https://doi.org/10.15388/vu.thesis.79
https://orcid.org/ 0000-0003-0297-8822

VILNIUS UNIVERSITY

Juozas
NAINYS

High-throughput single-cell sequencing and analysis

DOCTORAL DISSERTATION

Natural Sciences,
Biochemistry (N 004)

VILNIUS 2020

This dissertation was written between 2015 and 2019 in Vilnius University Life Sciences Center Institute of Biotechnology.

Academic supervisor:
Prof. dr. Linas Mažutis, (Vilnius University, Natural Sciences, Biochemistry - N 004).

https://doi.org/10.15388/vu.thesis.79
https://orcid.org/ 0000-0003-0297-8822

VILNIAUS UNIVERSITETAS

Juozas
NAINYS

Aukšto našumo pavienių ląstelių sekoskaita ir analizė

DAKTARO DISERTACIJA

Gamtos mokslai,
Biochemija (N 004)

VILNIUS 2020

Disertacija rengta 2015– 2019 metais Vilniaus universiteto Gyvybės mokslų centro Biotechnologijos institute

Mokslinis vadovas:

prof. dr. Linas Mažutis, (Vilniaus universitetas, gamtos mokslai, biochemija - N 004).

CONTENTS

ABBREVIATIONS

APC	antigen presenting cell
AT	archetype
BHB	barcoding hydrogel bead
bHLH	basic helix–loop–helix
BSA	bovine serum albumin
CSC	cancer stem cell
CTC	circulating tumor cell
DC	dendritic cell
DE	differential expression
DEG	differentially expressed gene
ECM	extracellular matrix
EDTA	ethylenediaminetetraacetic acid
EGF	epidermal growth factor
EMD	earth mover distance
EMT	epithelial to mesenchymal transition
ER	estrogen receptor
FACS	fluorescence-activated cell sorting
FGF	fibroblast growth factor
GSEA	gene set enrichment analysis
GWAS	genome-wide association study
HER2	human epidermal growth factor receptor 2
HGF	hepatocyte growth factor
HMLE	Human Mammary Epithelial
HVGs	highly variable genes
ICB	immune checkpoint blockade
IFCs	integrated fluidic circuits
IVT	*in vitro* transcription
K562	human immortalized myelogenous leukemia cells
MEGM	Mammary Epithelial Cell Growth Medium
MET	Mesenchymal-Epithelial Transition

MHC	major histocompatibility complex
NK cells	Natural killer cells
PBMC	peripheral blood mononuclear cell
PCA	principal component analysis
PCR	polymerase chain reaction
pDC	plasmacytoid dendritic cell
PDMS	poly(dimethylsiloxane)
PR	progesterone receptor
QC	quality control
qPCR	quantitative polymerase chain reaction
RT	reverse transcription
RTK	tyrosine kinase receptor
scRNA-Seq	single-cell transcriptome sequencing
SSS	second-strand synthesis
t-SNE	t stochastic neighbor embedding
TAM	tumor-associated macrophage
TF	transcription factor
Th cell	T helper cell
TIL	tumor-infiltrating leukocyte
TME	tumor microenvironment
TNBC	Triple-Negative Breast Cancer
Treg cell	T regulatory cell
UMAP	uniform Manifold Approximation and Projection
UMI	unique molecular identifier
WHO	world health organization

INTRODUCTION

It is widely accepted that cells are the basic functional units of life [1]. The sequencing of the human genome was a vital stepping stone towards the understanding of the biology of human cells [2]. However, it was clear from the outset that the genetic information alone will not be sufficient to fully understand the observed diversity in a complex organism as it will be identical between all cells of the same individual. On the other hand, studying the transcriptome is particularly useful for unraveling the cell identity. It can reveal the active functional elements of the genome and the molecular constituents shaping cell phenotypes, but is difficult to study in bulk when the cells are averaged. While bulk transcriptomic analysis was first demonstrated in 1995 [3], the notable breakthrough for studying the biology of complex organisms came in 2009 with the first method for analyzing the transcriptomes of single cells [4]. However, the initial attempts were slow and costly, restricted to only a handful of individual cells. As a result, broader use of single-cell transcriptome analysis was impractical and out of reach for many researchers. Ultimately, in 2015 a major technological breakthrough came in the form of droplet microfluidic platforms for high-throughput single-cell analysis, allowing for the analysis of tens of thousands of single cells in a single experiment [5, 6]. These technological advances have galvanized the research community and ushered in a new age of complex organism biology. New discoveries followed immediately after – a new type of dendritic cells (DCs) was identified, a new cell type responsible of cystic fibrosis was identified and a number of tissue atlases were constructed to reveal new rare cell types. Success of these initial efforts have proven the potential of this new technology. Its importance was further solidified by the establishment of the Human Cell Atlas consortium that aims to create a comprehensive reference map of all human cells [7, 8]. Naturally, given the early stage of the research field, the single-cell transcriptome analysis still faces many challenges on different fronts. From a technological viewpoint, single-cell transcriptome analysis needs to become more efficient, robust, and cheaper to enable its widespread use and applications beyond fundamental research. On the other hand, even in its imperfect form, the technology is indispensable for complex biological system analysis and will undoubtedly lead to meaningful discoveries.

Cells are not static but dynamic systems that can acquire different phenotypes. One of the well-studied and fascinating phenotypic transitions is the Epithelial-Mesenchymal Transition (EMT). During the EMT process, polarized epithelial cells progressively lose their attachment to each other and the basal membrane, assuming a spindle-like morphology and becoming motile. In the human body, EMT is observed in different biological contexts and is accordingly classified into three types [9, 10]. Type 1 EMT is associated with development processes. EMT in the context of wound healing and tissue regeneration in a fully developed organism is assigned to Type 2. In this setting, EMT is not only beneficial but can also lead to organ fibrosis. Finally, Type 3 EMT is observed in cancer. Together with genetic and epigenetic changes, EMT circuitry promotes tumor formation, survival, and is critical for metastasis formation. EMT has been studied for over 30 years. Many insights have already been gained about the EMT process and its significance in different biological processes [11]. Initially, the research focused on understanding the role of EMT in development. However, over the last two decades, increasing efforts have been devoted to studying EMT in the context of cancer. Given its significance in pathology, EMT is an attractive target for therapy. However, many open questions remain, and to date, it has been hard to capitalize on the extensive knowledge accumulated due to the complexity of the underlying biochemical circuitry.

Whereas phenotypic changes during EMT is a natural process important for healthy organism development and survival, phenotypic and genotypic changes during tumor development often lead to serious health consequences. Cancer cells, typically through the process of mutagenesis, acquire new phenotypes that can impede the survival of the entire organism. For example breast cancer is the most frequent cancer type found in women [12]. An estimated 2.1 million women were diagnosed in 2018, with over 600,000 patients succumbing to the disease [13]. The prevalence of breast cancer is on the rise, with a 3.1% yearly increase in cases globally [14]. Overall, breast cancer is a well-studied disease, and systematic treatment guidelines are established. As a result, 70-80% of patients with early-stage, non-metastatic disease are cured [12]. By contrast, patients with advanced (metastatic) disease are considered incurable, and their median survival is 2-3 years [15]. Thus, improvements in advanced breast cancer care are as relevant as ever.

On a molecular level, breast cancer is a highly heterogeneous disease defined by different genetic mutations and diverse tumor microenvironments. Recently, immune infiltration of breast cancer tumors has been receiving renewed attention, as it is becoming a clearly potent therapeutic target. However, clinical developments in breast cancer immunotherapy have been slow as compared to other tumor types. Historically, breast cancer tumors have been considered immunologically quiescent or ‘cold’. However, recent evidence shows that this is not true as a significant amount of immune cells is detected in most breast cancer tumors [16, 17]. Furthermore, in line with the observed heterogeneity of breast cancer, the immune cell subset also heavily depends on a particular tumor type [18]. Overall, while a lot is now known about separate immune cell types and signaling pathways operating in the breast TME, the full picture is far from clear, making it hard to draw general conclusions. Understanding the intricate interplay between different cell types in different breast tumor types is the key to the successful application of immunotherapy in breast cancer care.

Study goal

To optimize and apply high-throughput droplet microfluidics based single-cell transcriptome analysis platform to studying complex biological systems.

Objectives

- Describe a detailed protocol for implementing the high-throughput single-cell transcriptome analytical platform
- Optimize the single-cell RNA-Seq library preparation
- Characterize the EMT process using the HMLE cell model system
- Predict and validate the targets of the ZEB1 transcription factor, which plays a key role in EMT process
- Construct an immune cell atlas of breast cancer patients
- Characterize the immune cell infiltrate of breast cancer tumors

Scientific novelty

In this work, a detailed protocol for high-throughput single-cell transcriptome analysis (scRNA-Seq) using droplet microfluidics was described for the first time. This method is termed “inDrops” and the described procedure, which is reported in a high-tier journal, enables the non-expert users to conduct transcriptome studies on thousands of single-cells. Furthermore, the optimizations of the inDrops protocol described in this work have led up to 10-times more efficient capture of unique transcripts of the individual cells as compared to the previous protocol version. The optimizations also provide additional cost savings, an important consideration when working with multiple samples. Results presented in this thesis are relevant not only for the inDrops method but also for other single-cell transcriptome analysis techniques as many of them share at least some of the protocol steps.

In the second part of this work the described scRNA-Seq platform was used to study a complex biological process employing a well established model system. To this end, the Epithelial-Mesenchymal Transition was characterized at single-cell resolution for the first time using HMLE cell model. Results provide an unprecedentedly detailed view of the EMT process. Given the novelty of scRNA-Seq technology, the initial part of the analysis was focused on investigating whether the single-cell results capture the already known biological features such as activation of certain gene pathways. Results presented in this work show that single-cell transcriptome analysis can reliably uncover the inctricate biological insights precisely matching the cellular mechanisms characterized through decades of research. It proves that the droplet-based scRNA-Seq technology can provide unbiased and biologically relevant data. For example, in this study it was shown that epithelial cells undergoing EMT first revert to a stem-like state before acquiring the mesenchymal phenotype; a phenomena that has been questioned and debated for a long time. One drawback of scRNA-Seq technology, however, is the sparse nature of the data. Because of this, weak gene-gene corelations can be obscured in scRNA-Seq data. To overcome this limitation, we have shown how the so called zero-inflated distributions of gene-expression matrices can be addressed using imputation algorithms based on data diffusion. The imputed data was then used to accurately predict the targets of the master-regulator transcription factor of the EMT process – ZEB1, revealing a previously unapreciated extent of trancsriptional reprograming that occurs during the EMT. The results presented in this section of the thesis demonstrate how single-cell transcriptome analysis can be used to discover regulatory gene-gene relationships without the need for system perturbations. Such approach is particularly valuable for clinical sample analysis and could ease the discovery of rogue regulatory pathways in disease.

Having shown that droplet-based scRNA-Seq can successfully reconstitute complex cellular processes, the final part of this thesis focuses on clinical samples. In collaboration with MSKCC (USA) clinicians for the first time, an atlas of immune cells was constructed combining over 62000 individual immune cells isolated from eight patients and spanning normal and cancerous breast tissue, as well as peripheral blood and the lymph node. This atlas revealed a vast diversity in immune cells of both the adaptive and innate immune systems. Results confirm a high degree of variability between patients, as could be expected. Importantly, the diversity of T cell phenotypic states was observed to be significantly expanded in breast tumors as compared to normal breast tissue, indicating that the complex signaling and local niches in tumor microenvironment plays a significant role in shaping the host immune response. The top three components contributing to this phenotypic expansion are the T cell activation, terminal differentiation, and hypoxic response. The results presented in this study show gradual cell ordering along the activation component and argue against a prevalent view of activated T cells rapidly traversing through sparse transitional cell states toward a few predominant, discrete, and stable states, including Treg, effector, memory, and exhausted T cells. Similar results have also been recently reported in the context of autoimmune disease [19]. Finally, the results of this work, in concordance with several recent reports in the field, prove that macrophage activation states exist as continuum of states and not as mutually exclusive discrete states. Results presented in this work solidify and reinforce recently reported similar findings from bulk analyses of tumor-associated macrophages.

Defending statements

- The efficiency of droplet microfluidics based single-cell transcriptome analysis platform can be increased by optimizing individual steps in the workflow.
- Imputation algorithms are effective for recovering gene-gene relationship information in sparse single-cell transcriptomics data.
- Single-cell transcriptome analysis can accurately predict activation targets of transcription factors.
- Tissue microenvironment affects the diversity of immune phenotypic states.
- T cells in breast tumor span a phenotypic state continuum that is shaped by local niches of the tumor microenvironment.

1 LITERATURE REVIEW

1.1. Single-cell transcriptome sequencing technologies

1.1.1. The development of single-cell sequencing technologies

Technological advancements often lead to new biological discoveries. A perfect example is an invention of the microscope in the 17th century that led to the discovery of cells [20]. Today it is understood that cells are the basic functional units of life [1]. A major stepping stone towards the understanding of how a human cell functions was the sequencing of the human genome. However, the genetic information alone is not sufficient to understand the observed diversity in a complex organism. What is need is the study of the transcriptome, which could reveal the active functional elements of the genome and the molecular constituents of different cells [21]. Therefore it is not surprising that the development of single-cell transcriptome sequencing (scRNA-Seq) has transformed the analysis of complex biological systems [22, 23]. The power of scRNA-Seq was first demonstrated in 2009 and scaled rapidly over the next decade [4, 24]. Initial efforts relied on manual cell separation and focused on profiling a few cells at a high depth [25-27]. However, it was soon realized that the power of the technology lies in the sampling of many cells in parallel [28]. A substantial number of different methods have been published over the years [29-31]. Each of these has aimed to increase the throughput and decrease the cost of the analysis (Table 1.1) and a few different technologies were commercialized along the way (Table 1.2). The commercial systems offer a straightforward workflow and high-quality data, yet they tend to have a higher price tag as compared to the in-house protocols. Combining the diverse academic and industrial developments of the scRNA-Seq technology has grown in importance in many branches of life sciences. The formation of an international initiative best exemplifies this. The Human Cell Atlas consortium aims to create a comprehensive reference map of all human cells [7, 8]. To date, the initiative has over 1800 individual members from 71 countries. It has the potential to transform cell biology similarly to the Human Genome Project at the turn of the century [31, 32].

Table 1.1. Summary of single-cell transcriptome analysis technologies. * marks costs indicated in the original publication. • marks the „inDrops" method.

Year	Number of cells	Cell isolation technology	Cost per cell	Focus	Citation
2009	30	Manual	N/A	First demonstration	[4]
2011	92	Manual	50$*	Sample multiplexing	[28]
2013	91	Integrated fluidic circuits	9-25$ [29]	Automation	[33]
2014	1536	FACS, Liquid-handling robots	1.3$ [29]	Automation and throughput increase	[34]
2015	11 149	Droplet microfluidics•	0.1$ [35]	Cost reduction and throughput increase	[5]
2015	44,808	Droplet microfluidics	0.1$ [29]	Cost reduction and throughput increase	[6]
2017	14 218	Picowells	0.1$*	Ease of use	[36]
2017	42 035	In situ barcoding	0.03$*	Cost reduction and throughput increase	[37]
2018	156 049	In situ barcoding	0.01$*	Cost reduction and throughput increase	[38]

Table 1.2 Summary of commercial scRNA-Seq solutions. Costs as of April 2020. Costs do not include sequencing costs.

Company	Cell isolation technology	Cost per cell	Cells analyzed per run	Analysis type
10X Genomics	Droplet microfluidics	0.5-1$	Up to 24 000	3' counting WTA
Dolomite bio	Droplet microfluidics	0.25-0.4$	Up to 50 000	3' counting WTA
1cellbio	Droplet microfluidics	0.2$	Up to 40 000	3' counting WTA
Rio-rad	Droplet microfluidics	1$	Up to 1200	3' counting WTA
Celsee	Microwells	N/A	Up to 40 000	3' counting WTA
BD	Microwells	0.6$	Up to 10 000	3' counting WTA and targeted panels
Fluidigm	Integrated fluidic circuits	3$	Up to 800	3' counting or Full-length WTA
Split Bio	*In situ* barcoding	N/A	Up to 100 000	3' counting WTA

1.1.2. Single-cell transcriptome amplification strategies

scRNA-Seq protocols rely on reverse transcription (RT) reaction to barcode the cell transcriptome. Using specifically designed barcoded RT primers, the mRNA of individual cells is copied and converted to copy DNA (cDNA) molecules. The primers have a few distinct features: i) RNA capture sequence, ii) a UMI for digital transcript counting, iii) a barcode that will be common to all transcripts from a single cell (cell barcode), and iv) a standard sequencing adapter that can also be used as a PCR handle (Figure 1.1, panel A) [31]. Most protocols rely on having poly(T) sequences at 3’ end for capturing polyadenylated RNA, thus efficiently excluding rRNA and tRNA from further analysis. Specific protocols are available for the analysis of totalRNA [39, 40] or miRNAs [41]. Similarly, commercial options for targeted panel analysis at single-cell resolution are available (Table 1.2)

The amount of totalRNA in individual cells ranges from 1-50pg [42]. Such a low amount of material means that the barcoded cDNA needs to be amplified before the sequencing library is produced. There are two strategies for cDNA amplification: exponential amplification by PCR and linear amplification by IVT(Figure 1.1, panel B). Amplification by PCR requires a second adaptor sequence to be added to the cDNA molecule. This can be done by utilizing the intrinsic terminal transferase activity of the M-MuLV reverse transcription enzyme. The enzyme tends to add a few nucleotides (mostly cytosines) at the 3' end of the cDNA molecule [43]. This short sequence is then used as an annealing site for the template switching primer (termed TSO). This primer serves as a template for the reverse transcription enzyme to synthesize the end of cDNA molecule. Because TSO sequence will be identical for all cDNA molecules this site can be used as a primer binding site during PCR [28] (Figure 1.1, Panel B). Alternatively, poly(A) tail can be added to the 3' end of the cDNA molecule by terminal deoxynucleotidyl transferase. This tail is then used as a priming site for the second strand synthesis, during which a second PCR primer site is introduced [4, 44]. The linear amplification of cDNA by IVT requires a T7 promoter sequence to be included in the RT primer (Figure 1.1, Panel A). After producing a double-stranded cDNA molecule in the second strand synthesis reaction, the T7 protomer is used to produce multiple copies of antisense RNA (Figure 1.1, Panel B) [45, 46].

Regardless of the strategy used, amplification leads to noise and bias in the final data [29, 47]. Due to its linear nature, the IVT amplification strategy is less prone to produce noise [48]. However, this comes at the cost of additional downstream protocol steps, more hands-on time, and an overall longer protocol. On the other hand, limiting PCR cycles can help to reduce the noise during exponential amplification [29]. Furthermore, amplification noise and biases are corrected by UMI counting during the data processing step [49, 50]. To achieve this, the UMI needs to be sequenced together with the cDNA molecule. Such an approach analyzes the transcriptome by digital counting of 3' or 5' transcript ends, and full-length transcript analysis is not possible [31, 48]. While this is a cost-effective transcript quantification strategy, it means that largely no sequence information is retained. Detection of splice variants, alternative transcripts, single-nucleotide variants, and fusion transcripts is possible only with a full-length transcriptome sequencing protocol [48]. It is important to note that such an approach is significantly more expensive [29]. Utilizing long-read sequencing technologies allows retaining UMI for digital transcript counting and full-length sequence information [51]. However, the sequencing throughput of long-read technologies is not yet sufficient for transcriptome-wide quantification. A recently developed strategy allows researchers to utilize UMIs transcript counting and to partially reconstruct the sequenced transcripts *in silico* [52]. Such an approach enables digital transcriptome quantification as well as assigning particular transcripts to specific isoforms and allelic origin.

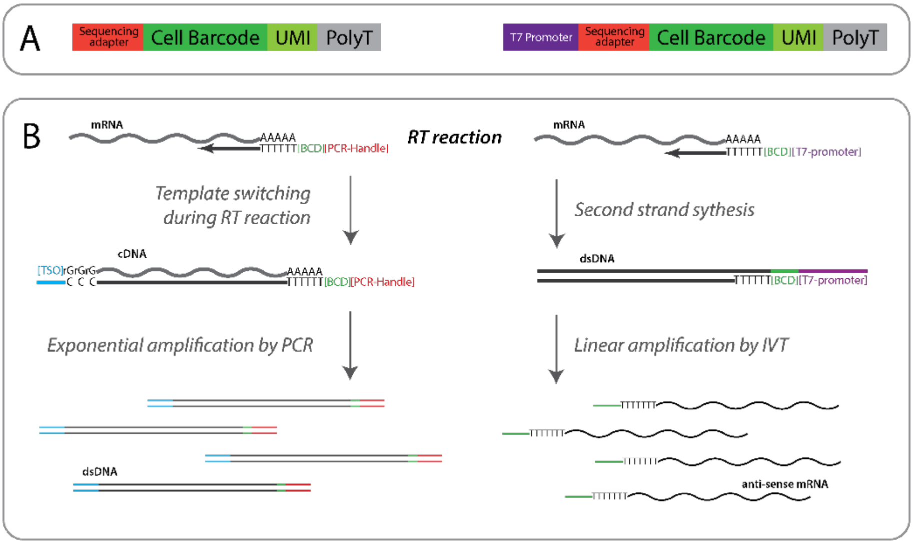


Figure 1.1. Panel A: RT primer structure schematics. Panel B: Library amplification strategies.

1.1.3. Single-cell transcriptome barcoding platforms

One of the main technical challenges of the scRNA-Seq process is the compartmentalization of the transcriptome barcoding reaction. It is a crucial step that, to a large extent, determines the throughput of the analysis. The aim is to compartmentalize the single-cell transcriptome barcoding reaction in such a way that only a single cell would get a single cell barcode (Figure 1.1). Several different strategies exist for the isolation of single cells. Cells can be sorted into microwell plates, isolated in integrated microfluidic circuits, encapsulated in microfluidic droplets, distributed in microfluidic picowells, or permeabilized and barcoded *in situ* (Figure 1.2) [31]. It is important to note that most scRNA-Seq methods require a single cell suspension as the input [31]. This means that tissues need to be dissociated into single cell suspensions before transcriptome barcoding. The preparation of high-quality single-cell suspensions is vital for high-quality data [48]. Tissue dissociation procedures can be a source of significant noise and bias in the final data and can lead to particular cell type loss [53-55]. One way to circumvent the fresh tissue handling issues is to use single nuclei from frozen tissues for transcriptome barcoding [53, 56-58]. Analyzing RNA from nuclei can provide the same cell type deconvolution information as fresh tissue sample analysis [59]. However, single nuclei sequencing typically results in reduced transcriptome capture efficiency [31, 48].

The first scRNA-Seq protocols relied on manual isolation of cells into individual tubes [4, 27]. Advances in sample multiplexing allowed to increase the scale of the analysis and to adopt microtiter plates [28]. Cells can be distributed into wells by limiting dilution. However, more often, fluorescence-activated cell sorting (FACS) is used to arrange single cells into microtiter plates or custom microwell arrays (Figure 1.2, Panel A) [34, 60-62]. Each well has a unique primer that contains a unique cell barcode sequence. Using FACS to isolate cells facilitates the exclusion of dead or damaged cells, and allows to enrich for target cell populations (e.g., through surface marker labeling). However, specialized flow cytometers and liquid handling robots required by such protocols are expensive and require dedicated staff [24]. Furthermore, relatively large reaction volumes mean that cost per cell is high due to the amount of consumables used and would it be prohibitively expensive to process large cell numbers (Table 1.1) [29]. Moreover, studies have shown that cell sorting can have a distinct effect on gene expression profiles [63].

Three different issues have to be addressed to increase the cell barcoding throughput: cell compartmentalization made fully automated, reaction volumes scaled-down, and the number of compartments drastically increased. Microfluidic technologies can be successfully employed to address these challenges [64]. Initially integrated fluidic circuits (IFCs) were used to automate cell capture (Figure 1.2, Panel B) [33, 65]. However, IFC did not see widespread adoption due to the limited throughput and high associated costs (Table 1.2) [24]. A real breakthrough in single-cell transcriptome barcoding came with the development of droplet microfluidics platforms that can simultaneously address all of the challenges specified above (Figure 1.2, Panel C) [5, 6]. Such platforms rely on microfluidic chips to combine aqueous phases with inert oil to produce monodisperse droplet emulsions. Typically a few different aqueous phases are infused into a chip containing: cells, barcoding beads (each bead has covalently attached unique primers), and RT reagents. Cell loading is governed by Poisson distribution, and only a fraction of droplets contain a cell. Using dilute cell suspensions ensures that double cell events (termed doublets) are rare. Depending on the particular platform, barcoding beads are introduced into droplets in a random - Poisson [6] or controlled manner [5]. Combined with the random cell loading, this means not every cell gets barcoded. Depending on the particular platform, only 2-4% [6] or over 75% [35] of cells will get a barcode. This limitation makes the droplet-based platforms less attractive when rare cell populations need to be sampled or when the sample size is limited (less than 2000 cells) [31]. On the other hand, droplet microfluidics offers unparalleled throughput allowing users to barcode over 50000 cells in a single experiment. Furthermore, several droplet microfluidic platforms were successfully commercialized (Table 1.2) and have been widely adopted by the research community [66].

As an alternative to droplet microfluidics, picoliter well arrays (picowell platforms) can be used towards the same goal (Figure 1.2, Panel D) [36, 67-69]. Separate wells on the array act as separate compartments for single-cell transcriptome barcoding. The same barcoded beads that are used in droplet microfluidic platforms are delivered into individual wells by gravitational settling. Occupancy of wells by beads is limited by geometry in such a way that only a single bead fits a single well, while Poisson statistics governs cell loading. Thus a diluted cell suspension is used to avoid doublets. One important feature of picowell platforms is that no special microfluidic equipment is required for the experiments [24, 36]. However, picowell platforms require more hands-on time. Moreover, the throughput of such platforms is limited to the array size. For example, the Seq-Well platform utilizes an 86000 well array, and it is possible to barcode up to 8000 cells in a single experiment [36].

Further barcoding throughput increase and cost reduction were enabled by *in situ* single-cell barcoding technologies (Figure 1.2, Panel E) [37, 38]. These platforms utilize a fixed and permeabilized cell or nuclei as a transcriptome barcoding compartment. Barcode diversity is generated by the split-and-pool strategy. Fixed single cells or nuclei are permeabilized and deposited into 96-well plate were each well contains a unique primer (unique cell barcode). Each well may get 10-100 cells, and transcripts are barcoded *in situ* by adding barcode sequences via RT or ligation reactions. Cells are then pooled and deposited into a different 96-well plate containing different barcodes, and the barcoding process is repeated. Such an approach allows to combinatorically scale the barcode diversity – 96^n (where n is the number of barcoding plates used). *In situ* barcoding, approaches allow to barcode over 100000 cells in a single run in the most cost-effective manner (Table 1.1) and do not require the use of any sophisticated equipment. However, no independent benchmarking of *in situ* barcoding platforms has been performed yet, and it remains unclear how well the protocols perform on different cell types and in the hands of independent users.

It is important to note that both the droplet and the picowell based platforms have a lower transcript recovery than the microplate-based protocols [29, 31, 48]. Most recent microtiter plate platform improvements allow to capture up to 80% of the cell transcriptome [52]. However, the sensitivity (ability to detect genes that are expressed at a low level – a few molecules per cell) is not dependent on the cell compartmentalization platform [30]. Therefore, choosing which platform to use depends on the aim of the study. If many cells need to be profiled, droplet or picowell based platforms are the most cost-effective solution. If, on the other hand, the cell sample size is limited, and transcriptomes need to interrogated at a high depth – microtiter plate based platforms are a better choice. Furthermore, *In situ* barcoding platforms hold much promise for ultra-high throughput scRNA-Seq applications. However, external benchmarking is needed before these platforms can be directly compared to other existing solutions.

Over 50 different single-cell transcriptome barcoding protocols and protocol variants exist [31]. Specific protocols differ in their features, throughput, workflow duration, and required equipment. Furthermore, protocols are continuously being updated and improved. High protocol diversity makes it difficult to point out the best one. Several detailed protocols have been published [35, 60, 70], and different commercial options are available (Table 1.2). Altogether, this makes single-cell transcriptome analysis widely accessible to biological researchers. Moreover, single-cell analysis is fast becoming a tool that is transforming the way complex biological systems are analyzed.

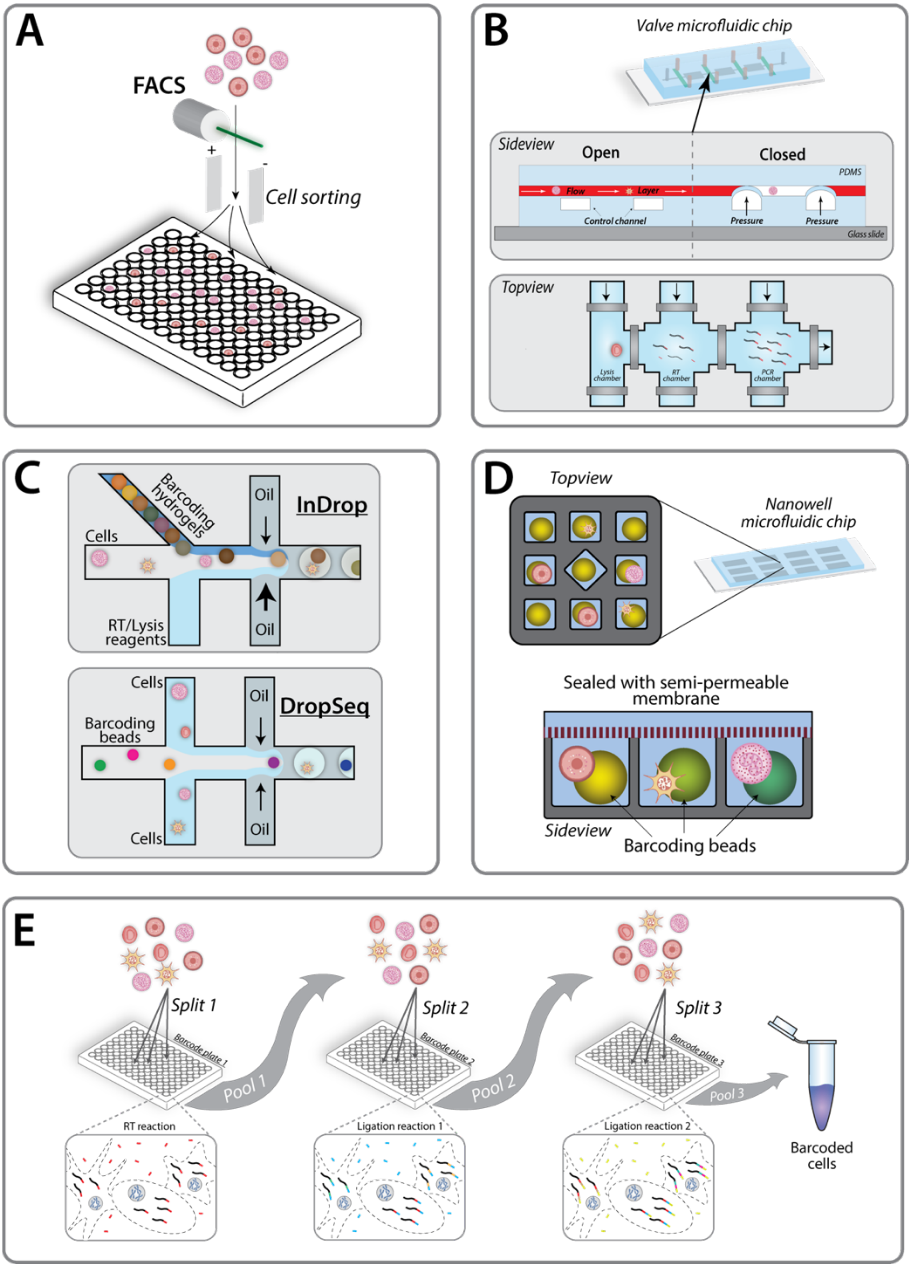
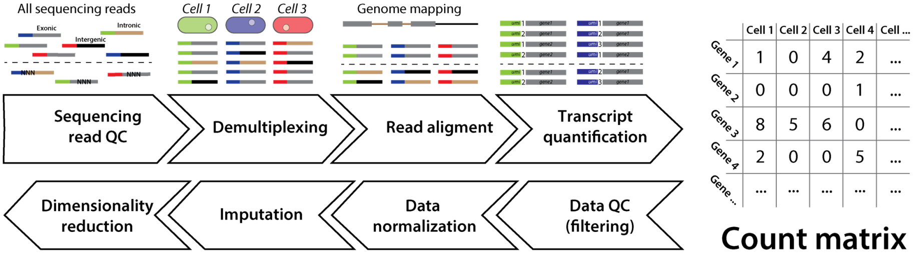


Figure 1.2. *Different single cell transcriptome barcoding platforms. Panel A: Cells are sorted into 96-well plates with FACS. Panel B: Valve microfluidics platform. Panel C: Droplet microfluidics platform. Panel D: Nanowell microfluidics platform. Panel E: Split-and-poll method for barcoding single-cell transcriptomes.*

1.2. Single-cell transcriptomics data analysis strategies

1.2.1. Overview of single-cell transcriptomics data analysis

Single-cell transcriptomics data analysis is a multistep process. It can be divided as follows: pre-processing, cleaning and normalization, imputation, dimensionality reduction and visualization, downstream analysis. Each step has different tools and considerations associated with it that are detailed in the following sections. Overall, scRNA-Seq data analysis is a fast-growing research field that, to date, lacks standardization. Over 600 different analysis tools exist, and they are implemented in a variety of programming languages [71, 72]. A typical single-cell analysis workflow consists of a collection of independently developed tools. However, integrated environments – analysis platforms have been developed to facilitate data movement between algorithms and improve user experience different [71]. A few of the most popular command-line platforms are Scater [73], Seurat [74], and Scanpy [75]. Graphical user interface platforms have also been developed – Granatum [76], ASAP [77], FASTGenomics [78]. While such applications are more convenient to use, they provided limited analysis flexibility, and their use is not widespread. Efforts to standardize scRNA-Seq data analysis are underway in the framework of the Human Cell Atlas project [79]. General guidelines are already being put in place [72]. However, new tools are constantly being developed while their benchmarking naturally lags behind. Often, results from benchmarking studies indicate that no single algorithm can provide the best result in all cases [80-82]. Therefore scRNA-Seq analysis will keep requiring highly skilled experts to deliver reliable results. A particular area of promise is deep learning algorithms. Having transformed fields like computer vision and natural language processing, these algorithms are starting to be increasingly applied in genomics and single-cell analysis [83, 84].



	Cell 1	Cell 2	Cell 3	Cell 4	Cell ...
Gene 1	1	0	4	2	...
Gene 2	0	0	0	1	...
Gene 3	8	5	6	0	...
Gene 4	2	0	0	5	...
Gene ...	...	...	...	...	...

Figure 1.3. *Data pre-processing and subsequent analysis workflow steps.*

1.2.2. Single-cell transcriptomics data pre-processing.

The output of scRNA-Seq experiments is raw sequencing data that needs to be processed before any meaningful insights can be gained. The first step of the analysis process is called data pre-processing. The goal is to generate matrices of transcript counts (count matrices) or read counts (read matrices), depending on the type of scRNA-Seq analysis – digital transcript counting with UMIs or full-length transcript sequencing. Pre-processing workflow has a few different steps (Figure 1.3). First, sequencing read quality is evaluated, and low-quality reads are removed. The most popular tool for this is the FastQC algorithm [85]. Sequencing reads of adequate quality are then sorted by unique cell barcodes (a process called "demultiplexing"). This step differs depending on the particular scRNA-Seq protocol used to barcode single cells. If a full-length transcriptome protocol is employed typically, the cell barcode will correspond to the library index read [25, 60]. For the digital transcript counting approaches demultiplexing process is more complicated because the cell barcode sequence comprises only part of sequencing read [48]. Usually, dedicated algorithms within the pre-processing pipelines perform demultiplexing without additional user input. If the expected barcode list is known, the process may involve sequencing error correction [5, 86].

Once reads are demultiplexed, they are then passed onto alignment algorithms. Alignment can be done to the genome or the transcriptome. It is recommended to map the reads to the genome as single-cell, and particularly single-nuclei transcriptomic data contains a high fraction of intronic and intergenic reads [31, 87]. Furthermore, mapping to the transcriptome alone has been shown to increase multimapping [88]. Typically 10-15% of the mapped reads span splicing junctions. Therefore splice-aware algorithms are preferable [31, 48]. Popular algorithms include TopHat [89] and STAR [90] aligners, both of which were developed for bulk RNA-Seq data. As the amount of data produced in scRNA-Seq experiments is increasing, the alignment speed is becoming an important parameter. Faster alignment algorithms that rely on pseudoalignment were recently developed to address the scaling issue [91, 92]. Overall, when correct transcript mapping parameters are used, the choice of a particular algorithm appears to have little impact on subsequent analysis [31, 88].

