Collective migration and phenotypic plasticity in cancer metastasis: conflicting views or complementary mechanisms?

Master thesis submitted to Delft University of Technology

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in Nanobiology

Faculty of Applied Sciences

by

Mathijs Verhagen

Student number: 4263642

To be defended in public on 7th of August 2019

Graduation committee

Chairperson:	Prof. Dr. R. Fodde, Department of Pathology (EMC)
Daily Supervisor:	M. Teeuwssen, MSc, Department of Pathology (EMC)
Second Supervisor:	Prof. Dr. G.W. Jenster, Department of Experimental Urological Oncology (EMC)
Third Supervisor:	Dr. M.E. van Royen, Optical Imaging Center (EMC)

Abstract

Cancer metastasis, the spread of cancer to distant organs, is the major cause of cancer-related mortality. Hence, understanding the mechanisms underlying cancer metastasis is crucial to improve clinical interventions. Despite intensive efforts, the driving mechanisms remain ill-understood due to the difficulties posed by studying the different steps of the metastatic cascade in patients. Two established models have been proposed to underlie the driving mechanisms of metastasis are phenotypic plasticity and collective migration. Phenotypic plasticity, i.e. the capacity of the migrating cancer cell to adapt to the different cellular contexts that it encounters *en route* to form a metastasis, revolves around reversable transitions from epithelial to mesenchymal (EMT and MET) identities. The collective migration model denotes migrating cancer cells can overcome barriers by coordinated cooperation. Recently, these views have been integrated in a model where partial (EMT) is believed to mediate collective migration.

Here, we will investigate critical assumptions of this integrated model by focusing on different steps along the invasion-metastasis cascade. Using an unsupervised approach based on complete transcriptomes, we unravel single cell EMT-related transcriptional differences in colorectal cancer cell lines and map different phenotypes on the EMT spectrum to identify E/M sub-states that might underlie collective invasion. Next, we have developed and evaluated 3D collagen models to facilitate collective migration studies both *in vitro* and *ex vivo*. Preliminary data obtained using this approach highlights how EMT induction can alter the dominating tumor migration type.

Taken together, our results support a case for phenotypic plasticity and collective migration as complementary and functionally correlated mechanisms, and could serve as point of engagement for further studies aimed at clarifying the role of partial EMT in collective migration.

Key words: single-cell RNA sequencing, epithelial-mesenchymal transition, E/M sub-state, collagen, collective migration, circulating tumor cell

Contents

List of abbreviations

CAF	cancer associated fibroblast
CIN	chromosomal instability
CRC	colorectal cancer
CTC	circulating tumor cell
E/M	epithelial/mesenchymal
ECM	extra-cellular matrix
EMT	epithelial-mesenchymal transition
FACS	fluorescent activated cell sorting
IPA	ingenuity pathway analysis
ISC	intestinal stem cell
MAGIC	markov affinity-based graph imputation of cells
MET	mesenchymal-epithelial transition
MIN	microsatellite instability
PC	paneth cell
PSF	phenotypic stability factor
qPCR	quantitative polymerase chain reaction
RBC	red blood cell
scRNAseq	single cell RNA sequencing
tSNE	t-distributed stochastic neighbor embedding
UMAP	uniform manifold approximation and projection
UMI	unique molecular identifier
WBC	white blood cell

List of figures

List of Tables

Chapter 1. Introduction

Colorectal cancer (CRC) is the third most common cancer worldwide, and its incidence is expected to rise in the coming years[1]. Despite advancements in treatment options and improvements in surgical procedures, CRC remains a lethal disease for a substantial group of patients. In particular, CRC becomes lethal when the cancer spreads over the body, referred to as metastatic CRC. About half of the cases involve metastatic CRC, with the liver and lungs as the main sites of metastasis formation. Patients suffering from metastatic CRC are difficult to treat, and have a poor prognosis with a five-year survival of only 15%[2].

The adenoma-carcinoma sequence in colorectal cancer

While the process of metastasis formation is still far from being fully understood, the cascade of genetic alterations ultimately resulting in colon carcinomas has been well characterized in the past decades and has served as a unique model to elucidate the molecular and cellular mechanisms underlying tumor onset and progression to malignancy[3,4]. In a normal, healthy, colon the intestinal epithelium is organized in crypts, and, in the small intestine, in villi and crypts , which provide the lower gastrointestinal tract a large surface for extraction and absorption of energy and nutrients. Along the crypt-villus axis, different cellular lineages with distinct functional roles are found at conserved positions. At the bottom of the crypts, intestinal stem cells (ISCs) are intermingled with secretory Paneth cells (PCs). Together, ISC and PCs constitute the stem cell niche responsible for the high regenerative turnover of the intestinal tract. The progeny of ISCs progressively moves upwards by first undergoing a proliferative burst (transient-amplifying or TA cells) to then eventually differentiate into enterocytes and 4 other specialized intestinal lineages, namely entero-endocrine, Tuft, goblet, and the above-mentioned secretory Paneth cells. At the top of the villus, these post-mitotic lineages detach from the epithelial lining when a programmed cell death occurs.

In the adenoma-carcinoma sequence of colon cancer, the normal epithelium develops into precursor lesions followed by increasingly more invasive stages through the accumulation of sequential mutations at tumor suppressors and oncogenes. Most of the sporadic colorectal cancer cases are initiated by loss of function mutations in the *APC* gene, found in over 80% of colon adenocarcinomas[5]. Loss of APC function leads to constitutive Wnt signaling activation, which provide the cell with a growth advantage and result in abnormal proliferation. As a consequence, an adenomatous polyp or adenoma can arise as first indication of abnormal cell growth.

The steps following adenoma onset underlie its growth and accordingly include mutations that result in growth stimulation and apoptosis inhibition. Gain of function mutations of the *KRAS* oncogene occur in 50% of colorectal adenomas larger than 1 cm^3, and result in activation of the Ras signaling pathway that stimulates transcriptional activation of genes involved in proliferation and apoptosis inhibition[6].

Further progression towards malignancy is accompanied by the loss of the *TP53* tumor suppressor gene, a crucial transcription factor involved in cell cycle arrest and apoptosis. Deletions leading to loss of *TP53* occur in over 75% of colorectal carcinomas[7], and embark the transition from benign to invasive cancer . While *APC*, *KRAS*, and *P53* play a central role in the adenoma-carcinoma sequence, they are not the only genes involved. They make part of a larger pool of genes that contribute to the adenoma-carcinoma sequence resulting in alternative genetic paths by which cancer can progress. Together, these genetic changes represent requirements for colon cancer onset and progression, and represent

the full set of hallmarks of cancer[8]. Interestingly, while these genetic alterations ultimately result in invasive carcinomas, they are unlikely to provide the cancer cells with the capacity to invade the stromal microenvironment and form distant metastases. In fact, the process of metastasis formation is highly inefficient and only a very small fraction of the primary cancer cells is capable of completing the multi-step route to metastasis formation[9]. Moreover, the multi-step genetic progression model carries a conceptual inconsistency when trying to explain the metastasis process, which has been debated as a progression puzzle[10]. That is, there is no reason to think that genes specifying the final step in tumor progression – metastasis – enable cells to proliferate more efficiently in the primary tumor. Hence, following Darwinian evolution, the fraction of cells in the tumor mass to acquire metastatic potential would remain rare. Given the extraordinary low success rate of individual cells undertaking metastasis, it is unlikely metastasis could ever proceed. So what makes some cells 'more equal than others': why do cells with identical genotype have different metastatic potential?

In the absence of genetic mutations explaining the transition from carcinoma at the primary site to the metastasis, it is plausible to think that epigenetic modifications underlie the dissemination of cancer cells into the tumor microenvironment and their long journey to a distant organ. From this perspective, phenotypic plasticity, i.e. the capacity of the migrating cancer cell to adapt to the different cellular contexts that it encounters en route to form a metastasis, represents a key feature and a central issue in this thesis.

The second aspect of colon cancer metastasis addressed in this thesis and highly relevant for the above-mentioned phenotypic plasticity, is collective cell migration. To date, it is not entirely clear whether metastases are preferentially initiated by single migrating cancer cells or by cell clusters shed from the primary tumor. Here, evidence will be highlighted in support of both views and an integrated model will be discussed that integrates phenotypic plasticity with the two migration modalities. This thesis will question several assumptions underlying the integrated model, and aims to contribute to the debate on whether phenotypic plasticity and collective migration are complementary mechanisms, or alternative views that compete to explain the process of cancer metastasis.

Phenotypic plasticity as the main hallmark of the metastasizing carcinoma cell

Phenotypic plasticity has been argued as the clinically most relevant hallmark of cancer cells[11]. The capacity to undergo transient and reversable changes in morphology and functionality can provide a cell with the necessary means to survive its journey to metastasis formation. In particular, the reversable transitions from epithelial to mesenchymal (EMT and MET) identities, representing loss of adhesive traits and the acquisition of migratory phenotypes, has gained a central role in the literature (Figure 1A). Different signaling cascades (Tgf-ß, Wnt, Notch) trigger the epigenetic activation of specific transcription factors (*ZEB1/2, SNAI1/2, TWIST1/2*), the so-called EMT-TFs, which control the expression if key downstream target genes regulating epithelial and mesenchymal cell identity.

Hence, epithelial cancer cells can undergo EMT to acquire mesenchymal phenotypes thus facilitating the process of dissemination and intravasation. Upregulation of mesenchymal-specific genes (e.g. vimentin, fibronectin, metalloproteinases) enable degradation of extracellular matrix (ECM) and changes in cell polarity and shape. Vice versa, downregulation of epithelial-specific genes (e.g. E-cadherin, claudins, and cytokeratins) affects cell adhesion and cytoskeletal organization. At the organ site of metastasis, the reverse process, mesenchymal to epithelial transition (MET), converts the cancer cell back to the epithelial phenotype, stimulating proliferation and enabling colonization.

Thus, the EMT/MET model of metastasis formation can explain the epithelial characteristics shared

between primary tumor and metastasis notwithstanding the inherently non-invasive features of epithelial cells. To bridge the gap, a migratory phenotype, which bears similarity to mesenchymal cells, is proposed to function as temporary state enabling invasion in the stroma, intra- and extravasation, and colonization of a distant organ site[12].

Evidence supporting a role for EMT in cancer metastasis has come from different studies. For example, conditional knock out of important EMT transcription factors, including Zeb1 and Snai1, has been shown to be critical for metastasis formation[13,14], while conditional knock out of miR-200s, known to target Zeb1/2, was found sufficient to drive metastasis[15]. Moreover, inducible overexpression of Twist1[16] and Snai1[14] was shown to be promote cells to undergo EMT and disseminate into the blood circulation, but that these transcription factors need to be turned off in order for disseminated cells to proliferate and form metastases.

Given the above, the question arises on how the EMT/MET model explains why only a small subpopulation of the cancer cells have the potential to form metastasis. Induction of EMT cannot be explained by the genetic alterations observed along the adenoma-carcinoma sequence. As example, deletion of the *APC* gene should switch on the Wnt signaling pathway (Figure 1B). APC is involved in the phosphorylation of ß-catenin which triggers its degradation and resembles the state of the Wnt pathway during homeostasis. When this process is hampered, ß-catenin is recruited to the nucleus where it can induce the expression of Wnt target genes. While deletion of the *APC* gene should result in accumulation of nuclear ß-catenin, this is only observed in cells at the invasive margin of the tumor, a phenomenon referred to as the ß-catenin paradox[17]. Hence, it was proposed that secreted factors from the tumor microenvironment are needed to elicit full blown Wnt signaling and trigger EMT[18]. EMT is a complex, non-autonomous, and context-dependent cellular program that can be triggered by different factors[12]. These observations were conceptualized in a model where stationary cancer stem cells could transiently develop into migratory cancer stem cells when localized in the proximity of EMT-inducing cues[19], and that the latter acquire stem-like features upon EMT activation[20].

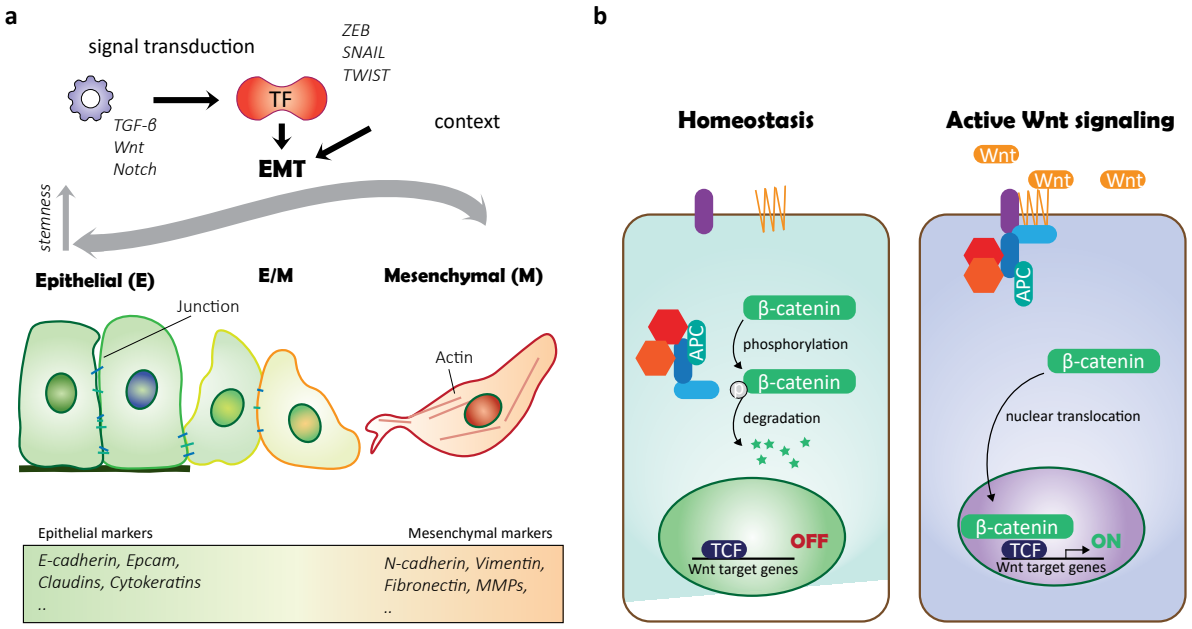


Figure 1. Schematic visualization of Epithelial-Mesenchymal transition and Wnt signaling pathway. (a) EMT can be induced by various signaling pathways via key EMT regulating transcription factors and is dependent on context. Epithelial (E) cells transition to Mesenchymal (M) phenotypes via an intermediate (E/M) state. (b) Wnt signaling is off during homeostasis, where β-catenin is degraded after phosphorylation. During active Wnt signaling, the complex of protein including APC are recruited to the membrane and β-catenin is no longer degraded. As a result, β-catenin is translocated to the nucleus where it will bind to TCF to activate Wnt target genes.

Collective migration and efficient metastasis formation

Although metastasis formation has been traditionally regarded as an event initiated by a single cell, evidence is accumulating in support of the view according to which cancer cells can invade, disseminate, and colonize collectively. The circulating tumor cell (CTC) cluster model in metastasis formation relies on the survival advantage of CTC clusters along the route to metastasis formation when compared to single CTCs. For example, CTCs clusters may avoid apoptotic programs due to loss of adhesion-dependent survival signals[21]. Moreover, cooperative efforts of cells in clusters could shield inner cells from immune assault or other toxic agent in circulation. It has also been hypothesized that heterotypic clusters, thus composed of different cell types, may benefit from different cellular functions[22]. Here below, selected examples in support of the CTC cluster model will be highlighted along the sequence of events in the invasion-metastasis cascade.

Experimental evidence has indicated that cells can invade in the tumor stroma in different migration modalities, separating single cell migration from multi-cellular strand formation and migration of cell clusters[23]. Collective migration is usually initiated by the mesenchymal-like tumor cells[24,25] and can be facilitated by cancer associated fibroblasts[26,27]. Clinical studies have suggested a link between clusters of tumor cells invading the stroma[28] and poor clinical outcome. Thus, the presence of de-attached cell clusters might result in more CTC clusters in the circulatory system, which resembles a subsequent step along the invasion-metastasis cascade.

In the circulatory system, the presence of CTC clusters has been associated to worse clinical outcome[29], which suggest their presence makes a distinct contribution to metastasis formation. By intravenous injections in mice, it was shown that CTC clusters were more efficient than single cells in metastasis formation. Moreover, using a mouse with mixed GFP/mCherry mammary tumors, Aceto et al. classified metastases as derived from single CTCs (GFP or mCherry) or as derived from CTC-clusters when the metastasis showed expression of both markers. Following this approach and normalizing to the number of CTCs and CTC clusters in the blood of mice, they computed that CTC clusters can have up to 50 fold increased metastatic potential[30].

Subsequent steps along the invasion-metastasis cascade include extravasation and colonization. If CTC clusters are to colonize collectively, one may expect multi-clonal heterogeneity in the metastasis. Considering this reasoning, studies have focused on the heterogeneity of the metastases. For example, by performing deep sequencing analyses of primary tumors and metastases, it was shown that metastases were frequently composed of multiple clones[31], indicating polyclonal seeding from the primary tumor. Similar results were obtained from studies that created multi-colored primary tumors, and observed multi-colored metastases at distant organs[30,32,33.] Taken together, these results suggest that CTC clusters could form an efficient path to metastases formation.

An integrated model for phenotypic plasticity and collective migration

At first, integration of EMT and collective migration may sound counterintuitive, because EMT is expected to result in cells with poor adhesive features. These models appear especially irreconcilable when EMT is seen as binary switch distinguishing two distinct cellular populations[34], and when collective migration is defined with requirements for stable physical contacts throughout migration.

However, recent studies have shown that different intermediate EMT stages exist[35.] In particular, hybrid E/M sub-states were studied along a broad spectrum of intermediate phenotypes[36,37,38]. Moreover, hybrid E/M states have been associated to increased stemness[39], plasticity[40] and metastatic potential[38], suggesting a prevalent role in metastasis formation. Of note, while full EMT was found to be irreversible, partial EMT was shown to be reversible. A partial EMT state was obtained by

exposing cells to a low dose of hTGFß-1 (0.5 ng/mL), but this state reverted back to epithelial when cells were recultered in absence of stimulus[37]. Furthermore, it was shown that these hybrid E/M-states can be stabilized by phenotypic stability factors such as, indicating distinct phenotypes along the EMT spectrum[41,42].

In parallel, the definition of collective cell migration has loosened into the more complex notion of 'fellow travelers'[43], which can cooperate and differ in adhesion, signaling and force-dependent interactions. Interestingly, it was shown that even mesenchymal cells can achieve collective migration through mutual chemotaxis[44], indicating the varying degrees of interactions that can keep cells together.

These developments opened room for integrated models where partial EMT and collective migration are proposed as mutually beneficial[12,45,46]. In these models, it was suggested that partial EMT could facilitate dissemination of cell clusters from primary tumors, as cells acquire more migratory capacities while maintaining a certain degree of adhesion. Hence, the combination of epithelial and mesenchymal features would provide the integrity of CTC clusters and promote their migratory capacity[46].

However, the notion of an integrated model does not exclude mechanisms based on one of the models without dependency of the other. Partial EMT could also occur at the single cell level, and similarly, cell clusters could also be obtained by 'leader/follower' invasive modes with stromal cells, and hence excluding the need for EMT. Overall, this combined view distinguishes three mechanisms of metastasis formation (Figure 2): 1) an EMT-independent dissemination of heterotypic cell clusters; 2) a single cell dissemination through EMT and MET and 3) collective cell migration facilitated by partial EMT.

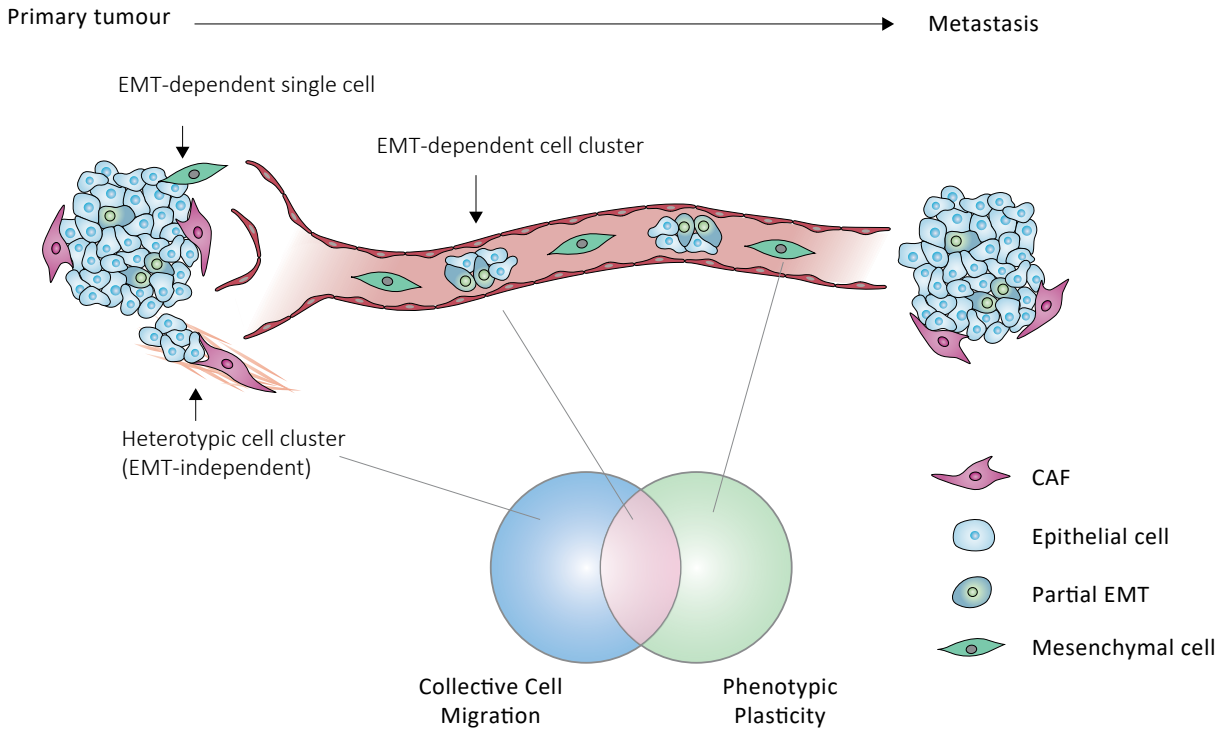


Figure 2. EMT and collective migration as complementary mechanisms. Schematic overview of the integrated model based on Nieto et al. (2017). Metastasis can proceed following an EMT-dependent manner, either via single cell or collective migration. Alternatively, metastasis can proceed according to an EMT-independent manner based on the heterotypic cell clusters.

Research objective and approach

The present study aims to contribute to the ongoing debate on the role of phenotypic plasticity and collective migration in cancer metastasis. The main research question is:

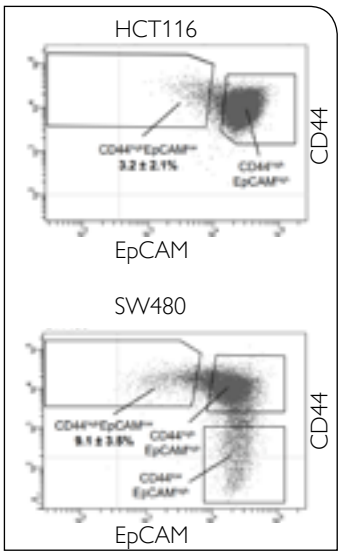
Are phenotypic plasticity and collective migration complementary and functionally correlated mechanisms during cancer metastasis?

We believe this is a relevant question in need of further research for different reasons. First, clarification of this question can influence clinical management of late stage cancer patients. If EMT can be held responsible for the modus operandi of tumor invasion, there is a potential to interfere with invasion mechanisms to hamper or delay a crucial step in the process of metastasis formation. Alternatively, if these mechanisms act independently, it is plausible that in specific cancer types or patient groups one mechanism is prevalent. Ultimately, this could be exploited to stratify patients, and adjust treatment decisions depending on the dominating mode of invasion. Second, from a scientific point of view, answering this question may result in crucial insights in the process of metastasis formation. Traditionally, major focus has been on stem cells, and their unique properties allowing them to overcome hurdles along the way to metastasis formation. However, if collective cell clusters can indeed benefit from an organized EMT hierarchy, we may have underestimated the complexity by which cancer spreads. To unravel this process at single cell resolution and elucidate whether migrating cancer cells overcome barriers by coordinated cooperation will substantially contribute to our understanding of metastasis.

To address this question, we will focus on the assumption that collective cell migration results from partial EMT. From this perspective, three sub-questions arise with distinct approaches for investigation (Figure 3). One of the key assumptions of the integrated model is that EMT progresses through intermediate E/M states which enable cells to simultaneously encompass both mesenchymal and epithelial characteristics. Until now, E/M sub-states have been identified based on expression of few EMT markers or based on comparisons between pools of cells. Here, we aim to contribute to the characterization of the EMT spectrum by comparing complete transcriptomes of single cells (scRNAseq).

1. *How is EMT controlled in colorectal cancer cell lines, and can partial EMT cells be identified based on single cell transcriptomes?*

To answer this question, we will take advantage of earlier work in the laboratory aimed at the characterization of $CD44^{high}EpCAM^{low}$ cells in comparison to $CD44^{high}EpCAM^{high}$ cells in conventional immortalized colon cancer cell lines (HCT116, SW480, SW620). The labels high and low refer to predefined gates used in the fluorescent activated cell sorter (FACS) that have been set to distinguish a mesenchymal ($EpCAM^{low}$) from an epithelial ($EpCAM^{high}$) state, both in presence of stem-like characteristics ($CD44^{high}$). To address the distinctive functional features of these distinct subpopulations, we showed that $CD44^{high}EpCAM^{low}$ cells are EMT-competent, invasive, chemo resistant, and highly metastatic in vivo (Box 1). Furthermore, $CD44^{high}EpCAM^{low}$ are plastic, as shown by their ability to reconstruct the cellular composition of the parental cell line. As such, these cell lines represent an appropriate model to study the dynamic and reversible EMT process at high resolution. We perform single-cell RNA sequencing (scRNA-seq) to investigate the EMT state across and within subpopulations of $CD44^{high}EpCAM^{low}$ cells and $CD44^{high}EpCAM^{high}$ cells.



Box 1. **Previous work in the lab showed that[1] ...**

Colorectal cancer cell lines encompass two distinct populations: CD44^{high}EpCAM^{high} cells and CD44^{high}EpCAM^{low} cells. When compared to CD44^{high}EpCAM^{high} cells, CD44^{high}EpCAM^{low} cells are …

- **highly motile**
  An assay based on the differential ability to migrate through a membrane, showed a significantly increased migratory (HCT116: P< 0.05; SW480: P < 0.05) ability of the CD44^{high}EpCAM^{low} cells in both colon cancer cell lines when compared with CD44^{high}EpCAM^{high} cells

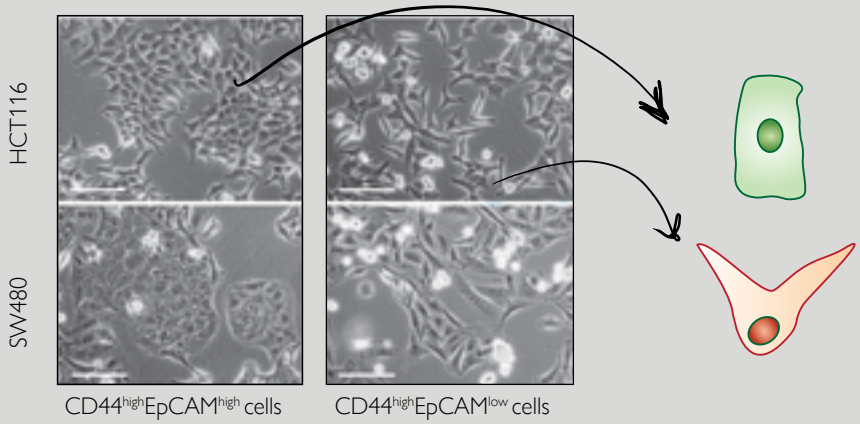
- **invasive**
  An assay based on the differential ability to migrate through an extracellular matrix (collagen)-coated membrane, showed a significantly increased invasive (HCT116: P< 0.05; SW480: P < 0.05) ability of the CD44^{high}EpCAM^{low} cells in both colon cancer cell lines when compared with CD44^{high}EpCAM^{high} cells.

- **chemoresistant**
  Oxaliplatin and 5-fluorouracil, among the most commonly employed chemotherapeutic drugs in the treatment of colon cancer, preferentially affect non-EMT cells while the EMT-competent and stem-like CD44^{high}EpCAM^{low} cells are resistant.

- **metastatic**
  In both HCT116 and SW480, injection of CD44^{high}EpCAM^{low} cells resulted in significantly more liver metastases than with CD44^{high}EpCAM^{high} cells. Notably, IHC analysis of the liver metastases revealed a heterogeneous pattern of intracellular β-catenin, with membranous and cytoplasmic localization in cells from within the center of the lesion, and nuclear β-catenin accumulation in cells localized in the periphery, thus recapitulating the situation in primary colon carcinomas.

- **phenotypically plastic**
  To investigate the stem-like properties of CD44^{high}EpCAM^{low} colon cancer cells, their capacity to differentiate into more epithelial cell types and reconstitute the heterogeneous composition of the parental cell lines was investigated. CD44^{high}EpCAM^{low} and CD44^{high}EpCAM^{high} cells from HCT116 and SW480 were sorted and grown separately under conventional culture conditions, and analyzed by FACS at different time points. After 1-3 months in culture, both subpopulations were capable of re-establishing the complex and heterogeneous composition of the parental cell lines.

- **predictive for CMS4 subtype**
  Expression of the RNA seq signatures derived from CD44^{high}EpCAM^{low} cells in HCT116 and SW480 correlate with the consensus molecular subtype 4 of human colon cancers, which has the greatest propensity to form metastases.



[1] Sacchetti, A., Teeuwssen, M., et al. Phenotypic plasticity of CD44^{high}EpCAM^{low} cells underlies local invasion and distant metastasis in colon cancer. *(In preparation)*

Next, we aim to investigate the effect of EMT on the tumor invasion modalities. It is widely accepted that a mesenchymal cell has enhanced migratory capacity when compared to its epithelial equivalent. However, whether induction of EMT will alter the modalities of tumor invasion in a dosage-dependent fashion still remains elusive. Our aim is to develop in vitro models that enable comparison of tumor migration strategies. Hence, the sub-question:

2. *Will ectopic expression of the EMT-TF ZEB1 affect the CTC migration modality?*

To date, most of the published work on tumor invasion has either been based on observational data from paraffin embedded sections of patient-derived cancers, or on in vitro (e.g. collagen; 3D vs. 2D) models taking advantage of immortalized cell lines, tumor-derived organoids or even freshly resected primary cancers. While in vitro models enable accurate spatiotemporal monitoring of migration mechanisms, they have poor representation of the in vivo situation. In our approach, we have combined in vitro models (i.e. collagen scaffolds) to study tumor migration *ex vivo*, balancing the level of control with the complexity found at tumors *in vivo*.

To this aim, we have employed mouse intestinal organoids carrying mutations (in *Apc, Kras, Tp53* genes) matching the later stages of the adenoma-carcinoma sequence, which can be transplanted orthotopically in the mouse caecum to develop invasive tumors in vivo. The corresponding organoid-derived tumors can be employed in ex vivo migration experiments in 3D collagen models. Previous studies have shown that ZEB1 is a crucial EMT inducer in many cancer types. Here, we have taken advantage of inducible *ZEB1* overexpression vectors to elicit the complete EMT program in the organoid-derived tumors and compare the migration strategies of these tumors to the ones without the EMT stimulus.

