Phenotypic plasticity underlies local invasion and distant metastasis in colon cancer.

Andrea Sacchetti[1*], Miriam Teeuwssen[1*], Mathijs Verhagen[1*], Rosalie Joosten[1], Tong Xu[1], Roberto Stabile[1], Berdine van der Steen[2], Martin M. Watson[1**], Alem Gusinac[1], Won Kyu Kim[3], Inge Ubink[4], Harmen J.G. van de Werken[5], Arianna Fumagalli[6], Madelon Paauwe[7], Jacco van Rheenen[8], Owen Sansom[7,9], Onno Kranenburg[4], Riccardo Fodde[1‡]

Department of [1]Pathology, [2]Otorhinolaryngology, Head and Neck Surgery, and [5]Cancer Computational Biology Center and Department of Urology; Erasmus University Medical Center, Rotterdam, The Netherlands. [3]Natural Product Research Center, Korea Institute of Science and Technology (KIST), Gangneung, Korea.[4]Department of Surgical Oncology, Cancer Centre, University Medical Centre Utrecht, Utrecht, The Netherlands. [6]Princess Máxima Center for Pediatric Oncology Utrecht, The Netherlands. [7]Cancer Research UK Beatson Institute, Glasgow, UK. [8]Department of Molecular Pathology, Oncode Institute, Netherlands Cancer Institute, Amsterdam, The Netherlands. [9]Institute of Cancer Sciences, University of Glasgow, Glasgow, UK.

Running Title: The cells of origin of colon cancer metastasis

Competing Interests Statement: The authors of this manuscript do not have any competing financial interests in relation to the work described.

*These authors equally contributed to this study.

**Present address: Gastrointestinal Translational Research Unit, Department of Gastrointestinal Surgery at Stavanger University Hospital, Norway.

‡To whom correspondence should be addressed at:

Erasmus MC, Department of Pathology, PO Box 2040, 3000 CA Rotterdam, The Netherlands

E-mail: r.fodde@erasmusmc.nl Tel: +31 10 7043896

Abstract

Phenotypic plasticity represents the most relevant hallmark of the carcinoma cell as it bestows it with the capacity of transiently altering its morphologic and functional features while *en route* to the metastatic site. However, the study of phenotypic plasticity is hindered by the rarity of these events within primary lesions and by the lack of experimental models. Here, we identified a subpopulation of phenotypic plastic colon cancer cells: EpCAM^{lo} cells are motile, invasive, chemo-resistant, and highly metastatic. EpCAM^{lo} bulk and single-cell RNAseq analysis indicated 1. enhanced Wnt/β-catenin signaling, 2. a broad spectrum of degrees of EMT activation including hybrid E/M states (partial EMT) with highly plastic features, and 3. high correlation with the CMS4 subtype, accounting for colon cancer cases with poor prognosis and a pronounced stromal component. Of note, a signature of genes specifically expressed in EpCAM^{lo} cancer cells is highly predictive of overall survival in tumors other than CMS4, thus highlighting the relevance of quasi-mesenchymal tumor cells across the spectrum of colon cancers.

Enhanced Wnt and the downstream EMT activation represent key events in eliciting phenotypic plasticity along the invasive front of primary colon carcinomas. Distinct sets of epithelial and mesenchymal genes define transcriptional trajectories through which state transitions arise. pEMT cells, often earmarked by the extracellular matrix glycoprotein SPARC together with nuclear ZEB1 and β-catenin along the invasive front of primary colon carcinomas, are predicted to represent the origin of these (de)differentiation routes through biologically-distinct cellular states, and to underlie the phenotypic plasticity of colon cancer cells.

Introduction

Cancers of epithelial origin such as breast, prostate, pancreas, lung, and colon carcinomas are thought to develop from normal tissues through a multistep sequence of genetic events from benign precursor lesions to increasingly more malignant stages. This is exemplary illustrated by the adenoma-carcinoma sequence in colon cancer where a stepwise buildup of genetic alterations in specific oncogenes and tumor suppressor genes underlies tumor initiation and progression[1]. These alterations result in well-defined cellular changes largely reflecting the so-called 'hallmarks of cancer', which provide different selective advantages to the developing tumor and represent essential requirements for carcinoma formation at the primary site[2]. However, with regard to the capacity to disseminate through the tumor microenvironment and establish metastases in distant organ sites, epigenetic changes, rather than genetic mutations, underlie what the most clinically relevant hallmark of cancer, namely phenotypic plasticity[3,4].

Malignant cells, and in particular those responsible for local dissemination and distant metastasis, are often endowed with the capacity to undergo transient and reversible morphological and functional changes. In particular, epithelial to mesenchymal transition (EMT), i.e. the progressive loss of epithelial features and the acquirement of a more migratory and mesenchymal phenotype[5], is regarded as a crucial event in tumor cell invasion and dissemination at the primary site. EMT bestows cancer cells with stem-like plastic characteristics[6] needed to acquire quasi-mesenchymal features at the invasive front of the primary tumor, disseminate and attain therapy resistance, and to revert back to more epithelial states (MET, mesenchymal to epithelial transition) at the organ site of metastasis[7]. Epigenetic activation and silencing of EMT-inducing transcription factors (EMT-TFs) underlies the transient nature of these cellular modifications[8]. Notwithstanding these ground rules, a very broad spectrum of molecular and cellular routes underlie EMT and the resulting phenotypic plasticity in context-dependent fashion[9].

The "migrating cancer stem cell" (mCSC) model has been first proposed for colon cancer by T. Brabletz[7], also as a solution to the so-called "β-catenin paradox"[10]. In the majority of sporadic colorectal cancer cases, the rate-limiting loss of the *APC* tumor suppressor is predicted to lead to nuclear β-catenin translocation and full-blown Wnt signaling activation. Notwithstanding these predictions, tumor cells with nuclear β-catenin represent only a small minority of the primary lesion and tend to cluster non-randomly at the invasive front of colon carcinomas where they gain

mesenchymal features to detach and disseminate into the adjacent stromal tissues. In view of these observations, it is plausible that cues secreted from the tumor microenvironment elicit EMT downstream of full blown Wnt signaling activation, earmarked by nuclear β-catenin, in a subset of cells located at the invasive front[7,10]. However, the molecular and cellular mechanisms underlying Wnt and EMT activation at the invasive front of colon cancers are yet largely unknown also due to a lack of robust *in vitro* and *in vivo* models.

Previously, it was shown that human immortalized breast cancer cell lines encompass different subpopulations of cells with distinct phenotypic states and functional characteristics maintained in a dynamic equilibrium through stochastic transitions between states[11]. Similar observations were made in oral squamous carcinoma cell lines where distinct CSC phenotypes are present: whereas non-EMT CSCs are proliferative and retain epithelial characteristics, the EMT-competent CSC fraction is (quasi)mesenchymal and of increased cellular motility[12]. As such, conventional immortalized cancer cell lines may offer a valid model to elucidate the mechanisms underlying phenotypic plasticity in cancer and to identify novel EMT/CSC-related therapeutic targets.

Here, we identified and extensively characterized a subpopulation of quasi-mesenchymal colon cancer cells endowed with phenotypic plasticity that underlie local invasion and distant metastasis, and whose expression signature is predictive of reduced disease-free survival among colon cancer patients.

Results

Conventional colon cancer cell lines encompass a subpopulation of ZEB1-driven quasi-mesenchymal, highly metastatic, and chemo-resistant cells.

To assess whether, as observed for breast cancer[11], commonly employed colon cancer cell lines encompass distinct differentiated and more stem-like subpopulations of cancer cells, we first analyzed a broad panel of cell lines by FACS with antibodies directed against different (cancer) stem cell markers (CD44, CD133, CD24, ALDEFLUOR) in combination with the epithelial marker EpCAM. As shown in **Fig. 1A and S1-1** for HCT116 and SW480, the CD44/EpCAM combination best highlighted the presence of distinct subpopulations with a predominant $CD44^{high}EpCAM^{high}$ cluster and a minority of $CD44^{high}EpCAM^{low}$ cells. The same $CD44^{high}EpCAM^{low}$ cells were identified at similarly variable percentages in other commonly employed colon cancer cell lines (**Fig. S1-2A**). Here, we mainly focused on the HCT116 and SW480 lines as they are representative of the two main colon cancer subtypes earmarked by MIN (microsatellite instability) and CIN (chromosomal instability; also known as MSS, microsatellite stable), respectively[13]. For the sake of clarity and readability, from this point on we will refer to the $CD44^{high}EpCAM^{low}$ and $CD44^{high}EpCAM^{high}$ subpopulations of colon cancer cells as $EpCAM^{lo}$ and $EpCAM^{hi}$, respectively.

To characterize the $EpCAM^{lo}$ and $EpCAM^{hi}$ colon cancer cells, they were sorted by FACS and shown to have distinct epithelial- ($EpCAM^{hi}$) and mesenchymal-like ($EpCAM^{lo}$) morphologies (**Fig. 1B**). Likewise, $EpCAM^{lo}$ cells were shown to have significantly increased migratory and invasive ability when compared with their $EpCAM^{hi}$ counterparts (**Fig. S1-3A**).

RT-qPCR analysis of the sorted subpopulations revealed differential expression of EMT-associated marker genes and transcription factors. Significantly reduced mRNA levels of *EPCAM* and E-cadherin (*CDH1*), and increased expression of vimentin (*VIM*) and of the EMT-transcription factor *ZEB1* were observed in $EpCAM^{lo}$ cells from both lines, as also confirmed by immunofluorescence analysis (**Fig. S1-3B-D**). Accordingly, knockdown of *ZEB1* expression by shRNA dramatically decreased the percentage of $EpCAM^{lo}$ cells in both lines (**Fig. S1-4A-C**).

Expression of *ZEB1* (and *ZEB2*) has been shown to be regulated by the miR-200 superfamily of micro RNAs which target specific 3'UTR sequences[14]. RT-qPCR analysis of sorted cells revealed significantly reduced expression levels of all five miR-200 members in $EpCAM^{lo}$ cells from both cell lines (**Fig. S1-4D**), i.e. in agreement with the observed

increase in *ZEB1* expression. Proliferation and cell-cycle analysis indicated decreased mitotic activity in EpCAM^{lo} cells from both cell lines (**Fig. S1-4E-F**)

In view of the well-established correlation between EMT and therapy resistance[15], EpCAM^{lo} cells were cultured in the presence of oxaliplatin and 5-fluorouracil (5-FU) and their viability compared with that of EpCAM^{hi} and bulk cells by metabolic activity assay (MTT). EpCAM^{lo} cells showed increased viability at all tested oxaliplatin (**Fig. S1-5A,C**, left panels) and 5-FU (**Fig. S1-5B,D**, left panels) concentrations. Likewise, re-growth assays revealed that EpCAM^{lo} cells from both cell lines are able to re-enter the cell cycle at a broad range of oxaliplatin and 5-FU concentrations when compared with EpCAM^{hi} cells (**Fig. S1-5**, right panels).

Last, to assess *in vivo* their capacity to form metastatic lesions in the liver, HCT116 and SW480 bulk and EpCAM$^{hi/lo}$ sorted cells were injected in the spleen of immune-incompetent recipient mice. EpCAM^{lo} cells from both lines resulted in significantly more liver metastases than with EpCAM^{hi} and bulk cells (**Fig. 1C-D**). Notably, IHC analysis of the resulting 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 what observed in primary colon carcinomas[10,16] (**Fig. S1-6A**). FACS analysis of the EpCAM^{lo}-derived liver metastases revealed predominant epithelial features with a vast majority of EpCAM^{hi} cells (>99%), thus highlighting their striking plasticity and the key role played by MET in metastasis formation **(Fig. S1-6B).**

In order to validate the role of *ZEB1*-driven EMT in the establishment of the EpCAM^{lo} subpopulation of colon cancer cells in a cell line-unrelated *in vivo* model mimicking the course of events observed in colon cancer patients, we employed mouse intestinal organoids carrying specific mutations at the *Apc*, *Kras*, and *Tp53* genes (*Apc*$^{fl/fl}$::*Kras*$^{G12D/+}$::*Trp53*$^{fl/R172H}$; AKP)[17]. Orthotopic transplantation of AKP organoids results in the establishment of primary carcinomas in the caecum and subsequent metastasis development at distant organ sites, mainly liver and lungs[18]. We further modified the AKP organoids by tagging them with GFP and a click beetle luciferase[19], and by making them doxycycline-inducible for the expression of mouse *Zeb1* (AKP-Z) (**Fig. S1-6C**). Upon orthotopic transplantation of the AKP-Z organoids and subsequent establishment of the primary tumor in the caecum, mice were administered doxycycline for one week in the drinking water to induce *Zeb1* expression. FACS analysis of the

primary tumor revealed an increase in EpCAM^{lo} cells from 4.8% in the non-induced tumors up to 22-76% upon dox-induction of *Zeb1* expression (**Figure S1-6D**). As expected, only a marginal increase of lung and liver metastases was observed in AKP-Z transplanted mice upon continuous dox-administration in the drinking water for 8 weeks when compared with control (no dox) animals (**Fig. S1-6E**), likely to result from the continuous induction of *Zeb1* expression and the consequent inhibition of mesenchymal-to-epithelial transitions essential for metastasis formation[7].

Overall, the results show the presence within colon cancer cell lines of EMT-driven, quasi-mesenchymal and therapy-resistant EpCAM^{lo} cells with increased invasive and metastatic capacity. *ZEB1* expression underlies *in vivo* the establishment and maintenance of the subpopulation of EpCAM^{lo} colon cancer cells, thereby contributing to increased dissemination along the invasion-metastasis cascade.

EpCAM^{lo} colon cancer cells are maintained in equilibrium with EpCAM^{hi} through stochastic state transitions

To further investigate on the plasticity of EpCAM^{lo} cancer cells, we assessed their capacity to differentiate into the more epithelial EpCAM^{hi} type and reconstitute the heterogeneous composition of the parental cell lines. Sorted EpCAM$^{lo/hi}$ cells from HCT116 and SW480 were grown separately under conventional culture conditions and analyzed by FACS at different time points. As shown in **Fig. 2A**, while the majority of EpCAM^{lo} cells from both cell lines revert to the epithelial phenotype, only a minority of EpCAM^{hi} cells switches to the more mesenchymal state. Accordingly, *CDH1* and *EPCAM* expression was significantly increased in 'late' (e.g. cultured for >60 and >100 days, respectively) vs. 'early' EpCAM^{lo} cells (e.g. collected after <7 days of culture), whereas *VIM* and *ZEB1* expression was decreased in 'late' vs. 'early' EpCAM^{lo} cells (**Fig. S2-1A**). In addition, the migratory capacity of 'late' EpCAM^{lo} HCT116 cells was reduced to levels comparable with those of 'early' EpCAM^{hi} cells (**Fig. S2-1B**).

To exclude cross-contamination between sub-populations, single EpCAM$^{hi/lo}$ cells from both cell lines were sorted into 96-well dishes, cultured up to 70-80 days, and analyzed by FACS at intermediate and end time points. As shown in **Fig. 2B and S2-1C**, the majority of EpCAM^{lo} single cells were capable of generating substantial percentages of EpCAM^{hi} progeny to eventually recapitulate the heterogeneous composition of the parental cell lines (e.g. **Fig. 2B**, clones 2F12 and 2E10,). A minority of the cells however, appears to have lost this plasticity and retains, even after extended culture, the EpCAM^{lo} phenotype (e.g. **Fig. 2B**, clone 2H12). In contrast, the majority of EpCAM^{hi} single cells retained their epithelial features with <1% switching to the EpCAM^{lo} state (**Fig. 2B and S2-1C**).

Based on these results, a two-state Markov model was developed to estimate the average probabilities to transition from one state to the other. First, the FACS data was employed to estimate the average doubling time of cells in both populations; slightly increased doubling times were reported for EpCAM^{lo} compared to EpCAM^{hi} cells in both lines (HCT116: 1.09 vs. 1.00 days; SW480: 1.86 vs. 1.68 days). Next, we employed least square optimization to estimate the transition probabilities that best fit the observed population dynamics. The fitted model predicts that both subpopulations have a high probability to retain their cell identity, with minor though significant likelihood to transit to the other state. Of note, EpCAM^{lo} cells show a higher transition probability compared to EpCAM^{hi} (8.1 and 4.0 fold in HCT116 and SW480, respectively)(**Fig. 2C)**.

Due to the observed differences in doubling times between the two states, subclones with a lower EpCAM$^{hi>lo}$ transition probability will experience a slight growth advantage which will become prevalent in the long run. We ran a simulation of this effect by starting from a culture with multiple subclones having distinct transition probabilities; the results indicate that subclones with lower plasticity gain dominance within a few months (**Fig. S1-2B**). Consequently, especially in the long run, the percentage of EpCAM^{lo} cells will decrease as observed in late cultures (**Fig. S1-2A**).

Overall, the above results highlight the high plasticity and stem-like features of EpCAM^{lo} cells in their ability to acquire epithelial features and reconstitute the heterogeneous composition of the parental cell line, independently of external factors other than the conventional culture conditions here employed.

Enhanced Wnt signaling activation underlies EMT and the establishment of EpCAM^{lo} colon cancer cells.

In order to elucidate the mechanisms underlying plasticity and EMT in EpCAM^{lo} colon cancer cells, RNAseq analysis was performed on the sorted subpopulations from the HCT116 and SW480 lines. Multidimensional scaling (MDS) showed a separation in the second dimension of EpCAM^{hi} and EpCAM^{lo} cells in both cell lines (**Fig. 3A**). A total of 152 and 353 differentially regulated genes were identified between the EpCAM^{hi} and EpCAM^{lo} cells in HCT116 and SW480, respectively (*P. adjusted* < 0.01, when applying a Log_2 fold change < -1.5 and > 1.5). Among these, only relatively few (n=34) were common to both lines (**Fig. S3-1A**). However, Ingenuity Pathway Analysis (IPA) revealed that genes differentially expressed in each cell line reflect common molecular and cellular functions including cellular motility, cellular assembly and organization, and drug metabolism (**Fig. S3-1B)**. Nonetheless, IPA of the combined

EpCAM^{lo} expression profiles highlighted significant associations with EMT regulation, Wnt/β-catenin signaling, human ESC pluripotency, IL8 signaling, and colorectal cancer metastasis (**Fig. 3B and S3-1B**).

The activation of canonical Wnt signaling in EpCAM^{lo} colon cancer cells is of interest in view of the fact that both cell lines harbor mutations (loss and gain of APC and β-catenin function in SW480[20] and HCT116[21], respectively) predicted to result in the constitutive activation of the pathway. Notwithstanding the latter, Wnt appears to be increased in EpCAM^{lo} cells, possibly due to the epigenetic activation/inhibition of synergistic/antagonistic loci. In view of these observations and of the established functional link between Wnt and EMT[22,23], we evaluated whether Wnt signaling 'super-activation' in the already Wnt-ON HCT116 and SW480 cell lines could expand the relative size of the EpCAM^{lo} subpopulations. Indeed, upon treatment with the glycogen synthase 3β (GSK3β) inhibitor CHIR99021 (Chiron), a robust Wnt signaling activation was observed in both cell lines by TopFLASH reporter assay (**Fig. 3C**). FACS analysis of the treated cell lines showed that the enhancement of Wnt signaling led to an approx. 3-fold increase of the EpCAM^{lo} population in the HCT116 cell line, though not in SW480 (**Fig. 3D**). However, IF analysis showed that Chiron treatment was consistently accompanied by an increase in *ZEB1* expression in both cell lines, in agreement with the role of *ZEB1* as a downstream Wnt target[24,25] (**Fig. 3E**).

To further validate the role of Wnt in the establishment of the EpCAM^{lo} subpopulation, we took advantage of the recently generated isogenic subclones of the HCT116 cell line with distinct β-catenin genotypes, namely wild type (HCT116-WT), hetero- (HCT116-P), and homozygous (HCT116-MT) for the Ser45del mutant allele[26]. FACS analysis of these cell lines revealed a progressive increase in the EpCAM^{lo} subpopulation from 3.9% in HCT116-WT, to 7.9% in HCT116-P, and 28.7% in HCT116-MT cells (**Fig. 3F**). The observed increase in EpCAM^{lo} cells in the isogenic HCT116 lines matches their morphological features ranging from a distinct epithelial-like morphology in HCT116-WT, to a progressively increased spindle-shaped and scattered patterns in HCT116-P and -MT cells, as originally reported by Kim et al. (2019)[26].