After the reads are demultiplexed and aligned, the transcripts need to be quantified. Typically only reads that map to exonic loci are counted [93]. However, recently it was demonstrated that intronic reads could be useful for downstream analysis [94]. Transcript counting algorithms are specific to the scRNA-Seq protocol used to barcode transcriptomes. For full-length transcript sequencing protocols, validated bulk RNA-Seq transcript counting algorithms can be applied. A few popular ones are RSEM [95], Cufflinks [96] and HTSeq [97]. For scRNA-Seq data that relies on UMIs for digital transcript counting, specialized tools need to be used, which can account for sequencing errors in the UMI [98]. While a lot of specific tools exist for individual processing steps, single-cell data processing pipelines have been developed to integrate and automate the process and have been widely adopted. Popular pipelines - Cell Ranger[99], indrops [5], SEQC [86], zUMIs [87] take raw sequencing data and return a count or a read matrix that has the dimensions of barcodes x number of transcripts.

Once the count or read matrix is generated, quality control needs to be performed to ensure that all cell barcodes correspond to viable cells. Not every barcode will correspond to an actual single cell. Double cell events, apoptotic cells, ambient RNA, and empty compartments are sources of considerable noise in the data [72, 100, 101]. Usually, QC is performed based on three metrics: the number for counts per barcode, the number of genes per barcode, and the fraction of mitochondrial genes per barcode (Figure 1.4) [100, 101]. Filtering is performed by manual thresholding, and it is important to consider all three metrics together as relying on only a single one may lead to incorrect cutoffs [72]. For example, cells with a low number of captured transcripts and few expressed genes may correspond to an inactive cell population. Similarly, a high mitochondrial gene fraction may be indicative of an active respiratory process in the population. The next QC step is to filter out the genes that have low expression in the dataset. Correct thresholding is essential as it may impact cell population detection [72]. If, for example, genes that are expressed in less than 20 cells are filtered out, then it will become difficult to detect a population consisting of fewer than 20 cells.

Specialized algorithms have also been developed to reduce particular noise in the data. A few different algorithms for doublets detection and removal have been recently published [102-104]. Similarly, different strategies were suggested for correcting ambient RNA introduced biases in droplet-based platforms [105-107]. Overall, QC filtering is an iterative process. As data quality cannot be determined a priori, thresholds differ between the experiments [108]. Typically, whether the data quality is sufficient or not is judged based on the performance of the downstream analysis. Accordingly, it is often necessary to adjust QC thresholds multiple times during the analysis. Also, care needs to be taken not to use QC filtering to improve the outcome of any statistical tests that are performed at a later stage.

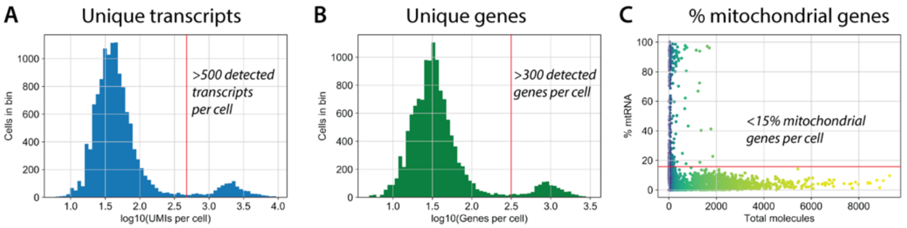


Figure 1.4. *Data filtering by different metrics. Panel A – data filtering by detected UMIs per cell. Red line indicates threshold. Panel B – data filtering by detected genes per cell. Red line indicates threshold. Panel C – data filtering by number of mitochondric transcripts. Red line indicates threshold.*

1.2.3.Single-cell transcriptomics data normalization and correction

Each entry in the count matrix represents a captured transcript. However, two identical cells may have very different values in the matrix due to the many technical variations in the protocols [31, 72]. For example, cell lysis, transcript capture, and cDNA synthesis efficiencies may differ between different compartments. Furthermore, different cells may be sequenced to a different depth. Such technical variations mean that differences between two cells in the matrix may have arisen only due to sampling effects. Data needs to be normalized to correct for these effects. It has been proven that data normalization is the most critical step in the single-cell data analysis process [109-111]. A standard normalization method is to divide counts for each gene in a cell by the total counts of that cell [31, 109]. This approach assumes that every cell in the dataset had the same number of mRNA molecules [112]. Such an assumption is false as samples are heterogeneous, and RNA amounts differ considerably between cells [113]. Furthermore, during the RT reaction, only part of the cell transcriptomes are captured due to the limited efficiency of the reaction [114]. This phenomenon is termed "dropout" and results in sparse data matrices [115]. Most values in such data matrices are 0, as no information for that particular transcripts is recorded. This makes algorithms used for bulk RNA-Seq normalization unsuitable for scRNA-Seq datasets [109, 111, 116]. Specialized algorithms have been developed for single-cell data normalization. Top among them is the Scran algorithm [111] that has been shown to be the best performer during independent benchmarking [72, 109]. The algorithm relies on estimating cell size factors based on a linear regression over genes after cells are pooled to avoid technical dropout effects. Non-linear normalization methods have also been developed and allow to account for more complex technical variations [117, 118]. These methods may be better suited for data where there are significant batch effects [72] [48]. For example, the microtiter plate based protocol exhibits a more significant variation in count depth between compartments than droplet-based platforms due to evaporation effects [30].

A single normalization method will not be appropriate for all single scRNA-Seq data types [72, 117, 119]. For example, full-length transcript sequencing data is more similar to bulk RNA-Seq data and can benefit from normalization methods that take gene length into account [120, 121]. On the other hand, digital transcript counting methods suffer from significant dropout effects and thus need specific algorithms to account for that. Therefore, care needs to be taken to select the most appropriate normalization algorithm for the given data. After normalization, the data matrix is typically log-transformed, which is useful for downstream applications – differential expression (DE) analysis [122] or batch correction algorithms[123].

Batch effects typically appear when cells are handled in distinct groups – different timepoints or different barcoding platforms. When integrating diverse data into a single dataset, it is important to remove the variation coming from purely technical handling differences. The best strategy for this is to avoid as much of technical variation as possible in the experiment design stage. Recently different methods have been proposed for sample pooling into a single barcoding experiment by tagging cells in a particular sample with a particular nucleotide sequence – i.e., a synthetic transcript. This can be achieved by tagging cells with antibodies [124], lipids [125], or by transfection [126]. However, this may not always be possible. For example, data from different experiments may need to be integrated in a retrospective manner. A widely used batch correction method for bulk RNA-Seq data – ComBat [127] can be successfully applied to scRNA-Seq data as well [123]. Several different algorithms for data integration specific to scRNA-Seq data have been developed [74, 128, 129]. Nevertheless, little systematic comparison between them exists, and general guidelines for their application have not been developed [72].

Data normalization aims to correct for differences in transcript sampling. It is a crucial step in data analysis workflow that can introduce biases if performed improperly [109, 112]. On the other hand, separate algorithms exist for addressing further sources of unwanted variation in the data – batch effects, cell cycle effects, and dropout [72]. It is important to note that these that it is not always appropriate to correct for these factors. The decision depends on the intended downstream analysis and overall experiment goal.

1.2.4. Single-cell transcriptomics data imputation

One of the most prominent sources of noise in scRNA-Seq data is dropout [111, 115, 130]. This term refers to 0 values in the data matrix for particular gene expression in a particular cell. There are two explanations of gene expression being zero. Either the gene was not being expressed at the time that the cell transcriptome was sampled (biological zero) or transcripts present in the cell were not detected due to technical limitations of scRNA-Seq protocols (technical zero). Correcting for this effect means determining which of the 0 values in the data matrix are true zeros and inferring the values lost due to technical noise. This process is called imputation. The concept of imputation is not new to scRNA-Seq. It has been successfully applied in GWAS data analysis to infer the missing single-nucleotide polymorphism values [131]. However, typically imputation algorithms rely on a reference dataset - 1000 Genomes project in the case of GWAS [132]. The single-cell analysis field does not have extensive reference datasets yet, which means that the imputation algorithms have to rely on the supplied data to infer the missing values in it. Imputation algorithms for scRNA-Seq can be categorized into three distinct groups [133]. The first group of methods aims to infer the probabilistic model that would describe the data generation step. After such a model is defined, it can be used to identify (probabilistically), which 0 values are technical zeros and need to be imputed. Imputation is then performed in a number of different ways depending on the particular algorithm – regression, similar cell identification through k-means clustering, or dimensionality reduction. Typically algorithms combine a few different techniques for imputation to get the best result [133]. Some of the more popular methods in this category are – SAVER [134], ScImpute [135], bayNorm [136], and VIPER [137].

The second group of imputation algorithms relies on data smoothing to infer missing values. Similar cells are usually identified by looking into local neighborhoods in the high-dimensional expression space. Once similar cells have been determined, expression values for every cell are adjusted based on the expression values of similar cells. This means all entries in the data matrix –biological zeros, technical zeros, and non-zero values get adjusted. Therefore, data smoothing acts as data denoising and can be useful in revealing gene-gene correlations [72]. Some of the more popular algorithms in this category are – MAGIC [138], kNN-smooth [139], netSmooth [140], and DrImptute [141].

The third group of algorithms aims to reconstruct the latent (feature) space of the cells. Essentially, this means capturing the underlying biological signals in the data. This can be done in two ways. Either through matrix factorization (for example, PCA) or using unsupervised machine learning algorithms (autoencoders, deep neural networks) [133]. Matrix factorizations are linear transformations that aim to decompose the observed data into a number of factors-components. Similar approaches are used for dimensionality reduction purposes. Once these factors are identified, they can be used to reconstruct the data - perform imputation. Popular algorithms for this approach are - ALRA [142], mcImpute [143] and PBLR[144]. More recently, unsupervised machine learning algorithms have emerged that use autoencoders to learn the latent space [84]. Autoencoders are artificial neural networks for unsupervised learning. The goal of these algorithms is to learn an efficient representation of the data by reconstructing the input dataset as accurately as possible under provided constraints [145]. By learning the underlying identity function, autoencoders are able to generate an imputed matrix that has the zero values filled in. Utilizing autoencoders allows performing not only data imputation but also denoising and batch effect correction [84]. Popular algorithms in the field are – AutoImpute [146], DCA [147], DeepImpute [148], scVI [149]. Finally, a few algorithms have also been developed that incorporate external information as a reference for imputation. ADImpute [150] utilizes gene regulatory network information, SAVER-X [151] uses information from atlas-type resources, and SCRABBLE [152] uses matched bulk RNA-Seq data for imputation.

The main problem with imputation algorithms is circularity. Most algorithms rely on internal information in the dataset to impute missing values. In turn, this tends to artificially amplify large signals present in the data and smooth over small differences. Imputation can introduce false-positive results in downstream data analysis, particularly in differential gene expression analysis [82, 153]. Imputation also shows limited benefit when considering other downstream analyses like clustering or trajectory inference [82]. On the other hand, imputation methods are useful for recovering bulk expression profiles and log fold changes of individual genes between cell types (without accounting for cell-cell variability). This means that for gene-gene correlation analysis, imputation algorithms provide substantial improvement [82]. A systematic evaluation of imputation algorithms has revealed that MAGIC, kNN-smoothing, and SAVER algorithms outperform most other methods [82]. On the other hand, the MAGIC algorithm performed worse than many others in terms of identifying differentially expressed genes (DEG) while taking cell variability into account. Therefore, it is important to note that there is no single algorithm that performs the best under all circumstances [82, 154]. Furthermore, imputation algorithms, especially deep learning ones, depend on the parameter choice [82, 84]. Overall, it is considered that imputation is more useful for exploratory purposes rather than hypothesis generation [72].

1.2.5. Single-cell transcriptomics data dimensionality reduction and visualization

In theory, human single-cell transcriptomics dataset can contain expression values for over 20000 genes present in the genome [155]. In practice, this number is lower, due to incomplete sampling and variations in gene expression. A typical scRNA-Seq dataset will have over 15000 dimensions [72]. This means that any given cell in a dataset has over 15000 values associated with it and exists in a high-dimensional space. Such data is difficult to interpret and computationally taxing to analyze. However, it can be simplified by taking advantage of *the manifold assumption* [156]. A manifold is a mathematical construct that represents a significantly lower-dimensional structure that exists in the high-dimensional space. This concept can be applied to single-cell data analysis because gene expression is not random, and cells exist in defined cellular states. Furthermore, transitions between cell states typically are smooth as gene expression gradually changes. Therefore, in a high-dimensional scRNA-Seq dataset space, there is a lower-dimensional structure – manifold along which all the cells are ordered [156, 157]. The true structure of the data can be accurately captured by a considerably smaller number of linear or non-linear vectors as compared to the original number of dimensions. The process of determining those vectors is called dimensionality reduction. Having fewer dimensions allows improving the computational performance of downstream data analysis algorithms significantly. Furthermore, dimensionality reduction is essential for data visualization algorithms that aim to reduce the dataset into two or three dimensions [72].

The first step in dimensionality reduction is called feature selection. Not all the genes will be informative of the variations in the dataset, and the gene list can be filtered to keep only those genes that are informative of the variability in the data [33]. The set of genes that is selected for the analysis is called – Highly Variable Genes (HVGs). Gene selection is carried out after normalization and technical noise correct as various effects, for example – batch effects, can significantly contribute to gene variation [122]. A few different algorithms exist for determining HVGs. All of them use the relationship between variance, or its variations, and the mean as an indicator

[158]. A popular method for selecting HVGs is binning them by their mean expression and then using the highest variance-to-mean ratio to select an HVG in each bin [72]. Downstream analysis is typically robust to the exact number of HVGs selected [5]. However, because it is the first step of dimensionality reduction, it is recommended to select more rather than fewer HVGs [72].

Once HVGs are selected, the reduced matrix is then subjected to dedicated dimensionality reduction algorithms. The most popular linear method for dimensionality reduction is principal component analysis [159]. It maximizes the captured residual variance in each further dimension. PCA reduces the dataset into N principal components, where N is determined by plotting components by explained variance and then manually selecting a threshold. Each separate component explains a specific amount of variance in the data. Due to the linear nature of PCA, distances in the reduced dimensions have a consistent interpretation throughout the low-dimensional space. Therefore, particular components can be correlated to particular sources of variance in the data. This can be useful when inspecting the performance of data correction steps [123] or showing the importance of particular genes in the dataset [160]. PCA allows to reduce dimensions from thousands down to hundreds or even less and is often used as a pre-processing step for many downstream algorithms (visualization, clustering, differential expression analysis).

Single-cell transcriptomics data is inherently non-linear. Therefore, linear dimensionality reduction algorithms like PCA cannot capture enough information in two or three components to be useful in visualization. Manifold-learning (non-linear) dimensionality reduction algorithms are needed for visualization purposes. One of the most popular methods for dimensionality reduction and visualization is t-distributed stochastic neighbor embedding (t-SNE) [161]. The algorithm captures local similarities in the manifold, but underrepresents the global structure. Visualization is generated by placing similar cells (based on localized similarity in their gene expressions) close to each other and placing dissimilar cells far away from each other in the visualization space. The algorithm is good at grouping similar cells into clusters, but their arrangement and distances between them are essentially meaningless [162]. As a result, t-SNE tends to fragment natural progressions-trajectories into separate unrelated clusters and makes it hard to interpret continuous biological processes [163]. Furthermore, the visualization depends on the choice of algorithm parameters, which makes it challenging to implement efficiently. More recently, alternative manifold-learning algorithms have been proposed for visualization purposes - Uniform Manifold

Approximation and Projection method (UMAP) [164] and a graph-based tool SPRING [165]. Both of these methods are substantially better than t-SNE at preserving global manifold structure [164, 166]. Furthermore, UMAP scales well on large datasets and thus has seen rapid adoption [72, 164].

Non-linear algorithms can also be applied for summarization purposes. An efficient manifold-learning method for preserving global manifold structure is the diffusion map [167, 168]. Diffusion maps can efficiently capture trends in the data and contain them in relatively few dimensions [169]. Main uses for this dimensionality reduction method are in data imputation and trajectory inference algorithms. The diffusion map cannot be readily used for visualization as the algorithm returns more than two dimensions. However, the concept was recently extended to visualization methods by the creation of the PHATE algorithm [163]. It takes into account both near and far manifold-intrinsic distances when plotting cells in two dimensions. As a result, this algorithm is efficient at visualizing not only cell clusters but also trajectories or progressions.

Dimensionality reduction is vital for most downstream analysis algorithms and different analysis methods may require different type of dimensionality reduction. A few tools discussed above (PCA, diffusion map, tSNE, UMAP) have been widely accepted and implemented [72]. Additionally, deep learning based algorithms for dimensionality reduction have been proposed [170-172]. However, these algorithms heavily depend on parameters and are not trivial to implement successfully [173]. A special case of dimensionality reduction are visualization algorithms that aim to accurately represent highly-dimensional scRNA-Seq data in two or three dimension. As a result, visualization is probably the most important tool for intuitively exploring the underlying biology. However, generated plots should not be used to make conclusions about the underlying biology. Rather, they should serve as a medium for exploring results of downstream analysis algorithm (clustering, trajectory inference, differential expression) and can help to understand technical variations in the data (batch effects) [48]. Recent years have seen efforts to create interactive visualization tools that would allow to share and explore scRNA-Seq data in a convenient and accessible way [174].

1.2.6. Single-cell transcriptomics downstream analysis

Downstream analysis methods are a group of diverse algorithms that are used to gain insights and understand the underlying biological processes. ScRNA-Seq primarily focuses on analyzing sample heterogeneity at a single-cell level. Therefore, identifying new cell types and states is one of the main applications of the technology. Cell populations are identified by clustering cells based on the similarity of their gene expression. Clustering is typically performed in an unsupervised manner using machine learning algorithms. Most clustering algorithms use data after dimensionality reduction, because distances between cells in the original high-dimensional space tend to be too small to identify cell grouping [81]. One of the most widespread clustering algorithms is *k*-means clustering. The algorithm iteratively identifies *k* cluster centers (centroids) and assigns each cell to the closest centroid. The standard method for performing *k*-means is Lloyd's algorithm [81, 175]. It scales linearly with the number of points, which means that the algorithm can easily be applied to large data sets. One drawback of *the k*-means clustering is that the expected cluster number needs to be supplied to the algorithm. This number is typically unknown and must be determined heuristically [72]. Furthermore, the algorithm has a bias towards generating equal size clusters. Because of this, rare cell types can be hidden within a larger cluster. A few methods have been developed that aim to overcome this limitation – RaceID [176] and SIMLR [177].

Another clustering strategy that can be applied to scRNA-Seq data is hierarchical clustering. The algorithm sequentially combines cells into larger clusters or alternatively divides existing clusters into smaller sub-clusters. Such a strategy improves the ability to identify small clusters [81]. However, the approach is not scalable to more extensive datasets. Due to limitations of *k*-means and hierarchical clustering strategies, it is becoming increasingly popular to use community detection strategies for cell clustering [72, 81]. This strategy relies on constructing a *k*NN-graph where each cell is a node and then identifying communities of densely connected nodes within the graph. The most popular algorithm for identifying communities in a graph is the Louvain algorithm [178]. It was first applied to scRNA-Seq data clustering by the development of the PhenoGraph algorithm [179]. This approach is easily scalable to large datasets and does not require the user to input the expected number of clusters. In recent benchmarking studies, community detection algorithms have been shown to outperform other clustering algorithms [180, 181]. It is essential to keep in mind that, depending on the results of the clustering algorithm, a particular cluster may not correspond to a particular cell type. A single cell type can be disturbed over separate clusters. Alternatively, a single cluster may be comprised of a few cell types [72, 81].

Some biological processes cannot be efficiently described by discrete classification of cells into clusters. Cell development, activation, differentiation are continuous transformations that require a different approach [182]. Typically such processes are analyzed by trajectory inference algorithms that order cells along the *pseudotime* variable that corresponds to transition time. A large number of trajectory inference algorithms exist, and no single method shows clear superiority [80]. The topology of trajectories can be very different – linear, tree-shaped, cyclical, or even discontinuous graphs. Therefore, it is not surprising that different methods perform better for different datasets. While early trajectory inference methods required users to fix the topology beforehand [183, 184], more recent algorithms attempt to infer the underlying topology [185, 186]. Detailed guidelines for trajectory inference analysis have been recently proposed [80]. To test the robustness of the hypotheses, it is essential to validate the results by at least a few different algorithms. Also, care needs to be taken, as inferred trajectories may not represent actual biological processes and only denote transcriptional similarity [72]. On the other hand, if a complex data topology is identified, it could indicate that the underlying biology is more complicated than anticipated by the user [80].

Once cells are grouped into clusters or ordered along the *pseudotime* trajectory, cell identities need to be determined. Typically expression profile of the cells in one cluster is compared to the expression profile of the rest of the cells. This analysis is called differential expression analysis. It used to reveal differentially expressed genes (upregulated and downregulated) in a particular cluster. DE algorithms are extensively used in bulk RNA-Seq analysis [187]. Building on similar ideas, algorithms specific for scRNA-Seq analysis have been developed [115, 122]. However, a recent benchmarking of both bulk and specific scRNA-Seq DE algorithms in single-cell transcriptomic analysis has revealed little difference between them [120]. A differentiating factor is the computational efficiency where single-cell specific algorithms show significant improvements over bulk counterparts. DE analysis returns a list of genes that are specific to a particular cell cluster – population. Simple statistical tests (Wilcoxon rank-sum or *t*-test) are sufficient to determine the most robust genes in the list [72]. These genes are called the maker genes and are used to determine cell identity. The process is typically performed by manually referencing genes and gene sets in the literature to assign a particular cell type. Alternatively, cell ontology analysis [188] can be performed to identify ontology terms associated with a particular cluster [189-191]. Such approaches are labor-intensive yet appear to yield consistent results [81]. As more atlas-type data is becoming available, it is becoming easier to annotate the clusters manually. Furthermore, recently automated annotation tools have been developed. Scmap [192] and Garnett [193] algorithms use a reference source to annotate a dataset under investigation. However, cell identities and clustering can differ between experiments due to batch effects [182]. Therefore, using automated solutions may not always be appropriate.

1.3. Epithelial to mesenchymal transition

1.3.1. EMT in biological processes

Epithelial-Mesenchymal Transition (EMT) is a biological process that enables epithelial cells to assume a mesenchymal cell phenotype. During the transition, polarized epithelial cells progressively lose their attachment to each other and the basal membrane, and assume a spindle-like morphology while becoming motile (Figure 1.5) [9, 194]. The concept of EMT was first described after the surprising observation that cultured epithelial cells under the influence of microenvironment stimulus could change their morphology and become migratory [195]. The first evidence of EMT *in vivo* came when studying developmental processes in the chicken embryo [196]. It is now known that EMT occurs under a wide range of circumstances and is critical for development and tissue homeostasis as well as in various pathological processes – fibrosis and cancer [9, 11, 194]. It is important to note that initially, EMT was considered as a "transformation" [197]. However, today it is viewed as a "transition" because the process is gradual, and cells undergo many intermediate states often not reaching the final fully mesenchymal phenotype (termed partial EMT) [11, 194]. Furthermore, the reverse process Mesenchymal-Epithelial Transition (MET) has also been well documented [198]. The many intermediate states and reversibility underscore the complexity of EMT.

Based on the biological context EMT is classified into three types [9, 10]. Type 1 EMT is associated with development processes – implantation, embryo formation, and organ development. EMT in the context of wound healing and tissue regeneration in a fully developed organism is assigned to Type 2. In this setting, EMT is driven by inflammation, and if it persists, EMT can lead to organ fibrosis. Finally, Type 3 EMT is observed in cancer. Together with genetic and epigenetic changes, EMT circuitry promotes tumors formation, survival, and is critical for metastasis formation.

While EMT in different contexts produces different results, the molecular biology machinery is common to all three types. A core set of transcription factors (TFs) controls the EMT process: SNAIL, SLUG, TWIST1, ZEB1, and ZEB2 [199, 200]. These TFs are called master regulators of the EMT process (EMT-TFs). They are non-redundant and can interact in a complex temporal manner and different combinations. Together they regulate the expression of hundreds of genes associated with the EMT process. A set of widely accepted markers also exists to track EMT. The epithelial state is described by the expression of E-cadherin, occludins, and cytokeratins, while the mesenchymal state is characterized by vimentin, fibronectin, and N-cadherin expression (Figure 1.5) [10, 201].

EMT has been studied for over 30 years [11]. Many insights have been gained about the EMT process and its significance in different biological processes. Initially, research focused on understanding the role of EMT in development. However, over the last two decades, increasingly more focus has been devoted to studying EMT in the context of cancer [202, 203]. Given its significance in pathology, EMT is an attractive target for therapy. However, many open questions remain, and to date, it has been hard to capitalize on the extensive knowledge accumulated due to the complexity of the underlying biochemical circuitry. [11, 194, 200].

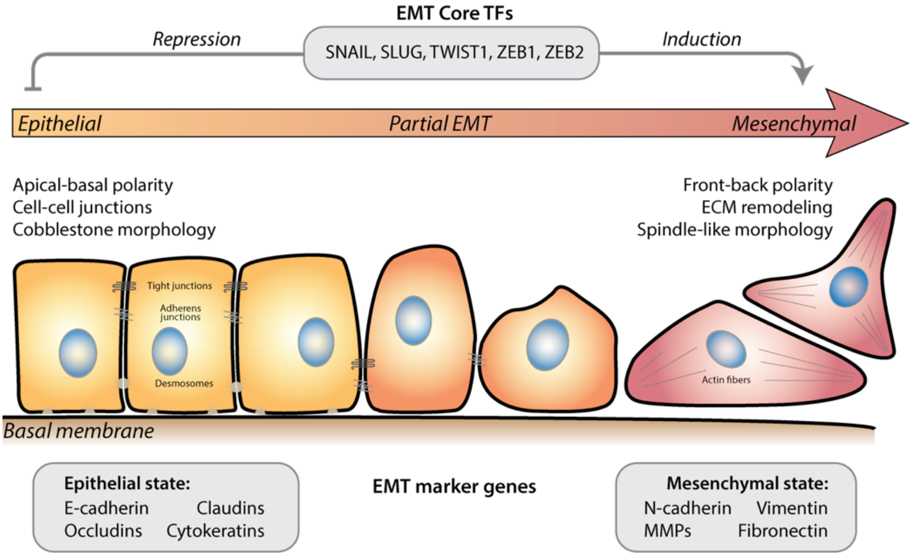


Figure 1.5. *EMT transition. Grey tables indicate important molecular markers of EMT process and core transcription factors.*

1.3.2.EMT gene regulatory networks

In all tissue contexts, EMT is characterized by a few key events: the dissolution of the epithelial cell-cell junctions, loss of apical-basal polarity, reorganization of the cytoskeletal architecture, increased cell motility, and, in many cases, the ability to remodel the extracellular matrix (ECM) [199]. The radical change in cell phenotype means that significant changes in gene expression occur throughout EMT. Initially, cells downregulate the expression of epithelial proteins, especially those included in cell-cell junction complexes [204]. The hallmark of EMT is the downregulation of E-cadherin, which leads to the dissolution of adherens junctions. Additionally, tight-junctions and desmosomes are also destabilized by repression of associated genes. These changes lead to the loss of epithelial barrier function and change in cell morphology [205]. Furthermore, it results in the loss of cell polarity, which is further supported by the repression of polarity complex genes [206].

E-cadherin is essential for epithelial phenotype, and its downregulation during EMT is balanced out by increased expression of mesenchymal cell adhesion molecule – N-cadherin. This change is called the ‘cadherin switch’ [207]. N-cadherin is essential for cell-cell interactions between mesenchymal cells and is utilized in various signaling pathways to facilitate cell migration [208]. Changes in cytoskeleton fiber composition, together with the repression of polarity complexes, is essential for enabling cell motility [199]. During the remodeling of internal cell structure, the composition of the intermediate filament changes: cytokeratin is repressed, and vimentin is activated. As this is an essential part of the process, vimentin is considered a marker for EMT progression [209]. The remodeling of ECM also plays a role in cell motility and is required for EMT progression. During EMT, the expression profile of integrin complexes that are responsible for cell interactions to the ECM changes to reflect changes in cell phenotype [210]. Epithelial integrins that mediate contact with basal membrane are downregulated [211], and integrins that promote cell migration are upregulated [212]. ECM remodeling and cell invasion are further enabled by increased expression of extracellular proteases [213]. These, together with integrins, act not only on the ECM but also on the EMT signaling pathways enabling cellular changes [199]. Finally, ECM remodeling is enhanced by the expression of ECM proteins, such as collagens and fibronectin, which is also an important marker of EMT [11].

Changes in gene expression are orchestrated by the core EMT transcription factors – SNAIL, SLUG, TWIST1, ZEB1, and ZEB2 [194, 200]. Their expression is activated through various signaling pathways, and often these

TFs function together [205]. Often targets of these TFs overlap, however they are not redundant and it has been demonstrated that each transcription factor has a distinct effector profile, and they all act both as activators and repressors (Table 1.3) [200]. SNAIL and SLUG belong to the same family of Snail transcription factors and repress epithelial genes by binding to the E-box sequences (CANNTG) through their C terminal zinc-finger domains [205]. The mechanism of the SNAIL TF effect on the E-cadherin promoter is well studied and provides an insight into how EMT is regulated [199, 200]. Upon binding the proximal promoter region, SNAIL recruits PRC2 and coordinates histone modifications. Interestingly, chromatin is marked by both active and repressive marks, which creates a poised state of the promoter. In the absence of activation signals, repression is maintained. However, upon appropriate signaling, gene activation can be rapidly achieved. Such control is common in many promoters in embryonic cells, and they are termed 'bivalent domains' [214]. Furthermore, other genes activated by SNAIL also exhibit similar poised promoter control [215]. Such a model of gene expression regulation contributes to the reversibility of EMT [199]. Besides acting directly on DNA, SNAIL and SLUG also cooperate with other transcription regulators. For example, SNAIL cooperates with ETS1 to activate protease expression [216]. Control over localization and degradation of Snail TFs is achieved through phosphorylation, which is tightly controlled by different signaling pathways active in EMT [199]. For example, p53 directly recruits SLUG for degradation, consequently preserving the epithelial phenotype in healthy adult tissue [217]. Furthermore, at translational level SNAIL and SLUG are repressed by a number of different miRNA [199, 218].

The next core EMT transcription factor TWIST1 belongs to the family of basic helix–loop–helix (bHLH) transcription factors. Similarly to Snail TFs, it downregulates epithelial and activates mesenchymal gene expression [219]. Functions of EMT-TFs are not redundant. For example, in cancer cells, TWIST1 can be induced under hypoxic conditions [220] and can repress E-cadherin independently from SNAIL [221]. Importantly, in the case of TWIST1 repression, the chromatin modification profile is different, and the E-cadherin promoter is fully repressed [222]. The precise mechanism of action of TWIST1 depends significantly on its dimer composition. It can form homodimers as well as heterodimers with other bHLH proteins [199]. The stability of TWIST1 is regulated by phosphorylation, albeit in a less complicated manner than SNAIL and SLUG [223].

Finally, ZEB1 and ZEB2 also bind E-boxes to repress or activate transcription through zinc finger domains [205]. Repression often involves the recruitment of a co-repressor - C-terminal-binding protein [224]. In the case of transcription activation, ZEB factors interact with transcriptional co-activators (p300 and PCAF) [225]. In different contexts, ZEB1 and ZEB2 have different effects [200], which could be explained by the structural differences between the two factors leading to a differential binding of co-activators and co-repressors [226]. Usually, ZEB TFs become active later in the EMT process. Research shows that SNAIL alone or in cooperation with TWIST directly activates ZEB1 expression [227]. Furthermore, ZEB1 promoter is controlled in bivalent manner described above, yet again underscoring the plasticity of EMT [228]. Little is yet known about ZEB TFs phosphorylation [229]. However, similar to Snail family proteins the translation of ZEB TFs is extensively controlled a network of miRNAs [199]. All five EMT master regulators have seemingly similar functions and can both repress epithelial genes and activate mesenchymal genes [199, 200]. However, they are not interchangeable, and the exact EMT molecular circuitry depends on the particular tissue type and signaling pathway involved [199]. Furthermore, a lot of additional transcription factors are involved in the regulation of EMT process in development and disease [199]. These transcription factors are often specific to a particular tissue or biological process and are not considered master regulators of the EMT process. Many separate studies on EMT-TFs have been performed, resulting in no small body of literature. However, there is only a limited amount of cases where EMT-TFs have been studied under the same conditions (Table 1.3). Therefore the full picture of the gene expression circuitry of the EMT remains far from fully understood [200].

Table 1.3. *EMT core transcription factors and their effector profiles. Adapted with permission from Nature Cell Biology [200].*

Target	Cell type	SNAIL	SLUG	TWIST1	ZEB1	ZEB2
ΔNp63	Mouse lung cancer cells	No effect	N/A	No effect	Represses	N/A
AXL	Breast cancer cells	N/A	No effect	N/A	Activates	N/A
CTGF	Breast cancer cells	N/A	No effect	N/A	Activates	N/A
CCL2	Human mammary epithelial cells	No effect	N/A	Activates	N/A	N/A
GRHL2	Human mammary epithelial cells	No effect	N/A	No effect	Represses	N/A
PTEN	Lung adenocarcinoma cells	N/A	N/A	N/A	Represses	No effect
VDR	Colorectal cancer cells	Represses	Represses	No effect	No effect	No effect

L1CAM	Endometrial carcinoma cells	Represses	Activates	N/A	N/A	N/A
SOD2	Transformed oesophageal epithelial cells	N/A	N/A	N/A	No effect	Activates

1.3.3.EMT signaling networks

Various signaling pathways can induce EMT (Figure 1.6). Usually, these pathways are activated by epithelial cell receptors binding ligands of stromal origin. Once EMT begins, cells can also be activated in an autocrine manner [11]. The main pathway of EMT activation is TGFβ signaling [11, 199]. TGFβ is a family of ligands, all of which activate specific transmembrane receptors (TGFβ receptors). These receptors have kinase activity, and once activated, they phosphorylate SMAD complexes. Depending on the particular ligand, either SMAD2 and SMAD3 or SMAD1 and SMAD5 are phosphorylated. Upon phosphorylation, in both cases, a trimetric complex with SMAD4 is formed (SMAD2–SMAD3–SMAD4 or SMAD1–SMAD5–SMAD4). Such complexes then migrate to the nucleus where they function as transcription factors and regulate a large number of genes. During EMT, SMAD complexes activate some mesenchymal genes directly (for example, vimentin and fibronectin) and also activate EMT master regulator TFs [219]. Interestingly, EMT-TFs can upregulate the expression of TGFβ ligands and form a positive feedback loop that helps to maintain EMT once it is induced [230]. TGFβ ligands can also directly regulate EMT by regulating EMT-TFs. For example, TGFβ can activate SNAIL by inducing the sumoylation of the protein, which is critical for its function in the EMT process [231, 232]. Additionally, TGFβ can induce EMT by regulating miRNAs and lncRNAs [233]. A prominent example is the miR-200 miRNA family, which inhibits the synthesis of ZEB1 protein [234]. By reducing the bioavailability of miR-200, TGFβ can promote EMT [235]. SMAD activation is considered the canonical TGFβ signaling pathway. However, depending on the particular ligand, TGFβ can also induce several other signaling pathways: ERK, p38 MAPK, PI3K–AKT, and RHOlike GTPases [199]. All of these pathways in different tissue contexts and under different circumstances can contribute to the EMT process.

EMT can also be activated through several other pathways – WNT, NOTCH, and through tyrosine kinase receptor (RTK) signaling [199]. The canonical WNT signaling pathway has long been studied in the context of EMT [236]. The WNT signaling pathway is critical during development - the deletion of WNT3A ligand disrupts embryogenesis [237]. The pathway begins by WNT ligands binding to Frizzled receptors, which triggers a series of events that lead to the nuclear translocation of β-catenin. It can act as a transcriptional cofactor and induces gene expression programs involved in differentiation, proliferation, and cell fate determination [238]. In adult tissues, the WNT pathway is active during wound healing related EMT [239]. Furthermore, the WNT pathway has been implicated in EMT related cancer progression and cancer stem cell (CSCs) formation [240-242].

The NOTCH pathway is another important EMT signaling circuit. NOTCH receptors bind the Delta-like or Jagged family ligands, and through proteolytic cleavage events, an intracellular fragment (NOTCH-ICD) is produced. This fragment translocates into the nucleus, where it promotes the expression of various gene programs related to differentiation and proliferation [243, 244]. Research shows that the NOTCH pathway is involved in EMT in the context of development [245]. Additionally, like the WNT pathway, the NOTCH signaling has been implicated in a number of different cancer contexts [246, 247].

Finally, a number of different signaling pathways activated by various growth factors through the tyrosine kinase receptors can also induce EMT. For example, the epidermal growth factor (EGF) can activate the MEK-ERK signaling pathway, which results in the reduction of E-cadherin expression [248]. Furthermore, EGF also activates JAK2 pathways, which leads to EMT through STAT3 activation in several cancer types [249, 250]. Similarly, fibroblast growth factor (FGF) and hepatocyte growth factor (HGF) have been implicated in EMT induction in the context of cancer [11].

