.654974

23. Vitak SA, Torkenczy KA, Rosenkrantz JL, Fields AJ, Christiansen L, Wong MH, et al. Sequencing thousands of single-cell genomes with combinatorial indexing. Nat Methods. 2017 Mar;14(3):302–8. doi:10.1038/nmeth.4154

24. Laks E, McPherson A, Zahn H, Lai D, Steif A, Brimhall J, et al. Clonal Decomposition and DNA Replication States Defined by Scaled Single-Cell Genome Sequencing. Cell. 2019 Nov 14;179(5):1207-1221.e22. doi:10.1016/j.cell.2019.10.026 PubMed PMID: 31730858.

25. Mulqueen RM, Pokholok D, Norberg SJ, Torkenczy KA, Fields AJ, Sun D, et al. Highly scalable generation of DNA methylation profiles in single cells. Nat Biotechnol. 2018 May;36(5):428–31. doi:10.1038/nbt.4112

26. Bartosovic M, Kabbe M, Castelo-Branco G. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. Nat Biotechnol. 2021 Jul;39(7):825–35. doi:10.1038/s41587-021-00869-9

27. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015 Jul;523(7561):486–90. doi:10.1038/nature14590

28. Tan L, Xing D, Chang CH, Li H, Xie XS. Three-dimensional genome structures of single diploid human cells. Science. 2018 Aug 31;361(6405):924–8. doi:10.1126/science.aat5641

29. Budnik B, Levy E, Harmange G, Slavov N. SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. Genome Biol. 2018 Oct 22;19(1):161. doi:10.1186/s13059-018-1547-5

30. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, et al. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods. 2017 Sep;14(9):865–8. doi:10.1038/nmeth.4380

31. Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, et al. Multiplexed quantification of proteins and transcripts in single cells. Nat Biotechnol. 2017 Oct;35(10):936–9. doi:10.1038/nbt.3973

32. Ma S, Zhang B, LaFave LM, Earl AS, Chiang Z, Hu Y, et al. Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. Cell. 2020 lapkričio;183(4):1103-1116.e20. doi:10.1016/j.cell.2020.09.056

33. Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nat Methods. 2016 Mar;13(3):229–32. doi:10.1038/nmeth.3728

34. Zhu C, Zhang Y, Li YE, Lucero J, Behrens MM, Ren B. Joint profiling of histone modifications and transcriptome in single cells from mouse brain. Nat Methods. 2021 Mar;18(3):283–92. doi:10.1038/s41592-021-01060-3

35. Tu AA, Gierahn TM, Monian B, Morgan DM, Mehta NK, Ruiter B, et al. TCR sequencing paired with massively parallel 3′ RNA-seq reveals clonotypic T cell signatures. Nat Immunol. 2019 Dec;20(12):1692–9. doi:10.1038/s41590-019-0544-5

36. Baysoy A, Bai Z, Satija R, Fan R. The technological landscape and applications of single-cell multi-omics. Nat Rev Mol Cell Biol. 2023 Oct;24(10):695–713. doi:10.1038/s41580-023-00615-w

37. Clark SJ, Argelaguet R, Kapourani CA, Stubbs TM, Lee HJ, Alda-Catalinas C, et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. Nat Commun. 2018 Feb 22;9(1):781. doi:10.1038/s41467-018-03149-4

38. Wang Y, Yuan P, Yan Z, Yang M, Huo Y, Nie Y, et al. Single-cell multiomics sequencing reveals the functional regulatory landscape of early embryos. Nat Commun. 2021 Feb 23;12(1):1247. doi:10.1038/s41467-021-21409-8

39. Hou Y, Guo H, Cao C, Li X, Hu B, Zhu P, et al. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. Cell Res. 2016 Mar;26(3):304–19. doi:10.1038/cr.2016.23

40. Mimitou EP, Cheng A, Montalbano A, Hao S, Stoeckius M, Legut M, et al. Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. Nat Methods. 2019 May;16(5):409–12. doi:10.1038/s41592-019-0392-0

41. Vandereyken K, Sifrim A, Thienpont B, Voet T. Methods and applications for single-cell and spatial multi-omics. Nat Rev Genet. 2023 Aug;24(8):494–515. doi:10.1038/s41576-023-00580-2

42. Home page [Internet]. [cited 2025 Jun 23]. Available from: https://www.humancellatlas.org/

43. HCA Data Portal [Internet]. [cited 2025 Sep 13]. Available from: https://data.humancellatlas.org/

44. Amit I, Ardlie K, Arzuaga F, Awandare G, Bader G, Bernier A, et al. The commitment of the human cell atlas to humanity. Nat Commun. 2024 Nov 20;15(1):10019. doi:10.1038/s41467-024-54306-x

45. Lafzi A, Moutinho C, Picelli S, Heyn H. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. Nat Protoc. 2018 Dec;13(12):2742–57. doi:10.1038/s41596-018-0073-y

46. Sant P, Rippe K, Mallm JP. Approaches for single-cell RNA sequencing across tissues and cell types. Transcription. 2023 Oct 20;14(3–5):127–45. doi:10.1080/21541264.2023.2200721 PubMed PMID: 37062951.

47. Oh JM, An M, Son DS, Choi J, Cho YB, Yoo CE, et al. Comparison of cell type distribution between single-cell and single-nucleus RNA sequencing: enrichment of adherent cell types in single-nucleus RNA sequencing. Exp Mol Med. 2022 Dec;54(12):2128–34. doi:10.1038/s12276-022-00892-z

48. Denisenko E, Guo BB, Jones M, Hou R, de Kock L, Lassmann T, et al. Systematic assessment of tissue dissociation and storage biases in

single-cell and single-nucleus RNA-seq workflows [Internet]. 2019 Nov 6. doi:10.1101/832444

49. O'Flanagan CH, Campbell KR, Zhang AW, Kabeer F, Lim JLP, Biele J, 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 Oct 17;20(1):210. doi:10.1186/s13059-019-1830-0

50. Van Den Brink SC, Sage F, Vértesy Á, Spanjaard B, Peterson-Maduro J, Baron CS, et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. Nat Methods. 2017 Oct;14(10):935–6. doi:10.1038/nmeth.4437

51. Lake BB, Chen S, Hoshi M, Plongthongkum N, Salamon D, Knoten A, et al. A single-nucleus RNA-sequencing pipeline to decipher the molecular anatomy and pathophysiology of human kidneys. Nat Commun. 2019 Dec;10(1):2832. doi:10.1038/s41467-019-10861-2

52. Wu H, Kirita Y, Donnelly EL, Humphreys BD. Advantages of Single-Nucleus over Single-Cell RNA Sequencing of Adult Kidney: Rare Cell Types and Novel Cell States Revealed in Fibrosis. J Am Soc Nephrol. 2019 Jan;30(1):23–32. doi:10.1681/ASN.2018090912

53. Slyper M, Porter CBM, Ashenberg O, Waldman J, Drokhlyansky E, Wakiro I, et al. A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. Nat Med. 2020 May;26(5):792–802. doi:10.1038/s41591-020-0844-1

54. Gupta A, Shamsi F, Altemose N, Dorlhiac GF, Cypess AM, White AP, et al. Characterization of transcript enrichment and detection bias in single-nucleus RNA-seq for mapping of distinct human adipocyte lineages. Genome Res. 2022 Feb 1;32(2):242–57. doi:10.1101/gr.275509.121

55. Guillaumet-Adkins A, Rodríguez-Esteban G, Mereu E, Mendez-Lago M, Jaitin DA, Villanueva A, et al. Single-cell transcriptome conservation in cryopreserved cells and tissues. Genome Biol. 2017 Mar 1;18(1):45. doi:10.1186/s13059-017-1171-9

56. Wohnhaas CT, Leparc GG, Fernandez-Albert F, Kind D, Gantner F, Viollet C, et al. DMSO cryopreservation is the method of choice to preserve cells for droplet-based single-cell RNA sequencing. Sci Rep. 2019 Jul 23;9(1):10699. doi:10.1038/s41598-019-46932-z

57. Alles J, Karaiskos N, Praktiknjo SD, Grosswendt S, Wahle P, Ruffault PL, et al. Cell fixation and preservation for droplet-based single-cell transcriptomics. BMC Biol. 2017 May 19;15(1):44. doi:10.1186/s12915-017-0383-5 PubMed PMID: 28526029; PubMed Central PMCID: PMC5438562.

58. Chen J, Cheung F, Shi R, Zhou H, Lu W, Candia J, et al. PBMC fixation and processing for Chromium single-cell RNA sequencing. J Transl Med. 2018 Jul 17;16(1):198. doi:10.1186/s12967-018-1578-4

59. Jiménez-Gracia L, Marchese D, Nieto JC, Caratù G, Melón-Ardanaz E, Gudiño V, et al. FixNCut: single-cell genomics through reversible tissue fixation and dissociation. Genome Biol. 2024 Mar 29;25(1):81. doi:10.1186/s13059-024-03219-5

60. Juzenas S, Goda K, Kiseliovas V, Zvirblyte J, Quintinal-Villalonga A, Siurkus J, et al. inDrops-2: a flexible, versatile and cost-efficient droplet microfluidic approach for high-throughput scRNA-seq of fresh and preserved clinical samples. Nucleic Acids Res. 2025 Jan 11;53(2):gkae1312. doi:10.1093/nar/gkae1312

61. Mazutis L, Gilbert J, Ung WL, Weitz DA, Griffiths AD, Heyman JA. Single-cell analysis and sorting using droplet-based microfluidics. Nat Protoc. 2013 May;8(5):870–91. doi:10.1038/nprot.2013.046

62. Zilionis R, Nainys J, Veres A, Savova V, Zemmour D, Klein AM, et al. Single-cell barcoding and sequencing using droplet microfluidics. Nat Protoc. 2017 Jan;12(1):44–73. doi:10.1038/nprot.2016.154

63. Pranauskaite E, Milkus V, Ritmejeris J, Zilionis R, Mazutis L. Increasing Fluid Viscosity Ensures Consistent Single-Cell Encapsulation. Anal Chem. 2024 May 7;96(18):6898–905. doi:10.1021/acs.analchem.3c05243

64. 10x Genomics [Internet]. [cited 2025 Jun 27]. Flex Gene Expression. Available from: https://www.10xgenomics.com/products/flex-gene-expression

65. Dal Molin A, Di Camillo B. How to design a single-cell RNA-sequencing experiment: pitfalls, challenges and perspectives. Brief Bioinform. 2019 Jul 19;20(4):1384–94. doi:10.1093/bib/bby007

66. Lebrigand K, Magnone V, Barbry P, Waldmann R. High throughput error corrected Nanopore single cell transcriptome sequencing. Nat Commun. 2020 Aug 12;11(1):4025. doi:10.1038/s41467-020-17800-6

67. Volden R, Vollmers C. Single-cell isoform analysis in human immune cells. Genome Biol. 2022 Feb 7;23(1):47. doi:10.1186/s13059-022-02615-z

68. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019 Jun;15(6):MSB188746. doi:10.15252/msb.20188746

69. Heumos L, Schaar AC, Lance C, Litinetskaya A, Drost F, Zappia L, et al. Best practices for single-cell analysis across modalities. Nat Rev Genet. 2023 Aug;24(8):550–72. doi:10.1038/s41576-023-00586-w

70. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods. 2015 Feb;12(2):115–21. doi:10.1038/nmeth.3252

71. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021 Jun 24;184(13):3573-3587.e29. doi:10.1016/j.cell.2021.04.048 PubMed PMID: 34062119.

72. scverse [Internet]. [cited 2025 Jul 2]. scverse. Available from: https://scverse.org/

73. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018 Dec;19(1):15. doi:10.1186/s13059-017-1382-0

74. Zhu X, Wolfgruber TK, Tasato A, Arisdakessian C, Garmire DG, Garmire LX. Granatum: a graphical single-cell RNA-Seq analysis pipeline for genomics scientists. Genome Med. 2017 Dec 5;9(1):108. doi:10.1186/s13073-017-0492-3

75. Gardeux V, David FPA, Shajkofci A, Schwalie PC, Deplancke B. ASAP: a web-based platform for the analysis and interactive visualization of single-cell RNA-seq data. Bioinforma Oxf Engl. 2017 Oct 1;33(19):3123–5. doi:10.1093/bioinformatics/btx337 PubMed PMID: 28541377; PubMed Central PMCID: PMC5870842.

76. Moreno P, Huang N, Manning JR, Mohammed S, Solovyev A, Polanski K, et al. User-friendly, scalable tools and workflows for single-cell RNA-seq analysis. Nat Methods. 2021 Apr;18(4):327–8. doi:10.1038/s41592-021-01102-w

77. Luecken MD, Gigante S, Burkhardt DB, Cannoodt R, Strobl DC, Markov NS, et al. Defining and benchmarking open problems in single-cell analysis. Nat Biotechnol. 2025 Jul 1;1–6. doi:10.1038/s41587-025-02694-w

78. Cao Y, Yu L, Torkel M, Kim S, Lin Y, Yang P, et al. The current landscape and emerging challenges of benchmarking single-cell methods. Brief Bioinform. 2025 Sep 1;26(5):bbaf380. doi:10.1093/bib/bbaf380

79. Kaminow B, Yunusov D, Dobin A. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data [preprint] [Internet]. Bioinformatics; 2021 May [cited 2021 Sep 6]. Report. Available from: http://biorxiv.org/lookup/doi/10.1101/2021.05.05.442755 doi:10.1101/2021.05.05.442755

80. Srivastava A, Malik L, Smith T, Sudbery I, Patro R. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. Genome Biol. 2019 Mar 27;20(1):65. doi:10.1186/s13059-019-1670-y

81. Melsted P, Booeshaghi AS, Liu L, Gao F, Lu L, Min KH (Joseph), et al. Modular, efficient and constant-memory single-cell RNA-seq preprocessing. Nat Biotechnol. 2021 Jul;39(7):813–8. doi:10.1038/s41587-021-00870-2

82. Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs. GigaScience. 2018 Jun 1;7(6):giy059. doi:10.1093/gigascience/giy059

83. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data [Internet]. [cited 2025 Jul 2]. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

84. Ziegenhain C, Vieth B, Parekh S, Hellmann I, Enard W. Quantitative single-cell transcriptomics. Brief Funct Genomics. 2018 Jul 1;17(4):220–32. doi:10.1093/bfgp/ely009

85. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016 Jan 26;17(1):13. doi:10.1186/s13059-016-0881-8

86. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan;29(1):15–21. doi:10.1093/bioinformatics/bts635

87. Guigó R. Genome annotation: From human genetics to biodiversity genomics. Cell Genomics. 2023 Aug 9;3(8):100375. doi:10.1016/j.xgen.2023.100375 PubMed PMID: 37601977; PubMed Central PMCID: PMC10435374.

88. Mudge JM, Carbonell-Sala S, Diekhans M, Martinez JG, Hunt T, Jungreis I, et al. GENCODE 2025: reference gene annotation for human and mouse. Nucleic Acids Res. 2025 Jan 6;53(D1):D966–75. doi:10.1093/nar/gkae1078

89. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016 Jan 4;44(D1):D733–45. doi:10.1093/nar/gkv1189

90. Almeida da Paz M, Warger S, Taher L. Disregarding multimappers leads to biases in the functional assessment of NGS data. BMC Genomics. 2024 May 8;25(1):455. doi:10.1186/s12864-024-10344-9

91. Pool AH, Poldsam H, Chen S, Thomson M, Oka Y. Recovery of missing single-cell RNA-sequencing data with optimized transcriptomic references. Nat Methods. 2023 Oct;20(10):1506–15. doi:10.1038/s41592-023-02003-w

92. Young MD, Behjati S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. GigaScience. 2020 Nov 30;9(12):giaa151. doi:10.1093/gigascience/giaa151

93. Fleming SJ, Chaffin MD, Arduini A, Akkad AD, Banks E, Marioni JC, et al. Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. Nat Methods. 2023 Sep;20(9):1323–35. doi:10.1038/s41592-023-01943-7

94. Yang S, Corbett SE, Koga Y, Wang Z, Johnson WE, Yajima M, et al. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. Genome Biol. 2020 Mar 5;21(1):57. doi:10.1186/s13059-020-1950-6

95. Janssen P, Kliesmete Z, Vieth B, Adiconis X, Simmons S, Marshall J, et al. The effect of background noise and its removal on the analysis of single-cell expression data. Genome Biol. 2023 Jun 19;24(1):140. doi:10.1186/s13059-023-02978-x

96. Caglayan E, Liu Y, Konopka G. Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets. Neuron. 2022 Dec 21;110(24):4043-4056.e5.

doi:10.1016/j.neuron.2022.09.010 PubMed PMID: 36240767; PubMed Central PMCID: PMC9789184.

97. Wolock SL, Lopez R, Klein AM. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Syst. 2019 Apr;8(4):281-291.e9. doi:10.1016/j.cels.2018.11.005

98. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Syst. 2019 Apr 24;8(4):329-337.e4. doi:10.1016/j.cels.2019.03.003 PubMed PMID: 30954475.

99. Xi NM, Li JJ. Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data. Cell Syst. 2021 Feb 17;12(2):176-194.e6. doi:10.1016/j.cels.2020.11.008 PubMed PMID: 33338399.

100. Qiu P. Embracing the dropouts in single-cell RNA-seq analysis. Nat Commun. 2020 Dec;11(1):1169. doi:10.1038/s41467-020-14976-9

101. Vieth B, Parekh S, Ziegenhain C, Enard W, Hellmann I. A systematic evaluation of single cell RNA-seq analysis pipelines. Nat Commun. 2019 Dec;10(1):4667. doi:10.1038/s41467-019-12266-7

102. Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC. Normalizing single-cell RNA sequencing data: challenges and opportunities. Nat Methods. 2017 Jun;14(6):565–71. doi:10.1038/nmeth.4292

103. L. Lun AT, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. 2016 Apr 27;17(1):75. doi:10.1186/s13059-016-0947-7

104. Lause J, Berens P, Kobak D. Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. Genome Biol. 2021 Sep 6;22(1):258. doi:10.1186/s13059-021-02451-7

105. Ahlmann-Eltze C, Huber W. Comparison of transformations for single-cell RNA-seq data. Nat Methods. 2023 May;20(5):665–72. doi:10.1038/s41592-023-01814-1

106. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007 sausio;8(1):118–27. doi:10.1093/biostatistics/kxj037

107. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and

microarray studies. Nucleic Acids Res. 2015 Apr 20;43(7):e47. doi:10.1093/nar/gkv007

108. Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, et al. Benchmarking atlas-level data integration in single-cell genomics. Nat Methods. 2022 Jan;19(1):1. doi:10.1038/s41592-021-01336-8

109. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol. 2018 May;36(5):5. doi:10.1038/nbt.4091

110. Polański K, Young MD, Miao Z, Meyer KB, Teichmann SA, Park JE. BBKNN: fast batch alignment of single cell transcriptomes. Bioinformatics. 2020 Feb 1;36(3):964–5. doi:10.1093/bioinformatics/btz625

111. Hie B, Bryson B, Berger B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat Biotechnol. 2019 Jun;37(6):6. doi:10.1038/s41587-019-0113-3

112. Xu C, Lopez R, Mehlman E, Regier J, Jordan MI, Yosef N. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. Mol Syst Biol. 2021 Jan;17(1):e9620. doi:10.15252/msb.20209620

113. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018 Dec;15(12):1053–8. doi:10.1038/s41592-018-0229-2

114. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018 May;36(5):411–20. doi:10.1038/nbt.4096

115. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019 Dec;16(12):1289–96. doi:10.1038/s41592-019-0619-0

116. Moon KR, Stanley JS, Burkhardt D, van Dijk D, Wolf G, Krishnaswamy S. Manifold learning-based methods for analyzing single-cell RNA-sequencing data. Curr Opin Syst Biol. 2018 vasario;• Future of systems biology• Genomics and epigenomics7:36–46. doi:10.1016/j.coisb.2017.12.008

117. Wattenberg M, Viégas F, Johnson I. How to Use t-SNE Effectively. Distill. 2016 Oct 13;1(10):e2. doi:10.23915/distill.00002

118. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv180203426 Cs Stat [Internet]. 2020 Sep 17 [cited 2021 Jul 5]. Available from: http://arxiv.org/abs/1802.03426

119. Weinreb C, Wolock S, Klein AM. SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. Bioinformatics. 2018 Apr 1;34(7):1246–8. doi:10.1093/bioinformatics/btx792

120. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2019 Jan;37(1):38–44. doi:10.1038/nbt.4314

121. Haghverdi L, Buettner F, Theis FJ. Diffusion maps for high-dimensional single-cell analysis of differentiation data. Bioinformatics. 2015 Sep 15;31(18):2989–98. doi:10.1093/bioinformatics/btv325

122. Wolf FA, Hamey FK, Plass M, Solana J, Dahlin JS, Göttgens B, et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol. 2019 Mar 19;20(1):59. doi:10.1186/s13059-019-1663-x

123. Moon KR, van Dijk D, Wang Z, Gigante S, Burkhardt DB, Chen WS, et al. Visualizing structure and transitions in high-dimensional biological data. Nat Biotechnol. 2019 Dec;37(12):1482–92. doi:10.1038/s41587-019-0336-3

124. Chari T, Pachter L. The specious art of single-cell genomics. PLOS Comput Biol. 2023 Aug 17;19(8):e1011288. doi:10.1371/journal.pcbi.1011288

125. Megill C, Martin B, Weaver C, Bell S, Prins L, Badajoz S, et al. cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices [Internet]. bioRxiv; 2021 [cited 2025 Jul 2]. p. 2021.04.05.438318. Available from: https://www.biorxiv.org/content/10.1101/2021.04.05.438318v1 doi:10.1101/2021.04.05.438318

126. Ouyang JF, Kamaraj US, Cao EY, Rackham OJL. ShinyCell: simple and sharable visualization of single-cell gene expression data. Bioinforma Oxf Engl. 2021 Oct 11;37(19):3374–6. doi:10.1093/bioinformatics/btab209 PubMed PMID: 33774659.

127. Single Cell Portal [Internet]. [cited 2025 Jul 2]. Available from: https://singlecell.broadinstitute.org/single_cell

128. MacQueen JB. Some methods for classification and analysis of multivariate observations. Multivar Obs.

129. Yu L, Cao Y, Yang JYH, Yang P. Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data. Genome Biol. 2022 Feb 8;23(1):49. doi:10.1186/s13059-022-02622-0

130. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet. 2019 May;20(5):273–82. doi:10.1038/s41576-018-0088-9

131. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008 Oct;2008(10):P10008. doi:10.1088/1742-5468/2008/10/P10008

132. Levine JH, Simonds EF, Bendall SC, Davis KL, Amir E ad D, Tadmor MD, et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell. 2015 Jul 2;162(1):184–97. doi:10.1016/j.cell.2015.05.047 PubMed PMID: 26095251.

133. Duò A, Robinson MD, Soneson C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. F1000Research. 2018;7:1141. doi:10.12688/f1000research.15666.3 PubMed PMID: 30271584; PubMed Central PMCID: PMC6134335.

134. Freytag S, Tian L, Lönnstedt I, Ng M, Bahlo M. Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. F1000Research. 2018;7:1297. doi:10.12688/f1000research.15809.2 PubMed PMID: 30228881; PubMed Central PMCID: PMC6124389.

135. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019 Mar 26;9(1):5233. doi:10.1038/s41598-019-41695-z

136. Clarke ZA, Andrews TS, Atif J, Pouyabahar D, Innes BT, MacParland SA, et al. Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. Nat Protoc. 2021 Jun;16(6):2749–64. doi:10.1038/s41596-021-00534-0

137. Domínguez Conde C, Xu C, Jarvis LB, Rainbow DB, Wells SB, Gomes T, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science. 2022 May 13;376(6594):eabl5197. doi:10.1126/science.abl5197

138. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014 Apr;32(4):381–6. doi:10.1038/nbt.2859

139. Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature. 2019 Feb 28;566(7745):496–502. doi:10.1038/s41586-019-0969-x

140. Haghverdi L, Büttner M, Wolf FA, Buettner F, Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching. Nat Methods. 2016 Oct;13(10):845–8. doi:10.1038/nmeth.3971

141. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics. 2018 Jun 19;19(1):477. doi:10.1186/s12864-018-4772-0

142. Setty M, Kiseliovas V, Levine J, Gayoso A, Mazutis L, Pe'er D. Characterization of cell fate probabilities in single-cell data with Palantir. Nat Biotechnol. 2019 Apr;37(4):451–60. doi:10.1038/s41587-019-0068-4

143. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, et al. RNA velocity of single cells. Nature. 2018 Aug;560(7719):494–8. doi:10.1038/s41586-018-0414-6

144. Bergen V, Lange M, Peidli S, Wolf FA, Theis FJ. Generalizing RNA velocity to transient cell states through dynamical modeling. Nat Biotechnol. 2020 Dec;38(12):1408–14. doi:10.1038/s41587-020-0591-3

145. Bergen V, Soldatov RA, Kharchenko PV, Theis FJ. RNA velocity—current challenges and future perspectives. Mol Syst Biol. 2021 Aug;17(8):e10282. doi:10.15252/msb.202110282

146. Lange M, Bergen V, Klein M, Setty M, Reuter B, Bakhti M, et al. CellRank for directed single-cell fate mapping. Nat Methods. 2022 Feb;19(2):2. doi:10.1038/s41592-021-01346-6

147. Weiler P, Lange M, Klein M, Pe'er D, Theis F. CellRank 2: unified fate mapping in multiview single-cell data. Nat Methods. 2024 Jul;21(7):1196–205. doi:10.1038/s41592-024-02303-9

148. Hou W, Ji Z, Ji H, Hicks SC. A systematic evaluation of single-cell RNA-sequencing imputation methods. Genome Biol. 2020 Aug 27;21(1):218. doi:10.1186/s13059-020-02132-x

149. Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat Commun. 2018 Mar 8;9(1):997. doi:10.1038/s41467-018-03405-7

150. Huang M, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, et al. SAVER: gene expression recovery for single-cell RNA sequencing. Nat Methods. 2018 Jul;15(7):539–42. doi:10.1038/s41592-018-0033-z

151. Linderman GC, Zhao J, Roulis M, Bielecki P, Flavell RA, Nadler B, et al. Zero-preserving imputation of single-cell RNA-seq data. Nat Commun. 2022 Jan 11;13(1):192. doi:10.1038/s41467-021-27729-z

152. van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, et al. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell. 2018 Jul;174(3):716-729.e27. doi:10.1016/j.cell.2018.05.061

153. Wagner F, Yan Y, Yanai I. K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data [Internet]. bioRxiv; 2018 [cited 2025 Jul 16]. p. 217737. Available from: https://www.biorxiv.org/content/10.1101/217737v3 doi:10.1101/217737

154. Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan CH, et al. Inference and analysis of cell-cell communication using CellChat. Nat Commun. 2021 Feb 17;12(1):1. doi:10.1038/s41467-021-21246-9

155. Vento-Tormo R, Efremova M, Botting RA, Turco MY, Vento-Tormo M, Meyer KB, et al. Single-cell reconstruction of the early maternal–fetal interface in humans. Nature. 2018 Nov;563(7731):347–53. doi:10.1038/s41586-018-0698-6

156. Efremova M, Vento-Tormo M, Teichmann SA, Vento-Tormo R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. Nat Protoc. 2020 Apr;15(4):1484–506. doi:10.1038/s41596-020-0292-x

157. Garcia-Alonso L, Handfield LF, Roberts K, Nikolakopoulou K, Fernando RC, Gardner L, et al. Mapping the temporal and spatial dynamics of the human endometrium in vivo and in vitro. Nat Genet. 2021 Dec;53(12):1698–711. doi:10.1038/s41588-021-00972-2

158. Garcia-Alonso L, Lorenzi V, Mazzeo CI, Alves-Lopes JP, Roberts K, Sancho-Serra C, et al. Single-cell roadmap of human gonadal development. Nature. 2022 Jul;607(7919):540–7. doi:10.1038/s41586-022-04918-4

159. Dimitrov D, Türei D, Garrido-Rodriguez M, Burmedi PL, Nagai JS, Boys C, et al. Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. Nat Commun. 2022 Jun 9;13(1):3224. doi:10.1038/s41467-022-30755-0

160. Žvirblytė J, Mažutis L. Microfluidics for Cancer Biomarker Discovery, Research, and Clinical Application. In: Caballero D, Kundu SC, Reis RL, editors. Microfluidics and Biosensors in Cancer Research: Applications in Cancer Modeling and Theranostics. Cham: Springer International Publishing; 2022. p. 499–524. (Advances in Experimental Medicine and Biology).

161. Dann E, Henderson NC, Teichmann SA, Morgan MD, Marioni JC. Differential abundance testing on single-cell data using k-nearest neighbor graphs. Nat Biotechnol. 2021 Sep 30;1–9. doi:10.1038/s41587-021-01033-z

162. Reshef YA, Rumker L, Kang JB, Nathan A, Korsunsky I, Asgari S, et al. Co-varying neighborhood analysis identifies cell populations associated with phenotypes of interest from single-cell transcriptomics. Nat Biotechnol. 2022 Mar;40(3):355–63. doi:10.1038/s41587-021-01066-4

163. Zhao J, Jaffe A, Li H, Lindenbaum O, Sefik E, Jackson R, et al. Detection of differentially abundant cell subpopulations in scRNA-seq data. Proc Natl Acad Sci. 2021 Jun;118(22):e2100293118. doi:10.1073/pnas.2100293118

164. Yi H, Plotkin A, Stanley N. Benchmarking differential abundance methods for finding condition-specific prototypical cells in multi-sample single-cell datasets. Genome Biol. 2024 Jan 3;25(1):9. doi:10.1186/s13059-023-03143-0

165. Nguyen HCT, Baik B, Yoon S, Park T, Nam D. Benchmarking integration of single-cell differential expression. Nat Commun. 2023 Mar 21;14(1):1570. doi:10.1038/s41467-023-37126-3

166. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015 Dec 10;16(1):278. doi:10.1186/s13059-015-0844-5

167. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. doi:10.1186/s13059-014-0550-8 PubMed PMID: 25516281; PubMed Central PMCID: PMC4302049.