In the majority of colon cancers, nuclear accumulation of β-catenin, is exclusively observed at the invasive front where tumor cells are more exposed to growth factors and cytokines from the stromal microenvironment likely to further enhance Wnt signaling in localized fashion thus triggering EMT, invasion, and distant metastasis[7,10]. We analyzed the invasive front of a cohort of colon carcinomas by IHC with antibodies directed against β-catenin and

ZEB1 in consecutive sections. As shown in **Figure S3-2**, co-localization of nuclear β-catenin and ZEB1 expression was found in 5 out of 25 cases investigated.

Overall, the results highlight the key role played by enhanced Wnt signaling activation in establishing and maintaining the EpCAM^{lo} subpopulation of colon cancer cells through *ZEB1* upregulation and EMT induction.

EpCAM^{lo} cells are associated with the CMS4 group of patients with shorter disease-free and overall survival.

Distinct recurrent gene expression patterns underlie the recently proposed classification of human colon cancers in four Consensus Molecular Subtypes (CMS1-4)[27]. Of these, the mesenchymal CMS4 subtype has the greatest propensity to form metastases. While fibrosis is a hallmark of CMS4 and a dominant contributor of mesenchymal gene expression, the cancer cells themselves can also express genes reflecting a (quasi-)mesenchymal state. Accordingly, the CMS4 subtype was identified in tumor-derived organoids and cell lines, suggesting that CMS4 is an intrinsic property of the mesenchymal colon cancer cell[28]. Therefore, we asked whether expression of the signatures derived from the RNAseq analysis of EpCAM^{lo} cells would correlate with the CMS classification of human colon cancers and cell lines. To this end, we employed a compiled data set encompassing expression data relative to 3232 human colon cancers classified as CMS1-4[27]. Expression of the HCT116 and SW480 signatures were highly correlated with each other (**Fig. 4A**), with the CMS4 signature genes (n=143) (**Fig. 4A-B**), and with the expression signature of colon cancer cell lines previously classified as mesenchymal-like (CCS3)[29] (**Fig. S4-1A**).

Next, we used both HCT116 and SW480 signatures from the bulk RNAseq to generate High, Intermediate, and Low expression subgroups in the CMS cohort through k-means clustering (**Fig. S4-1B**). High expression of the EpCAM^{lo} signatures from both cell lines identifies a group of primary colon cancer patients with a high propensity to form metastases and with significantly shorter overall survival (**Fig. 4C**).

Overall, the above data strongly link the expression of the EpCAM^{lo} signatures derived from common colon cancer cell lines to the CMS4 subtype and to shorter disease-free and overall survival.

scRNAseq analysis of EpCAM^{lo} colon cancer cells reveals high heterogeneity and partial EMT intermediate stages

To further elucidate the heterogeneity and molecular mechanisms underlying the phenotypic plasticity of EpCAM^{lo} colon cancer cells, single cell RNA sequencing (scRNAseq) was performed using chromium controller (10X Genomics) technology on sorted subpopulations from HCT116 and SW480. More than 1000 cells were analyzed from each subpopulation and sequenced to a depth of approx. 50000 reads each with the MiSeq System (Illumina). After dimension reduction with tSNE, the EpCAM^{lo} and EpCAM^{hi} cells clustered in separate groups in HCT116 (**Fig. 5A**), whereas the SW480 cells showed a partial overlap between the two subpopulations and a distinct EpCAM^{hi} cluster, identified as the non-adherent subpopulation within the SW480 cell line ("spheres")[30,31] (**Fig. S5-1A**). Substantial overlap between the EpCAM$^{hi/lo}$ subpopulations was retained upon subsequent exclusion of the non-adherent cells in SW480, attributed to additional variance in genes independent of EpCAM$^{hi/lo}$ differences (**Fig. S5-1B**). Next, the dimension reduction was repeated in supervised fashion by taking advantage of the publicly available EMT gene list from the Nanostring nCounter® PanCancer Progression Panel (n =180; 107 of which found to overlap with the scRNAseq data). Using this 'EMT signature', the two subpopulations were clearly resolved both in HCT116 and SW480 (**Fig. 5B**).

To rank and order the cells along the E-M axis, epithelial (E; n=51) and mesenchymal (M; n=56) gene sets were defined from the EMT signature based on the provided annotations (Nanostring), and evaluated by GSVA (Gene Set Variation Analysis). Next, an EMT score was computed by subtracting the E score from the M score (EMT = M – E), while a partial EMT score was obtained based on the co-expression of the E an M gene sets (pEMT = min[E,M]). The expression of genes previously found to be differentially expressed between EpCAM$^{hi/lo}$ by bulk RNAseq was then evaluated over the course of EMT progression (**Fig. 5C)**. While HCT116 showed a clear transition from EpCAM^{hi} to EpCAM^{lo} with a minority of "in-between" cells earmarked by high pEMT scores, in SW480 the transition appeared to be accompanied by higher levels of inter-cellular variance. We continued to use the EMT-axis in both cell lines to study gene expression trends. As expected, epithelial expression of *EPCAM* and *CDH1* gradually decreased over the EMT-axis, while mesenchymal genes such as *ZEB1* and *VIM*, and Wnt target genes such as *AXIN2* and *DKK1*, showed opposing behavior. More generically, the level of Wnt signaling ("Wnt.score"), measured by the activation of the WNT_PATHWAY from the PID database, increased over the EMT-axis (**Fig. 5D**). We evaluated the expression trends of other genes upregulated in EpCAM^{lo} cells and observed that *SPARC* (osteonectin) peaked in between the extremities of the EMT axis, indicative of a partial EMT state (**Fig. 5D**).

To further characterize the transition process between EpCAM^{lo} and EpCAM^{hi}, we analyzed the HCT116 scRNAseq data by RNA velocity, an algorithm that predicts the future cell state of individual cells on a short-term timescale (usually hours), based on the ratio between spliced and unspliced mRNA. Consistent with the fitted Markov Model, the majority of cells were moving in their respective state, and both states included source and sink points that could elicit or inhibit plasticity (**Fig. S2-1D**). On a population level, EpCAM^{lo} cells were more likely to transit to EpCAM^{hi} cells than vice versa (by partition-based graph abstraction or PAGA; **Fig. S2-1E**). Within the EpCAM^{lo} subpopulation, *CD44* and mesenchymal genes such as *VIM* and *ZEB1* were expressed with different dynamics (**Fig. S5-1C**). In fact, only a small population of cells was captured in the process of transitioning to the other identity (**Fig. 5E**). This apparently plastic population, earmarked by the opposing velocity arrows, showed the highest pEMT score (**Fig. 5F**), as well as high *SPARC* expression (**Fig. 5G)**.

Last, we performed cluster analysis on the HCT116 scRNAseq data. Unsupervised clustering using shared neighbor (SSN) modularity optimization revealed the presence of distinct subclusters (n=8): 3 of EpCAM^{hi} origin, and 5 EpCAM^{lo} (**Fig. S5-1D**). To further investigate the gene expression patterns underlying the different EMT states, we performed a cluster analysis where the EMT signature genes are grouped with k-means according to their average expression in the unsupervised clusters (**Fig. S5-1E**). This analysis revealed four sets of genes expressed in different combinations throughout the different clusters. Notably, gene set mes1, including *ZEB1, VIM* and *SNAI2,* is expressed throughout the EpCAM^{lo} clusters, while gene set mes2, including *SPARC*, *FN1* and *TWIST1* is mostly expressed in the pEMT EpCAM^{lo} cluster (#7). Of note, the partial EMT clusters #6 and #7 showed distinct expression of these gene sets indicating alternative activation of specific arrays through which partial EMT cells arise (**Fig. S5-1F**).

As *SPARC* earmarks pEMT states (**Fig. 5D**), we overexpressed it in the HCT116 and SW480 cell lines and observed a marked increase in the expression of EMT-TFs (*ZEB1* and *SNAI2*) and the corresponding up- and -down-regulation of mesenchymal and epithelial markers in EpCAM^{lo} cells and a corresponding increase in their motility and invasive capacity (**Fig. 5H**).

Taken together, our *in silico* analysis shows substantial heterogeneity within the EpCAM^{lo} subpopulation across colon cancer cell lines. Among EpCAM^{lo} cells, a minority exhibits partial EMT and underlies the stochastic EpCAM$^{lo\text{->}hi}$ and

EpCAM$^{hi->lo}$ transitions. Of note, pEMT is shown here not only as an '*in-between*' state earmarked by the co-expression of E- and M-specific genes, but also by specific genes like *SPARC* whose expression peaks at pEMT states.

Identification of EpCAM^{lo} colon cancer cells in primary colorectal tumors

As shown above, the EpCAM^{lo} gene signatures derived from both cell lines show high concordance with the CMS4 subtype, and are strongly associated with poor survival. Previously, it has been questioned to what extent the CMS4 profile reflects a *bona fide* clinical identity rather than being a representation of contamination from the tumor micro-environment[32,33]. We addressed this issue by evaluating the CMS4 signature (here referred to as "*CMS4_RF*") originally developed by Guinney et al.[27], and by taking advantage of a recent scRNAseq study on colon cancer resections from 29 patients[34]. Based on the latter study, as depicted in **Fig. 6A**, the CMS4_RF signature shows the highest association with normal and tumor stromal cells. At the bulk RNAseq level, using the larger (n= 3232 patients) cohort of colon tumors from Guinney et al.[27], the CMS4_RF signature is clearly enriched among the CMS4 tumors (**Fig. 6B**) and is likewise associated with poor survival (**Fig. 6C**). Hence, the CMS4_RF signature reflects the presence of tumor-associated stromal cells, likely to be enriched in patients with decreased overall survival. To question whether the presence of a subset of *bona fide* tumor cells in quasi-mesenchymal state (EpCAM^{lo}) may also represent a feature of colon cancers with poor prognosis, we derived a distinct signature ("*CMS4_TC*") by selecting genes correlated (Pearson >0.3) with the CMS4_RF signature within the tumor epithelial fraction from the Lee et al. study[34]. Accordingly, the CMS4_TC signature shows the highest association with epithelial tumor cells (**Fig. 6D**). Of note, at the bulk RNA level, the CMS4_TC signature shows increased association with both CMS1 and CMS4 tumors (**Fig. 6E**), and outperforms the CMS4_RF signature in stratifying patients on survival (**Fig. 6F**). The same CMS4_TC signature can clearly stratify patients from CMS groups other than CMS4, and in particular CMS1 and CMS3, on survival (**Fig. S6-1**).

Finally, in order to identify EpCAM^{lo} cells in primary colorectal tumors, we took advantage of the scRNAseq data from Lee et al.[34] to annotate the tumor cells with highest association to the CMS4_TC signature as mesenchymal-like (Mes-like = 9.7% of all tumor cells). In agreement with the observations made at the bulk RNA level, tumor cells from CMS1 and CMS4 patients showed the highest association to the CMS4_TC signature (**Fig. 6G**). Differential expression between the mesenchymal-like and bulk tumor cells revealed lower expression of *EPCAM*, while higher expression

of *CD44* and mesenchymal markers such as *VIM*, *COL1A1*, *MMP7*, *ECM1*, and *SPARC* (**Fig. 6H**), the latter previously identified to peak at intermediate EMT levels and predicted by RNAvelocity to earmark a subpopulation of plastic cells transitioning in between states (**Fig. 5D-G**).

Discussion

The progression from primary tumor to metastasis still represents an unsolved puzzle as genomic and gene expression profiles of primary cancers and their matched metastases are in general strikingly similar[35]. Colon cancer provides a unique example for this conundrum: although it is well-established that a sequence of gene mutations underlies the adenoma-carcinoma sequence at the primary site, no alterations have been identified in genes potentially able to underlie local invasion and distant metastasis. Hence, the capacity to metastasize could already be pre-determined by specific mutations acquired by tumor cells at relatively early stages of tumorigenesis. However, this does not yet explain why and how only rare cells within the primary lesion acquire a metastatic phenotype that endows them with increased cell motility and with the capacity to invade the stromal microenvironment to eventually home at distant organ sites and form metastases. From this perspective, phenotypic plasticity likely represents a relevant mechanism for metastasizing colon carcinoma cells to transiently and reversibly change their cellular identity along the invasion-metastasis cascade[3,4].

The vast majority of colon cancers are initiated by mutations at the *APC* gene predicted to lead to constitutive Wnt signaling by β-catenin nuclear translocation. However, IHC analysis has shown that nuclear accumulation of β-catenin is exclusively observed at the invasive front where tumor cells are exposed to growth factors and cytokines secreted by the stromal microenvironment likely to further enhance Wnt signaling in localized fashion[10]. As different levels of Wnt/β-catenin signaling are associated with distinct cellular outcomes[36], it is plausible to think that the 'full-blown' activation of this pathway, earmarked by nuclear β-catenin, may trigger EMT and endow phenotypic plasticity in a minority of colon cancer cells located at the invasive front. The transient and reversible nature of these cellular modifications is likely driven by epigenetic changes at EMT-TFs.

Here, we have taken advantage of the coexistence of biologically-distinct cellular phenotypes within commonly employed colon cancer cell lines to study the cellular and molecular mechanisms which underlie phenotypic plasticity and the acquisition of invasive and metastatic capacity. While the bulk of the cancer cells have a characteristic epithelial phenotype (here referred to as $EpCAM^{hi}$), a minority of cells with mesenchymal morphology and plastic features ($EpCAM^{lo}$) coexist in a dynamic equilibrium through stochastic state transitions with their more committed counterpart. Similar observations have been previously made in breast cancer and oral squamous cell

carcinoma cell lines[11,12] which underlines their relevance for the study of phenotypic plasticity in a broad spectrum of malignancies.

As predicted by their morphology, EpCAM^{lo} cells feature increased invasive and metastatic capacity and a distinct gene expression profile when compared to their epithelial counterpart. In particular, EMT activation resulting from enhanced Wnt signaling and *ZEB1* upregulation are distinctive features of the EpCAM^{lo} transcriptome. Of note, *ZEB1* plays many pleiotropic roles ranging from the modulation of oncogenic and tumor suppressive pathways, cell-fate determination, stemness, and cell plasticity[37], and is as such likely to be a determinant in colon cancer invasion and metastasis. From this perspective, the recent debate on EMT as an essential requirement for metastasis[38] likely mirrors the complexity of transcription factors and their downstream targets involved in these processes. Moreover, a recent scRNAseq comparative study of various time-course EMT models has revealed very limited overlap among differentially expressed genes indicative of the vastly context-dependent nature of these processes[9]. The here presented data emphasize the key role played by EMT and its reverse MET in colon tumor cell dissemination and distant metastasis formation.

The observed role of enhanced Wnt signaling in EpCAM^{lo} cells is of interest in view of the presence of *APC* and *CTNNB1* (β-catenin) gene mutations in these cell lines[20,21]. Further enhancement of Wnt signaling is likely to result in EMT activation through *ZEB1* upregulation and its conversion from a repressor to activator state[24]. On its turn, increased *ZEB1* expression was also shown to enhance Wnt, possibly in an autocrine stimulatory circuit[25]. Accordingly, Wnt 'super-activation' by GSK3β inhibition (i.e. Chiron) in the *APC*/*CTNNB1*-mutated colon cancer cell lines results in a significant expansion of EpCAM^{lo} cells and increased *ZEB1* expression. Vice versa, *ZEB1* downregulation by shRNA leads to a dramatic contraction of the EpCAM^{lo} subpopulation. Notably, isogenic subclones of HCT116 carrying wild type, heterozygous, and homozygous mutations at the β-catenin locus[26] show a progressive increase in the size of the EpCAM^{lo} subpopulation. These observations highlight the relevance of Wnt dosages in the downstream activation of EMT.

Transcriptional regulation of EMT-TFs has been shown to be controlled by miRNAs binding to specific 3'UTR sequences. In particular, members of the miR-200 family inhibit EMT and cancer cell motility by suppressing Wnt/β-catenin signaling and by directly interfering with the expression of EMT-TFs and in particular of *ZEB1/2*[14,23].

Accordingly, we have shown that the expression of different members of the miR-200 family is downregulated in EpCAM^{lo} colon cancer cells when compared with their epithelial counterpart. Moreover, the RKO cell line, composed entirely of EpCAM^{lo} cells, was previously reported to be characterized by epigenetic downregulation of miR-200a/b/c leading to high *ZEB1* expression levels[39] in the absence of Wnt-activating mutations[40]. As the epigenetic activation/silencing of the miR-200 locus is dynamically regulated by common EMT-inducers such as TGFβ and Wnt[23,39], it is likely that phenotypic plasticity is triggered *in vivo* in cells located at the invasive front of colon carcinomas as a result of a complex interplay between their specific mutation spectra and epigenetic profiles modulated by different signaling cues from the microenvironment. The consistent presence of cells with nuclear co-localization of ZEB1 and β-catenin located at the invasive front of patient-derived colon carcinomas validates these observations. Other EMT-TFs like ZEB2 are likely to be upregulated in ZEB1-negative cases[41].

The clinical relevance of our results, originated from the analysis of immortalized cancer cell lines, came from the bioinformatic mining of bulk and scRNAseq data from patient-derived colon cancers. First, EpCAM^{lo} expression signatures are highly correlated with the CMS4 signature[27] and with shorter disease-free and overall survival. The mesenchymal CMS4 subtype accounts for approx. 25% of the cases, and is associated with the greatest propensity to form metastases. While the CMS4 signature was initially attributed to the presence of tumor-associated stroma and not to a cancer cell-intrinsic property[32,33], it has been shown that *bona fide* tumor cells from stroma-rich cancers have a markedly mesenchymal gene expression pattern[28]. By taking advantage of the Lee et al. scRNAseq study[34], we derived a tumor-specific signature (*CMS4_TC*) by selecting epithelial-specific genes from the original CMS4_RF signature, that outperforms other commonly employed prognostic and predictive markers (e.g. *BRAF* mutation and MSI; *data not shown*) in stratifying CMS1, 3, and 4 colon cancer patients based on overall survival. These results highlight the relevance of our study in the identification of a quasi-mesenchymal cellular state with plastic, invasive, and metastatic properties predictive of poor prognosis in colon cancer patients regardless of their CMS classification.

Recent studies on the role of EMT in eliciting phenotypic plasticity in cancer cells have highlighted the relevance of intermediate cellular stages co-expressing both epithelial and mesenchymal genes for tumor progression and metastasis[42-44]. These hybrid E/M or partial EMT cells are thought to be endowed with increased invasive and metastatic capacity. Our scRNAseq analysis of EpCAM^{lo} colon cancer cells has revealed not only fully mesenchymal

but also hybrid E/M subclusters, the latter predicted *in silico* to underlie the observed transcriptional heterogeneity. Of note, transcriptional activation of specific arrays of E- (epi1/2) and M- (mes1/2) genes accompanies the transition between cellular states. Partial EMT cells, predicted by RNAvelocity to transit between states, are characterized not only by the co-expression of E- and M-specific genes at intermediate levels, but also by increased expression of specific genes like *SPARC* (Secreted Protein Acidic and Rich in Cysteine, also known as osteonectin), encoding for a matricellular protein involved in the modulation of cell–cell and cell–matrix interactions and known as a prognostic marker in colon cancer[45]. The role of SPARC in cancer is controversial as it has been shown to promote EMT and metastasis, but also to encompass tumor-suppressive functions in context-dependent fashion[46]. Importantly, SPARC triggers EMT through direct cell-to-cell contact and upregulation of other EMT-inducers like fibronectin (FN1)[47], reminiscent of the interactions occurring between parenchymal and stromal cells at the invasive front in colon cancer[10] where pEMT is expected to underlie plasticity and invasion through the ECM. Although SPARC is unique in its expression peaking at pEMT states and earmarking the transition between EpCAM^{lo} and EpCAM^{hi} in both cell lines, and in mes-like cells from patient-derived colon cancers, it seems unlikely that specific genes exist that can independently elicit pEMT. Instead, tumor-specific and context-dependent activation of subset of genes with distinct functions (e.g. *SPARC*, *FN1*, *MMP7*, *ZEB1*, *ECM1*) synergistically promoting, for example, collective cell migration upon interaction with the stromal microenvironment, may represent a more likely scenario.