Finally, we have focused on the subsequent step in the invasion-metastasis cascade, namely the survival of CTCs in circulation. The main hypothesis here is that collective migration results from partial EMT and that CTC clusters encompassing E/M cells have high phenotypic plasticity. In order to address this issue we question:

3. *Do circulating tumor cell clusters consist of E/M hybrid cells, and do they show an organized or stochastic arrangement of EMT-related inter-cellular heterogeneity?*

Attempts will be made to isolate CTC clusters from human patients and mouse models of metastatic colon cancer. In the past decades, CTCs have been isolated with increasing efficiency from the peripheral blood of patients affected by various cancer types. Some of the newly developed methods have even been standardized and commercialized, such as CellSearch and VyCAP. While these method have delivered valuable insights from liquid biopsies, they suffers from a number of pitfalls. The use of EpCAM as positive isolation marker has its limitations given that quasi-mesenchymal cells are expected to have low expression of this marker. Furthermore, these methods rely on single cells that are punched across microwells, which hampers detection of CTC clusters. More recently, novel microfluidic devices have been developed, such as the HB-chip[47] and its successor[48], which have been successfully employed to capture CTCs and CTC clusters. In contrast to these positive selection methods, we have mainly relied on negative depletion of white and red blood cells as purification step of the liquid biopsy.

Thesis structure

The remainder of this thesis will be structured as following. Chapter 2 describes the procedures and materials used for the experiments. In Chapter 3, results will be presented from the three projects described above. Finally, in Chapter 4, the results from the projects will be discussed to debate the research question and propose avenues for further research.

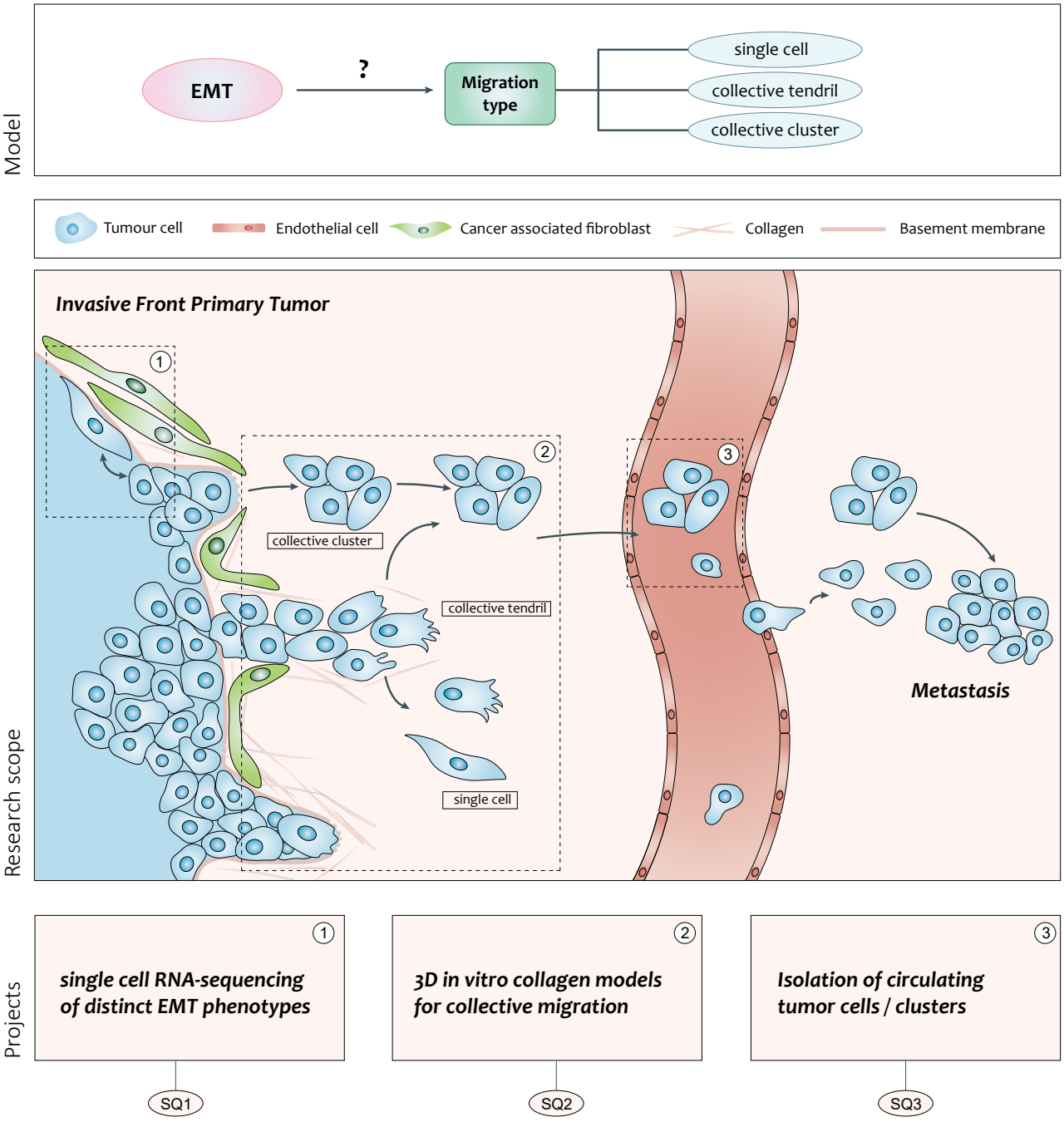


Figure 3. Schematic diagram of research model and scope. Figure based on Stuelten et al. (2018).

Chapter 2. Methods

Cell culture

Cell lines

Human colon cancer cell lines (ATCC) employed in this thesis are listed in Table 1. Cells were cultured in Dulbecco's Modified Eagle medium (DMEM, Gibco) with 10% fetal calf serum (FCS, Gibco), 100 U/mL penicillin (Invitrogen) and 100 µg/mL streptomycin (Invitrogen), 2mM L-Glutamine (Invitrogen). When the cells reached 80% confluency, cells were washed in PBS and trypsinized for 2 min. at 37C°. Subsequently, cells were passaged at 1:10 and incubated in a humidified atmosphere at 37C° and 5% CO_2. Medium was replaced once every 2/3 days. For experiments, cells were trypsinized at 80% confluency and counted using the Fuchs-Rosenthal counting chamber. Where indicated, cell aggregates were produced by incubating single cells in low-attachment plates for 3 days at 37C° and 5% CO_2.

Table 1. Overview of cell lines employed in this study.

Label	Type	Disease origin
HCT116	human colon cancer cell line	Colorectal carcinoma (primary)
SW480	human colon cancer cell line	Dukes' type B, colorectal adenocarcinoma (primary)
SW620	human colon cancer cell line	Dukes' type C, colorectal adenocarcinoma (lymph node metastasis)
HT29	human colon cancer cell line	Colorectal adenocarcinoma
HCT5.3	Human immortalized cancer associated fibroblasts (CAFs)	

Inactivation of CAFs

From a confluent layer of immortalized HCT5.3 cancer associated fibroblasts (CAFs), medium was removed and replaced with 10 µg/mL Mitomycin C in DMEM FCS for 2 hrs. in order to mitotically inactivate the cells. Next, cells were washed for 3 times in PBS. Trypsin was added for 2 min at 37C° to resuspend the cells. Hereafter, cells were counted using the Fuchs-Rosenthal chamber and frozen at $1.0x10^6$ cells/vial until use for subsequent experiments.

Organoids

In this thesis, the label organoids is used to refer to cells in spheroids, grape-like or dense aggregates, and aggregates with luminal organization. The different organoids used in this MSc thesis are listed in Table 2. Organoids were cultured in different media depending on their origin (Table 3). CSC08, DN08, G605, L145 and CRC48 were cultured in low attachment flasks and passaged 1:10 every 3 days. The patient-derived CMS4 organoids a nd the mouse AK/AKP organoids were incubated in droplets of 30-50 µl Matrigel. Matrigel was dissolved by adding Cell Recovery Solution (Corning) to the wells followed by 20 min. on ice. Subsequently, the organoids were resuspended and washed in 5 mL DMEM 10% FCS. Organoids were spun down at 1000 rpm for 3 min. and supernatant was removed before pipetting up and down 30-50 times with a p200 pipette to reduce the size of the organoids. Depending on the structure of the organoids (e.g. DN08), trypsin was added for 2 min. at 37C° to further dissociate the organoids. After dissociation, cells were passaged 1:10, resuspended in cold Matrigel and plated in droplets. Matrigel was polymerized at 37C° for 20 minutes before medium was added. Depending on the organoid type, medium was refreshed every 2-4 days, and organoids were passaged every 3-7 days.

Table 2. List of organoids employed in this study and their origin.

Label	Description	References	Medium
AK mOrg	Intestinal mouse organoids, mutant in *Apc* and *Kras*. (*Apc*$^{fl/fl}$::*Kras*$^{G12D/+}$)	49	A
AKP mOrg	Intestinal mouse organoids, mutant in *Apc, Kras*, and *Tp53* (*Apc*$^{fl/fl}$::*Kras*$^{G12D/+}$::*Trp53*$^{fl/R172H}$)	49	A
AKP-Zeb1 mOrg	Intestinal mouse organoids with APC and KRAS and P53 mutations, and conditional ZEB1 overexpression	Our lab	A
AKP-Zeb1-GFP mOrg	Intestinal mouse organoids with APC and KRAS and P53 mutations, conditional ZEB1 overexpression, and constitutive GFP expression	Our lab	A
TOR8	Human organoids derived from a CMS4 colon cancer patient	50	B
TOR9	Human organoids derived from a CMS4 colon cancer patient	50	B
TOR10	Human organoids derived from a CMS4 colon cancer patient	50	B
CSC08	Human organoids derived from a colon cancer patient	-	C
DN08	Human organoids derived from a colon cancer patient	51	C
G605	Human organoids derived from a colon cancer patient	52	C
CRC48	Human organoids derived from a colon cancer patient	53	C
L145	Human organoids derived from a colon cancer patient	53	C

Table 3. Composition of the culture media used in this study.

Medium A	Medium B	Medium C
DMEM/F12	DMEM/F12	1:7.5 H20
1:50 B27	1:50 B27	1:1.7 DMEM/F12
1:100 N2	1:1000 SB202180	1% Glucose
1:8 mL N	1:10000 A8301	0.023% Sodium Bicarbonate
1:2000 EGF	1:167 NAC	1 mM Hepes
	1:1000 Y	2 mM Glutamine
		3.5 nM Heparin
		60 nM BSA
		0.3 nM b-FGF
		3.0 nM EGF

FACS

For live cell sorting, cells were trypsinized for 2 min. at 37C°, washed and were stained with fluorescent antibodies diluted in PBS with 4% FCS for 30 min. on ice. After two washes in PBS with 4% FCS, DAPI (Sigma) was added at 1 µg/mL to distinguish dead from live cells. Samples were sorted in PBS with 4% FCS. Sorting and analysis was performed with a BD FACSAria III machine (BD Biosciences).

Immunohistochemistry (IHC)

4 µm sections were cut from formalin-fixed paraffin embedded tissues and deparaffinized Xylene at RT twice for 5 min.. Following hydration of the tissue by subsequent steps of 100%, 70% and 0% ethanol in water, antigen retrieval was performed by heating sections with the pressure cooker in either in 0.01M Tris or 0.001M EGTA depending on the antibody. After two washes in PBS-Tween, sections were incubated for 10 min. at RT in 3% peroxide to block endogenous peroxidase. Sections were washed twice in PBS-tween before incubation in 5% milk in PBS for 30 min. at RT. Cells were then washed twice in PBS-Tween and incubated overnight at 4°C with the primary antibody diluted in 5% milk PBS. After 2 washes in PBS-tween, sections were counterstained with polymer-HRP for 30 min at RT. Following two washes in PBS-tween, sections were stained by DAB in 1 mL DAB

substrate, washed in water, stained in hematoxylin for 10 sec. and washed again in water. Following dehydration by subsequent steps of 70% and 100% ethanol in water, sections were incubated 2x 5min. in Xylene and dried in Pertex with coverslip.

Immuno-fluorescence (IF)

Glass slides containing 4 µm thick tissue sections were circled with PAP Pen to earmark the samples. Next, samples were permeabilized by 1x Triton (0.25%) in PBS for 20 min at RT before being washed twice with PBS-Tween. Slides were blocked with 5% BSA in PBS-Tween for 30 min. at RT, to be then incubated overnight at 4˚C in primary antibodies in 5% BSA PBST. The next day, samples were washed twice in PBS-Tween, and secondary antibodies were added together with DAPI to stain the cells. Last, 1 droplet of VectaShield was added to the samples and slices were covered with glass slides. Prior to microscopy, slices were incubated for 30 min. at 37˚C to solidify the VectaShield and stabilize the coverslip.

Protein lysates and western blot

For western analysis, cultured cells (including organoids) were washed once in PBS and recovered from Matrigel using Cell Recovery Solution (Corning). Whole protein lysates were made in Laemli buffer, and boiled for 5 min. before gel separation using SDS-Page (SDS-Polyacrylamide Gel Electrophoresis). For each sample, 35 µl aliquots of the lysates were run at 100-150V and 15-20mA for 1.5 h. After separation, proteins were blotted to Polyscreen PVDF transfer membranes using a transfer apparatus according to the manufacturer's protocols (Bio-Rad). The membrane was then blocked in 5% milk in PBS, washed in PBS-Tween, and stained with the primary antibody diluted in 3% BSA-PBS overnight at 4˚C. Membranes were washed twice in PBS-Tween and incubated for 1h with diluted secondary antibody at RT.

RNA extraction and qPCR

RNA was extracted by resuspending the dissociated organoids or cells in Trizol (15596018, ThermoFisher Scientific) according to the manufacturer's instructions. RNA concentration was measured with the Nanodrop spectrophotometer (Thermo Scientific) at O.D. 260 (quality controls OD_{260}/OD_{280} and OD_{260}/OD_{240} ratios). Next, cDNA was synthetized by reverse transcription from random primers using the high-capacity cDNA reverse transcription Kit (4368814, Life Technologies). Hereafter, 500 ng cDNA was diluted in primer- and fast SYBR Green Master Mix (Applied Biosystems). Plates were centrifuged for 1 min at 1000 rpm before the start of the qPCR program (Applied Biosystems 7500 Fast) where 40 thermal cycles repeated denaturation (95 ˚C) and annealing (60 ˚C). The obtained dCt values were normalized against GAPDH and the control sample to show relative differences.

scRNA-seq

For the single-cell RNA seq experiment, three colorectal cancer cell lines were used: HCT116, SW480 and SW620. These cell lines were used during the characterization of $CD44^{high}EpCAM^{low}$ cells, which formed the basis of this experiment (Box 1). HCT116 is a model for microsatellite instability (MIN) with a near-diploid genotype, and carries mutation in mismatch repair genes (*MSH6*). In contrast, SW480 is a model for chromosomal instability (CIN), resulting in aneuploidy. SW480 carries mutations in the *APC* gene, which causes high Wnt-signaling. SW620 is derived from the same patient as SW480, but from a lymph node metastasis. Cell lines were cultured up to 60-70% confluency before being collected (by trypsinization) for single cell RNAseq analysis. For each sample, between 50k-100k $CD44^{high}EpCAM^{low}$ and $CD44^{high}EpCAM^{high}$ cells were FACS sorted and processed using the 10X Genomics Chromium Single Cell Controller according to the manufacturer's instructions. Samples were deep-sequenced (Illumina) to a depth ranging 49k-65k reads/cells (Table 4).

Bioinformatics

Bulk RNA seq

The gene-sample matrix containing four biological replicates for each cell line was imported in R for analysis. Reads were converted to counts per million and filtered based on a minimal expression in four samples using edgeR package[54]. The limma package was used to perform multidimensional scaling in scatter plots that approximate the typical log2 fold change between samples[55]. Heatmaps were made using stats package and gplots package[56]. Signatures were produced with edgeR differential expression analysis, setting absolute value of logFC > 1, and p.val < 0.01. Pathway analysis was performed using Enrichr based on KEGG 2019 Human pathways[57].

Single cell RNAseq

Gene-cell matrices were obtained by conversion of the raw data using the Cell Ranger pipeline[58]. Loupe cell browser was used to explore the samples and perform general quality checks including UMI distribution, reads per cell and sequencing saturation. Filtered gene-cell matrices were merged in R using the Seurat package[59]. Dimension reduction was performed using PCA and tSNE or UMAP[60]. CMS classification was done using the CMScaller Package[61], and epithelial and mesenchymal scores were computed using the Rmagic[62] and Gsva[63] packages. Heatmaps were created using the Pheatmap package[64]. Pseudotime analysis was performed in R using SCORPIUS[65] and Monocle3[66] and in Python using Palantir[67]. Kaplan Meier curves were created in R2 bioinformatics server[68]. Cluster specific signatures were used as input for k-means clustering on different cohorts of colon cancer bulk RNA seq data coupled to survival. Pathway analysis was done in IPA (Qiagen)[69].

Table 4. Sequencing information per sample.

Sample	Reads/cell	Seq Sat (%)	Median genes	Number of cells
HCT116 $CD44^{high}EpCAM^{low}$	51891	40.5	3941	1637
HCT116 $CD44^{high}EpCAM^{high}$	65584	42.7	4435	1246
SW480 $CD44^{high}EpCAM^{low}$	48996	40.5	4105	1069
SW480 $CD44^{high}EpCAM^{high}$	51446	39.2	4082	1596
SW620 Bulk	50172	43.5	3588	6676

Collagen models

For the collagen models, distinct protocols were used. In brief, rat tail collagen type 1 (8-10 mg/mL) was resuspended in 5x neutralization buffer A at 1:5. When indicated, collagen was diluted in PBS before neutralization to vary the final collagen concentration between 4 mg/mL – 8 mg/mL. After resuspension in neutralization buffer A (Table 5), organoids were mixed in the collagen slurry and collagen was plated in 30-50 µl drops on a preheated 24-well dish. Collagen was polymerized at 37˚C for 20 minutes before medium was added to the droplets. Dishes were then incubated at 37˚C / 5% CO_2 in culture medium refreshed every other day for 3-7 days.

Alternatively, the protocol had an additional pre-polymerization step to enhance the thickness of collagen fibers and thus the stiffness of the gel. In brief, 340 µl low-density rat tail collagen type 1 (3.52 mg/mL, Corning) was diluted in Neutralization buffer B to obtain a final volume of 600 µl and a concentration of 2mg/mL collagen. The slurry was carefully mixed with a p1000 avoiding the formation of bubbles and then incubated on ice for 2 hours. Cells were resuspended in the dish and 30 µl droplets were plated on preheated 24-well plates. Plates were incubated for 15 min at 37˚C, before the collagen was immersed in 0.5 mL DMEM-FCS.

Table 5. Collagen neutralization buffers

Neutralization buffer A (5x)	Neutralization buffer B
50 mg/mL αMEM powder	1.73x PBS
2% (wt/vol) $NaHCO_3$	0.03N NaOH
0.1M HEPES	1:4.5 H20
H_2O	1:1.7 DMEM-FCS
1:2000 EGF	1:167 NAC

For the freezing model, a mold was designed that enables temperature conductance from a single point (Figure 4). Molds were 3D-printed in stainless steel (Shapeways). Rat tail collagen type 1 (1% wt.) was mixed with 0.05 M acetic acid at 4˚C, and centrifuged for 3 min. at 1000 rpm to remove air bubbles. Next, 400 µl of homogenized collagen slurry was gently pipetted in the molds and covered with a glass coverslip. The molds were placed in a -20˚C cooling shelf for 1.5 hours to ensure complete freezing of the collagen slurry. Subsequently, scaffolds were gently loosened from the mold and placed in a vacuum freezer overnight to sublimate the ice. The next day, scaffolds were fully immersed in crossing-linking solution containing 33 mM 1-ethyl3-(3-dimethylaminopropyl)-carbodiimide and 6 mM N-hyrdroxysuccinimide for 20 minutes at RT. Scaffolds were washed in 70% ethanol and stored at RT until use for subsequent experiments.

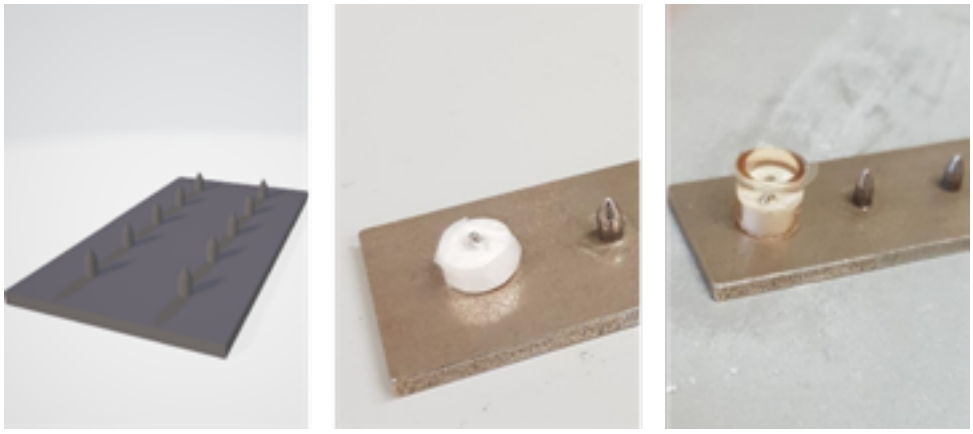


Figure 4. Fabrication of anisotropic collagen scaffolds.

Blood samples

Human blood sample was acquired from surgery. The sample containing 7 mL blood was deidentified prior to receipt and obtained with limited clinical information. Before the start of the experiment, the sample was maintained in EDTA tubes. Mouse blood was collected post-mortem using a syringe. In brief, red blood cells were lysed using 10 mL red blood cell lysis buffer containing 0.8% NH4Cl. Samples were rotated at 4˚C for 10 min and centrifuged for 3 min at 1000 rpm. Supernatant was removed and this procedure was repeated one more time. Following this, cells were stained in 200 µl with primary antibodies TER119 (1:100), CD31 (1:100), CD45 (1:200) and incubated for 10 min at 4C˚. Following this, samples were washed with 1 mL 2% PBS FCS and spun down at 1000 rpm to remove the supernatant. Hereafter, cells were resuspended in 1 mL PBS FCS with pre-washed 25 µl DynaBeads and incubated for 30 min at 4˚C. Magnetic beads were removed by carefully collecting the supernatant of the sample in a magnetic holder. Last, samples were fixed in 2% PFA and spotted on glass slides in 70% ethanol, to be used for IHC or IF.

Ex vivo migration assays

Human tumors

Human primary colorectal tumor or liver metastasis samples were acquired from the Pathology department after surgical resection. Each tumor sample was obtained anonymously with very limited clinical information. Before the start of the experiment, the tumor sample was maintained in Advanced DMEM/F12 (Gibco) at 4 ˚C for 1 hour. In brief, tumors samples were reduced in small pieces with a sterilized blade and washed three times in PBS. Following incubation in 300 µl collagenase diluted in 10 mL Advanced DMEM/F12 for 30 min at 37˚C, samples were resuspended in 50 mL Advanced DMEM/F12, and centrifuged for 5 min at 1000 rpm. Pellets were resuspended in 50 mL Advanced DMEM/F12 and either filtered using a 100 µm filter, or left for 30 min at RT to resolve the larger tumor fragments from the smaller ones by sedimentation. Subsequently, the supernatant was transferred to a different tube. Following this, the pellet, containing tumor fragments in the range of 100 µm, was resuspended in 0.5 mL Advanced DMEM/F12 in a FCS-coated low-binding Eppendorf tube. Samples were centrifuged for 5 min. at 1000 rpm and the pellet was resuspended in neutralized 4 mg/mL rat tail collagen type 1 before being plated in 15-30 µl droplets on pre-heated 24-well plates. After polymerization for 30 min at 37˚C, 500 µl of stem cell medium (1:10 N, 1:100 N2, 1:50 B27, 1:2000 EgF, diluted in Advanced DMEM/F12) was added to the wells. The culture medium was refreshed every other day until collagen droplets were fixed in 4% PFA and stored in at 4˚C in PBS for tissue processing and paraffin embedding.

Mouse tumors

AKP-ZEB1 mouse organoids were resuspended in 15 µl of neutralized high-concentration (8-10 mg/ mL) collagen type 1. Subsequently, collagen droplets were transplanted in the caecum of recipient mice according to procedures described elsewhere[70]. After 5-6 weeks, mice were euthanized by CO_2, and tumors were resected from the caecum and washed in PBS. Small fragments (+/- 1 mm^3) were cut from the tumor and placed on preheated (37˚C) collagen freeze scaffolds in medium (DMEM-FCS, gentamicin). Medium supplemented with doxycycline (1 µg/mL) was replaced every other day until samples were fixed in 4% PFA, and stored at 4˚C in PBS until used for clearing.

Clearing of collagen scaffolds

To facilitate the study migration behavior in 3D, a clearing protocol was used according to described procedures[71]. In brief, fixed scaffolds were dehydrated in methanol PBS gradients (1:0, 0.50:0.50, 0.25:0.75, 0:1 PBS in methanol), and blocked overnight in 5% H_2O_2 and 20% DMSO in methanol at 4˚C. Scaffolds were washed three times in methanol, and rehydrated up to 100% PBS before overnight incubation in 0.3M Glycine, 0.2% Triton X100 and 20% DMSO in PBS at 37˚C. Hereafter, scaffolds were blocked overnight in 3% milk, 0.2% Triton X100 and 20% DMSO in PBS at 37˚C. Following a 24 hrs. wash in PTwH (5% DMSO, 0.2% Tween-20 and 10 µg/ml Heparin in PBS) at 37 ˚C, scaffolds were stained with primary antibodies (ß-catenin, E-cadherin, α-SMA) diluted in PTwH-D-M (0.2% Tween, 5% DMSO, 10 µg/ml Heparin and 3% milk in PBS) and incubated for 3-5 days at 37 ˚C. This was followed by additional washes in PTwH for 24 hrs. and by incubation of secondary antibodies (anti-mouse 647, Anti-rabbit 647, Anti-mouse 488) diluted 1:200 and DAPI in PTwH-D-M for 1-4 days at 37 ˚C. When indicated, collagen was stained with 1 mg/mL Picro Sirius Red in picric acid (2,4,6-trinitrofenol, $C_6H_3N_3O_7$.) for 2h at RT and washed two times in 0.1% acetic acid. Subsequently, scaffolds were dehydrated again in methanol and incubated in 50% methanol 50% BABB (67% benzyl-benzoate, 33% benzyl-alcohol) for 20 min at RT and stored in 100% BABB at 4˚C until use for microscopy.

Microscopy

Fluorescent microscopy on glass slides was performed at 20X using a Zeiss LSM-700 confocal microscope. Collagen fibers were visualized by refraction imaging at 40X and 60X using a Zeiss LSM-880 confocal microscope. For 3D imaging, two approaches were used at 20x magnification: upright microscopy (Leica SP5 intravital) and inverted microscopy (Opera Phenix HCS system). Samples were fully immersed in BABB and mounted on a glass chamber. For the Leica SP5, freezing scaffolds were supported by the extremities of pipette tips and covered with glass slides and droplet of water. For the Opera Phenix HCS, a droplet of water was covered with glass slide before scaffolds were placed with the scaffold seeding area facing the glass slide. Time lapse imaging was done using the Opera Phenix HCS by measuring endogenous GFP present in AKP-*Zeb1*-GFP mouse organoids. In brief, organoids were seeded on fragments of collagen freezing scaffolds and incubated in DMEM-FCS at 37˚C. After 48 hrs., scaffolds were inverted and attached on glass by polymerization of a thin Matrigel coating for 10 min at 37˚C. In each well, 200 µl of DMEM-FCS and Dox (1 µg/mL) were added. Samples were imaged every 2 hours for a total duration of 48 hours.

Image analysis

Image analysis was done in ImageJ (version 1.52n). Hyperstacks were imported in Fiji and merged into multi-channel z-stack with corresponding colors. For 2D visualizations, contrast was adjusted for each channel for the specific z-slice. Multiple z-slices were aggregated using z-projection with either maximum or medium intensity as projection type. In order to obtain 3D visualization, 3D projection was used with interpolation and projection based on brightest point. Multichannel overviews were created using the 3D viewer plugin. Hyperstacks acquired with the Opera Phenix HCS were analyzed using accompanied Harmony 4.9 software (PerkinElmer). Overviews were presented as global max z-projections, and smaller areas were investigated with 3D projection or local max z-projection.

Chapter 3. Results

In this chapter, the experimental results of my MSc internship will be presented. The results are relative to 1. the *in silico* analysis of transcriptional heterogeneity of EMT-competent cells in cancer cell lines, and 2. the development and optimization of 3D collagen models for the *in vitro* and *ex vivo* study of collective vs. single cell migration in cancer invasion and dissemination. These models and the preliminary results obtained to date will form the basis for future experiments aimed at the elucidation of the role of EMT transcription factor ZEB1 in tumor invasion mechanisms. The results also include my efforts towards the isolation of circulating tumor cell clusters from liquid biopsies.

Resolving inter-cellular transcriptional heterogeneity along the EMT spectrum

To zoom in on the transcriptional heterogeneity in EMT, we have chosen as experimental model the subpopulations of colon cancer cell lines previously characterized for their EMT status and metastatic capacity (see Box 1; Chapter 1). Starting from bulk RNA sequencing, the major differences in gene expression among these subpopulations and their specific expression signatures have been determined. Given that the EMT-competent subpopulations are phenotypically plastic, i.e. they can convert to one another and reconstruct the heterogeneity of the parental cell line, it is likely that individual cells feature different sub-states reflecting this dynamic process. Single cell expression profiling is therefore necessary to dissect the heterogeneity within individual subpopulations and to identify intermediate E/M sub-states across the EMT spectrum.

Bulk RNA sequencing analysis of $CD44^{high}EpCAM^{low}$ and $CD44^{high}EpCAM^{high}$ populations

To investigate the gene expression differences between the CD44highEpCAMlow and $CD44^{high}EpCAM^{high}$ subpopulations, whole transcriptome RNA sequencing (RNAseq) was performed on samples sorted from the HCT116 and SW480 colon cancer cell lines by Illumina technology. For each of the HCT116 and SW480 cell lines, the distinct subpopulations were isolated by FACS resulting in 7 samples from which total RNA was isolated for sequencing purposes. The following were sorted: $CD44^{high}EpCAM^{low}$ cells (from here on referred to as "low"); $CD44^{high}EpCAM^{high}$ cells ("high"); $CD44^{low}EpCAM^{high}$ cells ("spheres"; specific for the SW480 cell line and consisting of cells which grow in suspension as small aggregates), and the "bulk" as control, i.e. the parental cell line processed and sorted by FACS in the same fashion as the subpopulation samples. RNAseq results were aggregated into a gene-sample matrix imported in R for analysis.

An overview of the variability across the samples was portrayed by performing multi-dimension scaling (MDS) as depicted in Figure 5A. MDS is an analysis technique that is based on principal component analysis (PCA) but includes an iterative algorithm which results in a more versatile mapping technique compared to PCA alone. As observed from the leading dimension (logFC dim1 axis) of the MDS plots, the variance between samples from different subpopulations was dominant over the variance between samples from the same subpopulation. Furthermore, the variance between cell lines was dominant over the variance across subpopulations. For HCT116, the bulk samples localize in between the low and high samples. For SW480, the bulk samples deviated strongly from both high and low samples though this is possibly due to the presence of the third subpopulation in this specific cell line, namely the $CD44^{low}EpCAM^{high}$ cells (spheres). Since these cells fell beyond the scope of our current study, we excluded this subpopulation in subsequent analyses.