168. Squair JW, Gautier M, Kathe C, Anderson MA, James ND, Hutson TH, et al. Confronting false discoveries in single-cell differential expression. Nat Commun. 2021 Sep 28;12(1):5692. doi:10.1038/s41467-021-25960-2

169. Sun W, Liu Z, Jiang X, Chen MB, Dong H, Liu J, et al. Spatial transcriptomics reveal neuron–astrocyte synergy in long-term memory. Nature. 2024 Mar;627(8003):374–81. doi:10.1038/s41586-023-07011-6

170. Mukamel EA, Yu Z. False positives in study of memory-related gene expression. Nature. 2025 Jun;642(8066):E1–3. doi:10.1038/s41586-025-08988-y

171. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005 Oct 25;102(43):15545–50. doi:10.1073/pnas.0506580102

172. Holland CH, Tanevski J, Perales-Patón J, Gleixner J, Kumar MP, Mereu E, et al. Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. Genome Biol. 2020 Feb 12;21(1):36. doi:10.1186/s13059-020-1949-z

173. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000 May;25(1):25–9. doi:10.1038/75556

174. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011 Jun 15;27(12):1739–40. doi:10.1093/bioinformatics/btr260

175. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017 Jan 4;45(D1):D353–61. doi:10.1093/nar/gkw1092 PubMed PMID: 27899662; PubMed Central PMCID: PMC5210567.

176. Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, et al. The reactome pathway knowledgebase 2022. Nucleic Acids Res. 2022 Jan 7;50(D1):D687–92. doi:10.1093/nar/gkab1028

177. Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J. Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. 2019 Aug 1;29(8):1363–75. doi:10.1101/gr.240663.118

178. Badia-i-Mompel P, Vélez Santiago J, Braunger J, Geiss C, Dimitrov D, Müller-Dott S, et al. decoupleR: ensemble of computational methods to infer biological activities from omics data. Bioinforma Adv. 2022 Jan 1;2(1):vbac016. doi:10.1093/bioadv/vbac016

179. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat Methods. 2016 Dec;13(12):966–7. doi:10.1038/nmeth.4077

180. Brendel M, Su C, Bai Z, Zhang H, Elemento O, Wang F. Application of Deep Learning on Single-Cell RNA Sequencing Data Analysis: A Review. Genomics Proteomics Bioinformatics. 2022 Oct 1;20(5):814–35. doi:10.1016/j.gpb.2022.11.011

181. Richter T, Bahrami M, Xia Y, Fischer DS, Theis FJ. Delineating the effective use of self-supervised learning in single-cell genomics. Nat Mach Intell. 2025 Jan;7(1):68–78. doi:10.1038/s42256-024-00934-3

182. Ma Q, Xu D. Deep learning shapes single-cell data analysis. Nat Rev Mol Cell Biol. 2022 May;23(5):303–4. doi:10.1038/s41580-022-00466-x

183. Kurts C, Panzer U, Anders HJ, Rees AJ. The immune system and kidney disease: basic concepts and clinical implications. Nat Rev Immunol. 2013 Oct;13(10):738–53. doi:10.1038/nri3523

184. He B, Chen P, Zambrano S, Dabaghie D, Hu Y, Möller-Hackbarth K, et al. Single-cell RNA sequencing reveals the mesangial identity and species diversity of glomerular cell transcriptomes. Nat Commun. 2021 Dec;12(1):2141. doi:10.1038/s41467-021-22331-9

185. Shankland SJ, Smeets B, Pippin JW, Moeller MJ. The emergence of the glomerular parietal epithelial cell. Nat Rev Nephrol. 2014 Mar;10(3):158–73. doi:10.1038/nrneph.2014.1

186. Chang AM, Ohse T, Krofft RD, Wu JS, Eddy AA, Pippin JW, et al. Albumin-induced apoptosis of glomerular parietal epithelial cells is modulated by extracellular signal-regulated kinase 1/2. Nephrol Dial Transplant. 2012 Apr 1;27(4):1330–43. doi:10.1093/ndt/gfr483

187. Smeets B, Kuppe C, Sicking EM, Fuss A, Jirak P, van Kuppevelt TH, et al. Parietal Epithelial Cells Participate in the Formation of Sclerotic Lesions in Focal Segmental Glomerulosclerosis. J Am Soc Nephrol. 2011 Jul;22(7):1262–74. doi:10.1681/ASN.2010090970

188. Dressler GR. The Cellular Basis of Kidney Development. Annu Rev Cell Dev Biol. 2006 Nov;22(1):509–29. doi:10.1146/annurev.cellbio.22.010305.104340

189. Perico L, Conti S, Benigni A, Remuzzi G. Podocyte–actin dynamics in health and disease. Nat Rev Nephrol. 2016 Nov;12(11):692–710. doi:10.1038/nrneph.2016.127

190. Jourde-Chiche N, Fakhouri F, Dou L, Bellien J, Burtey S, Frimat M, et al. Endothelium structure and function in kidney health and disease. Nat Rev Nephrol. 2019 Feb;15(2):87–108. doi:10.1038/s41581-018-0098-z

191. Avraham S, Korin B, Chung JJ, Oxburgh L, Shaw AS. The Mesangial cell — the glomerular stromal cell. Nat Rev Nephrol. 2021 Dec;17(12):855–64. doi:10.1038/s41581-021-00474-8

192. Schreibing F, Kramann R. Mapping the human kidney using single-cell genomics. Nat Rev Nephrol. 2022 Mar 17;18:347–60. doi:10.1038/s41581-022-00553-4

193. Lee JW, Chou CL, Knepper MA. Deep Sequencing in Microdissected Renal Tubules Identifies Nephron Segment–Specific Transcriptomes. J Am Soc Nephrol. 2015 Nov;26(11):2669–77. doi:10.1681/ASN.2014111067

194. Chen L, Clark JZ, Nelson JW, Kaissling B, Ellison DH, Knepper MA. Renal-Tubule Epithelial Cell Nomenclature for Single-Cell RNA-Sequencing Studies. J Am Soc Nephrol. 2019 Aug;30(8):1358–64. doi:10.1681/ASN.2019040415

195. Curthoys NP, Moe OW. Proximal Tubule Function and Response to Acidosis. Clin J Am Soc Nephrol. 2014 Sep 5;9(9):1627–38. doi:10.2215/CJN.10391012

196. Young MD, Mitchell TJ, Vieira Braga FA, Tran MGB, Stewart BJ, Ferdinand JR, et al. Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. Science. 2018 Aug 10;361(6402):594–9. doi:10.1126/science.aat1699

197. Subramanya AR, Ellison DH. Distal Convoluted Tubule. Clin J Am Soc Nephrol. 2014 Dec 5;9(12):2147–63. doi:10.2215/CJN.05920613

198. Pearce D, Soundararajan R, Trimpert C, Kashlan OB, Deen PMT, Kohan DE. Collecting Duct Principal Cell Transport Processes and Their Regulation. Clin J Am Soc Nephrol. 2015 Jan 7;10(1):135–46. doi:10.2215/CJN.05760513

199. Roy A, Al-bataineh MM, Pastor-Soler NM. Collecting Duct Intercalated Cell Function and Regulation. Clin J Am Soc Nephrol. 2015 Feb;10(2):305–24. doi:10.2215/cjn.08880914

200. Saxena V, Gao H, Arregui S, Zollman A, Kamocka MM, Xuei X, et al. Kidney intercalated cells are phagocytic and acidify internalized uropathogenic Escherichia coli. Nat Commun. 2021 Apr 23;12(1):1. doi:10.1038/s41467-021-22672-5

201. Dumas SJ, Meta E, Borri M, Luo Y, Li X, Rabelink TJ, et al. Phenotypic diversity and metabolic specialization of renal endothelial cells. Nat Rev Nephrol. 2021 Jul;17(7):441–64. doi:10.1038/s41581-021-00411-9

202. Zacchia M, capolongo giovanna, Rinaldi L, Capasso G. The importance of the thick ascending limb of Henle's loop in renal physiology and pathophysiology. Int J Nephrol Renov Dis. 2018 Feb;Volume 11:81–92. doi:10.2147/IJNRD.S154000

203. Kenig-Kozlovsky Y, Scott RP, Onay T, Carota IA, Thomson BR, Gil HJ, et al. Ascending Vasa Recta Are Angiopoietin/Tie2-Dependent Lymphatic-Like Vessels. J Am Soc Nephrol. 2018 Apr;29(4):1097–107. doi:10.1681/ASN.2017090962

204. Stewart BJ, Ferdinand JR, Young MD, Mitchell TJ, Loudon KW, Riding AM, et al. Spatiotemporal immune zonation of the human kidney. Science. 2019 Sep 27;365(6460):1461–6. doi:10.1126/science.aat5031

205. Stewart BJ, Ferdinand JR, Clatworthy MR. Using single-cell technologies to map the human immune system — implications for

nephrology. Nat Rev Nephrol. 2020 Feb;16(2):112–28. doi:10.1038/s41581-019-0227-3

206. Mantovani A, Sozzani S, Locati M, Allavena P, Sica A. Macrophage polarization: tumor-associated macrophages as a paradigm for polarized M2 mononuclear phagocytes. Trends Immunol. 2002 Nov 1;23(11):549–55. doi:10.1016/S1471-4906(02)02302-5

207. Liao J, Yu Z, Chen Y, Bao M, Zou C, Zhang H, et al. Single-cell RNA sequencing of human kidney. Sci Data. 2020 Dec;7(1):4. doi:10.1038/s41597-019-0351-8

208. Muto Y, Wilson PC, Ledru N, Wu H, Dimke H, Waikar SS, et al. Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney. Nat Commun. 2021 Dec;12(1):2190. doi:10.1038/s41467-021-22368-w

209. Lake BB, Menon R, Winfree S, Hu Q, Ferreira RM, Kalhor K, et al. An atlas of healthy and injured cell states and niches in the human kidney. Nature. 2023 Jul;619(7970):7970. doi:10.1038/s41586-023-05769-3

210. Abedini A, Levinsohn J, Klötzer KA, Dumoulin B, Ma Z, Frederick J, et al. Single-cell multi-omic and spatial profiling of human kidneys implicates the fibrotic microenvironment in kidney disease progression. Nat Genet. 2024 Aug;56(8):1712–24. doi:10.1038/s41588-024-01802-x

211. Dizman N, Philip EJ, Pal SK. Genomic profiling in renal cell carcinoma. Nat Rev Nephrol. 2020 Aug;16(8):435–51. doi:10.1038/s41581-020-0301-x

212. Hsieh JJ, Purdue MP, Signoretti S, Swanton C, Albiges L, Schmidinger M, et al. Renal cell carcinoma. Nat Rev Dis Primer. 2017 Dec 21;3(1):17009. doi:10.1038/nrdp.2017.9

213. Senkin S, Moody S, Díaz-Gay M, Abedi-Ardekani B, Cattiaux T, Ferreiro-Iglesias A, et al. Geographic variation of mutagenic exposures in kidney cancer genomes. Nature. 2024 May;629(8013):910–8. doi:10.1038/s41586-024-07368-2

214. Riazalhosseini Y, Lathrop M. Precision medicine from the renal cancer genome. Nat Rev Nephrol. 2016 Nov;12(11):655–66. doi:10.1038/nrneph.2016.133

215. Powles T, Albiges L, Bex A, Comperat E, Grünwald V, Kanesvaran R, et al. Renal cell carcinoma: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up☆. Ann Oncol. 2024 Aug

1;35(8):692–706. doi:10.1016/j.annonc.2024.05.537 PubMed PMID: 38788900.

216. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature. 2013 Jul;499(7456):43–9. doi:10.1038/nature12222

217. Semenza GL. HIF-1 mediates metabolic responses to intratumoral hypoxia and oncogenic mutations. J Clin Invest. 2013 Sep 3;123(9):3664–71. doi:10.1172/JCI67230

218. Dalgliesh GL, Furge K, Greenman C, Chen L, Bignell G, Butler A, et al. Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. Nature. 2010 Jan;463(7279):360–3. doi:10.1038/nature08672

219. Chen F, Zhang Y, Şenbabaoğlu Y, Ciriello G, Yang L, Reznik E, et al. Multilevel Genomics-Based Taxonomy of Renal Cell Carcinoma. Cell Rep. 2016 Mar;14(10):2476–89. doi:10.1016/j.celrep.2016.02.024

220. Gerlinger M, Endesfelder D, Stewart A, Tarpey P, McDonald NQ, Santos CR, et al. Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. N Engl J Med. 2012;10.

221. Kim KT, Lee HW, Lee HO, Song HJ, Jeong DE, Shin S, et al. Application of single-cell RNA sequencing in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma. Genome Biol. 2016 Dec;17(1):80. doi:10.1186/s13059-016-0945-9

222. Zhang Y, Narayanan SP, Mannan R, Raskind G, Wang X, Vats P, et al. Single-cell analyses of renal cell cancers reveal insights into tumor microenvironment, cell of origin, and therapy response. Proc Natl Acad Sci. 2021 Jun 15;118(24):e2103240118. doi:10.1073/pnas.2103240118

223. Peired AJ, Antonelli G, Angelotti ML, Allinovi M, Guzzi F, Sisti A, et al. Acute kidney injury promotes development of papillary renal cell adenoma and carcinoma from renal progenitor cells. Sci Transl Med. 2020 Mar 25;12(536):eaaw6003. doi:10.1126/scitranslmed.aaw6003

224. Webster WS, Lohse CM, Thompson RH, Dong H, Frigola X, Dicks DL, et al. Mononuclear cell infiltration in clear-cell renal cell carcinoma independently predicts patient survival. Cancer. 2006;107(1):46–53. doi:10.1002/cncr.21951

225. Chevrier S, Levine JH, Zanotelli VRT, Silina K, Schulz D, Bacac M, et al. An Immune Atlas of Clear Cell Renal Cell Carcinoma. Cell. 2017 May;169(4):736-749.e18. doi:10.1016/j.cell.2017.04.016

226. Fridman WH, Zitvogel L, Sautès–Fridman C, Kroemer G. The immune contexture in cancer prognosis and treatment. Nat Rev Clin Oncol. 2017 Dec;14(12):717–34. doi:10.1038/nrclinonc.2017.101

227. Braun DA, Bakouny Z, Hirsch L, Flippot R, Van Allen EM, Wu CJ, et al. Beyond conventional immune-checkpoint inhibition — novel immunotherapies for renal cell carcinoma. Nat Rev Clin Oncol. 2021 Apr;18(4):199–214. doi:10.1038/s41571-020-00455-z

228. Jansen CS, Prokhnevska N, Master VA, Sanda MG, Carlisle JW, Bilen MA, et al. An intra-tumoral niche maintains and differentiates stem-like CD8 T cells. Nature. 2019 Dec;576(7787):465–70. doi:10.1038/s41586-019-1836-5

229. Şenbabaoğlu Y, Gejman RS, Winer AG, Liu M, Van Allen EM, de Velasco G, et al. Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. Genome Biol. 2016 Dec;17(1):231. doi:10.1186/s13059-016-1092-z

230. Borcherding N, Vishwakarma A, Voigt AP, Bellizzi A, Kaplan J, Nepple K, et al. Mapping the immune environment in clear cell renal carcinoma by single-cell genomics. Commun Biol. 2021 Jan 27;4(1):1–11. doi:10.1038/s42003-020-01625-6

231. Krishna C, DiNatale RG, Kuo F, Srivastava RM, Vuong L, Chowell D, et al. Single-cell sequencing links multiregional immune landscapes and tissue-resident T cells in ccRCC to tumor topology and therapy efficacy. Cancer Cell. 2021 May;39(5):662-677.e6. doi:10.1016/j.ccell.2021.03.007

232. Braun DA, Street K, Burke KP, Cookmeyer DL, Denize T, Pedersen CB, et al. Progressive immune dysfunction with advancing disease stage in renal cell carcinoma. Cancer Cell. 2021 May;39(5):632-648.e8. doi:10.1016/j.ccell.2021.02.013

233. Bi K, He MX, Bakouny Z, Kanodia A, Napolitano S, Wu J, et al. Tumor and immune reprogramming during immunotherapy in advanced renal cell carcinoma. Cancer Cell. 2021 May;39(5):649-661.e5. doi:10.1016/j.ccell.2021.02.015

234. Li R, Ferdinand JR, Loudon KW, Bowyer GS, Laidlaw S, Muyas F, et al. Mapping single-cell transcriptomes in the intra-tumoral and associated territories of kidney cancer. Cancer Cell. 2022 Dec 12;40(12):1583-1599.e10. doi:10.1016/j.ccell.2022.11.001 PubMed PMID: 36423636.

235. Wu TD, Madireddi S, de Almeida PE, Banchereau R, Chen YJJ, Chitre AS, et al. Peripheral T cell expansion predicts tumour infiltration and clinical response. Nature. 2020 Mar;579(7798):274–8. doi:10.1038/s41586-020-2056-8

236. Li R, Ferdinand JR, Loudon KW, Bowyer GS, Laidlaw S, Muyas F, et al. Mapping single-cell transcriptomes in the intra-tumoral and associated territories of kidney cancer. Cancer Cell. 2022 Dec;40(12):1583-1599.e10. doi:10.1016/j.ccell.2022.11.001

237. McDermott DF, Huseni MA, Atkins MB, Motzer RJ, Rini BI, Escudier B, et al. Clinical activity and molecular correlates of response to atezolizumab alone or in combination with bevacizumab versus sunitinib in renal cell carcinoma. Nat Med. 2018 Jun;24(6):749–57. doi:10.1038/s41591-018-0053-3

238. Hakimi AA, Voss MH, Kuo F, Sanchez A, Liu M, Nixon BG, et al. Transcriptomic Profiling of the Tumor Microenvironment Reveals Distinct Subgroups of Clear Cell Renal Cell Cancer: Data from a Randomized Phase III Trial. Cancer Discov. 2019 Apr 1;9(4):510–25. doi:10.1158/2159-8290.CD-18-0957

239. Liu Y, Cao X. The origin and function of tumor-associated macrophages. Cell Mol Immunol. 2015 Jan;12(1):1–4. doi:10.1038/cmi.2014.83

240. Hu J, Chen Z, Bao L, Zhou L, Hou Y, Liu L, et al. Single-Cell Transcriptome Analysis Reveals Intratumoral Heterogeneity in ccRCC, which Results in Different Clinical Outcomes. Mol Ther. 2020 Jul;28(7):1658–72. doi:10.1016/j.ymthe.2020.04.023

241. Obradovic A, Chowdhury N, Haake SM, Ager C, Wang V, Vlahos L, et al. Single-cell protein activity analysis identifies recurrence-associated renal tumor macrophages. Cell. 2021 May;184(11):2988-3005.e16. doi:10.1016/j.cell.2021.04.038

242. Sjöberg E. Molecular mechanisms and clinical relevance of endothelial cell cross-talk in clear cell renal cell carcinoma. Ups J Med Sci. 2024 May 8;129:10.48101/ujms.v129.10632. doi:10.48101/ujms.v129.10632 PubMed PMID: 38863726; PubMed Central PMCID: PMC11165252.

243. Long Z, Sun C, Tang M, Wang Y, Ma J, Yu J, et al. Single-cell multiomics analysis reveals regulatory programs in clear cell renal cell carcinoma. Cell Discov. 2022 Jul 19;8(1):1. doi:10.1038/s41421-022-00415-0

244. Alchahin AM, Mei S, Tsea I, Hirz T, Kfoury Y, Dahl D, et al. A transcriptional metastatic signature predicts survival in clear cell renal cell carcinoma. Nat Commun. 2022 Sep 30;13(1):1. doi:10.1038/s41467-022-33375-w

245. Xu Y, Miller CP, Xue J, Zheng Y, Warren EH, Tykodi SS, et al. Single cell atlas of kidney cancer endothelial cells reveals distinct expression profiles and phenotypes. BJC Rep. 2024 Mar 14;2(1):23. doi:10.1038/s44276-024-00047-9

246. Mei S, Alchahin AM, Tsea I, Kfoury Y, Hirz T, Jeffries NE, et al. Single-cell analysis of immune and stroma cell remodeling in clear cell renal cell carcinoma primary tumors and bone metastatic lesions. Genome Med. 2024 Jan 29;16(1):1. doi:10.1186/s13073-023-01272-6

247. Tang H, Xu W, Lu J, Anwaier A, Ye D, Zhang H. Heterogeneity and function of cancer-associated fibroblasts in renal cell carcinoma. J Natl Cancer Cent. 2023 Jun 1;3(2):100–5. doi:10.1016/j.jncc.2023.04.001

248. Peng YL, Xiong LB, Zhou ZH, Ning K, Li Z, Wu ZS, et al. Single-cell transcriptomics reveals a low CD8+ T cell infiltrating state mediated by fibroblasts in recurrent renal cell carcinoma. J Immunother Cancer. 2022 Feb 4;10(2):e004206. doi:10.1136/jitc-2021-004206 PubMed PMID: 35121646; PubMed Central PMCID: PMC8819783.

249. Lu T, Zhang J, Lu S, Yang F, Gan L, Wu X, et al. Endosialin-positive tumor-derived pericytes promote tumor progression through impeding the infiltration of CD8+ T cells in clear cell renal cell carcinoma. Cancer Immunol Immunother. 2023 Jun 1;72(6):1739–50. doi:10.1007/s00262-023-03372-z

250. Anfray, Ummarino, Andón, Allavena. Current Strategies to Target Tumor-Associated-Macrophages to Improve Anti-Tumor Immune Responses. Cells. 2019 Dec 23;9(1):46. doi:10.3390/cells9010046

251. Weiss SA, Djureinovic D, Jessel S, Krykbaeva I, Zhang L, Jilaveanu L, et al. A Phase I Study of APX005M and Cabiralizumab with or without Nivolumab in Patients with Melanoma, Kidney Cancer, or Non–Small Cell Lung Cancer Resistant to Anti-PD-1/PD-L1. Clin Cancer Res. 2021 Sep 1;27(17):4757–67. doi:10.1158/1078-0432.CCR-21-0903

252. Underwood MA, Gilbert WM, Sherman MP. Amniotic Fluid: Not Just Fetal Urine Anymore. J Perinatol. 2005 May;25(5):5. doi:10.1038/sj.jp.7211290

253. Chaaban H, Burge K, McElroy SJ. Evolutionary bridges: how factors present in amniotic fluid and human milk help mature the gut. J Perinatol. 2024 Nov;44(11):1552–9. doi:10.1038/s41372-024-02026-x

254. Underwood MA, Sherman MP. Nutritional Characteristics of Amniotic Fluid. NeoReviews. 2006 Jun 1;7(6):e310–6. doi:10.1542/neo.7-6-e310

255. Crosland BA, Hedges MA, Ryan KS, D'mello RJ, Mccarty OJT, Malhotra SV, et al. Amniotic fluid: its role in fetal development and beyond. J Perinatol. 2025 Aug;45(8):1163–70. doi:10.1038/s41372-025-02313-1

256. Ross MG, Nijland MJM. Development of ingestive behavior. Am J Physiol-Regul Integr Comp Physiol. 1998 Apr;274(4):R879–93. doi:10.1152/ajpregu.1998.274.4.R879

257. Sherer DM. A Review of Amniotic Fluid Dynamics and the Enigma of Isolated Oligohydramnios. Am J Perinatol. 2002;19(5):253–66. doi:10.1055/s-2002-33084

258. Huri M, Di Tommaso M, Seravalli V. Amniotic Fluid Disorders: From Prenatal Management to Neonatal Outcomes. Children. 2023 Mar;10(3):561. doi:10.3390/children10030561

259. Slimani S, Hounka S, Mahmoudi A, Rehah T, Laoudiyi D, Saadi H, et al. Fetal biometry and amniotic fluid volume assessment end-to-end automation using Deep Learning. Nat Commun. 2023 Nov 3;14(1):7047. doi:10.1038/s41467-023-42438-5

260. Milunsky A, Sapirstein VS. Prenatal diagnosis of open neural tube defects using the amniotic fluid acetylcholinesterase assay. Obstet Gynecol. 1982 Jan;59(1):1–5. PubMed PMID: 6176922.

261. Zwemer LM, Bianchi DW. The Amniotic Fluid Transcriptome as a Guide to Understanding Fetal Disease. Cold Spring Harb Perspect Med. 2015 Apr 1;5(4):a023101–a023101. doi:10.1101/cshperspect.a023101

262. Orczyk-Pawilowicz M, Jawien E, Deja S, Hirnle L, Zabek A, Mlynarz P. Metabolomics of Human Amniotic Fluid and Maternal Plasma during Normal Pregnancy. PLOS ONE. 2016 Apr 12;11(4):e0152740. doi:10.1371/journal.pone.0152740

263. Levy HL, Montag PP. Free Amino Acids in Human Amniotic Fluid. A Quantitative Study by Ion-Exchange Chromatography. Pediatr Res. 1969 Mar;3(2):113–20. doi:10.1203/00006450-196903000-00002

264. Dawson EB, Evans DR, Van Hook JW. Amniotic fluid B12 and folate levels associated with neural tube defects. Am J Perinatol. 1998;15(9):511–4. doi:10.1055/s-2007-993975 PubMed PMID: 9890246.

265. Jang Y, Kim EK, Shim WS, Song KM, Kim SM. Amniotic fluid exerts a neurotrophic influence on fetal neurodevelopment via the ERK/GSK-3 pathway. Biol Res. 2015 Aug 5;48(1):44. doi:10.1186/s40659-015-0029-4

266. Hirai C, Ichiba H, Saito M, Shintaku H, Yamano T, Kusuda S. Trophic Effect of Multiple Growth Factors in Amniotic Fluid or Human Milk on Cultured Human Fetal Small Intestinal Cells. J Pediatr Gastroenterol Nutr. 2002;34(5).

267. Pierce J, Jacobson P, Benedetti E, Peterson E, Phibbs J, Preslar A, et al. Collection and characterization of amniotic fluid from scheduled C-section deliveries. Cell Tissue Bank. 2016 Sep 1;17(3):413–25. doi:10.1007/s10561-016-9572-7

268. Yoshio H, Tollin M, Gudmundsson GH, Lagercrantz H, Jörnvall H, Marchini G, et al. Antimicrobial Polypeptides of Human Vernix Caseosa and Amniotic Fluid: Implications for Newborn Innate Defense. Pediatr Res. 2003 Feb;53(2):211–6. doi:10.1203/01.PDR.0000047471.47777.B0

269. Mao Y, Pierce J, Singh-Varma A, Boyer M, Kohn J, Reems JA. Processed human amniotic fluid retains its antibacterial activity. J Transl Med. 2019 Mar 1;17(1):68. doi:10.1186/s12967-019-1812-8

270. Stinson LF, Boyce MC, Payne MS, Keelan JA. The Not-so-Sterile Womb: Evidence That the Human Fetus Is Exposed to Bacteria Prior to Birth. Front Microbiol. 2019 Jun 4;10:1124. doi:10.3389/fmicb.2019.01124

271. He Q, Kwok LY, Xi X, Zhong Z, Ma T, Xu H, et al. The meconium microbiota shares more features with the amniotic fluid microbiota than the maternal fecal and vaginal microbiota. Gut Microbes. 2020 Nov 9;12(1):1794266. doi:10.1080/19490976.2020.1794266 PubMed PMID: 32744162; PubMed Central PMCID: PMC7524391.

272. Kennedy KM, de Goffau MC, Perez-Muñoz ME, Arrieta MC, Bäckhed F, Bork P, et al. Questioning the fetal microbiome illustrates pitfalls of low-biomass microbial studies. Nature. 2023 Jan;613(7945):639–49. doi:10.1038/s41586-022-05546-8

273. Gomez-Lopez N, Romero R, Xu Y, Miller D, Leng Y, Panaitescu B, et al. The immunophenotype of amniotic fluid leukocytes in normal and complicated pregnancies. Am J Reprod Immunol. 2018;79(4):e12827. doi:10.1111/aji.12827

274. Marquardt N, Ivarsson MA, Sundström E, Åkesson E, Martini E, Eidsmo L, et al. Fetal CD103+ IL-17–Producing Group 3 Innate Lymphoid Cells Represent the Dominant Lymphocyte Subset in Human Amniotic Fluid. J Immunol. 2016 Oct 15;197(8):3069–75. doi:10.4049/jimmunol.1502204 PubMed PMID: 27591320.

275. Miller D, Gershater M, Slutsky R, Romero R, Gomez-Lopez N. Maternal and fetal T cells in term pregnancy and preterm labor. Cell Mol Immunol. 2020 Jul;17(7):7. doi:10.1038/s41423-020-0471-2

276. Gomez-Lopez N, Romero R, Xu Y, Leng Y, Garcia-Flores V, Miller D, et al. Are amniotic fluid neutrophils in women with intraamniotic infection and/or inflammation of fetal or maternal origin? Am J Obstet Gynecol. 2017 gruodžio;217(6):693.e1-693.e16. doi:10.1016/j.ajog.2017.09.013

277. Gomez-Lopez N, Romero R, Leng Y, Xu Y, Slutsky R, Levenson D, et al. The origin of amniotic fluid monocytes/macrophages in women with intra-amniotic inflammation or infection. J Perinat Med. 2019 Oct 1;47(8):822–40. doi:10.1515/jpm-2019-0262

278. Svenvik M, Jenmalm MC, Brudin L, Raffetseder J, Hellberg S, Axelsson D, et al. Chemokine and cytokine profiles in preterm and term labor, in preterm prelabor rupture of the membranes, and in normal pregnancy. J Reprod Immunol. 2024 Aug 1;164:104278. doi:10.1016/j.jri.2024.104278

279. Gomez-Lopez N, Romero R, Xu Y, Miller D, Arenas-Hernandez M, Garcia-Flores V, et al. Fetal T Cell Activation in the Amniotic Cavity during Preterm Labor: A Potential Mechanism for a Subset of Idiopathic Preterm Birth. J Immunol. 2019 Oct 1;203(7):1793–807. doi:10.4049/jimmunol.1900621

280. Gosden CM. AMNIOTIC FLUID CELL TYPES AND CULTURE. Br Med Bull. 1983;39(4):348–54. doi:10.1093/oxfordjournals.bmb.a071847

281. Von Koskull H, Aula P, Trejdosiewicz LK, Virtanen I. Identification of cells from fetal bladder epithelium in human amniotic fluid. Hum Genet. 1984 Jan;65(3):262–7. doi:10.1007/BF00286514

282. von Koskull H. Rapid identification of glial cells in human amniotic fluid with indirect immunofluorescence. Acta Cytol. 1984;28(4):393–400. PubMed PMID: 6205529.