The metastable and plastic features of EpCAM^{lo} cells was further highlighted by their striking capacity of giving rise to distant metastases reminiscent of the primary tumors both in the prevalence of EpCAM^{hi} cells and in the distinct patterns of β-catenin intracellular localization between the periphery and center of the lesions. These observations are in agreement with the key role of MET for the onset of distant metastasis[7]. The admittedly marginal increase in liver and lung metastases upon continuous induction of *ZEB1* expression is likely to result from MET inhibition. From this perspective, our results provide support to the "migrating CSC"[7] and "β-catenin paradox"[10] models and shed light in the cellular and molecular mechanisms underlying Wnt super-activation, (p)EMT induction, and cell dissemination along the invasive front of colon carcinomas.

The characterization of the cellular and molecular mechanisms underlying phenotypic plasticity in colon cancer cells along the invasion-metastasis cascade has great relevance for the development of future diagnostic and

therapeutic approaches based on circulating tumor cells (CTCs). Although current detection methodologies mainly rely on their epithelial features, more recent studies have indicated that CTCs with less pronounced epithelial characteristics are likely to underlie metastasis at distant organ sites. In breast and oral cancer, hybrid E/M tumor cells feature the highest degree of phenotypic plasticity coupling the capacity of undergoing EMT/MET with therapy resistance and other stem cells features[48,49]. Further single cell RNAseq and functional analyses of extensive cohorts of EpCAM^{lo} cells from matched primary tumors, CTCs, and the corresponding distant metastases will open new avenues for diagnostic and therapeutic applications.

Materials and Methods

Cell culture

Human colon cancer cell lines were acquired from ATCC (American Type Culture Collection) and cultured in DMEM medium (11965092, Thermo Fisher Scientific) [supplemented with 10% heat inactivated fetal bovine serum (FBS; Thermo Fisher Scientific), 1% Penicillin/Streptomycin (penicillin: 100 U/ml, streptomycin: 100 µg/ml; 15140122 Thermo Fisher Scientific)] in humidified atmosphere at 37°C and 5% CO_2. The identity of each cell line was confirmed by DNA fingerprinting with microsatellite markers (Amelogenin, CSF1PO, D13S317, D16S539, D5S818, D7S820, THO1, TPOX, vWA, D8S1179, FGA, Penta E, Penta D, D18S51, D3S1358, D21S11) and compared with the analogous data provided by ATCC, EACC, and https://web.expasy.org/cellosaurus/ (data not shown). The HCT116-P, HCT116-WT and HCT116-MT cell lines were kindly given by the laboratory of Hoguen Kim from the Yonsei University College of Medicine in Seoul, Korea.

Apc$^{fl/fl}$::*Kras*$^{G12D/+}$::*Trp53*$^{fl/R172H}$ (AKP) organoids organoids were grown in 50 µL droplets of Matrigel (Corning) covered with Advanced DMEM-F12 medium (12634028, Thermo Fisher Scientific) supplemented with 1x GlutaMAX (35050-061, Thermo Fisher Scientific), HEPES (15630-056, Thermo Fisher Scientific), gentamicin (15750060, Thermo Fisher Scientific), 100 ng/ml mEgf (PMG8041, Thermo Fisher Scientific), 2% B27 supplement (12587010, Thermo Fisher Scientific), mNoggin (self-produced according to protocol), and 100 µM Y-27632 (Y0503, Sigma Aldrich).

Chiron (CHIR99021, 4423, Tocris, dissolved at 20 mM in DMSO), 4 µM CHIR99021 was added to the culture medium every 48 hours for one week.

Flow cytometry and sorting

For flow cytometry and sorting purposes, cells were detached from the culture dishes using trypsin-EDTA (15400054, Thermo Fisher Scientific) and resuspended in staining buffer (PBS supplemented with 4% FCS). Antibody staining was performed on ice for 30' with CD44-APC (clone IM7, 559250, BD Pharmingen) and EpCAM-FITC (ESA-214, GTX30708, GeneTex). Cells were then washed and resuspended in PBS 4% FCS. Flow cytometry analysis and cell sorting were carried out with a FACSAria III Cell Sorter (BD Biosciences). Forward-scatter height (FSC-H) versus forward-scatter width (FSC-W), and side-scatter height versus side-scatter width were employed to eliminate cell aggregates and

ensure single cell sorting. Gates were defined as depicted and specified in Figure 1. The DAPI nuclear dye (D9542, Sigma-Aldrich) was used at 1 µg/ml to exclude dead cells. FITC and GFP were analyzed using 488 nm laser and 502LP and 530/30BP filters; APC and Sytox Red with a 633 nm laser and a 660/20BP filter; BV421 using a 405 nm laser and a 450/40BP filter; BV785 with a 405 nm laser and a 750LP and 780/60BP filter; PE with a 461 nm laser and a 582/15 BP filter.

For the clonogenicity assays, single cells were sorted into 96-well plates at 1 cell/well in culture medium. After sorting, the plates were checked under the microscope to confirm the presence of a single cell per well. Upon confluency, each single clone was FACS-analyzed for CD44/EpCAM at different time points.

For the flow cytometry analysis of mouse liver metastases and primary AKP-Z organoid derived tumors, freshly resected tissue samples were cut into small pieces, washed three times with ice-cold PBS, and subsequently digested in Advanced DMEM-F12 (12634028, Thermo Fisher Scientific) containing 50 mg/mL of collagenase A (11088793001, Roche) for 30' at 37°C, with vigorous pipetting every 10'. Residual tissue fragments were further treated with TrypLE Express (12605-010, Thermo Fisher Scientific), 0,25% Trypsin (Thermo Fisher Scientific) and 0.1 mg/mL DNaseI (DN25, Sigma-Aldrich) for 45' at 37°C. Samples were then washed and filtered in sequence through a 100 and 40 µm cell strainer (Corning). The HCT116 derived cells from the liver metastasis were identified and analyzed using the following antibodies for 30' on ice: HLA-A, B, C-biotin (clone W6/32, 311434, Biologend), Streptavidin-PE-Dazzle 549 (Biolegend), rat anti-mouse CD45-BV421 (clone 30-F11, 563890, BD Biosciences), rat anti-mouse CD31-BV421 (clone 390, 563356, BD Biosciences), rat anti-mouse TER-199-BV421 (clone TER-119, 563998, BD Biosciences), CD44-APC (clone IM7, 559250, BD Pharmingen) and EpCAM-FITC (ESA-214, GTX30708, GeneTex). AKP-Z tumor derived cells were stained using the following antibodies: rat anti-mouse CD45-BV421 (clone 30-F11, 563890, BD Biosciences), rat anti-mouse CD31-BV421 (clone 390, 563356, BD Biosciences), rat anti-mouse TER-199-BV421 (clone TER-119, 563998, BD Biosciences), anti-mouse EpCAM-PE (clone G8.8, Biolegend), anti-mouse/human CD44-APC (clone !M7, BD Biosciences), and DAPI. Intrinsic GFP expression was also used to select for organoid derived cells).

Cell proliferation assay and cell cycle analysis

For proliferation assays, sorted cells were seeded in 24-well dishes (2500 cells/well) and cultured in conventional medium. Cells were harvested and counted at the indicated time points. For cell cycle analysis, sorted cells were centrifuged, fixed in 70% ethanol and stained with 50 μg/ml propidium iodide (PI) (Sigma), and 0.2 mg/ml RNAse A (10109142001, Sigma-Aldrich). The stained cells were then FACS-analyzed using a 655LP and a 695/40 BP filter.

Cell viability assays

For the chemoresistance assays, cells were seeded in 96-well plates at 15000 cells/well and left overnight to adhere. Three technical replicates were plated per tested condition. Both oxaliplatin (Sigma-Aldrich) and 5-fluorouacil (5-FU; Sigma-Aldrich) were dissolved in DMSO. HCT116 cells were incubated for 3 or 5 days with 5-FU and oxaliplatin, respectively, whereas SW480 cells were treated with 5-FU or oxaliplatin for 7 days (calibrated on the response of the parental line in terms of viability). After removal of the chemotherapeutic drug, cells were washed with PBS and left to re-grow in standard culture medium. Cell viability was assessed using the MTT (3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide; Sigma-Aldrich) assay at time zero, i.e. upon removal of the drug, and after re-growth (range 1-36 days). Briefly, cells were incubated at 37°C, 5% CO_2 for 3 hrs. in culture medium supplemented with 0.45 mg/mL MTT. The 96-well plates were then centrifuged at 1,000 rpm for 5' and the culture medium removed. MTT formazan precipitates were solubilized with DMSO. O.D. reading was performed at 595 nm with a Microplate Reader (Model 550, Bio-Rad). Background measurements were subtracted from each data point. MTT-based results were also validated by manual count of trypan blue positive cells, using cells harvested from additional 96 wells. At least two biological experiments were performed for each individual cell line and drug.

Selection of oxaliplatin and 5-FU concentrations and administration times for chemo-resistance assays was as follows. In a pre-defined range of concentrations (1.5 -100 μg/ml for 5-FU, and 0.6-50 μg/ml for oxaliplatin) based on previous data, the maximal concentration and length of treatment was selected as the combination able to induce cell death in approx. 80% of $EpCAM^{lo}$ cells. From that concentration, the curve was scaled down by 50% at least three times to reach the lowest concentration to be tested. The minimal treatment length was 72 hrs, while the maximal duration was 1 week. Oxaliplatin was considerably slower than 5-FU in its cytotoxicity with HCT116 cells. The difference was less relevant in SW480 cells where numerous cells, although irreversibly damaged, remained viable for several days. For the latter reason, treatment was longer in SW480 cells.

Migration and invasion assays

For the 'transwell' (modified Boyden chamber) migration and invasion assays, cells were starved for 24 hrs. in DMEM supplemented with 1% FBS. For each tested subpopulation, 1.0 x 10^6 cells were harvested, resuspended in 1% FBS medium and plated in transwell tissue culture inserts (8 µm polycarbonate membrane, 24 well, 3428; Corning). For the invasion assays 1.5 µg/ml Matrigel dissolved in serum-free medium was pipetted on top of the transwell membrane and left to set at 37°C for several hours before plating the cells. DMEM medium with 10% FBS was then placed in the lower compartment of the transwell chamber. For both assays three technical replicates were plated for each subpopulation. Plates were incubated for 24 hrs. at 37°C/5% CO_2, after which the cells and membranes were fixed in methanol and stained with crystal violet. The non-migrated cells on the upper surface of the membrane were removed with a cotton-wool bud and the cells migrated to the lower surface of the membrane counted. For both assays two independent experiments were performed.

RNA isolation and qRT-PCR

RNA was isolated from cells using TRIzol™ Reagent (15596018, ThermoFisher Scientific) according to the manufacturer's instructions. RNA concentration was measured by NanoDrop. Subsequently, reverse transcription into cDNA was conducted using the High-Capacity cDNA Reverse Transcription Kit (4368814, Life Technologies) according to the manufacturer's instructions. RT-qPCR was performed with Fast SYBR Green Master Mix (4385617; ThermoFisher Scientific). PCR amplification and detection were implemented with the 7400 Fast Real-Time PCR System. Expression levels were normalized for the endogenous *GAPDH* reference gene. All PCR primers here employed are listed here below:

GENE	*FORWARD PRIMER*	*REVERSE PRIMER*
GAPDH	5'-ACCCAGAAGACTGTGGATGG-3'	5'-TCTAGACGGCAGGTCAGGTC-3'
EPCAM	5'-GCAGCTCAGGAAGAATGTG-3'	5'-CAGCCAGCTTTGAGCAAATGAC-3'
CDH1	5'-TGCCCAGAAAATGAAAAAGG-3'	5'-GTGTATGTGGCAATGCGTTC-3'
VIM	5'-GAGAACTTTGCCGTTGAAGC-3'	5'-GCTTCCTGTAGGTGGCAATC-3'
CDH2	5'-CAACTTGCCAGAAAACTCCAGG-3'	5'-ATGAAACCGGGCTATCTGCTC-3'
FN1	5'-CAGTGGGAGACCTCGAGAAG-3'	5'-TCCCTCGGAACATCAGAAAC-3'

ZEB1	5'-GCACAACCAAGTGCAGAAGA-3'	5'-CATTTGCAGATTGAGGCTGA-3'
ZEB2	5'-TTCCTGGGCTACGACCATAC-3'	5'-TGTGCTCCATCAAGCAATTC-3'
TWIST1	5'-GTCCGCAGTCTTACGAGGAG-3'	5'-GCTTGAGGGTCTGAATCTTGCT-3'
SNAI1	5'-GCGAGCTGCAGGACTCTAAT-3'	5'-CCACTGTCCTCATCTGACA-3'
SNAI2	5'-GGGGAGAAGCCTTTTTCTTG-3'	5'-TCCTCATGTTTGTGCAGGAG-3'
FoxC2	5'-GCCTAAGGACCTGGTGAAGC-3'	5'-TTGACGAAGCACTCGTTGAG-3'

miRNA

For miRNA experiments, sorted cell populations from HCT116 and SW480 cell lines were cultured up till 70% confluency in 6 well-multiwell plates. Total RNA was then isolated using the Trizol™ (Ambion) protocol. For *miR-200-family (miR-200a, miR-200b, miR-200c, miR-141, miR-429)*, reverse transcription was performed using the TaqMan MicroRNA RT Kit (Applied Biosystem), followed by qRT-PCR using TaqMan MicroRNA assay (Thermo Fisher scientific). *U6* snRNA was used as tanendogenous control. The expression of *miR-200-family* was analyzed by the ΔCT method. The experiment was repeated three times.

shRNA

in order to knock-down *ZEB1* expression, lentiviral inducible shRNA vectors encompassing control or *ZEB1* sequences were developed. Cloning was performed according to the manufacturer's instructions[50]. Briefly, Tet-pLKO-puro (gift from D. Wiederschain, Addgene plasmid # 21915) was digested with AgeI and EcoRI and isolated by gel purification (QIAEX II Gel Extraction Kit, Qiagen). The sequences of the control and *ZEB1* shRNA were as follows: shRNA control (shCT) (Addgene sequence #1864):

top 5'-CCGGCCTAAGGTTAAGTCGCCCTCGCTCGAGCGAGGGCGACTTAACCTTAGGTTTTTTG-3',

bottom: 5'-AATTCAAAAAACCTAAGGTTAAGTCGCCCTCGCTCGAGCGAGGGCGACTTAACCTTAGG-3',

shZEB1-A:

target sequence (Broad Institute): 5'-GCTGCCAATAAGCAAACGATT-3'

oligo sequence:

top: 5'-CCGGGCTGCCAATAAGCAAACGATTCTCGAGAATCGTTTGCTTATTGGCAGCTTTTT-3',

bottom: 5'-AATTAAAAAGCTGCCAATAAGCAAACGATTCTCGAGAATCGTTTGCTTATTGGCAGC-3',

shZEB1-B:

target sequence (Broad Institute): 5'-GTCTGGGTGTAATCGTAAATT-3'

oligo sequence:

top 5'-CCGGGTCTGGGTGTAATCGTAAATTCTCGAGAATTTACGATTACACCCAGACTTTTT-3',

bottom 5'- AATTAAAAAGTCTGGGTGTAATCGTAAATTCTCGAGAATTTACGATTACACCCAGAC-3'.

The "top" and "bottom" labels indicate the oligonucleotide that were annealed in 0.1 M NaCl, 10 mM TrisHCl, pH 7.4, after incubation at 95°C followed by a cooling down step until room temperature is reached. The digested vector (200 ng) was ligated with 1 µL of the oligonucleotides (0.45nmol/µl) using T4 DNA ligase (Promega) for 3 hrs. at room temperature. "One Shot Stbl3" chemically competent E. coli (Thermo Fisher Scientific) were transformed with the ligation product. Individual colonies were inoculated for mini-prep cultures and the presence of the insert was checked using the restriction enzyme XhoI and by sequencing.

For lentivirus production, the shCT or the pool of the two *ZEB1* shRNA constructs were packaged into second generation virus particles using psPAX2 (Addgene plasmid # 12260; gift from dr. Didier Trono), pMD2.G (Addgene plasmid # 12259) into HEK293T. Virus particles were titrated with the HCT116 cell line and a MOI of 0.5 was employed to produce the shRNA inducible HCT116 and SW480 cell lines. Lentivirus-infected cells were selected in medium containing 1 µg/mL puromycin (Dulbecco). shRNA induction was implemented by using 1 µg/mL doxycycline for 72 hrs. The extent of *ZEB1* downregulation was assessed by RT-qPCR.

Construction of Zeb1 inducible vector

A pORF mZeb1 plasmid (Cat, no. ORF062179) encompassing the cDNA of the *Zeb1* gene was employed for the construction of the lentiviral vector. The *Zeb1* gene was cloned into the pEN_TmiRc3 plasmid, gift from Iain Fraser (California Institute of Technology, CA). The *Zeb1* insert was transferred into a pSLIK-Hygro plasmid (#25737; Addgene, USA) by Gateway technology. All newly generated plasmids were sequence-verified by LGC Genomics (LGC Genomics GmbH, Germany). The pSLIK-Hygro plasmid was transiently transfected by Fugene HD (Promega) in HEK293T cells together with the VSV-G, MD and REV packaging plasmids. After 2 days, the culture medium was collected and viral particles were concentrated by ultracentrifugation. AKP organoids were transduced with the concentrated virus. After 24 hours, the transduced cells were selected by hygromycin B (Thermo Fisher Scientific,

The Netherlands) (60 µg/mL) for7 days. Expression of *Zeb1* was confirmed by qRT–PCR upon stimulation with doxycycline hyclate.

Animal experiments

All protocols involving animals were approved by the Dutch Animal Experimental Committee and were conformed to the Code of Practice for Animal Experiments in Cancer Research established by the Netherlands Inspectorate for Health Protections, Commodities and Veterinary Public health (The Hague, the Netherlands, 1999). Animals were bred and maintained in the Erasmus MC animal facility (EDC) under conventional specific pathogen-free (SPF) conditions.

Spleen transplantation assays were implemented on 6 to 8 week-old NOD.Cg-*Prkdc*^{scid} *Il2rg*^{tm1Wjl}/SzJ (NSG) male and female mice anesthetized intraperitoneally with ketamine (Ketalin®, 0.12 mg/ml) and xylazine (Rompun®, 0.61 mg/ml). Carpofen (Rimadyl®, 5 mg/ml) was given subcutaneously as analgesia. The spleen was exteriorized through a left lateral flank incision and 2.5x10^4 HCT116 and SW480 cells, resuspended in 50 µl of PBS, were injected into the spleen parenchyma using an insulin syringe. Fifteen minutes after injection, a splenectomy was performed with a fine tip cautery pen in order to remove spilled cells and ensure hemostasis. The peritoneum and skin were then sutured in two layers. Mice injected with HCT116 cells were sacrificed 4 weeks after injection for tumor collection; mice injected with SW480 cells were killed 8 weeks after spleen transplantation. Upon liver resection, individual macroscopic metastatic lesions were counted, dissected, and fixed in 4% paraformaldehyde (4% PFA). Residual liver tissue and the lungs were also fixed in 4% PFA for further immunohistochemical analyses.