We continued the analysis by focusing on the differences between the $CD44^{high}EpCAM^{low}$ and $CD44^{high}EpCAM^{high}$ subpopulations (Figure 5B). Differential expression analysis resulted in 228 and

465 differentially expressed genes for HCT116 and SW480, respectively (FDR < 0.01, abs(LogFC) > 1.0; Supplementary Table 1). The comparison performed with similar parameters between the spheres vs. the low and high SW480 subpopulations revealed 336 and 260 differentially expressed genes, respectively. To identify the genes with the highest expression difference between the subpopulations, we computed heatmaps showing the 15 most differentially expressed genes between high and low subpopulations. The top-15 differential expressed genes between high and low populations included both mesenchymal (e.g. *COL13A1, CDH11*) and epithelial (e.g. *EPCAM, KRT13*) genes, suggesting distinct EMT phenotypes in these subpopulations.

Next, we investigated the overlap among the different gene expression signatures of the various subpopulations (Figure 5C). To narrow down the list of genes characterizing high and low differences, we computed the intersection of three comparisons: 1) high vs. low in HCT116, 2) high vs. low in SW480, and 3) combined high vs. low from both cell lines. This analysis revealed 32 genes which most characterize the differences between the low and high populations (Supplementary Table 2), and included EMT-TF *ZEB1*, as well as epithelial genes, such as *EPCAM* and *CDH1*.

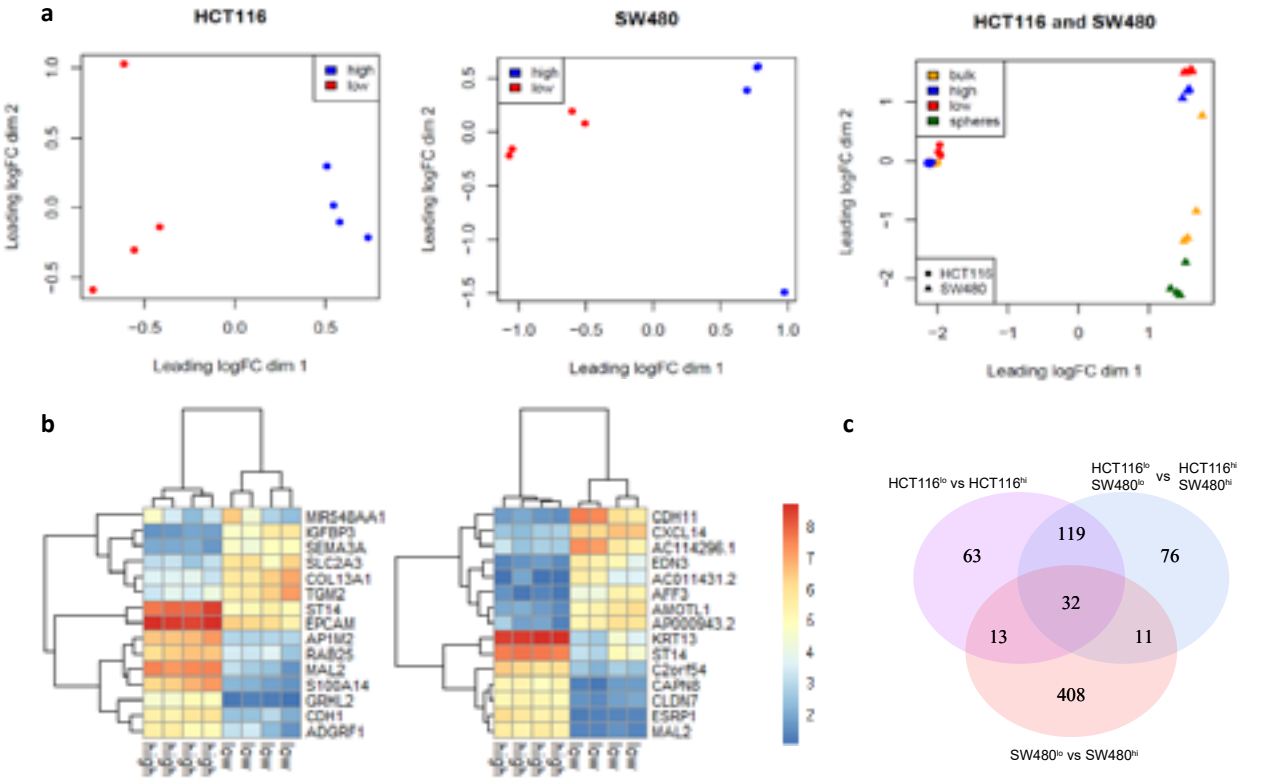


Figure 5. Bulk RNA sequencing analysis of CD44^{high}EpCAM^{low} versus CD44^{high}EpCAM^{high} populations. (a) multi dimension scaling plots of samples. (b) heatmaps showing top 15 DE genes between two subpopulations. (c) Venn diagram denoting the overlapping genes of DE analysis using abs(logFC) > 1.5, p.val < 0.01.

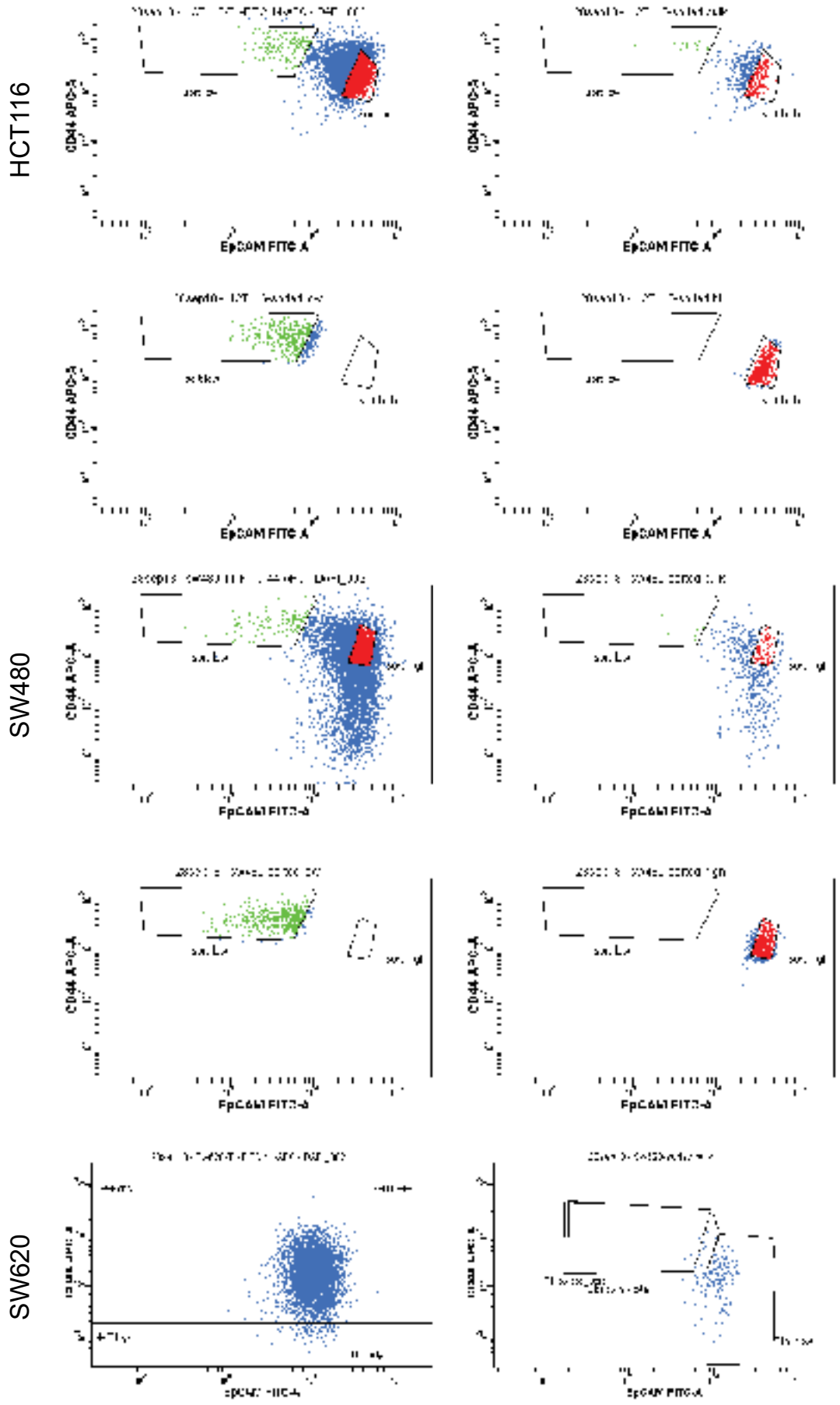


Figure 6. FACS plots showing the applied gating strategy for the isolation of subpopulations in HCT116, SW480 and SW620.

Single cell RNA sequencing of CD44^{high}EpCAM^{low} and CD44^{high}EpCAM^{low} populations

We performed single cell RNA sequencing using the chromium controller (10X Genomics) on FACS sorted subpopulations of CD44^{high}EpCAM^{low} cells and CD44^{high}EpCAM^{high} cells from HCT116 and SW480, together with the bulk population of SW620 (Figure 6). We analyzed >1000 cells for each subpopulation, and sequenced to a depth of approx. 50000 reads per cell with the MiSeq System (Illumina). Sequencing distributions appeared in accordance with good practice, evidenced by the knee-elbow plots of the samples (Supplementary Figure 1). Data was processed using the CellRanger Pipeline and imported in R for downstream analysis.

To compare variable expressed genes across subpopulations, we used the FindVariableGenes function in Seurat (`parameters: LogVMR, bottom cutoff average expression = 0.1, top cutoff average expression = 8, bottom cutoff dispersion = 1, top cutoff dispersion = Inf, num.bin = 20`). We detected 1717 and 2178 variable expressed genes across the CD44^{high}EpCAM^{low} cells compared with 1638 and 1783 genes across CD44^{high}EpCAM^{high} cells in the SW480 and HCT116 cell lines, respectively; instead, the bulk of SW620 showed 1019 variable expressed genes. Next, we compared differentially expressed genes between the CD44^{high}EpCAM^{low} and CD44^{high}EpCAM^{high} cells by performing differential expression analysis (`parameters: thresh.use = log(2), test: likelihood-ratio test for single cell gene expression`). This analysis revealed genes similar to those revealed by the bulk RNA-seq analysis (HCT116: V*IM, EPCAM, DKK1, RAB25, CLDN7, S100A14*; SW480: *WFDC2, LCN2, IFI27, CLDN7, ATP6AP1L*) as well as novel genes (HCT116, N = 53; SW480, N = 94) (Figure 7A). As validation, we computed box plots for *VIM* and *EPCAM*, and observed differential expression of these genes between CD44^{high}EpCAM^{low} and CD44^{high}EpCAM^{high} cells (HCT116: *VIM* $P < 0.001$, *EPCAM* $P < 0.001$; SW480: *VIM* $P < 0.001$, *EPCAM* $P < 0.001$)(Figure 7B).

To further investigate the heterogeneity within the CD44^{high}EpCAM^{low} cells and CD44^{high}EpCAM^{high} cells, we performed unsupervised clustering using shared neighbor (SSN) modularity optimization, which is a powerful clustering method that overcomes problems associated to finding clusters with different densities[72]. After dimension reduction using the first 30 principal components and computation of tSNE (Barnes-Hut implementation), the CD44^{high}EpCAM^{low} cells and CD44^{high}EpCAM^{high} cells clustered in separate groups in HCT116 (Figure 7C). While HCT116 cells clearly separated in two distinct populations, SW480 cells showed a partial overlap between the CD44^{high}EpCAM^{low} and CD44^{high}EpCAM^{high} cells, with a distinct CD44^{high}EpCAM^{high} subpopulation (Figure 7D). Using a signature derived from our bulk RNA-seq experiments, we identified that this distinct population represented the SW480 non-adherent subpopulation ("spheres") (Supplementary Figure 2A). Because the study of this particular subpopulation was beyond the scopes of this thesis and because it interfered with the dimension reduction of the CD44^{high}EpCAM^{low} and CD44^{high}EpCAM^{high} cells, this cluster was removed from the subsequent analyses. When dimension reduction was performed on the SW480 bulk RNAseq gene signature, the two partially overlapping clusters were now clearly resolved. These results are indicative of the presence of EMT-related transcriptional variance between these subpopulations (Supplementary Figure 2B).

As a strong correlation between the CD44^{high}EpCAM^{low} cells and the CMS4 subtype was observed in the bulk RNA-seq experiments, we were interested to investigate this association at the single-cell level. To this aim, we used a previously described CMS classifier to allocate each cell to a consensus molecular subtype[61]. The CMS classifier uses filtered signatures, intrinsic to the cancer cells (i.e. not immune and stromal compartments), derived from the CMS subtypes based on public bulk RNA seq databases, to apply a nearest template prediction and allocate CMS labels to input samples. To cope

with the low-quality data of single cells compared to bulk RNAseq samples, the algorithm was forced to allocate labels independent of significance test. While this increases the chance of false allocations, the predictions can still be indicative at a subpopulation level. The CD44^{high}EpCAM^{low} cells were more often classified as CMS4 (HCT116: 38%, SW480: 35%) compared to the CD44^{high}EpCAM^{high} cells (HCT116: 9%, SW480: 17%)(Figure 7E). Notably, 28% of the SW620 bulk cells were classified as CMS4, suggesting that a large fraction of the SW620 cell line share a similar expression signature with the CD44^{high}EpCAM^{low} cells. Indeed, when aggregating the SW480 CD44^{high}EpCAM^{low} and CD44^{high}EpCAM^{high} cells with the SW620 bulk, the SW480 CD44^{high}EpCAM^{low} cells overlapped with the cells of SW620 that were classified as CMS4 (Supplementary Figure 3). Thus, even at the single cell level, the expression of CD44^{high}EpCAM^{low} cells associates with expression profiles of colon tumors from patients with poor clinical outcome. Accordingly, we derived signatures from clusters of CD44^{high}EpCAM^{low} cells by doing differential expression analysis (`parameters: thresh.use = log(2), ident.1 = EpCAM_low_cluster_i, ident.2 = EpCAM_high_cluster_i:n, test: likelihood-ratio test for single cell gene expression`), and showed that they can be used to stratify patient tumors according to survival and consensus molecular subtype (Supplementary Figure 4).

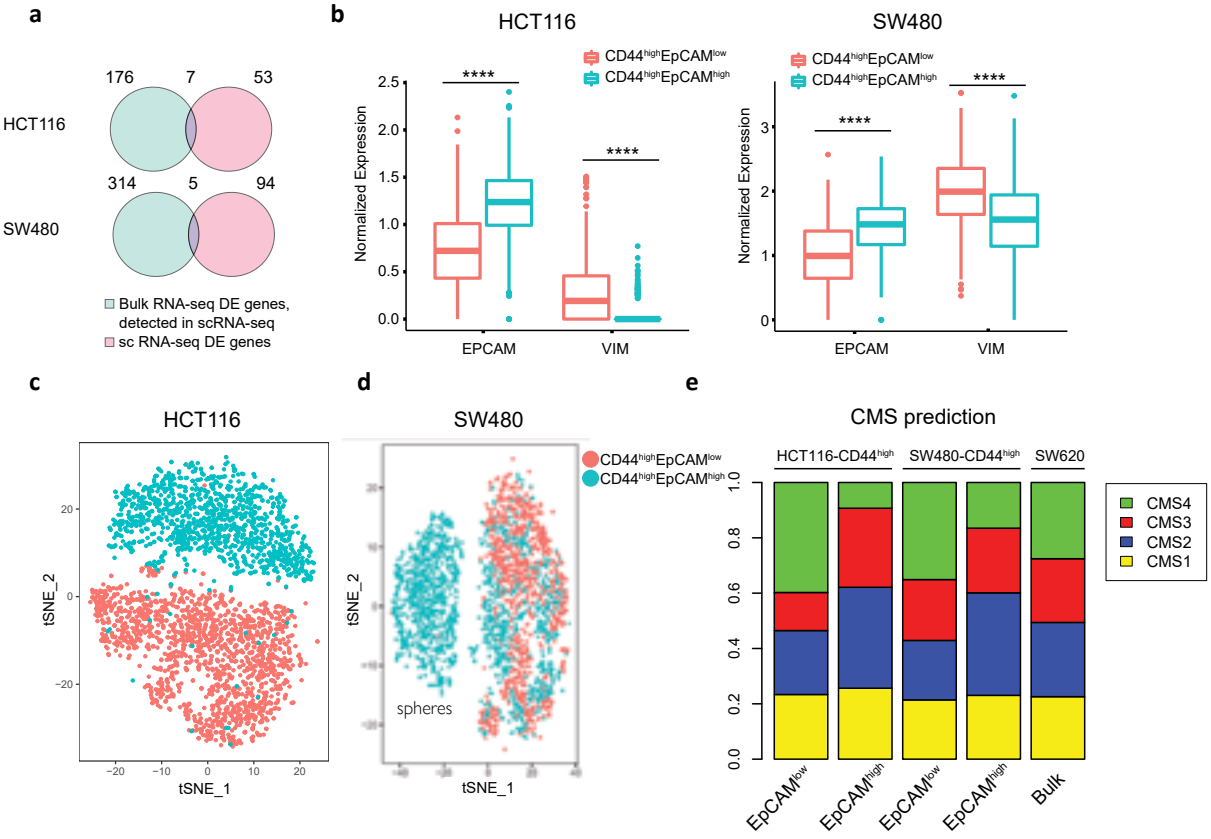


Figure 7. Overview of single cell RNA sequencing results. (a) Venn diagram denoting the number of DE genes between the subpopulations for scRNAseq versus bulk RNAseq. (b) Boxplots showing expression of *EPCAM* and *VIM* for both subpopulations in HCT116 and SW480. (c)(d) tSNE plots with annotation from FACS sorting. (e) CMS predictions for different subpopulations in HCT116, SW480 and SW620.

Competition on variance: cell cycle, apoptosis and batch effects

The overlap between SW480 CD44^{high}EpCAM^{high} cells and CD44^{high}EpCAM^{low} on the tSNE plot (Figure 7D) indicated the presence of strong EMT-independent transcriptional heterogeneity. Because, genes that do are not different between the low and high subpopulations appear to dominate in variance, resulting in a tSNE plot with overlapping populations. We first questioned whether the presence of the 'sphere' subpopulation (CD44^{low}EpCAM^{high} cells) interfered with the dimension

reduction. Given that cells from this subpopulation are so much different from the other cells, the principal components can be dominated by differences caused the sphere subpopulation, which would suppress more nuanced differences between $CD44^{high}EpCAM^{high}$ cells and $CD44^{high}EpCAM^{low}$ cells. Repeating the dimension reduction on SW480 without the spheres, however, did not reveal clear separation of the two subpopulations.

Next, we aimed to characterize the effect of other cellular programs on dimension reduction and unsupervised clustering results. To this aim, we compared the effect of corrections on the localization of cells on the tSNE plot of HCT116 while maintaining the cluster labels as provided by scaling without corrections. Differences in apoptotic states were compared by measuring the percentage of mitochondrial RNA expression. Cell cycle phase were estimated based on the expression of cycle-dependent genes[73], and cell doublets were assessed by comparing nUMI counts (because doublets are expected to yield a significantly higher number of detected RNA molecules). We used these proxies to investigate their effect on unsupervised clustering results using linear regression models (Figure 8). This procedure removes all signal associated with a transcriptional program by modeling the relationship with all genes to the score derived from a signature list and scaling the residuals in a corrected matrix. Since we observed strong association between the assigned cell-cycle phases and the unsupervised clustering results, we corrected for cell-cycle effects to assess the effect of cell cycle (Supplementary Figure 5). It appeared that, cluster 5 disappeared and distributed over the newly generated tSNE when performing this procedure, suggesting that this cluster was established by cell cycle differences. This procedure can be used to reveal EMT differences within subpopulations that remain hidden in presence of dominant cell cycle effects (Supplementary Figure 5). However, it should be noted that cell cycle can strongly be related to EMT[74]. Hence, diminishing cell cycle effects can also negatively impact downstream analysis as cell cycle correction can blur differences between EMT states. A way to overcome this limitation may be to look at cell cycle score differences (S.score – G2m.score): the signal from non-cycling and cycling cells is retained but differences among proliferating cells due to cell cycle will be reduced[73]. This diminished the relation between cluster 6 and cluster 7, but also caused the loss of cluster 5.

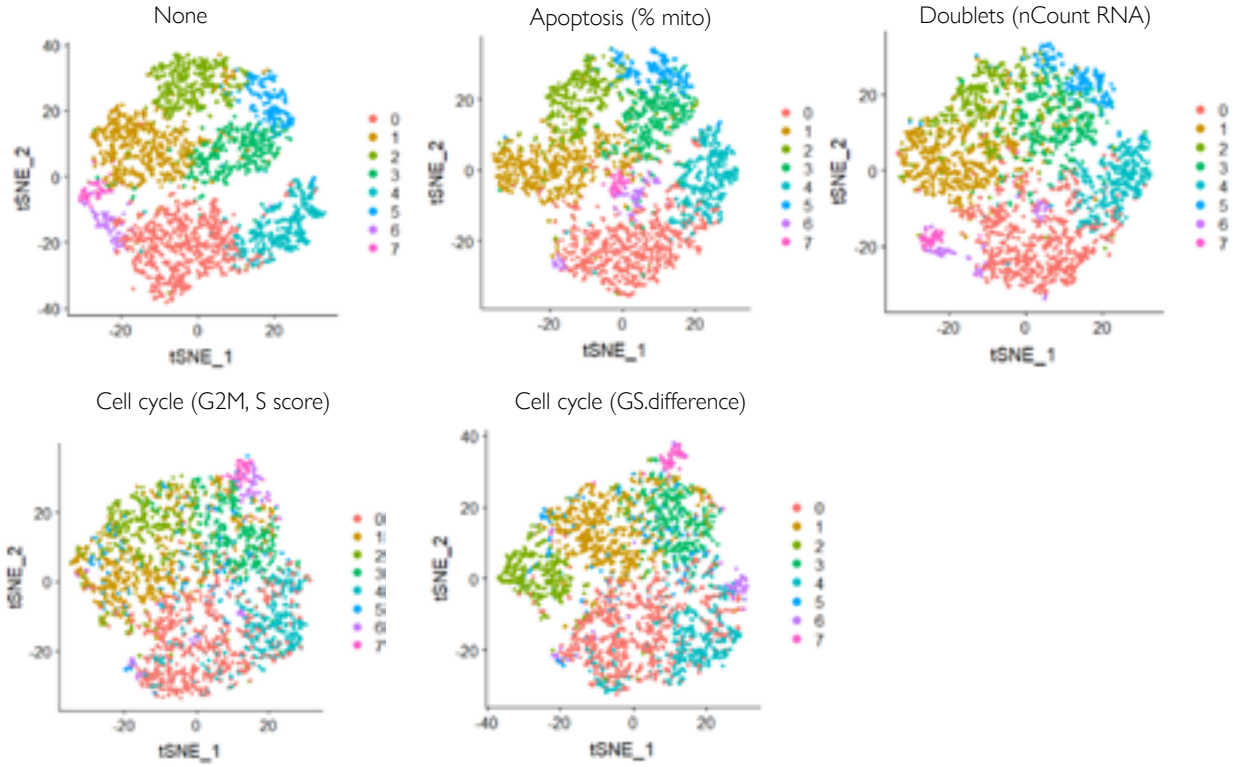


Figure 8. Effect of correction using linear regression models. Location of unsupervised cluster labels on the HCT116 tSNE plot after different types of corrections.

In view of the above, unsupervised clustering results appeared to be considerably robust after nUMI correction and apoptosis correction. Cell cycle effects play a role in the unsupervised clustering process, but correcting for this can also introduce bias. Because these effects did not explain the overlap of subpopulations in the samples of SW480, we concluded that these corrections were not imperative for the scaling procedure of our data. Therefore, in subsequent analyses, unsupervised clustering without correction is used as basis for subsequent steps.

A small cluster of HCT116 consists of E and M cells that associate with metastatic signature

From the tSNE plot of HCT116, it appears that cluster 7 encompasses cells originating from both $CD44^{high}EpCAM^{low}$ cells and $CD44^{high}EpCAM^{high}$ cells (Figure 9A). Because this E/M cluster may reflect a plastic state in between the other states, we went on to study this cluster in further depth.

Earlier, it was shown that the iso-clonal variants HCT116 and HCT116b, derived from the same primary tumor, have a significant difference in ability to form metastatic deposits *in vivo*[75]. Comparison of gene expression revealed striking difference in their gene signature, with 3587 genes being differentially expressed across the variants (1777 up and 1809 down in HCT116 compared to HCT116b)[76]. We questioned whether some cells of HCT116 have a gene signature shifted toward HCT116b, which could provide them with enhanced metastatic potential. From the 3587 differentially expressed genes, we identified 668 genes in the scRNA seq data (236 up and 432 down in HCT116b compared to HCT116). To investigate the association with the signature, a score was computed by subtracting the average scaled values of genes overexpressed in HCT116 from the set downregulated in HCT116 compared to HCT116b. Clusters containing cells from the $CD44^{high}EpCAM^{low}$ subpopulation, and especially cluster 6 and cluster 7, appeared to have most association with this signature (Figure 9B). To investigate the transcriptional differences of cluster 7 cells depending on the subpopulation origin, we dissected this cluster into two separate clusters, cluster 7 ($CD44^{high}EpCAM^{high}$ cells) and cluster 8 ($CD44^{high}EpCAM^{low}$ cells) (Figure 9C). Next, we computed a similar score from an expression signature previously obtained by profiling the invasive front of colon carcinomas (Supplementary table 3)[77]. Cluster 8 showed most association with this mesenchymal signature (Figure 9D).

Next, we questioned whether cells derived from cluster 8 can play a role in the observed plasticity of $CD44^{high}EpCAM^{low}$ cells. To this aim, we employed Palantir, an algorithm that models differentiation as a stochastic process where stem cells differentiate to terminally differentiated cells by a series of steps through a low dimensional phenotypic manifold, to align the cells along differentiation trajectories[67]. Cluster 8 localized at the corner of a tSNE in embedded space, and when appointed as approximate early cell, Palantir indicates cluster 8 as origin of a differentiation trajectory (Figure 9E).

Hence, cluster 8 shows characteristics of E/M cells and may have a role in the plasticity of the metastatic subpopulations of these colon cancer cell lines. Nevertheless, the cells from cluster 7 ($CD44^{high}EpCAM^{high}$) also show distinct expression of EMT genes, e.g. *EPCAM* and *VIM*. Therefore, we questioned whether these cells were truly E/M, or whether they are mesenchymal cells that cluster with epithelial cells because of other, EMT-independent, common features. To investigate this, we developed a computational approach to inspect the EMT-state for single cells and clusters.

Characterization of the EMT spectrum

While various EMT scores have been developed that can accurately distinguish EMT profiles from bulk RNA samples[78–80], these approaches cannot be reliably copied to single-cell RNAseq data. Most importantly, the technical noise and drop out, which is still an inherent feature of single cell RNAseq data, requires flexible approaches that do not rely too much on few key EMT markers.

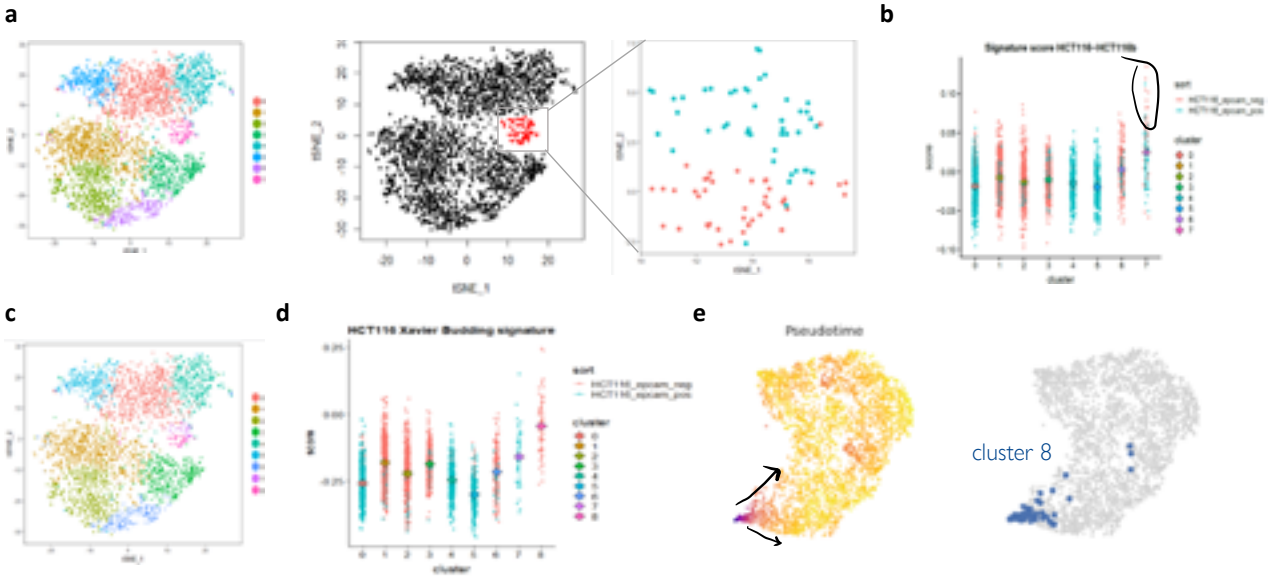


Figure 9. Characterization of cluster 7 in HCT116. (a) A small subpopulation clusters together, and contains cells from CD44^{high}EpCAM^{low} cells and CD44^{high}EpCAM^{high} cells. (b) Categorical scatter plot showing association between cluster 7 and HCT116-HCT116b signature. (c) tSNE plot after subdividing cluster 7 into cluster 7 (CD44^{high}EpCAM^{high}) and cluster 8 (CD44^{high}EpCAM^{low}). (d) Categorical scatter plot showing association with Xavier budding signature. (e) Palantir pseudo-time analysis pointing cluster 8 as the origin of a dynamic differentiation path.

Regression-based approaches based on ratio-metric values of key EMT regulators may be useful for stratifying distinct RNA pools, but are problematic at the single cell level.

Here, we have applied a mixed approach for the characterization of EMT sub-states using single-cell RNA sequencing data. Unsupervised clustering is used as input for two parallel computational approaches (Figure 10). Subsequently, three computational approaches are intertwined: 1) the mapping of single cells on epithelial and mesenchymal axes to obtain a general overview of the spectrum and assess the robustness of unsupervised clusters, and 2) the evaluation of cluster-specific differences across the input EMT genes to obtain cluster profiles and study genes with low expression, and 3) the ordering of cells along dynamic trajectories to study differentiation paths. Below, we will discuss our efforts on both approaches and use HCT116 as sample for comparison.