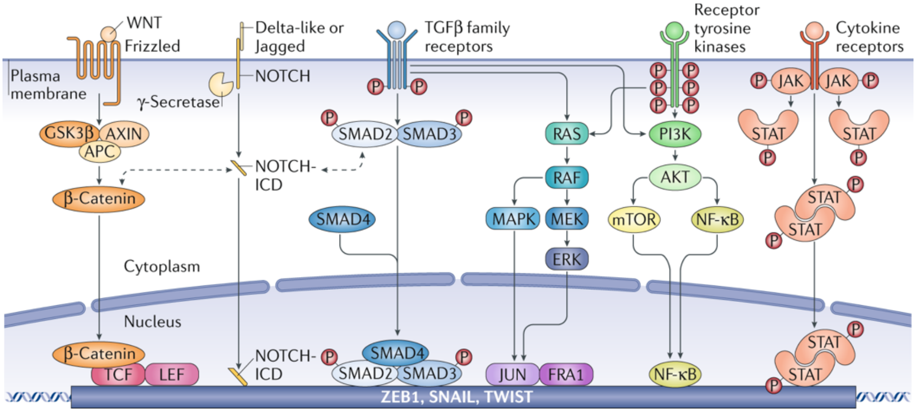


Figure 1.6. *Signaling pathways active in EMT. Reproduced with permission from Nature Reviews Molecular Cell Biology [11].*

The signaling pathways of EMT are diverse. Thus, it is important to note that often there is significant cross-talk and cooperation between these pathways. For example, during embryogenesis, EMT is regulated through WNT, TGFβ, and FGF signaling [251]. In the context of cancer, TGFβ, and RTK signaling pathway cooperation if often reported [252, 253]. The variety of EMT signaling and the cross-talk between the pathways can, in part, explain the plasticity of EMT. Signaling pathways have been extensively studied both in the context of development as well as cancer. However, it remains hard to link everything into a single framework due to the differences in tissues and biological processes.

1.3.4.EMT in development

EMT is an essential process in development and has been studied in great depth. Interestingly, both EMT and MET are extensively utilized during various stages of development. Four waves of EMT and MET occur throughout development resulting in the final differentiation of cell types and formation of organs [254]. Primary EMT occurs during mammalian implantation, gastrulation of metazoans , and neural crest formation in vertebrates [255]. The process of gastrulation has been extensively studied in several different model organisms. It is clear that in all cases, EMT is vital for gastrulation. Furthermore, the most important elements of the molecular machinery are conserved throughout different species [205, 255]. In particular, SNAIL and TWIST transcription factors take center stage in the process in sea urchin and fly, while for vertebrate embryogenesis, SNAIL and SLUG are most important [255]. A few different signaling pathways tightly control vertebrate gastrulation. Initially, WNT signaling confers competence to the cells, and EMT is then induced through TGFβ and maintained by FGF signaling. [251]. After gastrulation in vertebrates, the epidermal and neural regions are separated. Neural crest structure forms at the boundary of the region and cells in this structure undergo EMT. This allows the individual cells to migrate and, in turn, gives rise to different tissue types: craniofacial structures, most of the peripheral nervous system, some endocrine cells, and melanocytes [255]. The signaling pathways during neural crest EMT are analogous to the ones observed during gastrulation. The SNAIL TF is vital for all metazoans during gastrulation. However, in the context of EMT in the neural crest, SNAIL and SLUG are dispensable. Mice embryos can survive the deletion of these TFs during the neural crest EMT, albeit with some resulting defects [256]. Such results underscore the high spatiotemporal complexity of EMT regulation and the high degree of cooperation between different factors that drive EMT.

Primary EMT is followed by MET, which enables cells to differentiate into transient epithelial structures. These secondary epithelia then undergo the next round of EMT to generate mesenchymal cells with a more restricted differentiation potential [255]. A well-studied system that reveals differences between different rounds of EMT-MET process is heart development in vertebrates [194]. During gastrulation, two cardiogenic mesodermal layers form. The second cycle of EMT-MET creates the endothelial cell lining of the heart, and the third round forms the endocardial cushion and cardiac valves. Finally, the fourth wave of EMT-MET gives rise to various distinct cell populations in the heart [254]. Similarly, as discussed above, the second round of EMT is induced by TGFβ signaling and is not well studied [254]. The third round of EMT depends on the combination of TGFβ, Notch, and Erbb3 signaling. It is important to note that different TGFβ ligands are activating the pathways in the first and third rounds of EMT, resulting in different gene expression programs [257]. Finally, the fourth round of EMT relies on yet again, different pathways to control the EMT [258]. Neurofibromatosis type 1 (NF1) [259] and Wilms tumor gene 1 (Wt1) [260] genes are essential during this round of EMT. Interestingly, while the signaling is different, it still converges on the same EMT-TFs. In the case of Wt1, it can directly activate SNAIL and promotes EMT [261]. On the other hand, Wt1 has been shown to also activate the WNT pathway in the context of heart EMT [262].

Heart development is a perfect example of the complexity of the EMT process. It emphasizes how differences in signaling pathways result in different developmental outcomes. However, it is widely accepted that all pathways converge on the activation of different combinations of EMT-TFs [194]. While EMT in development has been studied extensively, further studies are needed to unravel the full picture of EMT mechanisms. Unsurprisingly, EMT studies in development have inspired cancer researchers to identify similar mechanisms in the development and progression of tumors. To some extent, pathological EMT can be considered as reactivation of developmental programs in the adult [255]. On the other and, it is clear that developmental EMT mechanisms are expanded on in the context of cancer.

1.3.5.EMT in tissue regeneration and fibrosis

Under normal circumstances in healthy adult tissues, the epithelial phenotype is protected by various mechanisms [194]. Such control is essential for maintaining tissue integrity and function. Some of the best-studied examples involve control by gene regulation. For example, transcription factor OVOL2 has been shown to repress mesenchymal genes and activate epithelial genes in two separate tissues [263, 264]. Similarly, the Elf5 transcription factor can inhibit EMT by repressing SLUG - a key TF in the EMT process [265]. Interestingly, p53 has been implicated in maintaining epithelial homeostasis. The loss of p53 leads to a decrease in miR-200c, which in turn increases ZEB1 protein [266]. Specific splicing [267] and epigenetic modification [268] mechanisms are also involved in maintaining the epithelial phenotype. Such evidence shows that epithelial phenotype is protected on many different levels. On the other hand, EMT can be induced by a variety of signaling molecules through many different pathways [11, 199]. Cell response ultimately depends on the balance of EMT promoting and suppressing mechanisms, and in healthy tissues, the epithelial phenotype will be generally preserved [194].

During wound healing, epithelial cells at the edge of the wound need to move into a damaged area and rebuild healthy tissue. This process is known as re-epithelialization, and it involves epithelial cells undergoing EMT to become migratory and increase their plasticity [269, 270]. Generally, EMT is thought to be induced by inflammation signaling associated with the wound, and once it ceases, so does EMT [9, 271]. It is important to note that in this scenario, cells undergo only partial EMT, which relies on the activation of the SLUG transcription factor under the tight control of the EGF receptor signaling pathway [272]. This notion is supported by the fact that the deletion of SLUG in mice impairs wound healing [273]. Similarly, lung wound repair has also been associated with the EMT process [274]. Basal cells in the airways undergo partial EMT that is characterized by a loss of cell-cell junctions, expression of vimentin, and a migratory phenotype. Wound healing is a beneficial process and demonstrates that EMT can be tightly controlled to achieve partial process activation. However, little is yet known about the regulatory circuits prevent cells from undergoing extensive EMT [270].

Defective wound healing is associated with sustained inflammation and leads to increased scarring, which is the cause of fibrosis [275]. Fibrosis is the hallmark of many chronic diseases and is on the increase globally [276]. EMT is thought to play a significant role in the development of fibrosis and has been proposed as a target for therapeutic strategies [194]. EMT is best studied in the context of Renal Fibrosis. Early studies showed that mice lacking the Smad3 gene were protected against fibrosis [277], while activation of SNAIL TF leads to fibrosis [278]. Both Smad3 and SNAIL are part of EMT regulatory circuits, which proves a direct link to fibrosis. However, it was also shown that renal epithelial cells do not undergo full EMT and never gain motility [279, 280]. Today it is thought that epithelial cells undergo only partial EMT and relay signals to promote fibrosis in a paracrine manner [194]. Cells in the state of partial EMT have been shown to secrete exosomes and cytokines, which promote activation of fibroblasts and recruit bone-marrow-derived mesenchymal cells that differentiate into myofibroblast [281]. The signaling also recruits macrophages that sustain inflammation in the tissue [282]. While less extensively studied, similar models can be proposed for liver [283] and lung [284] fibrosis processes. From a molecular viewpoint, the described fibrosis processes heavily rely on TGFβ signaling [194, 281]. Therefore, it is a clear target for fighting fibrosis. Therapeutic agents that hinder TGFβ signaling have been identified and show positive results in limiting fibrosis [285, 286]. Efforts to be more precise by targeting downstream effectors of EMT in fibrosis are also underway [287, 288].

1.3.6.EMT in cancer

EMT is considered to be one of the hallmarks of cancer [289] and is important for the survival and dissemination of cancer cells [11, 194, 202]. It is now widely accepted that EMT plays a role in the development and progression of most carcinomas (cancers originating from the epithelial cells) (Table 1.4) [11]. However, EMT activation is not homogeneous through the tumor and exists as a gradient [290]. The bulk of the tumor remains epithelial with little to no EMT activation. In contrast, the leading edge of the tumor exhibits substantial EMT activation and can be characterized by a mesenchymal phenotype. This gradient is more or less steep, depending on the particular genetic profile of the mutations [194]. Historically it has been challenging to link EMT signature to clinical prognosis owing to the broad diversity in tissue types and heterogeneity of tumor genetic composition. A system for quantifying EMT states has been proposed, and results appear to link more active EMT signatures to a worse prognosis for the patient [291]. EMT is thought to be induced by the tumor microenvironment. Research shows that both stromal and immune cells can induce EMT in cancer cells in a paracrine manner. For example, cancer-associated fibroblasts secrete an array of signaling molecules TGFβ, IL-6, EGF, VEGF, and HGF that all play a role in EMT signaling pathways [292, 293]. Similarly, tumor-associated macrophages (TAMs) promote inflammation and secrete TGFβ, among other signaling molecules that contribute to EMT activation [294]. Importantly, extensive EMT activation might be a local rather than global tumor event due to the diverse nature of the signaling in the microenvironment [11, 194].

EMT activation promotes tumor survival by conferring resistance to both chemotherapy and immunotherapy. EMT helps cells escape death in cancer and embryogenesis [255, 295]. Studies have established clear links between resistance to chemotherapy and EMT [203, 296]. For example, miR-200c can restore breast carcinoma susceptibility to chemotherapy by downregulating ZEB1 and ZEB2 expression [297]. In general, EMT-TFs promote chemoresistance by regulating genes involved in cell death and stem cell maintenance. EMT can give rise to cancer stem cell (CSCs) population. These cells form a minority subpopulation and have elevated tumor-initiating potential [298, 299]. They can self-renew and differentiate into non-CSC tumor cells in this way, helping the tumor to expand. The CSCs are thought to arise by cells undergoing partial EMT, although precise molecular circuits are not yet known [203, 300]. Furthermore, in some cases where tumor cells are destroyed by therapy, CSCs may survive, leading to relapse [203]. Cancer cells undergoing EMT are also less susceptible to immunotherapy [11, 194]. While no clear picture yet exists, it seems that EMT contributes to immunosuppression and immune evasion in several different ways. Firstly, cells undergoing EMT secrete large amounts of TGFβ, which is known to have immunosuppressive effects [301]. Similarly, cells during EMT can secrete other cytokines and chemokines that regulate immune cell activities. For example, melanoma cells under activation of SNAIL secrete CCL2 and LCN2 chemokines, which then activate dendritic cells to express PD-L1, which leads to attenuation of cytotoxic T cells [302, 303]. Furthermore, EMT leads to a reduction in the surface display of MHC class I molecules on cancer cells, which allows then to evade cytotoxic T cells [304, 305]. Finally, EMT can induce the expression of PD-L1 in cancer cells leading to immunosuppression [306]. Most of these observations have been made to a particular carcinoma type and may not be directly transferable to other cancer contexts. However, enough evidence exists to show that the EMT process is an essential player in tumor resistance to different therapies.

EMT has long been implicated in the process of metastasis formation in carcinomas [11, 202]. The cancer spread throughout the body is a multistep process that is termed 'invasion-metastasis cascade' [307]. It begins by tumor cells becoming motile and invasive enough to colonize the tumor surrounding tissue. Next comes the intravasation of tumor cells into blood vessels. At this point, circulating tumor cells (CTCs) are generated and can spread throughout the body by finally escaping into healthy tissue and colonizing it. Overall, the process is remarkably inefficient, and only a tiny number of carcinoma cells leaving the original tumor will be able to form macroscopic metastasis [307].

However, the study of this process is critical, as about 90% of cancer-related deaths occur due to metastasis rather than primary tumors [307]. It has been shown that EMT-TFs are directly related to the process of metastasis formation. For example, ZEB1 is required for efficient invasion and metastasis in a mouse model of pancreatic cancer [308]. Similarly, SNAIL is essential for the dissemination of mouse carcinoma cells [309], while SLUG significantly increases the metastatic potential of previously non-metastatic cells [310]. Given that the EMT process, in general, increases the motility, invasiveness, and stemness of cells, it is easy to see similar properties being important in the 'invasion-metastasis cascade'. However, the EMT program fails to explain the last step - distant tissue colonization. This step is of critical importance to metastasis formation, as only very few metastatic cells can survive in the new niche [311]. While little is yet known about the establishment of the metastatic colony, the plasticity and reversibility of the EMT process can offer some insights [11, 194]. Interestingly, CTCs often travel in clusters rather than alone, and the composition of the cluster is not homogenous with cells existing along the EMT spectrum [312]. As metastatic cells enter a new niche, the extracellular signaling drastically changes, and this disrupts the EMT, potentially inducing MET and allowing cells to colonize the new environment [313-315]. Thus parallels can be drawn between rounds of EMT-MET in development and the establishment of metastasis.

Given the importance of EMT in cancer biology, it is not surprising that it is an attractive target for cancer therapy [194]. A few different approaches have been proposed, and efforts are underway. TGFβ inhibitors are the most intensively investigated anti-EMT compounds and are showing promising results [194, 316]. On the other hand, targeting EMT may be more effective in combination with established therapies [317]. However, due to the heterogeneity of the EMT process, it is hard to pinpoint the best approach. Reversing EMT may not be the best strategy as that can lead to the enhancement of metastasis formation [307]. Targeting cells that underwent EMT may be a better option. However, there is a lack of known markers to specially target these cells [11]. It can be confidently said, that even though EMT has been extensively studied in the context of cancer, many unanswered questions remain [11]. Thus research effort continues and, with advancement of novel single-cell analysis tools recently taking center stage, new insights will undoubtedly follow [318, 319].

Table 1.4. *EMT links to cancer. Adapted with permission from Nature Reviews Molecular Cell Biology [11].*

Tumor type	Observations that link EMT to cancer
Breast	SNAIL expression is observed in invasive ductal carcinomas and correlates with lymph node metastasis TWIST1 promotes metastasis of mouse mammary carcinomas HER2-induced mammary tumors spontaneously express SNAIL and express features of EMT SNAIL expression is observed during carcinoma progression in an autochthonous model of breast cancer
Pancreatic	A switch from E-cadherin to N-cadherin shows significant associations with prostate cancer progression in patients Invasive carcinoma cells exhibit features of EMT in an autochthonous mouse model of pancreatic cancer ZEB1 strongly impacts tumor progression, invasion and metastasis in a mouse model of pancreatic cancer
Lung	The expression of EMT markers is tightly associated with disease progression in SCLCs EMT markers are expressed at the peripheral leading edge of NSCLCs, and marker presence is correlated with tumor progression
Colorectal	SLUG expression is correlated with tumor progression and is a marker for poor prognosis in patients ZEB2 is expressed at the invasive front, which correlates with tumor progression and is a prognostic marker for colorectal cancer N-cadherin drives malignant progression of colorectal cancer
Hepatocellular	Overexpression of TWIST induces EMT and promotes invasion and metastasis of hepatocellular carcinomas SNAIL induces EMT and promotes metastasis and tumor-initiating properties in hepatocellular carcinomas
Bladder	EMT markers are associated with tumors of high grade and stage SNAIL-induced EMT promotes metastasis in a xenograft model of bladder cancer E-cadherin is negatively correlated with, and SOX2 and NANOG are positively correlated with, tumor grade and stage in patients with invasive bladder carcinoma

1.4. Breast cancer

1.4.1. Overview

Breast cancer is the most frequent cancer in women [12]. An estimated 2.1 million women were diagnosed with it in 2018, with over 600,000 patients succumbing to the disease [13]. The prevalence of breast cancer is on the rise, with a 3.1% yearly increase in cases globally [14]. The increase can be partially attributed to widespread testing, as breast cancer surveillance has become a standard procedure over the years. This also, in part, explains why high-income regions report more cases than the low-income regions: 92 cases per 100,000 women in North America as compared to 27 cases per 100,000 women in middle Africa and eastern Asia [320]. Therefore, the actual number of patients globally is likely even higher. On the other hand, the increase in breast cancer incidence and differences between regions can also be, in part, traced back to the later age of first pregnancies in developed countries [321]. Furthermore, about 20% of the cases can be attributed to modifiable risk factors – obesity, alcohol use, and low physical activity [322]. Finally, hormonal contraceptives can also increase the risk of breast cancer occurrence [323].

On a molecular level, breast cancer is a highly heterogeneous disease and can be categorized by several different classification systems. The standard classification in current clinical practice relies on histological and molecular characterization [12]. Important molecular characteristics are the expression of estrogen receptor (ER), progesterone receptor (PR), and enrichment of human epidermal growth factor receptor 2 (HER2) [324]. Based on these markers, each patient case is assigned to one of five distinct types (Table 1.5). Furthermore, these markers are crucial for guiding therapy decision-making. From a histological viewpoint, breast cancer can be classified into 19 subtypes, according to WHO classification [325]. Based on which, the majority of tumors – 70-75% are ductal carcinomas (also referred to as carcinoma of no special type), while 10-15% of tumors are lobular carcinomas [12]. The remaining 17 types are rare and, in some cases, are associated with a very good or, on the contrary, a poor prognosis for the patient.

Breast cancer is a well-studied disease, and systematic treatment guidelines are established. As a result, 70-80% of patients with the early-stage, non-metastatic disease are cured [12]. By contrast, patients with the advanced (metastatic) disease are considered incurable, and their median survival is 2-3 years, with the spread of metastases being the dominant cause of death [15].

Furthermore, it is necessary to acknowledge that there are substantial differences in the quality of care globally, and typically in low-income countries, cancer patients have a worse prognosis and lower survival rates. Much of the focus recently has been devoted to studying advanced breast cancer cases, and a wide array of new drugs and therapies is under evaluation in clinical trials. Together with emerging better diagnostic tools, this gives hope to implement the precision medicine approach in breast cancer care in the not too distant future [12].

Table 1.5. *Types of breast cancer and defining molecular features.*

Breast cancer type	Molecular description	Prognosis
Luminal A-like	Strongly ER+ and/or PR+; HER2–	Good
Luminal B-like HER2-	ER+ and/or PR+, but lower expression than in luminal A-like; HER2–	Intermediate
Luminal B-like HER2+	ER+ and/or PR+, but lower expression than in luminal A-like; HER2+	Intermediate
HER2+ non-luminal	ER–, PR–, HER2+; non-luminal	Intermediate
Triple-negative (TNBC)	ER–, PR–, HER2–; non-luminal	Poor

1.4.2. Development of breast cancer

There are two different models for cancer establishment. The first is called the clonal evolution model. It proposes that mutations accumulate, and epigenetic changes occur with each successive cell division, and the 'fittest' cells survive to establish the tumor [326]. The second model suggests that only precursor cancer stem cells can initiate and sustain tumor progression [327]. In reality, both models play some part in breast cancer development as cancer stem cells can evolve in a clonal fashion [328]. Upon establishment, breast tumors can further develop in two divergent pathways. The first pathway is termed low grade-like pathway and applies primarily to luminal A type tumors. It is characterized by the loss of 16q and gain of 1q chromosome, and a gene expression signature is associated with the ER phenotype. The second pathway is termed high grade-like pathway and is characterized by the loss of 13q chromosome and amplification of 17q12 segment, which leads to amplification of HER2 expression. This type of progression is associated with lower genetic stability and increased tumor aggressiveness [328, 329]. All HER2+ and TNBC breast tumors develop according to the second pathway. Such divergence in development pathways further highlights the heterogeneity of breast cancer.

The exact mechanism by which breast cancer arises is unknown. However, hormone exposure is a major risk factor. The imbalance between estrogen and progesterone during menstrual cycles promotes cell proliferation [12]. Due to the repeated nature of the process, this can lead to DNA damage accumulation, which in turn results in cells becoming malignant. Extensive studies have revealed that most frequently mutated and/or amplified genes are TP53 (41% of tumors), PIK3CA (30%), MYC (20%), PTEN (16%), CCND1 (16%), ERBB2 (13%), FGFR1 (11%) and GATA3 (10%) [330]. All of these genes in one or another way are related to cell-cycle control and cell proliferation. For example, the ERBB2 gene encodes the HER2 protein. It is a receptor of human epidermal growth factor family and upon activation is involved in promoting cell proliferation, survival, metastasis, and adhesion. Different types of breast cancer have different genetic profiles, which can be determined with various diagnostic tests and have implications for treatment strategies. Luminal cancer types (ER+ and/or PR+) are frequently associated with PI3K-AKT pathway activation and inactivation of GATA3 and JUN kinase pathways. HER2+ tumors are associated with mutations in the ERBB2 gene, while TNBC frequently has TP53 mutations and extensive copy number variation [331]. Furthermore, to date, over 100 high-probability breast cancer driver genes have been identified [332]. However, the majority of the mutations affecting the driver genes are rare, and most cancers are caused by the accumulation of many low-penetrance mutations [12].

Family genetics is also a significant factor in breast cancer. About 10% of the cases are linked to family history. Out of those cases, about 30% can be explained due to mutations in particular genes - BRCA1, BRCA2, PTEN, TP53, CDH1, and STK11 [333]. Perhaps the best-studied is the case of BRCA1 and BRCA2. Protein products of these genes are involved in DNA damage repair and are considered tumor suppressor genes [334]. Mutations in BRCA genes are associated with an increased mutational load and a significant cumulative risk of developing breast cancer by the age of 80 - 69-72% [335]. Moreover, the risk of ovarian cancer is also significantly increased in the case of mutated BRCA genes. Over 2000 different mutations have been identified for these genes. Remarkably, only a few of them have been found repeatedly in unrelated families [12, 336]. Finally, it must be noted that cancer risk differs depending on the exact mutation profile of BRCA genes [335].

Primary breast tumors are well studied, while metastasis formation and genetic evolution are less understood [12, 331]. Metastases develop through the dissemination of CTCs, which are likely generated through partial EMT, as discussed in the previous chapter [337]. In breast cancer patients, metastases typically develop in bones, lungs, or liver [15]. Research shows that metastases disseminate late from the primary tumor, and up to 80% of driver mutations are preserved during dissemination [331]. However, new tumors often harbor 'private' mutations, resulting in subclonal diversification [338]. For example, 24% of tumors lose ER expression in the metastatic site, while 14% gain ER expression. Furthermore, some mutations may arise as a response to the treatment of the primary tumor [12]. Different metastatic sites in the same patient may also have divergent evolutionary pathways [331]. Such complexity of advanced breast cancer makes it challenging to treat it successfully.

1.4.3.Immune infiltration of breast cancer

The immune system plays an essential role in breast cancer development and progression (Figure 1.7). In a temporal sense, the interaction of the immune system and the tumor can be described by the "three Es" model: elimination, equilibrium, and escape [339]. Early in tumor development, the inflammatory environment and associated cytokines will promote tumor-suppression. Inflammation signaling will activate the innate immune system, and early cancer cells can be eliminated by NK cells [340, 341]. Furthermore, recent studies show that innate lymphoid cells may also be implicated in the early stages of breast cancer development, although their involvement is not yet fully understood [342]. If the early tumor control fails, tumors will progress into the stage of equilibrium with the immune system. During this stage, tumor growth will be slowed down, but ultimately resistance mechanisms will develop, and tumor cells will escape immunosurveillance [339]. At this stage the inflammation becomes chronic and it actively promotes tumor expansion [343, 344]. Inflammation signaling shapes the breast tumor microenvironment (TME) and also affects tumor cells directly by promoting their survival and proliferation. Furthermore, inflammation signaling leads to the downregulation of MHC class I and upregulation PD-L1 expression in tumor cells, which directly helps to escape killing by NK and cytotoxic T cells [340, 345]

Myeloid cells are critical players in the breast tumor microenvironment, and their population has been shown to increase in tumor tissue as compared to healthy breast tissue [346]. Myeloid cells are typically recruited and activated by tumor cell signaling and primarily contribute to tumor survival and immunosuppression [341, 347]. In the context of breast cancer, myeloid-derived suppressor cells (MDSCs) play an essential role in establishing the immunosuppressive environment [348, 349]. These cells are recruited by tumor cell signaling and will deplete nutrients needed by lymphocytes, generate oxidative stress, and interfere with lymphocyte trafficking [341]. The overall effect inhibits the functions of T cells, NK cells, and dendritic cells while promoting Th2 T cells, regulatory T cells (Tregs), and tumor-associated macrophages (TAMs) [348, 350]. Furthermore, research suggests that the level of MDSCs in peripheral blood directly correlates with disease burden and duration [351]. Another relevant category of myeloid cells in breast cancer is tumor-associated macrophages (TAMs) [352]. They can have different roles in the TME, depending on their polarization. M1-like TAMs are stimulated by Th1 cell cytokines and have antitumor properties. On the opposite end of the polarization spectrum, M2-like TAMs are activated by Th2 cell cytokines and display pro-tumor characteristics [353]. Due to the immunosuppressive and inflammatory environment, M2-like TAMs are the dominant type in breast tumors [352]. They are involved in mediating tumor growth and progression and can contribute to therapy resistance mechanisms [354]. M2-like TAMs secrete cytokines, which then act to recruit Tregs and suppress antigen-presenting cells (APCs) and CTLs. While not extensively studied in the context of breast cancer, high levels of TAMs in the tumor are associated with poor patient survival indicating their importance in breast tumor biology [352].

The adaptive immune system has a dual role in breast tumors. On the one hand, CD8+ T lymphocytes give rise to CTLs, which are the main effectors acting against tumor cells [355]. CTLs recognize specific antigens presented by tumor cells and can successfully destroy them. Harnessing this interaction has been the focus of immunotherapy and has been utilized in the treatment of various cancers [356]. In the context of breast cancer, tumor infiltration with CTLs is associated with a favorable prognosis [357]. On the other hand, the roles of CD4+ T cells in the TME are more diverse. Naïve CD4+ T cells can differentiate into a few different effector subtypes: mainly different T helper cells (Th1, Th2, and Th17) and T regulatory cells [358]. Th1 cells directly activate CTLs by secreting anti-tumor cytokines, thus playing an essential role in the anti-tumor response [345]. However, other CD4+ T cell subtypes have tumor-promoting properties. Th2 cells secrete a wide array of cytokines that contribute to the immunosuppressive environment in TME and, among other roles, have a direct effect on CTL suppression and activation of M2-like TAMs [359]. Treg cells can directly suppress CTL and Th1 cell functions and play a central role in immune suppression in breast cancer tumors [18, 360]. Accordingly, high breast tumor infiltration with Treg cells has been shown to correlate with a poor prognosis [357, 361]. Finally, the B cells also need to be recognized in the context of breast cancer. Much like CD4+ T cells, B cells have been shown to have both anti-tumor as well as tumor-promoting effects [362]. For example, B cells can produce antibodies against tumor cell antigens. This can induce CTL [363] as well as NK cell response [362] and contribute to tumor suppression. On the other hand, B cells can secrete cytokines that promote the immunosuppressive environment in the TME [364]. Furthermore, B cells can also directly activate Tregs [365]. Research shows that B cells have different effector phenotypes, and their prognostic value remains controversial [362].Overall, when considering tumor and immune system interaction, it is important to recognize that no cell type acts in an isolated manner. Instead, the whole TME is a single interconnected network where each cell type contributes to and is regulated by the diverse signaling networks.

Historically breast cancer tumors have been considered immunologically quiescent or ‘cold’. However, recent evidence shows that this is not true as a significant amount of immune cells is detected in most breast cancer tumors [16, 17]. Thus in the current view, most breast tumors are considered immunosuppressed rather than immunologically quiescent. Furthermore, it must be noted that in line with the heterogeneity of breast cancer, the immune cell subset will also heavily depend on the particular tumor type [18]. In particular, ER+ tumors display a broad diversity in TIL profiles. This may be partly explained by the effect of hormone signaling pathways on immune cells and their interactions [18]. Furthermore, differences between tumor types can, in part, be attributed to differences in genetic mutation profiles [366]. However, it must be noted that there is no consistent proof that the mutational load alone directly correlates to TIL levels in tumors [367].

Studying the role of the immune system in breast tumors is getting increasing attention in recent years as it is becoming clear that it is a potent therapeutic target [17, 341, 368]. Furthermore, the profile of tumor-infiltrating leukocytes (TILs) can potentially be a prognostic marker. While no definitive guidelines exist, literature meta-analysis has shown that for TNBC, overall high TIL infiltration correlates with a favorable prognosis. To a lesser extent, the same principles apply to HER2+ tumors. On the other hand, for tumors expressing hormone receptors (ER+ and/or PR+), immune infiltration does not correlate with patient prognosis [357]. It is important to note that the precise composition of TILs is also relevant for prognosis, as discussed above. However, most of the observations have been made independently of one another. Thus, it is likely that more prognostic power can come from analyzing specific TIL infiltration profiles and uniting them with existing biomarkers. Overall, while a lot is now known about separate immune cell types and signaling pathways operating in the breast TME, the full picture is far from clear, making it hard to draw general conclusions. Understanding the intricate interplay between different cell types in different breast tumor types is the key to the successful application of immunotherapy, which has so far seen only limited success in breast cancer [12, 368].

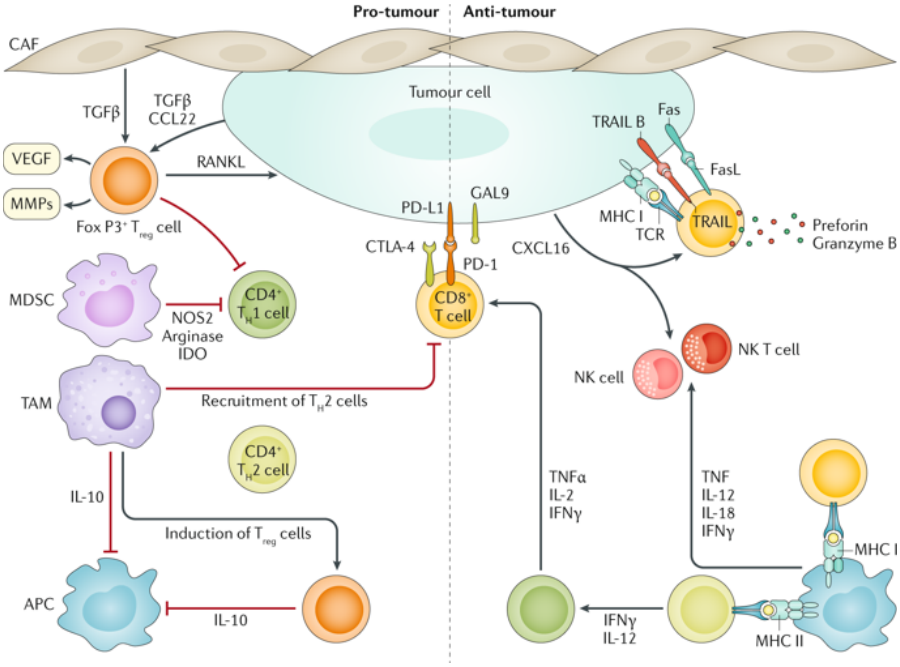


Figure 1.7. *Immune cell interactions in breast cancer. Reproduced with permission from Nature Reviews Disease Primers [12].*

1.4.4. Breast cancer treatment strategies

The first line of defense against breast cancer is population screening. This strategy aims to identify the disease at an early stage for which there is effective treatment available [12]. Mammography is the most widely used screening technique with exact guidelines differing between countries. Population screening reduces relative mortality risk by 20% and is most beneficial for women in the age group of 50-69 years old [369]. Diagnosis of breast cancer is based on three different tests – clinical examination, imaging, and a needle biopsy. According to internationally recognized guidelines, the determination of the ER, PR, and HER2 status is mandatory for all patients with invasive breast cancer [370]. Based on the presence of metastases, breast cancer is classified as either early (without metastasis) or advanced (if metastases are detected). Treatment strategies and outcomes differ significantly between the two cases [12].

For most early breast cancer patients, the tumor will be surgically removed if the procedure is possible [12]. However, most patients will also need some form of systemic therapy. In some cases, it can be administered before the surgery (neoadjuvant) if the tumor burden is high and needs to be reduced before the operation. Furthermore, neoadjuvant chemotherapy is also applied to patients that have an aggressive form of HER2+ or TBNC, and response to it serves as a prognostic marker [371]. After surgery, based on the risk of recurrence (as determined by biomarkers or response to neoadjuvant therapy), additional rounds of adjuvant therapy may be applied. In all cases of ER+ breast cancers, endocrine therapy is applied to block the ER activity. Additional diagnostic tests can be performed to aid in decision-making for adjuvant systemic therapy. Postoperative radiation therapy may also be applied according to guidelines and has been proven to benefit patient survival [372]. Overall, modern-day systemic therapies are highly effective, and when applied in an adjuvant manner, reduce mortality by one-third [373].

Advanced breast cancer is an incurable disease [12]. Several different treatment strategies exist that all aim to relieve patient symptoms and to prolong quality-adjusted life expectancy. Radiation therapy is often prescribed and has been proven to provide substantial benefit in the case of bone, brain, and soft tissue metastases [374]. Furthermore, radiation therapy might also induce a systemic immune response and has the potential to increase the efficacy of immunotherapy, as discussed below [375]. A large number of systemic therapies are available, and a lot of novel drugs are under investigation [12]. While general treatment guidelines exist, most cases are approached in an individualized manner, and treatment is adjusted based on the progression of the disease. Furthermore, due to the substantial heterogeneity and complexity of advanced breast cancer, the optimal treatment sequences are unknown in most cases and, in particular, in the case of TNBC.

Immunotherapy has shown great promise for cancer treatment over the last decade. However, in the case of breast cancer, clinical developments have been slow as compared to other tumor types [12]. The first treatment - immune checkpoint blockade (ICB) for metastatic TNBC was only approved in 2019 [368]. Historically breast cancers have been considered 'cold'. While this viewpoint is changing, it is also becoming apparent that different breast cancer types have different immune infiltration rates and profiles [376]. Furthermore, the breast tumor environment is often highly immunosuppressive [341]. A successful immune response to tumor cells consists of several steps. First, tumors cells need to release antigens upon their death, which must then be presented by antigen-presenting cells to prime and activate effector cells. Activated T cells must then infiltrate the tumor, recognize, and kill tumor cells [377]. For immunotherapy to be efficient, all these steps must function adequately, which does not seem to be the case in most breast cancer tumors. For example, the survival of the patients in the ICB clinical trials for advanced TNBC treatment was not longer than in chemotherapy control groups [368]. This means that only a fraction of patients benefited. Such results suggest that immune surveillance evasion by tumor cells is not the only problem in the immune response cascade. Similar observations have prompted an increased interest in combination approaches where immunotherapy is administered together with chemotherapy or radiotherapy [17, 368]. Conventional therapies can potentially prime the tumor by activating the initial stage of the immune response cascade and lead to more effective immunotherapy. Some preliminary results are available that support this approach [368]. As of April 2020, there are at least 230 active clinical trials involving immunotherapy in breast cancer treatment (clinicaltrials.gov database). This highlights increased interest in immunotherapy approaches for breast cancer treatment. Furthermore, it gives hope to see improvements in breast cancer care in the near future.

2 MATERIALS AND METHODS

2.1. Single-cell analysis

2.1.1. Single-cell transcriptome barcoding

Transcriptomes of single cells were barcoded using droplet microfluidic platform [35]. This procedure involves three distinct parts: (i) microfluidic chip manufacturing, (ii) barcoding hydrogel bead (BHB) synthesis, and (iii) single-cell transcriptome barcoding in microfluidic droplets. Microfluidic chip and BHB preparations were carried out in advance of single-cell transcriptome barcoding. Both the microfluidic chips and BHBs were prepared in larger batches and used in multiple experiments to minimize technical errors.

The microfluidic chips were fabricated using a previously described procedure [378]. Briefly, this multistep process involves using computer-assisted design software to draw the microfluidic device design, which is then printed on a transparent film to produce a mask. Such a mask is then used to fabricate a silicon master using soft photolithography. The finished master serves as a reusable mold to transfer the microfluidic chip pattern onto poly(dimethylsiloxane) (PDMS) slabs, which are then bonded to a glass slide to create the microfluidic chip. Microfluidic channels are treated with a hydrophobic coating to complete the microfluidic chip manufacturing. In this work, two different chip designs were used (Figure 2.1).