Orthotopic transplantation of intestinal mouse organoids was performed as previously described[18]. In brief, the day before transplantation AKP-Zeb1 organoids containing a GFP and click beetle luciferase vector (ATG-1929, a gift from Keith Wood, Addgene plasmid #108712) were collected and mechanically dissociated into small clumps. About 100.000 cells were plated in 10 µL drops neutralized Rat Tail High Concentrate Type I Collagen (Corning, cat. No. 354249) and let to recover overnight at 37°C in Advanced DMEM-F12 medium (12634028, Thermo Fisher Scientific) 100 ng/mL mEgf (PMG8041, Thermo Fisher), 2% B27 supplement (12587010, Thermo Fisher), mNoggin (self-produced according to protocol), 100 µM Y-27632 (Y0503, Sigma Aldrich). Caecum transplantation assays were implemented on 6 to 8-week-old male NOD.Cg-*Prkdc*^{scid} *Il2rg*^{tm1Wjl}/SzJ (NSG) mice anesthetized intraperitoneally with

ketamine (Ketalin®, 0.12 mg/ml) and xylazine (Rompun®, 0.61 mg/ml). Carpofen (Rimadyl®, 5 mg/ml) was given subcutaneously as analgesia. The caecum was exposed through a midline abdominal incision and a collagen drop containing tumor cells was surgically transplanted in the caecal submucosa. The peritoneum and skin were then sutured in two layers. Tumor growth was monitored by abdominal palpation. Mice were sacrificed 6-8 weeks after transplantation. Upon collection, primary caecal tumors were single cell digested and further analyzed by FACS. Mice used for quantification of liver and lung metastases were injected with Luciferin-D and imaged with an IVIS® Spectrum imaging system (Caliper Life Sciences, Hopkinton, MA, USA). After imaging tissues were fixed and cut into 500-micron slices using a Fibrotome, processed for IHC, stained for β-catenin visualize tumor cells, scanned using a NanoZoomer, and counted using NDP view software.

Immunofluorescence (IF) analysis

Coverslips containing a monolayer of cancer cells were fixed for 30' in 4% PFA at 37 °C and washed twice with PBS. Cells were first permeabilised for 15 minutes at room temperature with 0.2% of Triton X-100 and then incubated in blocking buffer (5% milk powder in PBS-Tween) for 1 hour at room temperature. Cells were then exposed overnight at 4 °C to primary antibodies against EpCAM (mouse, 1:250; sc-66020; Santa Cruz Biotechnology) and ZEB1 (rabbit, 1:200; sc-25388, Santa Cruz Biotechnology). After washing twice with PBS-Tween, coverslips were incubated for 1 hour at room temperature in blocking buffer containing the following secondary antibodies: Goat anti-Rabbit Alexa Fluor® 594 conjugate (1:250, #A-11037, Life Technologies) and Donkey anti-Mouse Alexa Fluor® 488 conjugate (1:250, #A-21202, Life Technologies). Cells were counterstained with DAPI to visualize the nuclei. Coverslips were mounted in VECTAHIELD HardSet Antifade Mounting Medium (#H-1400, Vector Labs) and imaged with a Zeiss LSM-700 confocal microscope. Images were processed with ImageJ (U.S. National Institutes of Health, Bethesda, MD, USA).

Immunohistochemistry (IHC) analysis

Tissues from animal experiments were fixed overnight in 4% PFA and embedded in paraffin. Paraffin blocks containing human colon cancer tissue were obtained from the department of Pathology at the Erasmus Medical Center in Rotterdam. Four µm sections were mounted on slides. Immunohistochemistry was performed using the

EnVision Plus-HRP system (Dako) and antibodies directed against β-catenin (1:200, 610154, BD Biosciences) and ZEB1 (1:200, HPA027524, Sigma-Aldrich). Briefly, paraffin embedded sections were dewaxed with Xylene and hydrated in 100% and 70% ethanol. Antigen retrieval was performed using pressure cooker pretreatment in a citrate buffer (pH 6.0) for ZEB1 and in a Tris-EDTA buffer (pH 9.0) for the anti-human-mitochondria and anti-β-catenin antibodies. Subsequently, slides were incubated at room temperature in 3% hydrogen peroxidase for 15' to block endogenous peroxidase activity. Tissue sections were washed and blocked with 5% BSA in PBS-Tween for 1 hour to then be incubated with the primary antibodies overnight at 4 °C. Slides were washed twice with PBS-Tween and incubated with Rabbit EnVision+ System-HRP (K4001, Dako) or Mouse EnVision+ System-HRP (K4007, Dako) for 30'. Subsequently, signal detection was and tissues were counterstained with Mayer's Hematoxylin. Dehydration was performed by incubation in 70% and 100% ethanol followed by Xylene before sleds were mounted using Pertex (00811, Histolab).

TOP-Flash reporter assay

For the β-catenin/TCF reporter assay (TOP-Flash reporter assay), cells were plated on 48-well dishes and cultured in medium with or without 4 μM CHIR99021. After 48 hrs., when 70% confluence was reached, cells were transfected by Fugene HD (Promega) with 125 ng of the TOP-Flash or FOP-Flash reporter constructs together with 25 ng of the Renilla luciferase vector for normalization purposes. Luciferase activity was measured using the Dual-Luciferase Reporter Assay System (Promega) 24 hrs. post-transfection. Luminesence was measured using a GloMax Luminometer.

Next Generation Sequencing (NGS): RNAseq

RNA quality and quantity was evaluated on a 2100 Bio-analyzer (Agilent) using the Agilent RNA 6000 Pico Kit. RNA samples were further processes according to the TruSeq Sample Preparation v.2 Guide (Illumina) and paired end-sequenced on the HiSeq 2500 (Illumina).

IIllumina paired-end reads of 76 bases were trimmed by removing the TrueSeq adapter sequences using Trimmomatic (v.0.33)[51]. Subsequently, the reads were mapped in a two-pass procedure to the human reference

genome build hg38 with the RNA-seq aligner STAR (v2.4.2a)[52] and the Homo sapiens GENCODE v23 annotation[53]. Raw counts were summed with the summarize overlaps function with union mode from the Bioconductor Genomic Alignments package[54] (v1.14.0). Genes were called differentially expressed with a generalized linear model using a negative binomial distribution with correcting for cell lines in multi-cell line comparisons. DESeq2 (v1.16.1) was used to perform these calculations[55]. We applied a Wald-test to identify statistical significant differently expressed genes. *P*-values were adjusted using the Benjamini-Hochberg[56] correction based on which a threshold value was set at < 0.01. Multidimensional scaling was performed after the read counts were normalized with blind variance stabilizing log2 transformation function of DESeq2. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) gene enrichment analyses were carried out as described previously[57]. R (v 3.4.0) (R Core Team, 2017; https://www.R-project.org/) was employed for statistical analysis and visualization of the data.

Bioinformatics* analysis *bulk RNAseq

For pathway analysis generated RNAseq datasets were uploaded into Ingenuity Pathway Analysis software (Qiagen). For other bioinformatics analyses, the generated datasets were uploaded into the R2 Genomics Analysis and Visualization Platform (http://r2.amc.nl). First, we used the 'differential expression between groups' option to identify the hundred genes with highest expression in the EpCAM^{lo} fraction in both the HCT116 and SW480 cells. These genes were saved as separate gene sets. Expression values of all genes in both gene sets was then assessed in the CMS-3232 composite cohort[27], as well as in a large cohort of profiled cell lines originating from the Broad institute. This yielded single meta-gene expression values per tumor or cell line. These gene set expression values were then stored as separate tracks and compared using the 'relate 2 tracks' option, yielding Pearson R values and accompanying p values.

The 'relate 2 tracks' option was also used to compare the 2 gene sets identifying the EpCAM^{lo} cells with gene sets positively identifying the 4 different molecular subtypes within the published 273-gene CMS classifier, again yielding r values and corresponding p-values.

The gene sets identifying the EpCAM^{lo} cells from both cell lines were also used to cluster the tumors in the CMS3232 cohort into low, intermediate and high expression groups by k-means clustering. The Kaplan-Meier method was subsequently used to assess significant differences in survival between the generated subgroups. As all tumors in

this cohort had previously been assigned to specific CMS subgroups, we then analyzed the contribution of each CMS subtype to each of the generated low, intermediate, high expression subgroups.

Single cell RNA seq

Cell lines were brought to 60-70% confluency before the start of the experiment. For each sample, between $5x10^4$ and $1x10^5$ $EpCAM^{lo}$ and $EpCAM^{hi}$ cells were FACS sorted and processed using the 10X Genomics Chromium Single Cell Controller. Samples were deep-sequenced (Illumina) to a depth ranging 49k-65k reads/cells. Gene-cell matrices were obtained by conversion of the raw data using the Cell Ranger pipeline. Filtered gene-cell matrices were merged in R, and processed for downstream analysis using the Seurat package. Dimension reduction was performed using PCA, tSNE and UMAP. Epithelial and mesenchymal scores were computed using the Rmagic (imputation) and GSVA (scoring) packages. RNA velocity analysis was done in Python using Velocyto and scVelo packages.

Data accessibility

The RNA-sequencing data from this study have been submitted to the Gene Expression Omnibus (GEO)[58] database under the accession number GSE154927 and GSE154930 for the bulk and single-cell RNAseq data, respectively. Other dataset referenced in this study are publicly available and can be accessed with GSE144735 and GSE132465 (Lee et al.), and at Synapse with syn2623706 (Guinney et al.).

Statistical analysis

For each experiment, data are shown as mean ±SD. IBM SPSS Statistics software was used for data analysis. The Mann-Whitney U test was used to analyze the difference between two groups of quantitative variables; α-value was set at 5%.

Acknowledgments

This study has been made possible by funding to RF from the Dutch Digestive Foundation (FP 15-09), Dutch Cancer Society, and the Erasmus MC (EMCR 2015-8090). And by funding to MP and OS from the Cancer Research UK core funding to the CRUK Beatson Institute (A17196) and to OS (A21139). The authors are grateful to Drs. R. Smits, M.

Trerotola, S. Alberti and Catherine Winchester for their advice, to Dr. Eric Bindels for assistance with RNAseq analysis, and to 10X Genomics for their support through their Pilot Award Grant Program.

Competing Interests Statement

The authors of this manuscript do not have any competing financial interests in relation to the work described.

References

1. Fearon, E.R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759-67 (1990).
2. Hanahan, D. & Weinberg, R.A. The hallmarks of cancer. *Cell* **100**, 57-70 (2000).
3. Varga, J. & Greten, F.R. Cell plasticity in epithelial homeostasis and tumorigenesis. *Nat Cell Biol* **19**, 1133-1141 (2017).
4. Teeuwssen, M. & Fodde, R. Cell Heterogeneity and Phenotypic Plasticity in Metastasis Formation: The Case of Colon Cancer. *Cancers (Basel)* **11**(2019).
5. Nieto, M.A., Huang, R.Y., Jackson, R.A. & Thiery, J.P. Emt: 2016. *Cell* **166**, 21-45 (2016).
6. Mani, S.A. *et al.* The epithelial-mesenchymal transition generates cells with properties of stem cells. *Cell* **133**, 704-15 (2008).
7. Brabletz, T., Jung, A., Spaderna, S., Hlubek, F. & Kirchner, T. Opinion: migrating cancer stem cells - an integrated concept of malignant tumour progression. *Nat Rev Cancer* **5**, 744-9 (2005).
8. Skrypek, N., Goossens, S., De Smedt, E., Vandamme, N. & Berx, G. Epithelial-to-Mesenchymal Transition: Epigenetic Reprogramming Driving Cellular Plasticity. *Trends Genet* **33**, 943-959 (2017).
9. Cook, D.P. & Vanderhyden, B.C. Context specificity of the EMT transcriptional response. *Nat Commun* **11**, 2142 (2020).
10. Fodde, R. & Brabletz, T. Wnt/beta-catenin signaling in cancer stemness and malignant behavior. *Curr Opin Cell Biol* **19**, 150-8 (2007).
11. Gupta, P.B. *et al.* Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell* **146**, 633-44 (2011).
12. Biddle, A. *et al.* Cancer stem cells in squamous cell carcinoma switch between two distinct phenotypes that are preferentially migratory or proliferative. *Cancer Res* **71**, 5317-26 (2011).
13. Lengauer, C., Kinzler, K.W. & Vogelstein, B. Genetic instability in colorectal cancers. *Nature* **386**, 623-7 (1997).
14. Brabletz, S. & Brabletz, T. The ZEB/miR-200 feedback loop--a motor of cellular plasticity in development and cancer? *EMBO Rep* **11**, 670-7 (2010).
15. Dean, M., Fojo, T. & Bates, S. Tumour stem cells and drug resistance. *Nat Rev Cancer* **5**, 275-84 (2005).
16. Kirchner, T. & Brabletz, T. Patterning and nuclear beta-catenin expression in the colonic adenoma-carcinoma sequence. Analogies with embryonic gastrulation. *Am J Pathol* **157**, 1113-21 (2000).
17. Fumagalli, A. *et al.* Genetic dissection of colorectal cancer progression by orthotopic transplantation of engineered cancer organoids. *Proc Natl Acad Sci U S A* **114**, E2357-E2364 (2017).
18. Fumagalli, A. *et al.* A surgical orthotopic organoid transplantation approach in mice to visualize and study colorectal cancer progression. *Nat Protoc* **13**, 235-247 (2018).
19. Hall, M.P. *et al.* Click beetle luciferase mutant and near infrared naphthyl-luciferins for improved bioluminescence imaging. *Nat Commun* **9**, 132 (2018).

20. Nishisho, I. *et al.* Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients. *Science* **253**, 665-9 (1991).

21. Ilyas, M., Tomlinson, I.P., Rowan, A., Pignatelli, M. & Bodmer, W.F. Beta-catenin mutations in cell lines established from human colorectal cancers. *Proc Natl Acad Sci U S A* **94**, 10330-4 (1997).

22. Lamouille, S., Xu, J. & Derynck, R. Molecular mechanisms of epithelial-mesenchymal transition. *Nat Rev Mol Cell Biol* **15**, 178-96 (2014).

23. Ghahhari, N.M. & Babashah, S. Interplay between microRNAs and WNT/beta-catenin signalling pathway regulates epithelial-mesenchymal transition in cancer. *Eur J Cancer* **51**, 1638-49 (2015).

24. Sanchez-Tillo, E. *et al.* beta-catenin/TCF4 complex induces the epithelial-to-mesenchymal transition (EMT)-activator ZEB1 to regulate tumor invasiveness. *Proc Natl Acad Sci U S A* **108**, 19204-9 (2011).

25. Sanchez-Tillo, E. *et al.* ZEB1 and TCF4 reciprocally modulate their transcriptional activities to regulate Wnt target gene expression. *Oncogene* **34**, 5760-70 (2015).

26. Kim, W.K. *et al.* beta-catenin activation down-regulates cell-cell junction-related genes and induces epithelial-to-mesenchymal transition in colorectal cancers. *Sci Rep* **9**, 18440 (2019).

27. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat Med* **21**, 1350-6 (2015).

28. Vellinga, T.T. *et al.* Collagen-rich stroma in aggressive colon tumors induces mesenchymal gene expression and tumor cell invasion. *Oncogene* **35**, 5263-5271 (2016).

29. De Sousa, E.M.F. *et al.* Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nat Med* **19**, 614-8 (2013).

30. Hirsch, D. *et al.* LGR5 positivity defines stem-like cells in colorectal cancer. *Carcinogenesis* **35**, 849-58 (2014).

31. Yi, H. *et al.* Integrative multi-omics analysis of a colon cancer cell line with heterogeneous Wnt activity revealed RUNX2 as an epigenetic regulator of EMT. *Oncogene* **39**, 5152-5164 (2020).

32. Calon, A. *et al.* Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nat Genet* **47**, 320-9 (2015).

33. Isella, C. *et al.* Stromal contribution to the colorectal cancer transcriptome. *Nat Genet* **47**, 312-9 (2015).

34. Lee, H.O. *et al.* Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nat Genet* **52**, 594-603 (2020).

35. Bernards, R. & Weinberg, R.A. A progression puzzle. *Nature* **418**, 823 (2002).

36. Gaspar, C. & Fodde, R. APC dosage effects in tumorigenesis and stem cell differentiation. *Int J Dev Biol* **48**, 377-86 (2004).

37. Caramel, J., Ligier, M. & Puisieux, A. Pleiotropic Roles for ZEB1 in Cancer. *Cancer Res* **78**, 30-35 (2017).

38. Maheswaran, S. & Haber, D.A. Cell fate: Transition loses its invasive edge. *Nature* **527**, 452-3 (2015).

39. Davalos, V. *et al.* Dynamic epigenetic regulation of the microRNA-200 family mediates epithelial and mesenchymal transitions in human tumorigenesis. *Oncogene* **31**, 2062-74 (2012).

40. Sparks, A.B., Morin, P.J., Vogelstein, B. & Kinzler, K.W. Mutational analysis of the APC/beta-catenin/Tcf pathway in colorectal cancer. *Cancer Res* **58**, 1130-4 (1998).

41. Kahlert, C. *et al.* Overexpression of ZEB2 at the invasion front of colorectal cancer is an independent prognostic marker and regulates tumor invasion in vitro. *Clin Cancer Res* **17**, 7654-63 (2011).

42. Jolly, M.K. *et al.* Implications of the Hybrid Epithelial/Mesenchymal Phenotype in Metastasis. *Front Oncol* **5**, 155 (2015).

43. Aiello, N.M. *et al.* EMT Subtype Influences Epithelial Plasticity and Mode of Cell Migration. *Dev Cell* **45**, 681-695 e4 (2018).

44. Pastushenko, I. *et al.* Identification of the tumour transition states occurring during EMT. *Nature* **556**, 463-468 (2018).

45. Kim, J.Y. *et al.* Expression of Secreted Protein Acidic and Rich in Cysteine in the Stroma of a Colorectal Carcinoma is Associated With Patient Prognosis. *Ann Coloproctol* **29**, 93-9 (2013).

46. Podhajcer, O.L. *et al.* The role of the matricellular protein SPARC in the dynamic interaction between the tumor and the host. *Cancer Metastasis Rev* **27**, 691-705 (2008).

47. Takigawa, H. *et al.* Mesenchymal Stem Cells Induce Epithelial to Mesenchymal Transition in Colon Cancer Cells through Direct Cell-to-Cell Contact. *Neoplasia* **19**, 429-438 (2017).

48. Yu, M. *et al.* Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. *Science* **339**, 580-4 (2013).

49. Biddle, A., Gammon, L., Liang, X., Costea, D.E. & Mackenzie, I.C. Phenotypic Plasticity Determines Cancer Stem Cell Therapeutic Resistance in Oral Squamous Cell Carcinoma. *EBioMedicine* **4**, 138-45 (2016).

50. Wiederschain, D. *et al.* Single-vector inducible lentiviral RNAi system for oncology target validation. *Cell Cycle* **8**, 498-504 (2009).

51. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-20 (2014).

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

53. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-74 (2012).

54. Gentleman, R.C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80 (2004).

55. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).

56. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Controlling the False Discovery Rate. . *Journal of the Royal Statistical Society.* **57**, 289-3000 (1995).

57. Meinders, M. *et al.* Sp1/Sp3 transcription factors regulate hallmarks of megakaryocyte maturation and platelet formation and function. *Blood* **125**, 1957-67 (2015).

58. Edgar, R., Domrachev, M. & Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207-10 (2002).

Figures and Legends

Figure 1

A. Flow cytometric analysis of the colon cancer cell lines HCT116 (left panel) and SW480 (right panel) with antibodies directed against CD44 and EpCAM. EpCAM/CD44 positive and negative regions were defined using multiple isotype controls, and are shown by the quadrants in the plots. Notably, both HCT116 and SW480 revealed a continuum of different EpCAM and CD44 expression levels with a large $CD44^{high}EpCAM^{high}$ ($EpCAM^{hi}$) cluster followed by a tail of gradually decreasing EpCAM and increasing CD44 levels. By applying specific gates, cells were divided in a large $EpCAM^{hi}$ cluster, together with a considerably smaller $CD44^{high}EpCAM^{low}$ ($EpCAM^{lo}$) subpopulation. To ensure good separation from the large $EpCAM^{hi}$ cluster and maximal sorting purity, $EpCAM^{lo}$ cells were gated as $CD44^{hi}$ events ≤60% of the EpCAM fluorescence intensity of the left border of the $EpCAM^{hi}$ gate, and sorted from ≤50% of that value. Variable percentages of $EpCAM^{lo}$ cells were found to feature the HCT116 (5.0±2.5%) and SW480 (16.7±13%) cell lines, respectively. For the sake of simplicity, gates are shown in the figure only if they encompass sizeable percentages of cells. Graphs show representative analysis of one experiment.