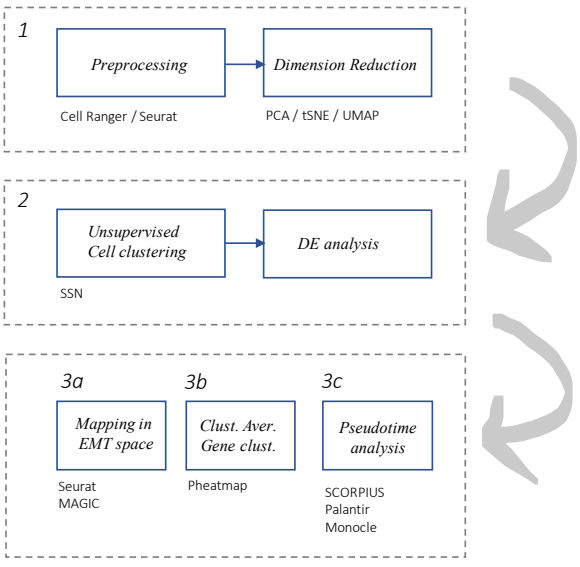


Figure 10. Computational pipeline for studying EMT at the single cell level.

Assigning EMT scores to single cells

First, we studied the principal components to identify the genes responsible for most variance (Figure 11, panel 1-2). As the first two principal components resolved the two subpopulations, the product of the two can be used to model the 'EMT axis'. Accordingly, we computed an EMT score for each cell by multiplying the principal component 1 (`PC_1`) values by principal component 2 (`PC_2`). This approach revealed cluster 6 as most mesenchymal and cluster 0 as most epithelial, with the other clusters in between (Figure 11A, panel 3). While this approach can be useful to identify markers associated with distinct phenotypes, it has several limitations. First of all, there is a non-linear contribution to the variance so that specific genes may highly contribute to variance while their contribution to EMT functionality is marginal. Furthermore, these highly-variable genes can differ across cell lines, which makes this approach sensitive to the input sample. In view of this, we opted for a different approach based on EMT input genes.

Genes were included based on a 40-genes EMT panel described previously[81]. From this list, we identified four mesenchymal genes (*VIM, ACTA2, FN1, COL1A1*) and twenty-two epithelial genes (*EPCAM, CDH, OCLN, TJP1-3, COL4A5, CLDN12/23, CLDN1/3-4/6/7/9, KRT7/8/10/15/18-20*). First we determined the average value of the epithelial and mesenchymal genes using the scaled expression matrix (Figure 11B, panel 2). The majority of $CD44^{high}EpCAM^{high}$ cells had a mesenchymal score equal to zero, which made it problematical to compare these cells based on their EMT score (Figure 11B, panel 3). To tackle this problem, we extended the approach with an imputation step during preprocessing of the data. Markov affinity-based graph imputation of cells (MAGIC), an algorithm that smooths the features and restores the structure of the data, was applied to the count expression matrix prior to calculation of the scores[62]. This resulted in a graph showing gradual transition of epithelial to mesenchymal phenotypes (Figure 11B, panel 4). Interestingly, cluster 6 and cluster 8, earlier shown to associate with mesenchymal (Figure 11A) and E/M characteristics (Figure 9), aligned with the diagonal of the axis suggesting co-expression of epithelial and mesenchymal markers (Figure 11B, panel 4-5). It must be noted that this approach is highly sensitive to input genes, and especially to genes with high expression.

A way to enhance the robustness of this approach is to extend the number of input genes and correct for the weights by which genes can contribute to the variation of the score. Following this reasoning, the input list was replaced with 180 EMT genes from the nCounter® PanCancer Progression Panel. From this list, 52 epithelial genes and 55 mesenchymal genes were identified resulting in a score composed of 107 different markers (Supplementary Table 4). In addition, prior to averaging the epithelial and mesenchymal values, genes were scaled between 0 and 1 to equalize the relative contribution of all the genes. This approach resulted in a continuous near-linear landscape of phenotypic states, where epithelial scores showed a clear negative correlation with mesenchymal scores (Pearson Correlation -0.75)(Figure 11C, panel 1-2). To assess the robustness of the procedure, we overlaid cluster labels and computed averages and standard deviations for each cluster. In support of the robustness of this approach, clusters mapped coherently on the EMT plot. Cluster 6 and cluster 7 appeared as most mesenchymal while cluster 4 appeared as most epithelial, matching the far opposite location of these clusters on the tSNE plot (Figure 11C, panel 3).

To highlight the E/M cells on a dimension reduction plot, we implemented an algorithm previously developed by Tan et al. (2014)[79] with some modifications. We used two gene sets, 52 and 55 epithelial mesenchymal genes respectively, to perform a gene set variation analysis and a two-sample Kolmogorov-Smirnov test (Figure 11D, panel 1). Then, the absolute difference of these scores was subtracted from one, resulting in 1D scores where values close to one denote an E/M state, and lower

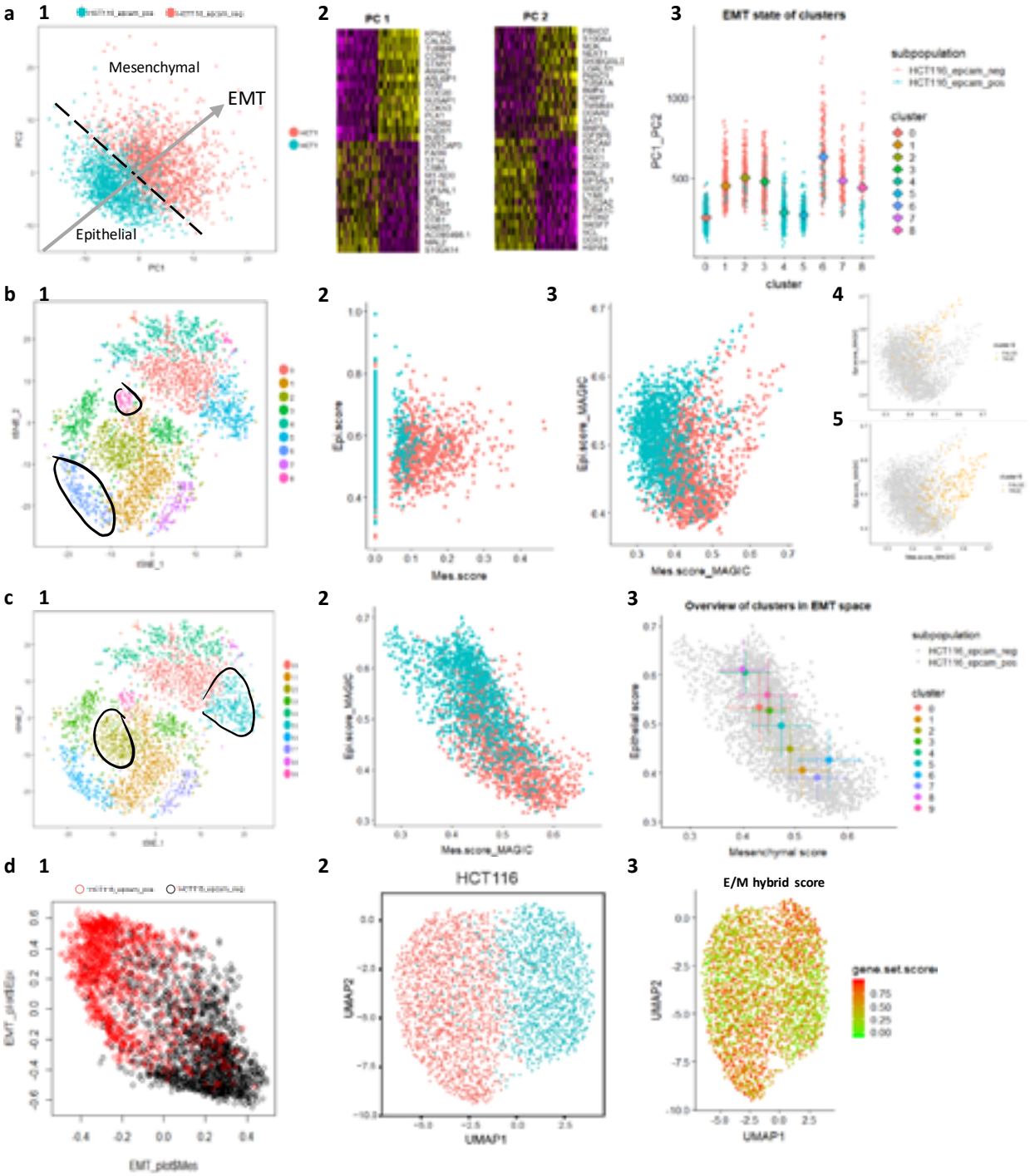


Figure 11. Computation of EMT scores in HCT116. (a) Score based on the principal components. (b) Score based on 40 input genes shows improvement after MAGIC, and points cluster 6 and 8 as hybrid E/M sub-states. (c) Score based on 100 input genes with MAGIC shows gradual transition in EMT with coherent mapping of unsupervised clusters. (d) Conversion to 1D score by gene-set-variation analysis and 2KS statistic to produce E/M hybrid scores on UMAP dimension reduction plot.

values correspond to either epithelial or mesenchymal states. Moreover, we extended the dimension reduction of HCT116 using a novel uniform manifold approximation and projection (UMAP) algorithm. When compared to tSNE, UMAP optimizes the layout of data in a low dimensional space to minimize the error between the two topological representations. In other words, the distance of the clusters have more accurate representation of the underlying variation, and thus may better reflect biological difference. As appeared from the UMAP plot, the E/M cells tend to cluster on opposite sides of the plot, suggesting distinct paths by which a E/M-state can be obtained (Figure 11D, panel 2-3).

This brought us to question whether distinct configurations of genes sets can result in similar E/M scores. And thus, whether E/M-states can be acquired through different transcriptional configurations. To examine this, the analysis was continued with cross-cluster comparison of EMT-related gene expression.

Comparing EMT profiles across clusters

Due to updates in software packages at this time of the analysis, including improvements in normalization and data processing, the tSNE plot as well as the unsupervised clusters changed slightly (Figure 12A). Unsupervised clustering revealed 7 clusters: 2 consisting of $CD44^{high}EpCAM^{high}$ cells; 4 consisting of $CD44^{high}EpCAM^{low}$ cells; and 1 composed of cells from either subpopulation. Evaluation of EMT marker expression revealed again that this E/M cluster showed dichotomous behavior relative to the expression of EMT genes, and thus we subdivided this cluster according to the $CD44^{high}EpCAM^{high}$ (cluster 6) and $CD44^{high}EpCAM^{low}$ (cluster 7) profiles.

To obtain a comparable value for EMT genes in each cluster, normalized expression values of the cells were averaged for the clusters. Following this, z-scores were computed for the EMT genes by subtracting the global gene average from the cluster gene average and dividing by the standard deviation. Next, complete clustering was used on Euclidian distances to reconstruct a hierarchical tree of the clusters and cluster the EMT genes. This procedure resulted in a heatmap showing the average expression values of EMT genes across the different clusters (Figure 12B). As shown in the heatmap, the EMT genes clustered in two sets of epithelial and two of mesenchymal genes. Surprisingly, cluster 5 and cluster 7, originating from the $CD44^{high}EpCAM^{low}$ cells, showed co-expression of an epithelial gene set and mesenchymal gene set, in opposite manner when compared to each other. This suggests that E/M-states can be obtained by two distinct configurations of epithelial and mesenchymal gene expression. In view of this, we computed average values of the z-scores for the gene sets, enabling comparison of gene set expression per cluster. As appears from the graph, cluster 1, cluster 2 and cluster 3 show high expression of mesenchymal gene sets, and lower expression of epithelial gene sets. In contrast, cluster 6 shows high expression of epithelial gene sets while low expression of mesenchymal gene sets (Figure 12C). Clusters 0, 4, and 5 show high expression of epithelial gene set 2 (epi2) and mesenchymal gene set 2 (mes2), with lower values of the other gene sets (epi1, mes1), while cluster 7 shows the opposite profile. Hence, the HCT116 E/M-states can result from different configurations of gene sets (Figure 12D).

To highlight the complexity of these configurations, we employed Ingenuity Pathway Analysis (IPA) to portray the underlying EMT signaling networks. As for the list of EMT genes here employed, we took the 2log-values of the ratio of z-scores for cluster 7 compared to the other clusters, and then imported this list in IPA. The top pathway based on this input "*cell-to-cell signaling and interaction*" was visualized in hierarchical manner, annotating the different components with their gene set origin (EPI1, EPI2, MES1, and MES2). As can be seen from the network, *ZEB1* fulfills a crucial role in this pathway and is upregulated in this cluster compared to others, as well as several transcription factors associated with both mesenchymal (*TWIST*) and epithelial (*GHRL2*) development, thus indicating a state of transition in the EMT program (Figure 12E).

Next, the above procedure was repeated for the SW480 and SW620 cell lines. In SW480, as mentioned, unsupervised clustering gave mixed clusters consisting of both $CD44^{high}EpCAM^{low}$ and $CD44^{high}EpCAM^{high}$ cells (Supplementary Figure 6A). Restricting the clustering to the EMT input list revealed separation of both populations and clusters, with prevalence for either $CD44^{high}EpCAM^{low}$ or $CD44^{high}EpCAM^{high}$ cells (Supplementary Figure 6B). The derived heatmap showed no clear segregation of epithelial and mesenchymal identities, with strong opposing expression values for

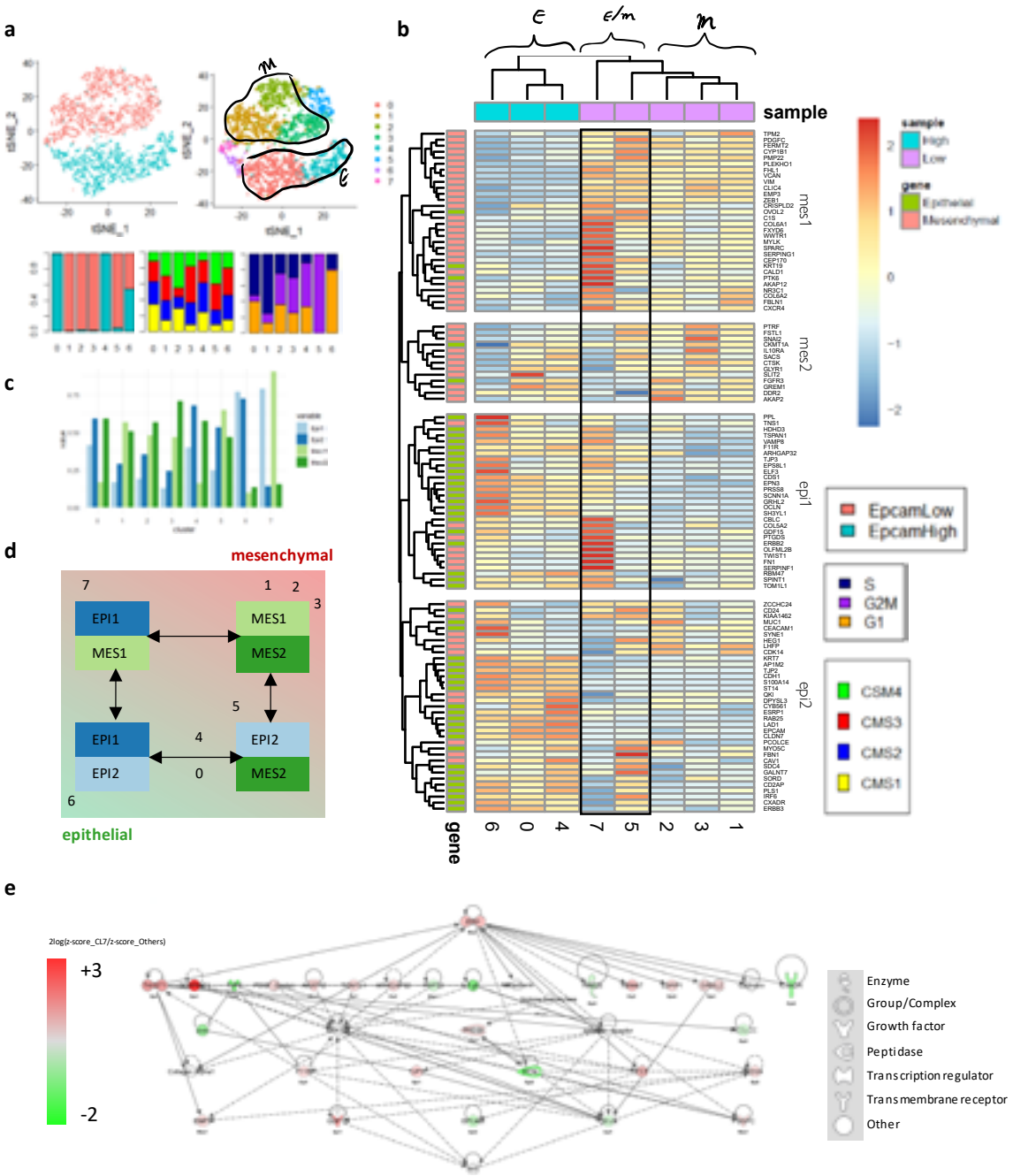


Figure 12. Comparison of cluster profiles on EMT genes. (a) Updated tSNE plot with clusters using Seurat V3. (b) Heatmap showing z-scores of cluster averages for the unsupervised clusters. Genes cluster together in two epithelial sets (epi1, epi2) and two mesenchymal sets (mes1, mes2). (c) barplot showing average z-scores for the distinct gene sets that clustered together in the heatmap. (d) schematic showing opposite configurations of gene sets for cluster 5 and cluster 7. (e) Top pathway IPA: Cell-to-Cell signaling and interaction. HCT116 Cluster 7 vs Others based on "EMT_genes". Values denote log2 values of z-score ratio: cluster 7 / rest.

cluster 5 and cluster 6 (Supplementary Figure 6C). We reasoned that these clusters interfered with the gene clustering, and thus repeated it using all cluster except clusters 5 and 6 (Supplementary Figure 6D). As shown in the heatmap, the remaining clusters showed improved separation of epithelial and mesenchymal genes, with cluster 0 and cluster 3 being more epithelial opposing the mesenchymal clusters 1, 2 and 4.

Overall, it appears that a gradual transition from epithelial to mesenchymal states depending on gene expression intensity of EMT genes is characteristic of the SW480 cell line. Moreover,

additional sub-populations (cluster 5 and cluster 6) accounting for 5% of the CD44^{high}EpCAM^{low} and CD44^{high}EpCAM^{high} cells (i.e. by excluding the CD44^{low}EpCAM^{high} cells), show deviating expression patterns with co-expression of epithelial and mesenchymal genes (Supplementary Figure 6E).

As for SW620, the mechanism by which E/M-states occur is comparable to SW480. From the 12 clusters identified using unsupervised clustering (Supplementary Figure 6F), all but one show similar gradual transition from epithelial expression to mesenchymal gene expression matching the increase of predicted CMS4 fractions (Supplementary Figure 6G). In contrast, cluster 11 is characterized by a striking co-expression of epithelial and mesenchymal genes, including *EPCAM, CLDN7, VIM, ZEB1*, and *SNAI2* (Supplementary Figure 6H).

In conclusion, this analysis revealed different mechanisms by which an E/M-state can be acquired: 1) through coordinated switch of gene set expression (HCT116); 2) through gradual shift in gene expression intensity from one state to the other (SW480, SW620) and 3) via co-expression of both mesenchymal and epithelial genes (SW480, SW620, HCT116).

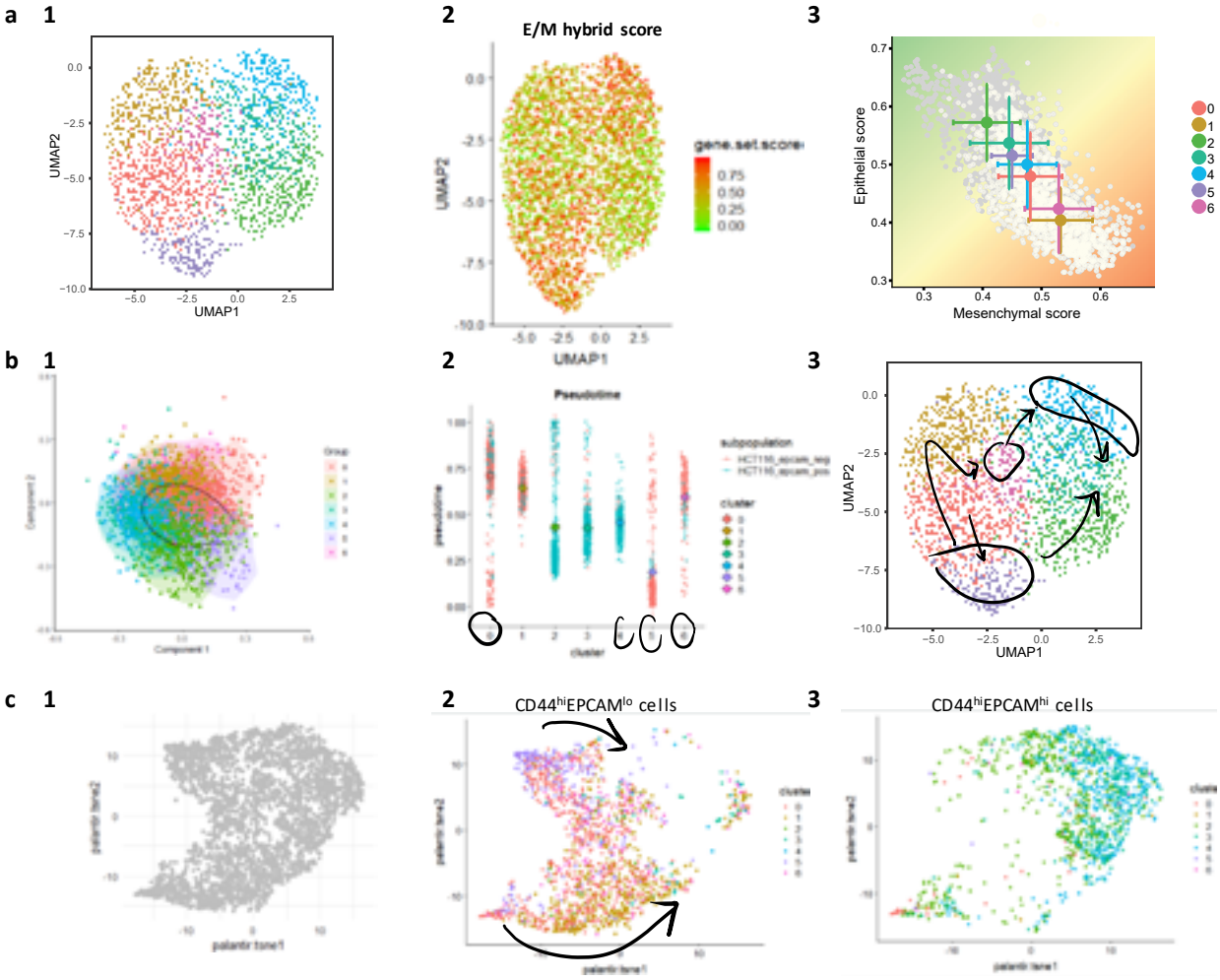


Figure 13. Pseudo-temporal analysis of HCT116. (a) UMAP plot with unsupervised clusters, and E/M hybrid score indicating two E/M sub-states. (b) SCORPIUS analysis reveals two paths by which CD44^{hi}EpCAM^{lo} cells transition into CD44^{hi}EpCAM^{hi} cells. (c) Palantir reveals two points with CD44^{hi}EpCAM^{lo} that can be the start of dynamic trajectories.

Pseudo-temporal ordering of single cells

Our findings suggest that E/M-states can be acquired from CD44^{high}EpCAM^{low} and CD44^{high}EpCAM^{high} cells via several ways of transcriptional alteration. Thus, we speculate that multiple differentially activated paths underlie the transcriptional heterogeneity observed in these cell lines. To elucidate on this reasoning, cells were aligned on hierarchical paths enabling comparison of progression along EMT across clusters. For this purpose, unsupervised trajectory inference was performed on all the cells using SCORPIUS[65]. This analysis was visualized for HCT116 using UMAP dimension reduction, with a clustering procedure slightly differing from former analyses (Figure 13A, panel 1). Four CD44^{high}EpCAM^{low} clusters (0, 1, 5, and 6) and three CD44^{high}EpCAM^{high} clusters (2, 3, and 4) were identified. As shown in the UMAP plot with E/M scores, cluster 0, 5, and 1 have intermediate E and M scores and correspond to E/M clusters (Figure 13A, panel 2-3).

Notably, the E/M-clusters from HCT116 CD44^{high}EpCAM^{low} cells (cluster 0 and cluster 5) appeared at both ends of the process, suggesting that E/M-hybrid states can initiate different paths (Figure 13B, panel 1-2). Cluster 4, the E/M clusters from HCT116 CD44^{high}EpCAM^{high} cells, mapped in between these end points, as well as clusters 3 and 2, that gradually transitioned in pseudo-time values to cluster 5. Thus, it seems that E/M-states from CD44^{high}EpCAM^{low} cells can initiate two trajectories that proceed across the different clusters from CD44^{high}EpCAM^{high} cells (Figure 13B, panel 3).

To test the robustness of these results, we repeated the trajectory analysis using Palantir, an algorithm that models trajectories of differentiating cells by treating cell fate as a probabilistic process and leverages entropy to measure cell plasticity along the trajectory[67,82]. We found again the two E/M-hybrid states (cluster 0 and cluster 5) from CD44^{high}EpCAM^{low} cells as initiation points of different dynamic paths (Figure 13C).

A similar analysis on the SW480 cell line (Supplementary Figure 7) revealed 6 clusters, of which E/M-cluster 2 mapped in between CD44^{high}EpCAM^{high} cluster 4 and CD44^{high}EpCAM^{low} cluster 3. Cluster 3 transitioned in cluster 6, 5, and 1. This suggests that CD44^{high}EpCAM^{low} cells can convert in CD44^{high}EpCAM^{high} cells and vice versa via distinct biological paths.

Altogether, our single cell RNAseq analysis reveals considerable heterogeneity in the CD44^{high}EpCAM^{low} and CD44^{high}EpCAM^{high} subpopulations across 3 colon cancer cell lines. While our results show a strong association between CD44^{high}EpCAM^{low} cells and the mesenchymal phenotype, they also indicate that within this subpopulation cells are present that co-express epithelial and mesenchymal genes. We identified E/M-states in all cell lines, and found they can occur by different mechanisms of transcriptional activity. Lastly, E/M-states could play an important role in phenotypic plasticity of the cell lines since they can be the start of different trajectories.

Screening and optimization of collagen models for collective migration in vitro

In this section, results will be presented on three different collagen models that we tested for the study of collective migration studies in vitro and ex vivo: 1) a collagen droplet model, 2) a collagen chamber model that enables directional incentives, and 3) a collagen freeze model with aligned pores.

Migration of cell aggregates in a collagen droplet model

Our initial experimental set up bears a close resemblance to approaches previously used in breast24 and brain cancer[62]. Starting from cellular aggregates of HCT116 and SW480, we monitored invasion in three-dimensional rat tail collagen type 1 (4µg/mL) droplets. After 2 days of culturing, cells started to invade the collagen, either as single cells detaching from the aggregate, or by budding sites as predominantly observed in SW480 (Figure 14A).

We used the same set up to study invasion of fragments of human primary colorectal tumors. To trigger EMT and invasion, fragments were exposed to compounds that have been shown to trigger migration or invasion: CHIR[63] (Wnt pathway), TGFß[64] (TGFß-pathway), SCF (SCF/c-KIT-pathway), JAG1[65] (Notch pathway), HGF[66] (HGF/c-MET pathway). This resulted in events where cells invaded in the collagen either as long, stretched single cells, or as multi-cellular strands (Figure 14B). We questioned whether the non-directional approach of this set up would hamper cells from directed invasion. A stimulus from a source can give the cells an incentive to invade in certain direction, which may increase the frequency of invasion, and could thus be an improvement of the set up. Since we did not observe any detachment of migrating cell clusters using this approach, we decided to change the experimental set up and include directional incentive.

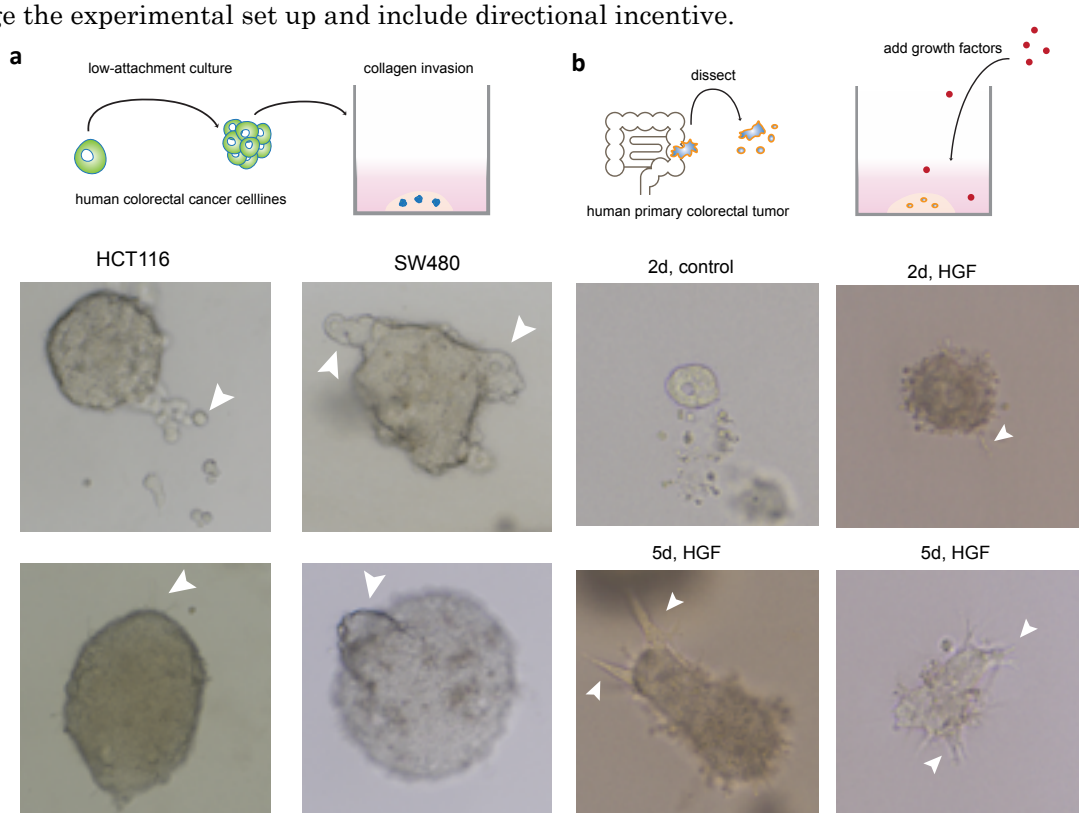


Figure 14. Collagen droplet model. (a) Invasion of cells from cellular aggregates of HCT116 and SW480. (b) Invasion of cells by fragments of human primary colorectal tumors.

Migration of organoid-derived multicellular layers in a collagen chamber model

We adapted an experimental set up previously applied in skin cancer[67] with some variations (Figure 15A). Human (CSC08) and intestinal mouse (APK) organoids were seeded on top of a collagen

gel at $0.5*10^6$ cells/well and placed in a Boyden chamber. Inactivated cancer associated fibroblasts (CAFs) were co-cultured in 2D monolayer on the bottom of the well to provide a continuous source of EMT induction. Following 7 days of incubation, gels were harvested, processed, and sectioned for immunohistochemistry and immunofluorescence.