Barcoding hydrogel beads were synthesized using a microfluidic chip (Figure 2.1, panel A). Monodisperse droplets were generated using acrylamide:bis-acrylamide solution supplemented with an acrydite-modified DNA primer (Table 2.1, HBH-Stub) to a final concentration of 50µM. The acrydite-modified DNA primer also contained a photo-liable linker that can be cleaved by a >350nm UV light. The photo-liable linker enables the DNA primer release from the hydrogel beads. All barcoding hydrogel beads synthesis steps were carried out in a red light environment to minimize the photo-liable linker's cleaving. Using a microfluidic chip (Figure 2.1, panel A, channel height 50 µm) monodisperse emulsion of 60 µm diameter droplets were generated. The flow rates used to operate the microfluidic chip are indicated in Table 2.2. Commercially available microfluidic oil was used for all experiments (RAN Biotechnologies, cat. no. 008-FluoroSurfactant-2wtH-50G). The emulsion was then transferred to 65°C for 4 hours. The acrylamide, bis-acrylamide solution polymerizes, and acrydite-modified DNA primers are covalently incorporated into the hydrogel mesh during the incubation. After polymerization, the hydrogel beads were washed using hexane (Sigma-Aldrich, cat. no. 208752) and buffer solution (Table 2.3, TBSET) to remove microfluidic oil and surfactants. Next, the hydrogel beads with attached DNA primers were barcoded in a combinatorial split-and-pool manner to generate a barcoding hydrogel bead library. Two successive barcoding rounds were performed, and each round contained 384-unique oligonucleotide sequences (Table 2.1, barcoding oligo 1 and barcoding oligo 2) resulting in 147 456 unique barcodes. During each barcoding round, hydrogel beads were barcoded by splitting the whole hydrogel pool between four 96-well plates (a total of 384 individual wells). Each well contained a unique primer that had a complementary region to the ssDNA attached to the hydrogels. After hybridization, the primers attached to the hydrogel beads were extended using isothermal primer extension with BST 2.0 polymerase (NEB, cat. no. M0537L). After each barcoding step, all hydrogels were pooled, and the double-stranded DNA (dsDNA) produced during primer extension was converted into a single-stranded form by alkaline denaturation (Table 2.3, Denaturation buffer). After two barcoding rounds, the remaining unextended primers attached to hydrogel beads were removed. This procedure involved protecting the fully extended primers by hybridization with a complementary probe (Table 2.1, Protection oligo) and digesting the barcoded hydrogel bead pool with Exo I exonuclease (Thermo Scientific, cat. no. EN0581). After completing the synthesis, the quality of BHBs was evaluated by fluorescence *in situ* hybridization (FISH) using a set of probes complementary to regions common to all the primers attached to the hydrogel beads (Table 2.1, FAM-PE1, FAM-W1, FAM-BA19). BHBs prepared in such a way were stored at +4°C (Table 2.3, Storage buffer) until used. Before being used in a single cell transcriptome barcoding experiment, the BHBs were washed in a buffer solution that contained RT reaction components (Table 2.3, barcoding buffer). This was done in order not to dilute the RT components in droplets.

Single cell transcriptomes were barcoded using microfluidic chip (Figure 2.1, panel B, channel height 80 µm). In this step, single cells are captured in droplets with RT–lysis reagents and BHBs. Accordingly, the microfluidic chip consists of two junctions: the first for bringing the RT–lysis reagents, cells, and barcoded beads together, and the second for cell and bead co-encapsulation, where droplet generation occurs. Due to laminar flow, the mixing of cells and reagents occurs only after encapsulation, preventing premature cell lysis. The flow rates used to operate the microfluidic chip are detailed in Table 2.2. Commercially available microfluidic oil was used for all experiments (RAN Biotechnologies, cat. no. 008-FluoroSurfactant-2wtH-50G). Using the described flow rates allows to barcode ~30000 cells per hour. The droplet volume can be precisely tuned by adjusting the flow rate of microfluidic oil. In this work, either 1.5nl or 3nl droplets were used to barcode single cell transcriptomes. The droplet size for each barcoding experiment is indicated in the sections below. cells were diluted and injected into the device at a concentration corresponding to one cell in every ~10 droplets to minimize cases in which two or more cells enter the same droplet. Using such dilute cell samples corresponds to Poisson λ=0.1, and under such condition, over 99% of droplets will contain one cell or no cells. Cell sample preparation is detailed in the relevant sections below. Monitoring the cell and bead co-encapsulation with the high-speed camera and adjusting the BHBs flow rate allowed to achieve a highly efficient BHBs loading into droplets. In this work, all single-cell transcriptome barcoding experiments were performed with 75–90% BHB loading efficiency. During encapsulation, the emulsion was collected on ice, and after encapsulation was completed, the emulsion was exposed to 350 nm UV-light for 5 min to release DNA barcoding primers attached to the hydrogel beads. Finally, the emulsion was transferred to a dry heat block, and the RT reaction was performed using conditions indicated in sections below.

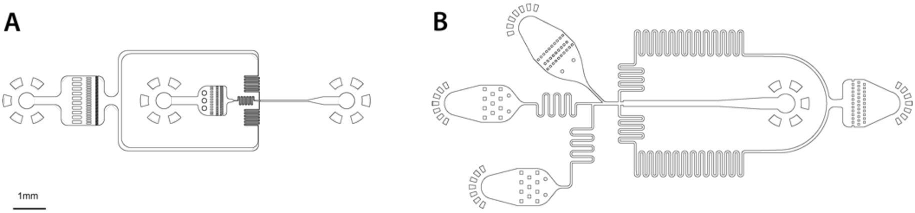


Figure 2.1. Microfluidic chip designs used in this work. Panel A – microfluidic chip used for barcoded hydrogel bead manufacturing. Panel B – microfluidic chip used for single-cell transcriptome barcoding.

Table 2.1. DNA oligo sequences used in single-cell transcriptome barcoding experiments. DNR modifications are highlighted in bold.

Name	Sequence
DNA-Stub	5’-/**acrydite/photocleavable spacer**/CGATGACGTAATACGACTCAC TATAGGGATACCACCATGGCTCTTTCCCTACACGACGCTCTTC-3′
Barcoding oligo 1	5′-GGCGTCACAAGCAATCACTC[Cellindex1]AGATCGGAAGAGC GTCGTGTAGGGAAAG-3′, where Cell-index1 is 384 unique sequences 8-11nt length.
Barcoding oligo 2	5’-BAAAAAAAAAAAAAAAAAAANNNNNNNN[Cellindex2]TTGG CGTCACAAGCAATCACTC-3′, where Cell-index2 is 384 unique sequences 8nt length.
Protection oligo	5′-BAAAAAAAAAAAAAAAAAAA-3′
FAM-PE1	5′-/**6-FAM**/AGATCGGAAGAGCGTCGTGTAGG GAAAGAG-3′
FAM-W1	5′-/**6-FAM**/AAGGCGTCACAAGCAATCACTC-3′
FAM-BA19	5′-/**6-FAM**/BAAAAAAAAAAAAAAAAAAA-3′
2nd RT oligo	5’- GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNN-3’
Final PCR oligo 1	5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACA CGA-3’
Final PCR oligo 2	5′-CAAGCAGAAGACGGCATACGAGAT[index]GTGACTGGAGTT CAGACGTGTGCTCTTCCGATCT-3’, where the index is: CGTGAT, ACATCG,GCCTAA, TGGTCA, CACTGT or ATTGGC.

Table 2.2. Flow rates used to operate microfluidic chips.

Component	Flow rate
Hydrogel bead synthesis	
Acrylamide–primer mix	900μl/hr
Microfluidic Oil	1800μl/hr
Single cell transcriptome barcoding	
Single-cell suspension	250μl/hr
RT–lysis reagents	250μl/hr
BHBs	80±10 μl/hr
Microfluidic Oil	700μl/hr - 900μl/hr

Table 2.3. Compositions of solutions used during barcoded hydrogel bead synthesis.

Name	Composition
TBSET	10mM Tris-HCl (pH 8.0), 137mM NaCl, 2.7mM KCl, 10mM EDTA, 0.1% (vol/vol) Triton X-100. Solution prepared in ddH_2O and filtered through a 0.2-μm membrane.
Denaturation buffer	150mM NaOH, 0.5% (wt/wt) Brij-35. The solution prepared in ddH_2O and filtered through a 0.2-μm membrane.
Storage buffer	10mM Tris-HCl (pH 8.0), 10 mM EDTA, 0.1% (vol/vol) Tween-20. The solution prepared in ddH_2O and filtered through a 0.2-μm membrane.
Barcoding buffer	50 mM Tris-HCl (pH 8.3 at 25°C), 75 mM KCl, 3 mM $MgCl_2$, 1% (vol/vol) Igepal-CA630. The solution prepared in ddH_2O,

2.1.2. Single-cell sequencing library preparation

Sequencing library preparation refers to a number for steps that need to be carried out to amplify the barcoded cDNA from single cells and prepare it for sequencing. This procedure was described in detail as part of this work [35]. First, the emulsion was aliquoted in such a way that each aliquot contained around 3000 cells based on observed actual cell concentration in a particular run. This was done to minimize cases in which two cells acquire the same barcode. Next, the droplets were broken by the addition of perfluoroctanol (Sigma-Aldrich, cat. no. 370533), and cDNA was cleaned to remove hydrogel beads, unextended primers and primer dimers generated during RT reaction (for a detailed protocol see section 2.2.3). After cDNA was purified, it was subjected to second strand synthesis reaction (for a detailed protocol see section 2.2.4) and then amplified by *in vitro* transcription (IVT) using T7 RNA polymerase (for a detailed protocol see section 2.2.5). The amplified antisense RNA was fragmented using zinc-ion-mediated cleavage (Ambion, cat. no AM8740), purified with AMPure reagent (Beckman Coulter, cat. no. A63881) using 1.2x volume ratio, and converted into a DNA library by a second RT reaction (Takara Clontec, cat. no. 2680A) using a random priming sequence (Table 2.1, 2^{nd} RT oligo). Next, the reaction product was purified with AMPure reagent (Beckman Coulter, cat. no. A63881) using a 1x volume ratio. Finally, the library was PCR amplified (Kapa Biosystems, cat. no. KK2601) using primers compatible with Illumina sequencing machines (Table 2.1, Final PCR oligo 1 and Final PCR oligo 2) and purified with AMPure reagent (Beckman Coulter, cat. no. A63881) using 0.8x volume ratio. A library index added during the final PCR allowed to pool multiple libraries in a single sequencing run. The described sequencing library preparation also contained two quality control steps were the fragment size distribution of the library was analyzed: i) after library amplification by IVT (Agilent, cat. no. 5067-1513) and ii) after final PCR amplification (Agilent, cat. no. 50674626). Only those libraries that passed the quality control were subjected to sequencing.

2.1.3. Single-cell library sequencing

Single-cell library sequencing was carried out on Illumina MiSeq, NextSeq 550, and HiSeq2500 sequencing instruments using Illumina sequencing reagent kits appropriate for each instrument. In all cases, sequencing was carried out in paired-end (PE) mode where Read1 contained cell barcode and UMI information, and Read2 contained transcript information. Because of differences between sequencing platforms, the read lengths varied between different experiments. In all cases, Read1 length was ≥51bp, and Read2 length was ≥35bp. Data that was used to investigate the EMT process and to create the breast cancer immune atlas was gathered on a HiSeq2500 machine using PE sequencing mode were Read1 length was 54bp and Read2 length was 66bp. All sequencing runs were set to “FastQ generation" mode, and no further preprocessing was done.

2.1.4. Single-cell data analysis

Single-cell data analysis performed in this work consisted of a few separate stages. First, the raw sequencing data was processed to produce a cell x gene matrix. Next, the matrix was manually thresholded to remove low-quality cells and noisy barcodes. After the matrix was prepared, it was subjected to downstream analysis by various algorithms.

All raw sequencing data in this work was processed using the SEQC pipeline [86]. The workflow is presented in Figure 2.2. Briefly, the pipeline begins by extracting the cell barcode and UMI from the forward read (Read1) and storing these data in the header of the reverse read (Read2). This produces a single FastQ file that contains alignable sequences and all relevant metadata. The merged file is then filtered for cell barcode substitution errors, broken barcodes, and low-complexity sequences (homopolymers) to eliminate errors early in the pipeline. Next, the filtered reads are aligned against the human genome using the STAR aligner [90]. After alignment, minimal representations of sequencing reads are translated into an hdf5 read store object, where cell barcodes are represented in reduced 3-bit coding. Reads are then annotated with a reduced set of exon and gene ids representing gene features — only the ones that are possible to detect with poly-A capture-based droplet RNA sequencing. The pipeline then attempts to resolve reads with multiple equal-scoring alignments. In cases where both genomic and transcriptomic alignments are present, only the transcriptomic alignments are retained. Unique alignments from the previous step are corrected for errors using an enhancement of a previously described method [34] - with an additional probability model to constrain the false positive rate. The error-reduced, uniquely-aligned data are grouped by cell, molecule, and gene annotation, and compressed into a final cell x gene matrix. The pipeline was run using virtual instances provided by Amazon Web Services (c4.4xlarge instance type).

The matrix produced by the pipeline was manually thresholded to remove dead cells (mtRNA >20%), low complexity cells (based on the low number of detected unique genes given the number of total molecules assigned to the cell) and barcodes that captured ambient RNA (barcodes were separated by finding the saddle point in the distribution of total molecule counts per barcode and excluding the mode with lower mean). Data processed in such a way was used to compare the effects of different sequencing library preparation optimizations or was subjected to further analysis using different computational algorithms. The summary of algorithms used for downstream analysis is presented in Table 2.4.

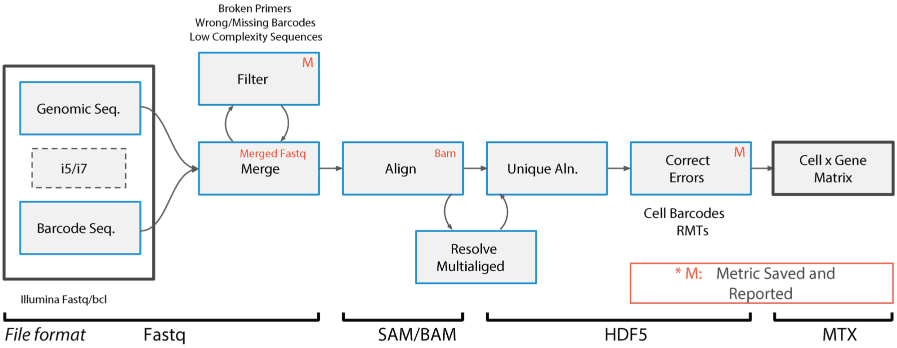


Figure 2.2. SEQC pipeline workflow.

Table 2.4. Summary of computational algorithms used in this work.

Algorithm	Brief description	Reference
	EMT data analysis	
MAGIC	Data imputation using data diffusion	[138]
Principal Convex Hull Analysis	Archetype identification using imputed data	[379]
kNN-DREMI	Gene-gene correlation quantification	[138]
DREVI	Gene set clustering and pseudotime ordering based on kNN-DREMI scores.	[138]
	Breast cancer data analysis	
Biscuit	Data imputation and clustering	[86]
Phenograph	Data clustering	[179]
t-SNE	Dimensionality reduction and data visualization	[161]
Diffusion maps	Dimensionality reduction and diffusion component identification	[167]

2.2. scRNA-Seq protocol optimization

2.2.1. qPCR

Transcript capture efficiency was evaluated using qPCR (Thermo Scientific, cat. no. K0222). Reaction was performed according to manufacturer's recommendations. Eight different genes were measured for each inhibitor (Table 2.5). Further comparison of two RNA-Seq inhibitors (Superase IN and RiboLock) was done in scRNA-Seq assay using K562 cells and comparing the transcript capture rates.

Table 2.5. Sequences of qPCR primers used to compare the performance of RNAse inhibitors. RNA levels specific for K562 cells. RNA levels data from Human Protein Atlas database.

Name of target gene	Forward primer sequence	Reverse primer sequence	RNA levels, FPKM
ActB	CGCCGCCAGCTCACC	TCTCCATGTCGTCCCAGTTG	1137
VIM	CGGGAGAAATTGCAGGAGGA	TCTTGGCAGCCACACTTTCA	508
B2M	CTCACGTCATCCAGCAGAGAA	TGCTTACATGTCTCGATCCCAC	185
TGFb1	TACCTGAACCCGTGTTGCTC	CCGGTAGTGAACCCGTTGAT	58
STAT3	GGAGAAACAGGATGGCCCAA	ACCTGCTCTGAAGAAACTGCT	52
AKT1	AAGTCATCGTGGCCAAGGAC	GTTCTCCAGCTTGAGGTCCC	42
SMAD2	GTTCCTTTCCTCCTCCGCTC	CTTGTATCGAACCTCCCGGC	30
EGFR	CGAATGGGCCTAAGATCCCG	CCCTTATACACCGTGCCGAA	0

2.2.2. Comparison of RT enzymes

In this work performance of three different RT enzymes for single cell transcriptome barcoding was evaluated: SuperScript III (Invitrogen, cat. no. 18080044), SuperScript IV (Invitrogen, cat. no. 8090010) and Maxima H minus (Thermo Scientific, cat. no. EP0751). RT reaction compositions and conditions are detailed in Table 2.6. Efficiency of RT enzymes was evaluated by performing scRNA-Seq assay using K562 cells and comparing the transcript capture rates.

Table 2.6. Summary of different RT reaction conditions used to barcode single-cell transcriptomes.

RT enzyme	Reaction composition in droplets	Thermal protocol
SuperScript III (Invitrogen, cat. no. 18080044)	37.6 mM KCl, 45 mM NaCl, 5.8 mM $MgCl_2$, 54 mM Tris-HCl [pH 8.0], 0.3 mM KH_2PO_4, 0.87 mM Na_2HPO_4, 0.4% (v/v) Igepal-CA630, 0.017% (v/v) BSA, 3.9% (v/v) Optiprep, 2.17 mM DTT, 0.44 mM dNTPs, 1.16 U/ml RiboLock RNAse inhibitor, and 10.4 U/ml SuperScript-III RT enzyme	*Initiation:* 1 min at 60°C *RT reaction:* 2 hours at 50°C *Inactivation:* 15 min at 70°C
SuperScript IV (Invitrogen, cat. no. 8090010)	9.8 mM KCl, 45 mM NaCl, 0.4 mM $MgCl_2$, 6.5 mM Tris-HCl [pH 8.0], 0.3 mM KH_2PO_4, 0.87 mM Na_2HPO_4, 0.4% (v/v) Igepal-CA630, 0.017% (v/v) BSA, 3.9% (v/v) Optiprep, 0.435 SSIV RT buffer, 2.17 mM DTT, 0.44 mM dNTPs, 1.16 U/ml RiboLock RNase Inhibitor, and 10.4 U/ml Maxima H minus RT enzyme	*RT reaction:* 60 min at 50°C *Inactivation:* 10 min at 80°C
Maxima H minus (Thermo Scientific, cat. no. EP0751)	43 mM KCl, 45 mM NaCl, 1.7 mM $MgCl_2$, 28 mM Tris-HCl [pH 8.0], 0.3 mM KH_2PO_4, 0.87 mM Na_2HPO_4, 0.4% (v/v) Igepal-CA630, 0.017% (v/v) BSA, 3.9% (v/v) Optiprep, 4.4 mM DTT, 0.44 mM dNTPs, 1.16 U/ml RiboLock RNase Inhibitor, and 10.4 U/ml Maxima H minus RT enzyme	*RT reaction:* 60 min at 50°C *Inactivation:* 10 min at 80°C

2.2.3. Comparison of cDNA cleanup strategies

After single-cell transcriptome barcoding, cDNA was cleaned up using two different approaches. The first step was the same for both protocols – hydrogel beads were removed by spinning down the post-RT reaction mix through a spin column (Zymo, cat. no. C1004-50) for 1 min at 1000 g. Next, the unused barcoding primers and primer dimers were removed by enzymatic digestion. The reaction was carried out in 80μl volume using 1μl of ExoI (Thermo Scientific, cat. no. EN0581), 2μl of HinFI (Thermo Scientific, cat. no.

FD0804) and 0.5μl of FastAP (Thermo Scientific, cat. no. EF0654) enzymes, and 0.5x FD buffer (Thermo Scientific, cat. no. B64). The digestion product was then purified with AMPure reagent (Beckman Coulter, cat. no. A63881) using a 1.2x volume ratio. Alternatively, no enzymatic digestion was carried out, and the post-RT reaction mix after hydrogel removal was diluted to 100μl with nuclease-free water and purified with AMPure reagent (Beckman Coulter, cat. no. A63881) using different volume ratios. Performance of cDNA cleanup strategies was evaluated by performing scRNA-Seq assay using K562 cells and comparing the transcript capture rates

2.2.4. Comparison of second strand synthesis reaction protocols

The second strand synthesis reaction was performed using two different protocols. One protocol relied on a commercially available reagent kit (NEB, cat. no. E6111S), and the reaction was performed in 20μl volume using 1μl of provided enzyme mix and 1x provided buffer solution. The other protocol relied on combining separate reaction components: the reaction was also performed in 20μl volume using 1x Second Strand Buffer solution (Thermo Scientific, cat. no. 10812014), 0.133μl of DNA Polymerase I enzyme (Thermo Scientific, cat. no. EP0041), 0.133μl of T4 DNA ligase enzyme (Thermo Scientific, cat. no. EL0012), 0.533μl of RNaseH enzyme (Thermo Scientific, cat. no. EN0202) and 200μM dNTP. Performance of SSS protocols was evaluated by performing scRNA-Seq assay using K562 cells and comparing the transcript capture rates.

2.2.5. Comparison of IVT reaction kits

Two different commercially available kits were used to perform IVT reaction: HiScribe T7 High Yield RNA Synthesis Kit (NEB, cat. no. E2040S) and TranscriptAid T7 kit (Thermo Scientific, cat. no. K0441. In both cases, the reaction was carried out in 80μl volume using 20μl of second strand synthesis reaction product. IVT reaction was performed at 37°C for 14 hours in a heated air thermostat. Afterward, the reaction product was purified with AMPure reagent (Beckman Coulter, cat. no. A63881) using a 1x volume ratio and eluted into 20μl of nuclease-free water. Performance of IVT kits was evaluated by performing scRNA-Seq assay using K562 cells and comparing the transcript capture rates.

2.3. Cell Samples

2.3.1. K562 cell culturing

A commercially available K562 cell line was used in this work (ATCC, cat. no. CCL-243). Cells were cultured using IMDM culturing media (Gibco, cat. no. 12440053) supplemented with 10% FBS (Gibco, cat. no. 10270-106) and 1x Pen-Strep (Gibco, cat. no. 15140122). Cells were cultured in 25cm^2 tissue culture flasks (Corning, cat. no. 430639) and split to a ratio of 1:6 every 2 to 3 days or once they reached 80% confluence.

2.3.2. HMLE cell culturing

HMLE Cell Lines used in this work were shared by the Robert Weinberg Lab (MIT). HMLE and all derived cell lines were cultured in MEGM (Mammary Epithelial Cell Growth Medium) media (Lonza, cat. no. CC-3051). Cells were cultured in round tissue culture dishes 10cm in diameter (Corning, cat. no. 430167) and split to a ratio of 1:7 every 2 to 3 days or once they reached 80% confluence on a plate. All cell dissociations were performed using 1X TrypLE Express reagent (Gibco, cat. no. 12604013).

EMT was induced in HMLE cells by the addition of Recombinant Human TGF-β1 (HEK293 cell-derived) (PeproTech, cat. no. 100-21) to the culture media to a final concentration of 5ng/ml. EMT was also induced by overexpression of the ZEB1 transcription factor. HMLE cells transfected with FUW plasmid, a tetracycline operator, and minimal CMV promoter were used, and the ZEB1 gene overexpression was induced by the addition of doxycycline (Sigma, cat. no. D3447) to the culture media to a final concentration of 1 μg/ml. All cells under induction were passaged once they reached 80% confluence.

2.3.3. K562 cell barcoding

To prepare the cells for scRNA-Seq experiments, they were cultured to 70% confluence, harvested, and kept at +4°C at all times. Three 1x PBS (Gibco, cat. no. 20012027) washes were performed on the cells, and cell viability was evaluated using trypan blue staining prior to scRNA-Seq. All single-cell transcriptome barcoding experiments were performed with cell viability exceeding 95%. The resulting suspension of single-cells was diluted to 140000 cells/ml and supplemented with 16% (v/v) Optiprep (Sigma-Aldrich, cat. no. D1556) and 0.05% (w/v) BSA (Carl Roth, cat. no. 8076.2) and encapsulated into 3 nL droplets. The cell encapsulation was set at ~30000 cells per hour, and over 75% of cells entering microfluidics chips were co-encapsulated with one DNA barcoding hydrogel bead. cDNA synthesis in droplets was performed under different reaction conditions, as described in section 2.2.2. The sequencing library was prepared using different reaction conditions outlined in section 2.2.2.

2.3.4.HMLE cell barcoding

To prepare the cells for scRNA-Seq experiments, they were cultured to 70% confluence and dissociated from the plate with the addition of 3ml of trypsin for 5 mins at 37°C. After dissociation, the cell samples were kept at +4°C at all times in MEGM-complete media. Three 1x PBS (Gibco, cat. no. 20012027) washes were performed on the dissociated cells, and cell viability was evaluated using trypan blue staining prior to scRNA-Seq. All single-cell transcriptome barcoding experiments were performed with cell viability exceeding 90%. The resulting suspension of single-cells was diluted to 140 000 cells/ml and supplemented with 16% (v/v) Optiprep (Sigma-Aldrich, cat. no. D1556) and 0.05% (w/v) BSA (Carl Roth, cat. no. 8076.2) and encapsulated into 3 nL droplets. The cell encapsulation was set at ~30000 cells per hour using, and over 75% of cells entering microfluidics chips were co-encapsulated with one DNA barcoding hydrogel bead. SuperScript III RT enzyme and RNAseOUT inhibitor were used for cDNA synthesis (Table 2.6, SuperScript III). The RT reaction was initiated by transferring the emulsion to 50°C for 1-hour and terminated by incubating for 15 min at 75°C. After the transcriptome barcoding reaction, the barcoded cDNA was subjected to enzymatic digestion and cleanup with magnetic beads (see section 2.2.3). Next, the second strand synthesis reaction was performed (NEB, cat. no. E6111S), and cDNA was amplified using IVT reaction (NEB, cat. no. E2040S). The library was then completed using the procedure described in section 2.1.2 and sequenced.

2.3.5.Breast tumor cell barcoding

Tissues were collected from women undergoing surgery for primary breast cancer. Healthy tissue was obtained from contralateral prophylactic mastectomies of the same cancer patients, and peripheral blood mononuclear cells (PBMCs) were obtained from patients prior to their surgical procedures. All samples were obtained after informed consent and approval from the Institutional Review Board (IRB) at Memorial Sloan Kettering Cancer Center. After tissue collection single-cell suspensions were prepared and isolated, FACS-sorted CD45+ cells were suspended in ice-cold 1X PBS (Gibco, cat. no. 20012027), diluted to 140000 cells/ml and supplemented with 16% (v/v)

Optiprep (Sigma-Aldrich, cat. no. D1556) and 0.05% (w/v) BSA (Carl Roth, cat. no. 8076.2), and encapsulated into 1.5 nL droplets together with custom-made BHBs and RT/lysis reagents. The microfluidics chip was operated at a throughput of ~30000 cells per hour, and over 75% of cells entering microfluidics chips were co-encapsulated with one DNA barcoding hydrogel bead. SuperScript III RT enzyme and RNAseOUT inhibitor were used for cDNA synthesis (Table 2.6, SuperScript III). The RT reaction was initiated by transferring the emulsion to 65°C for 1 min, followed by a 1-hour incubation at 50°C and 15 min at 75°C. After the transcriptome barcoding reaction, the barcoded cDNA was subjected to enzymatic digestion and cleanup with magnetic beads (see section 2.2.3). Next, the second strand synthesis reaction was performed (NEB, cat. no. E6111S), and cDNA was amplified using IVT reaction (NEB, cat. no. E2040S). The library was then completed using the procedure described in section 2.1.2 and sequenced.

3 RESULTS AND DISCUSSION

This dissertation consists of three major parts. In the first part the results of the optimization of the *inDrops* scRNA-Seq technology are presented. The second part describes the characterization of the epithelial to mesenchymal transition using HMLE cell model system. Finally, in the last part, the study of tumor infiltering immune cells is presented including the construction of the immune cell atlas infiltrating breast tumor. At the end of each section the results are discussed in a broader context.

3.1. Single-cell RNA-Seq protocol optimizations

3.1.1.Experimental approach of this study

Single-cell transcriptome analysis is a multistep process. Generally, the scRNA-Seq workflow consists of four distinct elements: 1) cell sample preparation, 2) single-cell transcriptome barcoding, 3) sequencing library preparation, and 4) next-gen sequencing and data analysis. Each of these elements also has a varying degree of technicalities. The overall performance of the assay will depend on the efficiency of each technical step. In the first part of this work the single-cell transcriptome barcoding and library preparation for sequencing were investigated. These two elements consist of seven separate molecular biology reactions that are performed in succession. The efficiency of these reactions determines the amount of information recovered from each cell as well as the level of noise in the data. The transcriptome barcoding and sequencing library preparation steps together form a single functional unit that remains unchanged between different biological samples and projects. Therefore, these two parts can be investigated independently of the sample type.

At the time of the start of PhD studies, high-throughput single-cell RNA-Seq was a novel technique, and robust protocols were lacking. Thus, the first part of this work was to validate and document the protocol for high-throughput single-cell RNA-Seq using a droplet microfluidics platform (*inDrops*). The result of this effort was protocol guidelines detailing single-cell transcriptome barcoding and next-generation sequencing [1]. These guidelines served as the basis for further optimizations and developments presented in this work.

After describing the detailed *inDrops* protocol efforts were focused on investigating different commercially available reagents for separate protocol steps and determining the optimal overall combination that would enable improved recovery of the unique transcripts of individual cells.

To prepare the barcoded transcriptomes for sequencing the single cells are first compartmentalized into droplets using a microfluidic chip, including the reagents needed for cell lysis, cDNA synthesis and barcoding (see section 2.1.1). Once encapsulated, the cells are lysed, and their mRNA is converted to barcoded-cDNA during the RT reaction (Figure 3.2). The barcoded-cDNA molecules from all cells are then pooled by breaking microfluidic droplets, purified, and amplified enzymatically. Finally, the DNA library is prepared for sequencing by fragmenting to the required size and adding sequencing adapters.

One could predict that inefficiencies in any of the steps between the transcriptome barcoding and cDNA amplification will lead to the loss of unique transcript molecules. Contrary, any losses of DNA material after library amplification will not lead to transcript loss as multiple copies of the same molecule already exist in the mix. Therefore, main efforts were focused on the workflow steps that may directly impact the transcript loss. These steps include: i) RT reaction, ii) cDNA cleanup, iii) second-strand synthesis, and iv) library amplification (Figure 3.2 and Figure 3.4 marked in red). The optimization strategy and results for each of these steps are discussed below individually.

Readout strategies also need to be considered. Three different analysis techniques were used throughout this work. Firstly, the transcript amount before library amplification was quantified by qPCR assay of eight different genes having different expression profiles (Table 2.5). Secondly, the yield and fragment size distribution of the amplified library after IVT reaction was analyzed. Finally, to directly compare tested conditions, barcoded single-cell RNA-Seq libraries were prepared and sequenced. After sequencing, the cell barcodes were deconvoluted and assigned to individual cells, the cDNA sequences were aligned to the genome to identify the active genes, and the relative number of transcripts of each active gene were then quantified by counting the abundance of molecular barcodes (known as unique molecular identifiers). After sequencing data processing, the median number of captured transcripts per cell was determined and used as a major metric for comparisons. A commercially available cultured cell line – K562 (human immortalized myelogenous leukemia cells) was used in the analysis, to keep the external factors as constant as possible.

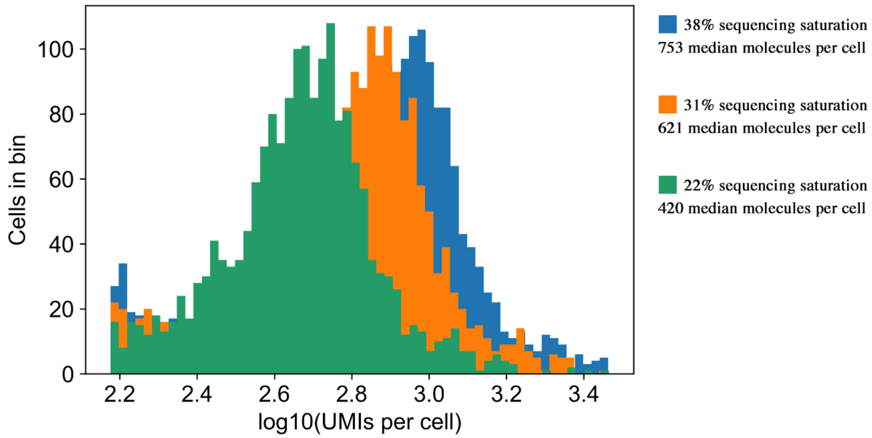


Figure 3.1. *Histogram indicating transcript capture at different sequencing depths. Original data (blue) was downsampled to simulate lower sequencing depths.*

Equation 3.1 *Sequencing saturation metric. Unique molecules are considered unique combinations of cell barcode, UMI and transcripts. Productive reads are reads that have an identifiable combination of cell barcode, UMI and transcripts*

$$sequencing\ staturation = 100 * (1 - \frac{unique\ molecules}{productive\ reads})$$

It is important to note that sequencing results are directly related to sequencing depth. Literature indicates that sequencing saturation is approached as 500000 to 1000000 reads per single cell is reached [30]. However, reaching such saturation is often prohibitively costly. Therefore, most of the sequencing is being conducted under so called shallow depth (20000 - 50000 reads per cell). Under shallow sequencing conditions (before reaching saturation) the recovery of unique transcripts is directly related to the sequencing depth, as shown in Figure 3.1. Therefore, the 'sequencing saturation' metric was used throughout the work to account for shallow sequencing. Sequencing saturation is based on the fraction of PCR duplicates in the sequencing data (Equation 3.1). For example, 75% sequencing saturation means that for every four sequencing reads, there will be one new unique molecule identified. Two samples can be directly compared with one another if they have the same sequencing saturation. Therefore, in the cases where sequencing saturation between the samples was different, the sample that had more reads associated with it was downsampled to match the saturation rates of the other sample(s). Downsampling was performed by randomly discarding reads from the read array associated with the particular sample. Due to economic considerations low sequencing saturation (typically below 20%) was used throughout this work. As a result, small differences in reaction efficiency may have been missed.

3.1.2. Single-cell transcriptome barcoding optimization

To perform the transcriptome barcoding step a suspension of single cells is compartmentalized in microfluidic droplets with barcoding hydrogels carrying barcoding primers as well as lysis and RT reaction reagents (Figure 3.2, panel A). Once encapsulated, cells are exposed to lysis reagent are lysed and the barcoding primers are released from the hydrogels by UV-light (365 nm), and their mRNA is converted to barcoded cDNA by reverse transcriptase (Figure 3.2, panels A and B). In the original protocol (baseline conditions), the RT reaction is performed using the SuperScriptIII RT enzyme and RiboLock RNAse inhibitor. In this work, the performance of three different commercially available RT enzymes was investigated: SuperScriptIII, SuperScriptIV, and Maxima H-. First, the stability of the enzymes on ice (4°C) was evaluated. This property of the enzyme is important because RT reaction reagents may spend up to two hours on ice before the transcriptome barcoding reaction begins, a time that is needed to isolate ~1 million single-cells. The aforementioned enzymes were incubated in either their native or in scRNA-Seq reaction buffer for two hours, and their residual activity was determined by performing RT-qPCR assay. All three enzymes retained 100% of their activity after being stored for two hours in supply as well as scRNA-Seq reaction buffers.

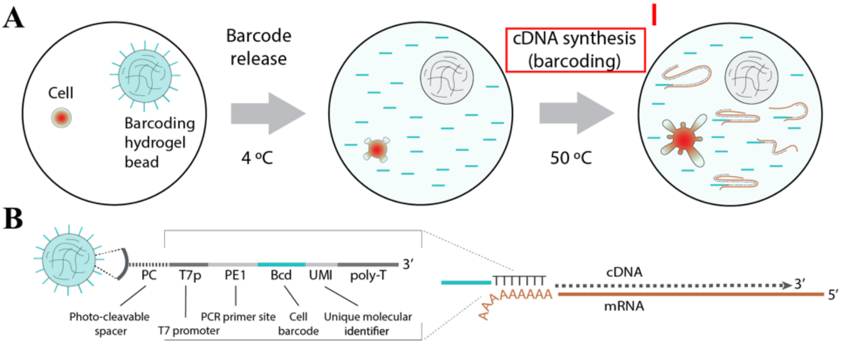


Figure 3.2. *Single-cell compartmentalization and transcriptome barcoding. Panel A: single cell lysis and barcoded cDNA synthesis in droplets. Panel B: structure of barcoding primers attached to hydrogel beads.*

Next, the performance of the enzymes was evaluated in the scRNA-Seq assay. Summarized results presented in Table 3.1 show that Maxima H- RT enzyme performs the best in the scRNA-Seq assay. For example, we found that it increases the capture rate of transcripts per cell by 3.4-times as compared to the SuperScriptIII enzyme. Interestingly, the SuperScriptIV enzyme was not the top performer even though it is the most advanced reverse transcription enzyme based on properties reported by the manufacturer. Differences in enzyme properties and in their amino acid sequences may explain the observed significant differences in enzyme efficiency. Even though all three enzymes are based on the M-MuLV RT enzyme, they have been engineered to have improved processivity, thermostability, and efficiency. Moreover, it is important to point out that the Maxima H- enzyme was not only the most efficient but also the cheapest of three RT enzymes tested.