B. Phase contrast microscopy images of sorted $EpCAM^{hi}$ and $EpCAM^{lo}$ cells from HCT116 (upper images) and SW480 (lower images) cells. While $EpCAM^{hi}$ cells formed compact colonies with characteristic epithelial morphology, $EpCAM^{lo}$ cells showed a more spindle- and mesenchymal-like appearance Scale bar: 100 µm.

C. Intrasplenic injection of bulk, $EpCAM^{hi}$, and $EpCAM^{lo}$ cells from HCT116 (left panel), and SW480 (right panel) in For each transplantation experiment, $2x10^4$ cells were injected in the spleen of a recipient NSG mouse. Four (HCT116) and 8 (SW480) weeks after injection mice were sacrificed and individual tumors counted. Single and double asterisks indicate significant differences ($P < 0.05$ and $P < 0.01$, respectively).

HCT116: bulk (n=8), $EpCAM^{hi}$ (n=9), and $EpCAM^{lo}$ (n=7). SW480: bulk (n=4), $EpCAM^{hi}$ (n=4), and $EpCAM^{lo}$ (n=4).

D. Images of mouse livers four (HCT116) and eight (SW480) weeks after orthotopic injection with 10^4 cells. Scale bar: 5 mm.

Figure S1-1

A. FSC-A/SSC–A, FSC-W/FSC-A, and SSC-W/SSC-A single cell gates (confirmed by gating on FSC-A/FSC-H). Purity of sorted single cells was confirmed by microscopy.

B. Acquisition parameters used for FACS analysis.

C. Gating strategy employed to select for live cells by DAPI staining (top: unstained cells; bottom: after DAPI staining).

D. Top: unstained and isotype controls, showing absence of relevant unspecific antibody binding. Bottom: Fluorescence Minus One (FMO) samples, showing the absence of uncompensated spillover between fluorescence channels. For the sake of clarity, a quadrant is placed to show the relative position of unstained and single stains in both cells lines. For the EpCAM-FITC antibody, a mouse IgG1-FITC isotype control S. Cruz sc-2855 was used; for the CD44-APC antibody, a Rat IgG2a-APC isotype control S. Cruz sc-2895 was used. Similar results were obtained with SW480 cells (not shown).

E. Full EpCAM/CD44 staining, also showing the rationale behind the definition of high and low referred to EpCAM and CD44 levels. Despite the differences in CD44 levels between EpCAM^{hi} and EpCAM^{lo} cells, for the sake of simplicity the populations were defined "low" for a defined marker if they were mainly covering a region within the 1st log above the negative gate, high if they were above the 1st log.

Figure S1-2

A. The percentage of CD44^{hi}/EpCAM^{lo} subpopulation was determined in a panel of commonly employed colon cancer cell lines by flow cytometric analysis. Please note that, within the same cell line, these percentages can vary depending on the passage number and culture conditions. If indicated by Lindeman et. al CMS classification of the cell lines is shown.

B. Simulation of the HCT116 two state Markov Model with a non-homogenous starting population. The starting population was defined as ten subclones with varying transition probability P_HL at equal proportion. The Markov Model was run till day 200 and shows that the total EpCAM^{lo} population decreases over time (top panel), because clones with lower plasticity will gain dominance in the long run (lower panels).

Figure S1-3

A. Top panel: transwell migration assay of EpCAM^{hi} (black bar) and EpCAM^{lo} (grey bar) cells from the HCT116 and SW480 lines. 10^5 cells were plated on TC-coated membrane in triplicate and left overnight before counting the number of migrated cells on the bottom side of the membrane. Each bar represents the mean ± SD of two

independent experiments. Asterisks indicate significant differences ($P < 0.05$). Bottom panel: invasion assay of EpCAM^{hi} (black bar) and EpCAM^{lo} (grey bar) cells from the HCT116 and SW480 lines. 10^5 cells were plated in triplicate on top of an extracellular matrix-coated coated membrane and left overnight before counting the number of cells migrated to other side of the membrane. Each bar represents the mean ±SD of two independent experiments. Asterisks indicate significant differences ($P < 0.05$).

B. RT-qPCR expression analysis of epithelial (*EPCAM* and *CDH1*) and mesenchymal (*VIM*) markers in sorted EpCAM^{hi} (black bars) and EpCAM^{lo} (grey bars) from the HCT116 and SW480 lines. *GAPDH* was employed for normalization purposes. Each bar represents the mean ±SD of three independent experiments. Asterisks indicate significant differences ($P < 0.05$).

C. RT-qPCR expression analysis of EMT transcription factors (*ZEB1, ZEB2, TWIST, FOXC2, SLUG* and *SNAIL*) in EpCAM^{hi} (black bars) and EpCAM^{lo} (grey bars) cells. Left panel: HCT116. Right panel: SW480. *GAPDH* was employed for normalization. Each bar represents the mean ± SD of three independent experiments. Asterisks indicate significant differences ($P < 0.05$).

D. Immunofluorescence (IF) analysis of EpCAM^{hi} and EpCAM^{lo} cells. Cells were sorted and directed plated on cover slips. After four days cells were fixed with 4% PFA and stained with antibodies against EpCAM (green) and ZEB1 (red). Nuclei were visualized by DAPI staining of DNA (blue). Scale bar: 50 µm.

Figure S1-4

A. qRT-PCR expression analysis of *ZEB1* in HCT116 and SW480 transduced with an inducible control (shCT) or ZEB1-shRNA (shZEB1) construct. shRNA expression was induced with 1 µg/mL of doxycycline. Each bar represents the mean ± SD of three independent experiments.

B. Bar graph of flow cytometric analysis (see **B**. here below). Each bar represents the relative mean ± SD of three independent experiments.

C. Representative analysis of the flow cytometric analysis of the shCT- and shZEB1-transfected HCT116 and SW480 cell lines using antibodies against CD44 and EpCAM. Cells were induced with 1 µg/mL doxycycline for 72 hrs. before analysis.

D. RT-qPCR expression analysis of the members of the miRNA 200 family (miR-200a, miR-200b, miR-200c, miR-141 and miR-429) in EpCAM^{hi} (black bars) and EpCAM^{lo} (grey bars) cells. Upper panel: HCT116. Bottom panel: SW480. *U6* was employed for normalization. Each bar represents the mean ± SD of three independent experiments. Single asterisks indicate significant differences of $P < 0.05$, double asterisks of $P < 0.01$, and triple asterisks of $P < 0.001$.

E. Cell proliferation assay. Sorted bulk, EpCAM^{hi} and EpCAM^{lo} cells were seeded in triplicate in plates and cultured in conventional medium. HCT116 and SW480 cells were harvested and number of cells was counted at 4 and 11 days, respectively. Each bar represents the mean ± SD three independent experiments.

F. Cell cycle analysis of EpCAM^{hi} and EpCAM^{lo} cells in HCT116 (upper panel) and SW480 (lower panel). Cells fractions were sorted and plated in culture. After 72 hours cells were fixed and stained with propidium iodide. Cell cycle distribution was assayed by flow cytometry. Graphs show representative analysis of one experiment. Tables demonstrate average and standard deviation of three independent experiments. White graph: EpCAM^{hi}, grey graph: EpCAM^{lo}. Asterisks show the significant ($P < 0.05$) differences between EpCAM^{hi} and EpCAM^{lo} cells in G_1 and G_2M-phases.

Figure S1-5

Bulk, EpCAM^{hi} and EpCAM^{lo} HCT116 and SW480 cells were sorted and plated to recover and expand for four days. 10^5 cells per well were plated in in triplicate in a 96-well plate and left to attach. After 24 h, standard cell culture medium was changed for medium containing chemotherapeutic drug oxaliplatin or 5-FU. HCT116 cells were left with 5-FU and oxaliplatin for respectively 3 and 5 days. For each drug SW480 cells were treated for 7 days. After removal of the chemotherapeutic drug, cells were washed with PBS and left to re-grow in standard culture medium. Cell viability was assessed using the MTT upon removal of the drug, and after re-growth (ranging 13-30 days). O.D. reading was performed at 595 nm with a Microplate Reader. Background measurements were subtracted from each data point. All experiments were performed at least in duplicate for each individual cell line and drug.

A-B. Chemo sensitivity assays using oxaliplatin (**A**) or 5-FU (**B**) in bulk (black bars), EpCAM^{hi} (dark grey bars) and EpCAM^{lo} (light grey bars) HCT116 cells. Bars represent the average ± SD of the O.D. corrected for the background measurements.

C-D. Chemo sensitivity assays using oxaliplatin (**C**) or 5-FU (**D**) in bulk (black bars), EpCAM^{hi} (dark grey bars) and EpCAM^{lo} (light grey bars) SW480 cells. Bars represent the average ± SD of the O.D. corrected for the background measurements.

Figure S1-6

A. Hematoxylin and eosin (H&E; first two panels) and IHC with antibody directed against β-catenin (third and fourth panel) in liver metastasis obtained four or eight weeks after intrasplenic injection with HCT116 (upper panels) and SW480 (lower panels) cells, respectively. Second and fourth panels show zoom of the marked area in respectively first and third panel. Scale bar first and third panels: 100 µm. Scale bar second and fourth panel: 50 µm.

B. FACS analysis of liver metastases obtained by spleen injection of HCT116 bulk, EpCAM^{hi}, and EpCAM^{lo} cells. Upper panel: representative FACS plots. Bottom table: quantification of EpCAM$^{hi/lo}$ percentages of all liver metastases as determined by FACS.

C. Microscopic (top, left panel) analysis of AKP-Z organoids tagged with GFP and click beetle luciferase. The graph (bottom, left panel) shows the RT-qPCR expression analysis of *Zeb1* in AKP (left) and AKP-Z (right) organoids upon *in vitro* doxycycline treatment for 48 hrs. Black bars: no doxycycline treatment; grey bars: 1 µg/mL doxycycline. Each bar represents the mean ± SD of three independent experiments.

D. Upon establishment of a primary tumor in the caecum, transplanted mice were administered doxycycline in the drinking water to induce Zeb1 expression. FACS analysis of the primary tumor was performed one week after the start of the doxycycline. The panels show representative FACS plots of a control and two dox-treated primary tumors.

E. Quantification of the number of lung and liver metastases in uninduced (black; n=4) and dox-induced (grey; n=5) AKP-Z transplanted mice. Liver tissue was cut into 500-micron slices, processed for IHC, stained for β-catenin to visualize tumor cells, scanned using a NanoZoomer, and counted using NDP view software. The area of tissue analyzed was used to normalize the data.

Figure 2

A. Analysis of plasticity of EpCAM^{hi} and EpCAM^{lo} cells from HCT116 (upper panel) and SW480 (lower panel). EpCAM^{hi} and EpCAM^{lo} cell fractions were sorted and plated in culture. At different time points, as indicated, cells were re-analyzed by flow cytometry for their levels of CD44 (*y*-axis) and EpCAM (*x*-axis) expression.

B. Flow cytometric analysis of single CD44^{hi}EpCAM^{hi} and CD44^{hi}EpCAM^{lo} HCT116 cells sorted by FACS and cultured for 72 days. Three representative individual single cell clones per cell fraction are shown.

C. Dynamics of the EpCAM^{hi} and EpCAM^{lo} sub-populations from the HCT and SW480 cell lines as measured by FACS (% of total) over time. Under each graph, a schematic shows the estimated transition probabilities from the fitted two-state Markov Model.

Figure S2-1

A. RT-qPCR expression analysis of *EPCAM*, *CDH1*, *VIM*, and *ZEB1* in 'early' (e.g. cells that were used shortly after FACS sorting) and 'late' (e.g. cells that were cultured for an extended period of time before performing the experiment) sorted EpCAM^{hi} and EpCAM^{lo} cell from HCT116 and SW480 cell lines. *GAPDH* was employed for normalization purposes. Each bar represents the mean ± SD of two independent experiments.

B. Transwell migration assay of 'early' and 'late' EpCAM^{hi} (black bar) and EpCAM^{lo} (grey bar) cultures in HCT116. 10^5 cells were plated in triplicate on TC-coated membrane and left overnight before counting the number of migrated cells on the bottom side of the membrane. Each bar represents the mean ± SD of two independent experiments. Asterisks indicate significant differences ($P < 0.05$).

C. Table with selection of individual single cell sorted EpCAM^{hi} and EpCAM^{lo} HCT116 and SW480 clones. Single cells were sorted, grown up and analyzed at two indicated times points by flow cytometry for their CD44 and EpCAM expression. Percentage of cells present in the set EpCAM^{hi} and EpCAM^{lo} gates are presented in the bar graph. Total success rate of clones grown out of single cells: HCT116 EpCAM^{hi} 15 clones out of 24 sorted cells, HCT116 EpCAM^{lo}: 31/48; SW480 EpCAM^{hi}: 5/36; SW480 EpCAM^{lo}: 11/36.

D. Analysis of the HCT116 scRNAseq data as a Markov diffusion process. Markov forward (left) and backward (right) diffusion indicating the presence of sink and source points in both EpCAM^{hi} and EpCAM^{lo} populations.

E. Partition-based graph abstraction (PAGA) velocity graph mapping out the direction of velocity on a subpopulation level in HCT116.

Figure 3

A. Multidimensional scaling (MDS) analysis of RNAseq profiles of EpCAM^{hi} and EpCAM^{lo} cells from the HCT116 and SW480 lines. Red: HCT116, black: SW480, circle: EpCAM^{hi}, triangle: EpCAM^{lo}.

B. Ingenuity Pathway Analysis (IPA) of the HCT116 and SW480 expression profiles from the multi-cell line analysis (*P. adjusted value* < 0.01; $\log_2$ fold change < -1.5 and > 1.5). Red marked pathways highlight the enhanced involvement of pathways involved in EMT, Wnt signaling and the formation of colon cancer metastasis in the EpCAM^{lo} subpopulation compared to EpCAM^{hi} cells.

C. TOP-Flash luciferase reporter analysis of Wnt signaling activity in colon cancer cell lines HCT116 and SW480 upon treatment with 4 µM Chiron for three days. Each bar represents the mean ± SD of two independent experiments.

D. Flow cytometric analysis using antibodies directed against CD44 and EpCAM of control and 4 µM Chiron-supplemented HCT116 (A) and SW480 (B) cultures. Graphs show percentage of cells within the CD44^{hi}EpCAM^{hi} and CD44^{hi}EpCAM^{lo} gates relative to the control. Each bar represents the mean ± SD of two independent experiments.

E. Immunofluorescence analysis of control and Chiron treated HCT116 (left panel) and SW480 (right panel) cells. After three days of treatment, cells were fixed with 4% PFA and stained with antibodies against EpCAM (green) and ZEB1 (red). Nuclei were visualized by DAPI staining of DNA (blue). Scale bar: 100 µm.

F. Flow cytometric analysis of three HCT116 cell lines with differential β-catenin mutation status, a parental HCT116 (HCT116-P) cell line harboring one WT and one mutant allele (Ser45 del), and two HCT116-WT and HCT116-MT cell lines harboring one WT or one mutant allele, respectively, generated by disruption of the other allele in the parental cell line[26].

Figure S3-1

A. Gene list of commonly differentially expressed genes in EpCAM^{lo} versus EpCAM^{hi} cells in HCT116 and SW480. A total of 152 and 353 differentially regulated genes were identified between the CD44^{hi}EpCAM^{hi} and CD44^{hi}EpCAM^{lo} cells in HCT116 and SW480, respectively (*P. adjusted* < 0.01, when applying a Log_2 fold change of < -1.5 and > 1.5).

B. Gene ontology (GO) of molecular and cellular functions in HCT116 (upper panel), SW480 (middle) and the combined (bottom panel) gene lists.

Figure S3-2

Hematoxylin and eosin (H&E) and IHC analyses with antibodies directed against β-catenin and ZEB1 in consecutive sections of colon cancers from three unrelated patients with sporadic colon cancer. For each case, areas from the tumor center and the invasive front are shown. Co-localization of nuclear β-catenin and ZEB1 expression was found in 5 out of 25 cases investigated. Scale bar: 50 µm.

Figure 4

A. Correlation of meta-gene expression values of the signatures derived from EpCAM^{lo} HCT116 and SW480 cells in the CMS-3232 human colon cancer cohort[27].

B. Correlation of meta-gene expression values of the signatures derived from EpCAM^{lo} HCT116 and SW480 cells with expression of CMS classifier genes positively identifying each of the four molecular subtypes.

C. Kaplan Meier analysis. The gene sets identifying the EpCAM^{lo} cells from both HCT116 and SW480 cell lines were used to cluster the tumors in the CMS3232 cohort into HIGH (purple), INTERMEDIATE (blue) and LOW (yellow) expression groups by k-means clustering. The Kaplan-Meier method was subsequently used to assess significant differences in relapse-free (left panels) and overall (right panels) survival between the generated subgroups.

Figure S4-1

A. Correlation of meta-gene expression values of the signatures derived from EpCAM^{lo} HCT116 and SW480 cells in two cohorts of colon cancer cell lines in the absence of a stromal component. CCS1: epithelial-like tumor lines; CCS2: tumor lines with microsatellite instability. CCS3: mesenchymal-like tumor cell lines[29].

B. The gene sets identifying the EpCAM^{lo} cells from both HCT116 and SW480 were used to cluster the tumors in the CMS3232 cohort into High, Intermediate, and Low expression groups by k-means clustering. As all tumors in this cohort had previously been assigned to specific CMS subgroups, and then analyzed the contribution of each CMS subtype to each of the generated High, Intermediate, and Low expression subgroups.

Figure 5

A. tSNE of HCT116 and SW480 cells based on the variable expressed genes across EpCAM^{lo} and EpCAM^{hi} populations.

B. tSNE of HCT116 and SW480 cells based on genes from the EMT signature (N = 107).

C. Heatmap of differentially expressed genes between EpCAM^{lo} and EpCAM^{hi} populations. Cells were ranked according to their EMT score (EMT = M – E).

D. Gene expression trends projected over the EMT axis. Expression values were imputed with MAGIC, scaled by their z-score and smoothened by general additive models to visualize the gene expression trend.

E. RNA velocity analysis of the HCT116 scRNAseq data. Cells from both populations were moving in their respective state (left panel) with a minority population of transitioning cells (right panel).

F. Projection of the pEMT score on the UMAP embedding of the HCT116 cell line.

G. The expression of SPARC on the UMAP embedding of the HCT116 cell line.

H. Left panels: qPCR analysis of SPARC overexpression in the subpopulations of HCT116 and SW480. Gene expression values are relative to GAPDH and normalized to the EpCAM^{hi} subpopulation. Right panel: quantification of the transwell migration assay upon overexpression of SPARC.

Figure 5S-1

A. tSNE of the SW480 cell line indicating an additional subpopulation in the EpCAM^{hi} population (left panel). Using a signature derived from the bulk RNAseq, this population was identified as the “sphere” population (middle panel) and annotated to be excluded for further analysis (right panel).

B. Left panel: the SW480 cell line, after exclusion of the “sphere” population, contains slightly higher variability compared to the HCT116 cell line, as evidenced by the variance of the top-50 principal components. Right panel: while in HCT116, most of the variable expressed genes are differentially expressed between the EpCAM^{hi} and EpCAM^{lo} population, this is not the case in SW480, where most of the highly variable genes do not differ between the two populations.

C. Top panels: expression values of *VIM*, *ZEB1* and *CD44* on the UMAP embedding of the HCT116 cell line. Lower panels: projection of the RNA velocity direction of the same genes.

D. HCT116 UMAP embedding annotated with the eight unsupervised clusters.

E. Heatmap of HCT116 with expression values of the EMT signature averaged by the eight clusters. Clusters were ranked according to their EMT score and genes were clustered in four distinct gene sets using k-means clustering.