Organoid-derived multilayers appeared on top of the collagen gel (Figure 15B). We did observe, although scarcely, events of single and collective cell invasion in the collagen (Figure 15C). When using immunofluorescence, it appeared that the multilayer from human organoids showed non-homogenous expression of E-cadherin. Surprisingly, cells on top of the multilayer showed perpendicular alignment of their nuclei and were characterized with lower levels of E-cadherin expression (Figure 15D).

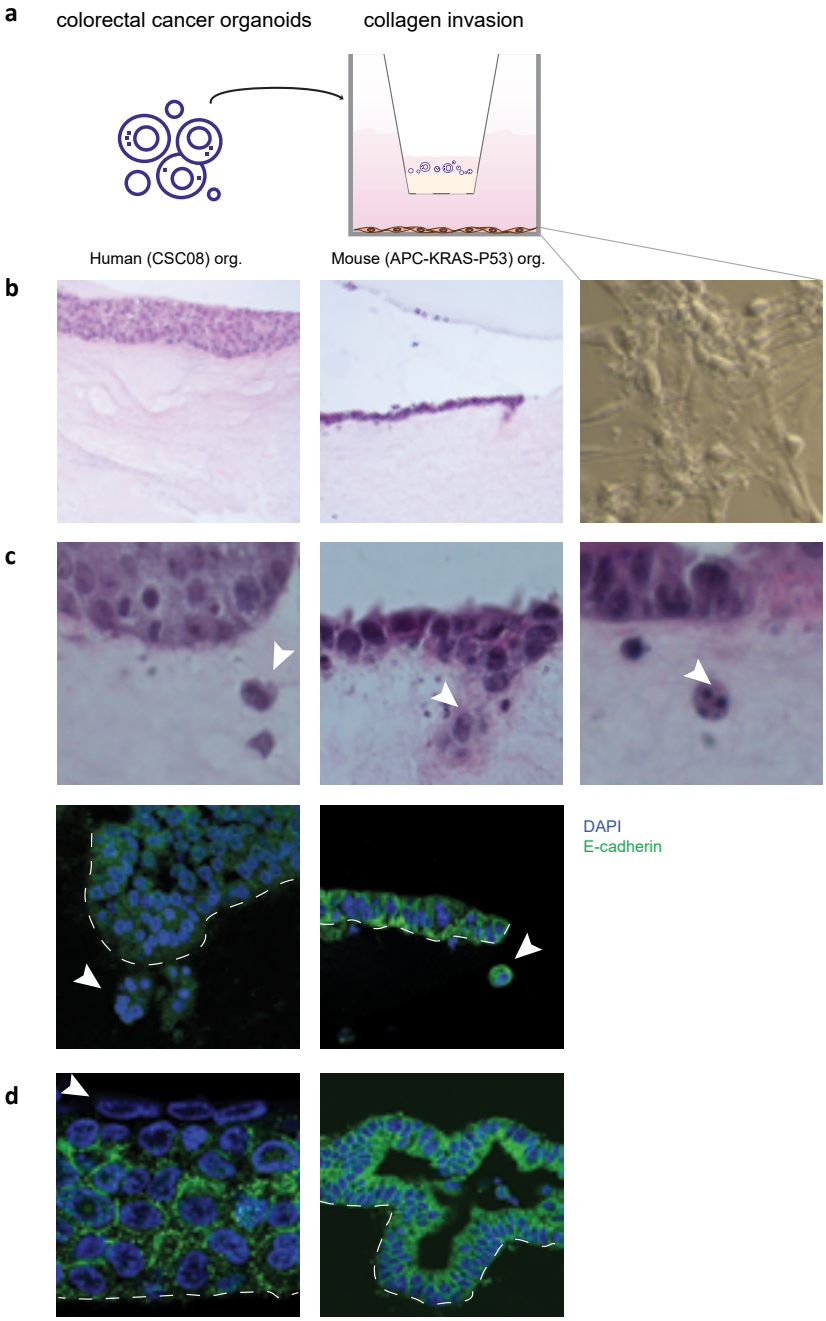


Figure 15. Collagen chamber model. (a) Schematic experimental set up of the collagen chamber model. (b) HE showing multi-cellular layers on top of the collagen gels. (c) Events of single/collective invasion in the collagen gel. (d) IF showing E-cadherin (green) and DAPI (blue).

Although we did see sporadic events of invasion as either single cells or small cell clusters, we considered this experimental set up as suboptimal for experiments assessing collective cell migration in qualitative and quantitative fashion. Next, we attempted to alter the architecture of the collagen matrix to facilitate and orient cell invasion/migration. Earlier studies have successfully shown that collagen stiffness, i.e. its relative concentration, can strongly influence the migration type[64]. Other parameters include polymerization temperature68, pre-polymerization steps[69], and aligned collagen fibers by mechanical stretching[70]. Recently, a novel approach has been developed in breast cancer based on a freezing step to create aligned fibers in the collagen[72,73]. Based the observed increased frequency of cell clusters and multi-cellular tendrils obtained with this approach, we decided to further implement this method.

Migration of mouse organoids in a collagen freeze model

We produced anisotropic collagen scaffolds according to procedures described elsewhere[73] (Figure 16A). In brief, collagen (10 mg/mL) was acidified and vortexed to produce a slurry that was frozen at -20C˚ using a mold enabling freezing from one focal point. Subsequently, ice crystals were removed by sublimation for 18 hours and scaffolds were cross-linked to enhance stability. When compared to the collagen matrices without freezing, clear differences in matrix architecture can be observed due to the freezing procedure. Using bright field imaging, we observed the presence of aligned pores (Figure 16B).

One of the key characteristics of collagen is its structural organization. The amino acids of three peptides interact to form a triple helix (tropocollagen, 300 nm) that gather in fibrils (1 µm) to form fibers (10µm), or bundles of fibers. To study the collagen organization at the µm scale, we used reflection imaging. By capturing fluorescent light, close the wavelength of excitation, an image can be produced that reflects the structure of the scaffolds. In collagen that is polymerized without freezing, fibers can be recognized in a isotropic manner (Figure 16C). In the scaffolds that were created using the freezing procedure, thick bundles were observed that were composed of aligned fibers (Figure 16D). Thus, the freezing procedure has a strong impact on the microscopic organization of the collagen.

To study the behavior of cells on the freezing scaffolds, mouse AKP organoids were seeded on top of the scaffold and incubated for 7 days. Next, 4µm sections were made, and stained for hematoxylin, eosin and Sirius red to stain the cells and collagen respectively. We observed a thick multi-cellular layer on top of the scaffold, where the morphology of the spheroids can still be recognized (Figure 16E). Deeper in the scaffold, multi-cellular strands were migrating along the thick bundles of collagen. Similar observations were made using immunofluorescence in confocal microscopy (Figure 16F).

Implementation of the embedding and sectioning procedures implied problems with sectioning due to the fragility of the scaffolds. Moreover, 2D planes can be misleading as cross-sections of cellular strands can give the impression of freely migrating cell clusters while in reality they were attached to the multi-cellular mass. Therefore, we put effort on fluorescent 3D imaging as method of choice for evaluation of migration.

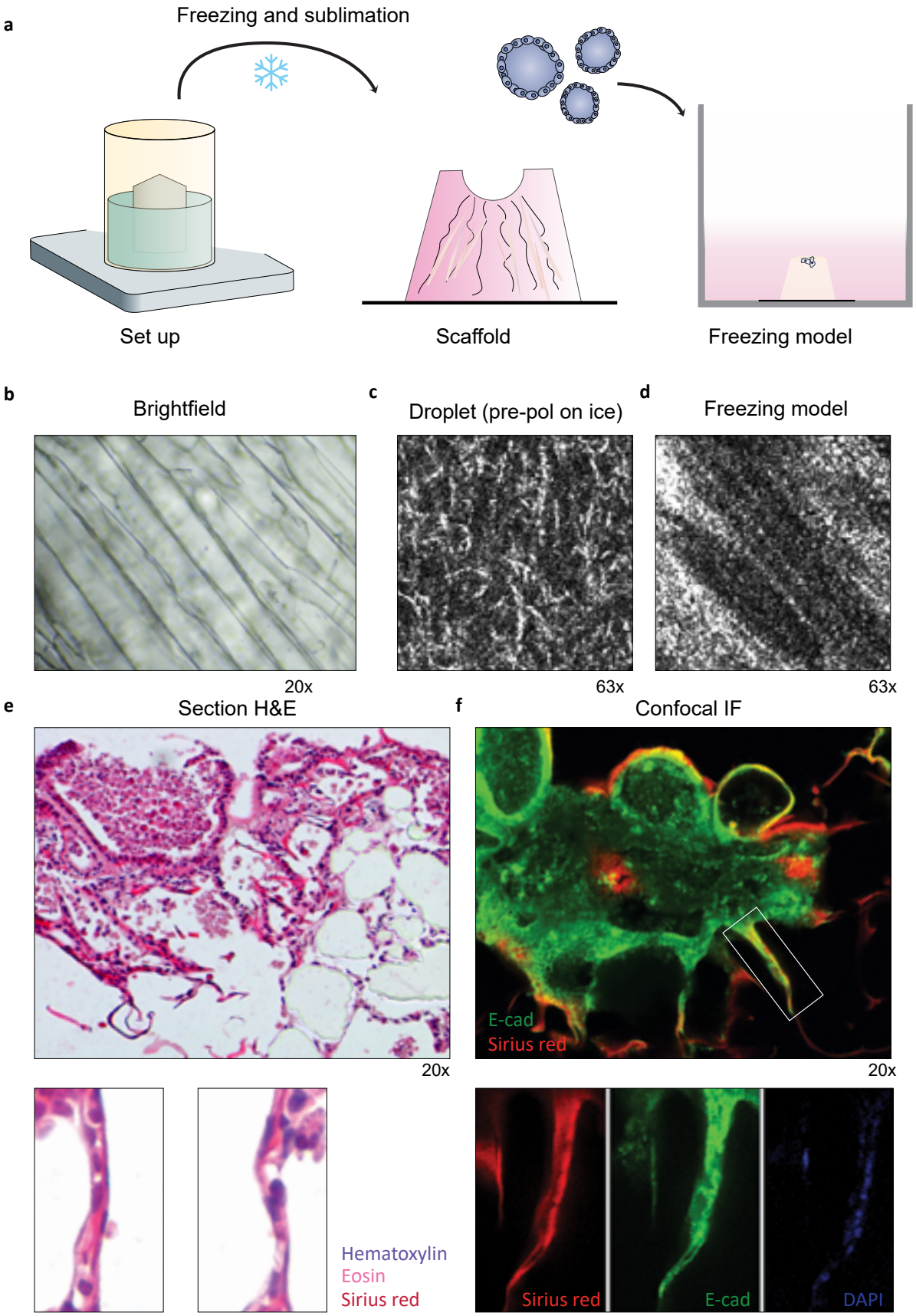


Figure 16. Collagen freeze model. (a) Schematic experimental set up of the collagen freeze model. (b) Brightfield imaging showing aligned bundles of collagen in the collagen freeze model. Reflection imaging showing difference in fiber organization in the (c) collagen droplet model with 2 hours pre-polymerization on ice and (d) the collagen freeze model. (e) HE sections after 7 days incubation of AKP organoids in the collagen freeze model showing multi-cellular strands migrating along the collagen bundles. (f) Immunofluorescence using confocal microscopy showing a section of the AKP organoids on the collagen freeze model.

3D imaging of organoids in the collagen freeze model

We repeated the migration assay with mouse AKP organoids incubated for 3, 5, and 7 days on the collagen scaffolds prepared according to the freeze protocol. Scaffolds were then fixed and processed using a clearing protocol[47]. The Leica SP5 intravital and the Opera Phenix were employed to visualize the samples, and Hyperstacks were acquired with a depth in the range of 400-1000 µm. We observed clear migration after 5- and 7-days incubation, predominantly by multi-cellular strand formation but also found examples of de-attached cell clusters (Figure 17). in view of these results, we decided to continue with the collagen freeze model as scaffold for subsequent experiments, and use clearing and 3D imaging as method of choice to evaluate migration events.

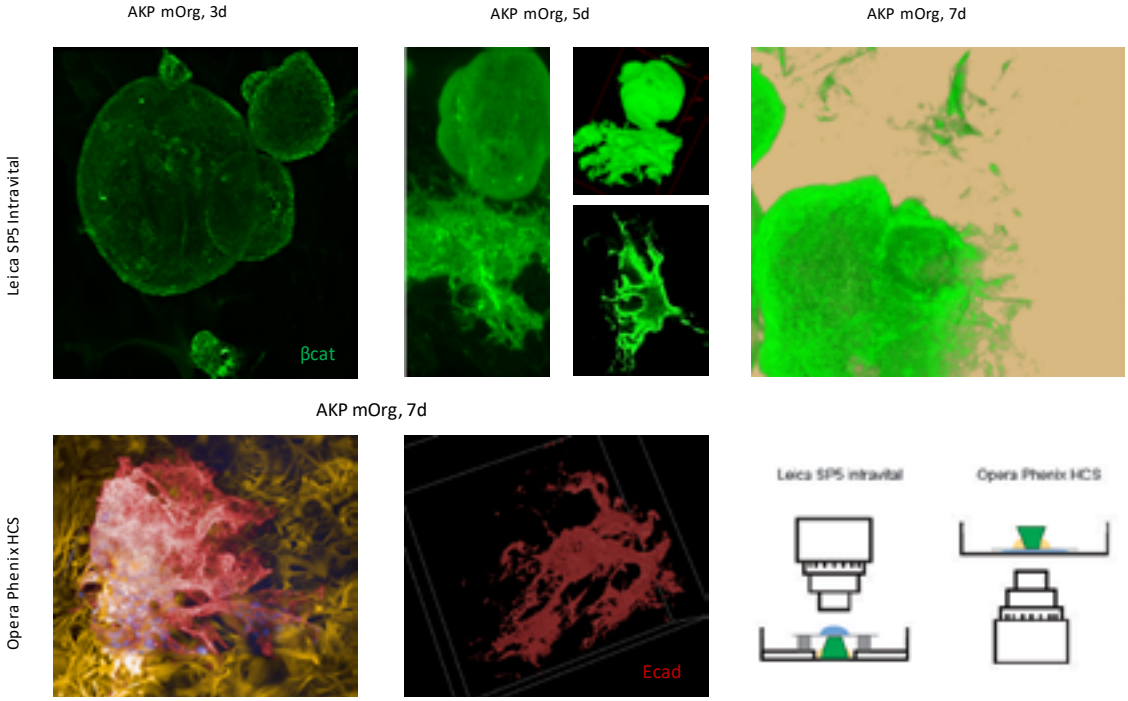


Figure 17. 3D imaging approaches for the collagen freeze model. Top: Leica SP5 acquisition for migration of AKP organoids in the collagen freeze model after 3, 5, and 7 days incubation. Bottom: Opera Phenix HCS acquisition of AKP organoids in the collagen freeze model after 7 days incubation

Evaluation of tumor migration mechanisms upon induction of *Zeb1*

Next, and illustrated in this section, exploratory experiments were performed aimed at understanding the role of EMT transcription factor ZEB1 in tumor invasion. The AKP mouse organoids was employed as experimental study model of migration and invasion when coupled with the collagen freeze model. The experiments below are anecdotical, and performed to show how the employed systems can be used to study migration modalities *ex vivo*.

Validation of conditional ZEB1 overexpression in mouse organoids

First, the mouse AKP organoids were transduced with a vector to conditionally induce mouse *Zeb1* expression (Supplementary Figure 8). Ten clones were evaluated for inducible *Zeb1* expression. To this aim, organoids were cultured in matrigel and exposed to doxycycline (dox, 1µg/mL) for 48 hrs. (Figure 18A). RTqPCR analysis revealed a *Zeb1* mRNA increase of 65-, 70-, and 40-fold for clones 4, 6 and 10, respectively (Figure 18B). Accordingly, western blot analysis showed an increase of Zeb1 protein in these clones upon dox-induction (Figure 18C).

To assess migratory changes upon dox-induction, we seeded AKP-*Zeb1*-GFP clone 10 organoids

on collagen freeze scaffolds and followed them over time by imaging every 2 hours over 48 hours. Unfortunately, the organoids died over time, probably due to the intensive imaging schedule (i.e. outside the incubator). Yet, some cells – both in the dox-induced and uninduced samples, can be recognized with mesenchymal-like morphology, as well as sharp deformations in the organoid suggesting initiating events of migration (Supplementary Figure 9).

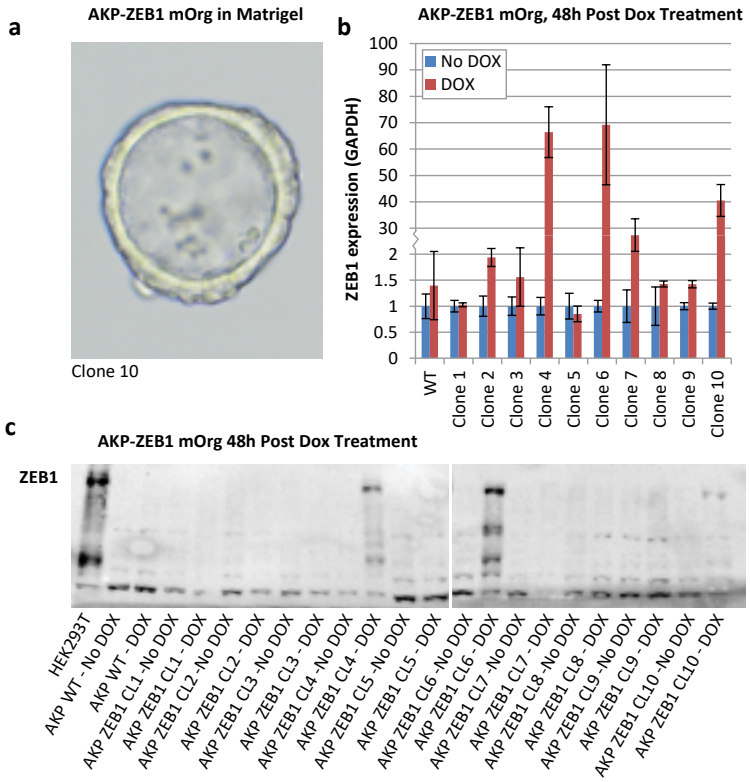


Figure 18. Validation of inducible *Zeb1* expression. (a) Representative image of AKP organoids in Matrigel. (b) qPCR of ZEB1 expression upon 48h exposure to DOX. (c) Western blot showing Zeb1 protein expression after 48h exposure to DOX.

Comparison of *ex vivo* tumor migration mechanisms upon induction of *Zeb1*

Ex vivo migration assays were implemented with ±1 mm^3 tumor fragments resected from immune-deficient mice orthotopically (i.e. caecum) transplantated with AKP-*Zeb1* mouse organoids (clones 4, 6, or 10). These fragments were seeded on collagen freeze models and incubated in DMEM-FCS for 7 days in the presence of gentamicin to prevent intestinal microbial contaminations, and in wells pre-coated with matrigel to improve attachment of the collagen scaffolds.

Scaffolds were imaged after 3 and 7 days of incubation. Sirius red was employed to stain collagen type I, while ß-catenin or E-cadherin antibodies were used to stain the tumor cells by IF. 3D visualization of tumor fragments was focused on regions of contact with the collagen scaffolds (Figure 19A). In the 3-day sample without doxycycline, multi-cellular strand formation was observed (Figure 19B), not observed in the 3-day sample exposed to doxycycline. In both samples, single cell migration along collagen bundles as well as small clusters consisting of few cells were observed (Figure 19C).

In the samples incubated for 7 days, we long strand formation was observed in the absence of doxycycline, which were not seen with doxycycline (Figure 19D). Interestingly, a cell cluster was found in the no-dox sample at approx. 2 mm from the seeding area. In the +dox sample, we did not observe similar events. Instead, a prevalence of single cell migration from the tumor center was noticed.

In order to quantify these migration modes, we extracted the center point for each nucleus (Figure 19E), and computed a 2D kernel density to represent the tumor mass of the samples, and annotate migratory cells (Figure 19F). Subsequently, k-nearest neighbor distance was computed for the migratory cells against the tumor mass to approximate the migrated distances. The sample exposed to Doxycycline had a higher number of single migratory cells (+dox: 184, -dox: 99), but where found closer to the tumor mass than the migrated cells from the sample without Doxycycline exposure (Figure 19G). Since we performed this experiment only once, it is impossible to dissect conclusions from the observations. Yet, this serves as anecdotical example to illustrate how this system can be employed to study and quantify migration modalities upon culturing tumor fragments *ex vivo*.

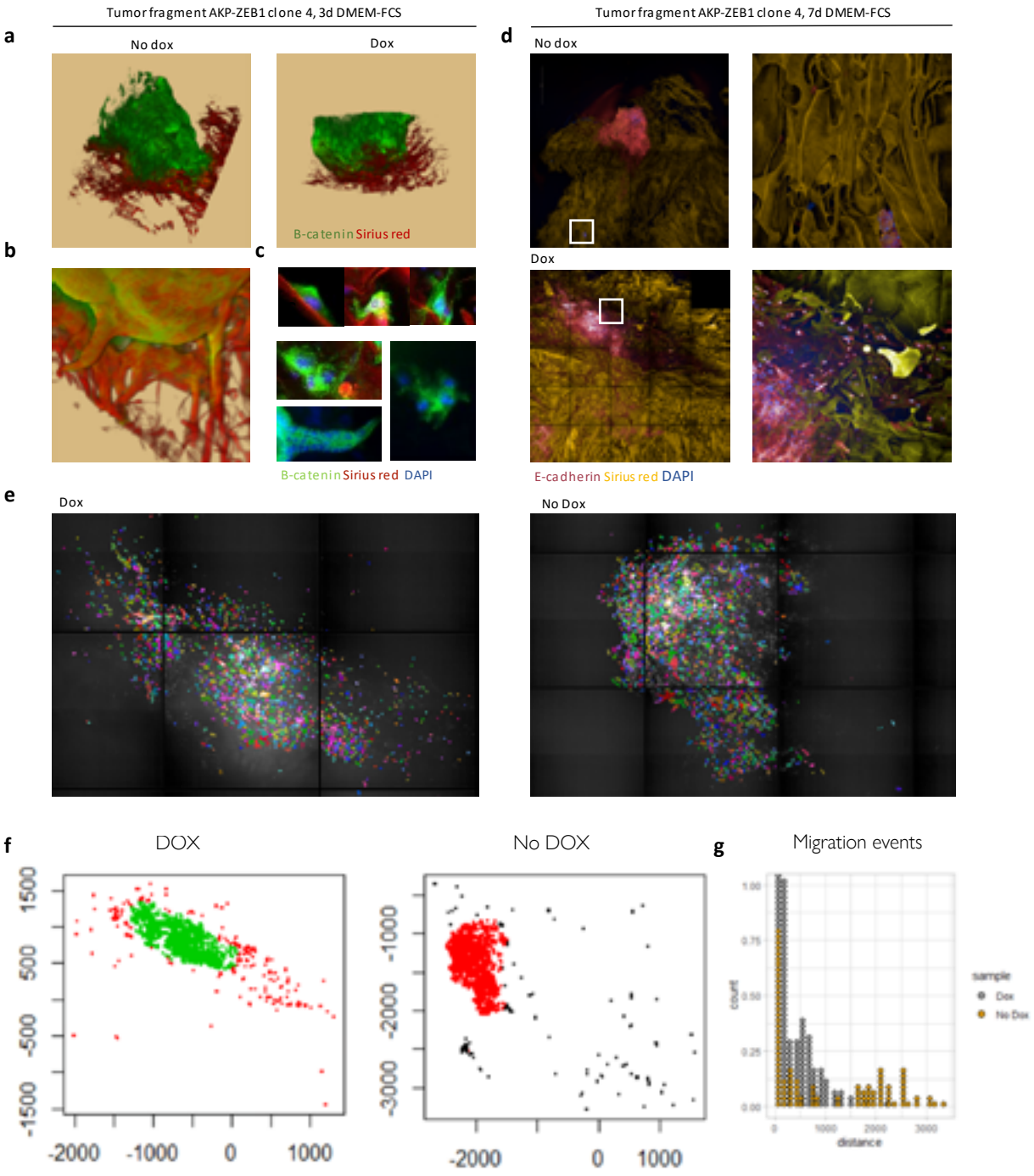


Figure 19. Tumor migration mechanism upon *Zeb1* induction. (a) 3D projections of organoid derived tumors seeded on the collagen freeze model. (b) Multi-cellular strand formation in the No Dox sample. (c) Examples of single cell migration and de-attached cell clusters in the collagen freeze model. (d) Comparison of Dox and No Dox sample after 7d incubation *ex vivo*. (e) Segmentation of nuclei based on relative DAPI intensities. (f) Quantification of single cell migration based on distance of cells to nearest cell that falls within the contour line.

Tumor invasion mechanisms *in vivo*

To compare the results of the *in vitro* models to the situation *in vivo*, we performed two exploratory experiments. First, we did a ß-catenin IHC staining to identify the invasion modalities in the orthotopic transplantated caecum tumor in mice. This revealed different migration types at the invasive front: i) single cells, 2) cell clusters and 3) multi-cellular strands (Figure 20).

Next, we aimed to reconstruct a 3D image of the tumor invasion *in vivo*. To this aim, we punched small cylinders of 1 mm in diameter out of the Paraffin Embedded tumor front. Following this, cylinders containing tumor tissue were cleared similar to the procedures used for the visualization of collagen freeze scaffolds. This was followed by a staining of E-cadherin and α-SMA to distinguish epithelial cells from fibroblasts. This experiment revealed that a considerable resolution, sufficient to distinguish single cells, was maintained until deep in the tumor tissue. To illustrate this, 3D projections of epithelial cells and fibroblasts versus only fibroblast shows the presence of fibroblast deep in the tumor (Supplementary Figure 10).

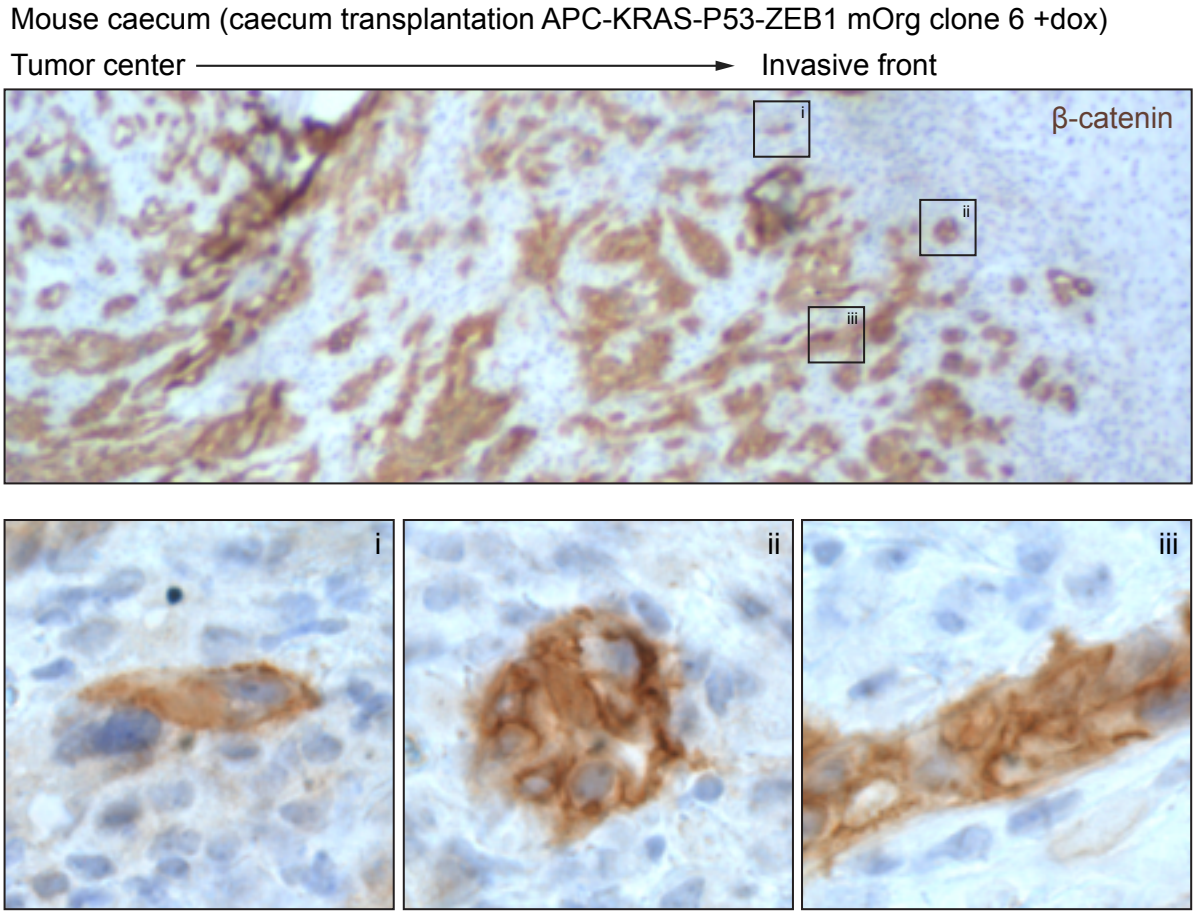


Figure 20. Invasive front of AKP mOrg derived ceacum tumor. Invasion can be observed by i) single cell migration, ii) de-attached cell clusters and, iii) multi-cellular strand formation.

Toward isolation of circulating tumor cell clusters from liquid biopsies

Last, and exemplified in this section, we analyzed blood samples with the aim to identify and quantify CTC and CTC clusters.

Prior to peripheral blood collection, 3 mice carrying organoid-derived (AKP-*Zeb1*-GFP mOrg) caecum tumors were imaged in the IVIS to identify potential metastases. Mouse MI-11574-03 appeared to have one metastasis, which was not apparent for mouse MI-11574-01 and MI-11574-07 based on this acquisition (Figure 21A). We isolated 240 µl, 900 µl, and 740 µl of blood from mouse MI-11574-01, MI-11574-03, and MI-11574-07 respectively. Next, we performed a red blood cell (RBC) lysis and PBS wash before fixation in 2% PFA in BSA for 20 min. at RT. Then, the remaining pool of cells were distributed over glass slides in 70% ethanol, and stained for DAPI and anti-GFP to visualize the CTCs.

As initial experiment, the OPERA Phenix was employed to scan four glass slides (2x MI-11574-03, 1x MI-11574-01 and 1x MI-11574-07). Whole glass slides were scanned at 20x magnification in two channels, 488 nm and 405 nm. Hereafter, the software was tasked to identify all nuclei on the glass slides. This revealed 10.000 cells on the glass slides from MI-11574-01, 75.000 and 49.000 cells from the samples of MI-11574-03, and 3.000 cells in the sample from MI-11574 (Figure 21B). The variability in these numbers indicated that this procedure requires optimization. As a reference, others have reported that on average, NOD scid-gamma (NSG) mice have a white blood cell count in the range of $1.0\text{-}7.0*10^3/\mu l$[94]. Hence, this would imply that on a slide representing few hundred µl of blood, in the range of $10^5\text{-}10^6$ cells should be identified. Thus, the samples of MI-11574-03 and MI-11574-01 approach the lower limit of what would be expected, but the sample of MI-1157407 is not representative and probably many cells were lost during the RBC lysis and washing procedures.