Table 3.1. *RT enzyme comparison in scRNA-Seq assay. Sequencing results***.** *Workflow conditions: SS – cDNA cleanup by size selection, NN – NEBNext SSS kit, HS – HiScribe IVT kit.*

RT enzyme	**Median transcripts per cell**	**Sequencing saturation**	**Workflow conditions**
SuperScript III	952	11.79 %	SS, NN, HS
SuperScript IV	2220	11.84 %	SS, NN, HS
Maxima H-	3304	10.68 %	SS, NN, HS

3.1.3. Library preparation optimization

After transcriptomes of single cells are barcoded in droplets, the material is pooled, and the remaining steps are performed in a single solution. Because each cDNA molecule has an associated cell barcode, it can be assigned to a particular single cell, and the UMI tags allow to count all unique molecules for any given cell. As discussed above, before the barcoded cDNA is amplified, every loss of material will result in information loss making the initial library preparation steps critical (Figure 3.3). The unoptimized protocol (baseline conditions) relied on enzymatic digestion to cleanup barcoded cDNA, second-strand synthesis (SSS) was performed using the NEBNext kit, and the library was amplified using HiScribe kit.

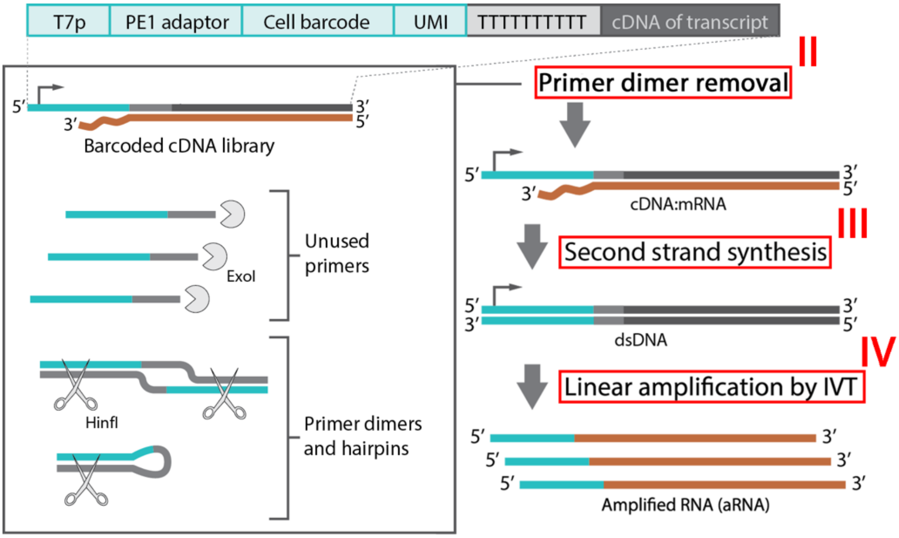


Figure 3.3. *Sequencing library preparation steps that flow the RT reaction in droplets. Red marks the separate protocol steps that were optimized in this work. II – primer dimer removal; III – second strand synthesis reaction; IV – library amplification by IVT.*

Firstly, the purification of barcoded cDNA was investigated. The cleanup of post-RT reaction mix is needed to remove primer dimers and the remaining unused barcoding primers. Primer dimers are dsDNA molecules that are generated during RT reaction if primers tend to form heteroduplex. Excess amount of barcoding primers can facilitate primer dimer formation. As a result, after the RT reaction is completed the reaction mix typically contains not only barcoded cDNA but also primer dimers and a significant amount of unused ssDNA primers. On capillary electrophoresis gel the unused primers and primer dimers form a distinct peak(s) in the 25-200nt region (Figure 3.4**,** red arrows). These relatively short DNA fragments can contribute to the noise in the sequencing data and reduce the fraction of useful reads, effectively increasing the sequencing cost. Two strategies for primer cleanup were investigated. The first strategy was dubbed "enzymatic digestion". Due to the specific sequence, the primer dimers (dsDNA) carries HinFI restriction endonuclease recognition site and thus can be hydrolyzed enzymatically. On another hand, unused ssDNA barcoding primers can be removed by single-stranded DNA exonuclease Exo I. Each of the enzymes was evaluated separately as presented in Figure 3.4**,** panel A. Successful cDNA cleanup should lead to the reduction of the 25-200nt fragments in the library. This can be evaluated by analyzing the fragment size distribution of the amplified library (Figure 3.4, red arrows). Ideally the amount larger size fragments should remain unchanged as they are the barcoded and amplified transcripts. Results show that using the enzymatic digestion cDNA cleanup strategy leads to the loss of material across the whole library (Figure 3.4, red trace). Furthermore, investigating the effect of each enzyme separately revealed that digestion with ExoI is the main cause of the library loss. Digesting the cDNA with HinFI alone does not significantly reduce the amount of the library but it also does not affect the 25-200nt fragments region.

An alternative strategy for barcoded cDNA cleanup is dubbed "size selection", where different size fragments are purified from solution using magnetic beads. Using different volume ratios of the reaction mixture to AMPure reagent leads to different fragments being bound to magnetic beads. The bound fragment can then be recovered, while unbound fragments will be left in solution. In this way the primer dimers and unused primers (25-200nt fragments) can be removed from reaction mix. Different ratios of AMPure reagent were investigated, to investigate the size selection cDNA cleanup approach (Figure 3.4**,** panel B). In accordance with the manufacturers' manual, results show that using lower AMPure reagent to reaction mixture volume ratios will lead to a more efficient removal of small fragments. Furthermore, as compared to enzymatic digestion approach, the size selection strategy preserves more of the large fragments (Figure 3.4**)** . Next, qPCR was performed to determine the optimal AMPure ratio for the size selection protocol. Results (not presented) revealed that no transcript loss occurs if the ratio of AMPure reagent to reaction mixture is equal or higher than 0.8x. Finally, to compare the two different strategies for the barcoded cDNA cleanup scRNA-Seq experiment was performed (Table 3.2). Sequencing data confirm that enzymatic digestion leads to a significant loss in library diversity. Using the size selection based barcoded cDNA cleanup strategy on average allows to recover three times more unique transcripts from each single cell. It is important to point out that chronologically this was the first step to be optimized. As a result, all subsequent optimizations of different protocol steps presented in this study used size selection strategy for barcoded cDNA cleanup.

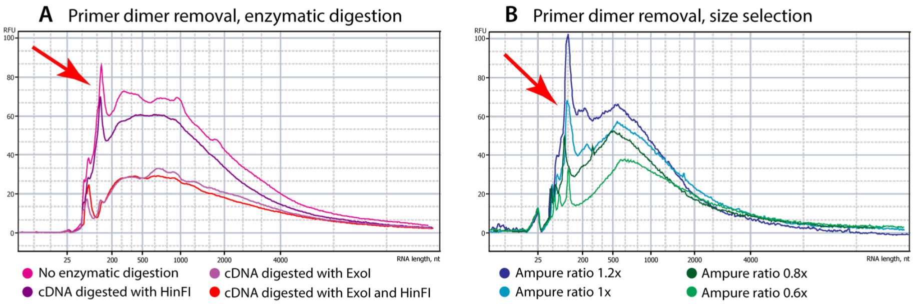


Figure 3.4. *Total library amount and fragment size distribution for different cDNA cleanup strategies. Red arrows mark the primer dimers. Panel A: results of primer dimer removal by enzymatic digestion. Panel B: results of primer dimer removal by size selection.*

Table 3.2. *Barcoded cDNA cleanup strategy comparison in scRNA-Seq assay. Sequencing results. Workflow conditions: SSIII – SuperScript III RT enzyme, NN – NEBNext SSS kit, HS – HiScribe IVT kit.*

Cleanup strategy	Median transcripts per cell	Sequencing saturation	Workflow conditions
Enzymatic digestion	312	11.46 %	SSIII, NN, HS
Size selection (*0.8x volume ratio*)	952	11.79 %	SSIII, NN, HS

After finding that the size selection protocol can be used to effectively clean up the barcoded cDNA , the second-strand synthesis (SSS) reaction was investigated. The SSS reaction consists of three separate enzymatic reactions and can be summarized in the following. First, RNAse H fragments the RNA strand in the RNA:cDNA duplex. The resulting single-strand breaks serve as priming sites for the second-strand DNA molecule synthesis by DNA polymerase I. Once the second DNA strand is synthesized, T4 ligase repairs any remaining single-strand breaks in the dsDNA molecule.

In this work, two commercially available reagent kits for second-strand synthesis were compared: NEBNext and SuperScript. The barcoded cDNA from single-cells was purified and then treated with the SSS reagents provided in each kit. As shown in Table 3.3, at 11% sequencing saturation, both kits produce very similar results with SuperScript showing a marginal increase in recovered transcript amount. The libraries were sequenced to a higher saturation to determine if the observed difference is significant. At 22% sequencing saturation the difference between SuperScript and NEBNext kit became even more pronounced. The recovery of barcoded transcripts was 1.2-

times higher when using SuperScript kit. Furthermore, it is likely that the barcoded transcript recovery efficiency may be further optimized by changing the amount of each enzyme in the reaction. Yet this possibility was not explored in this work because all the reagent kits were used according to the manufacturer's recommendations.

Table 3.3. *Second strand synthesis kit comparison in scRNA-Seq assay. Sequencing results. Data from two separate comparison experiments. Workflow conditions: SSIII – SuperScript III RT enzyme, SS – cDNA cleanup by size selection, HS – HiScribe IVT kit.*

SSS reagent kit	Median transcripts per cell	Sequencing saturation	Workflow conditions
Comparison experiment 1			
NEBNext	952	11.79 %	SSIII, SS, HS
SuperScript	1022	11.41 %	SSIII, SS, HS
Comparison experiment 2			
NEBNext	1392	22.77 %	SSIII, SS, HS
SuperScript	1689	22.91 %	SSIII, SS, HS

Finally, the library amplification reaction was investigated. After all previous steps are completed, the library is amplified during the IVT reaction. This amplification strategy is possible because the barcoded primers have a T7 promotor site, which becomes double-stranded after SSS reaction and can bind T7 RNA polymerase (Figure 3.2**,** panel B). Two different commercially available IVT kits were compared: HiScribe, TranscriptAid. Firstly, amplification efficiency was evaluated by analyzing the total library amount and fragment distribution (Figure 3.5). Results show that the HiScribe IVT kit performs better in terms of overall material yield. However, the amount of amplified library may not directly correlate to library diversity (captured transcript amount). To further compare library amplification kits scRNA-Seq assay was performed. Sequencing data agree with previous results indicating that the HiScribe kit is better than the TranscriptAid kit (Table 3.4). Due to low sequencing saturation, the exact difference in median transcript recovery between the two tested kits cannot be confidently determined. However, because the HiScribe kit was already included in the baseline protocol, further sequencing experiments were not performed.

Table 3.4. *Library amplification kit comparison in scRNA-Seq assay. Sequencing results. Workflow conditions: SSIII – SuperScript III RT enzyme, SS – cDNA cleanup by size selection, NN – NEBNext SSS kit.*

IVT reagent kit	Median transcripts per cell	Sequencing saturation	Reaction conditions
HiScribe	434	4.13 %	SSIII, SS, NN
TranscriptAid	388	4.58 %	SSIII, SS, NN

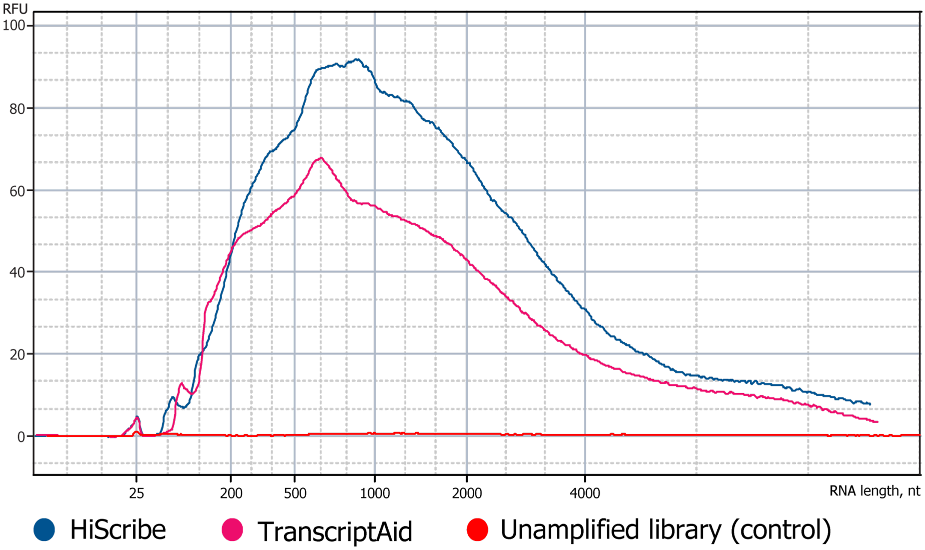


Figure 3.5. *Total library amount and fragment size distribution for different IVT kits.*

3.1.4.Discussion

In the first part of this thesis, a method for high-throughput single-cell transcriptome barcoding using droplet microfluidics termed *inDrops* was described in full detail in a publication [35]. Next, individual steps of the scRNA-Seq protocol were optimized to increase the overall capture of single-cell transcriptome. The overall increase in efficiency can be evaluated by comparing “Enzymatic digestion“ sample in Table 3.2 to the “Maxima H-“ sample in Table 3.1. This comparison reveals that improving the RT, cDNA cleanup, SSS and IVT reactions allowed to achieve up to ten times higher transcript detection as compared to the original *inDrops* protocol. However, it is important to point out that no sample in this study was sequenced to saturation. Therefore, the numbers presented in this study should not be used for direct comparison to other single-cell transcriptome analysis methods without accounting for sequencing saturation. In the context of existing literature, low sequencing saturation used in this work will, in most cases, prevent a direct comparison. It is also important to note that in this work, a cultured cell line (K562 cells) was used for the experiments. This allowed minimizing variations coming from differences in cell sample between the different experiments. However, the cell line used in this work is a homogenous cell population. Therefore it was not possible to evaluate if the increase in transcript recovery also improves the detection of cell types in the sample.

Independent protocol benchmarking studies have repeatedly shown the *inDrops* method to capture lower number of transcripts as compared to other scRNA-Seq platforms [380, 381]. However, in these and other studies, the original (unoptimized) version of the *inDrops* protocol was used. Interestingly, even though the comparison studies show the *inDrops* platform to have a low transcript recovery rate, it is as sensitive as other high-throughput methods in identifying cell types in a heterogeneous population [380]. Such observation raises the question of the importance of the transcript recovery metric. Overall, the low-throughput methods recover significantly more unique transcripts from single cells than the high-throughput platforms. However, rare cell types are typically not detected due to the limited number of cells being analyzed. For example, if a particular cell type makes up 1% of the total population, the probability of observing at least one cell of that type in a sample of 100 cells is only 0.63. Hence if a total of only a few hundred of cells were sampled at random from a large pool of cells the likelihood to miss these rare cell types is high. On the other hand, high-throughput scRNA-Seq platforms capture a sufficient number of transcripts to assign cell types and can leverage the large cell numbers to detect rare cell populations. Although transcript capture rate is important for studying genes that have low expression, however, not all studies require this, and in many cases, the ability to profile tens of thousands of individual cells outweighs the benefits of deep transcriptome profiling of just a handful individual cells. As a result, high-throughput platforms become more attractive when studying complex biological systems with many different cell types and states. ScRNA-Seq protocol and improvements presented in this work significantly increase the efficiency of the *inDrops* platform and make it an attractive method for single-cell transcriptome barcoding studies.

The two most critical steps for overall protocol efficiency, as determined by this study, are RT reaction and barcoded cDNA cleanup. However, some technical limitations of this work need to be taken into consideration. Firstly, not all commercially available reagents were investigated. For example, several available reverse transcription enzymes were not included in the study due to economic considerations. Therefore, it is possible that the protocol may further be optimized by testing other commercially available enzymes. Furthermore, the enzymes and reagents were used as detailed by the manufacturer's protocols. This means that potentially a further increase in transcript detection may be achieved by adjusting reaction conditions. For example, cell lysis reagents, barcoding primer structure and concentration in the reaction mix, are clear targets for further investigation. Similar to the work presented in this thesis, others have also shown that second-strand synthesis reaction optimization can improve single-cell transcriptomics protocol [60]. Interestingly, the study found that diluting the enzyme mix of NEBNext kit can significantly increase transcript recovery – an option that can be explored in the future.

It is important to separate two different scRNA-Seq workflow steps – transcriptome barcoding and sequencing library preparation. The transcriptome barcoding is directly related to cell compartmentalization as material from different cells can only be pooled after cDNA barcoding step. Although the components of the RT reaction were investigated in this work, the droplet size could be also important as it directly defines the reaction volume and system throughput. Smaller droplets would allow to increase the droplet generation rate and the experiments could be performed faster. It is also important to note, open systems such as presented in this work, are less user-friendly and require more hands-on time than commercial alternatives. However, the former offer significant cost savings and additional flexibility.

Two different strategies for barcoded cDNA amplification and sequencing library preparation exist, as discussed in the literature review section 1.1.2. The approach investigated in this study relies on *in vitro* transcription. The alternative strategy for library amplification utilizes the so called template switching property of the RT enzymes when three GGG nucleotides are introduced at the end of the transcripts. These triplets can then be used to introduce a DNA sequence that will be common to all 3‘ ends of barcoded cDNA molecules. The resulting library will have common sequences on both cDNA molecule ends and can be amplified via PCR. Therefore, scRNA-Seq protocols that use template switching approach do not require the SSS reaction and a lengthy IVT reaction reducing required hands-on time as well as overall library preparation time. Using the PCR based amplification, libraries can be typically prepared in the course of a single day while IVT based amplification requires two days. This is not very important when only a few samples are processed, yet for routine applications, the use of PCR-based amplification may lead to significant savings in time. Noteworthy, at this point no observations about protocol efficiency can be made as no direct comparison between two library amplification methods has been published.

Overall, the results of this section reveal improvements of the *inDrops* platform after careful optimization of individual steps in the scRNA-Seq library preparation protocol. Commercially available enzymes were investigated, and the combined effect of separate optimizations allowed to achieve a ten times higher transcript capture rate. This increase in the efficiency of the *inDrops* method is important because independent benchmarking studies have revealed that the original protocol suffered from poor performance. The presented results ask for a separate benchmarking study in order to evaluate the improved *inDrops* in the context of other available high throughput single-cell transcriptomics methods.

3.2. Studying EMT at the single-cell level

3.2.1.Characterizing the EMT process

The second part of this thesis investigated whether or not the data generated with scRNA-Seq *inDrops* platform can be used to uncover transcriptional networks that control a complex biological process. For this purpose, the EMT is particularly well suited. This process is a gradual cell state transition that is controlled by multiple signaling networks. During the EMT, cells lose epithelial markers (including E-cadherin, Epcam, and Epithelial Cytokeratins), and gain mesenchymal markers (including Vimentin, Fibronectin, and N-cadherin). Furthermore, it is known that the transcriptional changes are orchestrated by the core transcription factors: SNAIL, SLUG, TWIST1, ZEB1, and ZEB2. However, to date, the EMT process has been mainly studied by comparing the extreme states of EMT: the beginning (epithelial state) and the endpoint (mesenchymal state). Moreover, most studies to date have used bulk measurements that cannot reveal the subtle changes of individual cells. Therefore, while the initial and the final states of EMT are well characterized, little is still known about intermediate states and the temporal dynamics of the process.

Like most scRNA-Seq methods, the *inDrops* platform does not capture the full transcriptome of each single cell and as a result the data matrix that summarizes gene expression of each single cell is sparse. Most transcript count values in such data matrix are 0, as no information for that particular transcripts is recorded. This means that gene-gene correlation analysis could not be performed using the raw data and the missing values had to be inferred to reveal the EMT process. In particular, gene-gene correlation analysis was extensively used throughout this part of the thesis. To enable this an imputation algorithm had to be applied (Figure 3.6, panel A and B). The algorithm used in this study is called MAGIC (Markov Affinity-based Graph Imputation of Cells) and relies on data diffusion to impute the missing data (see literature review section 1.2.4). All analyses presented in this work have been carried out on imputed data unless specified otherwise. The importance and limitations of imputation are discussed below.

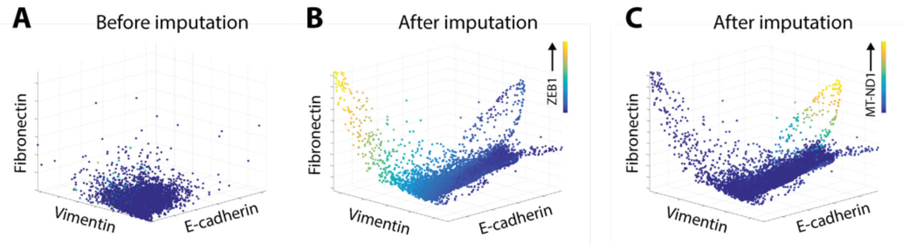


Figure 3.6. *3D scatterplots between canonical EMT genes E-cadherin, Vimentin, and Fibronectin. Each dot represents a single cell in transcriptional space. Pane A: before imputation. Panel B: after imputation with cells colored by the level of ZEB1. Panel C: after imputation with cells colored by the level of MT-ND1*

In this work, the EMT process was studied by stimulating transformed mammary epithelial cells (HMLE) with the TGF-β factor. Such treatment activates a canonical EMT pathway and provides a relevant model of the process. To thoroughly investigate cell transition, HMLE cells were continuously stimulated for 12 days. Throughout stimulation, samples were taken every two days, and single-cell transcriptomes were barcoded using the inDrops platform. The scRNA-Seq data analysis has revealed that the induction of EMT is asynchronous process, and each cell progresses through the transition at a different rate. As a result on days 8 and 10, cells occupy all states along the continuum of the EMT (Figure 3.6, B and C). The process is characterized by the decrease in E-cadherin (Epithelial state marker) and a simultaneous increase in Vimentin and Fibronectin (Mesenchymal state markers). Furthermore, expression of ZEB1, a key transcription factor for the EMT process responsible for the mesenchymal phenotype, progressively increases as expression of Vimentin and Fibronectin increase (Figure 3.6, panel B). The initial characterization fully agrees with the EMT model that is described in the literature. However, another progression was also revealed by scRNA-Seq analysis. It involves two branches that deviate from the main structure (Figure 3.6, C). These side branches display an increase in mitochondrial RNA, reflecting a progression into apoptosis (Figure 3.6, panel C, colored by mitochondrial gene expression). The apoptotic transition hypothesis is also supported by the expression of apoptotic markers in these cells (Figure 3.8). This observation also agrees with the reports in the literature as the role of TGF-β signaling in promoting apoptosis has been demonstrated to be important in the context of EMT [382].

Next, a more detailed characterization of cell states was performed. The transcriptome analysis revealed that most of the cells (79%) reside in an intermediate state that is neither epithelial nor mesenchymal. In the literature, this state is termed partial EMT and has been shown to be important in various biological processes, in particular, in the context of cancer. The intermediate cells are highly heterogeneous, and as a result, in the high-dimensional gene expression space, the data forms a multi-dimensional manifold that does not follow a simple one-dimensional progression. The archetypal analysis was used to characterize this structure and determine the main cell states of the EMT progression. The algorithm has identified 10 archetypes (AT) in the data, where each archetype corresponds to a particular cell state (Figure 3.7). Notably, not all cells can be confidently assigned to a particular cell state (Figure 3.7, panel B, grey color). It is likely that these cells were undergoing active transcriptional changes at the time of sampling and thus could not be assigned to a particular state.

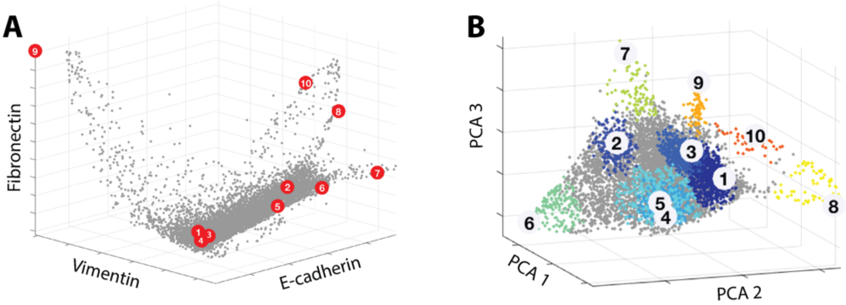


Figure 3.7. *EMT transition visualized by gene expression (A) and principal components (B). Red dots (A) and black numbers (B) represent each of the 10 archetypes in the data. Panel A: plotted by E-caherin, Vimentin, and Fibronectin. Panel B: plotted by PCA, colored by archetype. Grey cells are not associated with any archetype.*

Next, differential gene expression analysis was performed to characterize each of the identified cell states. The archetypes - states fall into the following categories: AT6, AT7 - 'epithelial,' AT1 to AT5 - 'intermediary,' AT9 - 'mesenchymal,' and AT8, AT10 - 'apoptotic'. Each archetype is characterized by the expression of particular genes, transcription factors, and chromatin modifiers (Figure 3.8). The epithelial cells (AT6 and AT7) are defined by strong epithelial marker expression, including CDH1, CDH3, MUC1, and CD24. Oppositely, mesenchymal state (AT9) is characterized by high expression of core EMT TFs – SNAIL, ZEB1, and TWIST1. The transcriptional profile of AT7 cells includes higher expression of ESR2 and GATA3 genes, commonly associated with the luminal mammary epithelial cells, and higher CD24 and CDH1 expression, suggesting a more differentiated epithelial phenotype than AT6 cells. Interestingly, both AT6 and AT7 cells express high levels of SOX4, which is an early master regulator of the TGFβ induced EMT [383]. This show that no cells in this study were fully epithelial at the time of sampling. Which is to be expected as cells were sampled after extended TGF-β stimulation (day 8 and day 10).

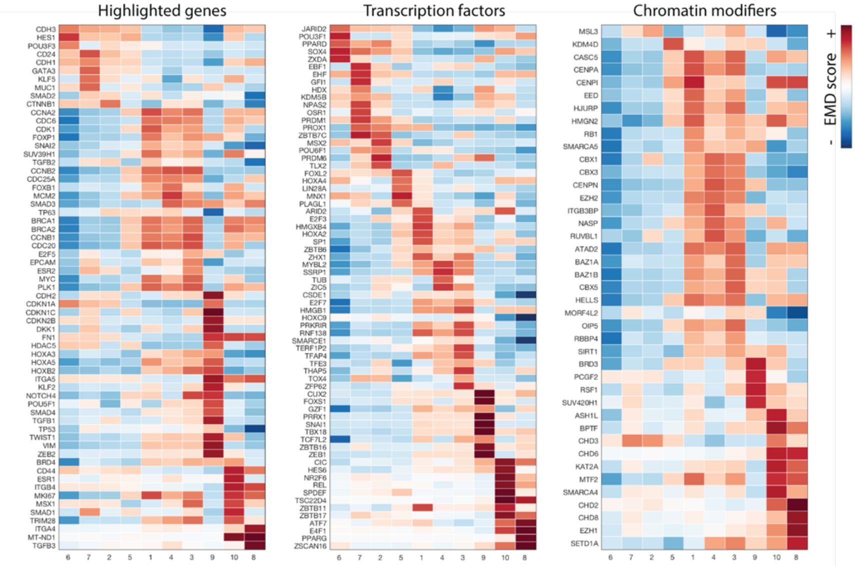


Figure 3.8. *Gene expression differences between different cell states. Each column represents the distinct cell state as defined by archetype analysis. To quantify gene expression differences between archetypes the earth mover distance (EMD) was used as defined previously [179] .*

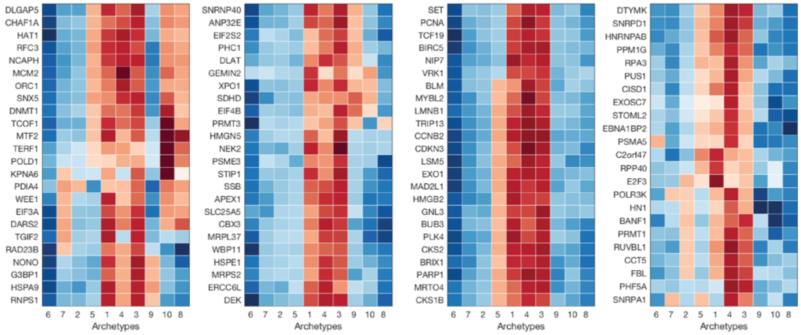


Figure 3.9. *A subset of differentially genes associated with embryonic development [384] for each cell state. Each column represents the distinct cell state as defined by archetype analysis.*

The analysis also identified five intermediate cell states (AT1–AT5), which fill the middle of EMT continuum. These cells vary between each other as they have undergone different extents of partial EMT. The fact that five distinct states can be identified supports the hypothesis that metastable cell states exist [194]. AT2 shows a similar gene expression profile as AT7, including the upregulation of SOX4 and is closest to the epithelial state. However, AT2 already expresses the KLF5 transcription factor, which, together with SOX4, acts to promote cell transition along the EMT [385]. Interestingly, if the cells do not express KLF5, they can undergo a SOX4 mediated apoptosis. This can also be observed in Figure 3.7 panel A where the AT8 apoptotic cells are branching out from AT2 cell population. Next, the AT3 is closest to the mesenchymal state. This archetype is characterized by increased expression of SMAD3 and an early mesenchymal phenotype regulator MSX1. It is also interesting to note that AT1, AT3, and AT4 cell states all express a large number of chromatin modifiers (Figure 3.8). This suggests that substantial chromatin remodeling takes place during cell reprogramming. Unsurprisingly, mesenchymal, and apoptotic cell states also have a distinct profile of chromatin-modifying gene expression. Notably, the intermediate state cells (AT1, AT3, AT4, and AT5) all show the increased expression of genes that are known to be active in embryonic stem cells (Figure 3.9) . It has been previously suggested that epithelial cells undergoing EMT may revert to a more primitive state before acquiring the ability to differentiate into a mesenchymal cell state [386]. This observation is supported by the results of this study.

The first part of the analysis highlights how single-cell transcriptomics can be used to study a complex biological process like EMT. The expected signals of the EMT process were observed, giving confidence that it was adequately sampled. Most strikingly, five distinct intermediate states were uncovered, suggesting the gradual nature of the EMT process that can be characterized by meta-stable states as suggested by others. The next part of the work focused on studying the temporal dynamics of gene expression during the EMT.

3.2.2.Regulatory networks and dynamics of the EMT process

The core gene regulatory circuitry of EMT is well defined. For example, it is known that ZEB1 and SNAIL are potent repressors of the epithelial phenotype and act both as transcriptional activators and repressors. However, the breadth of targets regulated by these EMT-TFs remains mostly unknown. Defining the EMT circuitry and the timing of different regulatory factors is vital for understanding how this cell state transition occurs. Data collected in this study spans the whole EMT continuum and is suitable to explore the temporal trends as cells progress from the epithelial to the mesenchymal state.

A pseudo-time of the cell state transition had to be established in order to analyze the gene expression dynamics. In this study the expression of VIM gene was used as a proxy for the EMT state as it gradually increases other the course of the transition. Thus, ordering cells based on their VIM gene expression allows to track position in the EMT continuum. The expression of each gene in the analysis can then be correlated to the expression of VIM to observe their individual temporal trends (Figure 3.10). As could be expected, several different gene expression dynamics were found. For example, the expression of genes can decrease or increase as EMT progresses (Figure 3.10, CDH1 and ZEB1). A considerable number of genes peak or are repressed at intermediate levels of VIM (Figure 3.10, SLUG, and ZMAT3). Such dynamics suggest that some genes are essential to the intermediate cell states – partial EMT. Furthermore, because their expression is similarly low/high in both the epithelial and mesenchymal states, these genes would be missed by studies that focus only on the end states.

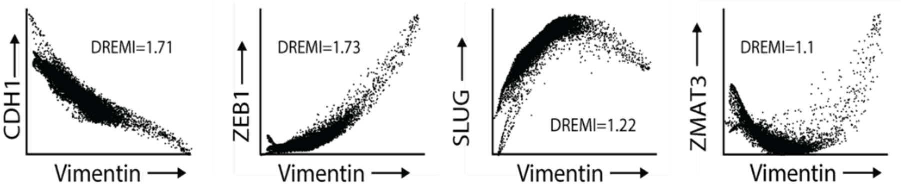


Figure 3.10. *2D scatterplots showing gene-gene relationships. Each dot represents a single cell. Cells are plotted based on their gene expression. Conditional probability score as calculated by DREMI algorithm is indicated in each case.*

Next, to systematically explore gene-gene interactions, a quantitative metric was used to score the statistical relationship between genes. An adaption of the DREMI algorithm was used [387]. The algorithm captures the functional relationship between two genes across all cells in the progression and calculates a score (conditional probability) (Figure 3.10). Using this metric, a genome-wide view of expression dynamics during the course of EMT could be constructed to uncover the transcriptional networks that govern the cell state transition. Firstly, apoptotic cells were filtered out (based on MT-ND1 expression). This was done in order to focus the analysis only on the EMT process. Next, the remaining cells were used to compute the DREMI score between VIM and each gene captured in the analysis. Interestingly, the majority of the genes demonstrated a temporal trend that follows VIM. Such a result reveals the vast extent of the cellular changes during EMT. For the subsequent analysis, 13,487 genes that have DREMI >0.5 with VIM were selected. Next, the genes were grouped based on the pattern and timing of their relationship with VIM. The grouping filters noise by averaging over trends with roughly similar shape and timing. The final output was 22 groups of genes with distinct temporal trends. Finally, gene groups were ordered based on their expression profile in the EMT pseudo-time. The result is a global map of the pseudo-temporal gene dynamics leading to the mesenchymal state (Figure 3.11). The same genome-wide analysis was also repeated with three other canonical markers of the mesenchymal state - CDH2, ITGB4, and CD44, to ensure the reliability of the analysis. The observed gene dynamics were both visually and quantitatively similar for all four markers of EMT progression, confirming that the output analysis are robust.

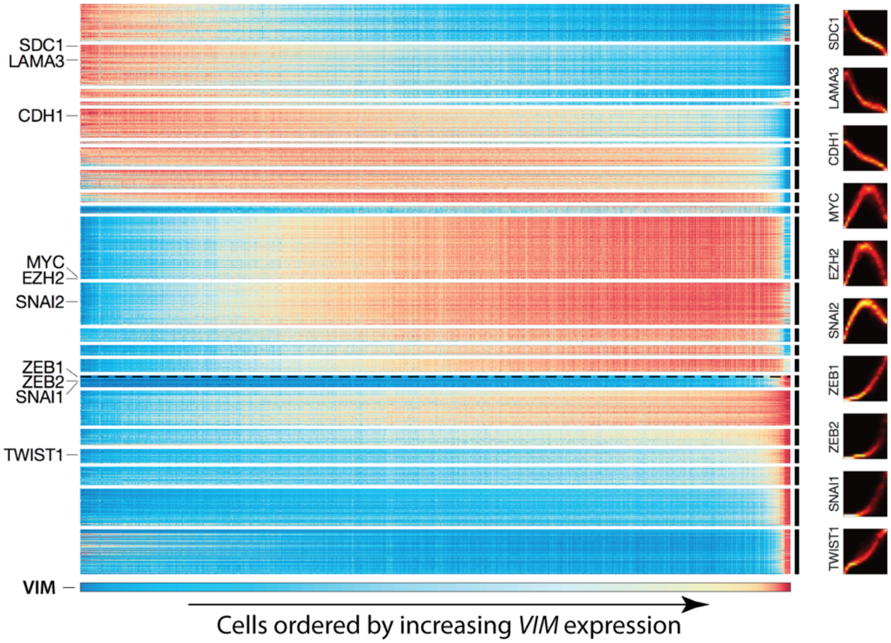


Figure 3.11. *Expression of genes (y axis) ordered by peak expression along VIM (x axis). ZEB1 is highlighted with dashed line. Representative DREMI plots with VIM shown to the right.*

Remarkably, the majority of the genes (2/3 of the genome) participate in EMT. Data shows that clusters of genes change expression in waves as EMT progresses (VIM expression rises). The expression of the first set of genes decreases with EMT progression. Examples of this dynamics are genes SDC1 and LAMA3, which are both involved in cell adhesion. These genes are mostly associated with the epithelial state, and their expression is gradually reduced as cell transition towards the mesenchymal state. Next set of genes shows an initial increase followed by a subsequent decrease in expression before cells enter the mesenchymal state. Examples of such genes include MYC and EZH2. As discussed above, these genes are likely related to the metastable cell states associated with partial EMT. Finally, as cells transition into the mesenchymal state, the expression of a large number of genes monotonically increases to define the mesenchymal state. The prime examples are the canonical EMT-TFs ZEB1, TWIST, and SNAIL.