F. Schematic diagram showing a transcriptional trajectory with distinct gene arrays through which pEMT cells arise.

Figure 6

A. Expression of the CMS4_RF signature in the scRNAseq data from Lee et al. (N = 91,103 cells) indicates high association to the stromal cells.

B. Expression of the CMS4_RF signature in the bulk RNAseq data from Guinney et al. (N = 3232 tumors) shows association to CMS4 tumors. Tumors were grouped in 3 equal groups according to their association with the CMS4_RF signature.

C. Kaplan-Meier plot of the 3 CMS4_RF groups shows significant differences in relapse-free survival.

D. Expression of the CMS4_TC signature in the scRNAseq data reveals high association to the tumor epithelial cells.

E. Expression of the CMS4_TC signature in the bulk RNAseq data shows association to CMS1 and CMS4 tumors. Tumors were grouped in 3 equal groups according to their association with the CMS4_TC signature.

F. Kaplan-Meier plot of the 3 CMS4_TC groups shows significant differences in relapse-free survival.

G. Expression of the CMS4_TC signature in the fraction of epithelial cells (N = 24707 cells). Cells from CMS1 and CMS4 tumors show the highest association to the CMS4_TC signature, and were annotated as mes-like.

H. Violin plots of normal, tumor and mes-like tumor cells showing expression patterns across different genes.

Figure S6-1

Survival analysis of the CMS4_RF (top panels) and CMS4_TC (lower panels) signatures across the different consensus molecular subtypes. Patients were divided into 2 equal groups based on their association to the signatures and Kaplan-Meier plots were made based on relapse-free survival.

Figure 1

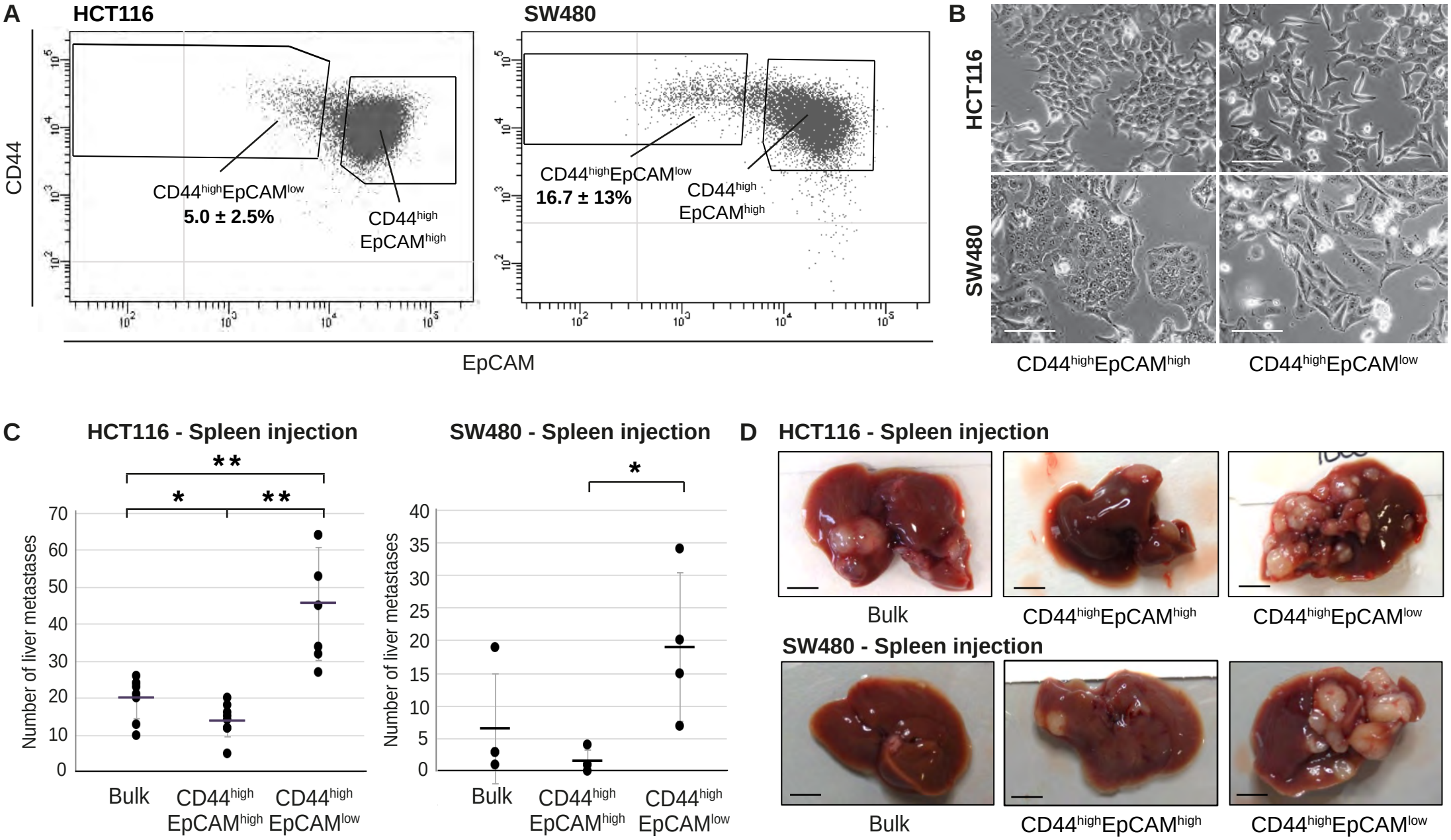

A
HCT116
SW480
CD44
EpCAM
CD44highEpCAMlow
5.0 ± 2.5%
CD44high
EpCAMhigh
CD44highEpCAMlow
16.7 ± 13%
CD44high
EpCAMhigh
B
HCT116
SW480
CD44highEpCAMhigh
CD44highEpCAMlow
C
HCT116 - Spleen injection
SW480 - Spleen injection
Number of liver metastases
Bulk
CD44high EpCAMhigh
CD44high EpCAMlow
D
HCT116 - Spleen injection
Bulk
CD44highEpCAMhigh
CD44highEpCAMlow
SW480 - Spleen injection
Bulk
CD44highEpCAMhigh
CD44highEpCAMlow

Figure 1 - Figure supplement 1

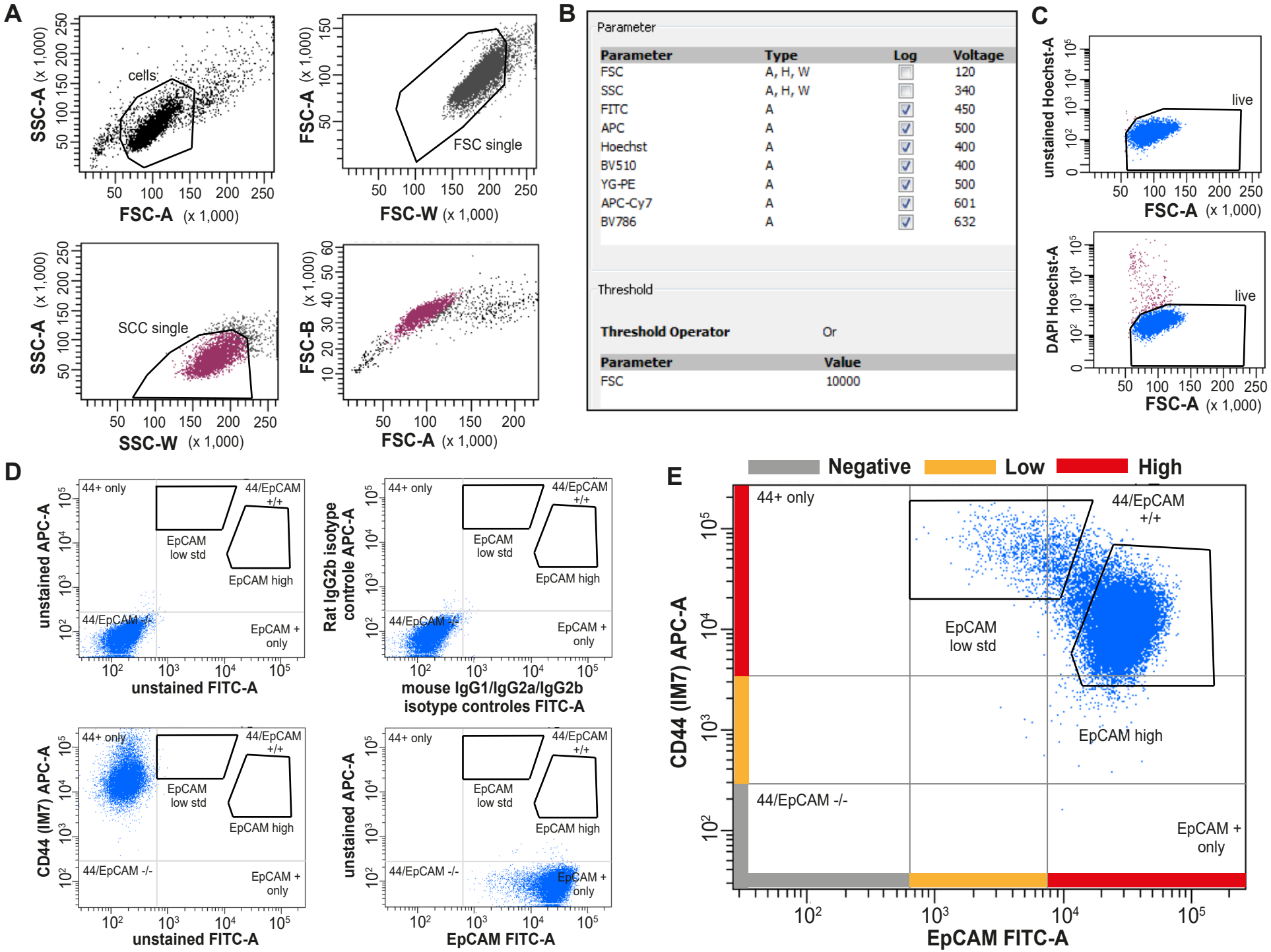


Figure 1 - Figure supplement 2

A

	% CD44^{high}EpCAM^{low}	Linnekamp et al.
HCT116 - early culture	5.0 ± 2.5	CMS1
HCT116 - late culture	0.32 ± 0.18	
SW480 - early culture	16.7 ± 13	
SW480 - late culture	0.34 ± 0.24	
RKO	100	CMS1
COLO320-HSR	99.6	CMS4
NCI-H716	66.0 ± 21.21	CMS4
SW620	12.0 ± 4.9	no subtype
CAR-1	0.15 ± 0.07	CMS4
HT29	0.1 ± 0.02	no subtype
T84	0.1	CMS2
COLO205	0.075	no subtype
LS123	0.06 ± 0.03	CMS4
HT55	0.05	CMS2
Hutu805	0.03 ± 0.0	CMS4
DLD1	0.024 ± 0.001	
LS174	0.007	
HT15	0.003	
CACO2	0.003 ± 0.001	
SW48	0.002 ± 0.001	CMS1
Oums23	0	CMS4
LOVO	0	CMS1

B

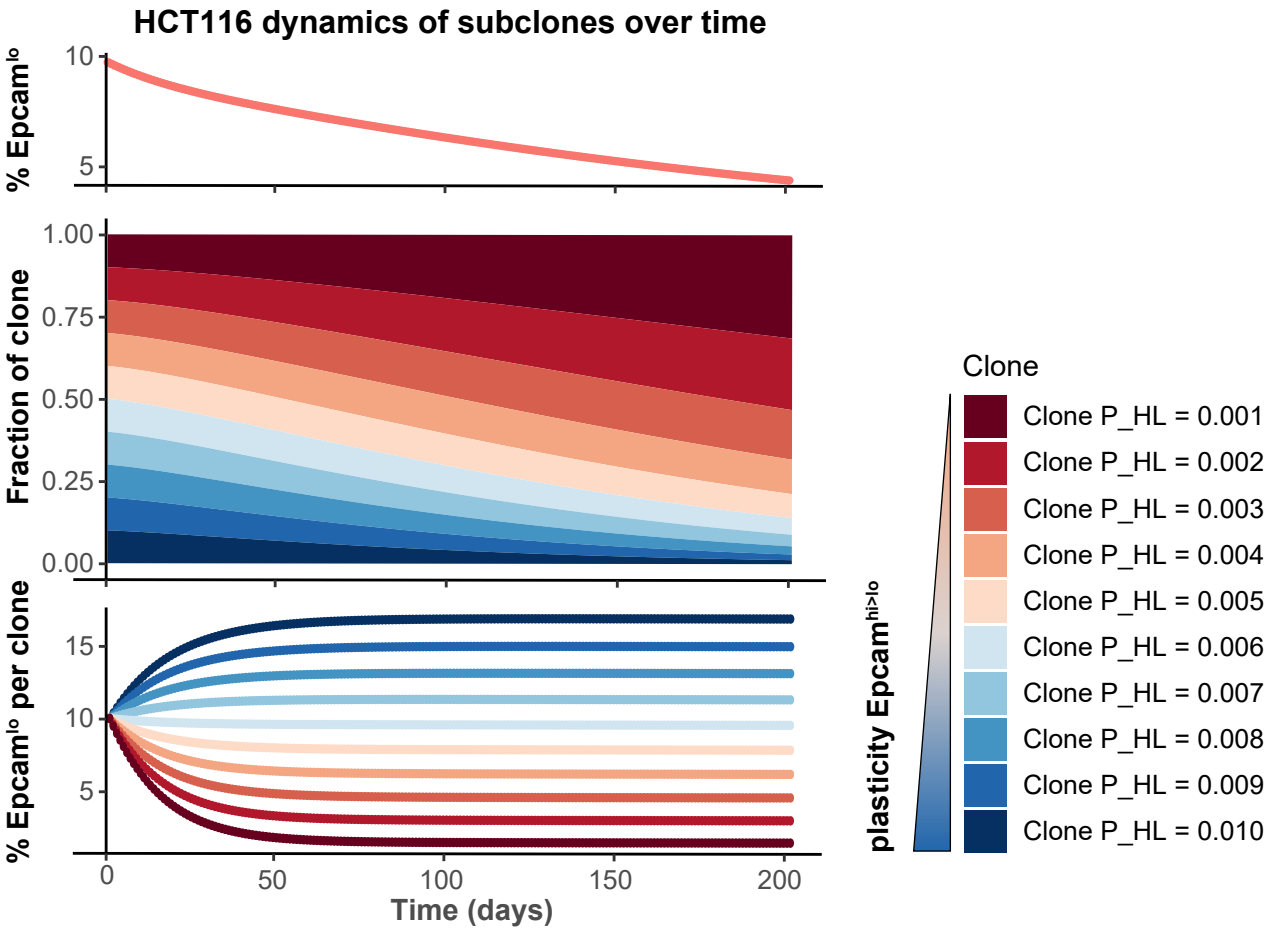


Figure 1 - Figure supplement 3

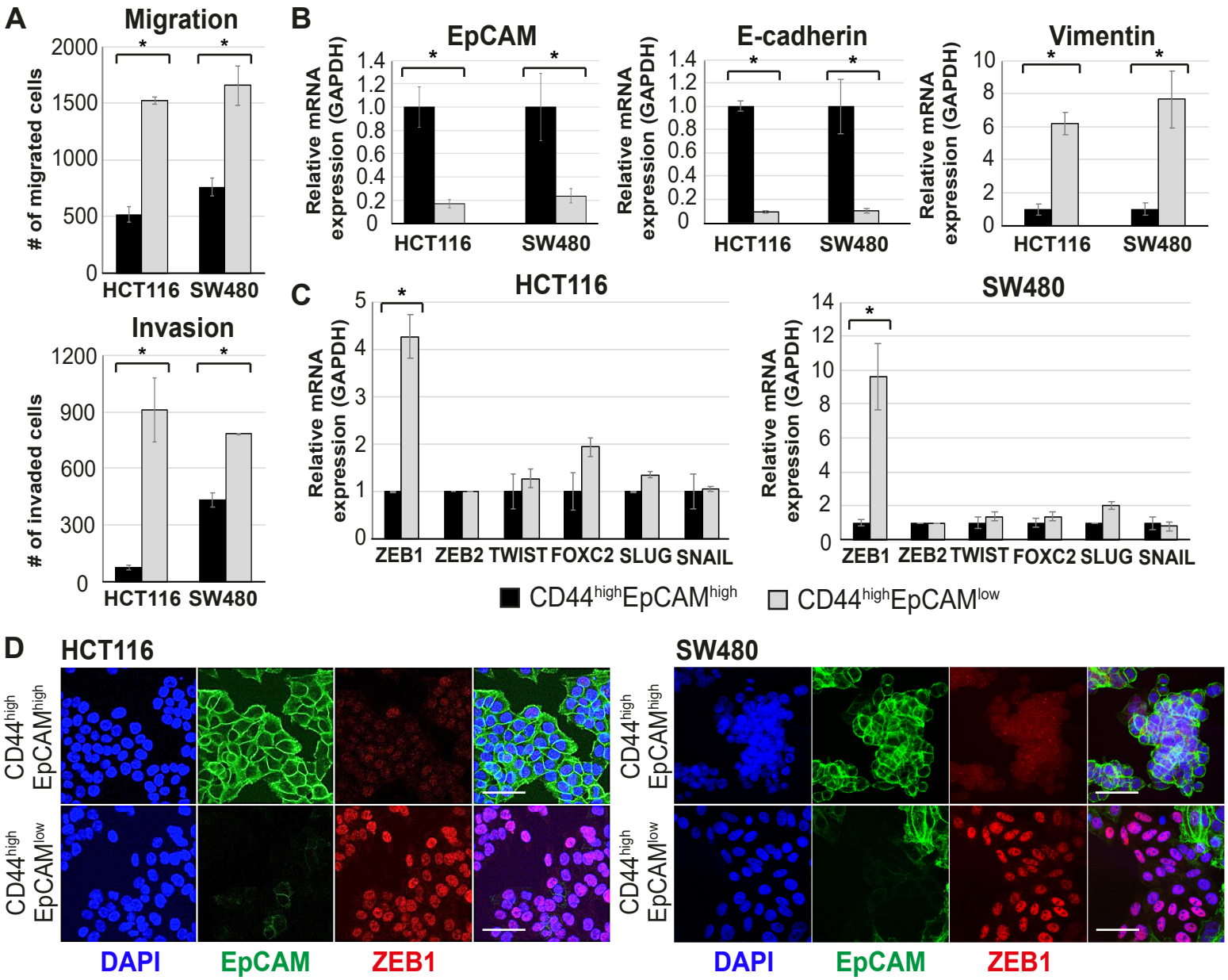
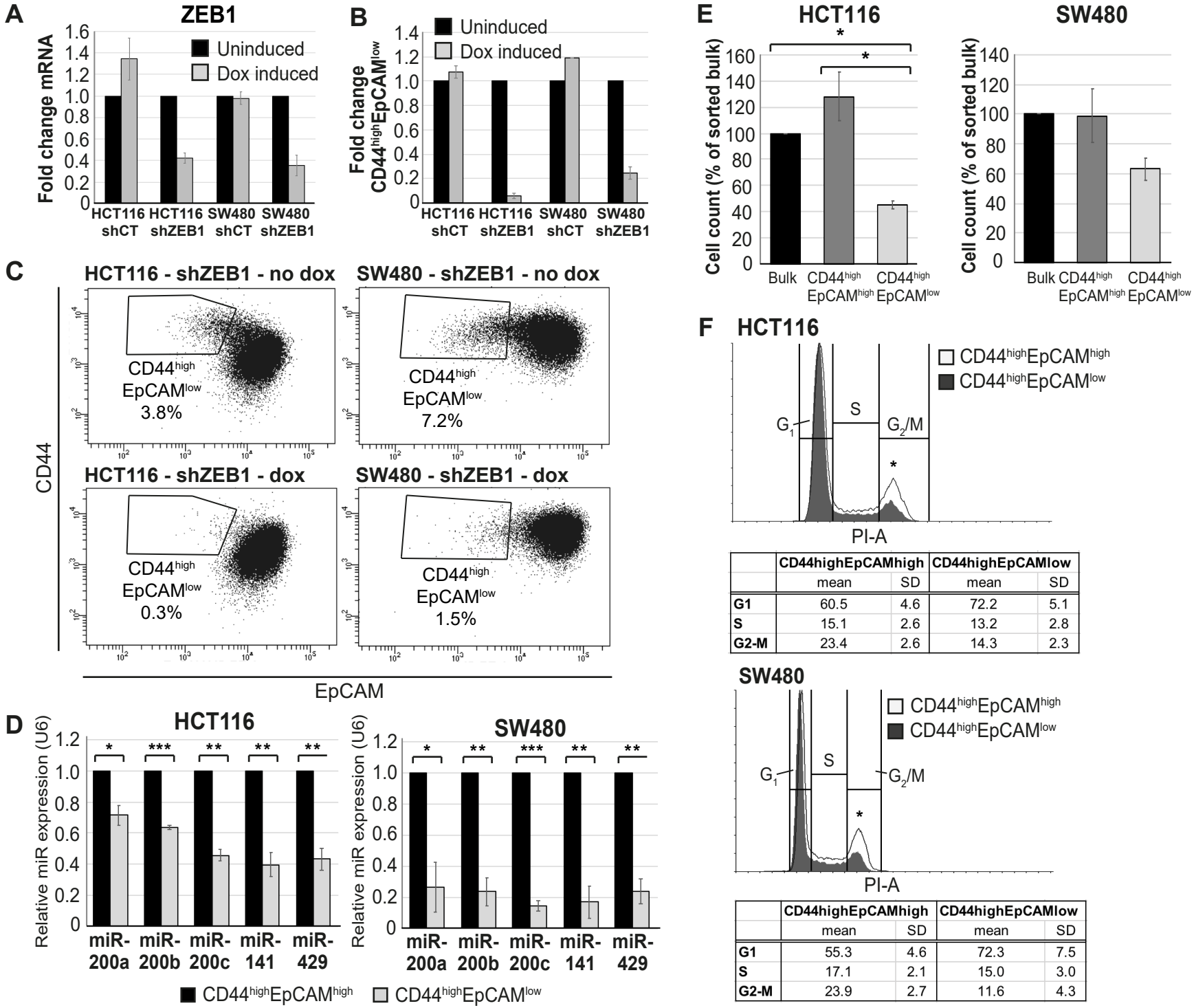