283. Chitayat D, Marion RW, Squillante L, Kalousek DK, Das KM. Detection and enumeration of colonic mucosal cells in amniotic fluid using a colon epithelial-specific monoclonal antibody. Prenat Diagn. 1990 Nov;10(11):725–32. doi:10.1002/pd.1970101106

284. in `t Anker PS, Scherjon SA, Kleijburg-van der Keur C, Noort WA, Claas FHJ, Willemze R, et al. Amniotic fluid as a novel source of mesenchymal stem cells for therapeutic transplantation. Blood. 2003 Aug 15;102(4):1548–9. doi:10.1182/blood-2003-04-1291

285. Kaviani A, Guleserian K, Perry TE, Jennings RW, Ziegler MM, Fauza DO. Fetal Tissue Engineering from Amniotic Fluid. J Am Coll Surg. 2003 Apr;196(4):592–7. doi:10.1016/S1072-7515(02)01834-3

286. Prusa A, Marton E, Rosner M, Bernaschek G, Hengstschläger M. Oct-4-expressing cells in human amniotic fluid: a new source for stem cell research? Hum Reprod. 2003 Jul 1;18(7):1489–93. doi:10.1093/humrep/deg279

287. Gerli MFM, Calà G, Beesley MA, Sina B, Tullie L, Sun KY, et al. Single-cell guided prenatal derivation of primary fetal epithelial organoids from human amniotic and tracheal fluids. Nat Med. 2024 Mar 4;1–13. doi:10.1038/s41591-024-02807-z

288. Babosova O, Weisz B, Rabinowitz G, Avnet H, Shani H, Schwartz A, et al. Amniotic Fluid Organoids As Personalized Tools For Real-Time Modeling Of The Developing Fetus [Internet]. bioRxiv; 2023 [cited 2023 Oct 25]. p. 2023.10.05.561078. Available from: https://www.biorxiv.org/content/10.1101/2023.10.05.561078v1 doi:10.1101/2023.10.05.561078

289. Di Bernardo J, Kunisaki SM. Amniotic Fluid Stem Cell Populations. In: Fauza DO, Bani M, editors. Fetal Stem Cells in Regenerative Medicine: Principles and Translational Strategies [Internet]. New York, NY: Springer; 2016 [cited 2025 Nov 14]. p. 167–79. Available from: https://doi.org/10.1007/978-1-4939-3483-6_9 doi:10.1007/978-1-4939-3483-6_9

290. Chen WW. Studies on the origin of human amniotic fluid cells by immunofluorescent staining of keratin filaments. J Med Genet. 1982 Dec;19(6):433–6. doi:10.1136/jmg.19.6.433

291. Roubelakis MG, Bitsika V, Zagoura D, Trohatou O, Pappa KI, Makridakis M, et al. In vitro and in vivo properties of distinct populations of amniotic fluid mesenchymal progenitor cells. J Cell Mol Med. 2011;15(9):1896–913. doi:10.1111/j.1582-4934.2010.01180.x

292. Rahman MS, Spitzhorn LS, Wruck W, Hagenbeck C, Balan P, Graffmann N, et al. The presence of human mesenchymal stem cells of renal origin in amniotic fluid increases with gestational time. Stem Cell Res Ther. 2018 Apr 25;9(1):113. doi:10.1186/s13287-018-0864-7

293. Trounson A. A fluid means of stem cell generation. Nat Biotechnol. 2007 Jan;25(1):62–3. doi:10.1038/nbt0107-62

294. Ditadi A, de Coppi P, Picone O, Gautreau L, Smati R, Six E, et al. Human and murine amniotic fluid c-Kit+Lin− cells display hematopoietic activity. Blood. 2009 Apr 23;113(17):3953–60. doi:10.1182/blood-2008-10-182105

295. De Coppi P, Bartsch G, Siddiqui MM, Xu T, Santos CC, Perin L, et al. Isolation of amniotic stem cell lines with potential for therapy. Nat Biotechnol. 2007 Jan;25(1):100–6. doi:10.1038/nbt1274

296. Antonucci I, Stuppia L, Kaneko Y, Yu S, Tajiri N, Bae EC, et al. Amniotic Fluid as a Rich Source of Mesenchymal Stromal Cells for Transplantation Therapy. Cell Transplant. 2011 Jul 1;20(6):789–96. doi:10.3727/096368910X539074

297. Dobreva MP, Pereira PNG, Deprest J, Zwijsen A. On the origin of amniotic stem cells: of mice and men. Int J Dev Biol. 2010;54(5):761–77. doi:10.1387/ijdb.092935md

298. Rosner M, Mikula M, Preitschopf A, Feichtinger M, Schipany K, Hengstschläger M. Neurogenic differentiation of amniotic fluid stem cells. Amino Acids. 2012 May 1;42(5):1591–6. doi:10.1007/s00726-011-0929-8

299. Sacco SD, Sedrakyan S, Boldrin F, Giuliani S, Parnigotto P, Habibian R, et al. Human Amniotic Fluid as a Potential New Source of Organ Specific Precursor Cells for Future Regenerative Medicine Applications. J Urol. 2010 Mar;183(3):1193–200. Located at: Philadelphia, PA. doi:10.1016/j.juro.2009.11.006

300. Lesage F, Zia S, Jiménez J, Deprest J, Toelen J. The amniotic fluid as a source of mesenchymal stem cells with lung-specific characteristics. Prenat Diagn. 2017;37(11):1093–9. doi:10.1002/pd.5147

301. Ryan JM, Pettit AR, Guillot PV, Chan JKY, Fisk NM. Unravelling the Pluripotency Paradox in Fetal and Placental Mesenchymal Stem Cells: Oct-4 Expression and the Case of the Emperor's New Clothes. Stem Cell Rev Rep. 2013 Aug 1;9(4):408–21. doi:10.1007/s12015-011-9336-5

302. Vlahova F, Hawkins KE, Ranzoni AM, Hau KL, Sagar R, De Coppi P, et al. Human mid-trimester amniotic fluid (stem) cells lack expression of the pluripotency marker OCT4A. Sci Rep. 2019 May 31;9(1):1. doi:10.1038/s41598-019-44572-x

303. Stoeckius M, Zheng S, Houck-Loomis B, Hao S, Yeung BZ, Mauck WM, et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol. 2018 Dec 19;19(1):224. doi:10.1186/s13059-018-1603-1

304. Bernstein NJ, Fong NL, Lam I, Roy MA, Hendrickson DG, Kelley DR. Solo: Doublet Identification in Single-Cell RNA-Seq via Semi-Supervised Deep Learning. Cell Syst. 2020 Jul 22;11(1):95-101.e5. doi:10.1016/j.cels.2020.05.010

305. Goveia J, Rohlenova K, Taverna F, Treps L, Conradi LC, Pircher A, et al. An Integrated Gene Expression Landscape Profiling Approach to Identify Lung Tumor Endothelial Cell Heterogeneity and Angiogenic Candidates. Cancer Cell. 2020 Jan 13;37(1):21-36.e13. doi:10.1016/j.ccell.2019.12.001

306. Elmentaite R, Kumasaka N, Roberts K, Fleming A, Dann E, King HW, et al. Cells of the human intestinal tract mapped across space and time. Nature. 2021 Sep;597(7875):7875. doi:10.1038/s41586-021-03852-1

307. Barnes JL, Yoshida M, He P, Worlock KB, Lindeboom RGH, Suo C, et al. Early human lung immune cell development and its role in epithelial cell fate. Sci Immunol. 2023 Dec 15;8(90):eadf9988. doi:10.1126/sciimmunol.adf9988 PubMed PMID: 38100545; PubMed Central PMCID: PMC7615868.

308. Fang Z, Liu X, Peltz G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. Bioinformatics. 2023 Jan 1;39(1):btac757. doi:10.1093/bioinformatics/btac757

309. Chan JM, Quintanal-Villalonga Á, Gao VR, Xie Y, Allaj V, Chaudhary O, et al. Signatures of plasticity, metastasis, and immunosuppression in

an atlas of human small cell lung cancer. Cancer Cell. 2021 Nov;39(11):1479-1496.e18. doi:10.1016/j.ccell.2021.09.008

310. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a Shared Vision for Cancer Genomic Data. N Engl J Med. 2016 Sep 22;375(12):1109–12. doi:10.1056/NEJMp1607591

311. Zvirblyte J, Nainys J, Juzenas S, Goda K, Kubiliute R, Dasevicius D, et al. Single-cell transcriptional profiling of clear cell renal cell carcinoma reveals a tumor-associated endothelial tip cell phenotype. Commun Biol. 2024 Jun 28;7(1):1–15. doi:10.1038/s42003-024-06478-x

312. Rozenblatt-Rosen O, Regev A, Oberdoerffer P, Nawy T, Hupalowska A, Rood JE, et al. The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. Cell. 2020 Apr 16;181(2):236–49. doi:10.1016/j.cell.2020.03.053 PubMed PMID: 32302568.

313. Zhang X, Li T, Liu F, Chen Y, Yao J, Li Z, et al. Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. Mol Cell. 2019 Jan 3;73(1):130-142.e5. doi:10.1016/j.molcel.2018.10.020 PubMed PMID: 30472192.

314. Hagemann-Jensen M, Ziegenhain C, Sandberg R. Scalable single-cell RNA sequencing from full transcripts with Smart-seq3xpress. Nat Biotechnol. 2022 Oct;40(10):1452–7. doi:10.1038/s41587-022-01311-4

315. Wulf MG, Maguire S, Humbert P, Dai N, Bei Y, Nichols NM, et al. Non-templated addition and template switching by Moloney murine leukemia virus (MMLV)-based reverse transcriptases co-occur and compete with each other. J Biol Chem. 2019 Nov 29;294(48):18220–31. doi:10.1074/jbc.RA119.010676 PubMed PMID: 31640989; PubMed Central PMCID: PMC6885630.

316. Son SM, Yun J, Lee SH, Han HS, Lim YH, Woo CG, et al. Therapeutic Effect of pHLIP-mediated CEACAM6 Gene Silencing in Lung Adenocarcinoma. Sci Rep. 2019 Sep 2;9(1):1. doi:10.1038/s41598-019-48104-5

317. Sui L, Wang S, Ganguly D, El Rayes TP, Askeland C, Børretzen A, et al. PRSS2 remodels the tumor microenvironment via repression of Tsp1 to stimulate tumor growth and progression. Nat Commun. 2022 Dec 27;13(1):1. doi:10.1038/s41467-022-35649-9

318. Habermann AC, Gutierrez AJ, Bui LT, Yahn SL, Winters NI, Calvi CL, et al. Single-cell RNA sequencing reveals profibrotic roles of distinct

epithelial and mesenchymal lineages in pulmonary fibrosis. Sci Adv. 2020 Jul 8;6(28):eaba1972. doi:10.1126/sciadv.aba1972

319. Sinjab A, Han G, Treekitkarnmongkol W, Hara K, Brennan PM, Dang M, et al. Resolving the Spatial and Cellular Architecture of Lung Adenocarcinoma by Multiregion Single-Cell Sequencing. Cancer Discov. 2021 Oct 1;11(10):2506–23. doi:10.1158/2159-8290.CD-20-1285

320. Bischoff P, Trinks A, Obermayer B, Pett JP, Wiederspahn J, Uhlitz F, et al. Single-cell RNA sequencing reveals distinct tumor microenvironmental patterns in lung adenocarcinoma. Oncogene. 2021 Dec;40(50):50. doi:10.1038/s41388-021-02054-3

321. Bensaad K, Favaro E, Lewis CA, Peck B, Lord S, Collins JM, et al. Fatty Acid Uptake and Lipid Storage Induced by HIF-1α Contribute to Cell Growth and Survival after Hypoxia-Reoxygenation. Cell Rep. 2014 Oct 9;9(1):349–65. doi:10.1016/j.celrep.2014.08.056 PubMed PMID: 25263561.

322. Shi Y, Zhang Q, Bi H, Lu M, Tan Y, Zou D, et al. Decoding the multicellular ecosystem of vena caval tumor thrombus in clear cell renal cell carcinoma by single-cell RNA sequencing. Genome Biol. 2022 Mar 31;23(1):87. doi:10.1186/s13059-022-02651-9

323. Balzer MS, Rohacs T, Susztak K. How Many Cell Types Are in the Kidney and What Do They Do? Annu Rev Physiol. 2022;84(1):507–31. doi:10.1146/annurev-physiol-052521-121841 PubMed PMID: 34843404.

324. Denisenko E, Guo BB, Jones M, Hou R, de Kock L, Lassmann T, et al. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. Genome Biol. 2020 Jun 2;21(1):130. doi:10.1186/s13059-020-02048-6

325. Rudman-Melnick V, Adam M, Potter A, Chokshi SM, Ma Q, Drake KA, et al. Single-Cell Profiling of AKI in a Murine Model Reveals Novel Transcriptional Signatures, Profibrotic Phenotype, and Epithelial-to-Stromal Crosstalk. J Am Soc Nephrol. 2020 Dec 1;31(12):2793–814. doi:10.1681/ASN.2020010052 PubMed PMID: 33115917.

326. Roumenina LT, Daugan MV, Noe R, Petitprez F, Vano YA, Sanchez-Salas R, et al. Tumor cells hijack macrophage-produced complement C1q to promote tumor growth. Cancer Immunol Res. 2019 Jun 4;canimm.0891.2018. doi:10.1158/2326-6066.CIR-18-0891

327. Pritykin Y, Veeken J van der, Pine AR, Zhong Y, Sahin M, Mazutis L, et al. A unified atlas of CD8 T cell dysfunctional states in cancer and infection. Mol Cell. 2021 Jun 3;81(11):2477-2493.e10. doi:10.1016/j.molcel.2021.03.045 PubMed PMID: 33891860.

328. Carosella ED, Gregori S, Roux DTL. HLA-G/LILRBs: A Cancer Immunotherapy Challenge. Trends Cancer. 2021 May 1;7(5):389–92. doi:10.1016/j.trecan.2021.01.004 PubMed PMID: 33563576.

329. Liu L, Zhang R, Deng J, Dai X, Zhu X, Fu Q, et al. Construction of TME and Identification of crosstalk between malignant cells and macrophages by SPP1 in hepatocellular carcinoma. Cancer Immunol Immunother. 2022 Jan 1;71(1):121–36. doi:10.1007/s00262-021-02967-8

330. Flieswasser T, Van den Eynde A, Van Audenaerde J, De Waele J, Lardon F, Riether C, et al. The CD70-CD27 axis in oncology: the new kids on the block. J Exp Clin Cancer Res. 2022 Jan 6;41(1):12. doi:10.1186/s13046-021-02215-y

331. Wang YH, Cheng TY, Chen TY, Chang KM, Chuang VP, Kao KJ. Plasmalemmal Vesicle Associated Protein (PLVAP) as a therapeutic target for treatment of hepatocellular carcinoma. BMC Cancer. 2014 Dec;14(1):815. doi:10.1186/1471-2407-14-815

332. De Palma M, Biziato D, Petrova TV. Microenvironmental regulation of tumour angiogenesis. Nat Rev Cancer. 2017 Aug;17(8):457–74. doi:10.1038/nrc.2017.51

333. Guzzi LM, Bergler T, Binnall B, Engelman DT, Forni L, Germain MJ, et al. Clinical use of [TIMP-2]•[IGFBP7] biomarker testing to assess risk of acute kidney injury in critical care: guidance from an expert panel. Crit Care. 2019 Dec;23(1):225. doi:10.1186/s13054-019-2504-8

334. Roudnicky F, Dieterich LC, Poyet C, Buser L, Wild P, Tang D, et al. High expression of insulin receptor on tumour-associated blood vessels in invasive bladder cancer predicts poor overall and progression-free survival. J Pathol. 2017;242(2):193–205. doi:10.1002/path.4892

335. Abe Y, Sakata-Yanagimoto M, Fujisawa M, Miyoshi H, Suehara Y, Hattori K, et al. A single-cell atlas of non-haematopoietic cells in human lymph nodes and lymphoma reveals a landscape of stromal remodelling. Nat Cell Biol. 2022 Apr;24(4):565–78. doi:10.1038/s41556-022-00866-3

336. Samuelson Bannow B, Recht M, Négrier C, Hermans C, Berntorp E, Eichler H, et al. Factor VIII: Long-established role in haemophilia A and

emerging evidence beyond haemostasis. Blood Rev. 2019 May 1;35:43–50. doi:10.1016/j.blre.2019.03.002

337. Su SC, Hu X, Kenney PA, Merrill MM, Babaian KN, Zhang XY, et al. Autotaxin–Lysophosphatidic Acid Signaling Axis Mediates Tumorigenesis and Development of Acquired Resistance to Sunitinib in Renal Cell Carcinoma. Clin Cancer Res. 2013 Dec 1;19(23):6461–72. doi:10.1158/1078-0432.CCR-13-1284

338. Sainson RCA, Johnston DA, Chu HC, Holderfield MT, Nakatsu MN, Crampton SP, et al. TNF primes endothelial cells for angiogenic sprouting by inducing a tip cell phenotype. Blood. 2008 May 15;111(10):4997–5007. doi:10.1182/blood-2007-08-108597

339. Lee WS, Yang H, Chon HJ, Kim C. Combination of anti-angiogenic therapy and immune checkpoint blockade normalizes vascular-immune crosstalk to potentiate cancer immunity. Exp Mol Med. 2020 Sep;52(9):9. doi:10.1038/s12276-020-00500-y

340. Rolny C, Mazzone M, Tugues S, Laoui D, Johansson I, Coulon C, et al. HRG Inhibits Tumor Growth and Metastasis by Inducing Macrophage Polarization and Vessel Normalization through Downregulation of PlGF. Cancer Cell. 2011 Jan;19(1):31–44. doi:10.1016/j.ccr.2010.11.009

341. Lee S, Kim H, Naidansuren P, Ham KA, Choi HS, Ahn H, et al. Peroxidasin is essential for endothelial cell survival and growth signaling by sulfilimine crosslink-dependent matrix assembly. FASEB J. 2020 Aug;34(8):10228–41. doi:10.1096/fj.201902899R

342. Yang X, Okamura DM, Lu X, Chen Y, Moorhead J, Varghese Z, et al. CD36 in chronic kidney disease: novel insights and therapeutic opportunities. Nat Rev Nephrol. 2017 Dec;13(12):769–81. doi:10.1038/nrneph.2017.126

343. Xu WH, Qu YY, Wang J, Wang HK, Wan FN, Zhao JY, et al. Elevated CD36 expression correlates with increased visceral adipose tissue and predicts poor prognosis in ccRCC patients. J Cancer. 2019;10(19):4522–31. doi:10.7150/jca.30989

344. Guessoum O, de Goes Martini A, Sequeira-Lopez MLS, Gomez RA. Deciphering the Identity of Renin Cells in Health and Disease. Trends Mol Med. 2021 Mar;27(3):280–92. doi:10.1016/j.molmed.2020.10.003

345. Moraes LA, Kar S, Foo SL, Gu T, Toh YQ, Ampomah PB, et al. Annexin-A1 enhances breast cancer growth and migration by promoting

alternative macrophage polarization in the tumour microenvironment. Sci Rep. 2017 Dec 20;7(1):1. doi:10.1038/s41598-017-17622-5

346. Araújo TG, Mota STS, Ferreira HSV, Ribeiro MA, Goulart LR, Vecchi L. Annexin A1 as a Regulator of Immune Response in Cancer. Cells. 2021 Sep;10(9):9. doi:10.3390/cells10092245

347. Salomé B, Sfakianos JP, Ranti D, Daza J, Bieber C, Charap A, et al. NKG2A and HLA-E define an alternative immune checkpoint axis in bladder cancer. Cancer Cell. 2022 Sep 12;40(9):1027-1043.e9. doi:10.1016/j.ccell.2022.08.005 PubMed PMID: 36099881.

348. André P, Denis C, Soulas C, Bourbon-Caillet C, Lopez J, Arnoux T, et al. Anti-NKG2A mAb Is a Checkpoint Inhibitor that Promotes Anti-tumor Immunity by Unleashing Both T and NK Cells. Cell. 2018 Dec 13;175(7):1731-1743.e13. doi:10.1016/j.cell.2018.10.014 PubMed PMID: 30503213; PubMed Central PMCID: PMC6292840.

349. Loukogeorgakis SP, De Coppi P. Concise Review: Amniotic Fluid Stem Cells: The Known, the Unknown, and Potential Regenerative Medicine Applications. STEM CELLS. 2017;35(7):1663–73. doi:10.1002/stem.2553

350. Savickiene J, Treigyte G, Baronaite S, Valiuliene G, Kaupinis A, Valius M, et al. Human Amniotic Fluid Mesenchymal Stem Cells from Second- and Third-Trimester Amniocentesis: Differentiation Potential, Molecular Signature, and Proteome Analysis. Stem Cells Int. 2015;2015(1):319238. doi:10.1155/2015/319238

351. Chakarov S, Lim HY, Tan L, Lim SY, See P, Lum J, et al. Two distinct interstitial macrophage populations coexist across tissues in specific subtissular niches. Science. 2019 Mar 15;363(6432):eaau0964. doi:10.1126/science.aau0964

352. Zhang L, Li Z, Skrzypczynska KM, Fang Q, Zhang W, O'Brien SA, et al. Single-Cell Analyses Inform Mechanisms of Myeloid-Targeted Therapies in Colon Cancer. Cell. 2020 Apr;181(2):442-459.e29. doi:10.1016/j.cell.2020.03.048

353. Gopee NH, Winheim E, Olabi B, Admane C, Foster AR, Huang N, et al. A prenatal skin atlas reveals immune regulation of human skin morphogenesis. Nature. 2024 Oct 16;1–11. doi:10.1038/s41586-024-08002-x

354. Valiulienė G, Zentelytė A, Beržanskytė E, Navakauskienė R. Metabolic Profile and Neurogenic Potential of Human Amniotic Fluid Stem Cells

From Normal vs. Fetus-Affected Gestations. Front Cell Dev Biol. 2021;9:1875. doi:10.3389/fcell.2021.700634

355. Gasiūnienė M, Zubova A, Utkus A, Navakauskienė R. Epigenetic and metabolic alterations in human amniotic fluid stem cells induced to cardiomyogenic differentiation by DNA methyltransferases and p53 inhibitors. J Cell Biochem. 2019 May;120(5):8129–43. doi:10.1002/jcb.28092

356. Ukita M, Hamanishi J, Yoshitomi H, Yamanoi K, Takamatsu S, Ueda A, et al. CXCL13-producing CD4+ T cells accumulate in the early phase of tertiary lymphoid structures in ovarian cancer. JCI Insight. 7(12):e157215. doi:10.1172/jci.insight.157215 PubMed PMID: 35552285; PubMed Central PMCID: PMC9309049.

357. Thommen DS, Koelzer VH, Herzig P, Roller A, Trefny M, Dimeloe S, et al. A transcriptionally and functionally distinct PD-1+ CD8+ T cell pool with predictive potential in non-small-cell lung cancer treated with PD-1 blockade. Nat Med. 2018 Jul;24(7):7. doi:10.1038/s41591-018-0057-z

358. Chen Y, McAndrews KM, Kalluri R. Clinical and therapeutic relevance of cancer-associated fibroblasts. Nat Rev Clin Oncol. 2021 Sep 6;1–13. doi:10.1038/s41571-021-00546-5

359. Mehner C, Radisky ES. Bad Tumors Made Worse: SPINK1. Front Cell Dev Biol. 2019 Feb 4;7(10). doi:10.3389/fcell.2019.00010

360. Xu L, Lu C, Huang Y, Zhou J, Wang X, Liu C, et al. SPINK1 promotes cell growth and metastasis of lung adenocarcinoma and acts as a novel prognostic biomarker. BMB Rep. 2018 Dec;51(12):648. doi:10.5483/BMBRep.2018.51.12.205 PubMed PMID: 30545439.

361. Kobayashi Y, Tata A, Konkimalla A, Katsura H, Lee RF, Ou J, et al. Persistence of a regeneration-associated, transitional alveolar epithelial cell state in pulmonary fibrosis. Nat Cell Biol. 2020 Aug;22(8):934–46. doi:10.1038/s41556-020-0542-8

362. Long Z, Sun C, Tang M, Wang Y, Ma J, Yu J, et al. Single-cell multiomics analysis reveals regulatory programs in clear cell renal cell carcinoma. Cell Discov. 2022 Jul 19;8(1):1. doi:10.1038/s41421-022-00415-0

363. Chauvin JM, Zarour HM. TIGIT in cancer immunotherapy. J Immunother Cancer. 2020 Sep 1;8(2):e000957. doi:10.1136/jitc-2020-000957 PubMed PMID: 32900861.

364. Hongu T, Pein M, Insua-Rodríguez J, Gutjahr E, Mattavelli G, Meier J, et al. Perivascular tenascin C triggers sequential activation of macrophages and endothelial cells to generate a pro-metastatic vascular niche in the lungs. Nat Cancer. 2022 Apr;3(4):4. doi:10.1038/s43018-022-00353-6

365. Peng YL, Xiong LB, Zhou ZH, Ning K, Li Z, Wu ZS, et al. Single-cell transcriptomics reveals a low CD8+ T cell infiltrating state mediated by fibroblasts in recurrent renal cell carcinoma. J Immunother Cancer. 2022 Feb 4;10(2):e004206. doi:10.1136/jitc-2021-004206 PubMed PMID: 35121646; PubMed Central PMCID: PMC8819783.

366. Baronas D, Norvaisis S, Zvirblyte J, Leonaviciene G, Mikulenaite V, Goda K, et al. High-throughput single cell omics using semipermeable capsules. Science. 2025 Dec 18;0(0):eady7227. doi:10.1126/science.ady7227

367. Lim AI, Li Y, Lopez-Lastra S, Stadhouders R, Paul F, Casrouge A, et al. Systemic Human ILC Precursors Provide a Substrate for Tissue ILC Differentiation. Cell. 2017 Mar 9;168(6):1086-1100.e10. doi:10.1016/j.cell.2017.02.021 PubMed PMID: 28283063.

368. He P, Lim K, Sun D, Pett JP, Jeng Q, Polanski K, et al. A human fetal lung cell atlas uncovers proximal-distal gradients of differentiation and key regulators of epithelial fates. Cell. 2022 Dec 8;185(25):4841-4860.e25. doi:10.1016/j.cell.2022.11.005 PubMed PMID: 36493756.

369. Kramann R, Kusaba T, Humphreys BD. Who regenerates the kidney tubule? Nephrol Dial Transplant. 2015 Jun 1;30(6):903–10. doi:10.1093/ndt/gfu281

370. Jardine L, Webb S, Goh I, Quiroga Londoño M, Reynolds G, Mather M, et al. Blood and immune development in human fetal bone marrow and Down syndrome. Nature. 2021 Oct;598(7880):327–31. doi:10.1038/s41586-021-03929-x

Supplementary Table S1. Available clinical information for lung carcinoma tissues.

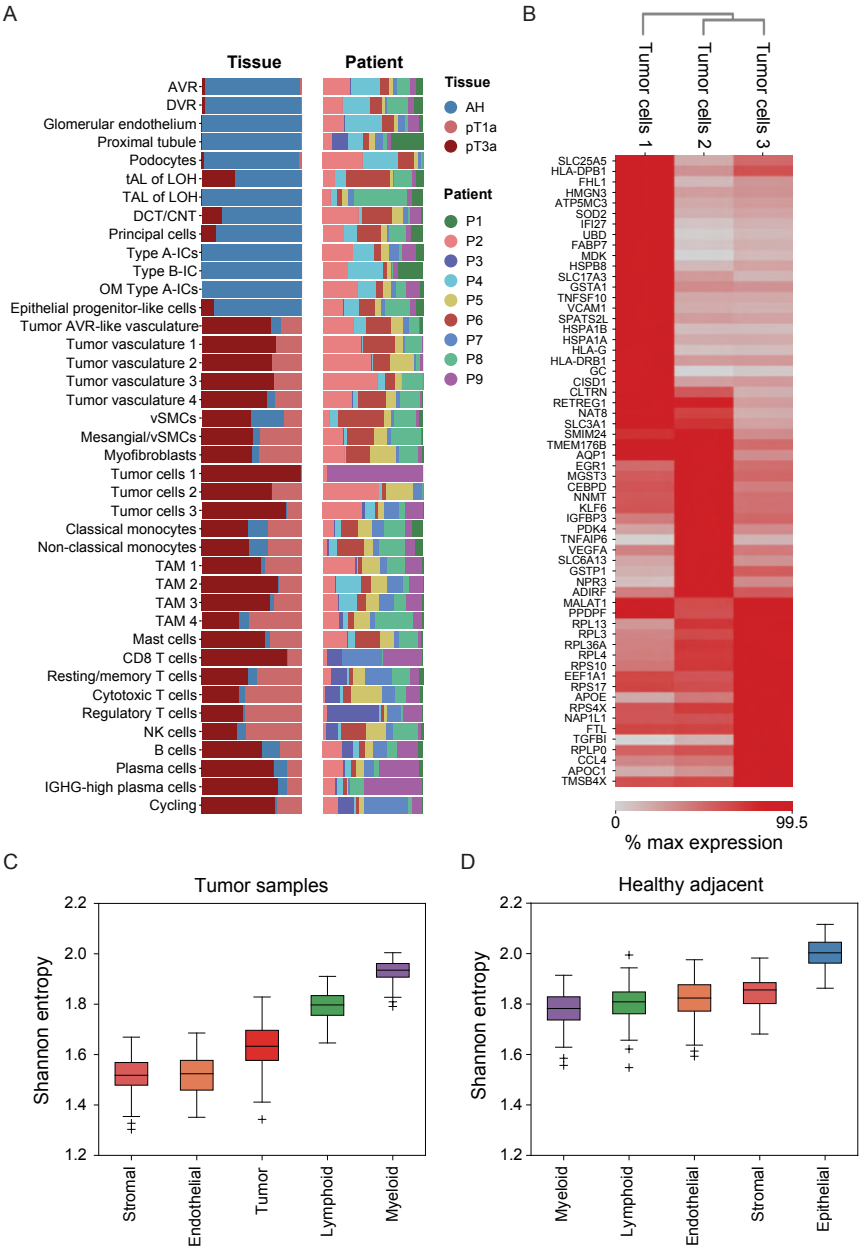
Sample ID	*Sex*	*Stage at diagnosis*	*Histology*	*Treatment status*	*Procedure*	*Smoking status*
P1	F	1A	Adeno	naïve	resection	former
P2	F	1B	Squamous	naïve	resection	former
P3	F	1A	Adeno	naïve	resection	never

Supplementary Table S2. Available clinical patient information for clear cell renal cell carcinoma (T) and healthy adjacent tissue (HA) specimens.

		Nephrectomy surgery						
Patient ID	*Samples*	*Radical (R) or partial (P)*	*Open (O) or laparascopic (L)*	*Sex*	*Age*	*pT stage*	*Tumor size, mm*	*Dif. grade*
P4	T, HA	R	O	M	62	pT3a	120	4
P2	T, HA	R	O	M	60	pT3a	75	2
P3	T, HA	P	O	M	63	pT1a	37	2
P1	HA	R	O	F	52	pT3a	13	4
P5	T, HA	R	O	F	65	pT1a	35	2
P6	T, HA	R	L	M	68	pT3a	30	2
P7	T, HA	R	L	M	43	pT3a	47	2
P8	T, HA	R	O	F	65	pT1a	33	2
P9	T, HA	R	L	F	61	pT3a	60	2

Supplementary Table S3. Available clinical patient information for amniotic fluid donors and samples. Y – yes, N – no, PCW – post-conception week.