The pseudo-temporal analysis presented in this section of the thesis was enabled by the asynchronous nature of the EMT progression. Because the cells transition through EMT at different rates sampling of only a few time points (day 8 and day 10) was enough to reveal the full cell state continuum. It is important to acknowledge that this may not be the case for other biological processes and sampling more time points would be needed to perform similar analysis. In the next part of study the EMT gene expression dynamics map was employed to predict activation targets of ZEB1 transcription factor.

3.2.3. Predicting and validating targets of ZEB1 TF

The core EMT transcription factors have some level of redundancy between them as their gene targets and the effects on them overlap. However, the core TFs are not interchangeable, and each plays a particular role in orchestrating EMT under different circumstances [200]. For example, one of the main transcription factors that is important for the establishment of the mesenchymal phenotype the ZEB1 transcription factor. However, even though this TF is a crucial regulator of EMT, its transcriptional targets remain poorly defined. We sought to address this issue by employing the gene expression dynamics map to study the transcriptional targets of ZEB1. The pseudo-time may be used to infer a causal relationship between gene expression. For example, the expression of the activation targets of a particular transcription factor should only peak after the expression of the TF. Following this logic a set of 4509 genes that peak along with or after ZEB1 was identified using the pseudo-time map described above. However, the fact that the expression of a particular gene peaks after the expression of ZEB1 is not enough to confirm their relationship. The expression of the regulator should also be related to the expression of its targets. This interaction can be quantified using the DREMI algorithm in the same fashion as described above. Therefore, out of the initial set of 4509 genes a subset of genes that had DREMI ≥ 1 with ZEB1 was determined. Using this strategy, a total of 1,085 potential target genes were identified that are likely to be either directly or indirectly activated by ZEB1.

To validate this prediction, an engineered HMLE cell line that had ZEB1 under a DOX-inducible promoter was used. Directly overexpressing ZEB1 TF should induce only some of the signaling pathways and transcriptional networks that are active in the TFGβ induced EMT. As a result, cell transformation will be different. However, due to the direct initiation of the ZEB1 transcriptional program, the target genes (both direct and indirect targets) will have a higher gene expression relative to the genes that are not targeted by ZEB1. Therefore, such a system should be suitable for validating the predicted ZEB1 targets.

After two days of continuous ZEB1 overexpression, cells were harvested, and approximately 3500 transcriptomes of individual cells were barcoded with the inDrops platform. Analysis revealed that ZEB1 induction strongly induced EMT as a significant number of the mesenchymal cells was observed (10% of the cells). Given this outcome, it can be expected that the expression of ZEB1 targets will be upregulated relative to other genes. To test for this hypothesis, an ‘impact score’ was used (Equation 3.2). This metric compares the relative ranking of gene expression between the two conditions – ZEB1 and TGF-β induced EMT. Briefly, genes are ranked from highest to lowest (based on mean expression) for each of condition. The impact score is then the average difference between the summed ranks of the two conditions, in N=1000 subsamples of gene set G (predicted ZEB1 activation targets) of fixed size S=200. This subsampling procedure controls for the size of G, as p values will be biased toward 0 given larger. The resulting impact score is the average difference between the summed ranks of the two conditions. A large impact score corresponds to an increase in the relative expression of the predicted targets under ZEB1 induction as compared to TFGβ induced EMT. Next, to compute the significance of the difference in gene expression, the impact score for a gene set of all genes involved in EMT (DREMI with VIM > 0.5) was computed in the same way. The p-value then is the fraction of subsamples that have equal or greater impact score than the predicted gene set G.

Equation 3.2. *Formula for the impact score calculation.* $\boldsymbol{r_z(g)}$ *is the rank of gene g in ZEB1 induction. .* $\boldsymbol{r_t(g)}$ *is the rank of gene g in TGFβ induction.* ***N*** *is the number of subsamples.* ***S*** *is the size of subsample from G set.*

$$impact\ score(G) = \frac{1}{N}\sum_{j=1}^{N}\left(\sum_{i=1}^{S} r_z\left(g_i^j\right) - \sum_{i=1}^{S} r_t\left(g_i^j\right)\right)$$

The predicted ZEB1 targets were confirmed to be upregulated under the ZEB1 overexpression conditions with a significance of $p = 3.1e^{-73}$, against all genes involved in EMT. Next, all genes that peak in the EMT pseudo-time with or after ZEB1 (4509 genes) were analyzed in the same way. These genes are involved in establishing the mesenchymal phenotype as they are expressed after ZEB1 transcription factor. Results show that upregulation of these genes under ZEB1 overexpression can still be observed yet at a substantially lower significance of p=0.004. The observed reduction in p-value reveals that there are additional regulatory networks besides ZEB1 transcriptional program that are at play in establishing the mesenchymal phenotype. Surprisingly, performing this analysis on a gene set where genes are selected only based on their relationship with ZEB1 and not taking in to account their pseudo-time ordering (DREMI with ZEB1 > 1, 1667 genes) leads to an impact score that is not significantly different between the two tested conditions (p=0.13). This outcome can be explained by the fact that ZEB1 is not only an activator but also a potent transcription repressor. Out of all genes related to ZEB1, approximately 1/3 are negatively correlated with it. However, the ZEB1 overexpression in HMLE cells experiment only allows to test for the genes that are directly or indirectly transcriptionally activated by ZEB1.

Unsurprisingly, the top predicted activation targets of ZEB1 transcription factor include many genes that are known to be involved in EMT. For example, SNAI1, ZEB2, BMP (bone morphogenic) antagonist family proteins, and MMP (matrix metalloproteinase) family proteins such as MMP3 can all be found in the list of top hits. Overall, various genes involved in the cell cycle, remodeling of the cell cytoskeleton, extracellular matrix remodeling, and cell migration were identified. This result agrees with the known mechanism of EMT, as discussed in the literature review section 1.3.2. Some of the identified targets are less known for their involvement in EMT. However, the phenotypic annotations of those genes match with known phenotypic changes involved in EMT. For example, RHOA is involved in the reorganization of the actin cytoskeleton and regulates cell shape, attachment, and motility, and CCBE1 is involved in extracellular matrix remodeling and migration. Some of the predicted targets were unexpected. For example, NTN4 is typically involved in neural migration, yet it seems to become active under ZEB1 induction in the context of EMT. In general, the accurate transcription factor activation target prediction demonstrated in this thesis reveals that single-cell transcriptomics can be used to analyze transcriptional networks in an unbiased global manner.

3.2.4.Discussion

Results presented in this work show how scRNA-Seq inDrops platform, in combination with computational biology methods, can be improve our understanding of the complex biological transformations. In this work, single-cell transcriptomics was used to reveal the underlying transcriptional changes that define the EMT process. It is a complex continuous shift of cell state involving extensive transcriptional and epigenetic changes as well as chromatin remodeling. The single-cell resolution of the analysis allowed to determine the intermediate cell states (as defined by archetypes) of the EMT process and reveal the temporal changes in transcriptional programs. Furthermore, the pseudo-time analysis allowed to accurately infer the regulatory relationships in gene expression. Alternatively, gene regulatory interactions can be studied in a high-throughput manner by combining scRNA-Seq with CRISPR technology [388, 389]. However, gene knockouts can disrupt the system in unintended ways that may not be representative of *in vivo* circumstances. Furthermore, such methods require considerable experimental efforts that are not always applicable. The approach described in this study does not require additional experimental manipulations and can be applied to primary tissue and clinical samples. This offers the possibility of discovering changes in gene regulatory pathways in disease (for example, cancer).

The gene expression analysis presented above relied on data imputation. While the development of the imputation algorithm (MAGIC) was outside of the scope of this study, it is important to discuss the impact of imputation and the limitations of the analysis. As discussed in the literature review section 1.2.4, scRNA-Seq data suffers from the significant dropouts. This means that for any given cell, the determined expression value for most genes will be 0. This data property does not significantly hinder many downstream analysis applications - cell type determination (clustering), visualization, and differential expression analysis. Typically, a large number of sampled cells provide enough information for transcriptome-wide analysis. However, dropout becomes a major limitation for gene-gene relationship analysis as most connections are lost due to the sparsity of the data (Figure 3.6, panel A). In the context of this work, the scoring of gene-gene interaction would not be possible without prior imputation. Consequently, the pseudo-time ordering, as well as TF target prediction, also would not be possible. The MAGIC algorithm relies on the diffusion of values between similar cells along an affinity-based graph structure. Such an approach averages over small gene expression differences between cells and will remove intrinsic noise (for example, transcriptional bursting) together with technical noise in the data. This means that fine structure in the data may be lost due to imputation, as intrinsic noise can be a meaningful biological signal. One way to account for this is to use more cells in the analysis. In essence the MAGIC algorithm tries to learn the manifold structure that is observed in the high-dimensional gene expression space (see literature review section 1.2.5). Thus, the more cells are sampled, the more accurately can the structure of the manifold can be reconstructed. However, the precise number of single-cells needed for this analysis cannot be determined *a priori* as the biological signals present in the biological system may not be known. On the other hand, results presented in this work show that less than ten thousand single-cells were enough to characterize the EMT process and determine transcriptional networks of the

ZEB1 TF. Furthermore, with the cost of the scRNA-Seq rapidly decreasing, the cell sample size is likely not to be a major limitation in future studies. In general, after applying the MAGIC algorithm, most cells no longer have zeros in their gene expression data. Instead, they may have very small values that should be interpreted as the probability that a cell is expressing the transcript. It is also important to keep in mind that imputation may introduce bias in the data. For example, algorithms for determining the differentially expressed genes (DEGs) assume sparsity and would likely over-estimate DEGs after imputation. When used appropriately, data diffusion based imputation can be a very useful tool for single-cell transcriptomics data analysis as presented in this work. Similarly, there are cases where imputation is not appropriate, as it would not benefit the analysis (for example, imputation will not benefit DEG analysis and clustering in most cases). Finally, regardless of the analysis goal, the hypothesis made using computational algorithms should be verified by follow up experiments.

In this work, EMT was studied using epithelial breast cancer cells. On the one hand, the use of a model system is convenient as it allowed to perform subsequent perturbations (ZEB1 TF overexpression) to verify the initial hypothesis. On the other hand, because a model system was used, insights produced from this study may not be directly translatable to *in vivo* biology. For example, while TFGβ has been shown to be important for breast tumor development, numerous other signaling pathways are active in the TME that contribute to EMT and overall tumor development. Furthermore, the EMT process is context-dependent and is controlled by many external and internal signaling factors. As a result, not all transcription factors or gene relationships identified in this study will be relevant in all biological contexts. However, the general observations produced by this study are relevant for understanding EMT biology. Perhaps the most important is the confirmation of many intermediate states of the EMT process. It is interesting to speculate what role do these partial states play *in vivo*. For example, partial EMT has been documented to be important in wound-healing (see literature review, section 1.3.5). Overall, the prevalent view is that *in vivo* cells rarely if ever reach the full mesenchymal state. The observations presented in this study support this a view. Even after extensive stimulation, the vast majority of the cells did not reach the mesenchymal states as defined by the transcriptome. Another interesting avenue for studies is the reverse process, MET, which may explain the plasticity during embryogenesis as well as metastasis establishment. Indeed, the MET could not be studied in this work as the cells were under the constant pressure of the EMT process. Nonetheless, the results presented in this work unequivocally prove that single-cell transcriptomics provide a valuable window to intricate and delicate mechanisms of biology. While work above used cell culture to characterize the intracellular biological programs driving cell response to stimuli the final part of this thesis will showcase an example of the single-cell transcriptomics study in clinical settings.

3.3. Immune cells of breast cancer patients

3.3.1. Constructing an immune cell atlas of breast cancer patients

Immunotherapy is a promising new strategy for cancer treatment. Recently, it was successfully employed to treat certain cancer types (melanoma, lung cancer, and kidney cancer), yet has been of limited benefit in treating others. Understanding the diversity and interplay of immune cells in tumors is critical for the successful application and continuous improvement of immunotherapy approaches. Breast cancer, in particular, displays significant heterogeneity in immune cell composition across tumor subtypes and patients. While a lot is now known about separate immune cell types and signaling pathways operating in the breast TME, the full picture is far from clear, making it hard to draw general conclusions. For example, literature meta-analysis has shown that for TNBC, overall high immune cell infiltration of tumor correlates with a favorable prognosis. Yet for tumors expressing hormone receptors (ER+ and/or PR+), immune infiltration does not correlate with patient prognosis. Furthermore, to date, immunotherapy approaches had a limited impact on breast cancer care. A successful immune response to tumor cells consists of several steps. First, tumors cells need to release antigens upon their death, which must then be presented by antigen-presenting cells to prime and activate effector T cells. For immunotherapy to be efficient, all these steps must function adequately, which does not seem to be the case in most breast cancer tumors. Yet the lack of unifiying knowledge about the interplay of immune cells and the tumor prevents us from making rapid progress in the advancment of breast cancer care. Thus, to further the understanding of the immune infiltration of breast cancer tumors, an immune cell atlas of breast cancer patients was constructed in this work. Results presented in this part of the thesis provide a high-resolution view of differences in immune cell populations between different breast cancer patients and reveal a previously unappreciated diversity of immune cell states within the tumor microenvironment.

In this study, samples from 8 treatment-naive patients were analyzed. Different subtypes of breast cancer as defined by canonical markers were included in the study to provide a broader picture. Furthermore, when possible, patient-matched immune cells from healthy breast tissue, peripheral blood, and lymph nodes were also analyzed (Table 3.5). Briefly, fresh surgical samples were enzymatically dissociated to single-cell suspensions and cells of hematopoietic origin were enriched by FACS (sorted by CD45+ staining). Next, the immune cell samples were subjected to scRNA-Seq using the inDrops platform and sequenced on Illumina HiSeq 2500 platform. In total 62024 transcriptomes obtained from single cells were analyzed. To control for technical errors, each sample had a minimum of 2 replicates. Furthermore, samples were sequenced to an average saturation of 91% to recover as much transcript diversity as experimentally possible. It is important to point out, that the inDrops platform allows to barcode the transcriptomes of up to 90% of the input cells. This efficiency enabled the deep sampling of immune cells even in cases where immune infiltration was relatively low. As a result, it was possible to compare Her2+ and ER+ tumors with TNBC samples, which in some cases had as few as 50000 tumor-infiltrating immune cells.

Table 3.5. *Summary of breast cancer patient samples analyzed in this work. For ER and PR markers the numbers in the table reflect the fraction of marker positive tumor cells as determined by established clinical practices.*

Patient	**Tissue**				**Marker**		
	Tumor	*Healthy*	*Blood*	*Lymph node*	*ER*	*PR*	*Her2*
BC1	+	+	+	-	**0.95**	**0.95**	-
BC2	+	+	-	+	**0.9**	0.1	-
BC3	+	+	-	-	0	0	-
BC4	+	-	+	-	**0.95**	**0.95**	-
BC5	+	-	-	-	0.05	0.01	-
BC6	+	-	-	-	**0.99**	0.01	-
BC7	+	-	-	-	0	0	+
BC8	+	-	-	-	0.2	0.05	-

First, each patient was analyzed separately to confirm whether the majority of expected immune cell types could be detected. Single cells from each patient were clustered separately and annotated using genome-wide correlations between cluster mean expression and previously characterized transcriptional profiles of sorted immune cell subset. Major immune cell types were detected in all patients (Figure 3.12, panel A). However, as expected, the distribution of cells varied significantly between patients. For example, the myeloid cell fraction varied in the range of 4%–55%, and T cell fraction varied in the range of 21%–96%. Such a result confirms the intertumoral heterogeneity of immune infiltration of breast cancer tumors. Next, metabolic signals relevant to cancer biology (hypoxia, fatty acid metabolism, glycolysis, and phosphorylation) were analyzed for each patient. For this purpose, the expression of genes associated with each signal (as defined by GSEA databases) was profiled. Interestingly, the expression of groups of genes contributing to signal was significantly different between different patients (Figure 3.12, panel B and C). This observation suggests that signaling in the tumor microenvironment (TME) was different between patients, and that may, at least in part, explain to differences in immune cell subsets.

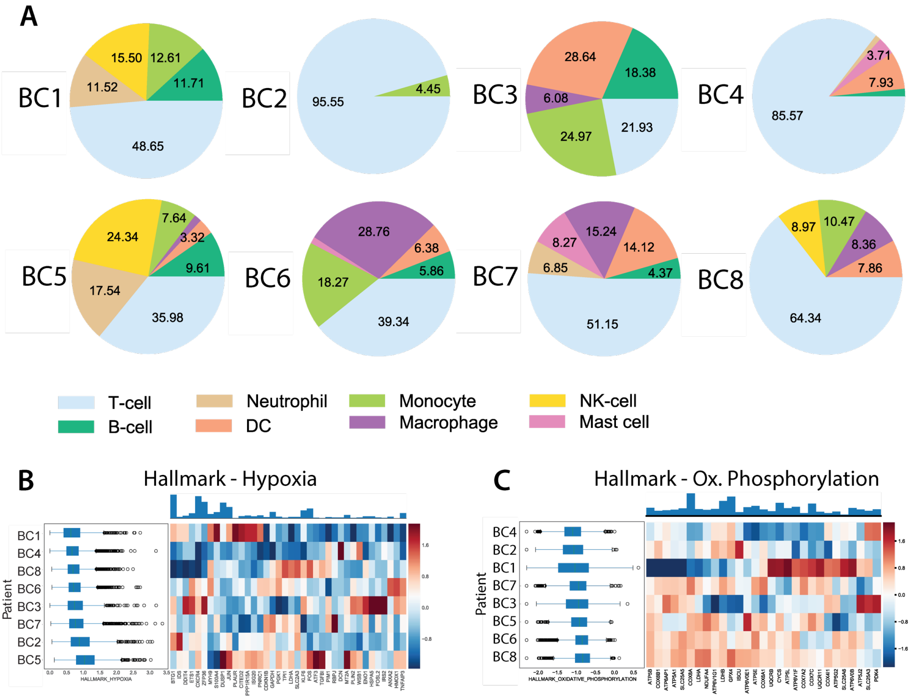


Figure 3.12. *Panel A: Pie charts of cell-type fractions for each patient's tumor-infiltrating immune cells, colored by cell type. Panel B: Left: Boxplots of expression of hallmark hypoxia signature (defined as the mean normalized expression of genes in the signature) across immune cells from each patient. Right: Heatmap of Z-scored mean expression of genes in signature. Top: Barplot of total expression of each gene, across all patients. Panel C: Left: Boxplots of expression of oxidative phosphorylation signature (defined as the mean normalized expression of genes in the signature) across immune cells from each patient. Right: Heatmap of Z-scored mean expression of genes in signature. Top: Barplot of total expression of each gene, across all patients*

Next, data from all cells had to be merged to enable a systematic comparison. However, initially, cells from the same patient appeared to be more similar than cells of the same lineage across patients indicating the batch effects that introduce technical variations due to different sample processing conditions (for example, differences in surgery conditions or sample handling time). Standard data normalization procedures do not account for this as they tend to conflate biological signals and technical differences. To correct for this, the combined data was normalized and imputed using the Biscuit algorithm [390]. Using this algorithm intrinsic biological variation is retained (for example, immune cell activation signal) while correcting for the technical noise in the library.

In total 62024 single-cells collected from all samples were used to construct the global atlas of immune cells in breast cancer patients. After normalization, imputation, and clustering, the initial atlas contained 95 cell clusters. Removing poor quality (low library size) cells from further analysis resulted in 57143 cells that had statistically significant cluster assignments. Next, each cluster was assigned to a known cell type by comparing mean gene expression of the cluster to sorted bulk datasets. In this process some of the clusters were identified as probable carcinoma or stromal cells. While these non-immune types may be of significant interest, they were out of the scope of this work and were excluded from the analysis. Thus, the final atlas contained 47,016 cells spanning four tissues from eight patients and was separated into 83 distinct transcriptionally-similar clusters (Figure 3.13, panel A). Overall, 38 T cell, 27 myeloid lineage cell, 9 B cell, and 9 NK cell clusters were identified.

Cluster annotations were further detailed using the expression of canonical markers. Based on them the T cell clusters were further separated into 15 CD8+ and 21 CD4+ clusters. Alternatively, T cells can be split into 9 naive, 7 central memory, 15 effector memory, and 5 Treg clusters. The myeloid cells were divided into 3 macrophage, 3 mast cell, 4 neutrophil, 3 dendritic cell, 1 plasmacytoid dendritic cell, and 13 monocytic clusters. Finally, 3 CD56– NK cell and 6 CD56+ NK cell clusters (2 of which are likely NK T cells) were identified. The set of well-established markers was enough to define the broad cell types. To rule out misidentification of the cell types it is important to point out that genome-wide profiles of each phenotypic state were used to confirm its identity. The biscuit algorithm identifies cell clusters based on both the mean expression and gene co-expression patterns (covariance patterns). The covariance was very important metric in defining the myriad of T cell clusters. Furthermore, significant differences between most clusters remained even after the mean gene expression is equalized giving confidence that the identified distinct cell types and states are indeed robust and are representative of the *in vivo* diversity. In the constructed atlas, T and myeloid cells represented the most abundant cell subsets. In the context of cancer treatment these two cell subsets are typically considered the most critical and therefore the subsequent analyses presented in this work focused on these two major cell types.

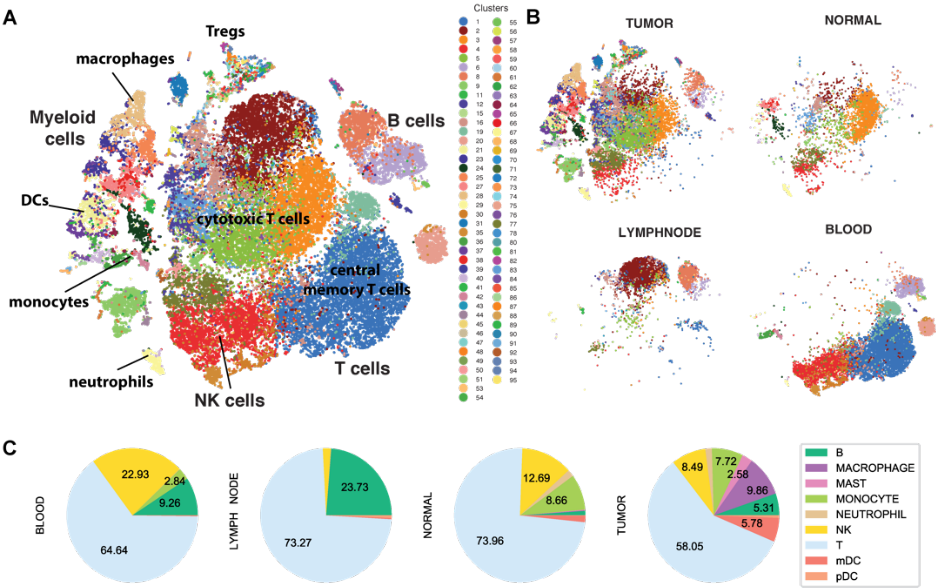


Figure 3.13. *Panel A: Breast immune cell atlas constructed from combining all patient samples (BC1-8) projected with t-SNE. Each dot represents a cell, colored by cluster. Panel B: Subsets of immune cells from Panel A represented on a t-SNE plot. Cells derived from different tissues are plotted on the same coordinates as Panel A to highlight the differences between tissue compartments. Panel C: Proportions of cell types across tissue types.*

3.3.2.Factors that shape the immune cell diversity

The first part of the analysis was focused on investigating the phenotypic overlap between different tissues. Apparent differences between tissue types can be observed in the atlas (Figure 3.13, panel B and C). Several definite conclusions could be drawn. For example, Naive T cells were strongly enriched in the blood (χ^2, p = 3x10^{-80}), while B cells were more prevalent in the lymph node (χ^2, p = 0.0). As can be seen from t-SNE plots, cells state diversity (clusters of the same color) exhibits significant overlap between tumor and healthy tissue (Figure 3.13, panel B). However, a substantial increase in phenotypic heterogeneity and expansion of cell populations in the tumor must also be acknowledged. Importantly, cytotoxic T cell clusters were more abundant in the tumor (χ^2, p = 3x10^{-25}), as were Treg clusters (χ^2, p = 5x10^{-91}) indicting a complex balance between immunosurveillance and immunosuppression. Moreover, while some myeloid clusters were shared between normal and tumor tissue, macrophages were specific to tumor likely indicating presence tumor-associated macrophages (χ^2, p = 0.0). A large number of tissue-resident immune cell states associated with healthy breast tissue (13 myeloid and 19 T cell clusters) were not observed in the blood or the lymph node tissue. Interestingly, the set of clusters found in healthy breast tissue cells represented a subset of those observed in the tumors. Furthermore, 14 myeloid and 17 T cell clusters were only found in the tumor while there were no clusters specific to healthy tissue. Overall, such results underscore the significance of tissue residence as a determinant of immune phenotype.

Table 3.6. *Most significant hallmark GSEA enrichment results on genes with the highest difference in variance in tumor T cells versus normal tissue T cells.*

Pathway	GSEA set size	Enrichment score
Oxidative phosphorylation	196	6.007957
INFγ Response	196	5.730341
Apoptosis	154	5.425722
INFα response	94	4.657201
TGFβ signaling	53	3.071116
Hypoxia	178	2.959150
Il2, Stat5 signaling	192	2.663728
Il6, Jak, Stat3 signaling	81	2.651983

Tumor tissue displays the largest phenotypic diversity and warrants further investigation. The increase in cell-type diversity is related to a significant overall increase in the variance of gene expression as compared to healthy tissue. In particular, specific signaling pathways were found to be activated in the tumor environment (Table 3.6) and coincidentally all of the identified pathways are known to be involved in tumor signaling. Thus, to further confirm the effect of these signaling pathways on immune cell type diversity increase, a specific metric was introduced - ''phenotypic volume''. It uses the covariance in gene expression to measure the relation of the distinct detected phenotypes. For example, if the covariance values between a *geneA* and other genes are very similar to that of another *geneB* and other genes, such that they are dependent, the *geneB* does not add to the volume metric. Assessment of the change in volume showed a significant increase in the phenotypic volume of all major cell types detected in the tumor compared to healthy breast tissue (U test, p = 0). Such results suggest that increased heterogeneity of cell states found within the tumor is indeed related to various signaling pathways active in the TME. Furthermore, an increase in phenotypic volume implies that diverse local niches within a single tumor most likely contribute to the expansions of the cell states.

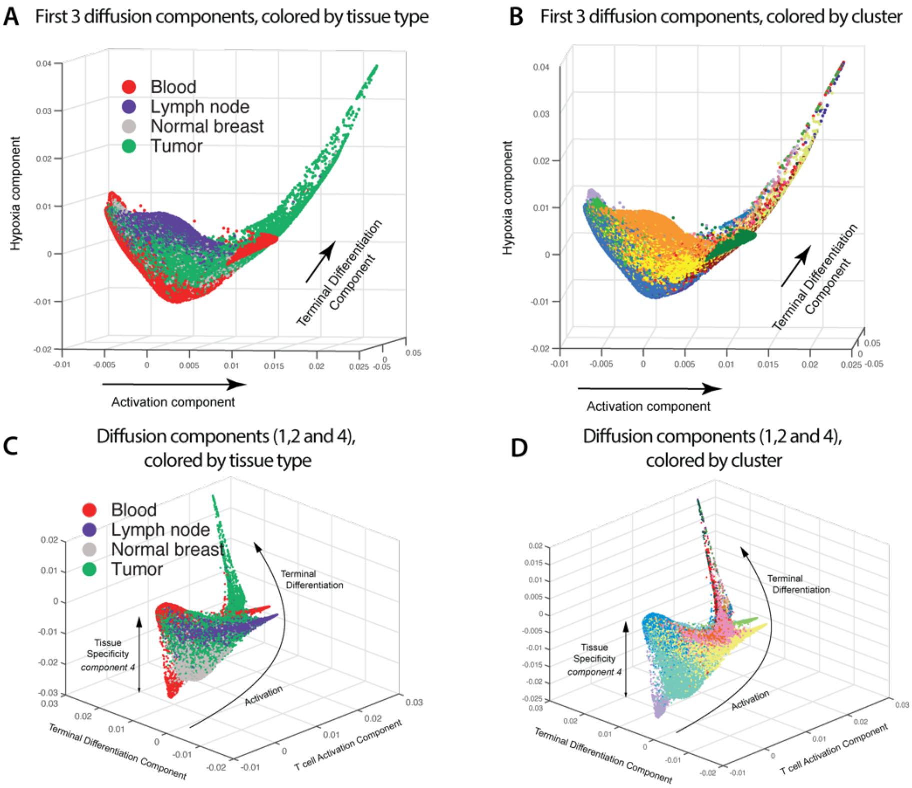


Figure 3.14. *3D plot visualization of all T cells using 3 diffusion components. The main trajectories are indicated with arrows and annotated with the signature most correlated with each component. Each dot represents a cell. Panel A: top 3 diffusion component, colored by cluster Panel B: top 3 diffusion component colored by tissue type. Panel C: 1,2 and 4 diffusion component, colored by cluster Panel D: 1,2 and 4 diffusion component colored by tissue type*

Next, the most significant sources of the observed phenotypic variation were determined. For this purpose, dimensionality reduction algorithm was applied. The first few components of the lower dimensional space reflect the most important sources of variation in the data. It is important to note that due to the non-linear nature of the data, a non-linear dimensionality reduction technique, diffusion maps, was used (see literature review section 1.2.5). The analysis was performed separately for T and myeloid cells as these cell types are intrinsically different. For T cells, the three most informative components as defined by the gene expression signatures correlated to the component where: activation, terminal differentiation, and hypoxia (Figure 3.14, panel A and B). The fourth diffusion component separated cells by their tissue specificity confirming the importance of tissue residence factor (Figure 3.14, panel C and D). The first component of variation (activation) was highly correlated with gene signatures of T cell activation and progressive differentiation, along with IFNγ signaling. Accordingly, tumor T cell populations (in particular, Treg and effector memory T cells) are enriched at the activated end of the component. In contrast, naive T cells from blood tissue can be found at the least activated terminus. It is important to note that while the mean expression levels of clusters gradually vary along the component, there is a wide range of activation states within each cluster. The next most informative component of variation was terminal differentiation. The genes correlated with it include co-stimulatory molecules as well as co-inhibitory receptors (CTLA-4 and TIGIT). Furthermore, genes characteristic of Treg cells (FOXP3, IL2RA, and ENTPD1 [391]) are also included in this component. There is also a moderate degree of overlap in the genes most correlated with the activation and terminal differentiation components, as observed in similar single-cell studies [392]. More importantly, visualizing the T cell activation and terminal differentiation components together revealed a single continuous trajectory (Figure 3.14). This indicates that T cells reside along a broad continuum of activation. Such a result suggests that conventional classification of T cells into relatively few discrete activation or differentiation subtypes may be an oversimplification of the phenotypic complexity present in tumor tissue.

The top components of variation do not fully explain cluster distinctness. Despite the continuum as visualized by the top components, each cluster appeared distinct when accounting for a combination of signatures associated with responses to diverse environmental stimuli (Figure 3.15, panel A). For example, CD4 effector and central memory clusters exhibit variable levels of gene expression involved in different pathways. Such results imply that he local microenvironment in the TME will have varying degrees of inflammation, hypoxia, and nutrient deprivation, thus creating many niches that in turn lead to the many observed phenotypic states.

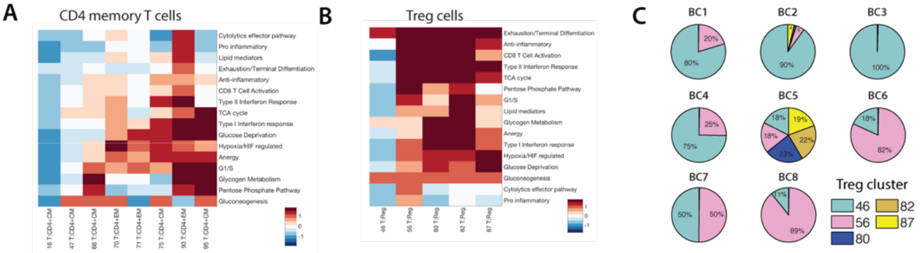


Figure 3.15. *Panel A: Heatmap of mean expression for a curated set of transcriptomic signatures for CD4 memory T cell. Only signatures with high expression in at least one T cell cluster are shown. Expression values are Z scored relative to all T cell clusters. Panel B: Heatmap of mean expression for a curated set of transcriptomic signatures for Treg cell. Only signatures with high expression in at least one T cell cluster are shown. Expression values are Z scored relative to all T cell clusters. Panel C: Proportion of Treg clusters in each patient, indicating that differences in covariance patterns between clusters translate to patients.*

It is important to note that TME signaling alone does not account for the observed cell state diversity in all cases. In particular, the majority of Treg clusters showed similar patterns for anti-inflammatory, exhaustion, hypoxia, and metabolism gene sets (Figure 3.15, panel B). Further analysis revealed that Treg clusters were differentiated by gene covariance. For example, two marker genes can exhibit similar mean expression in 2 different clusters, while the clusters show opposite signs in covariance between these genes. This can occur if the genes are being co-expressed in the same cells in one cluster, but expressed in a mutually exclusive manner in the other cluster. It is important to note that this is an oversimplified example as clustering is done based on the expression and covariance patterns of all detected genes. Interestingly, different proportions of Treg clusters were observed in individual patient samples, and the differences in gene co-expression were also present at the patient level (Figure 3.15, panel C). This result shows that the factors driving the differences in gene co-expression vary between as well as within individual tumors. One possible explanation is the large degree of interaction and cross-regulation displayed by immune cells in the TME. Thus different covariance patterns may be a result of complex signaling environments.

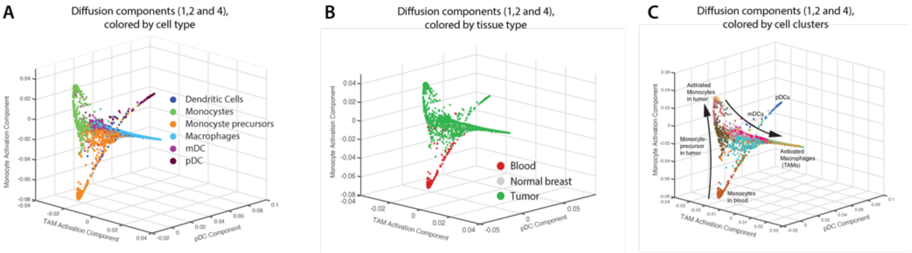


Figure 3.16. *3D plot visualization of all myeloid cells using 3 diffusion components (1, 2 and 4). The main trajectories are indicated with arrows and annotated with the signature most correlated with each component. Each dot represents a cell. Panel A: colored by cell type. Panel B: colored by tissue type (B). Panel C: colored by cluster.*

A similar analysis was also performed on the myeloid cell clusters. The top four diffusion components revealed variation along four major branches (Figure 3.16). Generally, myeloid cells displayed more distinct cell states, as explained by the diffusion components (Figure 3.16, panel A). The first component explains the activation of tumor-associated macrophages. The next two components together capture a more gradual trajectory from blood monocytes to tumor monocytes. Finally, the fourth component distinguishes plasmacytoid DCs (pDC) from the other monocytic cells. As described above, the components were characterized by the associated genes. For example, the first component (TAM activation) is characterized by APOE, CD68, TREM2, and CHIT1. These genes are related to the activation of either recruited or tissue-resident macrophages. Interestingly, the expression of genes associated with “alternatively activated” (M2) macrophages increased together with the genes associated with “classically activated” (M1) macrophages along the first component. All 3 of the TAM clusters had a high expression of the canonical M2 signature and were likewise high in the M1 signature. Furthermore, M1 and M2 gene signatures were positively correlated in the myeloid populations. This is a surprising result, yet it has been described before in the context of melanoma [393]. Thus, the observations presented in this study lend further support the idea that the prevalent alternative polarization model does not fully explain macrophage activation in the TME.

Collectively, analysis of both T cells and myeloid cells reveals that many diverse factors influence immune cell phenotypic states. Firstly, tissue of residence is a major determinant of immune cell state. Immune cells found in the blood and lymph node tissue differ significantly from the ones found in healthy breast and tumor tissues. Furthermore, a significant expansion of cell states is observed within the tumor. Diverse environmental signaling as well as immune cell interactions in the TME contributes to phenotypic cell state expansion. Results reveal that local niches within individual tumors are important for creating the observed diversity.

3.3.3. Discussion

In this part of thesis, an immune cell atlas of breast cancer patients is presented. ScRNA-Seq results revealed a remarkable diversity of immune cells across eight patients and four different tissues. One limitation of the study is that not all patient tumor tissue samples had matched peripheral tissue samples. In particular, only a single patient had a matched lymph node sample, and no patient had samples from all four tissues. Therefore, only limited conclusions can be drawn about the differences in immune cell composition between different tissues. However, it can still be confidently stated that little overlap exists between tissues in terms of cell states. Both the blood and lymph node tissue have distinctly different cell populations that the ones found in healthy and tumor breast tissue. Such observation suggests that biomarkers based on blood immune cells may not necessarily reflect immune cell composition in the tumor. However further studies are needed to explore this question. Interestingly, the immune cells found in the healthy tissue are a subset of cells that can be found in the tumor tissue, and no healthy tissue-specific cells were observed in this study. While caution must be taken due to a limited sample size (only three out of eight patients had a match healthy tissue sample), the results suggest that upon the establishment of a tumor, the immune cells experience a phenotypic expansion that is driven by diverse signaling in the TME.