Figure 1 - Figure supplement 4



	CD44highEpCAMhigh		CD44highEpCAMlow	
	mean	SD	mean	SD
G1	60.5	4.6	72.2	5.1
S	15.1	2.6	13.2	2.8
G2-M	23.4	2.6	14.3	2.3

	CD44highEpCAMhigh		CD44highEpCAMlow	
	mean	SD	mean	SD
G1	55.3	4.6	72.3	7.5
S	17.1	2.1	15.0	3.0
G2-M	23.9	2.7	11.6	4.3

Figure 1 - Figure supplement 5

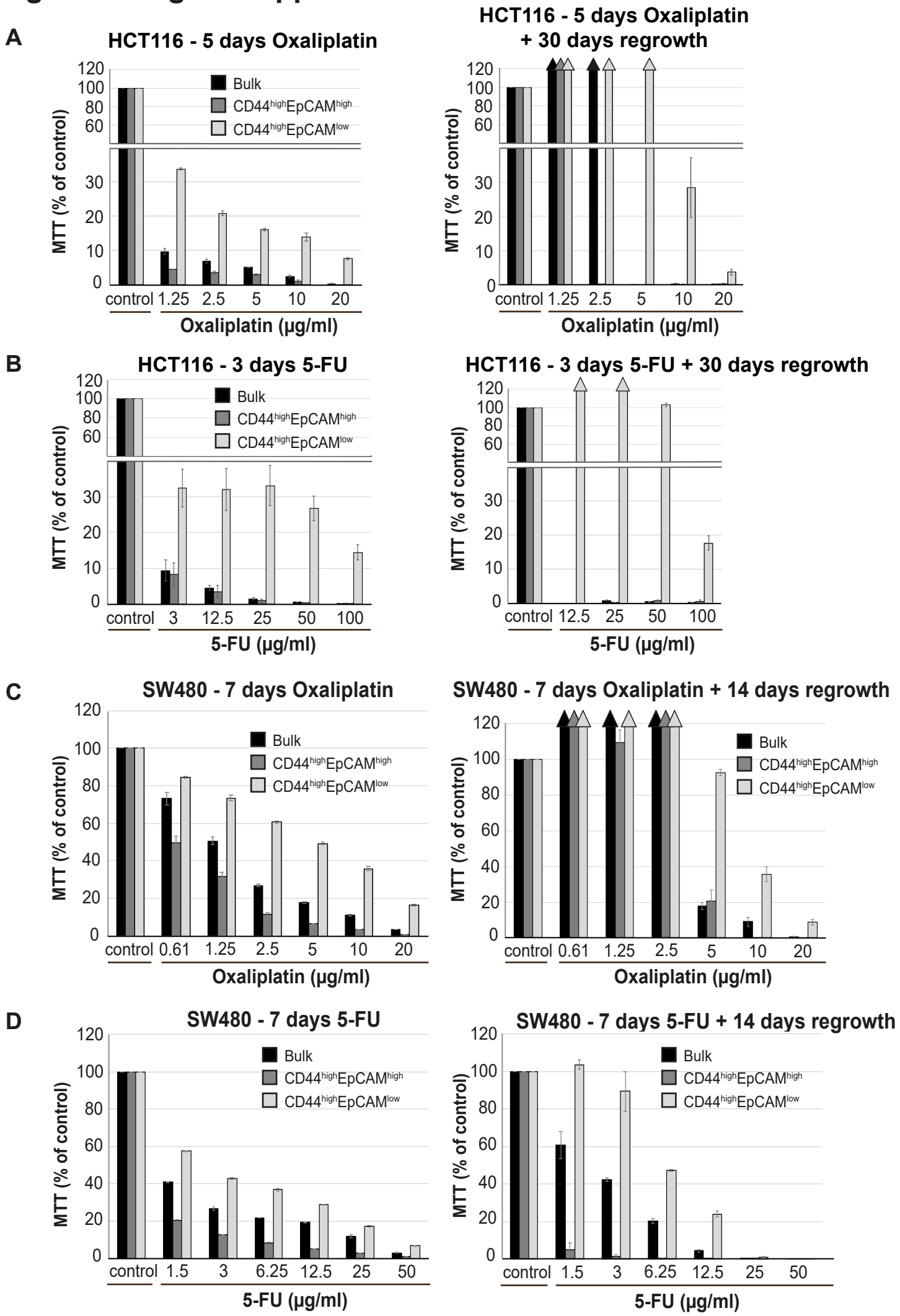
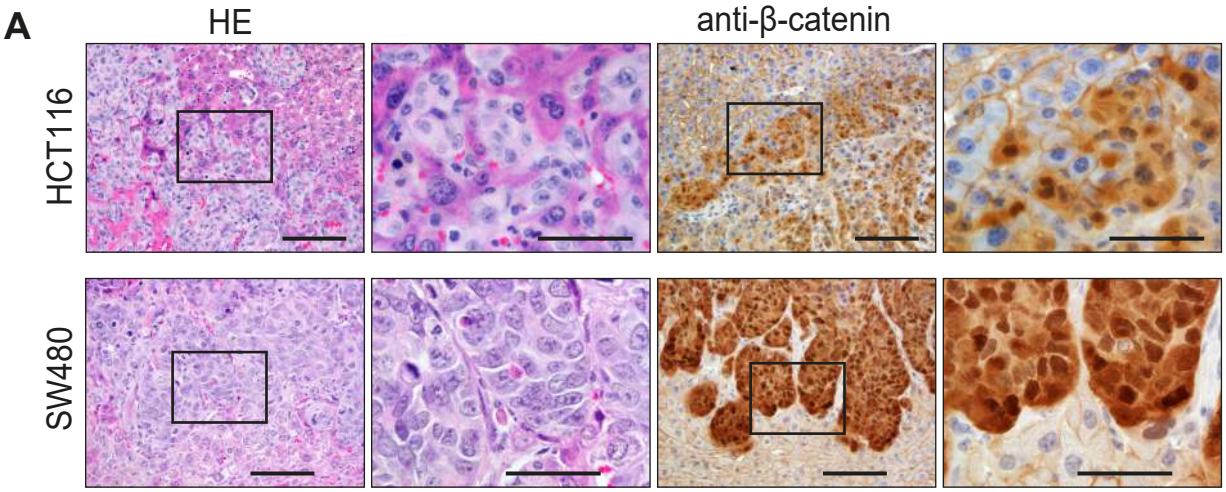
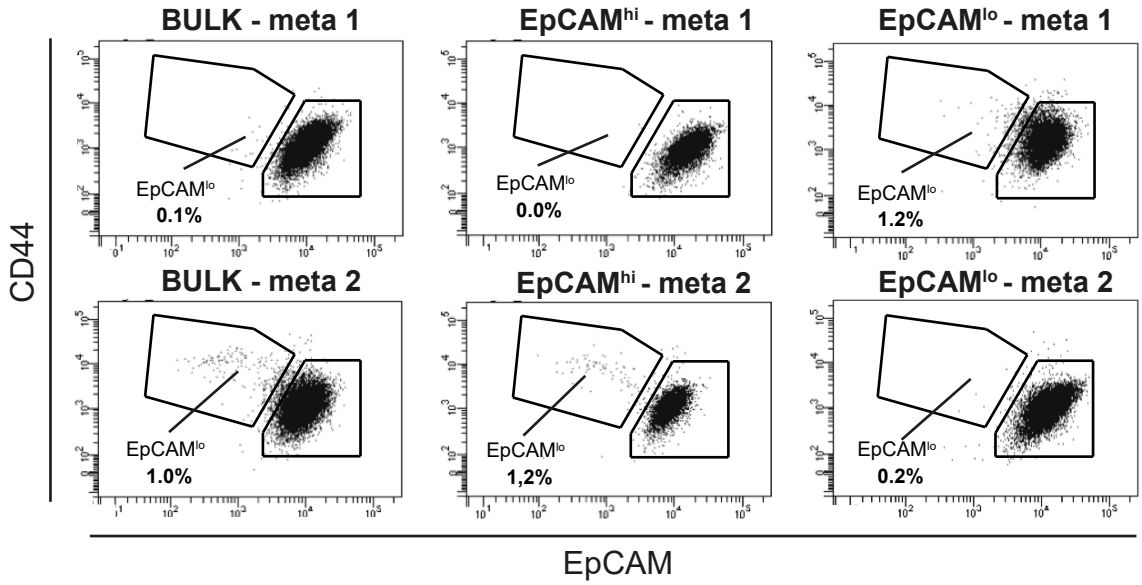
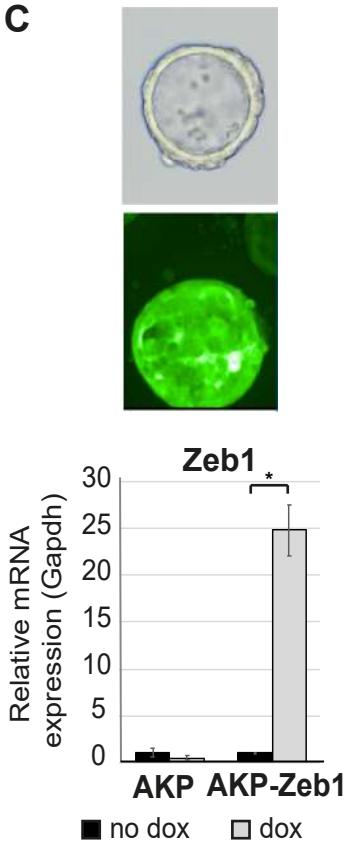


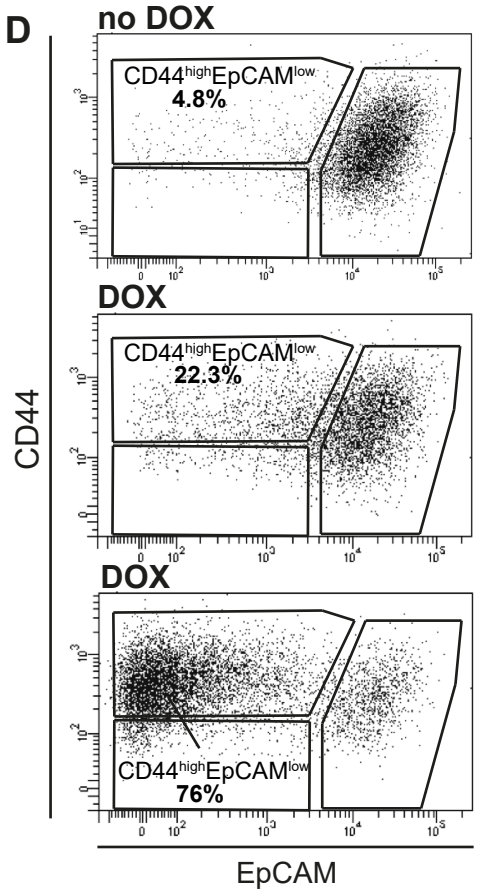
Figure 1 - Figure supplement 6





	bulk #1	bulk #2	bulk #3	bulk #4	bulk #5	EpCAM^{hi} #1	EpCAM^{hi} #2	EpCAM^{lo} #1	EpCAM^{lo} #2	EpCAM^{lo} #3	EpCAM^{lo} #4	EpCAM^{lo} #5
% EpCAM^{hi}	0.1	1.0	0.0	0.1	0.0	0.0	1.2	1.2	0.2	4.2	1.1	51.9
% EpCAM^{lo}	99.1	95.3	98.8	99.2	99.7	99.6	97.7	91.0	99.4	83.2	96.8	25.3





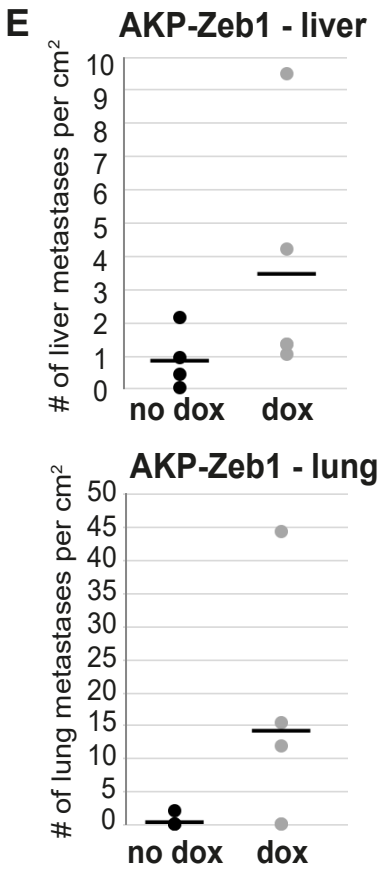


Figure 2

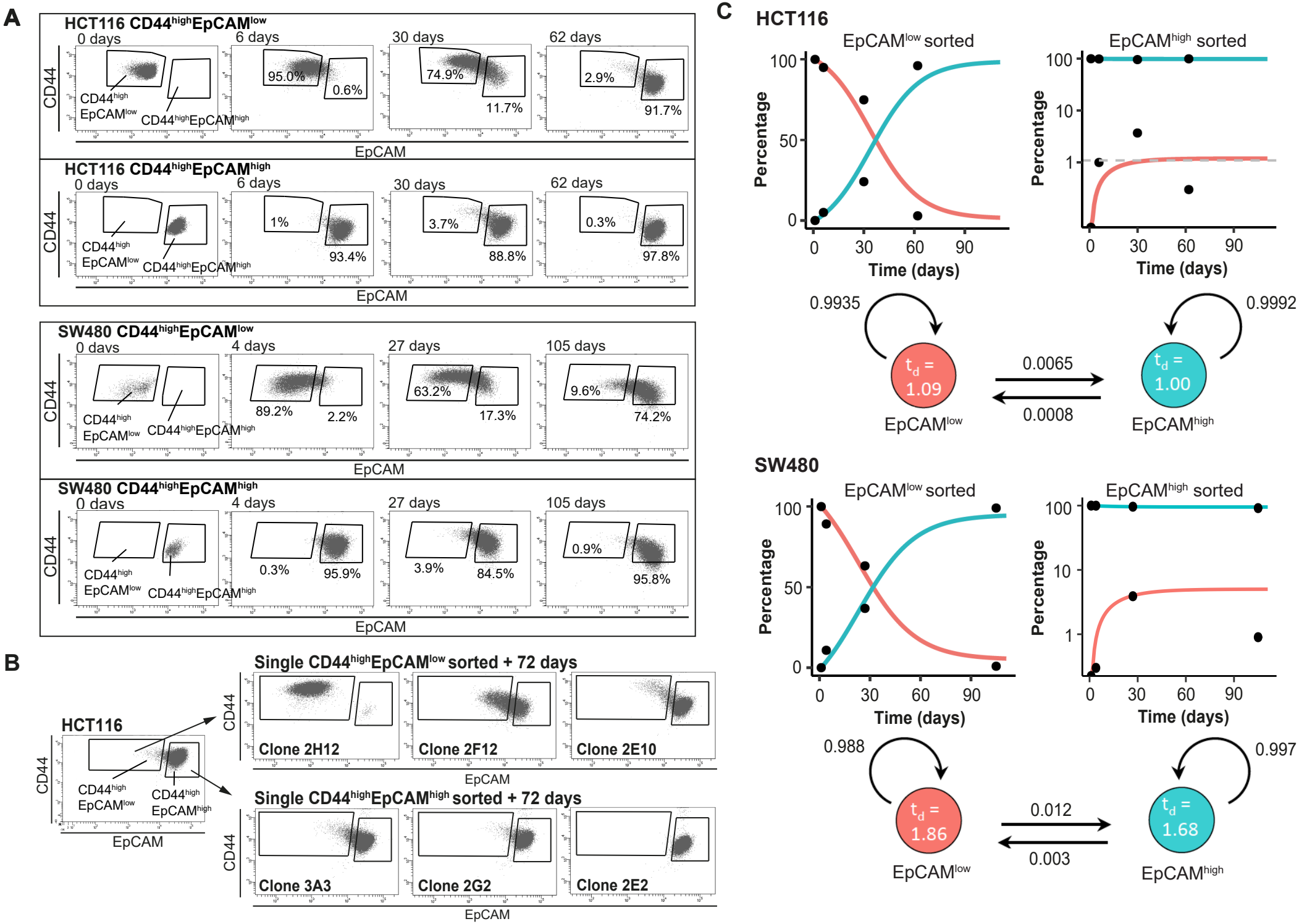
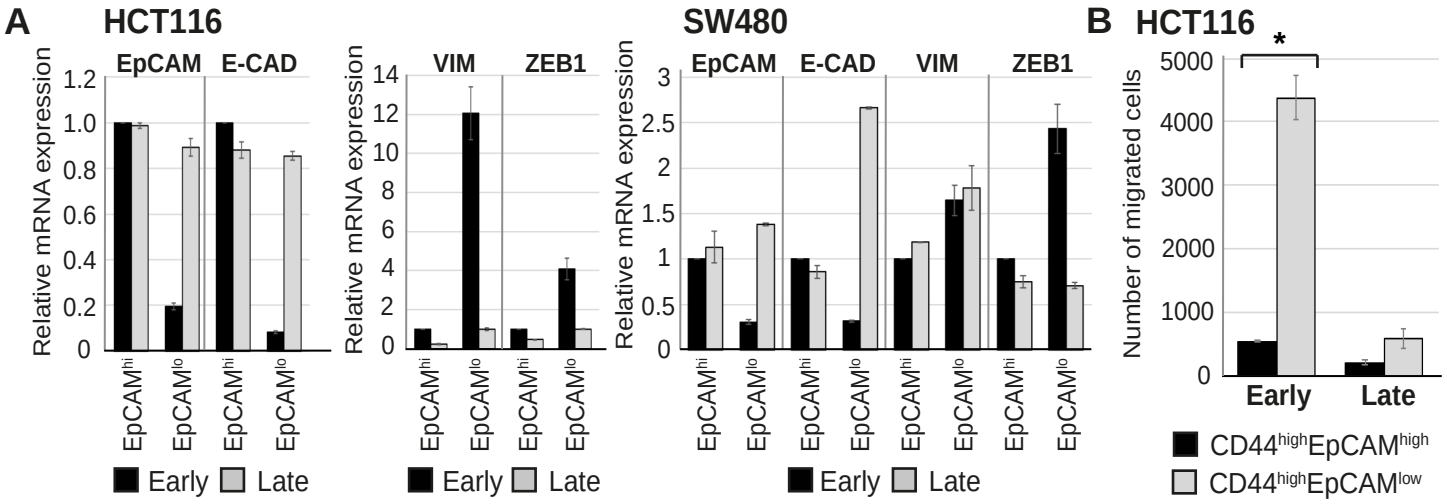