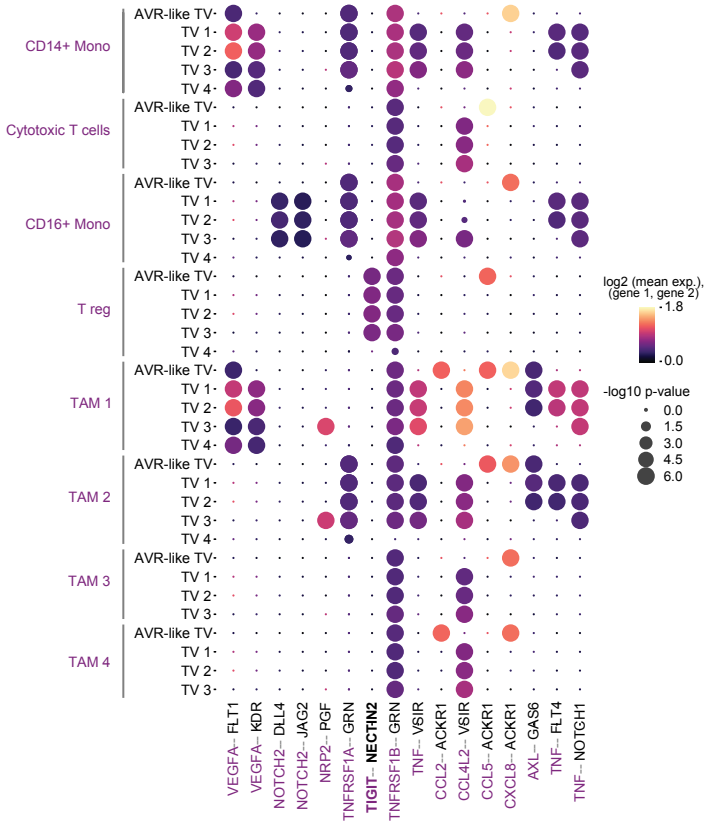
No.	*Sample ID*	*PCW*	*Method*	*Genetic test result*	*Maternal age above 35*	*Anomaly detected during ultrasound*
1	F01	16	qPCR, SNP array	Normal	N	Y
2	F02	20	qPCR, SNP array	Normal	Y	N
3	F04	16	qPCR	XXY	N	N
4	F10	16	qPCR	Trisomy 21	Y	N
5	F11	16	qPCR, SNP array	Normal	Y	N
6	F12	16	qPCR, SNP array	Normal	Y	N
7	F13	16	qPCR, SNP array	Normal	Y	N
8	F15	16	qPCR, SNP array	XXY	Y	N
9	F16	20	qPCR, SNP array	Normal	Y	Y
10	F17	20	qPCR, SNP array, *SOX9* seq, WES	Normal	N	Y
11	F29	16	qPCR, SNP array, MLPA	Normal	Y	N
12	F31	16	qPCR, SNP array, karyotyping	Trisomy 21	N	Y
13	F32	16	qPCR, SNP array	Normal	N	N
14	F34	16	qPCR	Trisomy 21	N	Y
15	F35	16	qPCR, SNP array	Normal	Y	N
16	F36	20	qPCR, SNP array	Normal	N	Y
17	F38	16	qPCR, SNP array	Normal	Y	N
18	F39	16	qPCR, SNP array	Normal	Y	N
19	F40	16	qPCR	45, X	N	Y
20	F41	16	qPCR, SNP array	Normal	Y	N
21	F45	20	qPCR, SNP array	Normal	N	Y
22	F46	16	qPCR	Triploidy	N	Y
23	F47	16	qPCR, SNP array	Part 4p duplicate	Y	Y
24	F48	16	qPCR, SNP array	Normal	Y	N
25	F49	20	qPCR, SNP array, *FGFR3* seq	Normal	N	Y
26	F50	20	qPCR, SNP array	Normal	N	Y



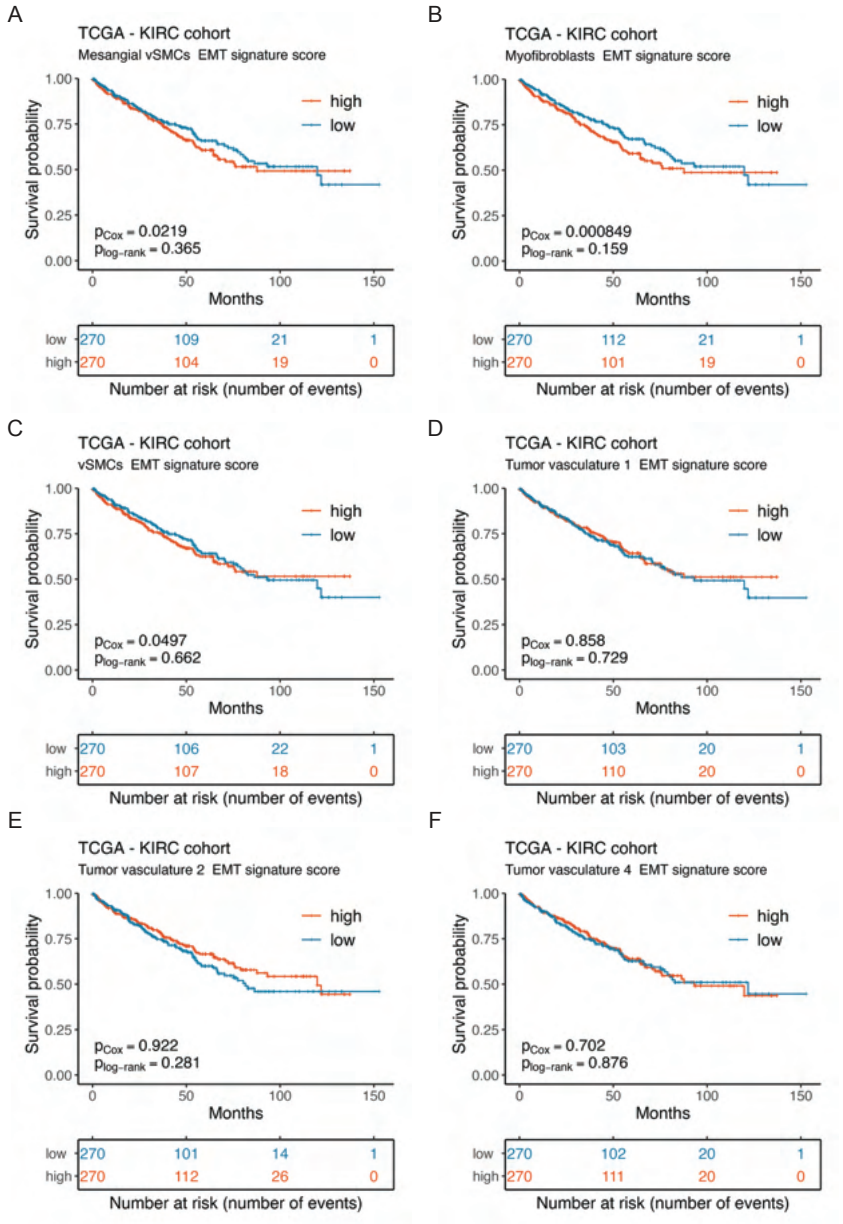
Supplementary Figure S1. Cell composition and inter-patient variability in ccRCC. **A** – cell composition by disease stage and patient ID. Specialized epithelial and endothelial cells mostly originated from the healthy-adjacent tissues, while immune, tumor, endothelial and stromal cells were enriched in the tumor samples. Cell phenotypes described were adequately represented by multiple samples, except for tumor cells 1 population, which was specific to patient P9. **B** – differential gene expression between tumor cell

subpopulations. Only genes with Benjamini-Hochberg adjusted p-value <0.05 are shown. Cptt-normalized expression level is showed, color saturates at 99.5th percentile of a given gene's expression level. **C**, **D** – tumor and healthy adjacent sample, respectively, heterogeneity for broad cell group as measured by Shannon entropy. Lower entropy values indicate higher sample heterogeneity. AH – adjacent healthy.

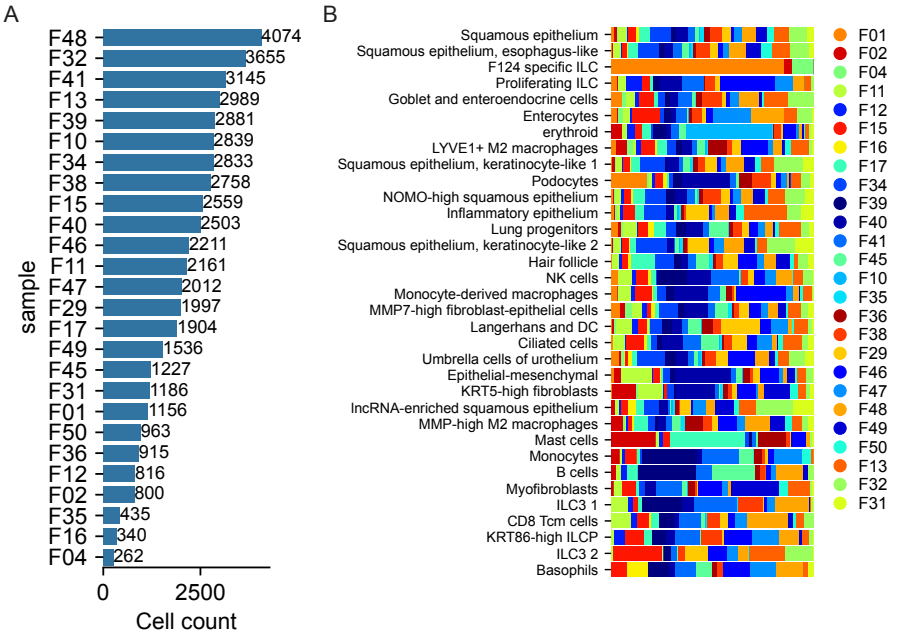
Supplementary Figure S2. Cptt-normalized expression of complement system molecules. C1QA, C1QB, C1QC was enriched in tumor-associated macrophages, while C1R and C1S were expressed by tumor and stromal cells. Color saturates at 99.5th percentile of a given gene's expression level.



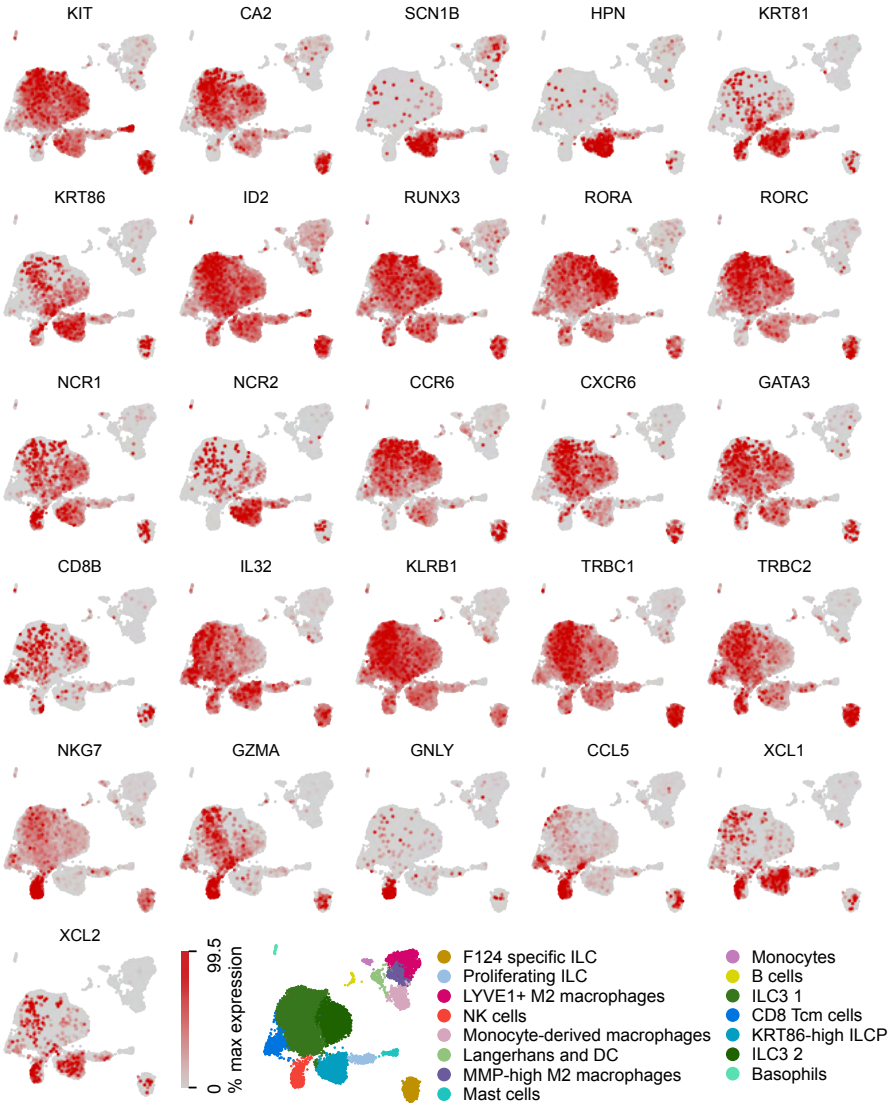
Supplementary Figure S3. Cell-cell communication analysis between immune cells and tumor vasculature populations. Notable immunosuppressive interaction was TIGIT and NECTIN2, predicted to appear between tumor vasculature and regulatory T cells.



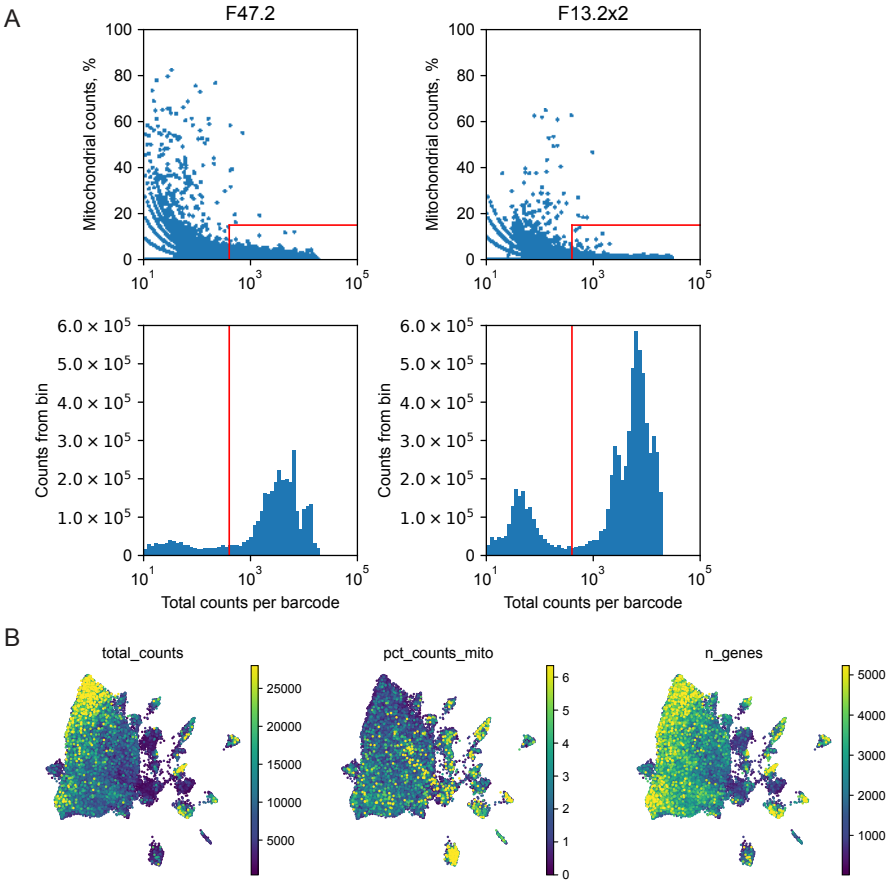
Supplementary Figure S4. Survival analysis of the TCGA KIRC cohort stratified by the expression of various stromal and endothelial subpopulation-enriched genes overlapping with the MSigDB EMT pathway. None of gene signatures have an effect, except for tumor vasculature 3 and AVR-like tumor vasculature populations, as shown in **Figure 3.18**.



Supplementary Figure S5. A – cell count per sample passing quality control, included in the global human AF cell atlas. **B** – cell population composition by samples profiled. All cell types were adequately represented by multiple samples and no sample-specific phenotypes were observed, aside from F01, F02 and F04-specific innate lymphoid cells.



Supplementary Figure S6. Human AF immune cell atlas colored by the cptt-normalized expression of selected ILCP, ILC, NK and T cell-associated genes, color saturating at 99.5th percentile of a given gene's expression level. Additionally, at the bottom, a UMAP colored by cell type annotations is provided for guidance.



Supplementary Figure S7. Quality metrics of amniotic fluid cell sequencing libraries and non-immune cells. **A** – representative libraries from AF sample experiments, mitochondrial gene count fraction and total counts per barcode distribution indicated adequate sample quality. **B** – a UMAP of AF non-immune cell atlas colored by total raw counts per barcode, mitochondrial gene count percentage and number of genes per cell. All three metrics showed good cell quality.

SANTRAUKA

Įvadas

Žmogaus audiniai pasižymi nepaprasta struktūrine ir funkcine įvairove – organizmo funkcijas ir homeostazę užtikrina koordinuota įvairių specializuotų ląstelių veikla. Nors somatinės ląstelės turi tą patį genomą, specializaciją apsprendžia preciziškai kontroliuojamas, nuo konteksto priklausantis selektyvus jo panaudojimas – genų raiška, kurios visuma ląstelėje vadinama transkriptomu. Ligos atveju, reguliacijai sutrikus, ir taip didelio masto ląstelių fenotipų įvairovė dar padidėja. Pavyzdžiui, vėžio kontekste ne tik atsiranda pačių vėžinių ląstelių, bet jos pertvarko naviko mikroaplinką, sukeldamos atsaką ir taip keisdamos aplinkinių nepiktybinių ar naviką infiltruojančių ląstelių fenotipą. Kita biologinė sistema, kurioje atsiskleidžia ypatingas ląstelių heterogeniškumas, yra organizmo vystymasis. Besiformuojančiame organizme galima aptikti įvairių nesubrendusių, tarpinių būsenų ląstelių, kurios nuolat kinta, kol galiausiai pasiekiama tinkama audinių struktūra ir funkcija.

Genų raiškos tyrimai, dar žinomi kaip transkriptomika, yra populiarus būdas analizuoti ląstelių būsenas ir skirtumus, ypač ligų kontekste. Vis dėlto, audinių tyrimai neatspindi ląstelių įvairovės, nes gaunamas suminis genų raiškos profilis, kuris nebūtinai yra būdingas konkrečiai ląstelei. Norint gilintis į ląstelių įvairovę reikalingi metodai, leidžiantys tirti viso genomo genų raišką vienos ląstelės lygmeniu. Pirmasis toks tyrimas pasirodė dar 2009-ais (1), tačiau nepaisant konceptualaus proveržio, pirminė technologija buvo brangi, reikalavo ilgų darbo valandų, ir tegalėjo tirti kelis šimtus ar kelis tūkstančius ląstelių. Sudėtingų audinių tyrimams reikalingas aukštas našumas ir didelis tiriamų ląstelių skaičius, be to, technologija turi būti įperkama, prieinama ir generuoti aukštos kokybės duomenis. Šių uždavinių sprendimas tapo tolesnių tobulinimų tikslu. 2015-ais tikru pavienių ląstelių RNR sekoskaitos proveržiu tapo lašeliais paremta ląstelių izoliavimo ir barkodavimo technologija. Dvi nepriklausomos grupės išpublikavo novatoriškus mikroskysčiais paremtus inDrops (2) ir Drop-seq (3) pavienių ląstelių sekoskaitos metodus, kurie žymiai padidino tiriamų ląstelių skaičių ir tokių eksperimentų prieinamumą. Šios inovacijos padėjo pamatus sparčiam pasauliniam pavienių ląstelių transkriptomikos srities išpopuliarėjimui, besitęsiančiam iki šiol.

Vienas pagrindinių pavienių ląstelių RNR sekoskaitos metodo privalumų yra galimybė tirti sudėtingas biologines sistemas neturint iš anksto žinomų žymenų, kurie dažniausiai naudojami klasikiniuose charakterizavimo

metoduose, tokiuose kaip imunocitochemija, citometrija ar tikro laiko PGR. Šis pranašumas buvo pademonstruotas atradus naujus, iki tol neaprašytus ląstelių tipus žmogaus organizme (4,5). Be to, įvairių sveikų ir ligos paveiktų audinių analizė vienos ląstelės lygmeniu atskleidė šių biologinių sistemų kompleksiškumą ir heterogeniškumą, bei ypač pagilino supratimą apie įvairių navikų mikroaplinką (6–8). Paskiri tyrimai greit tapo didelio masto tarptautinės tyrėjų iniciatyvos „Human Cell Atlas“ (*liet. žmogaus ląstelių atlasas*) dalimi, kurios tikslas – išsamiai kataloguoti visus žmogaus organizmo ląstelių tipus, per visą gyvenimo trukmę. Apibendrinant, išsamesnis nei kada nors ląstelių transkriptomo tyrimas, pasitelkiant pavienių ląstelių RNR sekoskaitos technologijas (scRNA-seq) jau atnešė svarų indėlį į fundamentalių biologinių procesų gilesnį supratimą bei atskleidė daugybę subtilių, nuo konteksto priklausomų ląstelinių pokyčių, keičiančių nusistovėjusią sampratą apie ląstelių tipus (10). Svarbiausia, kad šis metodas suteikė galimybę ne tik atsakyti, bet ir kelti gilius biologinius klausimus.

Šiame darbe pristatomų tyrimų ašis – pavienių ląstelių RNR sekoskaitos technologija, kuri pasitelkiama tirti įvairius sveikus ir ligos paveiktus žmogaus audinius. Pirmoje dalyje pristatoma patobulinta lašeliais paremta ląstelių mRNR barkodavimo sistema inDrops-2, kurios tinkamumas klinikinių mėginių analizei pademonstruojamas analizuojant plaučių karcinomos mėginius. Antroje dalyje išsamiai aprašomas sveikų inkstų ir šviesių ląstelių inkstų karcinomos ląstelių heterogeniškumas ir pristatomas naujas naviko endotelio ląstelių fenotipas. Galiausiai, pasitelkiant inDrops-2, sudarytas pirmasis nekultivuotų žmogaus vaisiaus vandenų ląstelių atlasas.

Tikslas

Apibūdinti ląstelių fenotipų įvairovę sveikų ir patologinių žmogaus audinių mėginiuose (plaučių, inkstų, vaisiaus vandenų), pasitelkiant pavienių ląstelių RNR sekoskaitos technologiją

Uždaviniai:

- Patvirtinti inDrops-2 tinkamumą klinikinių mėginių tyrimams atliekant fiksuotų plaučių karcinomos mėginių analizę
- Palyginti sveikų inkstų ir šviesių ląstelių inkstų karcinomos audinių ląstelinę sudėtį
- Apibūdinti endotelio ir imuninių ląstelių heterogeniškumą šviesių ląstelių inkstų karcinomos mėginiuose
- Sudaryti žmogaus vaisiaus vandenų ląstelių transkriptomų atlasą

- Aprašyti žmogaus vaisiaus vandenyse aptinkamų ląstelių fenotipus

Mokslinis naujumas

Šiandien, dėl paprasto naudojimo, rezultatų atkartojamumo ir aukštos duomenų kokybės, pavienių ląstelių transkriptomikos tyrimuose dažniausiai naudojamos komercinės platformos. Deja, tokioms sistemoms dažnai trūksta lankstumo, o kaina gali būti neprieinama daugeliui tyrėjų, ypač planuojant didelio masto projektus. Atviros prieigos (nekomerciniai) metodai, dėl savo pritaikomumo gali patenkinti specifinius eksperimentinius tikslus mažesnėmis sąnaudomis. Šiame darbe, tokios atviros platformos inDrops-2 privalumai pademonstruoti tiriant archyvuotus, chemiškai žymėtus ir fiksuotus plaučių karcinomos mėginius. Tokiu būdu ne tik patvirtinta komercinėms sistemoms nenusileidžianti gaunamų duomenų kokybė, bet ir atrasta įdomių, potencialiai kliniškai svarbių, retų ląstelių fenotipų plaučių karcinomos mėginiuose.

Taikant inDrops-2 sveikų inkstų ir šviesių ląstelių inkstų karcinomos (ccRCC) mėginių analizei, buvo sukurtas išsamus transkriptomų atlasas, suteikęs naujų įžvalgų apie naviko mikroaplinkoje esančius ląstelių fenotipus ir jų potencialų vaidmenį ligos kontekste. Svarbiausias šio darbo atradimas – viršūninių (*angl. tip cells*) naviko endotelio ląstelių fenotipas, iki šiol neaprašytas ccRCC kontekste. Atsižvelgiant į tai, kad pažengusios ir metastatinės ligos gydymui naudojama būtent angiogenezę slopinanti terapija, detalus naviko endotelio ląstelių ištyrimas, pristatytas šiame darbe, yra itin aktualus. Gauti rezultatai pagrindžia sampratą, kad su naviku susijęs endotelis skatina naviko augimą: pasižymi metastazavimui palankių genų bei specifinių užląstelinio užpildo komponentų raiška, bei galimai sąveikauja su naviką infiltruojančiomis imuninėmis ląstelėmis. Šios naviko augimui palankios sąveikos yra itin aktualios priešvėžinių terapijų kūrimo kontekste. Rezultatų mokslinį naujumą ir aktualumą patvirtina aukštas mokslinės bendruomenės susidomėjimas. Ypač džiugina tai, kad 2024 metų birželį pasirodžiusi publikacija jau buvo cituota kelias dešimtis kartų, o analizuoti duomenys aktyviai įtraukiami į kitus aukšto lygio mokslinius tyrimus.

Paskutinė darbo dalis, žmogaus vaisiaus vandenų transkriptomų atlaso sukūrimas, pabrėžia esminį pavienių ląstelių RNR sekoskaitos privalumą: galimybę tirti mažai aprašytas sistemas, potencialiai turinčias fenotipų, kuriems nėra žinomų žymenų. Nepaisant to, kad vaisiaus vandenys ilgus metus naudojami moksliniuose tyrimuose kaip kamieninių ląstelių šaltinis, didžioji dalis tyrimų vertina kultivuotų ląstelių fenotipą pagal

žinomus žymenis, pasitelkiant tokius metodus kaip citometrija ar tikro laiko PGR. Nekultivuotų vaisiaus vandenų ląstelių tyrimų apskritai trūksta, o esami taip pat apriboti naudojamų technologijų. Galimai dėl riboto mėginių prieinamumo ir techninių iššūkių, tokių kaip mažas ląstelių skaičius, vaisiaus vandenys taip pat nebuvo įtraukti į didelio masto projektus kataloguojančius žmogaus ląstelių tipus. Šiame darbe pristatomi rezultatai atskleidžia vaisiaus vandenų ląstelinę sudėtį dviem vystymosi momentais, 16 ir 20 savaitėmis po koncepcijos, ir pirmą kartą apibrėžia nuo vaisiaus audinių atkibusių ląstelių heterogeniškumą. Parodyta, kad vaisiaus vandenyse aptinkama įvairių fenotipų imuninių ir neimuninių ląstelių. Pastarąsias daugiausia sudaro epitelinės ląstelės, atkibusios nuo vaisiaus odos, žarnyno, inkstų ir plaučių, čia aptinkama ir pirmtakų, ir itin specializuotų ląstelių. Įdomu tai, kad aptinkama keletas tarpinių epitelinių-mezenchiminių fenotipų, mūsų žiniomis neturinčių publikuotų analogų kitose biologinėse sistemose. Imunines ląsteles sudaro tiek mieloidinės, tiek limfoidinės šakos ląstelės, kurios skiriasi gausumu tarp dviejų tirtų vaisiaus vystymosi stadijų. Apibendrinant, paskutinėje darbo dalyje pateikti rezultatai pabrėžia pavienių ląstelių RNR sekoskaitos naudą ir privalumus tiriant ląstelių įvairovę – sudarytas pirmasis išsamus vaisiaus audinių ląstelių, aptinkamų vaisiaus vandenyse, atlasas. Tikimės, kad greitu metu šie rezultatai taps analogų neturinčios mokslinės publikacijos dalimi.

Disertacijoje pristatomi teiginiai

- inDrops-2 pavienių ląstelių RNR sekoskaitos platforma tinkama aptikti retus ląstelių fenotipus klinikiniuose mėginiuose
- Šviesių ląstelių inkstų karcinomos naviko mikroaplinka gausiai infiltruota T limfocitų ir su vėžiu susijusių makrofagų
- Šviesių ląstelių inkstų karcinomos naviko endotelio ląstelės palaiko angiogenezę, yra heterogeniškos ir sudaro kelias populiacijas, tarp kurių ir viršūninio tipo ląstelės
- Žmogaus vaisiaus vandenyse aptinkama vaisiaus imuninių ir nuo plaučių, žarnyno ir inkstų atkibusių specializuotų ląstelių

METODAI

Bioetika

Visi eksperimentai su žmogaus ėminiais atlikti vadovaujantis Helsinkio Deklaracijos etikos standartais. Visi pacientai donavę mėginius, raštu išreiškė informuotą sutikimą. Plaučių karcinomos mėginiai surinkti operacijų Memorial Sloan Kettering Cancer Center (MSKCC) (Niujorkas, JAV) metu, vadovaujantis institucijos patvirtintu protokolu. Inkstų ir šviesių ląstelių inkstų karcinomos mėginiai surinkti Nacionaliniame vėžio institute (Vilnius, Lietuva), Vilniaus regioninio biomedicininių tyrimų komiteto leidimo Nr. 2019/2-1074-586. Vaisiaus vandenų mėginiai surinkti Vilniaus universiteto Santaros klinikose, Medicininės genetikos centre (Vilnius, Lietuva), Vilniaus regioninio biomedicininių tyrimų komiteto leidimo Nr. 2022/4-1429-900.

Mėginių surinkimas ir klinikiniai duomenys

Plaučių audiniai. Plaučių adenokarcinomos (n=2) ir plokščių ląstelių karcinomos (n=1) mėginai buvo paimti iš iki tol negydytų pacientų operacijos MSKCC metu. Gauta klinikinė informacija pateikta priede, lentelėje S1 (**Supplementary Table S1)**.