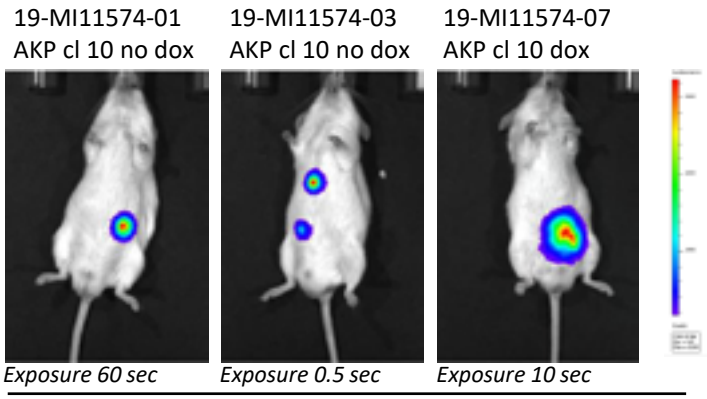
We continued by measuring the average GFP intensity in both the nuclei (DAPI segmentation) and the cytoplasm (DAPI region subtracted from GFP segmentation). Next, thresholds were set for both parameters ($4*10^3$ for nucleus, $4*10^3$ for cytoplasm) to separate white blood cells (WBC) from CTCs. In mouse MI-11574-03, we identified 283 and 225 CTCs in the two glass slides. We did not find CTCs in the other two mice.

To further investigate the CTC pool, we computed nearest-neighbor (NN) distances for all the cells that were annotated as GFP^+ (CTC). Then, we distinguished CTCs from CTC clusters by establishing a threshold for the NN distance. Cells with distances > 20µm were annotated as single cells, while cells with distances smaller or equal to 20 µm were labeled as part of CTC cluster. In the first acquisition of MI-11574-03, we observed 282 CTCs of which 30 cells (12%) were part of three different CTC clusters (Figure 21C). In the second acquisition of MI-11573-03, we observed 225 CTCs, which were all single cells.

To improve the resolution of the image, the glass slide that was found to contain CTC clusters was imaged under using the Zeiss LSM-700 microscope. We searched for one of the three CTC clusters and captured the DAPI and GFP channel at 63x magnification. One of the three cluster was composed of five cells (Figure 21C, left). As can be seen from the DAPI channel, the nuclei of the CTCs was bigger than the white blood cells, and showed occasionally a spotted pattern, different than the DAPI signal from the white blood cells (Figure 21D).

Taken together, we show that an automated imaging approach using the OPERA system can be suitable for the detection and quantification of CTCs and CTC clusters.

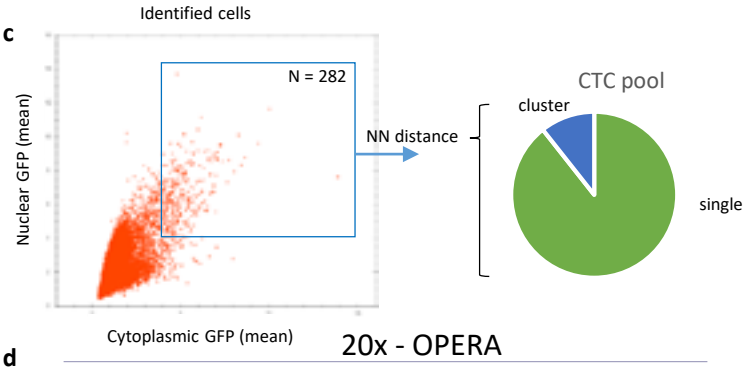
a



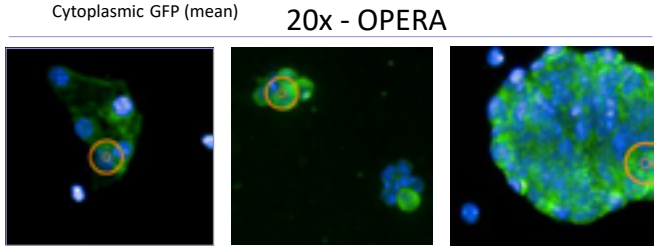
b

Mouse	Isolated blood volume (ul)	Distributed over slides	Opera	Number of cells	GFP+ cells	GFP+ cells in cluster	Estimated CTC/mL
1157401	240	3	Acquisition 1	10000	0	0	
1157403	900	3	Acquisition 1	75000	282	30	940
			Acquisition 2	49000	225	0	750
1157407	740	3	Acquisition 1	3000	0	0	

c



d



e

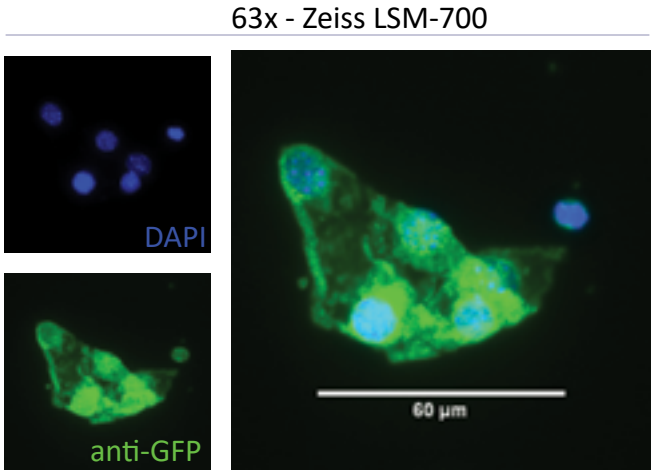


Figure 21. Approach for isolation of CTCs and CTC clusters. (a) Before collection of blood, mice were imaged under the IVIS to measure luciferase signal and detect sites of tumor and metastases. (b) Table summarizing the isolated blood volume per mice, and detected cells after red blood cell lysis, and staining for DAPI and GFP using the OPERA system. (c) Results of MI-11574-03 acquisition 1. From the 75.000 detected cells on the glass slide, 282 were annotated as GFP⁺ cells. From the pool of 282 CTCs, 30 cells were found to have a nearest-neighbor distance smaller than 20 µm, and thus were part of CTC clusters. (d) Examples of CTC clusters, imaged with the OPERA at 20x magnification. (e) Confocal image at 63x of one of the CTC clusters.

Chapter 4. Discussion

This thesis attempted to contribute to the current debate about the role of EMT in collective cell migration. Although there is an increasing number of studies indicating that E/M cells and CTC clusters play an important role in metastasis formation, a cause-effect relationship between the two, despite several conceptualizations, has not been demonstrated yet. In Chapter 1, we dissected this relationship into different questions demanding investigation. The first question revolved around EMT and the existence of E/M sub-states. We assessed this question experimentally by employing single cell RNA sequencing technology on $CD44^{high}EpCAM^{low}$ cells, a metastatic subpopulation of colorectal cell lines that was intensively characterized in the lab before the start of this thesis (see Chapter 1, Box 1). The two other questions focused on different steps along the invasion-metastasis cascade, i.e. invasion and survival in circulation, and need clarification in order to establish or deny a cause-effect relationship between partial EMT and collective migration. Efforts in this regard have been devoted to the optimization of approaches useful for studying the interplay of these models (see Chapter 2). While these approaches remain far from being 'state-of-the-art', we have showed that they can be used to address specific questions on phenotypic plasticity and the allegedly underlying partial EMT and collective migration.

Main findings

We zoomed in on subpopulations of $CD44^{high}EpCAM^{low}$ cells and $CD44^{high}EpCAM^{high}$ cells, which were characterized by several functional assays indicating that $CD44^{high}EpCAM^{low}$ cells have enhanced metastatic potential and stemness (see Chapter 1, Box 1). In HCT116, the transcriptional difference between these subpopulations was responsible for most of the variance, while in SW480 additional variance was present as evidenced by the overlap of populations in the dimension reduction plot (see Chapter 3, Figure 7D). This could relate to the genetic differences of the cell lines, as the chromosomal instability status of SW480 may result in increased inter-cellular epigenetic differences when compared to the HCT116 cell line that is classified with microsatellite instability[75].

We developed a computational pipeline suitable for investigating EMT at the single cell RNA level. This pipeline was used to show that colon cancer cell lines are composed of several subpopulations that map differently on the EMT spectrum. Our analysis indicates E/M sub-states based on the mapping of unsupervised clusters, that contribute to the plasticity of the cell line as they can initiate different paths of differentiation. In HCT116, these E/M sub-states proceeded by coordinated switch of gene sets via either high expression of *SNAI2* or *ZEB1*, which is in accordance to the notion of the distinct feedback loops of these transcription factors[36], and may relate to the two *ZEB1*-dependent steady-state attractor states as described before[76]. In SW480 and SW620, we observed E/M sub-states showing strong co-expression of epithelial and mesenchymal genes including key EMT regulators such as *ZEB1, SNAI2* and *TWIST*, suggesting a different dynamic transition by which EMT proceeds.

Overall, these results indicate that EMT states cannot be arrayed along a linear spectrum that starts with fully epithelial phenotype and ends with a fully mesenchymal one, but that states can branch off in different cell fates, which can be different for each cell line (Figure 22).

We did observe coherent mapping of unsupervised clustering on the EMT plots, and distinct EMT expression across clusters, suggesting preference for certain states along the possible phenotypes. However, one should be cautious with claims about the number of EMT sub-states, because the identified number of sub-states depends on the number of groups that were initially clustered.

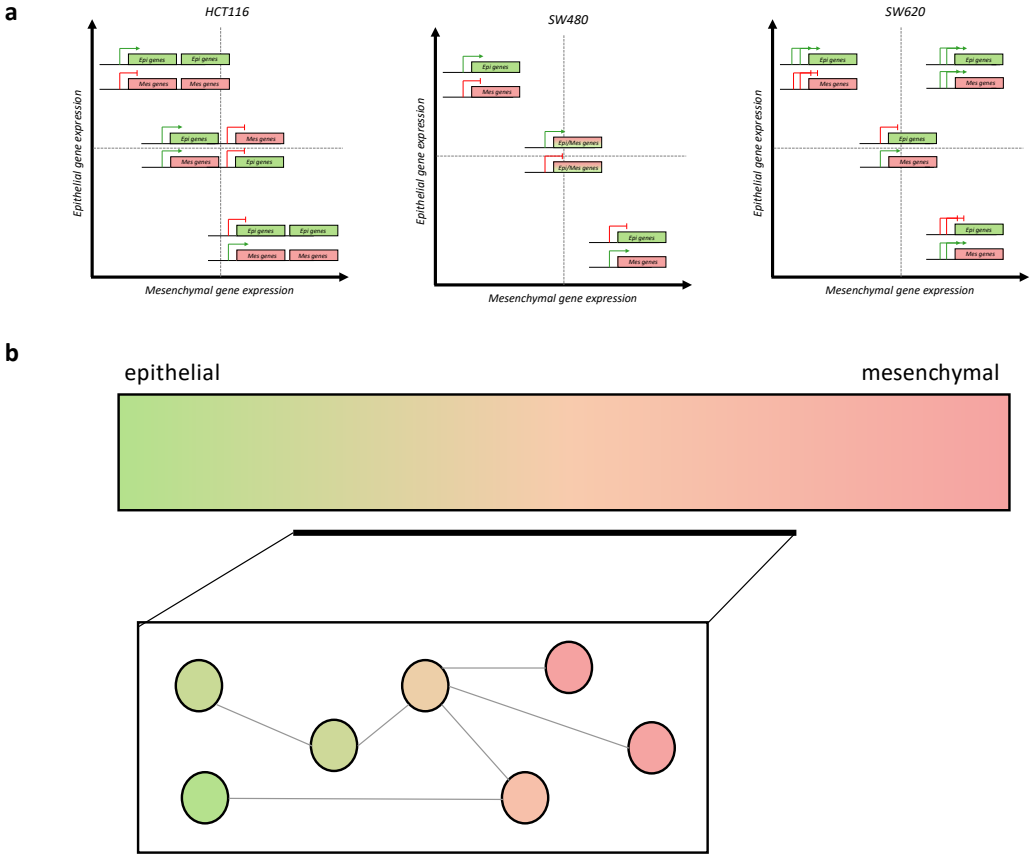


Figure 22. Schematic diagram proposing EMT progression along multiple paths. (a) Different mechanisms of E/M sub-state transition in colorectal cancer cell lines. (b) Schematic diagram showing the non-linear array of EMT phenotypes that are connected via multiple paths.

In the second part of this thesis, we explored approaches to study collective migration *in vitro/ex vivo* and ways to identify CTC clusters from liquid biopsies. We evaluated different collagen models with respect to their ability to capture collective migration events. Over the course of the project, we varied both collagen models and biological samples to increasingly complex experimental designs (Figure 23). Based on the exploratory experiments performed in this project, the results showed most resemblance to the situation *in vivo* when tumor fragments were seeded on the collagen freeze model. This set up resulted in events of single cell migration, as well as collective clusters and multi-cellular tendril formation, and was therefore selected for subsequent experiments.

To demonstrate the use of this experimental design in studying collective migration and EMT, we compared tumor migration *ex vivo* after 7 days of *Zeb1* ectopic expression. From this anecdotical experiment, we observed an apparent shift to single cell migration upon induction of *Zeb1*. Moreover, our results suggest a reduction of multicellular-tendril formation upon *Zeb1* induction. However, these results should be taken with caution given the low experimental numbers. Interestingly, in the sample without *Zeb1* induction, we observed a collective migrating cell cluster 2 mm away from the scaffold seeding area. This event illustrates the ability of this model, as well as its employed evaluation, to capture and study migration events at a relatively large scale.

Finally, we focused on the liquid biopsies to study the circulating tumor cells. A shift from single CTCs versus clusters, and the arrangement of EMT-related inter cellular heterogeneity in clusters, would provide ultimate proof for the association between E/M and collective migration. To study

these questions, we are in need of methods that can accurately identify CTC clusters from liquid biopsies. In this regard, we have chosen to deviate from standardized methods because they may preferentially detect single CTCs and hamper detection of CTC clusters. Our approach is rather simple, and is solely based on a red blood cell lysis and one wash with PBS. To aid in the identification of the "needle in the haystack", we have developed an automated approach to detect rare events. This approach was used to show that we could detect 3 CTC clusters in a pool of 75.000 white blood cells.

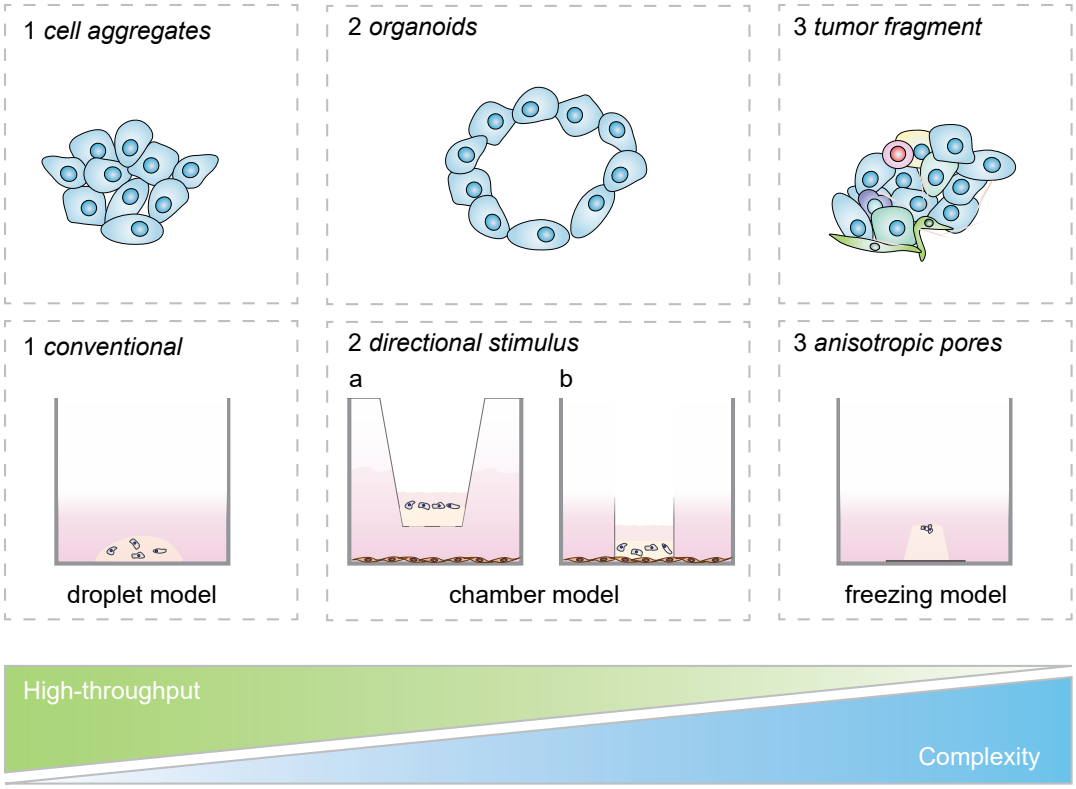


Figure 23. Overview of experimental designs used for migration studies.

Limitations

The presented approaches suffer from a number of pitfalls. We used cell lines to study the spectrum of EMT at the single cell RNA level. Indeed, a cell line is a strong simplification of the in vivo situation. This is of relevance since EMT has been described as context-dependent, non-autonomous cellular program that is greatly influenced by the tumor microenvironment[11]. The influence of these factors was neglected in our model.

As migration assays we used a collagen freeze model to assess collective migration. Due to the freezing procedure, anisotropic pores were produced that result in holes in the scaffolds. In contrast to the collagen droplet model and the collagen chamber model, this allows cells to migrate in the scaffolds without active degradation of the collagen fibers. While collagen architecture *in vivo* is also strongly heterogeneous with 30 to 50 µm thick sheets of collagen fibers and gap diameters often exceeding 10-20 µm[97], it remains unclear to what extent this feature influences experimental outcome. Further experiments need to assess the phenotype of the migratory cells and compare their features to invading cells *in vivo*.

In the organoid transplantation experiments, we induced constitutive *Zeb1* expression in the majority or possibly all cells. Considering that *in vivo Zeb1* expression is usually found in a minority of cells,

the ectopic expression may lead to artifacts. Given our interest in collective migration, which may be established by intersection of epithelial and mesenchymal-like tumor cells through a 'leader/follower'-mechanism[43], our approach may hamper us from identifying these events.

Avenues for future research

The current study highlight the need for future research. Below, suggestions for continuation will be discussed for the different aspects of this project.

CD44^{high}EpCAM^{low} project

In view of the CD44^{high}EpCAM^{low} project, a challenging but rewarding direction would be to couple the scRNA seq data to FACS. If E/M sub-states could be isolated from the pool of CD44^{high}EpCAM^{low} cells, there is an opportunity to further characterize partial EMT, and study the functional aspects of E/M sub-states in colorectal cancer cell lines. Our scRNA seq data can be used to identify candidate (membrane) markers for each cluster, that can be tested in FACS to see their profile in the different CD44^{high}EpCAM^{high} and CD44^{high}EpCAM^{low} populations (Supplementary Figure 11).

Further optimization of collective migration models

An interesting avenue would be the further development of collagen scaffold for the use of *ex vivo* tumor migration experiments. For example, and extension could be the pre-seeding of scaffolds with cancer associated fibroblasts (CAFs) to mimic tumor-stroma interactions. Alternatively, hybrid models could be explored to combine benefits from different models. For example, scaffolds from the collagen freeze model, perhaps even after pre-seeding with CAFs, could serve as molds for polymerization of the collagen droplet model to fill the pores. This could result in complex matrices with local architecture similar to the collagen droplet model, but a global architecture that provides structure and drives cells to invade in coordinated direction.

Alternatively, an intriguing approach could be to establish "mixed" organoids (Figure 23), containing labeled cells from both inducible and non-inducible organoids. This would enable induction in a fraction of the cells, increasing the likelihood of an induced cell to have non-induced neighbor cells. Hence, the induction of these organoids may reflect better the situation *in vivo*, and could provide examples of 'leader/follower' migration, where *Zeb1* induced cells initiate collective migration and are followed by non-induced cells.

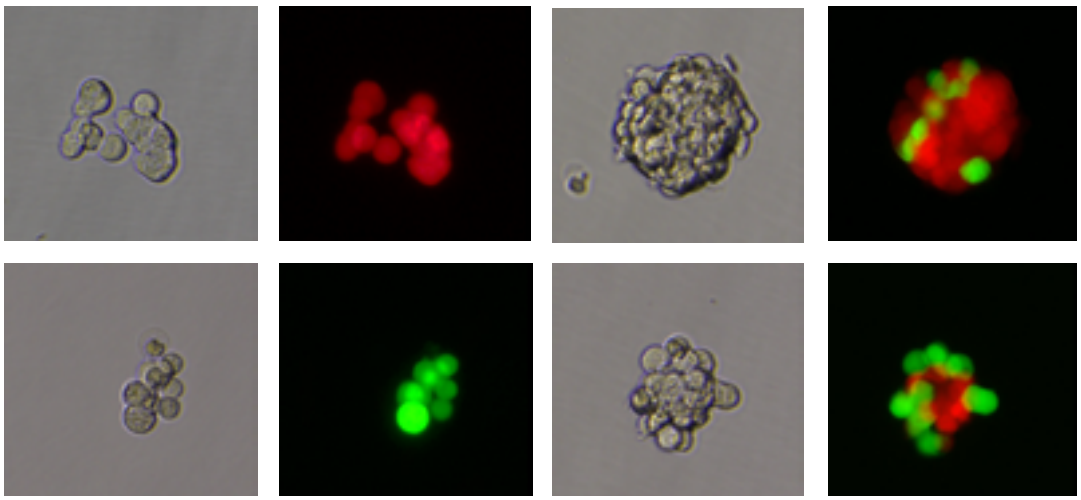


Figure 23. Example of bicolored organoids. Transduction of HCT116 with LegoGFP and mCherry. Cells can be used to form green, red or bicolored clusters by mixing different ratio's in low-attachment culture

Establishing a cause-effect relationship between partial EMT and collective migration

Our approaches for collective migration and CTC isolation provide a basis for further studies aimed at the establishment of a cause-effect relationship between partial EMT and collective migration. Subsequent studies could take advantage of phenotypic stability factors (PSF), such as NUMB, GRHL2, OVOL that have been described to promote and stabilize hybrid E/M sub-states[41,98]. Using these transcription factors, collective migration may be promoted as well as the presence of CTC clusters in the circulation. Furthermore, using a panel of EMT markers, the EMT related inter-cellular heterogeneity in CTC clusters could be studied using the approach described here.

Conclusion

Our results indicate the existence of E/M sub-states at the single cell level, which is in line with the notion that cells can remain in partial EMT, where they benefit from both epithelial and mesenchymal features. Moreover, our anecdotical results from *ex vivo* migration assay suggest that EMT induction can alter the migration modality, supporting a case for phenotypic plasticity and collective migration as mutually beneficial mechanisms.

While our results apply to context of colorectal cancer, the methodological approaches can be exploited to other types of solid cancer, and may provide points of engagements for further studies addressing the link between EMT and collective migration. Further studies need to address these issues in a systematic manner to clarify details about the complex mechanisms by which cancer metastasis proceeds.

Acknowledgments

I would first like to thank Prof. Riccardo Fodde, for giving me the opportunity to start in his lab and for his valuable guidance throughout this thesis project. And my daily supervisor, Miriam Teeuwssen, for teaching me the techniques in the lab and helping me with the interpretation of results. In addition, I want to thank my colleagues Rosalie, Roberto, Ting, Tong, and Andrea, for the joyful discussions and experimental help.

This project is the result of collaborative efforts, and I'm grateful to the people who contributed along the way. For the bioinformatics, my thanks goes to Prof. Onno Kranenburg, Jan Koster, and Wenjie Sun, who helped me during the analysis of the single cell RNA sequencing data. For the collagen models, I would like to thank Florian Markus, Antoine Khalil, Anke Husmann and Karel Bezstarosti, who contributed in different ways to the production and evaluation of anisotropic collagen scaffolds. For the imaging, my thanks goes to Esther Verhoef and Martin van Royen, for their help during 3D imaging.

I would also like to thank my assessment committee members, Prof. Guido Jenster and Martin van Royen, for taking time to evaluate the work.

References

1. International Agency for Research on Cancer, W. H. O. Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018. in PRESS RELEASE N° 263 (2018).
2. Howlader, N. et al. SEER Cancer Statistics Review 1975-2013. Natl. Cancer Inst. (2013).
3. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. Cell (1990). doi:10.1016/0092-8674(90)90186-I
4. Tariq, K. & Ghias, K. Colorectal cancer carcinogenesis: a review of mechanisms. Cancer Biol. Med. (2016). doi:10.20892/j.issn.2095-3941.2015.0103
5. Powell, S. M. et al. APC mutations occur early during colorectal tumorigenesis. Nature (1992). doi:10.1038/359235a0
6. Vogelstein, B. M. D. & Fearon, E. R. B. A. Genetic Alterations during Colorectal-Tumor Development — NEJM. N. Engl. J. Med. (1988).
7. Baker, S. J. et al. Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. Science (80-.). (1989). doi:10.1126/science.2649981
8. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. Cell 100, 57–70 (2000).
9. Luzzi, K. J. et al. Multistep nature of metastatic inefficiency: Dormancy of solitary cells after successful extravasation and limited survival of early micrometastases. Am. J. Pathol. (1998). doi:10.1016/S0002-9440(10)65628-3
10. Bernards, R. & Weinberg, R. A. A progression puzzle. Nature (2002). doi:10.1038/418823a
11. Varga, J. & Greten, F. R. Cell plasticity in epithelial homeostasis and tumorigenesis. Nat. Cell Biol. 19, 1133–1141 (2017).
12. Aiello, N. M. & Kang, Y. Context-dependent EMT programs in cancer metastasis. J. Exp. Med. (2019). doi:10.1084/jem.20181827
13. Krebs, A. M. et al. The EMT-activator Zeb1 is a key factor for cell plasticity and promotes metastasis in pancreatic cancer. Nat. Cell Biol. (2017). doi:10.1038/ncb3513
14. Tran, H. D. et al. Transient SNAIL1 expression is necessary for metastatic competence in breast cancer. Cancer Res. (2014). doi:10.1158/0008-5472.CAN-14-0923
15. Title, A. C. et al. Genetic dissection of the miR-200–Zeb1 axis reveals its importance in tumor differentiation and invasion. Nat. Commun. (2018). doi:10.1038/s41467-018-07130-z
16. Tsai, J. H., Donaher, J. L., Murphy, D. A., Chau, S. & Yang, J. Spatiotemporal regulation of epithelial-mesenchymal transition is essential for squamous cell carcinoma metastasis. Cancer Cell (2012). doi:10.1016/j.ccr.2012.09.022
17. Fodde, R. & Brabletz, T. Wnt/β-catenin signaling in cancer stemness and malignant behavior. Curr. Opin. Cell Biol. 19, 150–158 (2007).
18. Jung, H. Y., Fattet, L. & Yang, J. Molecular pathways: Linking tumor microenvironment to Epithelial-mesenchymal transition in metastasis. Clin. Cancer Res. (2015). doi:10.1158/1078-0432.CCR-13-3173
19. Brabletz, T., Jung, A., Spaderna, S., Hlubek, F. & Kirchner, T. Opinion: migrating cancer stem cells [mdash] an integrated concept of malignant tumour progression. Nat. Rev. Cancer 5, 744–749 (2005).
20. Mani, S. A. et al. The Epithelial-Mesenchymal Transition Generates Cells with Properties of Stem Cells. Cell 133, 704–715 (2008).
21. Yao, X. et al. Functional analysis of single cells identifies a rare subset of circulating tumor cells with malignant traits. Integr. Biol. (United Kingdom) (2014). doi:10.1039/c3ib40264a
22. Giuliano, M. et al. Perspective on circulating tumor cell clusters: Why it takes a village to metastasize. Cancer Res. 78, 845–852 (2018).
23. Friedl, P., Locker, J., Sahai, E. & Segall, J. E. Classifying collective cancer cell invasion. Nat. Cell Biol. (2012). doi:10.1038/ncb2548
24. Carey, S. P., Starchenko, A., McGregor, A. L. & Reinhart-King, C. A. Leading malignant cells initiate collective epithelial cell invasion in a three-dimensional heterotypic tumor spheroid model. Clin. Exp. Metastasis (2013). doi:10.1007/s10585-013-9565-x
25. Cheung, K. J., Gabrielson, E., Werb, Z. & Ewald, A. J. Collective invasion in breast cancer requires a conserved basal epithelial program. Cell (2013). doi:10.1016/j.cell.2013.11.029
26. Gaggioli, C. et al. Fibroblast-led collective invasion of carcinoma cells with differing roles for RhoGTPases in leading and following cells. Nat. Cell Biol. (2007). doi:10.1038/ncb1658
27. Labernadie, A. et al. A mechanically active heterotypic E-cadherin/N-cadherin adhesion enables fibroblasts to drive cancer cell invasion. Nat. Cell Biol. (2017). doi:10.1038/ncb3478
28. Sun, Y. et al. Prognostic value of poorly differentiated clusters in invasive breast cancer. World J. Surg. Oncol. (2014). doi:10.1186/1477-7819-12-310
29. Hou, J. M. et al. Clinical significance and molecular characteristics of circulating tumor cells and circulating tumor microemboli in patients with small-cell lung cancer. J. Clin. Oncol. (2012). doi:10.1200/JCO.2010.33.3716
30. Aceto, N. et al. Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. Cell (2014). doi:10.1016/j.cell.2014.07.013
31. Gundem, G. et al. The evolutionary history of lethal metastatic prostate cancer. Nature (2015). doi:10.1038/nature14347
32. Maddipati, R. & Stanger, B. Z. Pancreatic cancer metastases harbor evidence of polyclonality. Cancer Discov. (2015). doi:10.1158/2159-8290.CD-15-0120
33. Cheung, K. J. et al. Polyclonal breast cancer metastases arise from collective dissemination of keratin 14-expressing tumor cell

clusters. Proc. Natl. Acad. Sci. (2016). doi:10.1073/pnas.1508541113

34. Nieto, M. A., Huang, R. Y. Y. J., Jackson, R. A. A. & Thiery, J. P. P. EMT: 2016. Cell (2016). doi:10.1016/j.cell.2016.06.028
35. Jordan, N. V., Johnson, G. L. & Abell, A. N. Tracking the intermediate stages of epithelial-mesenchymal transition in epithelial stem cells and cancer. Cell Cycle (2011). doi:10.4161/cc.10.17.17188
36. Huang, R. Y. J. 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 Dis. (2013). doi:10.1038/cddis.2013.442
37. Zhang, J. et al. TGF-β-induced epithelial-to-mesenchymal transition proceeds through stepwise activation of multiple feedback loops. Sci. Signal. (2014). doi:10.1126/scisignal.2005304
38. Pastushenko, I. et al. Identification of the tumour transition states occurring during EMT. Nature (2018). doi:10.1038/s41586-018-0040-3
39. Vergara, D. et al. Epithelial-mesenchymal transition in ovarian cancer. Cancer Lett. (2010). doi:10.1016/j.canlet.2009.09.017
40. Latil, M. et al. Cell-Type-Specific Chromatin States Differentially Prime Squamous Cell Carcinoma Tumor-Initiating Cells for Epithelial to Mesenchymal Transition. Cell Stem Cell (2017). doi:10.1016/j.stem.2016.10.018
41. Jolly, M. K. et al. Stability of the hybrid epithelial/mesenchymal phenotype. Oncotarget (2016). doi:10.18632/oncotarget.8166
42. Hong, T. et al. An Ovol2-Zeb1 Mutual Inhibitory Circuit Governs Bidirectional and Multi-step Transition between Epithelial and Mesenchymal States. PLoS Comput. Biol. (2015). doi:10.1371/journal.pcbi.1004569
43. Rørth, P. Fellow travellers: emergent properties of collective cell migration. EMBO Rep. (2012). doi:10.1038/embor.2012.149
44. Haeger, A., Wolf, K., Zegers, M. M. & Friedl, P. Collective cell migration: Guidance principles and hierarchies. Trends in Cell Biology (2015). doi:10.1016/j.tcb.2015.06.003
45. Jolly, M. K., Mani, S. A. & Levine, H. Hybrid epithelial/mesenchymal phenotype(s): The 'fittest' for metastasis? Biochimica et Biophysica Acta - Reviews on Cancer (2018). doi:10.1016/j.bbcan.2018.07.001
46. Campbell, K. & Casanova, J. A common framework for EMT and collective cell migration. Development (2016). doi:10.1242/dev.139071
47. Stott, S. L. et al. Isolation of circulating tumor cells using a microvortex-generating herringbone-chip. PNAS 107, 18392–7 (2010).
48. Au, S. H. et al. Microfluidic isolation of circulating tumor cell clusters by size and asymmetry. Sci. Rep. (2017). doi:10.1038/s41598-017-01150-3
49. Fumagalli, A. et al. Genetic dissection of colorectal cancer progression by orthotopic transplantation of engineered cancer organoids. Proc. Natl. Acad. Sci. (2017). doi:10.1073/pnas.1701219114
50. Kranenburg, O., Maurice, M., Beekman, J. & Brousali, A. Utrecht Platform for Organoid Technology. (2019). Available at: https://www.uu.nl/en/research/life-sciences/research/hubs/utrecht-platform-for-organoid-technology.
51. Francescangeli, F. et al. Proliferation state and polo-like kinase1 dependence of tumorigenic colon cancer cells. Stem Cells (2012). doi:10.1002/stem.1163
52. Vermeulen, L. et al. Wnt activity defines colon cancer stem cells and is regulated by the microenvironment. Nat. Cell Biol. (2010). doi:10.1038/ncb2048
53. Emmink, B. L. et al. Differentiated human colorectal cancer cells protect tumor-initiating cells from irinotecan. Gastroenterology (2011). doi:10.1053/j.gastro.2011.03.052
54. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (2009). doi:10.1093/bioinformatics/btp616
55. Ritchie, M. E. et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. (2015). doi:10.1093/nar/gkv007
56. Team, R. D. C. & R Development Core Team, R. R: A Language and Environment for Statistical Computing. R Found. Stat. Comput. (2016). doi:10.1007/978-3-540-74686-7
57. Chen, E. Y. et al. Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics (2013). doi:10.1186/1471-2105-14-128
58. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. Nat. Commun. (2017). doi:10.1038/ncomms14049
59. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. (2018). doi:10.1038/nbt.4096
60. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. J. Open Source Softw. (2018). doi:10.21105/joss.00861
61. Eide, P. W., Bruun, J., Lothe, R. A. & Sveen, A. CMScaller: An R package for consensus molecular subtyping of colorectal cancer pre-clinical models. Sci. Rep. (2017). doi:10.1038/s41598-017-16747-x
62. van Dijk, D. et al. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell (2018). doi:10.1016/j.cell.2018.05.061
63. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: Gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics (2013). doi:10.1186/1471-2105-14-7
64. Kolde, R. pheatmap: Pretty Heatmaps. Available at: https://github.com/raivokolde/pheatmap/issues.
65. Cannoodt, R. et al. SCORPIUS improves trajectory inference and identifies novel modules in dendritic cell development. bioRxiv (2016).