The breadth of the cell phenotypic diversity in the tumor tissue is perhaps the most striking observation of this study. Defining cells based on conventional cell types appears to be too coarse as almost in all cases at least a few clusters could be assigned to the same cell type. Thus, the term “cell state” has been used throughout this part of the thesis. It is possible that some of the observed states may be transient and could potentially rapidly change upon changes in the TME signaling. However, this study does not have a temporal dimension to provide further insights. Analysis presented in this work has revealed that the diversity of cell states is shaped by multiple different signaling pathways present in the TME. In particular, the presence of local niches in the TME can, in part, account for the observed diversity. Moreover, gene co-expression patterns are also critical in defining the cell states as well as their effector phenotypes. Particularly noteworthy was co-expression of checkpoint receptor genes (CTLA-4, TIGIT, and GITR and other co-receptors) in some Treg subpopulations as compared to mutually exclusive expression of the same genes in other Treg clusters, suggesting that these populations may occupy different functional niches. For example, cells co-expressing CTLA-4 and TIGIT have been shown to inhibit pro-inflammatory Th1 and Th17 responses selectively but not Th2 responses, promoting tissue remodeling [394]. On the other hand, it is possible that in some cases, similar cell states may have similar effector phenotypes. Thus, every hypothesis needs to be considered and tested separately.

The diversity of T cells and myeloid cells was analyzed separately and revealed surprising observations in each case. Firstly, the phenotypic expansion of T cell states was determined to be mainly driven by three separate factors - T cell activation, terminal differentiation, and hypoxic response. The activation component contributes the most to the expansion (it is the first component after dimensionality reduction). Furthermore, cells appear to span the continuum of this component gradually. In particular, cells from the same cluster (same cell state) can be observed throughout the continuum of T cell activation and terminal differentiation signals. Such observation challenges the view of activated T cells rapidly traversing through sparse transitional cell states toward a few predominant, discrete, and stable states (for example, Treg, effector, memory, and exhausted T cells). On the other hand, myeloid cells exhibited a more clearly defined cell state separation as revealed by the top components. This difference between T cells and myeloid cells can be explained in part by the establishment of cell heterogeneity during myeloid cell development [395]. Surprisingly, in TAMs, both M1 and M2 associated genes were frequently expressed in the same cells and positively correlated with one another along the same activation trajectory. Such results challenge the customary model of macrophage polarization wherein M1 and M2 activation states exist as mutually independent discrete states. Similar findings have also been previously reported in lung and kidney cancers [396, 397]. Unexpected observations in both the T cell and the myeloid cell case underscore the importance of single-cell transcriptomics as a tool for unraveling the complex biology of the tumor environment. However, caution must be taken before drawing general conclusions, and findings should be further validated in independent studies.

The immune cell atlas of breast cancer patients presented in this work should also serve as a resource for the breast cancer research community. While the results of this study confirm the known high variation of immune cell subsets between patients, it goes further to detail the considerable phenotypic cell state expansion within each patient. The observed high diversity may, in part, explain why only a limited response to immunotherapy treatments has been observed in breast cancer patients. For example, the considerably different proportions of Treg clusters across patients, and potential differences in their effector phenotypes suggest that multi-dimensional profiling might be necessary to personalize future therapies. It is well established that immune cells in the tumor environment play an essential role in promoting as well as opposing tumor progression. However, the precise balance of these effects remains to be understood. While data presented in this study can shed some light on the complex interactions in the TME, it falls short from providing a full picture. An exciting avenue for future research will be analyzing multiple samples from patients throughout treatment. Furthermore, investigating the interaction of immune cells with tumor cells is another important direction that warrants further exploration.

CONCLUSIONS

- The transcript capture of the *inDrops* scRNA-Seq method was increased approximately 10-times by optimizing the library preparation steps before cDNA amplification.
- Diffusion map based imputation algorithm (MAGIC) recapitulates biologically meaningful gene-gene interactions that otherwise remain obscured due to dropouts.
- It was found that in TGFβ induced EMT process, the transcription factor ZEB1 activates 1085 genes in both direct and indirect manner.
- A high degree of heterogeneity of immune cell subsets infiltrating breast cancer tumors is observed between different patients: myeloid cell fraction varies in the range of 4%–55%, and T cell fraction varies in the range of 21%–96%. This heterogeneity can be, in part, explained by differences in TME signaling.
- The phenotypic diversity of T cell states is significantly expanded in breast tumors as compared to normal breast tissue. The three top components contributing to this phenotypic expansion are T cell activation, terminal differentiation, and hypoxic response.

LIST OF PUBLICATIONS AND PERSONAL CONTRIBUTIONS

Publications included in this thesis

Zilionis R, **Nainys J**, Veres A, Savova V, Zemmour D, Klein AM, Mazutis L. Single-cell barcoding and sequencing using droplet microfluidics. *Nature protocols. 12*(1) (2017) 44.

In this work I have performed part of protocol optimization experiments: optimized RT reaction conditions, optimized cDNA cleanup conditions, optimized IVT reaction conditions. I prepared some of the figures, wrote and edited parts of the manuscript.

Van Dijk D, Sharma R, **Nainys J**, Yim K, Kathail P, Carr AJ, Burdziak C, Moon KR, Chaffer CL, Pattabiraman D, Bierie B, Mazutis L, Wolf G, Krishnaswamy S, Pe'er D. Recovering gene interactions from single-cell data using data diffusion. *Cell. 174*(3) (2018) 716-729.

In this work I have performed cell culture experiments, optimized conditions to achieve epithelial-mesenchymal cell transition, performed single-cell transcriptomics experiments. To a limited degree I also have participated in data analysis and contributed to the first draft of the manuscript.

Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, Prabhakaran S, **Nainys J,** Wu K, Kiseliovas V, Setty M, Choi K, Fromme RM, Dao P, McKenney PT, Wasti RC, Kadaveru K, Mazutis L, Rudensky AY, Pe'er D. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell. 174*(5) (2018) 1293-1308.

In this work I was responsible for scRNA-Seq experiments and data collection that was used to build computational algorithms. I have also performed breast tumor patient sample single-cell barcoding and sequencing optimization and data collection experiments.

Publications not included in this thesis

- Leonavicius K, Nainys J, Kuciauskas D, Mazutis L. Multi-omics at single-cell resolution: comparison of experimental and data fusion approaches. *Current opinion in biotechnology*. *55* (2019) 159-166.
  I collected references related to single-cell -omics, prepared the figures, wrote and edited parts of the manuscript.

- Nainys J, Milkus V, Mažutis L. Single-cell screening using microfluidic systems. In *Microfluidics for Pharmaceutical Applications* (2019) 353-367 William Andrew Publishing.
  I collected references related to single-cell -omics technologies, prepared the figures, wrote and edited parts of the manuscript.

- Kaliniene L, Šimoliūnas E, Truncaitė L, Zajančkauskaitė A, Nainys J, Kaupinis A, Valius M, Meškys R. Molecular analysis of Arthrobacter myovirus vB_ArtM-ArV1: we blame it on the tail. *Journal of virology*, *91*(8) (2017) e00023-17
  I have performed transmission electron microscopy imaging of the purified viral particle samples.

List of conferences and seminars

Oral presentations at international conferences:

- 2017-06. VIB Novel tools for transcriptional profiling, Belgium. Title: “Single-cell barcoding and sequencing using droplet microfluidics”
- 2017-12, Bioateitis, Lithuania. Title: “Studying cancer metastasis models with single-cell transcriptomics”
- 2018-02, The COINS, Lithuania. Title: “Full-stack single-cell transcriptomics”

Poster presentations at international conferences:

- 2016-02, Genomics and Personalized Medicine, Canada. Title: “Single-cell barcoding and sequencing using droplet microfluidics”
- 2017-06, Single Cell Omics, Sweden. Title: “Single-cell barcoding and sequencing using droplet microfluidics”

AKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor prof. dr. Linas Mažutis, who has played the main role in guiding me through my PhD studies and making it a wonderful and enriching experience. I am immensely grateful for all the trust, support, and the possibilities provided to me. I would also like to thank prof. dr. Dana Pe‘er for the opportunity to work in the exhilarating environment on world-class scientific projects.

One thing that I have learned during my PhD studies is that science is a team effort. Thus I would like to thank the lab members in the sector of microtechnology in Vilnius University. In particular I would like to thank Rapolas, Vaidotas, Karolis S. and Karolis G. with whom it was a pleasure to work side-by-side, tackle technical issues and discuss new ideas. I would also like to thank all the members of Mažutis Lab for making the lab feel like a second home. Likewise I am grateful to Pe’er Lab members for making me feel welcome in a new and unknown environment and helping me glimpse the world computational biology. I particular I would like to thank David and Ambrose for the insights and enriching discussions. I am also like grateful to prof. Alma Gedvilaitė and EGIS Lab members for the ABCs of science and the opportunity to learn so many crucial skills that I have put to use in the lab and outside of it.

I am grateful to the members for my defense panel, Kęstutis Sužiedelis, Sonata Jarmalaitė, Urtė Neniškytė, Edita Kriukienė and Artūras Petronis for accepting to evaluate my doctoral thesis. A special thank you to Edita Kriukienė and Artūras Petronis for reviewing my thesis and providing valuable feedback. I would also like to thank VU Life Sciences Center team for helping me build my interpersonal skills and the support I have received on various projects.

I am grateful for my extended support network of both friends and family. It is easy to go forward in life even in difficult times when you have such amazing friends and family. In particular I would like to thank my parents Jurga and Tauras for all the love, the never ending support and the many many happy moments in my life. My deepest gratitude goes to both of you. I would also like to thank my grandmother Rasa for teaching me the value of discussions and always looking out for me. I would like to thank my sister Rasa for being there for me in both happy and difficult moments. Finally, I am grateful to my fiancé Aidai for the love, the joy, the support and the inspiration.

BIBLIOGRAPHY

1. Arendt, D., et al., *The origin and evolution of cell types.* Nature Reviews Genetics, 2016. **17**(12): p. 744.
2. Collins, F.S., M. Morgan, and A. Patrinos, *The Human Genome Project: lessons from large-scale biology.* Science, 2003. **300**(5617): p. 286-290.
3. Velculescu, V.E., et al., *Serial analysis of gene expression.* Science, 1995. **270**(5235): p. 484-487.
4. Tang, F., et al., *mRNA-Seq whole-transcriptome analysis of a single cell.* Nature methods, 2009. **6**(5): p. 377.
5. Klein, A.M., et al., *Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells.* Cell, 2015. **161**(5): p. 1187-1201.
6. Macosko, E.Z., et al., *Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets.* Cell, 2015. **161**(5): p. 1202-1214.
7. Rozenblatt-Rosen, O., et al., *The Human Cell Atlas: from vision to reality.* Nature News, 2017. **550**(7677): p. 451.
8. Regev, A., et al., *Science forum: the human cell atlas.* Elife, 2017. **6**: p. e27041.
9. Kalluri, R. and R.A. Weinberg, *The basics of epithelial-mesenchymal transition.* The Journal of clinical investigation, 2009. **119**(6): p. 1420-1428.
10. Zeisberg, M. and E.G. Neilson, *Biomarkers for epithelial-mesenchymal transitions.* The Journal of clinical investigation, 2009. **119**(6): p. 1429-1437.
11. Dongre, A. and R.A. Weinberg, *New insights into the mechanisms of epithelial–mesenchymal transition and implications for cancer.* Nature reviews Molecular cell biology, 2019. **20**(2): p. 69-84.
12. Harbeck, N., et al., *Breast cancer.* Nature Reviews Disease Primers, 2019. **5**(1): p. 66.
13. Bray, F., et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA: a cancer journal for clinicians, 2018. **68**(6): p. 394-424.
14. Bray, F., et al., *Cancer I ncidence in F ive C ontinents: Inclusion criteria, highlights from Volume X and the global status of cancer registration.* International journal of cancer, 2015. **137**(9): p. 2060-2071.
15. Cardoso, F., et al., *4th ESO–ESMO international consensus guidelines for advanced breast cancer (ABC 4).* Annals of Oncology, 2018. **29**(8): p. 1634-1657.
16. Thorsson, V., et al., *The immune landscape of cancer.* Immunity, 2018. **48**(4): p. 812-830. e14.

17. Gatti-Mays, M.E., et al., *If we build it they will come: targeting the immune response to breast cancer.* NPJ breast cancer, 2019. **5**(1): p. 1-13.
18. Segovia-Mendoza, M. and J. Morales-Montor, *Immune tumor microenvironment in breast cancer and the participation of estrogens and its receptors into cancer physiopathology.* Frontiers in Immunology, 2019. **10**.
19. Gaublomme, J.T., et al., *Single-cell genomics unveils critical regulators of Th17 cell pathogenicity.* Cell, 2015. **163**(6): p. 1400-1412.
20. Gest, H., *The discovery of microorganisms by Robert Hooke and Antoni Van Leeuwenhoek, fellows of the Royal Society.* Notes and records of the Royal Society of London, 2004. **58**(2): p. 187-201.
21. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nature Reviews Genetics, 2009. **10**(1): p. 57-63.
22. *Method of the Year 2013.* Nature Methods, 2014. **11**(1): p. 1-1.
23. Shapiro, E., T. Biezuner, and S. Linnarsson, *Single-cell sequencing-based technologies will revolutionize whole-organism science.* Nature Reviews Genetics, 2013. **14**(9): p. 618-630.
24. Svensson, V., R. Vento-Tormo, and S.A. Teichmann, *Exponential scaling of single-cell RNA-seq in the past decade.* Nature protocols, 2018. **13**(4): p. 599-604.
25. Ramsköld, D., et al., *Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells.* Nature biotechnology, 2012. **30**(8): p. 777.
26. Tang, F., et al., *Deterministic and stochastic allele specific gene expression in single mouse blastomeres.* PloS one, 2011. **6**(6).
27. Tang, F., et al., *Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis.* Cell stem cell, 2010. **6**(5): p. 468-478.
28. Islam, S., et al., *Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq.* Genome research, 2011. **21**(7): p. 1160-1167.
29. Ziegenhain, C., et al., *Comparative analysis of single-cell RNA sequencing methods.* Molecular cell, 2017. **65**(4): p. 631-643. e4.
30. Svensson, V., et al., *Power analysis of single-cell RNA-sequencing experiments.* Nature methods, 2017. **14**(4): p. 381.
31. Ziegenhain, C., et al., *Quantitative single-cell transcriptomics.* Briefings in functional genomics, 2018. **17**(4): p. 220-232.
32. Human Cell Atlas: HCA. *HCA Members.* [cited 2020 March 24]; Available from: www.humancellatlas.org/join-hca.
33. Brennecke, P., et al., *Accounting for technical noise in single-cell RNA-seq experiments.* Nature methods, 2013. **10**(11): p. 1093.

34. Jaitin, D.A., et al., *Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types.* Science, 2014. **343**(6172): p. 776-779.

35. Zilionis, R., et al., *Single-cell barcoding and sequencing using droplet microfluidics.* Nature protocols, 2017. **12**(1): p. 44.

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

37. Cao, J., et al., *Comprehensive single-cell transcriptional profiling of a multicellular organism.* Science, 2017. **357**(6352): p. 661-667.

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

39. Hayashi, T., et al., *Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs.* Nature communications, 2018. **9**(1): p. 1-16.

40. Avital, G., et al., *scDual-Seq: mapping the gene regulatory program of Salmonella infection by host and pathogen single-cell RNA-sequencing.* Genome biology, 2017. **18**(1): p. 200.

41. Faridani, O.R., et al., *Single-cell sequencing of the small-RNA transcriptome.* Nature biotechnology, 2016. **34**(12): p. 1264.

42. Boon, W.C., et al., *Increasing cDNA yields from single-cell quantities of mRNA in standard laboratory reverse transcriptase reactions using acoustic microstreaming.* JoVE (Journal of Visualized Experiments), 2011(53): p. e3144.

43. Zhu, Y., et al., *Reverse transcriptase template switching: A SMART™ approach for full-length cDNA library construction.* Biotechniques, 2001. **30**(4): p. 892-897.

44. Sasagawa, Y., et al., *Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads.* Genome biology, 2018. **19**(1): p. 29.

45. Hashimshony, T., et al., *CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification.* Cell reports, 2012. **2**(3): p. 666-673.

46. Baugh, L., et al., *Quantitative analysis of mRNA amplification by in vitro transcription.* Nucleic acids research, 2001. **29**(5): p. e29-e29.

47. Parekh, S., et al., *The impact of amplification on differential expression analyses by RNA-seq.* Scientific reports, 2016. **6**: p. 25533.

48. Lafzi, A., et al., *Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies.* Nature protocols, 2018. **13**(12): p. 2742-2757.

49. Islam, S., et al., *Quantitative single-cell RNA-seq with unique molecular identifiers.* Nature methods, 2014. **11**(2): p. 163.

50. Kivioja, T., et al., *Counting absolute numbers of molecules using unique molecular identifiers.* Nature methods, 2012. **9**(1): p. 72-74.

51. Karlsson, K. and S. Linnarsson, *Single-cell mRNA isoform diversity in the mouse brain.* BMC genomics, 2017. **18**(1): p. 126.

52. Hagemann-Jensen, M., et al., *Single-cell RNA counting at allele-and isoform-resolution using Smart-seq3.* bioRxiv, 2019: p. 817924.

53. Wu, H., et al., *Advantages of single-nucleus over single-cell RNA sequencing of adult kidney: rare cell types and novel cell states revealed in fibrosis.* Journal of the American Society of Nephrology, 2019. **30**(1): p. 23-32.

54. van den Brink, S.C., et al., *Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations.* Nature methods, 2017. **14**(10): p. 935.

55. Poulin, J.-F., et al., *Disentangling neural cell diversity using single-cell transcriptomics.* Nature neuroscience, 2016. **19**(9): p. 1131.

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

57. Lacar, B., et al., *Nuclear RNA-seq of single neurons reveals molecular signatures of activation.* Nature communications, 2016. **7**(1): p. 1-13.

58. Grindberg, R.V., et al., *RNA-sequencing from single nuclei.* Proceedings of the National Academy of Sciences, 2013. **110**(49): p. 19802-19807.

59. Bakken, T.E., et al., *Single-nucleus and single-cell transcriptomes compared in matched cortical cell types.* PloS one, 2018. **13**(12).

60. Keren-Shaul, H., et al., *MARS-seq2. 0: an experimental and analytical pipeline for indexed sorting combined with single-cell RNA sequencing.* Nature protocols, 2019. **14**(6): p. 1841.

61. Bagnoli, J.W., et al., *Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq.* Nature communications, 2018. **9**(1): p. 1-8.

62. Hochgerner, H., et al., *STRT-seq-2i: dual-index 5′ single cell and nucleus RNA-seq on an addressable microwell array.* Scientific reports, 2017. **7**(1): p. 1-8.

63. Richardson, G.M., J. Lannigan, and I.G. Macara, *Does FACS perturb gene expression?* Cytometry Part A, 2015. **87**(2): p. 166-175.

64. Prakadan, S.M., A.K. Shalek, and D.A. Weitz, *Scaling by shrinking: empowering single-cell'omics' with microfluidic devices.* Nature Reviews Genetics, 2017. **18**(6): p. 345.

65. Zeisel, A., et al., *Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq.* Science, 2015. **347**(6226): p. 1138-1142.

66. 10X Genomics. *Publications*. [cited 2020 March 24th]; Available from: https://www.10xgenomics.com/resources/publications/.

67. Hughes, T.K., et al., *Highly efficient, massively-parallel single-cell RNA-seq reveals cellular states and molecular features of human skin pathology.* bioRxiv, 2019: p. 689273.

68. Bose, S., et al., *Scalable microfluidics for single-cell RNA printing and sequencing.* Genome biology, 2015. **16**(1): p. 120.

69. Han, X., et al., *Mapping the mouse cell atlas by microwell-seq.* Cell, 2018. **172**(5): p. 1091-1107. e17.

70. Proserpio, V., *Single Cell Methods: Sequencing and Proteomics.* 2019: Springer.

71. Zappia, L., B. Phipson, and A. Oshlack, *Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database.* PLoS computational biology, 2018. **14**(6): p. e1006245.

72. Luecken, M.D. and F.J. Theis, *Current best practices in single-cell RNA-seq analysis: a tutorial.* Molecular systems biology, 2019. **15**(6).

73. McCarthy, D.J., et al., *Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R.* Bioinformatics, 2017. **33**(8): p. 1179-1186.

74. Butler, A., et al., *Integrating single-cell transcriptomic data across different conditions, technologies, and species.* Nature biotechnology, 2018. **36**(5): p. 411-420.

75. Wolf, F.A., P. Angerer, and F.J. Theis, *SCANPY: large-scale single-cell gene expression data analysis.* Genome biology, 2018. **19**(1): p. 15.

76. Zhu, X., et al., *Granatum: a graphical single-cell RNA-Seq analysis pipeline for genomics scientists.* Genome medicine, 2017. **9**(1): p. 108.

77. Gardeux, V., et al., *ASAP: a web-based platform for the analysis and interactive visualization of single-cell RNA-seq data.* Bioinformatics, 2017. **33**(19): p. 3123-3125.

78. Scholz, C.J., et al., *FASTGenomics: An analytical ecosystem for single-cell RNA sequencing data.* bioRxiv, 2018: p. 272476.

79. Regev, A., et al., *The human cell atlas white paper.* arXiv preprint arXiv:1810.05192, 2018.

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

81. Kiselev, V.Y., T.S. Andrews, and M. Hemberg, *Challenges in unsupervised clustering of single-cell RNA-seq data.* Nature Reviews Genetics, 2019. **20**(5): p. 273-282.

82. Hou, W., et al., *A Systematic Evaluation of Single-cell RNA-sequencing Imputation Methods.* bioRxiv, 2020.

83. Webb, S., *Deep learning for biology.* Nature, 2018. **554**(7693).

84. Zheng, J. and K. Wang, *Emerging deep learning methods for single-cell RNA-seq data analysis.* Quantitative Biology, 2019. **7**(4): p. 247-254.

85. Bioinformatics, B., *FastQC: a quality control tool for high throughput sequence data.* Cambridge, UK: Babraham Institute, 2011.

86. Azizi, E., et al., *Single-cell map of diverse immune phenotypes in the breast tumor microenvironment.* Cell, 2018. **174**(5): p. 1293-1308. e36.

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

88. Conesa, A., et al., *A survey of best practices for RNA-seq data analysis.* Genome biology, 2016. **17**(1): p. 13.

89. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-1111.

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

91. Bray, N.L., et al., *Near-optimal probabilistic RNA-seq quantification.* Nature biotechnology, 2016. **34**(5): p. 525-527.

92. Patro, R., et al., *Salmon provides fast and bias-aware quantification of transcript expression.* Nature methods, 2017. **14**(4): p. 417.

93. Hwang, B., J.H. Lee, and D. Bang, *Single-cell RNA sequencing technologies and bioinformatics pipelines.* Experimental & molecular medicine, 2018. **50**(8): p. 1-14.

94. La Manno, G., et al., *RNA velocity of single cells.* Nature, 2018. **560**(7719): p. 494-498.

95. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC bioinformatics, 2011. **12**(1): p. 323.

96. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nature biotechnology, 2010. **28**(5): p. 511.

97. Anders, S., P.T. Pyl, and W. Huber, *HTSeq—a Python framework to work with high-throughput sequencing data.* Bioinformatics, 2015. **31**(2): p. 166-169.

98. Smith, T., A. Heger, and I. Sudbery, *UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy.* Genome research, 2017. **27**(3): p. 491-499.

99. Zheng, G.X., et al., *Massively parallel digital transcriptional profiling of single cells.* Nature communications, 2017. **8**(1): p. 1-12.

100. Ilicic, T., et al., *Classification of low quality cells from single-cell RNA-seq data.* Genome biology, 2016. **17**(1): p. 29.

101. Griffiths, J.A., A. Scialdone, and J.C. Marioni, *Using single-cell genomics to understand developmental processes and cell fate decisions.* Molecular systems biology, 2018. **14**(4).

102. DePasquale, E.A., et al., *DoubletDecon: Deconvoluting Doublets from Single-Cell RNA-Sequencing Data.* Cell reports, 2019. **29**(6): p. 1718-1727. e8.

103. McGinnis, C.S., L.M. Murrow, and Z.J. Gartner, *DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors.* Cell systems, 2019. **8**(4): p. 329-337. e4.

104. Wolock, S.L., R. Lopez, and A.M. Klein, *Scrublet: computational identification of cell doublets in single-cell transcriptomic data.* Cell systems, 2019. **8**(4): p. 281-291. e9.

105. Heaton, H., et al., *souporcell: Robust clustering of single cell RNAseq by genotype and ambient RNA inference without reference genotypes.* BioRxiv, 2019: p. 699637.

106. Lun, A.T., et al., *EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data.* Genome biology, 2019. **20**(1): p. 63.

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

108. Plasschaert, L.W., et al., *A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte.* Nature, 2018. **560**(7718): p. 377-381.

109. Vieth, B., et al., *A systematic evaluation of single cell RNA-seq analysis pipelines.* Nature communications, 2019. **10**(1): p. 1-11.

110. Stegle, O., S.A. Teichmann, and J.C. Marioni, *Computational and analytical challenges in single-cell transcriptomics.* Nature Reviews Genetics, 2015. **16**(3): p. 133-145.

111. Lun, A.T., K. Bach, and J.C. Marioni, *Pooling across cells to normalize single-cell RNA sequencing data with many zero counts.* Genome biology, 2016. **17**(1): p. 75.

112. Evans, C., J. Hardin, and D.M. Stoebel, *Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions.* Briefings in bioinformatics, 2018. **19**(5): p. 776-792.

113. Marinov, G.K., et al., *From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing.* Genome research, 2014. **24**(3): p. 496-510.

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

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

116. Vallejos, C.A., et al., *Normalizing single-cell RNA sequencing data: challenges and opportunities.* Nature methods, 2017. **14**(6): p. 565.

117. Cole, M.B., et al., *Performance assessment and selection of normalization procedures for Single-Cell RNA-Seq.* Cell systems, 2019. **8**(4): p. 315-328. e8.

118. Mayer, C., et al., *Developmental diversification of cortical inhibitory interneurons.* Nature, 2018. **555**(7697): p. 457-462.

119. Vieth, B., et al., *powsimR: power analysis for bulk and single cell RNA-seq experiments.* Bioinformatics, 2017. **33**(21): p. 3486-3488.

120. Soneson, C. and M.D. Robinson, *Bias, robustness and scalability in single-cell differential expression analysis.* Nature methods, 2018. **15**(4): p. 255.

121. Patel, A.P., et al., *Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma.* Science, 2014. **344**(6190): p. 1396-1401.

122. Finak, G., et al., *MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data.* Genome biology, 2015. **16**(1): p. 278.

123. Büttner, M., et al., *A test metric for assessing single-cell RNA-seq batch correction.* Nature methods, 2019. **16**(1): p. 43-49.

124. Stoeckius, M., et al., *Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics.* Genome biology, 2018. **19**(1): p. 1-12.

125. McGinnis, C.S., et al., *MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices.* Nature methods, 2019. **16**(7): p. 619.

126. Shin, D., et al., *Multiplexed single-cell RNA-seq via transient barcoding for simultaneous expression profiling of various drug perturbations.* Science advances, 2019. **5**(5): p. eaav2249.

127. Johnson, W.E., C. Li, and A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical Bayes methods.* Biostatistics, 2007. **8**(1): p. 118-127.

128. Haghverdi, L., et al., *Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors.* Nature biotechnology, 2018. **36**(5): p. 421-427.

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

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

131. Visscher, P.M., et al., *10 years of GWAS discovery: biology, function, and translation.* The American Journal of Human Genetics, 2017. **101**(1): p. 5-22.

132. Chou, W.-C., et al., *A combined reference panel from the 1000 Genomes and UK10K projects improved rare variant imputation in European and Chinese samples.* Scientific reports, 2016. **6**: p. 39313.
133. Laehnemann, D., et al., *12 Grand challenges in single-cell data science*. 2019, PeerJ Preprints.
134. Huang, M., et al., *SAVER: gene expression recovery for single-cell RNA sequencing.* Nature methods, 2018. **15**(7): p. 539-542.
135. Li, W.V. and J.J. Li, *An accurate and robust imputation method scImpute for single-cell RNA-seq data.* Nature communications, 2018. **9**(1): p. 1-9.
136. Tang, W., et al., *bayNorm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data.* Bioinformatics, 2020. **36**(4): p. 1174-1181.
137. Chen, M. and X. Zhou, *VIPER: variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies.* Genome biology, 2018. **19**(1): p. 1-15.
138. Van Dijk, D., et al., *Recovering gene interactions from single-cell data using data diffusion.* Cell, 2018. **174**(3): p. 716-729. e27.
139. Wagner, F., Y. Yan, and I. Yanai, *K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data.* BioRxiv, 2017: p. 217737.
140. Ronen, J. and A. Akalin, *netSmooth: Network-smoothing based imputation for single cell RNA-seq.* F1000Research, 2018. **7**.
141. Gong, W., et al., *DrImpute: imputing dropout events in single cell RNA sequencing data.* BMC bioinformatics, 2018. **19**(1): p. 220.
142. Linderman, G.C., J. Zhao, and Y. Kluger, *Zero-preserving imputation of scRNA-seq data using low-rank approximation.* bioRxiv, 2018: p. 397588.
143. Mongia, A., D. Sengupta, and A. Majumdar, *McImpute: Matrix completion based imputation for single cell RNA-seq data.* Frontiers in genetics, 2019. **10**: p. 9.
144. Zhang, L. and S. Zhang, *PBLR: an accurate single cell RNA-seq data imputation tool considering cell heterogeneity and prior expression level of dropouts.* bioRxiv, 2018: p. 379883.
145. Bourlard, H. and Y. Kamp, *Auto-association by multilayer perceptrons and singular value decomposition.* Biological cybernetics, 1988. **59**(4-5): p. 291-294.
146. Talwar, D., et al., *AutoImpute: Autoencoder based imputation of single-cell RNA-seq data.* Scientific reports, 2018. **8**(1): p. 1-11.
147. Eraslan, G., et al., *Single-cell RNA-seq denoising using a deep count autoencoder.* Nature communications, 2019. **10**(1): p. 1-14.
148. Arisdakessian, C., et al., *DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data.* Genome biology, 2019. **20**(1): p. 1-14.

149. Lopez, R., et al., *Deep generative modeling for single-cell transcriptomics.* Nature methods, 2018. **15**(12): p. 1053-1058.

150. Leote, A.C., X. Wu, and A. Beyer, *Network-based imputation of dropouts in single-cell RNA sequencing data.* bioRxiv, 2019: p. 611517.

151. Wang, J., et al., *Data denoising with transfer learning in single-cell transcriptomics.* Nature methods, 2019. **16**(9): p. 875-878.

152. Peng, T., et al., *SCRABBLE: single-cell RNA-seq imputation constrained by bulk RNA-seq data.* Genome biology, 2019. **20**(1): p. 88.

153. Andrews, T.S. and M. Hemberg, *False signals induced by single-cell imputation.* F1000Research, 2018. **7**.

154. Zhang, L. and S. Zhang, *Comparison of computational methods for imputing single-cell RNA-sequencing data.* IEEE/ACM transactions on computational biology and bioinformatics, 2018.

155. Pertea, M., et al., *CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise.* Genome biology, 2018. **19**(1): p. 1-14.

156. Moon, K.R., et al., *Manifold learning-based methods for analyzing single-cell RNA-sequencing data.* Current Opinion in Systems Biology, 2018. **7**: p. 36-46.

157. Heimberg, G., et al., *Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing.* Cell systems, 2016. **2**(4): p. 239-250.

158. Yip, S.H., P.C. Sham, and J. Wang, *Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data.* Briefings in bioinformatics, 2019. **20**(4): p. 1583-1589.

159. Pearson, K., *LIII. On lines and planes of closest fit to systems of points in space.* The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 1901. **2**(11): p. 559-572.

160. Chung, N.C. and J.D. Storey, *Statistical significance of variables driving systematic variation in high-dimensional data.* Bioinformatics, 2015. **31**(4): p. 545-554.

161. Maaten, L.v.d. and G. Hinton, *Visualizing data using t-SNE.* Journal of machine learning research, 2008. **9**(Nov): p. 2579-2605.

162. Wattenberg, M., F. Viégas, and I. Johnson, *How to use t-SNE effectively.* Distill, 2016. **1**(10): p. e2.

163. Moon, K.R., et al., *Visualizing structure and transitions in high-dimensional biological data.* Nature Biotechnology, 2019. **37**(12): p. 1482-1492.

164. Becht, E., et al., *Dimensionality reduction for visualizing single-cell data using UMAP.* Nature biotechnology, 2019. **37**(1): p. 38.

165. Weinreb, C., S. Wolock, and A.M. Klein, *SPRING: a kinetic interface for visualizing high dimensional single-cell expression data.* Bioinformatics, 2018. **34**(7): p. 1246-1248.

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

167. Coifman, R.R., et al., *Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps.* Proceedings of the national academy of sciences, 2005. **102**(21): p. 7426-7431.

168. Haghverdi, L., F. Buettner, and F.J. Theis, *Diffusion maps for high-dimensional single-cell analysis of differentiation data.* Bioinformatics, 2015. **31**(18): p. 2989-2998.

169. Haghverdi, L., et al., *Diffusion pseudotime robustly reconstructs lineage branching.* Nature methods, 2016. **13**(10): p. 845.

170. Ding, J., A. Condon, and S.P. Shah, *Interpretable dimensionality reduction of single cell transcriptome data with deep generative models.* Nature communications, 2018. **9**(1): p. 1-13.

171. Wang, D. and J. Gu, *VASC: dimension reduction and visualization of single-cell RNA-seq data by deep variational autoencoder.* Genomics, proteomics & bioinformatics, 2018. **16**(5): p. 320-331.

172. Pierson, E. and C. Yau, *ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis.* Genome biology, 2015. **16**(1): p. 241.

173. Hu, Q. and C.S. Greene. *Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics*. in *PSB*. 2019. World Scientific.

174. Çakır, B., et al., *Comparison of visualisation tools for single-cell RNAseq data.* BioRxiv, 2020.

175. Lloyd, S., *Least squares quantization in PCM.* IEEE transactions on information theory, 1982. **28**(2): p. 129-137.

176. Grün, D., et al., *Single-cell messenger RNA sequencing reveals rare intestinal cell types.* Nature, 2015. **525**(7568): p. 251-255.

177. Wang, B., et al., *Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning.* Nature Methods, 2017. **14**(4): p. 414-416.

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

179. Levine, J.H., et al., *Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis.* Cell, 2015. **162**(1): p. 184-197.

180. Duò, A., M.D. Robinson, and C. Soneson, *A systematic performance evaluation of clustering methods for single-cell RNA-seq data.* F1000Research, 2018. **7**.

181. Freytag, S., et al., *Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data.* F1000Research, 2018. **7**.

182. Tanay, A. and A. Regev, *Scaling single-cell genomics from phenomenology to mechanism.* Nature, 2017. **541**(7637): p. 331-338.

183. Bendall, S.C., et al., *Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development.* Cell, 2014. **157**(3): p. 714-725.

184. Setty, M., et al., *Wishbone identifies bifurcating developmental trajectories from single-cell data.* Nature biotechnology, 2016. **34**(6): p. 637.

185. Qiu, X., et al., *Reversed graph embedding resolves complex single-cell trajectories.* Nature methods, 2017. **14**(10): p. 979.

186. Street, K., et al., *Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics.* BMC genomics, 2018. **19**(1): p. 477.

187. Scholtens, D. and A. Von Heydebreck, *Analysis of differential gene expression studies*, in *Bioinformatics and computational biology solutions using R and Bioconductor*. 2005, Springer. p. 229-248.

188. Bard, J., S.Y. Rhee, and M. Ashburner, *An ontology for cell types.* Genome biology, 2005. **6**(2): p. R21.

189. Ashburner, M., et al., *Gene ontology: tool for the unification of biology.* Nature genetics, 2000. **25**(1): p. 25-29.

190. Bakken, T., et al., *Cell type discovery and representation in the era of high-content single cell phenotyping.* BMC bioinformatics, 2017. **18**(17): p. 559.

191. Aevermann, B.D., et al., *Cell type discovery using single-cell transcriptomics: implications for ontological representation.* Human molecular genetics, 2018. **27**(R1): p. R40-R47.

192. Kiselev, V.Y., A. Yiu, and M. Hemberg, *scmap: projection of single-cell RNA-seq data across data sets.* Nature methods, 2018. **15**(5): p. 359.

193. Pliner, H.A., J. Shendure, and C. Trapnell, *Supervised classification enables rapid annotation of cell atlases.* Nature methods, 2019. **16**(10): p. 983-986.

194. Nieto, M.A., et al., *Emt: 2016.* Cell, 2016. **166**(1): p. 21-45.

195. Greenburg, G. and E.D. Hay, *Epithelia suspended in collagen gels can lose polarity and express characteristics of migrating mesenchymal cells.* The Journal of cell biology, 1982. **95**(1): p. 333-339.