Inkstai ir šviesių ląstelių inkstų karcinoma. Švieži navikų (n=8) ir sveiko gretimo audinio (n=9) ėminiai rinkti iš iki tol sistemiškai negydytų pacientų, rezekcijos arba pilnos nefrektomijos metu (atvira arba laparoskopinė operacija) Nacionaliniame vėžio institute. Mėginiai laikyti lede ir skubiai (< 1 val.) transportuoti į laboratoriją disociacijai. Naviko mėginys T1 (iš paciento P1) pasižymėjo nekroze ir mažu gyvybingų ląstelių skaičiumi, todėl į analizę nebuvo įtrauktas. Klinikinė informacija pateikta priede, lentelėje S2 (**Supplementary Table S2)**.

Vaisiaus vandenys. Švieži vaisiaus vandenų ėminiai (n=26) surinkti amniocentezės metu Vilniaus universiteto Santaros klinikose, Medicininės genetikos centre. Amniocentezės procedūra atlikta dėl paskirtų genetinių vaisiaus tyrimų, o moksliniams tyrimams naudota 1,5-3 ml ėminio. Didžioji dalis tirtų ėminių neturėjo genetinių pakitimų. Šiame darbe tirtų mėginių klinikinė informacija pateikta priede, lentelėje S3 (**Supplementary Table S3**).

Mėginių paruošimas pavienių ląstelių RNR sekoskaitai

Plaučių karcinoma. Kiekvienas mėginys (n=3) buvo padalintas į tris ~5-10 mm^3 dalis. Tuomet audiniai buvo susmulkinti skalpeliu ir disocijuoti 15 min 37°C temperatūroje, GentleMACS Octo Dissociator with Heaters (Miltenyi)

prietaise, naudojant Human Tumor Dissociation Kit (Miltenyi Biotec, Kat. Nr. 130-095-929) fermentų kokteilį pagal gamintojo rekomendacijas. Po to, ląstelių suspensija buvo perleista per 35 µm Cell Strainer Snap Cap (TFS, Kat. Nr. 08-771-23) sietelį. Tuomet pašalinti eritrocitai, inkubuojant ląsteles eritrocitų lizės tirpale (ACK buffer, Lonza, Kat. Nr. BP10-548E) 2 minutes kambario temperatūroje. Vienas plaučių adenokarcinomos ląstelių mėginys buvo suspenduotas PBS (Gibco, Kat. Nr. 20012027) papildytame 0.04% (w/v) BSA (Roth, Kat. Nr. 8076.20) ir šviežias naudotas pavienių ląstelių inkapsuliacijai. Kitų mėginių ląstelių suspensijos buvo nudažytos gyvų ląstelių dažu kalceinu (Calcein AM, Invitrogen, Kat. Nr. C3009) bei su fikoeritrinu konjuguotu antikūnu pries žmogaus CD45 baltymą (BioLegend, Kat. Nr. 368510) ir BD FACS Aria II ląstelių sorteriu išskirstytos į CD45 teigiamas ir neigiamas frakcijas. Tuomet ląstelės centrifuguotos 5 min, 300g greičiu 4°C temperatūroje centrifugoje su svyruojančiais laikikliais, ir suspenduotos 90% metanolyje. Metanolyje užfiksuotos ląstelės perkeltos į −80°C šaldiklį. Po 30 dienų, šios archyvuotos ląstelės buvo naudojamos pavienių ląstelių inkapsuliacijai. Trumpai, mėgintuvėliai su metanoliu 15 min laikyti ant ledo, tuomet nucentrifuguoti 1000g greičiu, 10 min iki 4°C atvėsintoje centrifugoje su svyruojančiais laikikliais. Supernatantas pašalintas, paliekant apie 50 µl. tuomet, ląstelės suspenduotos 400 µl of ledo šaltumo Rehidratacijos buferyje 1 (sudėtis: 3× SSC (Invitrogen, Kat. Nr. 15557044), 80 mM DTT (TFS, Kat. Nr. R0861), 0.2% BSA, 1 U/µl RiboLock RNazių slopiklio (TFS, Kat. Nr. EO0381)) ir perkeltos ant filtro centrifugavimui (Millipore, Kat. Nr. UFC30DV25), prieš tai praplauto 1% BSA tirpalu. Filtrinė kolonėlė centrifuguota 50g greičiu 45 s, 4°C temperatūroje. Tuomet pratekėjęs tirpalas buvo pašalintas, o ant filtro likusios ląstelės (apie 50 µl tūrio) dar du kartus analogiškai praplautos Rehidratacijos buferiu 1, ir kartą Rehidratacijos buferiu 2 (sudėtis: 1× SSC, 40 mM DTT, 0.1% BSA, 1 U/µl RiboLock RNazių slopiklio). Galiausiai ląstelės surinktos nuo filtro, suskaičiuotos ant standartinio hemocitometro ir suspenduotos 1X DPBS (Gibco, Kat. Nr. 14190144) papildytame 0.04% (w/v) BSA ir 16% OptiPrep (Sigma-Aldrich, Kat. Nr. D1556).

Inkstai ir šviesių ląstelių inkstų karcinoma. Navikai buvo susmulkinti skalpeliu ir disocijuoti gentleMACS Octo Dissociator with Heaters (Miltenyi Biotec) instrumente naudojant Tumor Dissociation Kit (Miltenyi Biotec, Kat. Nr. 130-095-929) fermentų kokteilį pagal gamintojo instrukcijas. Gretimi sveiki audiniai disocijuoti Tissue Dissociation Kit I (Miltenyi Biotec, Kat. Nr. 130-110-201) kokteiliu pagal gamintojo instrukcijas. Po to pašalinti eritrocitai naudojant lizės reagentą (Miltenyi Biotec, Kat. Nr.130-094-183) pagal gamintojo instrukcijas. Tuomet ląstelės tris kartus praplautos ledo šaltumo 1X

DPBS, 500g greičiu 5 min, iki 4°C atvėsintoje centrifugoje. Ląstelių kiekis ir gyvybingumas įvertintas naudojant Trypan Blue dažą (Gibco, Kat. Nr. 15250061) ir standartinį hemocitometrą. Joks ląstelių praturtinimas nebuvo atliktas, ir ląstelės suspenduotos 1X DPBS papildytame 0.04% (w/v) BSA ir 15% OptiPrep.

Vaisiaus vandenys. Mėginiai buvo gauti ant ledo, ir iškart nucentrifuguoti 300g greičiu 5 minutes iki 4°C atvėsintoje centrifugoje. Supernatantas perkeltas į Protein LoBind mėgintuvėlius (Fisher scientific, Kat. Nr. 05414206) ir saugojamas -80°C šaldiklyje. Ląstelės du kartus 1 ml 1X DPBS papildyto 0.04% BSA centrifuguojant 300g greičiu, 5 min, 4°C temperatūroje. Ląstelių skaičius ir gyvybingumas įvertintas naudojant Trypan Blue dažą vaizdinant ant standartinio hemocitometro. Galiausiai, inkapsuliacijai, ląstelės suspenduotos 1X DPBS papildyto 0.04% BSA ir 10% 500K MW Dekstrano (SERVA Feinbiochemica, Kat. Nr. 18695) (mėginiai F1-4), arba 15% OptiPrep (mėginiai F10-17), arba 0.02% ksantano dervos (Sigma Aldrich, Kat. Nr. G1253) (mėginiai F29-50).

Visiems RNR barkodavimo eksperimentams ląstelės praskiestos iki 400 tūkst./ml galutinės koncentracijos, kad inkapsuliacijos λ vertė siektų ~0.2.

Pavienių ląstelių RNR sekoskaita

Lašeliais paremtos inDrops platformos naudojimas pavienių ląstelių sekoskaitai detaliai aprašytas Zilionis et al. (62) ir Juzenas et al. (60), ir atliktas vadovaujantis šiais šaltiniais. Trumpai, procedūra susideda iš ląstelių ir reagentų paruošimo, inkapsuliacijos į nanolitrų tūrio lašelius, atvirkštinės transkripcijos (AT) reakcijos lašeliuose ir sekoskaitos bibliotekų paruošimo.

Barkoduojančių rutuliukų dizainas ir paruošimas

Barkoduojančių hidrogelio rutuliukų, nešančių AT pradmenis, sintezė ir paruošimas detaliai aprašytas Zilionis et al. (62). Prieš inkapsuliaciją, rutuliukai penkis kartus praplauti 1 ml rutuliukų plovimo tirpale (sudėtis: 1X Maxima H minus RT buffer (TFS, Kat. Nr. EP0751), 1% Igepal CA-630 (Sigma-Aldrich, Kat. Nr. I8896-50ML)) naudojant stalinę centrifugą. Tuomet rutuliukai sukoncentruoti pašalinant kiek įmanoma daugiau supernatanto ir patalpinti į 0.56 mm vidinio skersmens PTFE mikroskysčių žarnelę (Atrandi Biosciences, Kat. Nr. MAN-TUB2), kuri užmaunama ant 1ml švirkšto (Fisher Scientific, Kat. Nr. 1482330) pripildyto 500 µl HFE-7500 alyvos (nuo šiol minima kaip alyva; 3M, Kat. Nr. 98-0212-2929-3). Žarnelė su rutuliukais papildomai apsaugota nuo šviesos įmaunant į šviesiai nepralaidžią didesnio diametro žarnelę. Šiame darbe naudoti hidrogeliniai rutuliukai, prie kurių

prikabinti skirting dizaino AT pradmenys, pateikti **1 lentelėje**. inDrops-2 TS-v2020 dizaino rutuliukai įsigyti iš kompanijos Atrandi Biosciences (Kat. Nr. DG-BHB-C).

1 lentelė. AT reakcijos mRNR sugavimo pradmenys ir matrices keitimo pradmuo (TSO). Pabraukta seka žymi T7 promotorių; skaičiai žymi ląstelių barkodo seką.

Pavadinimas	Kam naudota	Seka (5'→3')
inDrops-2 *in vitro* transkripcijos (IVT) ir matricos keitimo (TS) palyginimas	plaučių adenokarcinoma (IVT vs TS)	CGATGACGTAATACGACTCACTATAGGG ATACCACCATGGCTTTCCCTACACGACGCTCT TCCGATCT[12345678901]GAGTGATTGCTTGTG ACGCCAA[12345678]NNNNNNNN TTTTTTTTTTTTTTTTTTV; Antrojo AT žingsnio pradmuo IVT protokolui: GTGACTGGAGTTCAGACGTGTGCTCTTCCGA TCTNNNNNN
inDrops-2 TS_v1	plaučių karcinoma	CTACACGACGCTCTTCCGATCT[12345678]CAT G[12345678]NNNNNNNN TTTTTTTTTTTTTTTTTT
inDrops-2 TS_v2020	inkstai, šviesių ląstelių inkstų karcinoma, vaisiaus vandenys	TACGGCGACCACCGAGATCTACAC[12345678] ACACTCTTTCCCTACACG[12345678]NNNNNN TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
TSO	visi mėginiai	AAGCAGTGGTATCAACGCAGAGTACATrGrGr G

Atvirkštinės transkripcijos mišinio paruošimas

Ląstelių barkodavimo eksperimentams buvo paruoštas atvirkštinės transkripcijos reakcijos mišinys, kurį sudarė 1X AT buferis, 25μM TSO (**1 lentelė**), 0.5 mM dNTP mišinio (Thermo Scientific, Kat. Nr. R0192), 10 U/μl Maxima H Minus AT fermento (Thermo Scientific, Kat. Nr. EP0751), 1 U/μl RiboLock RNazių slopiklio (Thermo Scientific, Kat. Nr. EO0382) ir 0.3% Igepal CA-630 (Sigma-Aldrich, Kat. Nr. I8896-50ML). Pateikta galutinė koncentracija lašelyje.

inDrops eiga

Po to kai buvo paruoštos ląstelės, barkoduojantys rutuliukai ir AT mišinys, ląstelių inkapsuliacija atlikta naudojant mikroskysčių lustą (Atrandi Biosciences, Kat. Nr. MCN-C5), kurio schema pateikta **Figure 2.1**. Ląstelių suspensija ir AT mišinys buvo suleisti į atšaldytus 1ml švirkštus ant 500 μl alyvos, įstatyti į infuzines pompas ir prijungti prie lusto mikroskysčių žarnele.

1 ml švirkštas buvo užpildytas lašelius stabilizuojančia alyva (Atrandi Biosciences, Kat. Nr. MON-DSO2), įstatytas į infuzinę pompą ir taip pat prijungtas prie lusto. Kai visi reikalingi reagentai buvo sujungti į lustą, inkapsuliacija vykdyta šiais tėkmės greičiais: ląstelės ir AT mišinys – po 250 µl/val; barkoduojantys rutuliukai – 100-150 µl/val, kontroliuojant, kad būtų aptinkami 80-90% lašelių; lašelius stabilizuojanti alyva – 700 µl/val. AT mišinys ir ląstelių suspensija inkapsuliacijos metu buvo šaldomi ledo pripildytomis guminėmis pirštinėmis. Procesas buvo stebimas mikroskopu (Nikon Eclipse Ti) naudojant Phantom greitaeigę kamerą. Emulsija buvo renkama į 1.5 ml Eppendorf DNA LoBind mėgintuvėlį (Fisher Scientific, Kat. Nr. 13698791) ant ledo.

Atvirkštinė transkripcija

Barkoduojantys pradmenys buvo nukabinti nuo rutuliukų apšvietus mėgintuvėlį su emulsija 350nm šviesa, naudojant arba LED įrenginį (Droplet Genomics, MHT-LAS1) 20 sekundžių, arba UV lempą (UVP, cat. no. 95–0127-01) 5 minutes. Surinktos emulsijos buvo išdalintos į PGR mėgintuvėlius po 1000-5000 ląstelių. Tuomet mėgintuvėliai perkelti į termociklerį AT reakcijai: IVT ir TS metodų palyginimui, AT vykdyta 42°C temperatūroje 90 min; visais kitais atvejais – 42°C temperatūroje 60 min, po to 5 min 85°C temperatūroje. Tokia kDNR emulsija gali būti saugojama -20°C temperatūroje iki sekoskaitos bibliotekų ruošimo ilgą laiką.

Sekoskaitos bibliotekų paruošimas

Eksperimento, kuriame lyginti IVT ir TS paremti sekoskaitos bibliotekos paruošimo metodai, detalus aprašymas pateiktas Juzenas et al. (60). Šviesių ląstelių inkstų karcinomos, inkstų ir vaisiaus vandenų sekoskaitos bibliotekų paruošimo eiga pateikta žemiau. Naudoti pradmenys pateikti **2 lentelėje**. Emulsijos buvo suardomos pridedant iki 10% (v/v) emulsijos suardymo reagento (Atrandi Biosciences, Kat. Nr. MON-EB1), ir trumpai nusukus 30 s 300g greičiu, supernatantas buvo perneštas ant kolonėlės filtro (Zymo, Kat. Nr. C1004-250). Pro filtrą praėjęs tirpalas, kuriame yra kDNR, buvo surenkamas į naują DNA LoBind mėgintuvėlį, centrifuguojant 1 min, 10 000g greičiu. Barkoduota kDNR buvo du kartus gryninta 0.8X AMPure XP magnetiniais rutuliukais (BeckMan Coulter, Kat. Nr. A63881) pagal gamintojo instrukcijas. Tuomet kDNR buvo padauginta PGR reakcijos metu, naudojant KAPA HiFi Hot Start Ready Mix (Roche, Kat. Nr. KK2601) reagentus ir 0.5µM koncentracijos pradmenis. PGR temperatūrinis režimas pateiktas **3 lentelėje**. DNR fragmentavimui ir ligavimui naudoti reagentai ir instrukcijos iš NEBNext® Ultra™ II FS DNA Library Prep Kit (NEB, Kat.

Nr. E7805S) rinkinio. Padauginta DNR buvo fragmentuojama naudojant using NEBNext Ultra II FS reakcijos buferį ir fermentų mišinį, 8 min 37°C temperatūroje, po kurios sekė 30 min inkubacija 65°C temperatūroje. Po to fragmentuota DNR buvo išgryninta naudojant abipusę fragmentų atranką 0.6X-0.8X AMPure magnetiniais rutuliukais. Adapterių ligavimas atliktas naudojant NEBNext Ultra II Ligation Master Mix ir Enhancer, bei 0.05 μM galutinės koncentracijos ligavimo adapterį (seka pateikta **2 lentelėje**), 15 min 20°C temperatūroje. Tuomet produktas išgrynintas 0.8X AMPure. Galiausiai, bibliotekos padaugintos atliekant indeksavimo PGR reakciją (**4 lentelė**), kurią sudarė 1X KAPA HiFi Hot Start Ready Mix (Roche, Kat. Nr. KK2601) bei 0.5 μM koncentracijos p5 ir p7 indeksai (**2 lentelė**). Reakcijos produktas išgrynintas atliekant abipusę atranką 0.6X-0.8X AMPure magnetiniais rutuliukais ir eliuotas į vandenį. Bibliotekų kokybė po pirmosios PGR reakcijos ir po indeksavimo PGR buvo vertinama naudojant Bioanalyzer DNA High Sensitivity lustą (Agilent, Kat. Nr. 50674626).

2 lentelė. Sekoskaitos bibliotekų ruošimo pradmenys. *- nurodo fosfotioatinę jungtį. REV – atvirkštinis, FWD – tiesioginis.

Pavadinimas	**Seka (5'→3')**
kDNR padauginimo pradmenys	
REV cDNA pradmuo	AAGCAGTGGTATCAACGCAGAG
FWD cDNA pradmuo	TACGGCGACCACCGAGATC
Ligavimo adapteris	
Ligavimo adapteris	/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGT CAC/3ddC /5AmMC6/GCTCTTCCGATCT
Indeksavimo PGR pradmenys	
FWD PCR indekso pradmuo p5	AATGATACGGCGACCACCGAGATCTACA*C
p7 indeksas 1	CAAGCAGAAGACGGCATACGAGAT AACCTG GTGACTGGAGTTCAGACGTG*T
p7 indeksas 2	CAAGCAGAAGACGGCATACGAGAT CCAAGT GTGACTGGAGTTCAGACGTG*T
p7 indeksas 3	CAAGCAGAAGACGGCATACGAGAT GGTTCA GTGACTGGAGTTCAGACGTG*T
p7 indeksas 4	CAAGCAGAAGACGGCATACGAGAT TTGGAC GTGACTGGAGTTCAGACGTG*T
p7 indeksas 5	CAAGCAGAAGACGGCATACGAGAT ACCACT GTGACTGGAGTTCAGACGTG*T

p7 indeksas 6	CAAGCAGAAGACGGCATACGAGAT CAGTGG GTGACTGGAGTTCAGACGTG*T
p7 indeksas 7	CAAGCAGAAGACGGCATACGAGAT GTTGTC GTGACTGGAGTTCAGACGTG*T
p7 indeksas 8	CAAGCAGAAGACGGCATACGAGAT TGACAA GTGACTGGAGTTCAGACGTG*T

3 lentelė. kDNR padauginimo PGR temperatūrinis režimas.

Etapas	**Temperatūra**	**Laikas**
Pirminė denatūracija	98°C	00:03:00
Denatūracija	98°C	00:00:15
Pradmenų prisijungimas	67°C	00:00:20
DNR Sintezė	72°C	00:01:00
Grįžti į 2 žingsnį, 15 ciklų (iš viso 16)		
Baigiamoji sintezė	72°C	00:01:00
Laikymas	4°C	∞

4 lentelė. Indeksavimo PGR temperatūrinis režimas.

Etapas	**Temperatūra**	**Laikas**
Pirminė denatūracija	98°C	00:00:45
Denatūracija	98°C	00:00:20
Pradmenų prisijungimas	54°C	00:00:30
DNR Sintezė	72°C	00:00:20
Grįžti į 2 žingsnį, 10 ciklų (iš viso 11)		
Baigiamoji sintezė	72°C	00:01:00
Laikymas	4°C	∞

Sekoskaita

Galutinės inDrops-2 (IVT) bibliotekos sekoskaita atlikta naudojant NextSeq550 and HiSeq2500 (Illumina) instrumentą (1 nuskaitymas: 54 ciklai; i7: 8 ciklai, 2 nuskaitymas: 35 ciklai ar daugiau). inDrop-2 (TS) bibliotekų sekoskaita atlikta naudojant MiSeq, HiSeq2500, NextSeq550 ir NovaSeq6000 (Illumina) instrumentus, be PhiX sekų. Sekoskaitos parametrai buvo: 1 nuskaitymas: 28 ciklai; i7 nuskaitymas: 8 ciklai, 2 nuskaitymas: tarp 35 ir 92 ciklų (priklausomai nuo sekoskaitos instrumento ir reagentų rinkinio).

Inkstų ir šviesių ląstelių karcinomos mėginių sekoskaita atlikta Illumina NextSeq 550 instrumentu, keliais pakartojimais naudojant arba NextSeq 500/550 High Output Kit v2.5 (75 ciklai) (Illumina, Kat. Nr. 20024906) reagentų rinkinį, arba NextSeq 500/550 High Output Kit v2.5 (150 ciklų) (Illumina, Kat. Nr. 20024907) rinkinį. Kadangi eksperimentams naudotos dvi barkoduojančių rutuliukų versijos, ir sekoskaita vykdyta skirtingai, pirmajai

versijai nuskaitymų ilgiai buvo: 1 nuskaitymas: 51 ciklas, 2 nuskaitymas: 35 ciklai, i7 nuskaitymas: 6 ciklai. Antrajai versijai – 1 nuskaitymas: 16 ciklų, 2 nuskaitymas: 62 ciklai, i7 nuskaitymas: 6 ciklai ir i5 nuskaitymas: 8 ciklai.

Vaisiaus vandenų bibliotekų sekoskaita atlikta naudojant Illumina NextSeq 2000 instrumentą ir NextSeq™ 1000/2000 P2 XLEAP-SBS™ Reagent Kit (100 ciklų) reagentus su šiais nustatymais: 1 nuskaitymas: 16 ciklų, 2 nuskaitymas: 108 ciklai, i7 nuskaitymas: 6 ciklai, i5 nuskaitymas: 8 ciklai.

Duomenų analizė

Pirminis duomenų apdorojimas

Pirminio sekoskaitos duomenų apdorojimo tikslas – barkodų x genų matricos sugeneravimas, kurioje nurodytos kiekvieno geno raiškos vertės kiekvienam iš aptiktų ląstelės barkodų. Šis procesas susideda iš: 1) nuskaitymų surūšiavimo pagal ląstelės barkodus ir mėginio indeksus (*angl. demultiplexing*), taip pat atliekant barkodų korekciją; 2) nuskaitymų palyginimo su atskaitos genomu; 3) aptiktų genų priskyrimo nuskaitymams; ir 4) UMI deduplikacija ir kiekybinis įvertinimas. Šiame darbe naudota solo-in-drops duomenų operacijų grandinė, (https://github.com/jsimonas/solo-in-drops), kuri matricos konstravimui naudoja STARsolo (79) algoritmą. MultiQC įrankis naudotas įvertinti sekoskaitos *fastq* failų kokybę bei STARsolo išdavos statistikos suvestinę.

inDrops-2 IVT ir TS metodų palyginimo eksperimento duomenų apdorojimui, STAR v2.7.10a algoritmas naudotas palyginti nuskaitymus su GRCh38 žmogaus genomu (GENCODE v41 anotacija). STARsolo parametrai buvo: --soloFeatures *GeneFull*; --soloType *CB_UMI_Complex*, --soloCBmatchWLtype *EditDist_2*, --soloUMIdedup *Exact.*

Metanolyje fiksuotų ir chemiškai mėginio barkodais žymėtų plaučių karcinomos bibliotekų sekoskaitos duomenų apdorojimui naudota SEQC (6) ir CITE-seq-Count (303) algoritmų kombinacija. SEQC naudotas su numatytaisiais parametrais sugeneruoti barkodų x genų matricas, o CITE-seq-Count algoritmas, taip pat numatytaisiais parametrais ir --*no_umi_correction* naudotas kiekybiškai įvertinti cheminio žymėjimo sekas.

Inkstų ir ccRCC sekoskaitos duomenims, STAR v2.7.6a algoritmas naudotas su šiais parametrais: --soloMultiMappers *Uniform*, --soloType *CB_UMI_Simple*, --soloUMIfiltering *MultiGeneUMI*, ir --soloCBmatchWLtype *1MM*. Sekų palyginimui naudotas žmogaus genomas GRCh38 su Ensembl v93 anotacija.

Vaisiaus vandenų sekoskaitos duomenų apdorojimui, STAR v2.7.10a naudotas palyginti nuskaitymus su GRCh38 žmogaus genomu (GENCODE v41 anotacija). STARsolo algoritmas naudotas su šiais parametrais: --soloMultiMappers *Uniform*, --soloType *CB_UMI_Complex*, --soloUMIfiltering *MultiGeneUMI*, --soloUMIdedup *Exact* ir --soloCBmatchWLtype *EditDist_2*.

Kokybės kontrolė ir hibridinių transkriptomų pašalinimas

Sukonstravus barkodų x genų matricas, visa tolesnė analizė atlikta programuojant Python kalba ir naudojantis scanpy (73) paketo įrankių rinkiniu. Trumpai, žemos kokybės ląstelės buvo pašalintos taikant slenkstines vertes pagal bendrą transkriptų kiekį (UMI) barkodui ir mitochondrinių genų transkriptų dalį, įvertinus šių verčių pasiskirstymą, kaip aprašyta ir parodyta literatūros apžvalgoje (**Figure 1.3)**. IVT ir TS metodo palyginimui, plaučių adenokarcinomos mėginiui šios vertės buvo bent 400 UMI per ląstelę ir ne daugiau kaip 15% mitochondrinių genų transkriptų. Plaučių karcinomos mėginiams, barkodų filtravimas buvo atliktas automatiškai dar duomenų apdorojimo metu, o hibridiniai transkriptomai (įvykiai, kai į lašelį inkapsuliuojama daugiau nei viena ląstelė) pašalinti įvertinus ar ląstelės barkodui priskirti nuskaitymai turėjo daugiau nei vieną mėginiui specifinį barkodą (*angl. hashtag*). Šis filtravimas atliktas HashSolo (304) algoritmu. Inkstų ir šviesių ląstelių inkstų karcinomos mėginiams, slenkstinės UMI ir mitochondrinių genų transkriptų dalies vertės buvo 400 UMI ir 20%, išskyrus bibliotekas T3.1, T9.1, N3.3, N4.3, N2.3, kur taikyta 300 UMI slenkstinė vertė. Vaisiaus vandenų mėginiams (iš viso 61 sekoskaitos biblioteka), naudotos 15% mitochondrinių genų transkriptų ir 400 UMI slenkstinės vertės, išskyrus tam tikras bibliotekas (pavadinimai sutartiniai): F10.1, F15.2x2 – 500 UMI, F34.3 – 800 UMI, ir F34.2, F34, F47.1, F49.1, F49.2, F50.2, F13_1x3, F32, F29.2x2, F31.1_2x2, F46.1x2, F48.2x2 – 1000 UMI per ląstelę.

Hibridiniai transkriptomai (*angl. doublets*) susidaro kai į lašelį atsitiktinai inkapsuliuojama daugiau nei viena ląstelė, o visi transkriptai pažymimi tuo pačiu ląstelės barkodu. Tokių nepageidaujamų transkriptomų identifikavimui ir pašalinimui naudotas Scrublet (v0.2.3) (97) algoritmas, kuriam naudotos tos pačios principinės komponentės (iš principinių komponenčių analizės, PCA) kaip ir pirminiam grafo ir UMAP projekcijos sudarymui (bendri šio proceso žingsniai aprašyti kitame skyrelyje). Scrublet naudotas kiekvienai emulsijai atskirai. Procesas susidėjo iš šių žingsnių: 1) naudojant Scrublet apskaičiuotas kiekvieno barkodo hibridinio transkriptomo tikimybės rodiklis (*angl. doublet score*); 2) atliktas barkodų grupavimas naudojant Louvain algoritmą su labai

didele rezoliucijos parametro verte (*resolution*=60 inkstams ir ccRCC, *resolution*=40 vaisiaus vandenims); 3) įvertintas kiekvieno klasterio bendras tikimybės rodiklis ir potencialių hibridų dalis; 4) interaktyvioje SPRING aplikacijoje (119) rankiniu būdu patikrinta genų raiška aukščiausias šių rodiklių vertes turinčiuose klasteriuose ir 5) patvirtinus transkriptomus kaip hibridinius, tokie klasteriai pašalinti. Inkstų ir ccRCC atlasui, procedūra pakartota du kartus ir pašalinta 913 ląstelių. Papildomai, šiam atlasui pašalinti barkodai, turintys daugiau nei 1% hemoglobino genų (*HBB*, *HBA1*, *HBA2*, *HBD*) transkriptų (47 ląstelės). Vaisiaus vandenų atlasui, hibridinių transkriptomų pašalinimo procedūra atlikta vieną kartą ir pašalintos 325 ląstelės, taip pat pašalintas žemos kokybės transkriptomų klasteris, kurį sudarė 843 ląstelės. Plaučių adenokarcinomos duomenims, naudotiems IVT ir TS metodų palyginimui, hibridinių transkriptomų šalinimas neatliktas. Plaučių karcinomos mėginiams, kaip minėta, tai atlikta naudojant HashSolo.

Duomenų paruošimas vizualizacijai ir klasterizavimas

Čia aprašyta bendra daugiadimensių pavienių ląstelių sekoskaitos duomenų atvaizdavimo UMAP algoritmu procedūra, o konkretūs mažo dimensiškumo projekcijų sudarymui naudoti parametrai pateikti **5 lentelėje**. Šis procesas kartojamas keletą kartų su skirtingomis parametrų vertėmis, kol surandami konkrečiam duomenų rinkiniui geriausiai tinkantys parametrai. Nepakeičiamas įrankis šiam darbui buvo interaktyvi duomenų analizės platforma SPRING (119), kurioje galima patogiai įvertinti vizualizacijos kokybę, klasterizavimo rezultatus, konkrečių žymenų raišką, atlikti diferencinės raiškos analizę ir kita. Toks duomenų tyrinėjimas suteikia daug žinių ir leidžia parinkti geriausiai duomenis atspindinčius parametrus.