66. Qiu, X. et al. Single-cell mRNA quantification and differential analysis with Census. Nat. Methods (2017). doi:10.1038/nmeth.4150

67. Setty, M. et al. Palantir characterizes cell fate continuities in human hematopoiesis. bioRxiv (2018). doi:10.1101/385328

68. Koster, J. R2: Genomics Analysis and Visualization Platform. (2019). Available at: http://r2.amc.nl.

69. Krämer, A., Green, J., Pollard, J. & Tugendreich, S. Causal analysis approaches in ingenuity pathway analysis. Bioinformatics (2014). doi:10.1093/bioinformatics/btt703

70. Fumagalli, A. et al. A surgical orthotopic organoid transplantation approach in mice to visualize and study colorectal cancer progression. Nat. Protoc. (2018). doi:10.1038/nprot.2017.137

71. van Royen, M. E. et al. Three-dimensional microscopic analysis of clinical prostate specimens. Histopathology (2016). doi:10.1111/his.13022

72. Xu, C. & Su, Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics (2015). doi:10.1093/bioinformatics/btv088

73. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science (80-.). (2016). doi:10.1126/science.aad0501

74. Vega, S. et al. Snail blocks the cell cycle and confers resistance to cell death. Genes Dev. (2004). doi:10.1101/gad.294104

75. Rajput, A. et al. Characterization of HCT116 Human Colon Cancer Cells in an Orthotopic Model. J. Surg. Res. (2008). doi:10.1016/j.jss.2007.04.021

76. Chowdhury, S. et al. Intra-Tumoral Heterogeneity in Metastatic Potential and Survival Signaling between Iso-Clonal HCT116 and HCT116b Human Colon Carcinoma Cell Lines. PLoS One (2013). doi:10.1371/journal.pone.0060299

77. De Smedt, L. et al. Expression profiling of budding cells in colorectal cancer reveals an EMT-like phenotype and molecular subtype switching. Br. J. Cancer (2017). doi:10.1038/bjc.2016.382

78. George, J. T., Jolly, M. K., Xu, S., Somarelli, J. A. & Levine, H. Survival outcomes in cancer patients predicted by a partial EMT gene expression scoring metric. Cancer Res. (2017). doi:10.1158/0008-5472.CAN-16-3521

79. Tan, T. Z. et al. Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. EMBO Mol. Med. (2014). doi:10.15252/emmm.201404208

80. Jia, D. et al. Testing the gene expression classification of the EMT spectrum. Phys. Biol. (2019). doi:10.1088/1478-3975/aaf8d4

81. Li, H. et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat. Genet. (2017). doi:10.1038/ng.3818

82. Setty, M. et al. Characterization of cell fate probabilities in single-cell data with Palantir. Nat. Biotechnol. (2019). doi:10.1038/s41587-019-0068-4

83. Gritsenko, P., Leenders, W. & Friedl, P. Recapitulating in vivo-like plasticity of glioma cell invasion along blood vessels and in astrocyte-rich stroma. Histochem. Cell Biol. (2017). doi:10.1007/s00418-017-1604-2

84. Guo, L., Chen, D., Yin, X. & Shu, Q. GSK-3β Promotes Cell Migration and Inhibits Autophagy by Mediating the AMPK Pathway in Breast Cancer. Oncol. Res. Featur. Preclin. Clin. Cancer Ther. (2018). doi:10.3727/096504018x15323394008784

85. Plou, J. et al. From individual to collective 3D cancer dissemination: roles of collagen concentration and TGF-β. Sci. Rep. (2018). doi:10.1038/s41598-018-30683-4

86. Tan, Y., Peng, J., Wei, D., Chen, P. & Zhao, Y. Effect of Jagged1 on the proliferation and migration of colon cancer cells. Exp. Ther. Med. (2012). doi:10.3892/etm.2012.549

87. Ren, Y. et al. Hepatocyte growth factor promotes cancer cell migration and angiogenic factors expression: A prognostic marker of human esophageal squamous cell carcinomas. Clin. Cancer Res. (2005). doi:10.1158/1078-0432.CCR-04-2553

88. Wang, X. et al. Girdin/GIV regulates collective cancer cell migration by controlling cell adhesion and cytoskeletal organization. Cancer Sci. (2018). doi:10.1111/cas.13795

89. Doyle, A. D. Generation of 3D collagen gels with controlled diverse architectures. Curr. Protoc. Cell Biol. (2016). doi:10.1002/cpcb.9

90. Sung, K. E. et al. Control of 3-dimensional collagen matrix polymerization for reproducible human mammary fibroblast cell culture in microfluidic devices. Biomaterials (2009). doi:10.1016/j.biomaterials.2009.05.043

91. Chaubaroux, C. et al. Cell Alignment Driven by Mechanically Induced Collagen Fiber Alignment in Collagen/Alginate Coatings. Tissue Eng. Part C Methods (2015). doi:10.1089/ten.tec.2014.0479

92. Campbell, J. J., Husmann, A., Hume, R. D., Watson, C. J. & Cameron, R. E. Development of three-dimensional collagen scaffolds with controlled architecture for cell migration studies using breast cancer cell lines. Biomaterials (2017). doi:10.1016/j.biomaterials.2016.10.048

93. Hume, R. D. et al. Tumour cell invasiveness and response to chemotherapeutics in adipocyte invested 3D engineered anisotropic collagen scaffolds. Sci. Rep. (2018). doi:10.1038/s41598-018-30107-3

94. Knibbe-Hollinger, J. S. et al. Influence of age, irradiation and humanization on NSG mouse phenotypes. Biol. Open (2015). doi:10.1242/bio.013201

95. Ahmed, D. et al. Epigenetic and genetic features of 24 colon cancer cell lines. Oncogenesis (2013). doi:10.1038/oncsis.2013.35

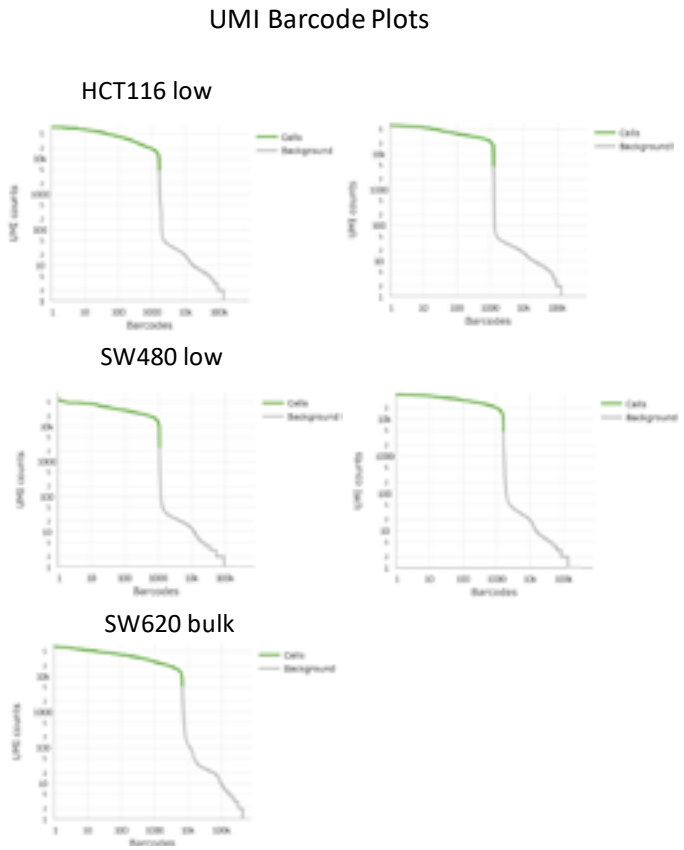
96. Li, C. & Balazsi, G. A landscape view on the interplay between EMT and cancer metastasis. npj Syst. Biol. Appl. (2018). doi:10.1038/s41540-018-0068-x

97. Wolf, K. et al. Collagen-based cell migration models in vitro and in vivo. Seminars in Cell and Developmental Biology (2009). doi:10.1016/j.semcdb.2009.08.005

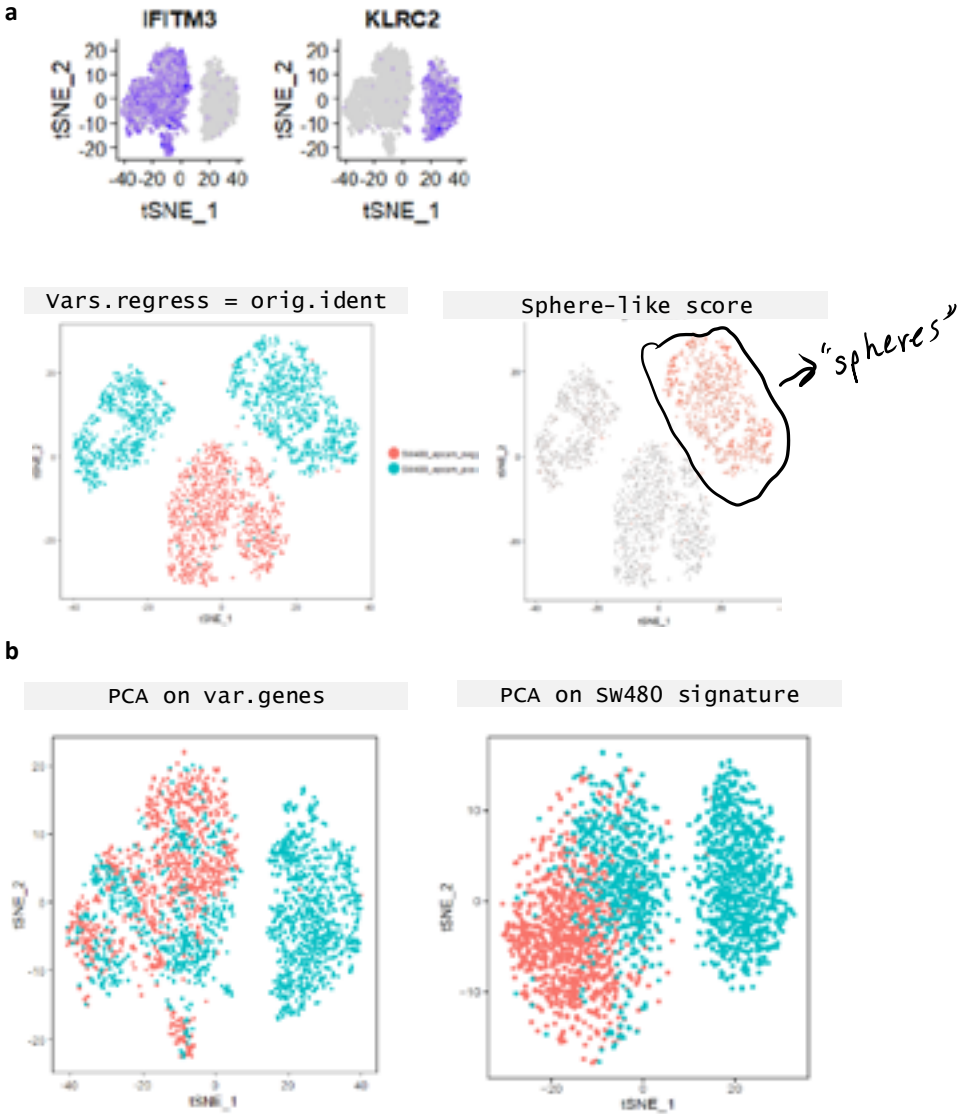
98. Bocci, F. et al. Numb prevents a complete epithelial-mesenchymal transition by modulating Notch signaling. J. R. Soc. Interface (2017). doi:10.1098/rsif.2017.0512

Appendix

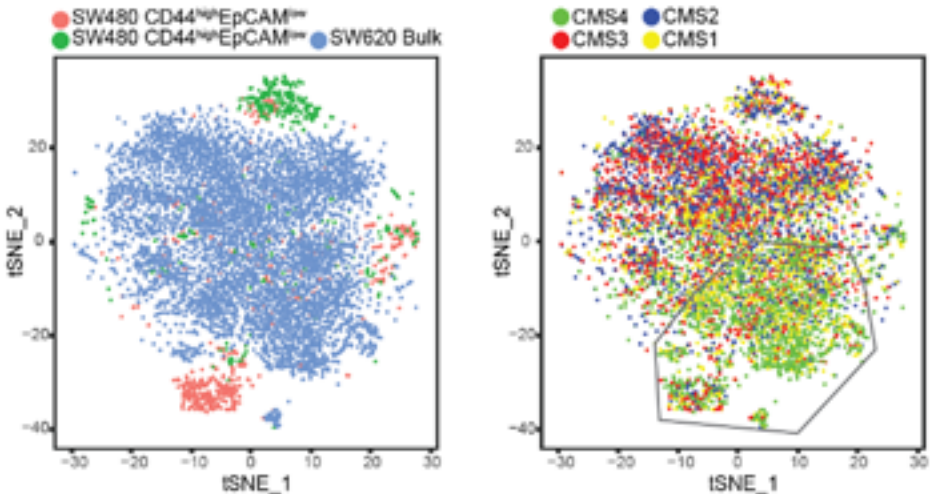
Supplementary Figures



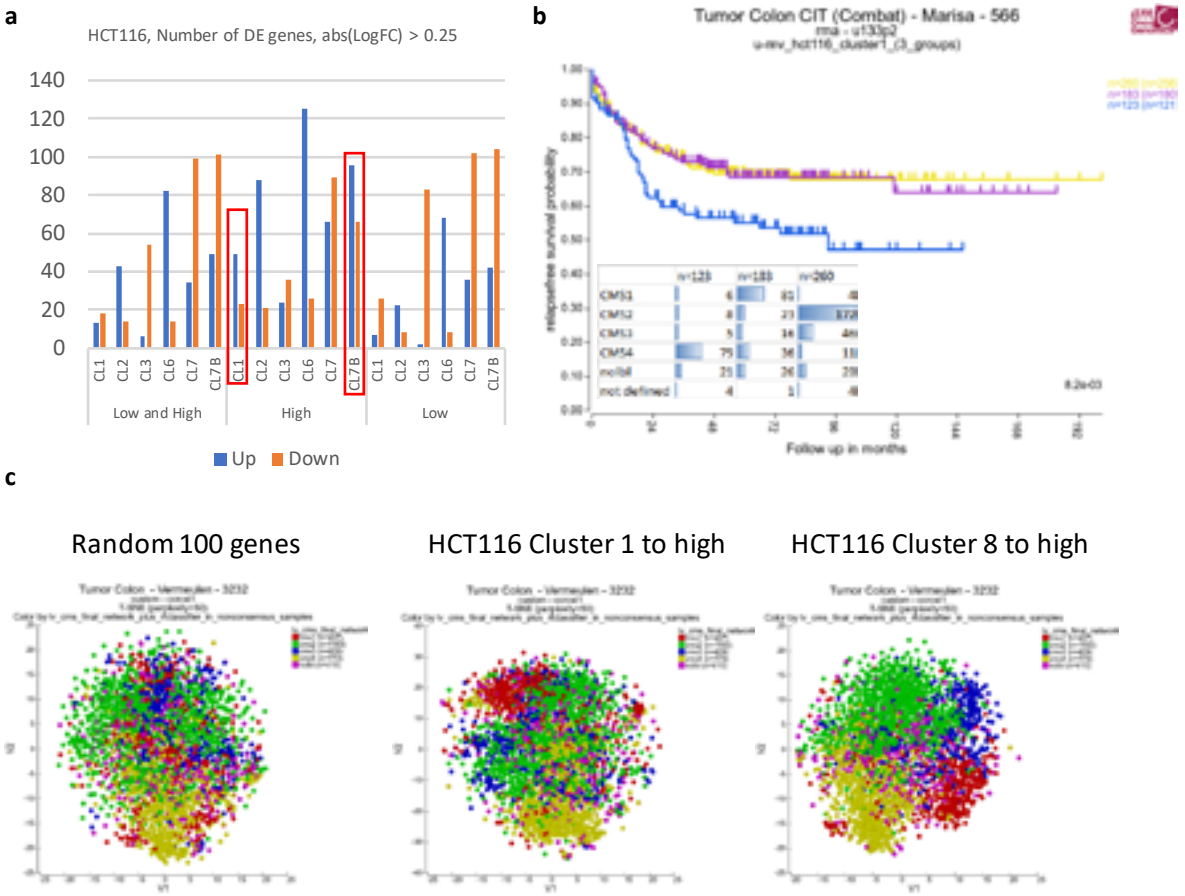
Supplementary Figure 1. UMI barcode plots of scRNAseq results.



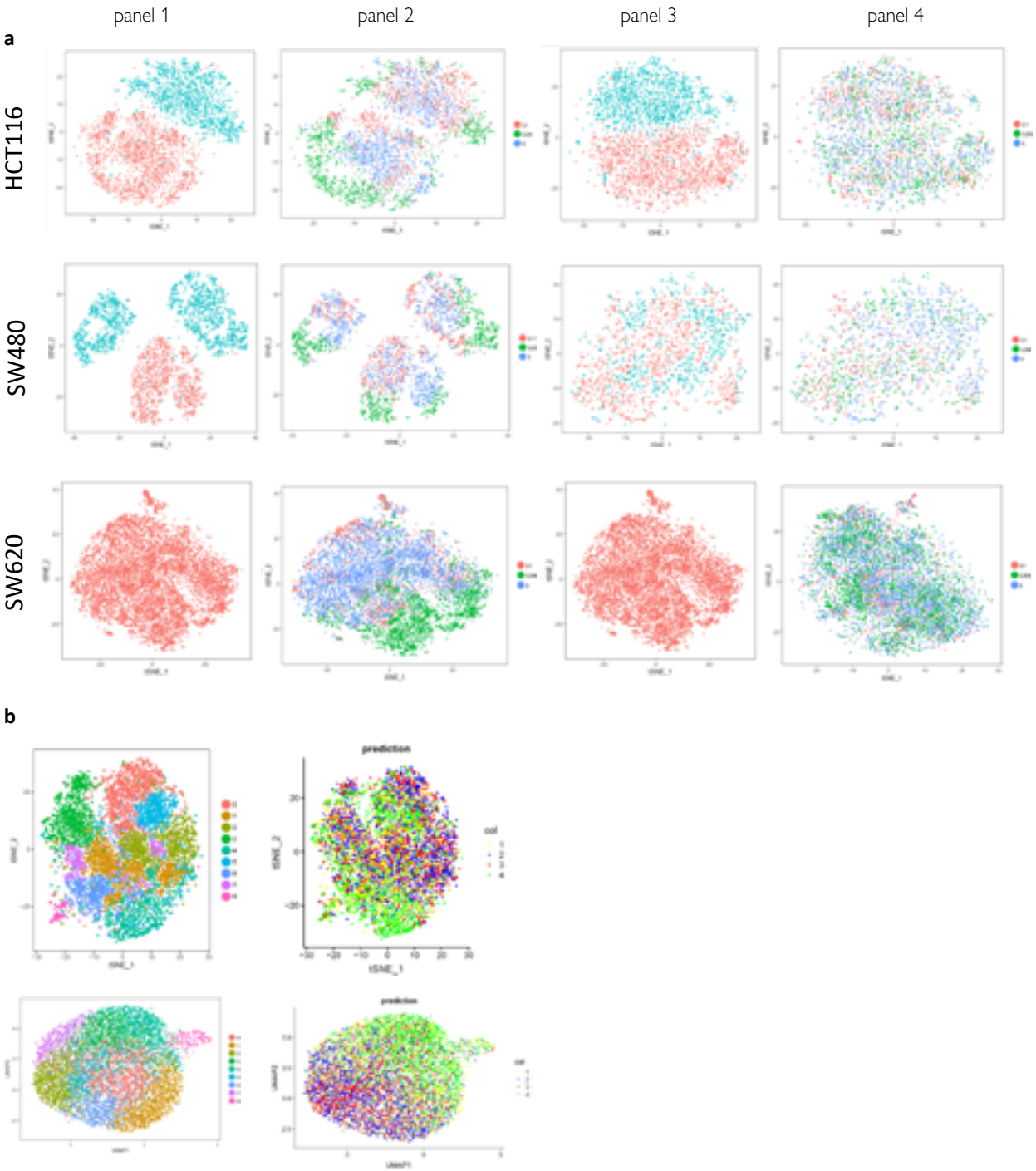
Supplementary Figure 2. Identification of "spheres" in SW480 scRNAseq. (a) Top: two markers for adherent cells and "sphere" population derived from DE analysis in the bulk RNA seq. Down: gene set score based on: Av_score = Average(UP_markers) - Average(DOWN_markers) for all markers found by DE analysis between spheres and adherent cells. (b) $CD44^{high}EpCAM^{low}$ cells and $CD44^{high}EpCAM^{high}$ cells are resolved once the dimension reduction (here PCA) is performed on the input list of DE markers derived from the bulk RNA seq by comparing $CD44^{high}EpCAM^{low}$ cells to $CD44^{high}EpCAM^{high}$ cells.



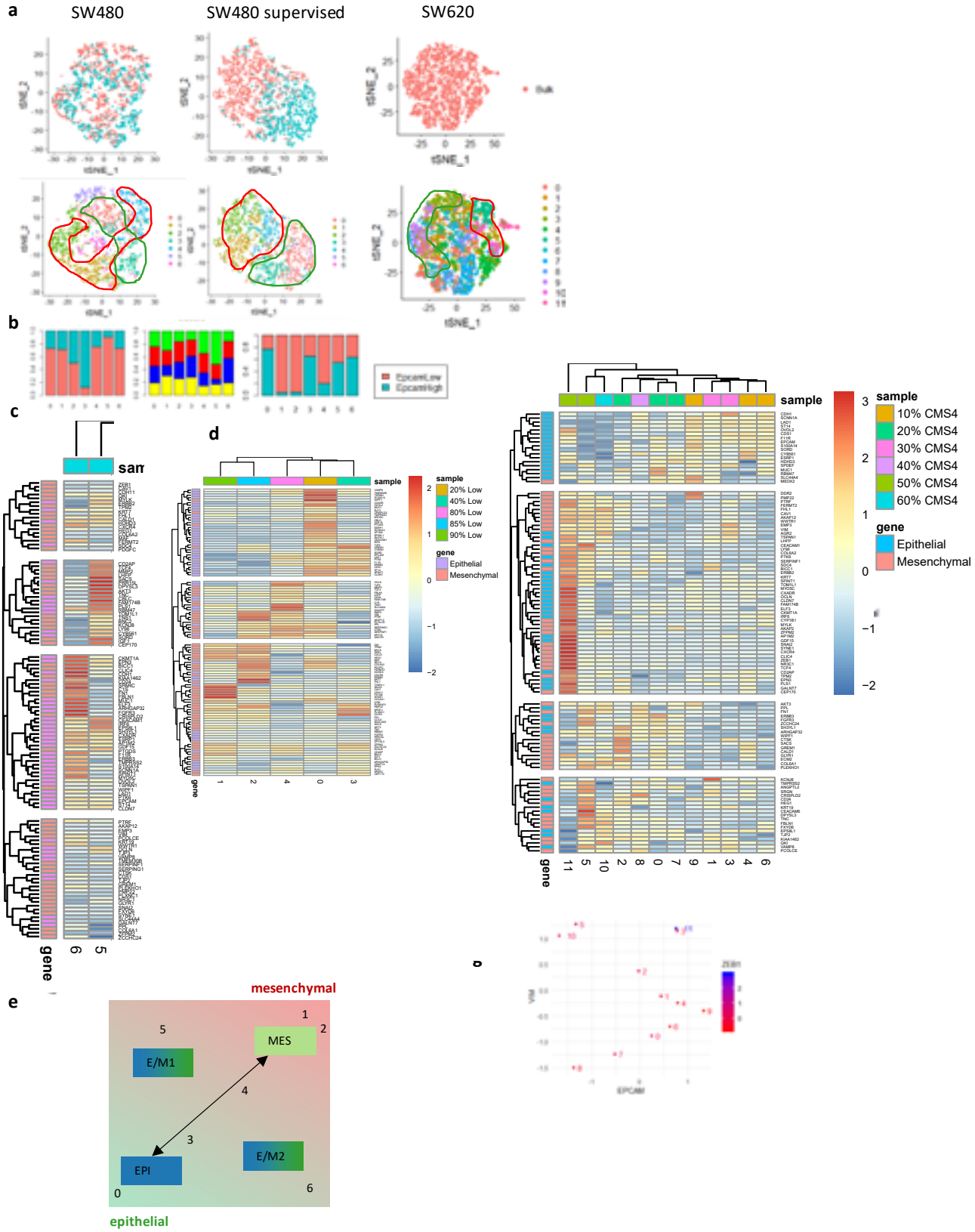
Supplementary Figure 3. Aggregated tSNE plot. SW480 CD44^{high}EpCAM^{low} cells locate in proximity of SW620 cells that were predicted as CMS4 using the CMScaller algorithm.



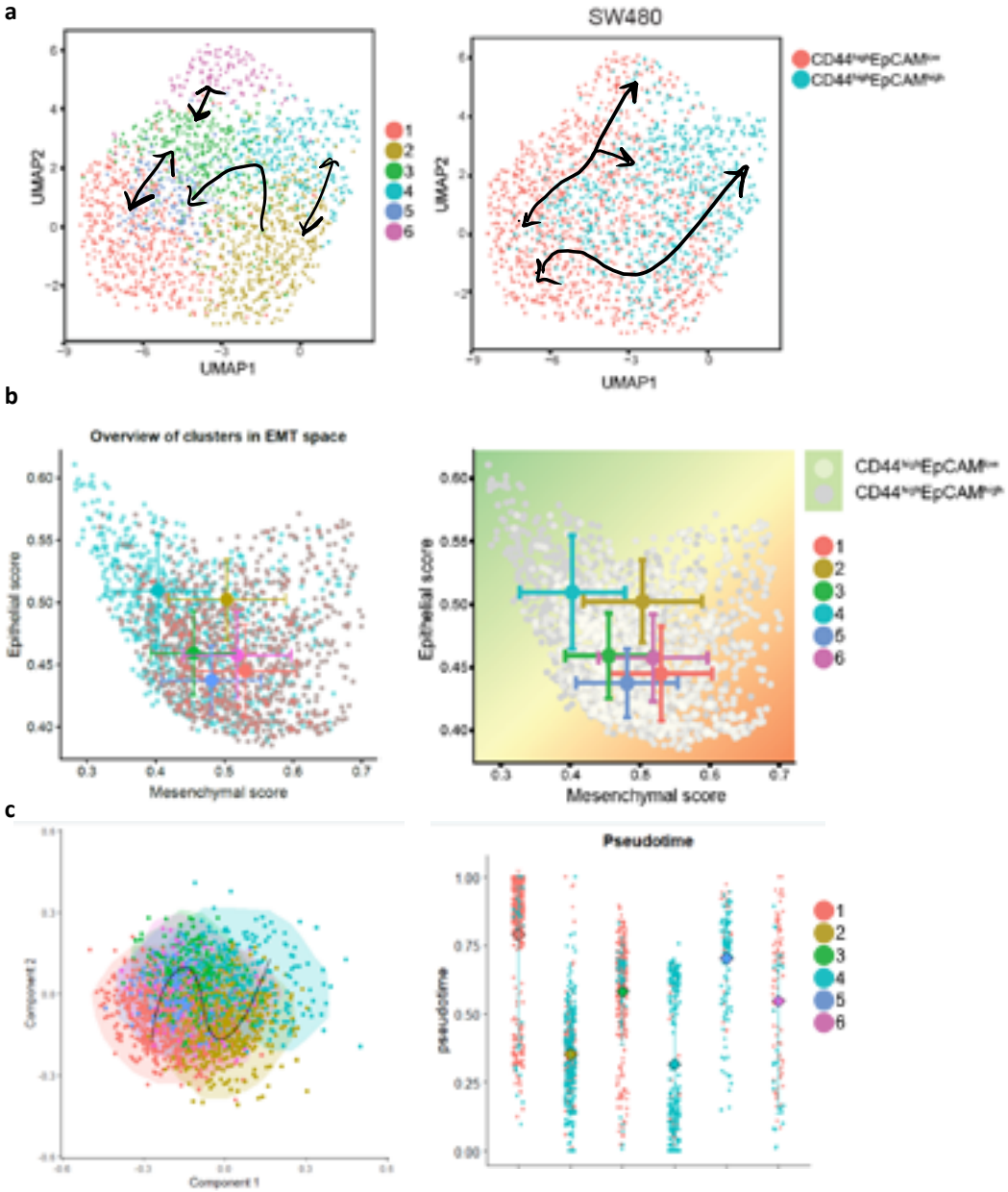
Supplementary Figure 4. ScRNAseq data and patient RNA data. (a) Barplot showing number of up- and down-regulated genes for different clusters among the pool of HCT116 cells (comparison: cluster versus low_and_high, low, and high). (b) Example of Kaplan-Meijer curve showing relapse-free survival for patient groups in the Vermeulen cohort (accessed in R2 server) that were stratified using k-means clustering based on an input signature from $CD44^{high}EpCAM^{low}$ cluster 1. (c) tSNE plot showing patients with their classification of CMS1-4. Dimension reduction was done on i) random 100 genes (left), ii) the input signature derived from $CD44^{high}EpCAM^{low}$ cluster 1 (middle), and iii) the input signature derived from $CD44^{high}EpCAM^{low}$ cluster 8 (right).