196. Nieto, M.A., et al., *Control of cell behavior during vertebrate development by Slug, a zinc finger gene.* Science, 1994. **264**(5160): p. 835-839.

197. Hay, E.D., *An overview of epithelio-mesenchymal transformation.* Cells Tissues Organs, 1995. **154**(1): p. 8-20.

198. Chaffer, C.L., E.W. Thompson, and E.D. Williams, *Mesenchymal to epithelial transition in development and disease.* Cells Tissues Organs, 2007. **185**(1-3): p. 7-19.

199. Lamouille, S., J. Xu, and R. Derynck, *Molecular mechanisms of epithelial–mesenchymal transition.* Nature reviews Molecular cell biology, 2014. **15**(3): p. 178.

200. Stemmler, M.P., et al., *Non-redundant functions of EMT transcription factors.* Nature cell biology, 2019. **21**(1): p. 102-112.

201. Thiery, J.P. and J.P. Sleeman, *Complex networks orchestrate epithelial–mesenchymal transitions.* Nature reviews Molecular cell biology, 2006. **7**(2): p. 131-142.

202. Brabletz, T., et al., *EMT in cancer.* Nature Reviews Cancer, 2018. **18**(2): p. 128.

203. Shibue, T. and R.A. Weinberg, *EMT, CSCs, and drug resistance: the mechanistic link and clinical implications.* Nature reviews Clinical oncology, 2017. **14**(10): p. 611.

204. Huang, R.Y.-J., P. Guilford, and J.P. Thiery, *Early events in cell adhesion and polarity during epithelial-mesenchymal transition.* 2012, The Company of Biologists Ltd.

205. Peinado, H., D. Olmeda, and A. Cano, *Snail, Zeb and bHLH factors in tumour progression: an alliance against the epithelial phenotype?* Nature reviews cancer, 2007. **7**(6): p. 415-428.

206. Moreno-Bueno, G., F. Portillo, and A. Cano, *Transcriptional regulation of cell polarity in EMT and cancer.* Oncogene, 2008. **27**(55): p. 6958-6969.

207. Wheelock, M.J., et al., *Cadherin switching.* Journal of cell science, 2008. **121**(6): p. 727-735.

208. Theveneau, E. and R. Mayor, *Cadherins in collective cell migration of mesenchymal cells.* Current opinion in cell biology, 2012. **24**(5): p. 677-684.

209. Mendez, M.G., S.-I. Kojima, and R.D. Goldman, *Vimentin induces changes in cell shape, motility, and adhesion during the epithelial to mesenchymal transition.* The FASEB Journal, 2010. **24**(6): p. 1838-1851.

210. Yilmaz, M. and G. Christofori, *EMT, the cytoskeleton, and cancer cell invasion.* Cancer and Metastasis Reviews, 2009. **28**(1-2): p. 15-33.

211. Yang, X., et al., *Regulation of β4-integrin expression by epigenetic modifications in the mammary gland and during the epithelial-to-mesenchymal transition.* Journal of cell science, 2009. **122**(14): p. 2473-2480.

212. Kim, Y., et al., *Integrin α3β1–dependent β-catenin phosphorylation links epithelial Smad signaling to cell contacts.* Journal of Cell Biology, 2009. **184**(2): p. 309-322.

213. Nisticò, P., M.J. Bissell, and D.C. Radisky, *Epithelial-mesenchymal transition: general principles and pathological relevance with special emphasis on the role of matrix metalloproteinases.* Cold Spring Harbor perspectives in biology, 2012. **4**(2): p. a011908.
214. Bernstein, B.E., et al., *A bivalent chromatin structure marks key developmental genes in embryonic stem cells.* Cell, 2006. **125**(2): p. 315-326.
215. Xi, Q., et al., *A poised chromatin platform for TGF-β access to master regulators.* Cell, 2011. **147**(7): p. 1511-1524.
216. Jordà, M., et al., *Upregulation of MMP-9 in MDCK epithelial cell line in response to expression of the Snail transcription factor.* Journal of cell science, 2005. **118**(15): p. 3371-3385.
217. Wang, S.-P., et al., *p53 controls cancer cell invasion by inducing the MDM2-mediated degradation of Slug.* Nature cell biology, 2009. **11**(6): p. 694-704.
218. Lamouille, S., et al., *Regulation of epithelial–mesenchymal and mesenchymal–epithelial transitions by microRNAs.* Current opinion in cell biology, 2013. **25**(2): p. 200-207.
219. Xu, J., S. Lamouille, and R. Derynck, *TGF-β-induced epithelial to mesenchymal transition.* Cell research, 2009. **19**(2): p. 156-172.
220. Yang, M.-H., et al., *Direct regulation of TWIST by HIF-1α promotes metastasis.* Nature cell biology, 2008. **10**(3): p. 295-305.
221. Yang, M.-H., et al., *Bmi1 is essential in Twist1-induced epithelial–mesenchymal transition.* Nature cell biology, 2010. **12**(10): p. 982-992.
222. Yang, F., et al., *SET8 promotes epithelial–mesenchymal transition and confers TWIST dual transcriptional activities.* The EMBO journal, 2012. **31**(1): p. 110-123.
223. Hong, J., et al., *Phosphorylation of serine 68 of Twist1 by MAPKs stabilizes Twist1 protein and promotes breast cancer cell invasiveness.* Cancer research, 2011. **71**(11): p. 3980-3990.
224. Sanchez-Tillo, E., et al., *ZEB1 represses E-cadherin and induces an EMT by recruiting the SWI/SNF chromatin-remodeling protein BRG1.* Oncogene, 2010. **29**(24): p. 3490-3500.
225. Postigo, A.A., et al., *Regulation of Smad signaling through a differential recruitment of coactivators and corepressors by ZEB proteins.* The EMBO journal, 2003. **22**(10): p. 2453-2462.
226. Gheldof, A., et al., *Evolutionary functional analysis and molecular regulation of the ZEB transcription factors.* Cellular and Molecular Life Sciences, 2012. **69**(15): p. 2527-2541.
227. Dave, N., et al., *Functional cooperation between Snail1 and twist in the regulation of ZEB1 expression during epithelial to mesenchymal transition.* Journal of Biological Chemistry, 2011. **286**(14): p. 12024-12032.

228. Chaffer, C.L., et al., *Poised chromatin at the ZEB1 promoter enables breast cancer cell plasticity and enhances tumorigenicity.* Cell, 2013. **154**(1): p. 61-74.

229. Llorens, M.C., et al., *Phosphorylation regulates functions of ZEB1 transcription factor.* Journal of cellular physiology, 2016. **231**(10): p. 2205-2217.

230. Dhasarathy, A., et al., *The transcription factors Snail and Slug activate the transforming growth factor-beta signaling pathway in breast cancer.* PloS one, 2011. **6**(10).

231. Gudey, S.K., et al., *Pro-invasive properties of Snail1 are regulated by sumoylation in response to TGFβ stimulation in cancer.* Oncotarget, 2017. **8**(58): p. 97703.

232. Ye, X. and R.A. Weinberg, *The sumo guards for snail.* Oncotarget, 2017. **8**(58): p. 97701.

233. Grelet, S., et al., *A regulated PNUTS mRNA to lncRNA splice switch mediates EMT and tumour progression.* Nature cell biology, 2017. **19**(9): p. 1105-1115.

234. Korpal, M. and Y. Kang, *The emerging role of miR-200 family of microRNAs in epithelial-mesenchymal transition and cancer metastasis.* RNA biology, 2008. **5**(3): p. 115-119.

235. Grelet, S., et al., *Pleiotropic roles of non-coding RNAs in TGF-β-mediated epithelial-mesenchymal transition and their functions in tumor progression.* Cancers, 2017. **9**(7): p. 75.

236. Savagner, P., *Leaving the neighborhood: molecular mechanisms involved during epithelial-mesenchymal transition.* Bioessays, 2001. **23**(10): p. 912-923.

237. Liu, P., et al., *Requirement for Wnt3 in vertebrate axis formation.* Nature genetics, 1999. **22**(4): p. 361-365.

238. Clevers, H., *Wnt/β-catenin signaling in development and disease.* Cell, 2006. **127**(3): p. 469-480.

239. Arwert, E.N., E. Hoste, and F.M. Watt, *Epithelial stem cells, wound healing and cancer.* Nature Reviews Cancer, 2012. **12**(3): p. 170-180.

240. Tammela, T., et al., *A Wnt-producing niche drives proliferative potential and progression in lung adenocarcinoma.* Nature, 2017. **545**(7654): p. 355-359.

241. e Melo, F.d.S., et al., *A distinct role for Lgr5+ stem cells in primary and metastatic colon cancer.* Nature, 2017. **543**(7647): p. 676-680.

242. Batlle, E. and H. Clevers, *Cancer stem cells revisited.* Nature medicine, 2017. **23**(10): p. 1124-1134.

243. Osborne, B.A. and L.M. Minter, *Notch signalling during peripheral T-cell activation and differentiation.* Nature Reviews Immunology, 2007. **7**(1): p. 64-75.

244. Bray, S.J., *Notch signalling in context.* Nature reviews Molecular cell biology, 2016. **17**(11): p. 722.

245. Timmerman, L.A., et al., *Notch promotes epithelial-mesenchymal transition during cardiac development and oncogenic transformation.* Genes & development, 2004. **18**(1): p. 99-115.

246. Yuan, X., et al., *Notch signaling and EMT in non-small cell lung cancer: biological significance and therapeutic application.* Journal of hematology & oncology, 2014. **7**(1): p. 87.

247. Tang, Y., Y. Tang, and Y.-s. Cheng, *miR-34a inhibits pancreatic cancer progression through Snail1-mediated epithelial–mesenchymal transition and the Notch signaling pathway.* Scientific reports, 2017. **7**(1): p. 1-11.

248. Tashiro, E., et al., *Involvement of the MEK/ERK pathway in EGF-induced E-cadherin down-regulation.* Biochemical and biophysical research communications, 2016. **477**(4): p. 801-806.

249. Lo, H.-W., et al., *Epidermal growth factor receptor cooperates with signal transducer and activator of transcription 3 to induce epithelial-mesenchymal transition in cancer cells via up-regulation of TWIST gene expression.* Cancer research, 2007. **67**(19): p. 9066-9076.

250. Colomiere, M., et al., *Cross talk of signals between EGFR and IL-6R through JAK2/STAT3 mediate epithelial–mesenchymal transition in ovarian carcinomas.* British journal of cancer, 2009. **100**(1): p. 134-144.

251. Heisenberg, C.-P. and L. Solnica-Krezel, *Back and forth between cell fate specification and movement during vertebrate gastrulation.* Current opinion in genetics & development, 2008. **18**(4): p. 311-316.

252. Uttamsingh, S., et al., *Synergistic effect between EGF and TGF-β1 in inducing oncogenic properties of intestinal epithelial cells.* Oncogene, 2008. **27**(18): p. 2626-2634.

253. Shirakihara, T., et al., *TGF-β regulates isoform switching of FGF receptors and epithelial–mesenchymal transition.* The EMBO journal, 2011. **30**(4): p. 783-795.

254. Lim, J. and J.P. Thiery, *Epithelial-mesenchymal transitions: insights from development.* Development, 2012. **139**(19): p. 3471-3486.

255. Thiery, J.P., et al., *Epithelial-mesenchymal transitions in development and disease.* cell, 2009. **139**(5): p. 871-890.

256. Murray, S.A. and T. Gridley, *Snail family genes are required for left–right asymmetry determination, but not neural crest formation, in mice.* Proceedings of the National Academy of Sciences, 2006. **103**(27): p. 10300-10304.

257. Mercado-Pimentel, M.E. and R.B. Runyan, *Multiple transforming growth factor-β isoforms and receptors function during epithelial-mesenchymal cell transformation in the embryonic heart.* Cells Tissues Organs, 2007. **185**(1-3): p. 146-156.

258. Krainock, M., et al., *Epicardial epithelial-to-mesenchymal transition in heart development and disease.* Journal of clinical medicine, 2016. **5**(2): p. 27.

259. Baek, S.T. and M.D. Tallquist, *Nf1 limits epicardial derivative expansion by regulating epithelial to mesenchymal transition and proliferation.* Development, 2012. **139**(11): p. 2040-2049.

260. Moore, A.W., et al., *YAC complementation shows a requirement for Wt1 in the development of epicardium, adrenal gland and throughout nephrogenesis.* Development, 1999. **126**(9): p. 1845-1857.

261. Martínez-Estrada, O.M., et al., *Wt1 is required for cardiovascular progenitor cell formation through transcriptional control of Snail and E-cadherin.* Nature genetics, 2010. **42**(1): p. 89.

262. von Gise, A., et al., *WT1 regulates epicardial epithelial to mesenchymal transition through β-catenin and retinoic acid signaling pathways.* Developmental biology, 2011. **356**(2): p. 421-431.

263. Kitazawa, K., et al., *OVOL2 maintains the transcriptional program of human corneal epithelium by suppressing epithelial-to-mesenchymal transition.* Cell reports, 2016. **15**(6): p. 1359-1368.

264. Watanabe, K., et al., *Mammary morphogenesis and regeneration require the inhibition of EMT at terminal end buds by Ovol2 transcriptional repressor.* Developmental cell, 2014. **29**(1): p. 59-74.

265. Chakrabarti, R., et al., *Elf5 inhibits the epithelial–mesenchymal transition in mammary gland development and breast cancer metastasis by transcriptionally repressing Snail2.* Nature cell biology, 2012. **14**(11): p. 1212-1222.

266. Chang, C.-J., et al., *p53 regulates epithelial–mesenchymal transition and stem cell properties through modulating miRNAs.* Nature cell biology, 2011. **13**(3): p. 317-323.

267. Warzecha, C.C. and R.P. Carstens. *Complex changes in alternative pre-mRNA splicing play a central role in the epithelial-to-mesenchymal transition (EMT)*. in *Seminars in cancer biology*. 2012. Elsevier.

268. Abell, A.N., et al., *MAP3K4/CBP-regulated H2B acetylation controls epithelial-mesenchymal transition in trophoblast stem cells.* Cell stem cell, 2011. **8**(5): p. 525-537.

269. Shaw, T.J. and P. Martin, *Wound repair: a showcase for cell plasticity and migration.* Current opinion in cell biology, 2016. **42**: p. 29-37.

270. Haensel, D. and X. Dai, *Epithelial-to-mesenchymal transition in cutaneous wound healing: Where we are and where we are heading.* Developmental dynamics, 2018. **247**(3): p. 473-480.

271. Eming, S.A., P. Martin, and M. Tomic-Canic, *Wound repair and regeneration: mechanisms, signaling, and translation.* Science translational medicine, 2014. **6**(265): p. 265sr6-265sr6.

272. Arnoux, V., et al., *Erk5 controls Slug expression and keratinocyte activation during wound healing.* Molecular biology of the cell, 2008. **19**(11): p. 4738-4749.

273. Hudson, L.G., et al., *Cutaneous wound reepithelialization is compromised in mice lacking functional Slug (Snai2).* Journal of dermatological science, 2009. **56**(1): p. 19-26.

274. Vaughan, A.E. and H.A. Chapman, *Regenerative activity of the lung after epithelial injury.* Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 2013. **1832**(7): p. 922-930.

275. Qian, L.W., et al., *Exacerbated and prolonged inflammation impairs wound healing and increases scarring.* Wound repair and regeneration, 2016. **24**(1): p. 26-34.

276. Hutchinson, J., et al., *Global incidence and mortality of idiopathic pulmonary fibrosis: a systematic review.* European Respiratory Journal, 2015. **46**(3): p. 795-806.

277. Sato, M., et al., *Targeted disruption of TGF-β1/Smad3 signaling protects against renal tubulointerstitial fibrosis induced by unilateral ureteral obstruction.* The Journal of clinical investigation, 2003. **112**(10): p. 1486-1494.

278. Boutet, A., et al., *Snail activation disrupts tissue homeostasis and induces fibrosis in the adult kidney.* The EMBO journal, 2006. **25**(23): p. 5603-5613.

279. Humphreys, B.D., et al., *Fate tracing reveals the pericyte and not epithelial origin of myofibroblasts in kidney fibrosis.* The American journal of pathology, 2010. **176**(1): p. 85-97.

280. LeBleu, V.S., et al., *Origin and function of myofibroblasts in kidney fibrosis.* Nature medicine, 2013. **19**(8): p. 1047.

281. Borges, F.T., et al., *TGF-β1–containing exosomes from injured epithelial cells activate fibroblasts to initiate tissue regenerative responses and fibrosis.* Journal of the American Society of Nephrology, 2013. **24**(3): p. 385-392.

282. Grande, M.T., et al., *Snail1-induced partial epithelial-to-mesenchymal transition drives renal fibrosis in mice and can be targeted to reverse established disease.* Nature medicine, 2015. **21**(9): p. 989.

283. Rowe, R.G., et al., *Hepatocyte-derived Snail1 propagates liver fibrosis progression.* Molecular and cellular biology, 2011. **31**(12): p. 2392-2403.

284. Marmai, C., et al., *Alveolar epithelial cells express mesenchymal proteins in patients with idiopathic pulmonary fibrosis.* American Journal of Physiology-Lung Cellular and Molecular Physiology, 2011. **301**(1): p. L71-L78.

285. Chen, Y., et al., *Sorafenib ameliorates bleomycin-induced pulmonary fibrosis: potential roles in the inhibition of epithelial–mesenchymal*

transition and fibroblast activation. Cell death & disease, 2013. **4**(6): p. e665-e665.

286. Wang, C., et al., *Low-dose paclitaxel ameliorates pulmonary fibrosis by suppressing TGF-β1/Smad3 pathway via miR-140 upregulation.* PLoS One, 2013. **8**(8).

287. Pedroza, M., et al., *STAT-3 contributes to pulmonary fibrosis through epithelial injury and fibroblast-myofibroblast differentiation.* The FASEB Journal, 2016. **30**(1): p. 129-140.

288. Palumbo-Zerr, K., et al., *Orphan nuclear receptor NR4A1 regulates transforming growth factor-β signaling and fibrosis.* Nature medicine, 2015. **21**(2): p. 150.

289. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* cell, 2011. **144**(5): p. 646-674.

290. Huang, R.Y., et al., *An EMT spectrum defines an anoikis-resistant and spheroidogenic intermediate mesenchymal state that is sensitive to e-cadherin restoration by a src-kinase inhibitor, saracatinib (AZD0530).* Cell death & disease, 2013. **4**(11): p. e915-e915.

291. Tan, T.Z., et al., *Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients.* EMBO molecular medicine, 2014. **6**(10): p. 1279-1293.

292. Yu, Y., et al., *Cancer-associated fibroblasts induce epithelial–mesenchymal transition of breast cancer cells through paracrine TGF-β signalling.* British journal of cancer, 2014. **110**(3): p. 724-732.

293. Shintani, Y., et al., *IL-6 secreted from cancer-associated fibroblasts mediates chemoresistance in NSCLC by increasing epithelial-mesenchymal transition signaling.* Journal of Thoracic Oncology, 2016. **11**(9): p. 1482-1492.

294. Bonde, A.-K., et al., *Intratumoral macrophages contribute to epithelial-mesenchymal transition in solid tumors.* BMC cancer, 2012. **12**(1): p. 35.

295. Vega, S., et al., *Snail blocks the cell cycle and confers resistance to cell death.* Genes & development, 2004. **18**(10): p. 1131-1143.

296. Singh, A. and J. Settleman, *EMT, cancer stem cells and drug resistance: an emerging axis of evil in the war on cancer.* Oncogene, 2010. **29**(34): p. 4741-4751.

297. Jurmeister, S., et al., *MicroRNA-200c represses migration and invasion of breast cancer cells by targeting actin-regulatory proteins FHOD1 and PPM1F.* Molecular and cellular biology, 2012. **32**(3): p. 633-651.

298. Mani, S.A., et al., *The epithelial-mesenchymal transition generates cells with properties of stem cells.* Cell, 2008. **133**(4): p. 704-715.

299. Morel, A.-P., et al., *Generation of breast cancer stem cells through epithelial-mesenchymal transition.* PloS one, 2008. **3**(8).

300. Chaffer, C.L., et al., *EMT, cell plasticity and metastasis.* Cancer and Metastasis Reviews, 2016. **35**(4): p. 645-654.

301. Wrzesinski, S.H., Y.Y. Wan, and R.A. Flavell, *Transforming growth factor-β and the immune response: implications for anticancer therapy.* Clinical cancer research, 2007. **13**(18): p. 5262-5270.

302. Kudo-Saito, C., et al., *CCL2 is critical for immunosuppression to promote cancer metastasis.* Clinical & experimental metastasis, 2013. **30**(4): p. 393-405.

303. Hsu, D.S.-S., et al., *Acetylation of snail modulates the cytokinome of cancer cells to enhance the recruitment of macrophages.* Cancer cell, 2014. **26**(4): p. 534-548.

304. Fruci, D., et al., *Major histocompatibility complex class i and tumour immuno-evasion: how to fool T cells and natural killer cells at one time.* Current Oncology, 2012. **19**(1): p. 39.

305. Blum, J.S., P.A. Wearsch, and P. Cresswell, *Pathways of antigen processing.* Annual review of immunology, 2013. **31**: p. 443-473.

306. Chen, L., et al., *Metastasis is regulated via microRNA-200/ZEB1 axis control of tumour cell PD-L1 expression and intratumoral immunosuppression.* Nature communications, 2014. **5**(1): p. 1-12.

307. Lambert, A.W., D.R. Pattabiraman, and R.A. Weinberg, *Emerging biological principles of metastasis.* Cell, 2017. **168**(4): p. 670-691.

308. Krebs, A.M., et al., *The EMT-activator Zeb1 is a key factor for cell plasticity and promotes metastasis in pancreatic cancer.* Nature cell biology, 2017. **19**(5): p. 518-529.

309. Ye, X., et al., *Distinct EMT programs control normal mammary stem cells and tumour-initiating cells.* Nature, 2015. **525**(7568): p. 256-260.

310. Guo, W., et al., *Slug and Sox9 cooperatively determine the mammary stem cell state.* Cell, 2012. **148**(5): p. 1015-1028.

311. Chambers, A.F., A.C. Groom, and I.C. MacDonald, *Dissemination and growth of cancer cells in metastatic sites.* Nature Reviews Cancer, 2002. **2**(8): p. 563-572.

312. Aceto, N., et al., *En route to metastasis: circulating tumor cell clusters and epithelial-to-mesenchymal transition.* Trends in cancer, 2015. **1**(1): p. 44-52.

313. Pattabiraman, D.R., et al., *Activation of PKA leads to mesenchymal-to-epithelial transition and loss of tumor-initiating ability.* Science, 2016. **351**(6277): p. aad3680.

314. Schmidt, J.M., et al., *Stem-cell-like properties and epithelial plasticity arise as stable traits after transient Twist1 activation.* Cell reports, 2015. **10**(2): p. 131-139.

315. Alderton, G.K., *Epithelial to mesenchymal and back again.* Nature Reviews Cancer, 2013. **13**(1): p. 3-3.

316. Rodon, J., et al., *First-in-human dose study of the novel transforming growth factor-β receptor I kinase inhibitor LY2157299 monohydrate in patients with advanced cancer and glioma.* Clinical Cancer Research, 2015. **21**(3): p. 553-560.

317. Meidhof, S., et al., *ZEB1-associated drug resistance in cancer cells is reversed by the class I HDAC inhibitor mocetinostat.* EMBO molecular medicine, 2015. **7**(6): p. 831-847.

318. McFaline-Figueroa, J.L., et al., *A pooled single-cell genetic screen identifies regulatory checkpoints in the continuum of the epithelial-to-mesenchymal transition.* Nature genetics, 2019. **51**(9): p. 1389-1398.

319. Karacosta, L.G., et al., *Mapping lung cancer epithelial-mesenchymal transition states and trajectories with single-cell resolution.* Nature communications, 2019. **10**(1): p. 1-15.

320. Torre, L.A., et al., *Global cancer incidence and mortality rates and trends—an update.* Cancer Epidemiology and Prevention Biomarkers, 2016. **25**(1): p. 16-27.

321. Britt, K., A. Ashworth, and M. Smalley, *Pregnancy and the risk of breast cancer.* Endocrine-related cancer, 2007. **14**(4): p. 907-933.

322. Danaei, G., et al., *Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors.* The Lancet, 2005. **366**(9499): p. 1784-1793.

323. Mørch, L.S., et al., *Contemporary hormonal contraception and the risk of breast cancer.* New England Journal of Medicine, 2017. **377**(23): p. 2228-2239.

324. Perou, C.M., et al., *Molecular portraits of human breast tumours.* nature, 2000. **406**(6797): p. 747-752.

325. Board, W.C.o.T.E., *Breast Tumours*. 2019: International Agency for Research on Cancer.

326. Greaves, M. and C.C. Maley, *Clonal evolution in cancer.* Nature, 2012. **481**(7381): p. 306.

327. Shackleton, M., et al., *Heterogeneity in cancer: cancer stem cells versus clonal evolution.* Cell, 2009. **138**(5): p. 822-829.

328. Bombonati, A. and D.C. Sgroi, *The molecular pathology of breast cancer progression.* The Journal of pathology, 2011. **223**(2): p. 308-318.

329. Lopez-Garcia, M.A., et al., *Breast cancer precursors revisited: molecular features and progression pathways.* Histopathology, 2010. **57**(2): p. 171-192.

330. Nik-Zainal, S., et al., *Landscape of somatic mutations in 560 breast cancer whole-genome sequences.* Nature, 2016. **534**(7605): p. 47-54.

331. Yates, L.R., et al., *Genomic evolution of breast cancer metastasis and relapse.* Cancer cell, 2017. **32**(2): p. 169-184. e7.

332. Yates, L.R. and C. Desmedt, *Translational genomics: practical applications of the genomic revolution in breast cancer*. 2017, AACR.

333. Shiovitz, S. and L.A. Korde, *Genetics of breast cancer: a topic in evolution.* Annals of Oncology, 2015. **26**(7): p. 1291-1299.

334. Huen, M.S., S.M. Sy, and J. Chen, *BRCA1 and its toolbox for the maintenance of genome integrity.* Nature reviews Molecular cell biology, 2010. **11**(2): p. 138-148.

335. Kuchenbaecker, K.B., et al., *Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers.* Jama, 2017. **317**(23): p. 2402-2416.

336. Daly, M.B., et al., *Genetic/familial high-risk assessment: breast and ovarian, version 2.2015.* Journal of the National Comprehensive Cancer Network, 2016. **14**(2): p. 153-162.

337. Mego, M., S.A. Mani, and M. Cristofanilli, *Molecular mechanisms of metastasis in breast cancer—clinical applications.* Nature reviews Clinical oncology, 2010. **7**(12): p. 693.

338. Aurilio, G., et al., *A meta-analysis of oestrogen receptor, progesterone receptor and human epidermal growth factor receptor 2 discordance between primary breast cancer and metastases.* European journal of cancer, 2014. **50**(2): p. 277-289.

339. Dunn, G.P., L.J. Old, and R.D. Schreiber, *The three Es of cancer immunoediting.* Annu. Rev. Immunol., 2004. **22**: p. 329-360.

340. Waldhauer, I. and A. Steinle, *NK cells and cancer immunosurveillance.* Oncogene, 2008. **27**(45): p. 5932-5943.

341. Edechi, C.A., et al., *Regulation of Immunity in Breast Cancer.* Cancers, 2019. **11**(8): p. 1080.

342. Dadi, S., et al., *Cancer immunosurveillance by tissue-resident innate lymphoid cells and innate-like T cells.* Cell, 2016. **164**(3): p. 365-377.

343. Landskron, G., et al., *Chronic inflammation and cytokines in the tumor microenvironment.* Journal of immunology research, 2014. **2014**.

344. Grivennikov, S.I., F.R. Greten, and M. Karin, *Immunity, inflammation, and cancer.* Cell, 2010. **140**(6): p. 883-899.

345. Palucka, A.K. and L.M. Coussens, *The basis of oncoimmunology.* Cell, 2016. **164**(6): p. 1233-1247.

346. Toor, S.M. and E. Elkord, *Comparison of myeloid cells in circulation and in the tumor microenvironment of patients with colorectal and breast cancers.* Journal of immunology research, 2017. **2017**.

347. Qian, B.-Z. and J.W. Pollard, *Macrophage diversity enhances tumor progression and metastasis.* Cell, 2010. **141**(1): p. 39-51.

348. Markowitz, J., et al., *Myeloid-derived suppressor cells in breast cancer.* Breast cancer research and treatment, 2013. **140**(1): p. 13-21.

349. Shou, D., et al., *Suppressive role of myeloid-derived suppressor cells (MDSCs) in the microenvironment of breast cancer and targeted immunotherapies.* Oncotarget, 2016. **7**(39): p. 64505.

350. Welte, T., et al., *Oncogenic mTOR signalling recruits myeloid-derived suppressor cells to promote tumour initiation.* Nature cell biology, 2016. **18**(6): p. 632-644.

351. Diaz-Montero, C.M., et al., *Increased circulating myeloid-derived suppressor cells correlate with clinical cancer stage, metastatic tumor burden, and doxorubicin–cyclophosphamide chemotherapy.* Cancer immunology, immunotherapy, 2009. **58**(1): p. 49-59.

352. Qiu, S.-Q., et al., *Tumor-associated macrophages in breast cancer: Innocent bystander or important player?* Cancer treatment reviews, 2018. **70**: p. 178-189.

353. Mosser, D.M. and J.P. Edwards, *Exploring the full spectrum of macrophage activation.* Nature reviews immunology, 2008. **8**(12): p. 958-969.

354. Mantovani, A., et al., *Tumour-associated macrophages as treatment targets in oncology.* Nature reviews Clinical oncology, 2017. **14**(7): p. 399.

355. Appay, V., D.C. Douek, and D.A. Price, *CD8+ T cell efficacy in vaccination and disease.* Nature medicine, 2008. **14**(6): p. 623.

356. Ribas, A. and J.D. Wolchok, *Cancer immunotherapy using checkpoint blockade.* Science, 2018. **359**(6382): p. 1350-1355.

357. Mao, Y., et al., *The prognostic value of tumor-infiltrating lymphocytes in breast cancer: a systematic review and meta-analysis.* PloS one, 2016. **11**(4).

358. Luckheeram, R.V., et al., *CD4+ T cells: differentiation and functions.* Clinical and developmental immunology, 2012. **2012**.

359. DeNardo, D.G., et al., *CD4+ T cells regulate pulmonary metastasis of mammary carcinomas by enhancing protumor properties of macrophages.* Cancer cell, 2009. **16**(2): p. 91-102.

360. Gupta, S., et al., *Intratumoral FOXP3 expression in infiltrating breast carcinoma: Its association with clinicopathologic parameters and angiogenesis.* Acta oncologica, 2007. **46**(6): p. 792-797.

361. Khaja, A.S.S., et al., *Preferential accumulation of regulatory T cells with highly immunosuppressive characteristics in breast tumor microenvironment.* Oncotarget, 2017. **8**(20): p. 33159.

362. Shen, M., J. Wang, and X. Ren, *New insights into tumor-infiltrating B lymphocytes in breast cancer: clinical impacts and regulatory mechanisms.* Frontiers in immunology, 2018. **9**: p. 470.

363. Carmi, Y., et al., *Allogeneic IgG combined with dendritic cell stimuli induce antitumour T-cell immunity.* Nature, 2015. **521**(7550): p. 99-104.

364. Yanaba, K., et al., *A regulatory B cell subset with a unique CD1dhiCD5+ phenotype controls T cell-dependent inflammatory responses.* Immunity, 2008. **28**(5): p. 639-650.

365. Zhang, Y., et al., *B lymphocyte inhibition of anti-tumor response depends on expansion of Treg but is independent of B-cell IL-10 secretion.* Cancer Immunology, Immunotherapy, 2013. **62**(1): p. 87-99.

366. Smid, M., et al., *Breast cancer genome and transcriptome integration implicates specific mutational signatures with immune cell infiltration.* Nature communications, 2016. **7**(1): p. 1-9.

367. Luen, S.J., et al., *Tumour-infiltrating lymphocytes and the emerging role of immunotherapy in breast cancer.* Pathology, 2017. **49**(2): p. 141-155.

368. Adams, S., et al., *Current landscape of immunotherapy in breast cancer: a review.* JAMA oncology, 2019. **5**(8): p. 1205-1214.

369. Marmot, M.G., et al., *The benefits and harms of breast cancer screening: an independent review.* British journal of cancer, 2013. **108**(11): p. 2205-2240.

370. Cardoso, F., et al., *Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up.* Annals of Oncology, 2019. **30**(8): p. 1194-1220.

371. Cortazar, P., et al., *Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis.* The Lancet, 2014. **384**(9938): p. 164-172.

372. McGale, P., et al., *Effect of radiotherapy after mastectomy and axillary surgery on 10-year recurrence and 20-year breast cancer mortality: meta-analysis of individual patient data for 8135 women in 22 randomised trials*. 2014, Elsevier.

373. Peto, R., et al., *Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100,000 women in 123 randomised trials.* Lancet, 2012. **379**(9814).

374. Thavarajah, N., et al., *Continued success in providing timely palliative radiation therapy at the Rapid Response Radiotherapy Program: a review of 2008–2012.* Current Oncology, 2013. **20**(3): p. e206.

375. Bernier, J., *Immuno-oncology: Allying forces of radio-and immuno-therapy to enhance cancer cell killing.* Critical reviews in oncology/hematology, 2016. **108**: p. 97-108.

376. Luen, S., et al., *The genomic landscape of breast cancer and its interaction with host immunity.* The Breast, 2016. **29**: p. 241-250.

377. Chen, D.S. and I. Mellman, *Elements of cancer immunity and the cancer–immune set point.* Nature, 2017. **541**(7637): p. 321-330.

378. Mazutis, L., et al., *Single-cell analysis and sorting using droplet-based microfluidics.* Nature protocols, 2013. **8**(5): p. 870.

379. Mørup, M. and L.K. Hansen, *Archetypal analysis for machine learning and data mining.* Neurocomputing, 2012. **80**: p. 54-63.

380. Ding, J., et al., *Systematic comparison of single-cell and single-nucleus RNA-sequencing methods.* Nature Biotechnology, 2020: p. 1-10.

381. Zhang, X., et al., *Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-seq systems.* Molecular cell, 2019. **73**(1): p. 130-142. e5.

382. Song, J. and W. Shi, *The concomitant apoptosis and EMT underlie the fundamental functions of TGF-β.* Acta biochimica et biophysica Sinica, 2018. **50**(1): p. 91-97.

383. Tiwari, N., et al., *Sox4 is a master regulator of epithelial-mesenchymal transition by controlling Ezh2 expression and epigenetic reprogramming.* Cancer cell, 2013. **23**(6): p. 768-783.

384. Wong, D.J., et al., *Module map of stem cell genes guides creation of epithelial cancer stem cells.* Cell stem cell, 2008. **2**(4): p. 333-344.

385. David, C.J., et al., *TGF-β tumor suppression through a lethal EMT.* Cell, 2016. **164**(5): p. 1015-1030.

386. Ben-Porath, I., et al., *An embryonic stem cell–like gene expression signature in poorly differentiated aggressive human tumors.* Nature genetics, 2008. **40**(5): p. 499.

387. Krishnaswamy, S., et al., *Conditional density-based analysis of T cell signaling in single-cell data.* Science, 2014. **346**(6213): p. 1250689.

388. Dixit, A., et al., *Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens.* Cell, 2016. **167**(7): p. 1853-1866. e17.

389. Datlinger, P., et al., *Pooled CRISPR screening with single-cell transcriptome readout.* Nature methods, 2017. **14**(3): p. 297.

390. Prabhakaran, S., et al. *Dirichlet process mixture model for correcting technical variation in single-cell gene expression data.* in *International Conference on Machine Learning*. 2016.

391. Josefowicz, S.Z., L.-F. Lu, and A.Y. Rudensky, *Regulatory T cells: mechanisms of differentiation and function.* Annual review of immunology, 2012. **30**: p. 531-564.

392. Tirosh, I., et al., *Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq.* Science, 2016. **352**(6282): p. 189-196.

393. Müller, S., et al., *Single-cell profiling of human gliomas reveals macrophage ontogeny as a basis for regional differences in macrophage activation in the tumor microenvironment.* Genome biology, 2017. **18**(1): p. 234.

394. Joller, N., et al., *Treg cells expressing the coinhibitory molecule TIGIT selectively inhibit proinflammatory Th1 and Th17 cell responses.* Immunity, 2014. **40**(4): p. 569-581.

395. Perdiguero, E.G. and F. Geissmann, *The development and maintenance of resident macrophages.* Nature immunology, 2016. **17**(1): p. 2.

396. Chevrier, S., et al., *An immune atlas of clear cell renal cell carcinoma.* Cell, 2017. **169**(4): p. 736-749. e18.

397. Lavin, Y., et al., *Innate immune landscape in early lung adenocarcinoma by paired single-cell analyses.* Cell, 2017. **169**(4): p. 750-765. e17.

NOTES

Vilniaus universiteto leidykla
Saulėtekio al. 9, III rūmai LT-10222 Vilnius
El. p. info@leidykla.vu.lt,
www.leidykla.vu.lt
Tiražas 15 egz.