Trumpai, išvalytos daugiadimensės duomenų matricos buvo ruošiamos atvaizdavimui UMAP algoritmu. Šis procesas susidėjo iš: 1) transkriptų kiekio normalizavimo iki 10 000 (CPTT), logaritminės transformacijos ir standartizavimo z transformacija; 2) variabilių genų parinkimo pagal Fano faktorių kaip aprašyta Klein et al. (2); 3) principinių komponenčių analizės; 4) eksperimentinių artefaktų pašalinimo (*angl. batch correction*) naudojant Harmony (115) algoritmą; 5) k-artimiausių kaimynų grafo sudarymo ir 6) UMAP atvaizdavimo. Grafo sudarymui atlikta genų atranka: po CPTT normalizacijos, atrinkti genai turintys bent *n_counts* transkriptų ne mažiau kaip *n_cells* ląstelėse, ir pašalinti mitochondriniai ir ribosominiai genai. Toliau, pagal Fano faktorių, parinktas konkretus skaičius *n_var* labiausiai variabilių genų, kurie naudoti principinių komponenčių analizei. Ši analizė kartota 10 kartų, vis sumaišant duomenis – tokiu būdu nustatyta, koks kiekis

principinių komponenčių, *num_PCs*, paaiškina daugiau nei atsitiktinę variaciją duomenyse, kaip aprašyta (2). Techninių artefaktų pašalinimui naudota scanpy.external.pp.harmony_integrate() funkcija, nurodant techninį aspektą kaip *batch_variable*. Tuomet, k-artimiausių kaimynų grafas sudarytas naudojant funkciją sc.pp.neighbors() parenkant k skaičių (*n_neighbors*) ir atvaizduotas UMAP naudojant sc.tl.umap() funkciją su parametru *min_dist*.

Grafo klasterizavimas atliktas naudojant arba spektrinio klasterizavimo algoritmą (sklearn.cluster.SpectralClustering() funkcija), arba scanpy paketo Leiden algoritmą (sc.tl.leiden() funkcija), arba PhenoGraph Leiden algoritmą (sc.external.tl.phenograph() funkcija). Spektrinio klasterizavimo algoritmas padalina grafą į iš anksto nustatytą grupių skaičių, o Leiden algoritmo grupių kiekį kontroliuoja rezoliucijos parametras. Abi šios vertės **5 lentelėje** pateiktos skiltyje *resolution*.

5 lentelė. Grafo sudarymo, UMAP vizualizacijos ir klasterizavimo parametrai.

Pavadinimas	***n_counts***	***n_cells***	***n_var***	***num_PCs***	***batch_variable***	***n_neighbors***	***min_dist***	***Clustering approach***	***resolution***
Plaučių adenokarcinoma IVT vs TS **Figure 3.4**	10	10	2000	28	library	20	0.5	leiden	0.6
Plaučių karcinoma **Figure 3.5**	10	10	2000	55	-	30	0.5	PhenoGraph leiden	2
Plaučių neimuninės last. **Figure 3.6**	5	10	2000	48	-	50	0.3	PhenoGraph leiden	0.6
Plaučių mieloidinės last. **Figure 3.7**	5	10	2000	36	-	50	0.3	PhenoGraph leiden	0.5
Plaučių limfoidinės last. **Figure 3.8**	5	10	2000	21	-	50	0.3	PhenoGraph leiden	0.8
Inkstai ir ccRCC **Figure 3.10**	15	25	2000	71	beads	30	0.4	spectral	43
Vaisiaus vandenys, visos last.	10	20	2000	128	library	30	0.5	leiden	1

Figure 3.21									
Vaisiaus vandenų imuninės last. **Figure 3.22**	10	10	2000	43	library	20	0.6	spectral	16
Vaisiaus vandenų neimuninės last. **Figure 3.24**	10	10	2000	151	library	20	0.5	spectral	20

Diferencinės raiškos analizė ir ląstelių anotavimas

Siekiant nustatyti konkrečiai ląstelių grupei (klasteriui) būdingą genų raišką (*angl. marker genes*) atlikta diferencinės raiškos analizė, lyginant pasirinktą klasterį su visomis likusiomis ląstelėmis (Mann Whitney U testas su Bonferoni-Hochberg p verčių korekcijos procedūra). Prieš statistinį testą, genai filtruoti pagal raišką naudojant tas pačias *n_counts* ir *n_cells* vertes kaip ir grafo sudarymui (**5 lentelė**). Vaisiaus vandenų analizei, taip pat pašalinti mitochondriniai genai. Labiausiai praturtinti 50 genų kiekvienoje grupėje (koreguota p vertė < 0.05) naudoti ląstelių fenotipo nustatymui, atliekant išsamią literatūros analizę.

CellTypist automatinis anotavimas

Inkstų ir ccRCC duomenims, automatinio transkriptomų anotavimo algoritmas CellTypist (137) naudotas įvertinti priskirtų ląstelių fenotipų panašumą publikuotiems. Tam atlikti, viešai prieinami duomenys naudoti CellTypist modelių apmokymui didelio našumo skaičiavimo kompiuteryje, pagal instrukcijas pateiktas https://www.celltypist.org/. Endotelio ląstelių įvertinimui pagal Goveia et al. (305), šioje publikacijoje pateiktų endotelio ląstelių genų raiškos matrica atsisiųsta iš https://endotheliomics.shinyapps.io/lung_ectax/, tuomet atlikta normalizacija ir logaritminė transformacija, pašalintos ne naviko endotelio ir vienam pacientui specifinės ląstelės. Tuomet, modelis apmokytas ant šių duomenų be genų filtravimo, ir anotacijos priskirtos mūsų naviko endotelio log-transformuotiems ir normalizuotiems duomenims naudojant funkciją celltypist.annotate() ir parametrą *majority_voting=True*. Analogiškai, kitas modelis apmokytas Zhang et al. (222) publikacijos duomenimis, juos atsisiuntus iš GEO (GSE159115). Šie duomenys buvo apdoroti pašalinant ne inkstų ar ccRCC epitelio ląsteles ir panaudoti modelio apmokymui be genų filtravimo. Tuomet modelis naudotas priskirti anotacijas inkstų ir ccRCC epitelinėms ląstelėms, naudojant log-normalizuotą genų raiškos matricą. Vaisiaus vandenų imuninių ląstelių anotavimui, CellTypist modeliai buvo

apmokyti ir naudoti analogiškai, naudojant viešai prieinamus duomenis iš žarnyno atlaso (306) ir fetalinių plaučių leukocitų (307).

Genų rinkinių praturtinimo analizė

Inkstų ir ccRCC duomenims, genų rinkinių praturtinimo analizė (*angl. gene set over-representation analysis*) atlikta vertinant 100 genų turėjusių aukščiausias diferencinės genų raiškos analizės vertes (*angl. fold-change*) buvimą Hallmark Pathways genų rinkiniuose iš MSigDB v7.5.1 duomenų bazės. Atliktas hipergeometrinis testas naudojant enrichGO() funkciją iš clusterProfiler R paketo, pasirinkus visus ekspresuotus (>0 UMI) genus kaip foną (*angl. background reference*). Signaliniai keliai, kurių praturtinimo p-vertės po Benjamini-Hochberg korekcijos buvo <0,05, laikyti reikšmingai praturtintais.

Vaisiaus vandenų neimuninių ląstelių genų rinkinių praturtinimo analizei naudoti top 200 diferencinės raiškos genų, kurių įsitraukimas biologiniuose procesuose ir signaliniuose keliuose vertintas keliose duomenų bazėse, Gene Ontology Biological Process 2023, MSigDB Hallmark 2020 ir Reactome 2022, prieinamose gseapy pakete (308) naudojant funkciją gseapy.enrichr(). Ši funkcija taip pat atlieka hipergeometrinį testą su Benjamini-Hochberg korekcija, naudojant visus genų rinkiniuose esančius genus kaip foną. Praturtinti biologiniai procesai ir signaliniai keliai pasirinkti pagal koreguotą p vertę <0,05 buvo atvaizduoti naudojant gseapy dotplot() funkciją.

Kiekybinis heterogeniškumo įvertinimas

Inkstų ir šviesių ląstelių inkstų karcinomos duomenų rinkiniui, mėginių fenotipų heterogeniškumas buvo įvertintas kiekybiškai pagal Shannon entropiją, kaip aprašyta Chan et al. (309). Trumpai, kiekvienoje ląstelių grupėje (stromos, endotelio, vėžinių, limfoidinių, mieloidinių, epitelinių ir proliferuojančių ląstelių) apskaičiuotos entropijos vertės nurodančios, kiek skirtingų mėginių ląstelių šiuos fenotipus sudaro. Siekiant atsižvelgti į ląstelių kiekio skirtumus grupėse, iš kiekvienos grupės 100 kartų buvo paimta po 100 ląstelių atsitiktiniu būdu su pakeitimu, ir Shannon entropija buvo apskaičiuota naudojant scipy.stats.entropy() funkciją iš scipy paketo. Ląstelės anotuotos kaip “Tumor cells 1” nevertintos, nes buvo specifinės vienam pacientui.

Receptorių – ligandų sąveikos analizė

Inkstų ir ccRCC duomenų rinkiniui, tarpląstelinės sąveikos buvo numatomos naudojant CellphoneDB v.2.0.0 (156) duomenų bazę ir metodą “statistical_analysis” su numatytaisiais parametrais. Duomenys filtruoti pašalinant sveiko audinio epitelio ir proliferuojančias ląsteles, ir normalizuota

bei log-transformuota likusių ląstelių matrica naudota tarpląstelinių sąveikų prognozavimui. Rezultatai peržiūrėti rankiniu būdu ir pasirinktos reikšmingos (p vertė < 0.05) sąveikos vizualizuotos (**Figures 3.14, A; 3.16, A; 3.20, A, Supplementary Figure S3**). Išgyvenamumo analizei (kaip parodyta **Figure 3.14, B**) naudoti ląstelių tarpusavio sąveikos genų rinkiniai (*angl. signatures*) sudaryti imant receptorių ir ligandų genus.

Išgyvenamumo analizė

Išgyvenamumo analizė pasirinktiems inkstų ir ccRCC genų rinkiniams atlikta naudojant TCGAbiolinks R paketą ir TCGA KIRC kohortos RNR sekoskaitos duomenis (normalizuoti viršutinės kvartilės FPKM) bei klinikinę informaciją, gautą iš NCI GDC Duomenų portalo (310). Pasirinkto genų rinkinio praturtinimo vertė šiuose duomenyse apskaičiuota kaip z-transformuotų raiškos verčių vidurkis. Ryšys tarp genų rinkinio raiškos ir išgyvenamumo nustatytas atliekant Kaplan-Meier ir daugiamatės Cox regresijos analizę. Logaritminio rango kriterijus ir Wald testas, atitinkamai, naudoti nustatyti statistiniam reikšmingumui. Kaplan-Meier analizei, genų rinkinių mėginiuose raiška stratifikuota į aukštą (daugiau arba lygu raiškos medianai) ir žemą (mažiau nei raiškos mediana). Daugiamatės Cox regresijos analizei naudotos tolydžios genų rinkinio raiškos vertės, nurodant amžių ir lytį kaip papildomus kintamuosius. Išgyvenamumo analizė atlikta naudojant survival ir survminer R paketus.

Duomenų ir kodo prieinamumas

Pavienių ląstelių RNR sekoskaitos duomenys sugeneruoti inDrops-2 metodo kūrime (plaučių karcinoma) (60) patalpinti į Europos nukleotidų archyvą (ENA), pasiekiami pagal registracijos numerį PRJEB71611. Pirminiai ir išanalizuoti inkstų ir šviesių ląstelių inkstų karcinomos pavienių ląstelių sekoskaitos duomenys, publikuoti Zvirblyte et al. (311) patalpinti į archyvą Gene Expression Omnibus (GEO), pasiekiami numeriu GSE242299. Viešai prieinami duomenys, naudoti inkstų atlaso analizei buvo parsisiųsti iš GEO (GSE159115) ir https://endotheliomics.shinyapps.io/lung_ectax/. Vaisiaus vandenų duomenys bus patalpinti į vieną iš archyvų po publikavimo. Viešai prieinami duomenys, naudoti vaisiaus vandenų analizei, atsiųsti iš https://www.gutcellatlas.org/ ir https://fetal-lung-immune.cellgeni.sanger.ac.uk/.

Kodas naudotas inkstų ir ccRCC analizei bei vizualizavimui pateiktas Jupyter notebook formatu https://github.com/zvirblyte/2023_ccRCC. Didžioji dalis Python kodo buvo paimta ir pritaikyta analizei iš

https://github.com/AllonMKlein/Pfirschke_et_al_2021 (8). Šis kodas taip pat naudotas plaučių karcinomos ir vaisiaus vandenų analizei.

REZULTATAI

Šioje disertacijoje pateikti rezultatai susideda iš trijų dalių, vienijamų pavienių ląstelių RNR sekoskaitos metodo naudojimo. Pirmoje dalyje pristatoma patobulinta pavienių ląstelių RNR sekoskaitos platforma inDrops-2 ir pademonstruojamas jos tinkamumas klinikinių mėginių tyrimams. Antroje dalyje, metodas naudojamas išsamiai ištirti inkstų ir šviesių ląstelių inkstų karcinomos mėginius vienos ląstelės lygmeniu. Paskutinėje dalyje pristatomas žmogaus vaisiaus vandenų ląstelių transkriptomų atlasas.

InDrops-2: patobulintas pavienių ląstelių RNR sekoskaitos metodas

Komercinės pavienių ląstelių sekoskaitos sistemos patrauklios dėl aukštos duomenų kokybės ir atkartojamumo, bet nekomercinės sistemos lankstesnės ir pigesnės, kas itin aktualu tiriant didelį ląstelių skaičių. Deja, pastarosios dažnai nusileidžia pagaunamų transkriptų ir genų skaičiumi (313). Dėl šios priežasties, kuriant inDrops-2 siekta pagerinti metodo jautrumą, įdiegti vartotojams patogesnį ir greitesnį sekoskaitos bibliotekų paruošimo protokolą, be to, pritaikyti metodą fiksuotų ir archyvuotų ląstelių analizei.

Pirminės inDrops metodo versijos bibliotekų paruošimas buvo paremtas *in vitro* transkripcija (IVT), o atnaujintoje versijoje įdiegtas matricos keitimu paremtas greitesnis ir patogesnis protokolas, išnaudojantis atvirkštinės transkriptazės fermentui būdingą kelių nuo sekos nepriklausančių nukleotidų pridėjimą sintetinamos kDNR gale. Ši sritis tuomet išnaudojama universalios sekos, naudojamos PGR padauginimui, pridėjimui. Analizuojant plaučių adenokarcinomos klinikinį mėginį parodyta, kad toks atnaujintas protokolas nenusileidžia transkriptų ir genų pagavimu (**Figure 3.2, A**), ir, nors duomenyse aptinkama daugiau trumpesnių genų transkriptų lyginant su IVT protokolu (**Figure 3.3, A**), tai neturi įtakos transkriptomų atlaso kūrimui bei ląstelių tipų identifikavimui (**Figure 3.4**).

inDrops-2-TS leidžia aptikti retus fenotipus klinikiniuose mėginiuose

Daugumai pavienių ląstelių RNR sekoskaitos protokolų reikalingas greitas šviežių audinių paruošimas ir barkodavimas, kas ne tik sukelia logistinių iššūkių, bet ir gali turėti neigiamos įtakos duomenų kokybei. Dėl šios priežasties, tobulinant inDrops-2 platformą sukurtas efektyvus pirminių ląstelių fiksavimo metanoliu protokolas, leidžiantis ilgalaikį audinių

archyvavimą, o papildomai pasitelkiant „ClickTag“ barkodavimo strategiją, galima ženkliai padidinti eksperimento mastą. „ClickTag“ barkodai nespecifiškai prikabinami prie fiksuotų ląstelių paviršiaus baltymų, ko pasekoje, skirtingomis sekomis pažymėtus mėginius galima sumaišyti ir juos barkoduoti vieno eksperimento metu. Tokia strategija pademonstruota tiriant plaučių karcinomos mėginius. Trumpai, plaučių karcinomos mėginiai (n=3) surinkti operacijos metu buvo padalinti į tris dalis, disocijuoti fermentiniu būdu ir ląstelės išrūšiuotos į CD45 teigiamas ir neigiamas (naudojant FACS), tuo pačiu iškart fiksuojant jas metanolyje. Ląstelių suspensijos (n=18) tuomet perkeltos į -80°C šaldiklį, o po 30 dienų pažymėtos „ClickTag“, sumaišytos ir barkoduotos naudojant inDrops-2-TS (matricos keitimo protokolą) (**Figure 3.5, A**).

Atlikus sekoskaitą ir duomenų filtravimą, gautos 32 937 aukštos kokybės ląstelės – vidutinis UMI ir genų skaičius buvo 6959 ir 1966, atitinkamai, panašiai kaip anksčiau publikuotame šviežių plaučių vėžio mėginių ląstelių atlase (309). Navikų mikroaplinkoje aptikta epitelinių, mezotelinių, endotelinių bei stromos ląstelių, taip pat pastebėta gausi imuninių ląstelių infiltracija (**Figure 3.5, B**). Skirtingų pacientų imuninės ląstelės buvo panašaus fenotipo, o neimuninių ląstelių tarpe pastebėta pacientams specifinių fenotipų (**Figure 3.5, D**). Tolesnė analizė neimuninėms, limfoidinės ir mieloidinės kilmės imuninėms ląstelėms atlikta atskirai.

Trumpai, neimuninių ląstelių atlase (n=12 521) aptikta specializuotų plaučių ląstelių: alveolinių epitelio, klubinių, žiuželiuotųjų, neuroendokrininių ir pamatinių epitelio ląstelių. Taip pat aptikta trečiajam pacientui būdingų *SPINK1* geno raiška pasižymėjusių klubinių ląstelių, bei *MMP7* ir *PRSS2* raiška pasižymėjusių alveolinių epitelio ląstelių. Pastarieji fenotipai, pagal literatūros duomenis, galimai susiję su ligos progresija (316) ir invazinėmis savybėmis (317). Taip pat aptiktos limfinio ir naviko endotelio, lygiųjų raumenų bei mezotelio ląstelės, bei dvi genų raiška besiskyrusios fibroblastų populiacijos. Viena pasižymėjo aukšta komplemento sistemos molekulių raiška, o kita buvo pažymėta *HAS1* bei citokinų *CXCL1*, *CXCL2*, *IL6* raiška, taigi, šios dvi populiacijos galimai dalyvauja naviko mikroaplinkoje vykstančiuose uždegiminiuose procesuose (**Figure 3.6 A, B**).

Visi tirti naviko audiniai pasižymėjo aukšta imuninių ląstelių infiltracija, ir nors tai dalinai susiję su neimuninių ląstelių jautrumu disociacijai, gautas pasiskirstymas buvo panašus į kitus plaučių navikų atlasus (7,319,320). Mieloidines ląsteles (n=9921) sudarė putliosios, į monocitus panašios dendritinės (kaip aprašyta (8)), pirmo tipo dendritinės, aktyvuotos dendritinės ląstelės, monocitai, alveolių makrofagai ir kelios grupės su naviku susijusių makrofagų (*angl. tumor associated macrophages, TAM*) (**Figure 3.7, A**).

Pastarieji pasižymėjo nevienodais M1 ir M2 poliarizacijos genų raiškos profiliais (**Figure 3.7, B**). Limfoidines ląsteles (n=10 495) sudarė įgimtos limfoidinės ląstelės (*angl. innate lymphoid cells, ILC*), plazmocitoidinės dendritinės ląstelės, natūralūs žudikai, B ir plazminės ląstelės, ir didelė grupė įvairių CD4 ir CD8 T ląstelių, įskaitant proliferuojančius T limfocitus (**Figure 3.8, A**). CD4 teigiamų T ląstelių grupę sudarė reguliacinių, naivių ir *CXCL13* raiška pasižymėjusių ląstelių fenotipai, o CD8 T limfocitų grupėje aptiktos citotoksinio ir atminties efektorinio fenotipo ląstelės (**Figure 3.8, B**). Mieloidinės ląstelės skirtingų pacientų mėginiuose buvo nevienodai gausios (**Figure 3.7, D**), tuo tarpu limfoidinių ląstelių pasiskirstymas buvo mažiau variabilus (**Figure 3.8, D**).

Apibendrinant, plaučių karcinomos analize parodyta, kad inDrops-2-TS metodas tinkamas klinikinių mėginių atlasų sudarymui, o sukurtas ilgalaikio mėginių saugojimo metanolyje protokolas suderinamas su barkodavimu – duomenų kokybė nenusileidžia šviežio audinio tyrimams. Be to, parodyta, kad inDrops-2-TS metodu galima aptikti retas, potencialiai kliniškai aktualias ląstelių populiacijas, tokias kaip su uždegimu susiję fibroblastai ir *CXCL13* raiška pasižymėję CD4 T limfocitai. Šios populiacijos anksčiau nebuvo aprašytos plaučių karcinomos kontekste ir galėtų būti įdomios tolesniems tyrimams.

Inkstų ir šviesių ląstelių inkstų karcinomos atlasas atskleidė naviko mikroaplinkos sudėtį ir retą ląstelių fenotipą sveikuose audiniuose

Šviesių ląstelių inkstų karcinomos (ccRCC) pavienių ląstelių RNR sekoskaitos atlasai suteikė vertingų žinių apie navikinių ląstelių pirmtakus (196), piktybėjimo metu vykstančius transkripcinius pokyčius (243), imuninių ląstelių populiacijas (230,234), jų fenotipinius pokyčius ligos eigoje (232) ir gydymo metu (233). Šie tyrimai daugiausia dėmesio skyrė vėžinėms ir imuninėms ląstelėms, o stromos ir endotelio ląstelės, kurios itin svarbios ccRCC ligos kontekste, nebuvo detaliai analizuotos. Šiame disertacijos skyriuje pagrindinis dėmesys skirtas būtent šių ląstelių ištyrimui. Be to, sudarytas gretimų sveikų inkstų audinių atlasas, kuriame aptikta retų ląstelių tipų, ne tik parodė ryškius ląstelinius ccRCC sukeltus pokyčius, bet ir pademonstravo inDrops-2 platformos galimybes.

Naudojant inDrops-2-TS platformą, tirti švieži navikų (n=8) ir sveikų gretimų inkstų audinių (n=9) mėginiai. Siekiant sutrumpinti manipuliacijas su ląstelėmis ir ištirti kiek įmanoma daugiau fenotipų, atsisakyta bet kokio ląstelių praturtinimo. Atlikus sekoskaitą ir kokybės kontrolę, sudarytas

bendras ląstelių (n=50 236) atlasas ir atlikus diferencinės genų raiškos bei išsamią literatūros analizę anotuoti ląstelių fenotipai (**Figure 3.9**).

Naviko audiniuose (pT1a ir pT3a stadijos) nustatyta gausi imuninių ląstelių infiltracija, aptikta stromos ląstelių (miofibroblastai, kraujagyslių lygiųjų raumenų ląstelės, mezanginės-lygiųjų raumenų), gausi endotelio ląstelių grupė, sudaryta iš kelių skirtingų fenotipų; vėžinės ląstelės, bet beveik neaptikta specializuotų nefrono epitelinių ląstelių (**Figure 3.10, A, B**). Vėžinėse ląstelėse, kurios sudarė tris populiacijas, aptikta kanoninių ccRCC žymenų raiška (*CA9*, *NDUFA4L2*, *VEGFA*). Viena šių populiacijų (Tumor cells 1) buvo specifinė vienam pacientui ir išsiskyrė žymenų, susijusių su lipidų metabolizmu (*FABP7* (321)) ir ląstelių-pirmtakių fenotipu (*VCAM1*, *SLC17A3*) raiška. Verta pažymėti, kad naviko mėginių ląstelinė sudėtis ženkliai skyrėsi tarp pacientų (**Figure 3.10, C**), o kiekybinis heterogeniškumo įvertinimas parodė didžiausią variaciją stromos, endotelio ir vėžinių ląstelių grupėse (**Supplementary Figure S1, C**).

Tirtuose navikui gretimuose sveikuose mėginiuose aptiktos visos pagrindinės inkstams būdingos epitelinės ir endotelinės ląstelės, anotuotos nefrono segmentų ir endotelio specializacijos detalumu (51,204,323). Eksperimentinis dizainas, atsisakius ląstelių praturtinimo, pasiteisino – aptikta ląstelių tipų, kurie žinomi kaip itin jautrūs manipuliacijoms (324). Pavyzdžiui, nustatytos abi, kylančioji (*DNASE1L3*) ir nusileidžiančioji (*AQP1*, *SLC14A1*) *vasa recta* kraujagyslės dalys, taip pat, glomerulės endotelis (*IGFBP5*, *SOST*). Be to, aptikti abu surenkamojo kanalėlio įsiterpiančiųjų ląstelių tipai (A ir B, pasižymintys atitinkamai *ATP6V1G3* ir *SLC26A4* raiška) bei podocitai (*NPHS2*, *PODXL*) (**Figure 3.9, B; Figure 3.10, A**). Įdomu tai, kad sveikame audinyje nustatyta reta epitelinių ląstelių populiacija (n=321), pasižymėjusi ccRCC pirmtakėms būdingu genų raiškos profiliu (*VCAM1*+, *SLC17A3*+, *SLC7A13*-), kaip aprašyta Young et al. (196), nors *SLC17A3* raiška pasižymėjo tik kelios ląstelės (**Figure 3.11, A**). Be to, ši populiacija ekspresavo genus, susijusius su dediferenciavusiu inkstų epiteliu (*PROM1* ir *ITGB8* (192)), bei nefrono atsaku pažeidimams (*CD24* ir *SOX4* (325)). Siekiant nustatyti šios ląstelių populiacijos panašumą su ccRCC pirmtakėmis, proksimalinio kanalėlio PT-B ląstelėmis, aprašytomis Zhang et al. (222), naudojantis viešai prieinamais šio tyrimo duomenimis buvo apmokytas CellTypist automatinio anotavimo modelis. Rezultatai ne tik patvirtino vėžinių ląstelių kilmę iš proksimalinio kanalėlio, bei kitų nefrono segmentų teisingą anotavimą, bet ir parodė, kad minėtos į pirmtakes panašios populiacijos transkriptomas išties panašiausias PT-B fenotipui (**Figure 3.11, B**). Taigi, tikėtina, kad sveikuose audiniuose aptikta reta epitelio populiacija yra dediferencijuoto, į ccRCC pirmtakes panašaus fenotipo.

Šviesių ląstelių inkstų karcinoma gausiai infiltruota su vėžiu susijusių makrofagų, slopinančių imuninį atsaką

Šviesių ląstelių inkstų karcinoma yra gausiai imuninių ląstelių infiltruotas navikas, o neseniai parodyta, kad imuninių ląstelių fenotipiniai pokyčiai ligos (232) ir gydymo eigoje (233) turi įtakos pacientų išgyvenamumo prognozei. Turint omenyje, kad pažengusios ir metastatinės ligos gydymui naudojami imuninės patikros slopikliai (215), naviką infiltruojančių imuninių ląstelių ištyrimas vienos ląstelės lygmeniu itin aktualus.

Inkstų ir ccRCC atlase identifikavome pagrindines ccRCC būdingas tiek limfoidinės, tiek mieloidinės kilmės imunines ląsteles, kaip ir anksčiau publikuotuose ccRCC atlasuose (230–232). Trumpai, aptikta putliųjų, plazminių, B ląstelių, natūralių žudikų, klasikinių ir neklasikinių monocitų, bei didelė populiacija įvairių T limfocitų bei su vėžiu susijusių makrofagų (**Figure 3.10, A**). Pastarieji sudarė keturias subpopuliacijas (TAM 1-4) (**Figure 3.12, A**), pasižymėjusias heterogeniška poliarizacijos žymenų raiška. Pavyzdžiui, TAM 1 ir TAM 2 aiškiai atitiko M1 ir M2 poliarizacijos profilį, atitinkamai, tuo tarpu TAM 4 pasižymėjo aukščiausia komplemento sistemos komponentų raiška (C1q genai) (**Figure 3.12, B**). Literatūroje yra duomenų, kad sąveikaujant su naviko ląstelių gaminamomis komplemento sistemos molekulėmis (pvz. C1s, C1r, C3) tokie makrofagai prisideda prie naviko augimo (326). Taigi, parodyta, kad naviko mikroaplinkoje dominuoja imunosupresinio fenotipo makrofagai.

T limfocitų grupę sudarė CD8+, atminties, citotoksinės ir CD4+ reguliacinės T ląstelės (**Figure 3.13, A**). Šios ląstelės pasižymėjo nevienoda išsekimo žymenų raiška (**Figure 3.13, B**). Kadangi tiek su vėžiu susijusių makrofagų, tiek T limfocitų fenotipai atitiko literatūroje išsamiai aprašytus (231,232,327), tolesnėje analizėje siekėme įvertinti tarpląstelines sąveikas su naviko ląstelėmis. Naudojant CellPhoneDB tarpląstelinių sąveikų duomenų bazę ir sąveikų prognozės statistinį modelį nustatyta, kad imuninės ir vėžinės ląstelės pasižymėjo receptorių ir ligandų raiška, susijusia su imuninio atsako slopinimu, chemokinų apdorojimu bei vėžinių ląstelių išgyvenamumo palaikymu (**Figure 3.14, A**). Pavyzdžiui, prognozuota vėžinių ląstelių ir makrofagų imuninės patikros sąveika per *HLA-G - LILRB1/2.* Yra duomenų, kad ši sąveika skatina imunosupresinę M2 makrofagų poliarizaciją ir leidžia vėžinėms ląstelėms išvengti imuninės kontrolės (328). Taip pat, parodyta, kad CD8+ ir reguliacinės T ląstelės sąveikauja su vėžinėmis per *CD27 - CD70,* o tai, literatūros duomenimis, skatina T limfocitų išsekimą ir su vėžiu susijusių makrofagų infiltraciją (330). Toliau, naudojant viešai prieinamus The Cancer

Genome Atlas (TCGA) ccRCC (KIRC imtis) RNR sekoskaitos duomenis, įvertintas sąveikų ryšys su klinikiniais parametrais. Nustatyta, kad bendra sąveikų (receptorių ir ligandų) genų rinkinio raiška koreliuoja su prastu pacientų išgyvenamumu (**Figure 3.14, B**) bei tolygiai kyla ligai progresuojant (**Figure 3.14, C**). Apibendrinant, gauti rezultatai parodo imuninių ir vėžinių ląstelių galimą kooperaciją sukuriant imunosupresinę naviko mikroaplinką, palankią naviko augimui ir išgyvenimui.