Supplementary Figure 5. Effect of cell cycle correction on scRNA seq results. (a) Panel 1: tSNE plots showing orig.ident for each of the cell lines. Panel 2: assigned cell cycle phases based on Seurat cell cycle scoring. Panel 3: tSNE after cell cycle correction, and for SW480 removal of "sphere" population. Panel 4: assigned cell cycle phases after cell cycle correction. (b) Top: unsupervised clustering results of SW620 before cell cycle correction, and prediction of CMS. on tSNE Bottom: unsupervised clustering results for SW620 after cell cycle correction and prediction of CMS on UMAP. CMS4 cells cluster together after cell cycle correction.

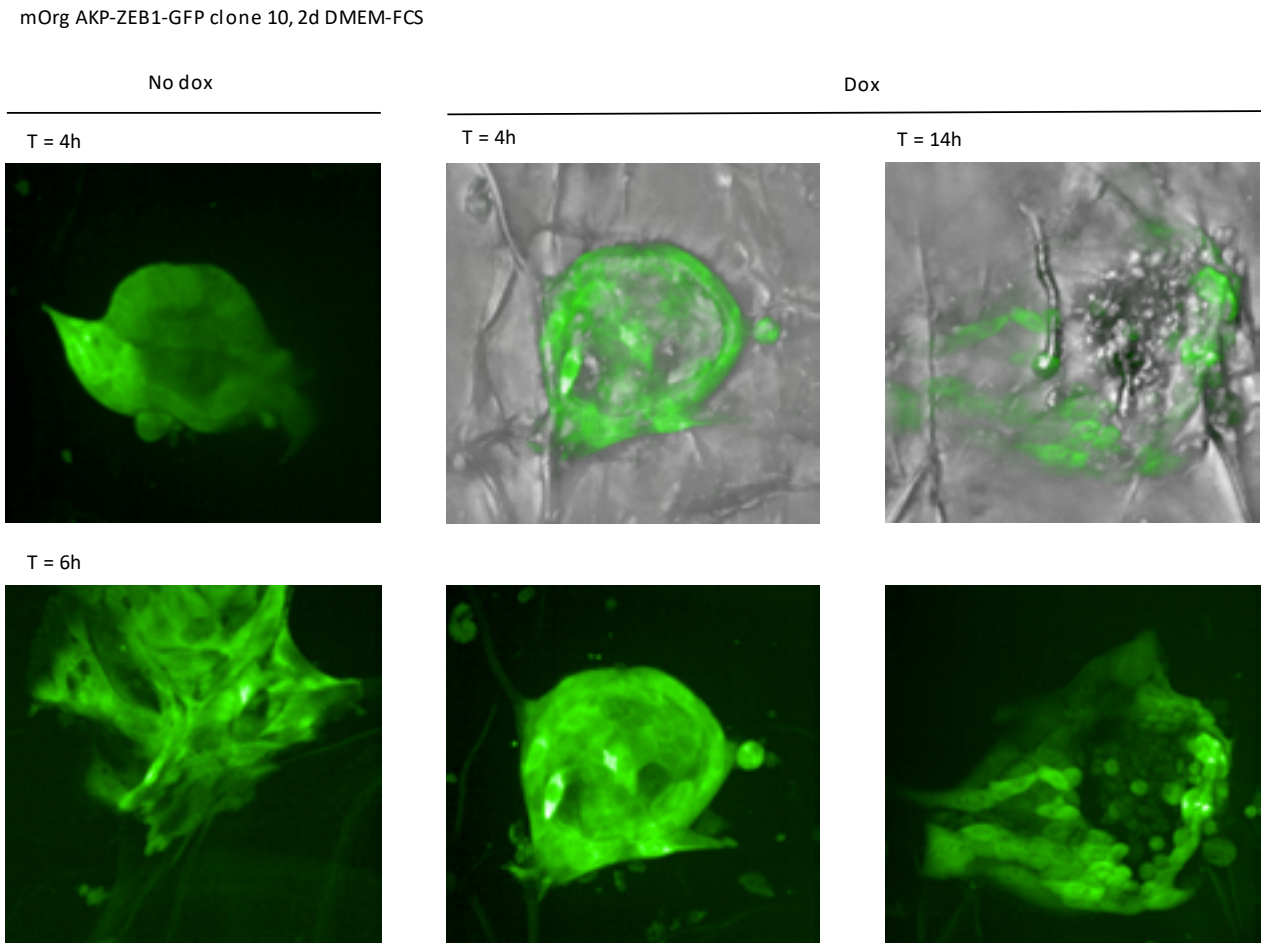


Supplementary Figure 6. Cross-cluster comparison in SW480 and SW620 (a) Unsupervised clustering results for SW480 and SW620 using Seurat v3. (b) Distribution of orig.ident and CMS predictions across unsupervised clusters. (c) Heatmap for SW480, focused on cluster 5 and cluster 6. (d) Heatmap of SW480 clusters without cluster 5 and cluster 6. (e) Schematic diagram for SW480 cluster profiles across EMT. (f) Heatmap showing SW620 clusters for the EMT gene list. (g) Scatterplot showing ZEB1 expression in clusters versus their VIM and EPCAM expression.

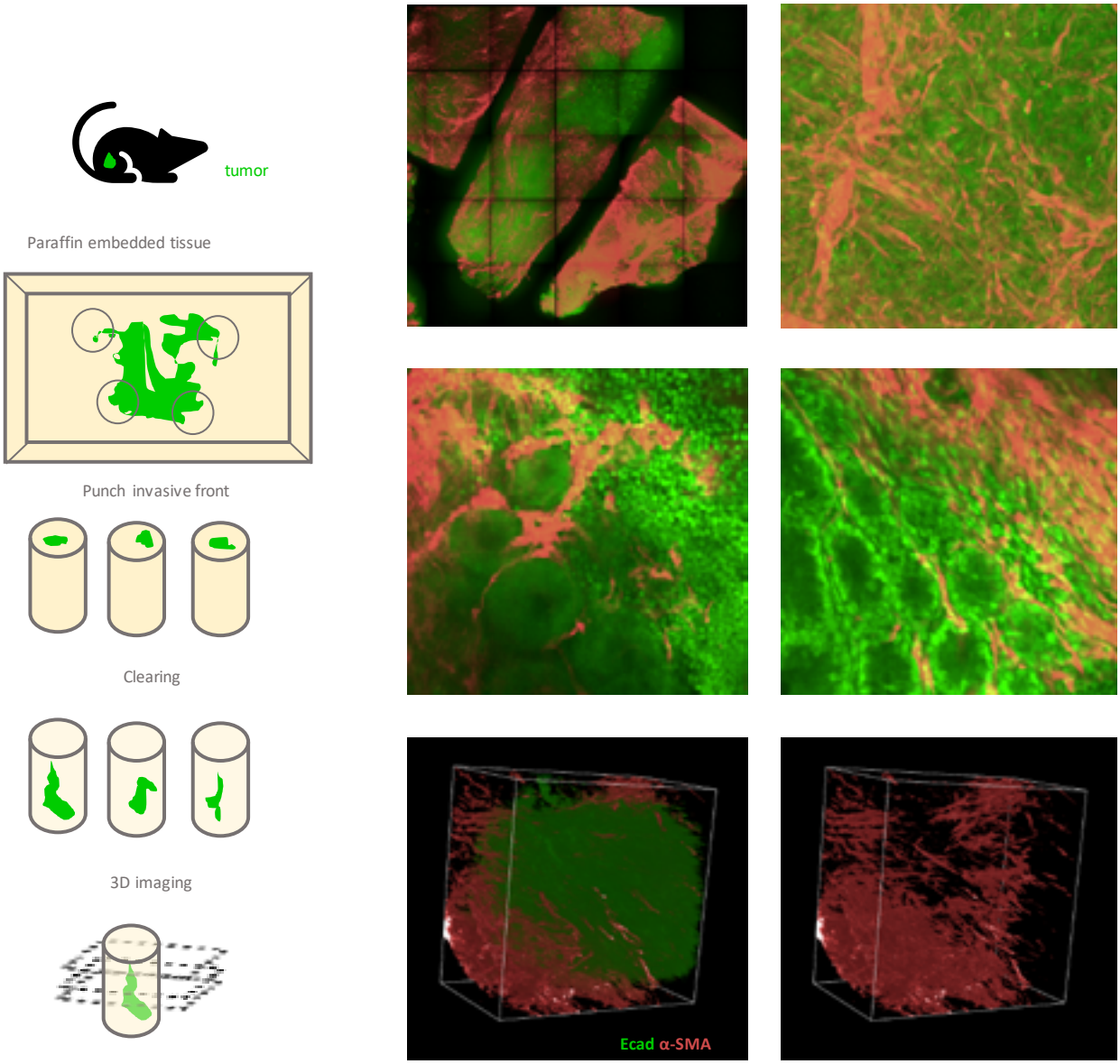


Supplementary Figure 7. EMT and pseudotemporal analysis of SW480. (a) Unsupervised clustering results for SW480 and the orig.ident of CD44^{high}EpCAM^{high} and CD44^{high}EpCAM^{low} cells. (b) EMT plot for SW480 showing individual cells with their orig.ident (left) and mapping of unsupervised clustering results (right). Here, cluster 2 appears in E/M-hybrid state, cluster 4 appears most epithelial and cluster 1 most mesenchymal. (c) In pseudo-time, cluster 2 maps in between cluster 4 and the other, more mesenchymal clusters. Left: trajectory plot from Scorpius. Right: Pseudotimes for individual cells, categorized according to unsupervised clusters.

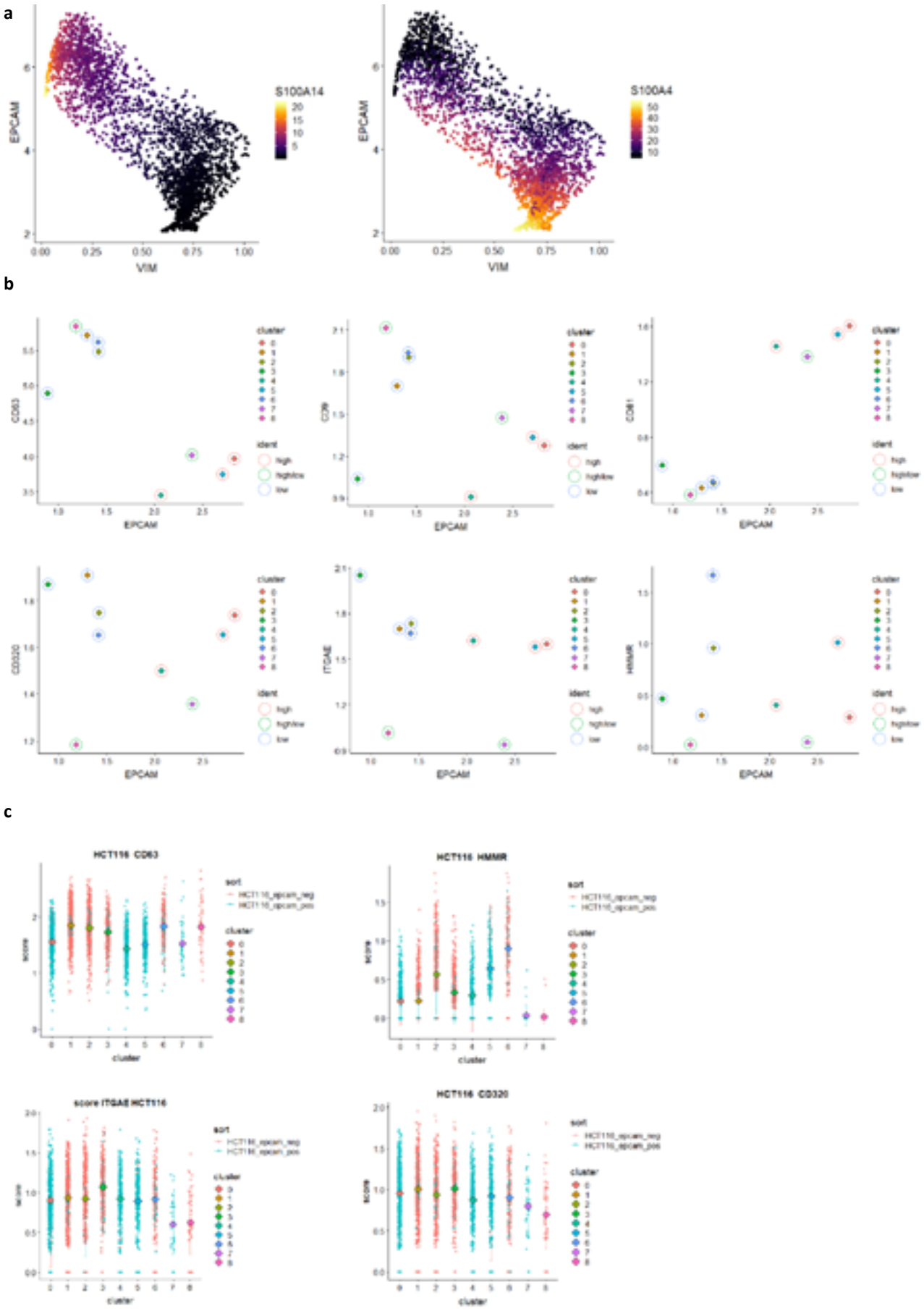
Supplementary Figure 8. Lentiviral vector for Tet-based inducible *Zeb1*.



Supplementary Figure 9. Snapshots from 48h timelapse imaging of AKP-*Zeb1*-GFP mOrg on the collagen freeze scaffold.



Supplementary Figure 10. Experimental approach for 3D imaging of paraffin embedded tissue. Cylinders were punched out of the invasive front of a paraffin embedded tumor. Cylinders were cleared and stained for E-cad and aSMA to seperate epithelial cells (green) from fibroblasts (red).



Supplementary Figure 11. Three approaches to visualize candidate markers, specific for one of the unsupervised clusters (in this case cluster 8, HCT116). (a) Scatterplots with imputed values using MAGIC show a switch between S100A14 and S100A4 over CD44^{high}EpCAM^{low} and CD44^{high}EpCAM^{high} cells. (b) Average cluster values for cluster 8 specific markers. (c) Categorical scatter for candidate markers to isolate cluster 8.

Supplementary Tables

SW480 Low vs High DE genes FDR < 0.01, LogFC > 1.5

ABCB1	AL121944.1	CDH11	FGF19	KRT32	MUCL1	RF00019	STEAP2
ABCB4	AL163952.1	CDH3	FGF3	KRT35	MYO16-AS1	RF00019	STYK1
ABCG2	AL354861.2	CDS1	FHDC1	KRT36	MYO1F	RF00019	SUSD2
ABHD12B	AL357033.2	CFAP161	FOXA2	KRT37	MYO5B	RF00190	SYDE1
AC002076.2	AL390719.1	CHN2	FP236240.1	KRT38	NACAP1	RFTN1	SYK
AC003958.2	AL391840.1	CHST2	FRMD6	KRT4	NANOS1	RGL3	SYT12
AC004066.2	AL512488.1	CLDN7	FRY	KRT41P	NCAM1	RGS5	SYT13
AC004540.1	AL683807.1	CMTM3	FRZB	KRT43P	NIPAL1	RGS7	SYT8
AC004540.2	ALK	CNTNAP3	GABRE	KRT71	NPM1P38	RIMS4	TBC1D30
AC004830.1	ALOXE3	CNTNAP3B	GFI1	KRTAP2-5P	NPTX2	RIPK3	TC2N
AC004947.1	ALPK2	CNTNAP3P2	GGT5	LAD1	NREP	RIPK4	TCAF2
AC005865.1	ALPP	COBL	GJB2	LAMA3	NT5DC4	RLBP1	TCIM
AC005865.2	ALPPL2	COBLL1	GJB3	LAMB3	NTRK2	RN7SL678P	TESC
AC006042.1	ALS2CL	COL1A1	GLS2	LAMC2	NTSR1	RNA5SP479	TESMIN
AC006511.1	AMOTL1	CPA4	GNAI1	LARGE2	NUAK1	RNU2-32P	TG
AC007952.2	ANP32BP3	CPA5	GNAL	LCK	OGFRL1	RNU4ATAC18I	TGFB2
AC007952.6	AP000943.1	CPQ	GNAO1	LCN12	OLR1	RNU6-1153P	TGFB2-AS1
AC008013.1	AP000943.2	CPVL	GNG11	LHX1	OVOL2	RNU6-1161P	TH
AC008610.1	AP001631.1	CPXM1	GPRC5A	LIMS2	PARM1	RNU6-1238P	THAP12P8
AC008957.1	AP002800.1	CRB3	GULP1	LINC00460	PDE10A	RNU6-1318P	TKTL1
AC010503.4	AP1M2	CRISPLD2	GYG2	LINC01116	PDGFRB	RNU6-531P	TMC4
AC010768.4	APBA1	CST1	HCG9P5	LINC01173	PDZRN3	RNU6-80P	TMEM125
AC022126.1	AR	CST6	HDAC9	LINC01356	PELI2	RNU6-91P	TMEM30B
AC026316.2	AREG	CTTNBP2	HKDC1	LINC02041	PGM5	RPL32P33	TMEM52B
AC034206.1	ARHGAP8	CXCL14	HLA-V	LINC02320	PHACTR3	RPS14P4	TNNC1
AC068491.2	ARHGEF16	CYP24A1	HMGCS2	LINC02331	PHLDA3	SCNN1A	TNNT2
AC068491.3	ARHGEF38	DAB2	HMX3	LINC02457	PIP5K1B	SDR16C5	TNS1
AC068580.2	ARL15	DACT1	HOOK1	LLGL2	PKP3	SEL1L3	TPD52L1
AC080037.1	ASCL2	DCDC2	HOXD10	LOXL3	PLA2G4D	SEMA5A	TPK1
AC084346.1	ASTN2	DENND1C	HOXD8	LTBP1	PLCE1-AS1	SERPINB5	TPRXL
AC092138.2	ATP6AP1L	DMBX1	HSD11B2	LYG1	PLEK2	SERPINF2	TRBC2
AC092299.1	B3GALT5	EDN3	HSD17B2	MAL2	PLEKHN1	SGPP2	TRBV30
AC092807.2	B3GNT3	EFNA5	HSH2D	MAOA	PLPPR4	SH2D3A	TREML2
AC092807.3	BDNF	ELMO3	ICA1	MAP7	PMEPA1	SHANK2	TRIM15
AC093162.2	BICDL2	ELOVL7	IFIH1	MAPK13	PPARG	SIGIRR	TSPAN1
AC093673.2	BSPRY	ENTPD2	IFITM1	MARVELD3	PPM1H	SIGLEC6	TSPAN11
AC098934.3	BST2	ENTPD8	IGFBP7	MBNL3	PRICKLE1	SLA	TSPAN15
AC106017.2	BTBD16	EPCAM	IGFBPL1	MBP	PRKAR2B	SLC12A7	TUBBP5
AC114296.1	BX322635.1	EPHA1	IL23A	MCTP2	PRRG2	SLC1A3	VIL1
AC114550.2	C1orf210	EPHB6	ILDR1	MDFI	PRRG4	SLC22A17	VWA2
AC119427.1	C1S	EPPK1	INHBB	MELTF	PRSS16	SLC22A20P	VWDE
AC121764.3	C2orf54	ERBB3	IQANK1	MIR10B	PRSS22	SLC25A48	WFDC2
AC124319.1	C4BPB	ESAM	IRF6	MIR200CHG	PRSS23	SLC40A1	WNT10A
AC127521.1	C6orf132	ESPN	ISM1	MIR2355	PRSS56	SLCO4A1-AS1	WNT7A
AC215522.2	C9orf84	ESPNP	ITGB4	MIR3179-2	PSG9	SNHG18	XDH
AC231657.2	CACNG6	ESRP1	ITGB6	MIR3179-4	PTK6	SNORA22C	YPEL2
ACOXL	CADPS	ESRP2	ITGB8	MIR3677	QPRT	SNORD117	Z69720.2
ACP7	CADPS2	F11R	ITIH3	MIR429	RAB17	SPARC	ZAP70
ADAMTS8	CALD1	F2RL1	KCND3	MIR4477A	RAPGEF5	SPINT1	ZDHHC20-IT1
ADAP1	CAMK1D	FA2H	KDF1	MIR4635	RASAL1	SPINT1-AS1	ZEB1
ADGRG1	CAPN8	FAM131B	KITLG	MISP	RASGEF1C	SPNS2	ZNF165
ADORA2B	CASC18	FAM157A	KLF7	MITF	RASGRF1	SPTLC3	ZNF204P
AF064858.2	CAV1	FAM183A	KLK10	MPP4	RF00019	ST14	ZNF521
AFF3	CAVIN2	FAM84B	KRT13	MPZL2	RF00019	ST3GAL5	ZSCAN12P1
AL021920.2	CD70	FBLN1	KRT15	MROH6	RF00019	ST6GAL1	
AL034376.1	CDC42BPG	FBN3	KRT16P6	MRPL35P2	RF00019	ST6GALNAC3	
AL049836.1	CDH1	FGD4	KRT19	MST1R	RF00019	STARD4-AS1	

Supplementary Table 1A. SW480 DE gene list bulk RNA seq.

HCT116 Low vs High DE genes FDR < 0.01, LogFC > 1.5							
ABCG1	AL590399.4	CNTNAP3P2	FXYD3	LAMB1	MIR6730	RGS6	ST14
AC005046.1	ALOXE3	COL13A1	GALNT3	LAMC2	MIR6856	RNA5SP152	ST8SIA6
AC009237.6	ANK1	COL9A3	GLIS3	LIMS2	MORC4	RNU4-78P	SUSD2
AC009237.9	ANK3	CR2	GPR176	LINC01405	MRC2	RNU5E-10P	SYDE1
AC009238.1	ANKRD1	CRIP2	GRB14	LINC01468	MSRB3	RNU6-268P	SYK
AC009238.2	AP000346.3	CRYBG1	GRHL2	LMO7	MT-TF	RNU6-757P	SYNM
AC018761.1	AP1M2	CRYBG2	GRIK2	LOXL2	NEBL	RNU6-833P	SYTL2
AC026468.1	ARHGEF6	CTGF	GRPR	LOXL4	NEURL1B	RNU6-840P	TGFB1I1
AC027338.2	ARL4C	DACT1	HAS3	MACC1	NMRAL2P	RPL7AP49	TGM2
AC068580.2	ATP2C2	DCLK1	HEG1	MAL2	OLFML2A	S100A14	THBS1
AC078993.1	ATP5F1AP1	DDIT4L	HMCN1	MAL2-AS1	OVOL1	SAMD3	THSD4
AC084033.3	ATP5F1AP10	DENND5B	HMGN1P12	MAMDC2	PALM3	SAMD4A	TMC4
AC087857.1	ATP5F1AP7	DKK1	HOXB-AS2	MAML2	PBX1	SAMD5	TMEM125
AC090617.9	ATP5F1AP8	DNM3	HTR1B	MAP1B	PDLIM5	SEMA3A	TMEM200A
AC108174.1	ATP8B1	DRAXIN	IFIT1	MAP7	PGM5P2	SERPINA1	TMTC1
AC108463.1	BCL2L15	EBF4	IGFBP3	MARVELD3	PMEPA1	SERPINA5	TNFRSF19
AC108463.2	BICDL2	EFR3B	IL32	MDGA1	PNPLA5	SERPINE1	TNFSF18
AC109309.2	BSPRY	EPAS1	IQANK1	MEF2C	PROM2	SESN3	TNFSF4
AC123788.2	C1orf116	EPCAM	IQGAP2	MGAT5B	PRSS22	SGK1	VCAN-AS1
ACOXL	C1orf210	EPHA1	ITGB8	MIR10B	PRSS33	SLC16A6	VIM
ACSL5	CASC10	EPN3	KCND3	MIR1915	PRSS8	SLC1A3	VIM-AS1
ADAMTS2	CD99L2	ESRP1	KDF1	MIR196A1	RAB25	SLC2A14	ZCCHC12
ADGRF1	CDH1	ETV1	KLK10	MIR320C1	RBM24	SLC2A3	ZCCHC24
AFAP1L2	CDH3	FAM83B	KRT13	MIR4701	RF00019	SLC4A8	ZEB1
AL049555.1	CLDN2	FAT4	KRT32	MIR4766	RF00019	SNORA79	ZNF608
AL157786.1	CLDN7	FGF18	KRTCAP3	MIR544B	RF00019	SP5	
AL353763.2	CNTNAP3	FGFBP1	LAD1	MIR548AK	RF00019	SPARC	
AL354718.1	CNTNAP3B	FN1	LAMA3	MIR5692A1	RGS5	SPINT1-AS1	

Supplementary Table 1B. HCT116 DE gene list bulk RNA seq.

List of overlapping genes from bulk RNA seq.

ACOXL, ALOXE3, AP1M2, BICDL2, BSPRY, C1orf210, CDH1, CDH3, CLDN7, EPCAM, EPHA1, ESRP1, IQANK1, ITGB8, KCND3, KDF1, KLK10, KRT32, LAD1, LIMS2, MAL2, MAP7, MARVELD3, PMEPA1, SPARC, SPINT1-AS1, ST14, SYDE1, SYK, TMC4, TMEM125, ZEB1

Supplementary Table 2. Intersection gene list from bulk RNA seq.

List of identified genes from Xavier budding signature.

Genes up
A4GALT, AHNAK2, AKAP12, ANXA1, APC, ARHGAP29, ARNTL2, CALD1, CD109, CDC42BPA, CEP170, DCBLD2, ELK3, FAT1, FER, FERMT2, FGFR1, GLTSCR2, GULP1, KIFC3, KLF6, MALT1, MAP1B, MEIS2, NIN, PALLD, PCDH7, PDE4B, PDLIM7, PHLDB2, PLK2, PLK3, PPFIBP1, RAI14, RGS2, SVIL, TAGLN, TLN1, TUBA1A, VIM, WWC2
Genes down
BDH1, BIRC5, CDCA7, CDX2, CKAP2, EBP, FGFR4, GGCT, GPR160, H2AFX, HMMR, HOOK1, KHK, KIF11, MGST1, MLXIPL, MMAB, POC1A, PPARG, RAVER2, RPL14, RPL36A, SAPCD2, SCD, SLC25A5, SNRPF, SRI, SUCLG1, TFAP4, THRA, TOP1MT, TPD52

Supplementary Table 3. Gene list in Xavier budding signature.

Gene list Nano String EMT signature

Epithelial
AGR2, AP1M2, ARHGAP32, BCAS1, CBLC, CD24, CD2AP,FGFR3, FUT3, GALNT7, GDF15, GPR56, GRHL2, HDHD3, IRF6, KRT19, KRT7, LAD1, MUC1, MYO5C, OCLN, CDH1, CDS1, CEACAM1, CEACAM5, CEACAM6,CKMT1A, CLDN7, CXADR, CYB561, ELF3, EPCAM, EPN3, EPS8L1, ERBB2, ERBB3, ESRP1, F11R, FAM174B,OVOL2, PLS1, PPL, PRR15L, PRSS8, PTK6, RAB25, RBM47, S100A14, SCNN1A, SDC4, SH3YL1, SLC44A4,SORD, SPDEF, SPINT1, ST14, TJP2, TJP3, TMEM30B, TMPRSS2, TMPRSS4, TOM1L1, TSPAN1, VAMP8, VAV3
Mesenchymal
AKAP12, AKAP2, AKT3, ANGPTL2, ASPN, BGN, BICC1, BNC2, C1S, CALD1, CAV1, CCL8, CD163, CDH11, CDH2, CDK14, CEP170, CHRDL1, CLEC2B, CLIC4, COL5A2, COL6A1, COL6A2, CRISPLD2, CSF2RB, CTSK, CXCL12, CXCL13, CXCR4, CYP1B1, DCN, DDR2, DPT, DPYSL3, ECM2, EMP3, ENPP2, EVI2A, FAP, FBLN1,FBN1, FERMT2, FGL2, FHL1, FLI1, FN1, FSTL1, FXYD6, GIMAP4, GIMAP6, GLYR1, GREM1, GZMK, HEG1,IGF1, IL10RA, ISLR, ITM2A, JAM2, JAM3, KCNJ8, KIAA1462, LHFP, LOX, LY96, MAF, MEOX2, MFAP4, MMP2, MPDZ, MRC1, MS4A4A, MS4A6A, MYLK, NAP1L3, NR3CSFRP1, SLIT2, SNAI2, SPARC, SPARCL1, SRGN, SYNE1, TCF4, TNC, TNS1, TPM2, TWIST1, VCAM1, VCAN,1, OLFML2B, PCOLCE, PDGFC, PLEKHO1,PLXNC1, PMP22, PTGDS, PTGIS, PTPRC, PTRF, PTX3, QKI, RUNX1T1, SACS, SAMSN1, SERPINF1, SERPING1,VIM, VSIG4, WIPF1, WWTR1, ZCCHC24, ZEB1, ZEB2, ZFPM2

Detected genes from Nano String EMT signature in scRNAseq

Epithelial
ERBB2, CEACAM1, CXADR, F11R, SPINT1, OVOL2, FGFR3, PRSS8, CDH1, DC4, ARHGAP32, CDS1, CYB561, ESRP1, GALNT7, LAD1, OCLN, PPL, PTK6, S100A14, SCNN1A, ST14, TJP2, TJP3, TSPAN1, VAMP8, CD24, CD2AP, EPCAM, MUC1, RAB25, ERBB3, CLDN7, AP1M2, CKMT1A, EPN3, EPS8L1, HDHD3, KRT19, KRT7, MYO5C, PLS1, RBM47, SH3YL1, TOM1L1, ELF3, GRHL2, CBLC, GDF15, SORD, IRF6
Mesenchymal
PDGFC, FN1, CLIC4, QKI, CAV1, GREM1, ERPINF1, SLIT2, SPARC, HEG1, SERPING1, NR3C1, TWIST1, CYP1B1, COL6A1, COL5A2, COL6A2, CRISPLD2, FBLN1, FBN1, FXYD6, IL10RA, LHFP, OLFML2B, PCOLCE, PLEKHO1, PTRF, SYNE1, CXCR4, AKAP12, DDR2, CDK14, EMP3, FHL1, PMP22, VCAN, CALD1, VIM, AKAP2, C1S, CEP170, CTSK, FSTL1, GLYR1, KIAA1462, PTGDS, SACS, TPM2, ZCCHC24, DPYSL3, MYLK, WWTR1, SNAI2, FERMT2, TNS1, ZEB1

Supplementary Table 4. EMT genes nanostring panel.