Naviko endotelio ląstelės yra heterogeniškos ir skiriasi nuo sveiko endotelio ląstelių, aptikta reta viršūninių ląstelių populiacija

Dažniausios ccRCC navikų mutacijos sukuria pseudo-hipoksines aplinkos sąlygas, lemiančias angiogenezę skatinančių veiksnių gamybą ir gausią naviko vaskuliarizaciją (211,216), todėl angiogenezė išlieka pagrindiniu ccRCC terapiniu taikiniu. Naviko endotelio ląstelių fenotipai ir jų potencialūs vaidmenys ccRCC naviko mikroaplinkoje yra aktualūs terapiniame kontekste, bet tebėra menkai aprašyti. Kadangi pavyko aptikti nemažą naviko endotelio ląstelių grupę, tolesnės analizės tikslas buvo nustatyti šių ląstelių fenotipus ir tarpląstelines sąveikas.

Identifikuotos penkios naviko endotelio ląstelių populiacijos (**Figure 3.15, A**). Diferencinės genų raiškos analizė parodė, kad naviko endotelis ženkliai skiriasi nuo sveiko endotelio ir pasižymi *PLVAP*, *VWF*, *ANGPT2*, *SPARC*, *HSPG2*, *IGFBP7*, *INSR*, IV tipo kolageno ir kitų genų raiška (**Figure 3.15, B**). Nemaža dalis šių genų yra susiję su ligos progresavimu. Pavyzdžiui, literatūros duomenimis, fenestracijos žymuo *PLVAP* yra daug žadantis terapinis taikinys hepatoceliulinės karcinomos gydymui, nes jo blokada slopina naviko augimą (331); *ANGPT2* susijęs su angiogeneze ir imunosupresinių makrofagų infiltracija (332); insulino receptorius *INSR* skatina naviko endotelio migraciją ir yra susijęs su prastesne šlapimo pūslės vėžio išgyvenamumo prognoze bei atsparumu anti-angiogenezės terapijai (334). Šie rezultatai apibrėžia ccRCC navikų kraujagyslių transkripcinį profilį ir gali būti naudingi kaip žymenys tolesniuose tyrimuose.

Keturios iš penkių aptiktų populiacijų, įskaitant į *vasa recta* panašų fenotipą (222), buvo aptiktos ir ankstesniuose tyrimuose (196,222,243), tačiau viena nedidelė (n=151 ląstelė) populiacija (Tumor Vasculature 3, TV 3), mūsų žiniomis, neturėjo publikuotų analogų ccRCC kontekste. Diferencinės genų raiškos analizė tarp visų endotelio ląstelių parodė, kad skirtingos naviko endotelio subpopuliacijos pasižymi subtiliais genų raiškos skirtumais (**Figure 3.15, C**). Pavyzdžiui, TV 3 populiacija ekspresavo viršūninių ląstelių (*angl. tip cells*) žymenis *LOX*, *PXDN*, *LY6H* ir *PGF* (305,335); TV 1, TV 4 ir TV 3

pasižymėjo užląstelinio užpildo komponentų, įskaitant angiogenezę skatinantį IV tipo kolageną ir *HPSG2*, raiška; o TV 2 populiacija išsiskyrė su angiogeneze (*FLT1*, *ANGPT2*, *KCNE3*, *ESM1* (201,336)) ir naviko augimu susijusių (*ENPP2* (337)) genų raiška. Tarpląstelinių sąveikų analizė parodė, kad naviko endotelis, kaip ir vėžinės ląstelės, palaiko imuninio atsako slopinimą naviko mikroaplinkoje (pvz. *TIGIT* - *NECTIN2*, *HLA-F* - *LILRB1/2, SCGB3A1* - *MARCO* sąveikos) ir skatina angiogenezę (**Figure 3.16, A**). Be to, nustatyta, kad bendra sąveikų (receptorių ir ligandų) raiška koreliuoja su prastu išgyvenamumu TCGA KIRC kohortoje (**Figure 3.16, B**).

Viršūninio fenotipo naviko endotelio ląstelės, TV 3 populiacija, pasižymėjo naviko augimą skatinančių genų *LOX*, *PXDN*, *LY6H* ir *PGF* raiška (**Figure 3.17, A**). Pavyzdžiui, žinoma, kad peroksidazė *PXDN* ir liziloksidazė *LOX* sutvirtina IV tipo kolagenu praturtintą užląstelinį užpildą, ir taip skatina endotelio ląstelių proliferaciją ir išgyvenamumą (341). Plaučių navikuose nustatyta, kad *LOX* slopinimas (*angl. knock-down*) sutrikdo endotelio ląstelių migraciją (305). Pasitelkiant CellTypist ląstelių anotavimo modelį ir viešai prieinamus Goveia et al. (305) duomenis parodyta, kad TV 3 populiacijos genų raiškos profilis išties atitinka viršūninės naviko endotelio ląsteles (**Figure 3.17, B**).

Signalinių kelių praturtinimo analizė parodė, kad endotelio ir stromos ląstelių populiacijų žymenys, nustatyti diferencinės raiškos analizės metu, yra susiję su epiteliniu-mezenchiminiu virsmu, nors konkretūs sutampantys genai skyrėsi tarp populiacijų (**Figure 3.18, A**). Toliau siekta nustatyti šių genų raiškos ryšį su pacientų išgyvenamumu TCGA KIRC kohortoje. Įdomu tai, kad tik TV 3 ir į *vasa recta* panašios populiacijos žymenų, susijusių su epiteliniu-mezenchiminiu virsmu, raiška koreliavo su prastesniu išgyvenamumu (**Figure 3.18, B, C, Supplementary Figure S4**).

Taigi, rezultatai pagrindžia sampratą, kad naviko endotelio ląstelės aktyviai dalyvauja navikui palankios mikroaplinkos kūrime: pasižymi specifinių užląstelinio užpildo komponentų raiška, skatina angiogenezę bei sąveikauja su imuninėmis ląstelėmis kuriant imunosupresinę nišą. Be to, rasta iki šiol ccRCC kontekste neaprašyta reta TV 3 populiacija yra viršūninio fenotipo, ir, atsižvelgiant į panašumą plaučių navikuose aptiktai populiacijai, tikėtina, skatina naviko progresavimą.

Šviesių ląstelių inkstų karcinomos stromos ląstelės prisideda prie imuninio atsako slopinimo

Nors stromos ląstelės pastebėtos kaip svarbi ccRCC navikų mikroaplinkos sudedamoji dalis (322), ankstesniuose tyrimuose jos sulaukė mažai dėmesio.

Inkstų ir ccRCC atlase identifikuotos trys negausios stromos ląstelių populiacijos: kraujagyslių lygiųjų raumenų ląstelės (vSMCs), miofibroblastai ir mezanginės/vSMC ląstelės, kurių tikslus anotavimas buvo komplikuotas dėl abiejų fenotipų žymenų raiškos (**Figure 3.19, A, B**). Pastaroji populiacija, aptikta tiek sveikuose inkstuose, tiek navikuose, skyrėsi genų raiška. Naviko mėginiuose ji pasižymėjo vėžinių ląstelių žymens *NDUFA4L2,* mezanginių ląstelių streso žymens renino *REN* (344) bei su inkstų nepakankamumu ir prasta ccRCC prognoze susijusio *CD36* raiška (342,343) (**Figure 3.19, C**). Šie transkripciniai pokyčiai galimai atspindi ląstelių atsaką į naviko sukeltus mikroaplinkos pokyčius.

Atlikus ląstelių komunikacijos analizę nustatytos su stromos ląstelių išgyvenamumu, proliferacija bei imuninių ląstelių slopinimu susijusios sąveikos (**Figure 3.20, A**). Prognozuota, kad su naviku susiję makrofagai ir stromos ląstelės, sąveikavo per *ANXA1-FPR1,* susijusią su imunosupresine makrofagų poliarizacija ir naviko progresija (345,346). Įdomi sąveika nustatyta tarp stromos ir citotoksinių T ląstelių, *HLA-E-KLRC1,* susijusi su T ląstelių išsekimu (347) ir pasiūlyta kaip daug žadantis terapinis taikinys (348). Stromos ir imuninių ląstelių sąveikų genų rinkinys koreliavo su prastu pacientų išgyvenamumu TCGA KIRC kohortoje (**Figure 3.20, B**) bei ligos stadija (**Figure 3.20, C**). Apibendrinant, gauti rezultatai parodo, kad stromos ląstelės taip pat gali būti aktyvios navikui palankios mikroaplinkos dalyvės.

Žmogaus vaisiaus vandenų ląstelių transkriptomų atlasas

Vaisiaus vandenys yra dinamiška biologinė sistema, ne tik apsauganti besivystantį vaisių nuo mechaninių pažeidimų, bet ir prisidedanti prie žarnyno ir kvėpavimo sistemų vystymosi. Žinoma, kad vaisiaus vandenyse yra nuo vaisiaus atkibusių ląstelių, kurios naudojamos prenatalinėje diagnostikoje. Taip pat manoma, kad maža (<1%) dalis jų yra kamieninės ląstelės, pasižyminčios c-kit paviršiaus žymens (295), bei mezenchiminių ir pluripotentiškumo žymenų raiška (349,350). Vis dėlto, tyrimams dažniausiai naudojamos kultivuotos ląstelės, o publikuoto nekultivuotų ląstelių transkriptomo tyrimo šiai dienai nėra. Taigi, paskutinėje disertacijos dalyje pristatomas pirmas nekultivuotų žmogaus vaisiaus vandenų ląstelių transkriptomų atlasas, sukurtas naudojant inDrops-2-TS platformą.

Trumpai, tirti švieži amniocentezės metu surinkti vaisiaus vandenų ėminiai, apimantys dvi vystymosi stadijas – 16 (n=19) ir 20 (n=9) savaitę po koncepcijos (PCW). Didžiajai daliai pacienčių, vaisiaus genetinių sutrikimų nenustatyta (klinikinė informacija pateikta **Supplementary Table S3**). Atlikus sekoskaitą ir bioinformatinę analizę, sudarytas visų ląstelių

transkriptomų atlasas, kurį sudarė 50 157 ląstelės (42 472 ląstelės PCW 16; 7685 ląstelės PCW 20), padalintos į imuninių, neimuninių ir eritroidinės linijos ląstelių grupes (**Figure 3.21, A**). Šios grupės, išskyrus eritrocitus, detaliau analizuotos atskirai. Naudojant Freemuxlet algoritmą, vertinantį vieno nukleotido polimorfizmą sekoskaitos nuskaitymuose, nustatyta, kad absoliuti dauguma (94.6%) ląstelių yra vaisiaus kilmės, o mažai (<1%) ląstelių daliai, kurios buvo imuninės, priskirta motinos kilmė galimai atspindi procedūros (adatos dūrio) metu sugautas motinos ląsteles (**Figure 3.21, B**). Taigi, tolesnė analizė atlikta neatsižvelgiant į prognozuotą ląstelių kilmę ir laikyta, kad ląstelės yra vaisiaus kilmės.

Mėginių sudėties analizė parodė tendenciją, kad ankstyvesnio vystymosi laikotarpio mėginiuose daugiau imuninių, o vėlyvesnio – neimuninių ląstelių (**Figure 3.21, C**). Tikėtina, kad tai atspindi laikui bėgant nuo vaisiaus audinių atkimbančių ląstelių akumuliaciją. Nors tirtų skirtingų mėginių bendras ląstelių kiekis varijavo nuo kelių šimtų iki kelių tūkstančių ląstelių (**Supplementary Figure S5, A**), konkretiems mėginiams specifinių fenotipų nenustatyta (**Supplementary Figure S5, B**).

Gausiausios vaisiaus vandenų imuninės ląstelės yra įgimtos limfoidinės ląstelės ir makrofagai

Ankstesniuose tėkmės citometrija paremtuose tyrimuose nustatyta, kad vaisiaus vandenyse, net nesant infekcijos, reziduoja įgimtos limfoidinės ląstelės, vienbranduoliiai fagocitai, neutrofilai ir mažas kiekis natūralių žudikų bei T ir B ląstelių (273). Vis dėlto, tokie tyrimai apriboti ląstelių fenotipavimu pagal kelis iš anksto žinomus žymenis.

Identifikavus imunines ląsteles, sukurtas atskiras jų transkriptomų atlasas (n=16 942 ląstelės) detalesnei analizei. Gausiausias fenotipas buvo 3 tipo įgimtos limfoidinės ląstelės (ILC3, **Figure 3.22, A**), pasižymėjusios *RORC* ir *KIT* raiška, ir atsiskyrusios į dvi grupes (ILC3 1 ir ILC3 2). Taip pat aptikta ILC pirmtakių (ILCP, žymenys *HPN*, *SCN1B*), proliferuojančių ILC ir mėginiams F01, F02 bei F04 specifinė ILC populiacija. Didžiojoje dalyje mėginių detektuota ir natūralių žudikų (*GNLY*, *NKG7*), centrinės atminties CD8 T ląstelių (*CD27*, *CCR7*, *CD8B*) ir maža B ląstelių (*CD79A*, *IGKC*) populiacija (**Figure 3.22, A, B**). Iš mieloidinių ląstelių, aptikta maža populiacija bazofilų (*CLC*, *HDC*), putliųjų ląstelių (*TPSB2*, *TPSAB1*), monocitų (*S100A9*, *FCN1*) ir didesnė grupė antigeną prezentuojančių ląstelių, tokių kaip Langerhans/dendritinių, pasižymėjusių MHC II klasės genų ir *CD207* raiška, bei makrofagų (*CD68*). Makrofagai buvo heterogeniški ir sudarė tris populiacijas. Viena jų, tikėtina, iš monocitų kilę makrofagai,

pasižymėję monocitų genų *FTL*, *AIF1*, *S100A9* raiška; kitos dvi populiacijos buvo M2 poliarizacijos (*MRC1*), tai *LYVE1+* makrofagai bei matrikso metaloproteinazes (MMP) gausiai ekspresavusi populiacija (**Figure 3.22, A**). *LYVE1+* makrofagai žinomi kaip reziduojantys audiniuose (pvz. širdyje, plaučiuose) prie kraujagyslių (351), o vaisiaus vandenyse aptikta populiacija pasižymėjo genų *RNASE1*, *SPP1* ir *VSIG4* raiška, būdinga neseniai aprašytiems vaisiaus žarnyno makrofagams (306). Įdomu tai, kad nors šie makrofagai pasižymėjo su priešuždegimine poliarizacija siejamų žymenų *MRC1*, *CD163*, *DAB2* raiška, taip pat stebėta ir chemokino CCL2 raiška (**Figure 3.22, B**). Pastebėta, kad vaisiaus vandenų imuninių ląstelių fenotipų gausumas kinta vystymosi metu – 16 savaitės ėminiuose dominavo limfoidinės, o 20 savaitės – mieloidinės ląstelės (**Figure 3.22, C**).

Kadangi tikėtinas imuninių ląstelių šaltinis vaisiaus vandenyse yra besivystantys plaučiai ir žarnynas (274), naudojant viešai prieinamus vaisiaus žarnyno (306) ir vaisiaus plaučių leukocitų atlasus (307), apmokytas ir anotavimui panaudotas CellTypist modelis. Analizė parodė, kad nustatyti imuninių ląstelių fenotipai atitinka publikuotus minėtuose vaisiaus audinių atlasuose (**Figure 3.23, A, B**). Įdomu tai, kad virškinimo trakto atlaso modelis abiem ILC3 populiacijoms priskyrė natūralaus citotoksiškumo receptoriaus (NCR) neigiamo limfoidinio audinio induktoriaus fenotipą, o daliai ILCP ir ciklinių ILC ląstelių – teigiamo NCR fenotipą. Taip pat, abu modeliai daliai ILC3 1 populiacijos ląstelių priskyrė T ląstelių tapatybę – folikulines pagalbines ir NK T ląsteles pagal žarnyno modelį (**Figure 3.23, A**, šviesiai violetinė spalva) ir 3 tipo įgimtas T ląsteles pagal plaučių modelį (**Figure 3.23, B**, šviesiai rožinė spalva). Su ILCP, ILC, T ląstelėmis ir natūraliais žudikais siejamų žymenų analizė parodė vaisiaus vandenų įgimtų limfoidinių ląstelių heterogeniškumą ir atskleidė skirtumus tarp ILC3 1 ir ILC3 2 populiacijų (**Figure 3.23, C**). Trumpai, abi ILC3 grupės pasižymėjo ILC žymenų, bet ne *NCR2* raiška, tačiau ILC3 2 populiacijoje ILC žymenys buvo labiau išreikšti (genai *RUNX3*, *RORA*, *RORC*, *CCR6*). Įdomu tai, kad ILC3 1 populiacija turėjo didesnę su T ląstelėmis susijusių (*KLRB1*, *TRBC1*, *TRBC2*) ir su NK ląstelėmis susijusių citotoksiškumo genų (*NKG7*, *GZMA*) raišką (**Figure 3.23, C**). Taigi, ILC3 1 populiacijos raiškos profilis, kaip ir prognozuota modelių, buvo panašus į T ląstelių, o ILC3 2 ląstelės greičiausiai atspindi grynesnį ILC fenotipą. Verta paminėti, kad ILC pirmtakės pasižymėjo *NCR1* ir *NCR2*, bei chemoatraktantų *XCL1*, *XCL2* raiška, nors šiš genų raiška įprastai siejama su aktyvuotomis T ląstelėmis ir natūraliais žudikais. Be to, nustatyta, kad plaukų keratinai *KRT86* ir *KRT81* buvo gausiai išreikšti ILC pirmtakėse, ir nors jų nežymią raišką galima aptikti ILC ląstelėse kituose vaisiaus audinių atlasuose, literatūroje duomenų apie šių genų raišką ir funkcijas ILCP ląstelėse nėra.

Apibendrinant, vaisiaus vandenų imuninių ląstelių analizė ne tik patvirtino jau žinomą makrofagų ir įgimtų limfoidinių ląstelių buvimą, bet ir atskleidė iki šiol neaprašytą vaisiaus vandenų imuninių ląstelių fenotipinę įvairovę.

Vaisiaus vandenyse yra specializuotų vaisiaus audinių ląstelių

Vaisiaus vandenyse plūduriuoja įvairios nuo vaisiaus audinių natūraliai atkibusios ląstelės, kurių absoliuti dauguma (95-99%) laikoma žuvusiomis, ir aptinkama maža dalis gyvų mezenchiminių kamnieninių, ar kaip neseniai publikuota, epitelinių pirmtakių ląstelių (287). Nepaisant to, vaisiaus vandenyse esančios neimuninės ląstelės, neskaitant kamieninių, išlieka menkai aprašytos. Išsamesnei neimuninių ląstelių analizei sudarytas atskiras atlasas (n=26 160 ląstelių). Nors ruošiant mėginius nebuvo vykdytas gyvų ląstelių praturtinimas, įprastai pavienių ląstelių RNR sekoskaitos duomenų analizėje naudojami rodikliai, skirti pašalinti žuvusias ar žūstančias ląsteles (bendras transkriptų skaičius ir mitochondrinių transkriptų dalis), neindikavo didelio žuvusių ląstelių skaičiaus.

Gausiausias ląstelių tipas, kaip ir tikėtasi, buvo plokščiojo (*angl. squamous*) epitelio ląstelės (**Figure 3.24, A**). Šią didelę ląstelių grupę sudarė aukšta metalotioneinų (*MT1X*, *MT2A*) raiška pasižymėjusios plokščiojo epitelio ląstelės; dvi į keratinocitus panašios populiacijos, pažymėtos *KRT6A* ir keratinizuoto apvalkalo žymens *SPRR2A* raiška; stemplės tipo plokščiasis epitelis (*S100A14*, *CEACAM6, CAPN14*); ilgosios nekoduojančios RNR ir NOMO (*NOMO1, NOMO2, NOMO3*) raiška praturtintas epitelis (**Figure 3.24, A, B**). Įdomu tai, kad aptikta nedidelė (n=346 ląstelės) populiacija epitelio, pasižymėjusio su audinio pažeidimu ir uždegimu susijusių *CCL20*, *CXCL2*, *CXCL3*, *CXCL8* raiška. Taip pat aptikta ląstelių, galimai atkibusių nuo plauko folikulo vidinės šaknies apvalkalo (*angl. inner root sheath*) ar lydinčiojo sluoksnio (*angl. companion layer*) – stebėta žymenų *KRTDAP* ir *LY6G6C* iš odos vystymosi atlaso (353) raiška. Identifikuota epitelinių (*CDH1*+) ląstelių, pažymėtų su plastiškumu siejamų *SOX4* ir *PLCG2* (309) bei plaučių ląstelių pirmtakių transkripcijos veiksnio *ID2* raiška. Dėl šios priežasties joms preliminariai priskirtas plaučių pirmtakių fenotipas.

Neimuninių ląstelių atlase pastebėta keletas brandžių, specializuotų vaisiaus audinių ląstelių populiacijų. Tai urotelio skėtinės ląstelės (praturtintos uroplakinais, pvz. *UPK1A*, *UPK2*); inkstų podocitai (*PODXL*, *NPHS1*); plaučių blakstienuotosios ląstelės (*CAPS*, *TMEM190*); žarnyno taurinės ir enteroendokrininės ląstelės (*FCGBP*, *ZG16*, *CHGA*); enterocitai (*APOA1*, *SELENOP*, *ANPEP*) (**Figure 3.24, B**).

Įdomu tai, kad kelias ląstelių populiacijas buvo sunku anotuoti dėl stebėtos tiek epitelinių, tiek mezenchiminių ląstelių žymenų raiškos (**Figure 3.25**). Pavyzdžiui, populiacija anotuota kaip epitelinė-mezenchiminė, ekspresavo mezenchiminį žymenį *VIM* kartu su epitelio keratinais (*KRT8*, *KRT18*). Kita populiacija pasižymėjo aukšta *MMP7* raiška, *VIM*, *CD44* ir transgelinais, nors ir ekspresavo kanoninius epitelio žymenis, tokius kaip *EPCAM*, *CDH1*, *KRT7.* Taip pat aptikti fibroblastai, pažymėti pamatinio epitelio žymens *KRT5* ir kitų keratinų raiška (**Figure 3.25**).

Siekiant patvirtinti priskirtas anotacijas, atlikta diferencinės genų raiškos analizės metu nustatytų žymenų ir funkciškai anotuotų genų rinkinių persidengimo analizė. Tam naudotos kelios duomenų bazės (Gene Ontology Biological Process 2023, MSigDB Hallmark 2020 ir Reactome 2022). Rezultatai pagrindė priskirtas anotacijas. Pavyzdžiui, blakstienuotųjų ląstelių genai persiklojo su blakstienėlių surinkimo ir judesio genų rinkiniais; plaukų folikulo ląstelių genai buvo susiję su atsaku estrogenui, kuris reguliuoja plaukų augimo fazes; podocitų genai persiklojo su podocitų vystymosi ir diferenciacijos signaliniais keliais (**Figure 3.26**). Tuo tarpu į keratinocitus panašaus plokščiojo epitelio genai buvo reikšmingai susiję su keratinocitų difereciacija (**Figure 3.27**). Įdomu tai, kad aptiktų tarpinio, epitelinio-mezenchiminio fenotipo ląstelių genai siejosi su epitelinio-mezenchiminio virsmo signaliniu keliu (**Figure 3.28**).

Remiantis pateiktais rezultatais galima teigti, kad vaisiaus vandenyse aptinkama nuo vaisiaus odos, žarnyno, plaučių ir inkstų atkibusių subrendusių, specializuotų ląstelių. Be to, aptiktos mūsų žiniomis iki šiol neaprašytos tarpinio fenotipo ląstelės.

Vaisiaus vandenų ląstelių fenotipai: kamieninių ląstelių paieška

Pavienių ląstelių RNR sekoskaitos technologija jau ne kartą reikšmingai pasitarnavo itin retų ląstelių aptikimui įvairiuose audiniuose. Vertinama, kad vaisiaus vandenų kamieninės ląstelės (AFSCs) sudaro, apytikriai, mažiau nei 1% visų ląstelių vaisiaus vandenyse. Siekiant identifikuoti šias ir kitas ląsteles-pirmtakes, tolesnės neimuninių ląstelių analizės tikslas buvo įvertinti mezenchiminių AFSCs žymenų (286,295,350,354,355) ir įvairių organų epitelinių ląstelių pirmtakių žymenų (287,288) raišką aptiktose populiacijose.

Parodyta, kad pluripotentiškumo žymenų Oct-4 (genas *POU5F1*), *SOX2*, *NANOG*, Zfp42 (genas *REX1*) kaip ir su AFSC susijusio paviršiaus žymens c-kit (dar žinomo kaip CD117, genas *KIT*) raiška, o ypač, viso šio genų rinkinio raiška kartu, neaptikta (**Figure 3.29**). Šie rezultatai netikėti, nes literatūroje yra tyrimų, kuriuose c-kit naudotas praturtinti kamieninėms ląstelėms. Žinoma, verta pastebėti, kad transkripto nebuvimas neparodo, kad

nėra baltymo. Vis dėlto, atsižvelgiant į kitų su AFSCs siejamų žymenų raiškos rezultatus, galima teigti, kad tokio fenotipo, kuriuo pasižymi kultivuotos AFSCs, vaisiaus vandenyse nėra. Tuo tarpu mezenchiminių žymenų *CD44*, CD90 (genas *THY1*), CD73 (genas *NT5E*), CD105 (genas *ENG*) ir CD29 (genas *ITGB1*) raiška stebėta miofibroblastų populiacijoje. Taip pat, *MMP7* raiška pasižymėjusi tarpinio fenotipo populiacija ekspresavo *CD44*, o integrinas β1 (*ITGB1*), susijęs su ląstelių adhezija, buvo išreikštas beveik visose populiacijose.

Atitinkamai, nei viena populiacija nepasižymėjo bendra genų rinkinių, siejamų su audiniui specifinėmis ląstelėmis pirmtakėmis (pagal Gerli et al. (287) ir Babosova et al. (288)), raiška (**Figure 3.30**). Populiacija anksčiau įvardinta kaip plaučių pirmtakės, pasižymėjo *ID2*, *PLCG2* ir *SOX4*, bet ne kitų plaučių pirmtakių žymenų raiška, taigi, galimai ši anotacija nėra tiksli. Įdomu tai, kad *MMP7*-praturtintos epitelinės-mezenchiminės ląstelės pasižymėjo *PAX2*, *PAX8*, *EMX2*, *LHX1*, *GATA3* ir *POU3F3* raiška (**Figure 3.30**). Šie genai koduoja transkripcijos veiksnius susijusius su inkstų vystymųsi ir, literatūros duomenimis, žymi inkstų ląsteles-pirmtakes. Taigi, tikėtina, kad *MMP7*-praturtintos epitelinės-mezenchiminės ląstelės yra kilusios iš vaisiaus inkstų ir gali turėti kamieninių ar pirmtakinių savybių, bet nėra pluripotentinės.

Apibendrinant, rezultatai rodo, kad kultivuotoms vaisiaus vandenų kamieninėms ląstelėms ekvivalentaus fenotipo šviežiuose vaisiaus vandenų mėginiuose nėra. Taip pat nepavyko aptikti ir kitų organų pirmtakinių ląstelių, o plaučių pirmtakėmis anotuotas fenotipas gali būti netikslus, ir bus reikalinga tolesnė analizė. Vis dėlto, atrastas įdomus, galimai inkstų ląstelių pirmtakių fenotipas, pasižymintis *MMP7* bei epitelinių ir mezenchiminių žymenų raiška.

DISKUSIJA

Pirmojoje rezultatų dalyje pristatomas patobulintas lašeliais paremta pavienių ląstelių RNR sekoskaitos platforma inDrops-2, pasižyminti didesniu, komercinėms sistemoms prilygstančiu jautrumu, bei patogiu matricos keitimu paremtu sekoskaitos bibliotekų paruošimo protokolu. Naudojant atnaujintą metodą buvo atlikta fiksuotų ir archyvuotų plaučių karcinomos audinių (n=18) analizė ir aprašyti ne tik pagrindiniai žinomi plaučių ląstelių tipai (7,319,320), bet ir keli įdomūs, potencialiai klinikinės reikšmės turintys fenotipai. Pavyzdžiui, aptikta chemokino CXCL13 raiška pasižymėjusi CD4 T ląstelių populiacija. Žinoma, kad šis chemokinas dalyvauja tretinių limfoidinių struktūrų formavimesi pritraukiant B ląsteles (356), o nesmulkialąstelinio plaučių vėžio kontekste CXCL13 bei *PDCD1* ekspresuojančių CD8 T ląstelių

kiekis turi teigiamą prognostinę reikšmę pacientams gydomiems PD-1 blokada (357). Aprašyta CXCL13 CD4 T ląstelių populiacija taip pat pasižymėjo *PDCD1* raiška, taigi, gali būti, kad šie fenotipai yra panašūs. Kita įdomi populiacija – trečiajam pacientui būdingi *HAS1* raiška pasižymėję fibroblastai. Neseniai parodyta, kad idiopatinės plaučių fibrozės mėginiuose būtent *HAS1* fibroblastai yra invazyviausi, ir prisideda prie patologinio užląstelinio užpildo gamybos (318). Iki šiol plaučių karcinomos audiniuose tokia populiacija nebuvo aptikta, o atsižvelgiant į augantį susidomėjimą stromos ląstelių įtaka navikams (358), šios ląstelės galėtų būti patrauklus tolesnių tyrimų objektas. Dar vienas įdomus fenotipas – *MMP7* ir *PRSS2* raiška pasižymėjusios alveolių epitelio ląstelės. *MMP7* naudojamas kaip plaučių fibrozės žymuo, o *PRSS2* raiška siejama su invazyviomis ir metastazėms palankiomis savybėmis (317). Be to, neseniai abiem šiais žymenimis pasižymėjusi populiacija aprašyta idiopatinės plaučių fibrozės kontekste, kaip tarpinis, plastiškas prieš-alveolinis fenotipas (361). Taigi, šis galimai plastiškas fenotipas galėtų būti įdomus ir plaučių karcinomos kontekste. Verta paminėti, kad plaučių karcinomos mėginių tyrimas šiame darbe apsiriboja ganėtinai paviršutiniška transkriptomo analize, ir aptartos ląstelių savybės nėra patvirtintos funkciniais *in vitro* ar *in vivo* tyrimais. Vis dėlto, ši analizė atskleidė inDrops-2 metodo tinkamumą retų, potencialiai kliniškai svarbių populiacijų aptikimui. Be to, ląstelių fiksavimo ir cheminio žymėjimo protokolas, suderinamas su ilgalaikiu mėginių saugojimu biobankuose, yra svarbus proveržis pavienių ląstelių tyrimuose, dažnai apsunkintuose dėl logistinių iššūkių. Taigi, aukšto našumo pavienių analizė inDrops-2 platforma yra lanksti, nebrangi ir prieinama, o gauti duomenys nei kokybe, nei biologine svarba nenusileidžia komercinėms alternatyvoms.

Antroje rezultatų dalyje išsamiai aprašomas šviesių ląstelių inkstų karcinomos (ccRCC) ir gretimų sveikų inkstų mėginių pavienių ląstelių transkriptomų atlasas. Pavienių ląstelių barkodavimui naudota inDrops-2 platforma ir, kitaip nei ankstesniuose tyrimuose, atsisakyta ląstelių praturtinimo, o tai leido aptikti retus ir jautrius ląstelių fenotipus. Kadangi imuninių ląstelių fenotipai sutapo su anksčiau detaliai ištirtais (230–234,244), pagrindinis dėmesys buvo skirtas mažiau ištirtoms naviko endotelio ir stromos ląstelių populiacijoms ir jų sąveikoms su kitomis ląstelėmis naviko mikroaplinkoje.

Naviko endotelio ląstelės yra itin svarbios šviesių ląstelių inkstų karcinomos vystymesi ir iki šiol išlieka pagrindiniu terapijų taikiniu pažengusios ir metastatinės ligos gydyme (211). Šiame darbe išsamiai aprašytos naviko endotelio ląstelės, o svarbiausias atradimas – nauja

viršūninio fenotipo (*angl. tip cells*) naviko endotelio ląstelių populiacija. Ankstesniuose ccRCC pavienių ląstelių tyrimuose endotelio ląstelės dažniausiai skirstytos tik į du fenotipus. Pavyzdžiui, Long et al. aprašė *VCAM1*+ ir *VCAM1*- naviko kraujagysles, o Zhang et al. padalino naviko endotelio ląsteles į *ACKR1*+ ir *EDNRB*+ grupes (222,362). Atitinkamai, šiame darbe pateiktame atlase aptikta į *vasa recta* panaši populiacija, pasižymėjusi *ACKR1* ir *VCAM1* raiška, o *EDNRB* raiška stebėta likusiose populiacijose, išskyrus viršūninį fenotipą. Taigi, inkstų vėžio kontekste šios endotelio (*PECAM1*+) ląstelės anksčiau nebuvo aprašytos. Panašus viršūninių endotelio ląstelių fenotipas, pasižymėjęs žymenų *LOX*, *PXDN*, *PGF*, *LXN*, bei IV tipo kolageno raiška kaip ir šiame darbe, detaliai aprašytas plaučių vėžio kontekste, kur koreliavo su prastu pacientų išgyvenamumu (305). Be to, autoriai pademonstravo, kad *LOX* slopinimas (*angl. knock-down*) sutrikdo endotelio ląstelių migraciją ir kraujagyslių šakojimąsi. Taigi, naujai aprašyta inkstų navikų endotelio populiacija galėtų būti įdomi tolesniems tyrimams ar net kaip potencialus terapinis taikinys.

Inkstų navikų mikroaplinkos tyrimo rezultatai leidžia kelti hipotezę, kad naviko endotelis dalyvauja navikui palankios mikroaplinkos kūrime dviem pagrindiniais būdais. Pirma, šios ląstelės pertvarko užląstelinį užpildą ir gamina navikui palankius jo komponentus; antra – sąveikauja su kitomis ląstelėmis, įskaitant imunines, ir taip prisideda prie imuninio atsako slopinimo bei angiogenezės skatinimo. Ankstesniuose tyrimuose parodyta, kad kolageną gaminančios endotelio ląstelės aptinkamos naviko kraštuose, kur gausu makrofagų ir epitelinio-mezenchiminio virsmo kelio (EMT) genų raiška pasižyminčių vėžinių ląstelių (234). Gauti rezultatai taip pat indikuoja, kad naviko endotelis gali būti susijęs su EMT ir sąveikauti su makrofagais. Bioinformatinė tarpląstelinių sąveikų prognozė atskleidė įdomių, kliniškai svarbių sąveikų tarp naviko endotelio, stromos bei infiltruojančių imuninių ląstelių. Pavyzdžiui, 2021 metais prasidėjo klinikinis tyrimas, tiriantis LILRB1 ir LILRB2 blokados veiksmingumą pažengusių ir metastazavusių navikų, įskaitant ccRCC, gydymui kaip monoterapija, ar kombinuojant su PD-1 imuninės patikros slopikliais (tyrimo ID: NCT04913337). Žinoma, kad LILRB2 slopinimas perprogramuoja mieloidines ląsteles į uždegiminį fenotipą, o LILRB1 – šia linkme keičia tiek mieloidinių, tiek limfoidinių ląstelių fenotipus. Įdomu tai, kad, mūsų duomenimis, *LILRB1/2*+ imuninės ląstelės per šiuos slopinančius receptorius sąveikauja ne tik su vėžinėmis, bet ir naviko endotelio ląstelėmis. Sąveika tarp T ląstelių receptoriaus TIGIT ir NECTIN2, tiriama daugybėje klinikinių tyrimų (363), taip pat prognozuota būtent tarp T ląstelių ir naviko endotelio, o ne pačių vėžinių ląstelių. Be to, parodyta, kad ir stromos ląstelės galimai dalyvauja imuninio atsako slopinime,

pavyzdžiui, per *HLA-E-KLRC1* sąveiką, susijusią su T ląstelių išsekimu (347) ir laikomą daug žadančiu terapiniu taikiniu (348). Yra duomenų, kad atsinaujinančioje inkstų karcinomoje padaugėja stromos ląstelių, o jų gaminamo galektino-1 slopinimas, pelių modeliuose, sumažina naviką ir gerina atsaką į PD-1 blokadą (365). Atsižvelgiant į tai, kad sąveikų genų rinkinių raiška koreliavo su prastu išgyvenamumu, o stromos ląstelių atveju ir su ligos stadija, galima manyti, kad naviko endotelio ir stromos ląstelės išties yra svarbios navikui palankios mikroaplinkos kūrėjos ir dalyvės.

Kaip ir bet kuris kitas, šis pavienių ląstelių RNR sekoskaitos tyrimas turi trūkumų, kuriuos svarbu aptarti. Pavyzdžiui, yra žinoma, kad adhezinės ląstelės, tokios kaip epitelinės ar vėžinės, sunkiau nei imuninės išgyvena audinio disocijavimo procedūrą (324). Todėl, nors gausi imuninių ląstelių infiltracija yra pripažinta ccRCC navikų savybė, mėginių ląstelinė sudėtis šiame ir kituose panašiuose tyrimuose (231,244,362) priklauso nuo eksperimentinio dizaino ir nėra tiksli. Kitas iššūkis, būdingas visiems pavienių ląstelių RNR tyrimams, yra nedidelis sugaunamos mRNR kiekis, lemiantis duomenų praradimą (*angl. dropout*). Šiuo aspektu analizuotas duomenų rinkinys nebuvo itin geros kokybės, todėl bioinformatinių metodų pasirinkimas buvo apribotas, ir kai kurių (pvz. pseudolaiko) metodų taikymas nebuvo informatyvus. Taip pat, analizė apsiribojo transkriptomo ištyrimu, ir papildomų *in vitro* ar *in vivo* tyrimų nebuvo atlikta. Nepaisant to, šis tyrimas svariai papildo šviesių ląstelių inkstų karcinomos navikų mikroaplinkos apibūdinimą vienos ląstelės lygmeniu – aprašyta imuninių, naviko endotelio ir stromos ląstelių įvairovė, aptikta viršūninio fenotipo endotelio populiacija bei jų tarpusavio sąveikos.

Paskutinėje rezultatų dalyje pristatomas pirmasis išsamus nekultivuotų vaisiaus vandenyse aptinkamų ląstelių transkriptomų atlasas. Jo sukūrimui pasitelkta inDrops-2 platforma ir ištirti 26 švieži amniocentezės metu surinkti vaisiaus vandenų ėminiai. Trumpai, aprašyta imuninių ir epitelinių ląstelių įvairovė, be to, aptikti įdomūs fenotipai, pasižymintys tiek epitelinių, tiek mezenchiminių žymenų raiška.

Šiuo metu yra publikuotas tik vienas tyrimas (Gerli et al. (287)), kuriame pateikiamas vaisiaus vandenų (n=12, nuo 15 iki 34 vystymosi savaitės) pavienių ląstelių RNR sekoskaitos duomenų rinkinys, nors pagrindinis šio tyrimo tikslas buvo epitelinių organoidų sukūrimas, paveikus vaisiaus vandenų ląsteles specifinėmis kultivavimo sąlygomis. Pateiktas atlasas nėra tinkamai anotuotas – pateikti ląstelių tipai užrašyti paveiksle, bet jų nėra duomenyse (NCBI GEO duombazė, nr. GSE220994); neatlikta diferencinės raiškos analizė, kas apsunkina prasmingą palyginimą su mūsų gautais

duomenimis. Didžioji dalis ląstelių, kaip ir mūsų atlase, pažymėtos kaip epitelinės; be to, pavaizduotos kelios imuninių ląstelių populiacijos: makrofagai, monocitai, neutrofilai, T ir B ląstelės, natūralūs žudikai ir eritroblastai. Įdomu tai, kad neaptikta įgimtų limfoidinių ląstelių, kurios dominavo mūsų atlase ir, literatūros duomenimis, laikomos gausiausia vaisiaus vandenų imuninių ląstelių populiacija (273). Taigi, nors šiame darbe pristatomas tyrimas nėra pirmasis bandymas tirti žmogaus vaisiaus vandenis naudojant pavienių ląstelių RNR sekoskaitos technologiją, atlikta išsamesnė duomenų analizė atskleidė iki šiol neaprašytą ląstelių įvairovę.

Imuninių ląstelių analizė parodė įgimtų limfoidinių ląstelių (ILC) transkriptomų heterogeniškumą – identifikuotos pirmtakių, proliferuojančių ląstelių, bei dvi 3 tipo ILC populiacijos, besiskyrusios su T ląstelėmis siejamų genų raiška. Įdomu tai, kad ILC pirmtakės pasižymėjo natūralaus citotoksiškumo receptoriaus NCR2 raiška, kitaip nei vaisiaus žarnyno (306), ar vaisiaus kepenų (367) ILC pirmtakės. Literatūros duomenimis, vaisiaus vandenyse aptinkama CD4 reguliacinių T ląstelių (275), o šiame atlase iš T ląstelių aptikta tik centrinės atminties (*CD27*, *CCR7*) CD8 T ląstelių. Aprašyta ir makrofagų įvairovė – galimai iš monocitų kilę makrofagai, *LYVE1* teigiami bei matrikso metaloproteinazių raiška (bei *TREM2*) pasižymėję M2 poliarizacijos makrofagai. *LYVE1+* makrofagai reziduoja audiniuose (pvz. širdyje, plaučiuose) šalia kraujagyslių (351), taip pat aprašyti ir vaisiaus žarnyno atlase (306). Neseniai pasirodęs vaisiaus odos pavienių ląstelių tyrimas taip pat aprašė tiek *LYVE1+* tiek *TREM2+* makrofagų buvimą ir rolę odos angiogenezėje (353). Nors šios ląstelės reziduoja dermyje, atsižvelgiant į didelį kiekį nuo odos atkibusių ląstelių, aptiktų vaisiaus vandenyse, gali būti, kad stebėti makrofagai patenka į vandenis tiek iš vaisiaus žarnyno, tiek iš odos. Nors imtis per maža daryti bet kokias išvadas, buvo pastebėta, kad 21-os chromosomos trisomija paveiktų vaisių vandenyse imuninių ląstelių kiekis ženkliai mažesnis, kas galimai atspindi sutrikusį imuninės sistemos vystymąsi (370). Detalesniam ištyrimui, tęsiame tokių mėginių rinkimą. Vis dėlto, vien tik iš transkriptominės analizės sunku spekuliuoti, ar imuninės ląstelės vaisiaus vandenyse yra funkcionalios.

Kalbant apie neimunines vaisiaus vandenų ląsteles, netikėta tai, kad ir be gyvų ląstelių praturtinimo (kaip atlikta minėtame Gerli et al. tyrime (287)) duomenų kokybės rodikliai nerodė su ląstelių žūtimi susijusių pokyčių. Aprašyta gausi specializuotų ląstelių įvairovė: podocitai, enterocitai, žarnyno taurinės ir enteroendokrininės ląstelės, blasktienuotosios plaučių ir skėtinės urotelio ląstelės. Kadangi šie audiniai kontaktuoja su vaisiaus vandenimis, tikėtina, patenka į juos tiesiogiai vos atkibus. Nors tokių ląstelių buvimas

pastebėtas jau prieš kelis dešimtmečius (280–282), šiame darbe pristatomas pirmasis jų transkriptomų tyrimas.

Įvertinus kanoninių vaisiaus vandenų kamieninių ląstelių (AFSCs) žymenų raišką, nepavyko nustatyti populiacijos atitinkančios šį kultivuotą fenotipą. Prusa et al. parodė, kad tik ~0.1-0.5% vaisiaus vandenų ląstelių pasižymi pluripotentiškumo žymens Oct-4, c-kit bei vimentino raiška (286), be to, jos aptiktos ne visuose mėginiuose. Taigi, gali kilti pagrįstų klausimų dėl šių retų ląstelių aptikimo naudojant pavienių ląstelių RNR sekoskaitos metodą. Vis dėlto, atsižvelgiant į adekvačią imtį (n=26), ir tai, kad aptikta itin mažų, vos 38 ląstelių dydžio populiacijų (bazofilai ir B ląstelės, 0.076% visų tirtų ląstelių), mažai tikėtina, kad AFSCs neaptikome vien dėl atsitiktinės atrankos (*angl. random sampling*). Oct-4 koduojantis genas *POU5F1* generuoja kelias transkriptų izoformas, iš kurių tik viena transliuojama į pluripotentiškumą lemiantį Oct-4A transkripcijos veiksnį; be to, kitoje chromosomoje esančio pseudogeno produktas – net 96% homologiškas baltymas, neatskiriamas nuo Oct-4A komerciniais antikūniais. Parodyta, kad tyrimai, skelbiantys Oct-4 raišką vaisiaus kamieninėse ląstelėse naudoja nespecifinius pradmenis ir antikūnius (301), o kitame tyrime pademonstruota, kad naudojant specifinius pradmenis, ir AFSCs nepasižymi Oct-4A raiška (302). Be to, ir Gerli et al. publikuotame atlase, nors ir tinkamai neanotuotame, neužsimenama apie tai, kad AFSCs neaptikta (287), nors vienas iš autorių dar 2007-aisiais paskelbė žymų tyrimą apie c-kit+ plataus potentiškumo ląsteles vaisiaus vandenyse. Taigi, atsižvelgiant į gautus rezultatus ir aptartas metodines problemas, galima teigti, kad vaisiaus vandenyse nėra fenotipo atitinkančio kultivuotas AFSCs.

Ypač netikėtas rezultatas buvo vaisiaus vandenų atlase aptiktos tarpinio fenotipo ląstelės: *KRT5* ekspresavę fibroblastai, epitelinės-mezenchiminės, ir *MMP7* pasižymėjusios ląstelės, pavadintos fibroblastais-epitelinėmis, dėl abiejų fenotipų žymenų raiškos. Svarbu pabrėžti, kad analogiškų fenotipų nepavyko aptikti kituose nei vaisiaus, nei suaugusio žmogaus audinių atlasuose, todėl anotacijos yra preliminarios ir bus tikslinamos atliekant tolesnę analizę. Viena iš šių įdomių populiacijų pasižymėjo MMP7, mezenchiminių žymenų *VIM*, *CD44* ir transgelinų raiška kartu su kanoniniais epitelio žymenimis *EPCAM*, *CDH1*, *KRT7*. Be to, dalis šios populiacijos ląstelių pasižymėjo su vystymųsi ir inkstų pirmtakėmis susijusių transkripcijos veiksnių *PAX2*, *PAX8*, *EMX2*, *LHX1*, *GATA3* ir *POU3F3* raiška. Literatūroje yra duomenų, kad CD44 raiška gali pasižymėti pažeistos, dediferenciavusios inkstų epitelio ląstelės (369). Taigi, tikėtina, kad MMP7 fibroblastai-epitelinės ląstelės yra vis tik epitelinės ląstelės, o stebėta genų raiška atspindi dediferenviavusių ar ląstelių pirmtakių fenotipą. Be to,

atsižvelgiant į su inkstais susijusių žymenų raišką, tikėtina, kad šių ląstelių šaltinis vaisiaus vandenyse yra besivystantys inkstai.

Apibendrinant, gauti rezultatai suteikia pamatinių žinių apie žmogaus vaisiaus vandenų ląstelinę sudėtį ir iki šiol neaprašytus ląstelių fenotipus. Be abejo, atlasas dar nėra galutinis – šiuo metu, vykdome tolesnį ėminių barkodavimą, ypač pirmenybę teikiant 21-osios chromosomos trisomija paveiktų vaisių ėminiams. Šiame darbe pateikta bioinformatinė analizė yra gana primityvi, todėl surinkus galutinę imtį planuojame atlikti gilesnę analizę bei panaudoti įvairesnius algoritmus (pvz., atlikti ląstelių komunikacijos, pseudolaiko analizę). Be to, siekiant išsiaiškinti plačiai kultivuojamų AFSCs kilmę, pradėjome kultivavimo ir pavienių ląstelių RNR sekoskaitos eksperimentus. Manau, kad šie tyrimai turi potencialo suteikti fundamentinių žinių ne tik apie vaisiaus vandenų ląsteles, bet ir bendrai apie ląstelių plastiškumą, peržengiant konkrečios biologinės nišos ribas.

IŠVADOS

1. inDrops-2-TS ir inDrops-2-IVT pavienių ląstelių RNR sekoskaitos metodai pasižymi panašiu pagaunamų genų ir transkriptų kiekiu
2. inDrops-2-TS metodas įgalina retų ląstelių fenotipų aptikimą
3. Pavienių ląstelių RNR sekoskaita inkstų audiniuose atskleidė visus pagrindinius ląstelių tipus, įskaitant specializuotas nefrono kanalėlių ląsteles, tuo tarpu šviesių ląstelių inkstų karcinomos mėginiai pasižymėjo gausia imuninių ląstelių infiltracija bei specifinėmis naviko endotelio ląstelėmis
4. Šviesių ląstelių karcinomos navikų T ląstelės yra išsekusio fenotipo, makrofagai tiek imunosupresinio, tiek uždegiminio fenotipo, o naviko endotelio ląstelėse aptikta viršūninio fenotipo populiacija. Visi šie fenotipai galimai prisideda prie navikui palankios mikroaplinkos palaikymo
5. Vaisiaus vandenų imunines ląsteles didžiąja dalimi sudaro įgimtos limfoidinės ląstelės ir makrofagai
6. Vaisiaus vandenyse aptinkama specializuotų odos, inkstų, žarnyno ir plaučių ląstelių bei tarpinių fenotipų, pasižyminčių tiek epitelinių, tiek mezenchiminių žymenų raiška
7. Kultivuotoms vaisiaus vandenų kamieninėms ląstelėms priskiriamo fenotipo nekultivuotuose vaisiaus vandenų ėminiuose neaptikta

ACKNOWLEDGMENTS

I have written and rewritten this section of my thesis countless times in my head – arguably more than any other part. It is the most important part, as I am filled with gratitude toward many people who helped me, challenged me, supported me and cared for me throughout the chaotic times of pursuing a PhD.

First and foremost, I would like to thank all the patients who donated their samples for research. Your contribution is essential for the progress of biomedical science. Anytime I look at the many cells scattered on my UMAPs I am fascinated by the fact there are very real people behind it whose cells I am looking at. I would also like to thank our collaborators in the clinics and beyond (especially dr. Laima Ambrozaitytė) for the acquisition of the samples, belief in our ideas and fruitful outcomes.

This chapter of my life would not have been completed without the defense panel, I am grateful to dr. Daiva Dabkevičienė, assoc. prof. dr. Monika Mozerė, prof. dr. Saulius Serva, prof. dr. Kęstutis Sužiedėlis and assoc. prof. dr. Johan Henriksson for their time evaluating my thesis and participating in the defense. I am particularly grateful to the reviewers prof. dr. Kęstutis Sužiedėlis and dr. Johan Henriksson for thoroughly examining my work and providing valuable comments. Additionally, I would like to thank the staff at VU Life Sciences Center for their help with administrative tasks and study requirements. A special thank you goes to PhD studies administrator Kristina Slavuckytė, BTI personnel specialist Janina Žiūkaitė and senior administrator Danutė Noreikienė.

Next, I would like to thank my wonderful advisors, supervisor Linas and consultant Rapolas. Rapolai – since the moment I met you, when I was still a clueless bachelor's student, you have always been patient, kind and a constant source of inspiration and support for me. Everything I know about data analysis, and the high standards I strive to maintain as a researcher, are a result of your influence and effort. You were also the bridge that brought me back to Lithuania and into Mažutis' lab, which was arguably one of the best decisions of my life. Linai – I cannot put into words how much I appreciate and value the bond we have built over my almost *seven* years in this lab. From the never-ending (and occasionally heated) discussions on science, standards, quality, attitude and life itself, to the jokes and smiles and even tears you have witnessed – I have always felt your support and trust. Despite the plentiful challenges that came in the way, I would not have asked for a better supervisor. Even though you joke about me being 'self-supervised', I am

eternally grateful for the opportunity I was given to become independent, confident, albeit sometimes *slightly* stressed, researcher.

I would also like to thank the 'old' Mažutis' lab – dr. Juozas, Emilis, dr. Greta and others. Juozai – I am very lucky my first contact with high-level science was under your supervision and help. Emili – it's always nice to have someone appreciate the inner sense of quality that we share. Greta – you are a role model, your work has always been pristine and genuine, and I appreciate the sincere conversations we had over the years. What you guys do now at Atrandi is beyond amazing and something to look up to.

All members of the lab – I would like to thank you for the atmosphere, acceptance, the unpredictable mix of ridiculous and very existential lunch talks, as well as the opportunity to appreciate the outstanding biodiversity of people and ways of thinking that make science move forward. A special thanks goes to Karolis, Vincenta, Simonas senior and the Irish squad. Vincenta – you never stop surprising me in all the good ways. Simonai – thank you for the reminders of me having limbs, but all jokes aside, for the lighthearted and positive attitude that you foster. Karoli – I guess we *got dropleted* – it's truly a pleasure to work alongside you, you have taught me so much and none of this would have been possible without your extensive *(emulsional)* support. Julija *O'Cackles* – thank you for the contagious cackles, the good times and the energy. Dominykai *McShenanigans* – for the 20/80, the wooden dad jokes, the listening, the cheering, and many more. Mindaugai *O'Nonsense* – my beloved kimchi master, my twin brother – for the sense, the antisense, the nonsense, the bluetooth and always being there for me. You all bring me so much joy every day, you are such great colleagues and true friends.

Finally, I would like to thank my family and friends. My sisters, Ilma and Vaiva – if not for your childish games, I would not be where I am today and I am immensely grateful for the world you have shown me throughout the years. Vytai – thank you for paving the way of being the dr. in the family. My dear parents Vitalija and Valdas – thank you for the values you ingrained and the unconditional love and support. My closest friends Rugilė, Evita, Kamilė, Karolis, Lorenzo, Ieva Š., Laurynas, Ieva Marija S.– I thank you for being the unpaid psychologists, for the perspectives you bring, the fulfilling talks and the shared experiences that I will forever cherish. My dear Kipras – thank you for being a light in my life and my personal anti-stress remedy. Lastly, my fluffball Stela, for the joy that she unknowingly radiates.

LIST OF PUBLICATIONS AND CONTRIBUTIONS

1. **Zvirblyte J**, Nainys J, Juzenas S, Goda K, Kubiliute R, Dasevicius D, et al. Single-cell transcriptional profiling of clear cell renal cell carcinoma reveals a tumor-associated endothelial tip cell phenotype. Commun Biol. 2024 Jun 28;7(1):1–15.

   https://doi.org/10.1038/s42003-024-06478-x

In this work, I performed the majority of single-cell profiling experiments, sequencing library preparation and sequencing, formulated ideas, analyzed the literature, analyzed and interpreted the data, prepared the material and figures, wrote the paper in its entirety, handled the publishing and peer-review process.

2. Juzenas S, Goda K, Kiseliovas V, **Zvirblyte J**, Quintinal-Villalonga A, Siurkus J, et al. inDrops-2: a flexible, versatile and cost-efficient droplet microfluidic approach for high-throughput scRNA-seq of fresh and preserved clinical samples. Nucleic Acids Res. 2025 Jan 11;53(2):gkae1312. https://doi.org/10.1093/nar/gkae1312

In this work, I analyzed the lung carcinoma scRNA-seq data, wrote that section of the results and discussion, heavily contributed to the preparation and editing of the manuscript.

3. **Žvirblytė J**, Mažutis L. Microfluidics for Cancer Biomarker Discovery, Research, and Clinical Application. In: Caballero D, Kundu SC, Reis RL, editors. Microfluidics and Biosensors in Cancer Research: Applications in Cancer Modeling and Theranostics. Cham: Springer International Publishing; 2022 p. 499–524. https://doi.org/10.1007/978-3-031-04039-9_20

For this book chapter, I collected and analyzed relevant literature, generated ideas, prepared the figures and wrote the initial manuscript.

Other publications:

1. Baronas D, Norvaisis S, **Zvirblyte J**, Leonaviciene G, Mikulenaite V, Goda K, et al. High-throughput single cell omics using semipermeable capsules. Science. 2025 Dec 18;0(0):eady7227.

2. Juran CM, **Zvirblyte J**, Cheng-Campbell M, Blaber EA, Almeida EAC. Cdkn1a deletion or suppression by cyclic stretch enhance the osteogenic potential of bone marrow mesenchymal stem cell-derived cultures. Stem Cell Res. 2021 Oct;56:102513.

3. Juran CM, **Zvirblyte J**, Almeida EAC. Differential Single Cell Responses of Embryonic Stem Cells Versus Embryoid Bodies to Gravity Mechanostimulation. Stem Cells Dev. 2022 Jul;31(13–14):346–56.

LIST OF CONFERENCES

1. International conference The Coins, April 24th-27th, 2023. Vilnius, Lithuania. Poster presentation.
   Justina Žvirblytė, Juozas Nainys, Simonas Juzėnas, Raimonda Kubiliūtė, Marius Kinčius, Albertas Vėželis, Albertas Ulys, Sonata Jarmalaitė and Linas Mažutis. "Single-cell RNA sequencing of clear cell renal cell carcinoma tumor microenvironment reveals novel tumor endothelium subpopulation". 1st place award in biochemistry and molecular biology section.

2. International Jędrzej Sniadecki memorial conference Frontiers in Molecular Life Sciences, May 23rd-25th, 2023. Vilnius, Lithuania. Poster presentation.
   Justina Žvirblytė, Juozas Nainys, Simonas Juzėnas, Raimonda Kubiliūtė, Marius Kinčius, Albertas Vėželis, Albertas Ulys, Sonata Jarmalaitė and Linas Mažutis. "Profiling the tumor microenvironment of clear cell renal cell carcinoma using single cell RNA sequencing".

3. International conference Single Cell Genomics 2024, September 16th-18th, 2024. Corinthia, Greece. Poster presentation.
   Karolis Goda, Simonas Juzėnas, Vaidotas Kiseliovas, **Justina Žvirblytė**, Alvaro Quintinal-Villalonga, Juozas Nainys and Linas Mažutis. "inDrops-2: a flexible, versatile and cost-efficient droplet microfluidics approach for high-throughput scRNA-seq of fresh and preserved clinical samples".

4. International conference Single Cell Genomics 2025, September 15th-17th, 2025. Stockholm, Sweden. Poster presentation.
   Justina Žvirblytė, Karolis Goda, Mindaugas Sinis, Juozapas Ivanauskas, Eglė Mazgelytė, Eglė Benušienė, Laima Ambrozaitytė and Linas Mažutis. "Single-cell transcriptomic profiling of uncultured human amniotic fluid cells".

CURRICULUM VITAE

Name, Surname: Justina Žvirblytė
Address: Life Sciences Center, Saulėtekio av.7, Vilnius, Lithuania, LT-10257

Education:

2021-2026: PhD in Biochemistry, Vilnius university, Lithuania
2019-2021: Master's in Molecular Biology, Cum Laude, Vilnius University, Lithuania
2015-2019: Bachelor's in Biophysics, Magna Cum Laude, Vilnius University, Lithuania

Relevant experience:

2021-present: Junior Research Scientist, Single Cell Analysis Laboratory, Institute of Biotechnology, Life Sciences Center, Vilnius University, Vilnius, Lithuania

2022 December – 2023 March: Research Visitor, Computational Systems Biology of Cancer (SysBio) group at U900, Institut Curie, Paris, France

2019 June – 2019 September: Research Associate, Young Scientist Program, Blue Marble Space Institute of Science, NASA Ames Research Center, California, United States of America

2018 August – 2018 December: Intern, International Internships program, Bone and Signaling Lab, NASA Ames Research Center, California, United States of America

2016 – 2018: Internship at the Department of Neurobiology and Biophysics, Institute of Biosciences, Vilnius University, Vilnius, Lithuania

Courses and awards:

2025: Research Council of Lithuania Scholarship for academic achievements during PhD studies

2024: Debut of the Year, Issued by Ministry of Culture of the Republic of Lithuania for debut poetry book (J.Žvirblytė, *Mikrosfera*, Lithuanian Writers' Union Publishing House, 2023, 84 p., ISBN: 978-609-480-382-6)

2023: EMBL-EBI PerMedCoE summer school: from pathway modeling tools to cell-level simulations, Barcelona, Spain

2023: NGSchool2023: Advances in Computational Biology, Otwock, Poland

2021: Best Master Thesis in Medicine and Health Sciences, issued by Lithuanian Junior Scientist Association

2020, 2024: VU Life Sciences Center Nominal Scholarship

2019: Vilnius University Rector's acknowledgment for promoting the name of University

Project participation:

- Specialist at "Microfluidic technologies for single-cell geno- and phenotyping research", No.09.3.3-LMT-K-712-01-0056, 2019 November – 2021 July
- Specialist at "Establishment of parallel single cell transcriptomics-genomics laboratory" No. 01.2.2-LMT-K-718-04-0002, 2020 July – 2023 August
- Junior researcher at "Single cell multi-omics: model studies", No. S-ERC-24-3, 2024 September – 2025 December
- Junior researcher at "Single cell transcriptomics of fetus derived cells", No. S-MIP-24-93, 2024 September – 2026 August
- Junior researcher at "Implementation of Mission-Based Science and Innovation Programmes" No. 02-002-P-0001 subproject "Innovation for Health" – "Translational Center for Gene Technology (TRACEGET)", 2024 February – 2026 April

NOTES

Vilniaus universiteto leidykla
Saulėtekio al. 9, III rūmai, LT-10222 Vilnius
El. p. info@leidykla.vu.lt, www.leidykla.vu.lt
Tiražas 13 egz.