A
HCT116 CD44highEpCAMlow
0 days
6 days
30 days
62 days
CD44
EpCAM
CD44high EpCAMlow
CD44high EpCAMhigh
95.0%
0.6%
74.9%
11.7%
2.9%
91.7%
HCT116 CD44highEpCAMhigh
1%
93.4%
3.7%
88.8%
0.3%
97.8%
SW480 CD44highEpCAMlow
4 days
27 days
105 days
89.2%
2.2%
63.2%
17.3%
9.6%
74.2%
SW480 CD44highEpCAMhigh
0.3%
95.9%
3.9%
84.5%
0.9%
95.8%
B
HCT116
Single CD44highEpCAMlow sorted + 72 days
Clone 2H12
Clone 2F12
Clone 2E10
Single CD44highEpCAMhigh sorted + 72 days
Clone 3A3
Clone 2G2
Clone 2E2
C
HCT116
EpCAMlow sorted
EpCAMhigh sorted
Percentage
Time (days)
0.9935
0.9992
td = 1.09
td = 1.00
0.0065
0.0008
EpCAMlow
EpCAMhigh
SW480
0.988
0.997
td = 1.86
td = 1.68
0.012
0.003

Figure 2 - Figure supplement 1



C HCT116 - single cell clones

HCT116 - EpCAMlow			EpCAMlow	EpCAMhigh
success rate = 31 clones/48	2F12	18 days	97,2	0
		75 days	28,2	46,7
	2 E10	18 days	97,5	0,2
		72 days	5,1	90,8
	2 E11	18 days	98,8	0
		72 days	43,4	23
	2H11	18 days	98,4	0,02
		72 days	9,3	84,6
	2D10	20 days	97,9	0,6
		72 days	33	47,2
	2B10	20 days	87	2,4
		72 days	44,7	42,9
	2B11	20 days	99,4	0,02
		72 days	57,7	36,2
	2B12	20 days	71,2	6,1
		72 days	17	66,2
	2F11	20 days	98,8	0
		72 days	61,5	25,3
	2G12	20 days	98,7	0,2
		72 days	93	4
	2H12	20 days	98,6	0
		72 days	99,1	0,7
	2A11	20 days	98	0,2
		72 days	18,3	68,2

HCT116 - EpCAMhigh			EpCAMlow	EpCAMhi
success rate = 15 clones/24	2A3	18 days	34,8	36,2
		72 days	0,5	97,3
	2C1	18 days	0,9	93,4
		72 days	0,3	98,1
	3C8	18 days	4	83,6
		72 days	0,24	98,8
	2F2	18 days	0,01	99,7
		72 days	0,02	99,3
	2D1	20 days	0,1	97,9
		72 days	0,6	96,9
	2H2	20 days	2,2	84,3
		72 days	0,04	99,3
	2B2	20 days	41,2	26,1
		72 days	6,5	79,1
	2D2	20 days	0,1	98,7
		72 days	0,03	99,3
	2D3	20 days	8	73,7
		72 days	0,6	97,3
	2 E2	20 days	23,5	48,3
		72 days	0,02	99,3
	2G1	20 days	0,5	95,8
		72 days	0,6	94,6
	2G2	20 days	3,7	84,2
		72 days	0,9	96,4
	3A8	20 days	8,8	64,5
		72 days	3,2	89,7
	3A9	20 days	37,3	37,6
		72 days	0,1	99,2
	2C3	20 days	0,2	94,7
		72 days	0,8	96
	2A2	20 days	0,02	99,1
		72 days	0	99,2

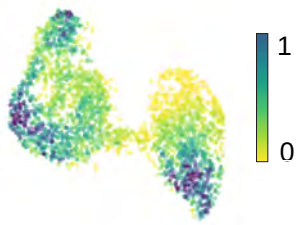
SW480 - single cell clones

SW480 - EpCAMlow			EpCAMlow	EpCAMhigh
success rate = 5 clones/36	A1	25 days	38,5	34
		75 days	15,4	59,6
	D5	25 days	33,8	32,9
		75 days	16,5	71,1
	F1	25 days	75,2	11,5
		75 days	41,7	47,1
	H4	28 days	91,7	1,6
		82 days	95,5	2
	F2	28 days	87,1	2,2
		82 days	79,8	10,2

SW480 - EpCAMhigh			EpCAMlow	EpCAMhi
success rate = 11 clones/36	A11	28 days	2,7	90,2
		82 days	1,5	92,3
	B12	28 days	0,2	96,1
		82 days	0,2	98,1
	B8	28 days	39	31,8
		82 days	26,3	51
	E10	28 days	0,2	97,5
		82 days	3,8	91,6
	F11	28 days	1	92,6
		82 days	2,1	90,8
	F12	28 days	3,2	83,5
		82 days	6,2	83,3
	G11	28 days	8,9	71,1
		82 days	35,8	47,8
	G8	28 days	2,3	90,2
		82 days	1,8	93,3
	H12	82 days	0,1	98,4
	B7	28 days	0,2	98,3
		82 days	0,1	98,6
	E12	28 days	0,1	97,9
		82 days	0,4	97,1

D Markov forward diffusion

Markov backward diffusion



E PAGA velocity graph

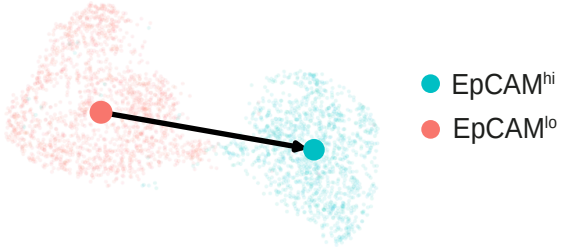


Figure 3

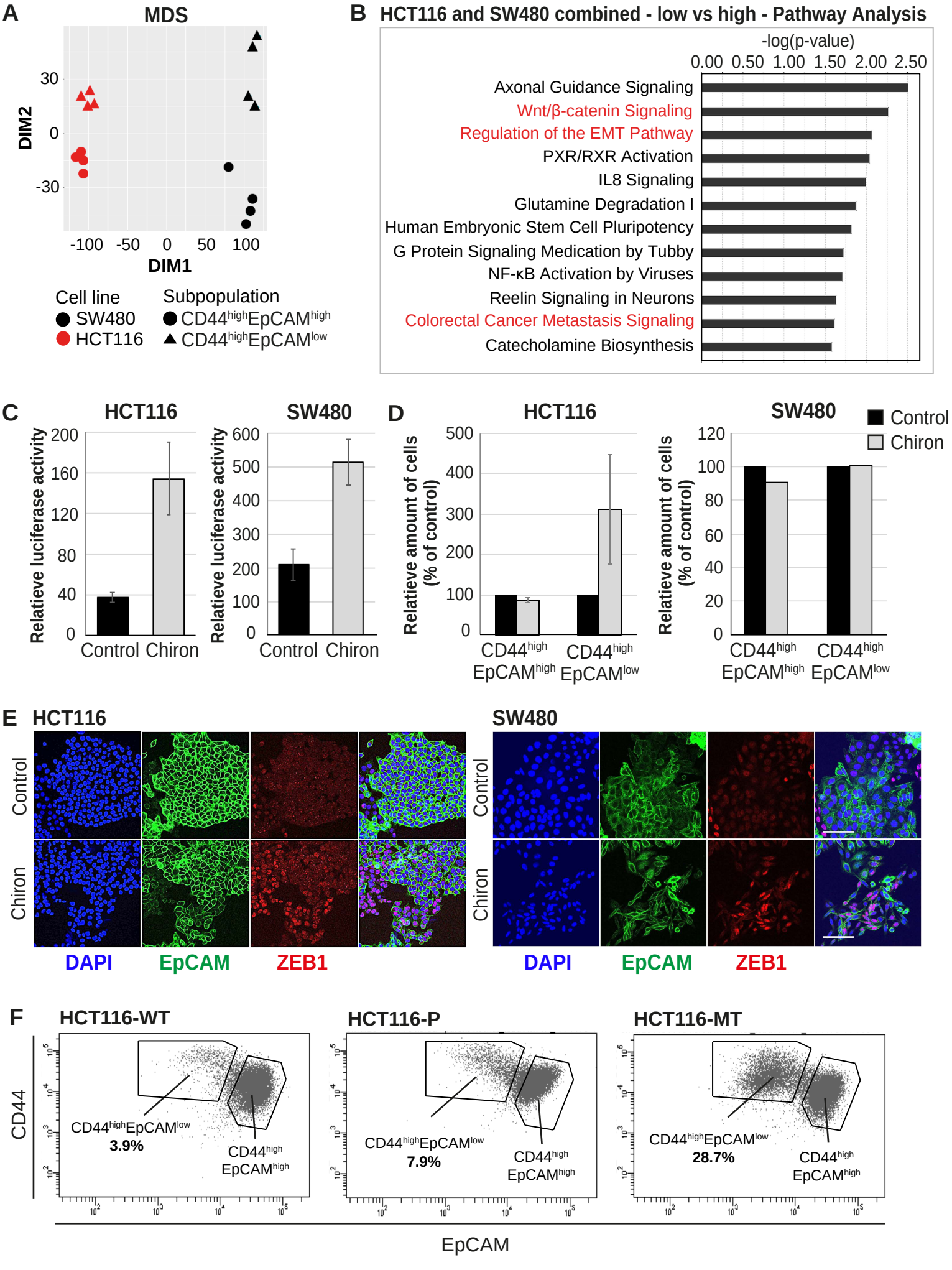

A
MDS
DIM2
30
0
-30
-100 -50 0 50 100
DIM1
Cell line
SW480
HCT116
Subpopulation
CD44highEpCAMhigh
CD44highEpCAMlow
B
HCT116 and SW480 combined - low vs high - Pathway Analysis
-log(p-value)
0.00 0.50 1.00 1.50 2.00 2.50
Axonal Guidance Signaling
Wnt/β-catenin Signaling
Regulation of the EMT Pathway
PXR/RXR Activation
IL8 Signaling
Glutamine Degradation I
Human Embryonic Stem Cell Pluripotency
G Protein Signaling Medication by Tubby
NF-κB Activation by Viruses
Reelin Signaling in Neurons
Colorectal Cancer Metastasis Signaling
Catecholamine Biosynthesis
C
HCT116
SW480
Relatieve luciferase activity
Control Chiron
D
HCT116
SW480
Relatieve amount of cells (% of control)
Control
Chiron
CD44high EpCAMhigh
CD44high EpCAMlow
E
HCT116
SW480
Control
Chiron
DAPI
EpCAM
ZEB1
F
HCT116-WT
HCT116-P
HCT116-MT
CD44
EpCAM
CD44highEpCAMlow
3.9%
7.9%
28.7%
CD44high EpCAMhigh

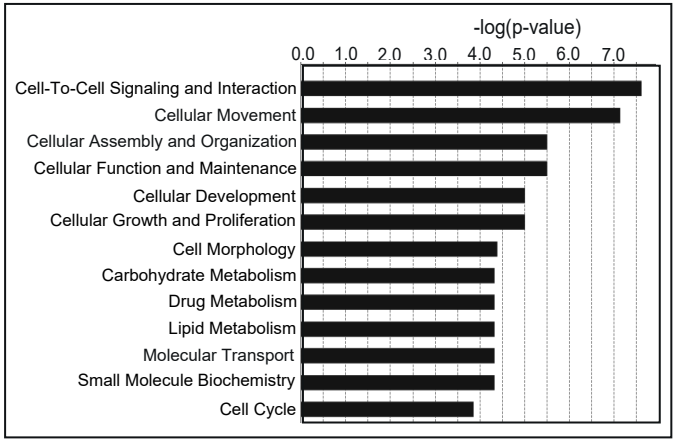
Figure 3 - Figure supplement 1

A

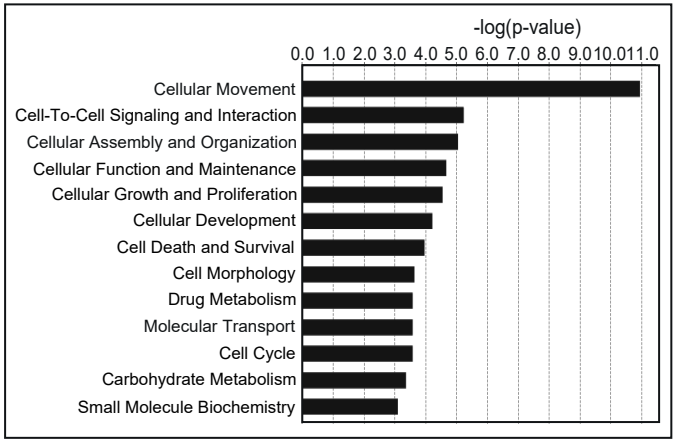
Gene symbol	Descrption
ACOXL	acyl-CoA oxidase-like
AP1M2	adaptor-related protein complex 1
BSPRY	B-box and SPRY domain containing
C1orf210	chromosome 1 open reading frame 210
CCDC64B	coiled-coil domain containing 64B
CDH1	cadherin 1, type 1
CDH3	cadherin 3, type 1,
CLDN7	claudin 7
CNTNAP3	contactin associated protein-like 3B
CNTNAP3B	contactin associated protein-like 3B
DACT1	dishevelled-binding antagonist of beta-catenin 1
EFNA5	ephrin-A5
EPCAM	epithelial cell adhesion molecule
EPHA1	EPH receptor A1
ESRP1	epithelial splicing regulatory protein 1
KDF1	keratinocyte differentiation factor 1
KLK10	kallikrein-related peptidase 10
LAD1	ladinin 1
LAMA3	laminin, alpha 3
LAMC2	laminin, gamma 2
MACC1	metastasis associated in colon cancer 1
MAL2	mal, T-cell differentiation protein 2 (gene/pseudogene)
MAP7	microtubule-associated protein 7
MARVELD3	MARVEL domain containing 3
NTNG2	netrin G2
PMEPA1	prostate transmembrane protein, androgen induced 1
RGS5	regulator of G-protein signaling 5
SPARC	secreted protein acidic and rich in cysteine (osteonectin)
ST14	suppression of tumorigenicity 14
SUSD2	sushi domain containing 2
SYT11	synaptotagmin XI
TMC5	regulator of G-protein signaling 5
TMEM125	transmembrane protein 125
ZEB1	zinc finger E-box binding homeobox 1

B

HCT116 - Molecular and Cellular Functions



SW480 - Molecular and Cellular Functions



HCT116 and SW480 - Molecular and Cellular Functions

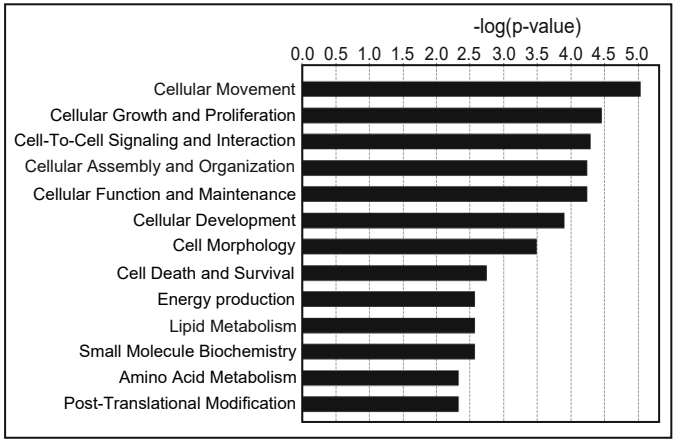


Figure 3 - Figure supplement 2

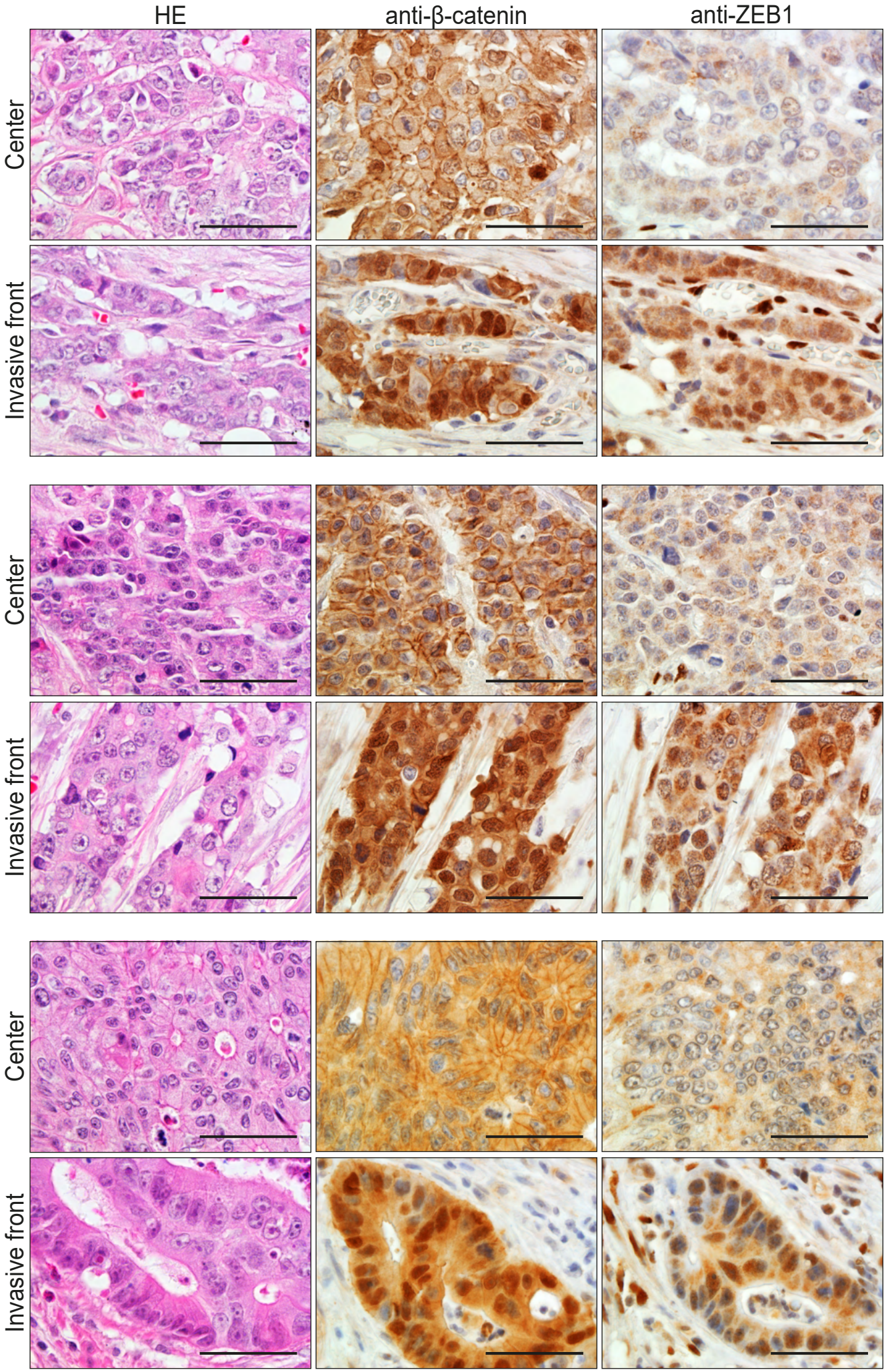


Figure 4

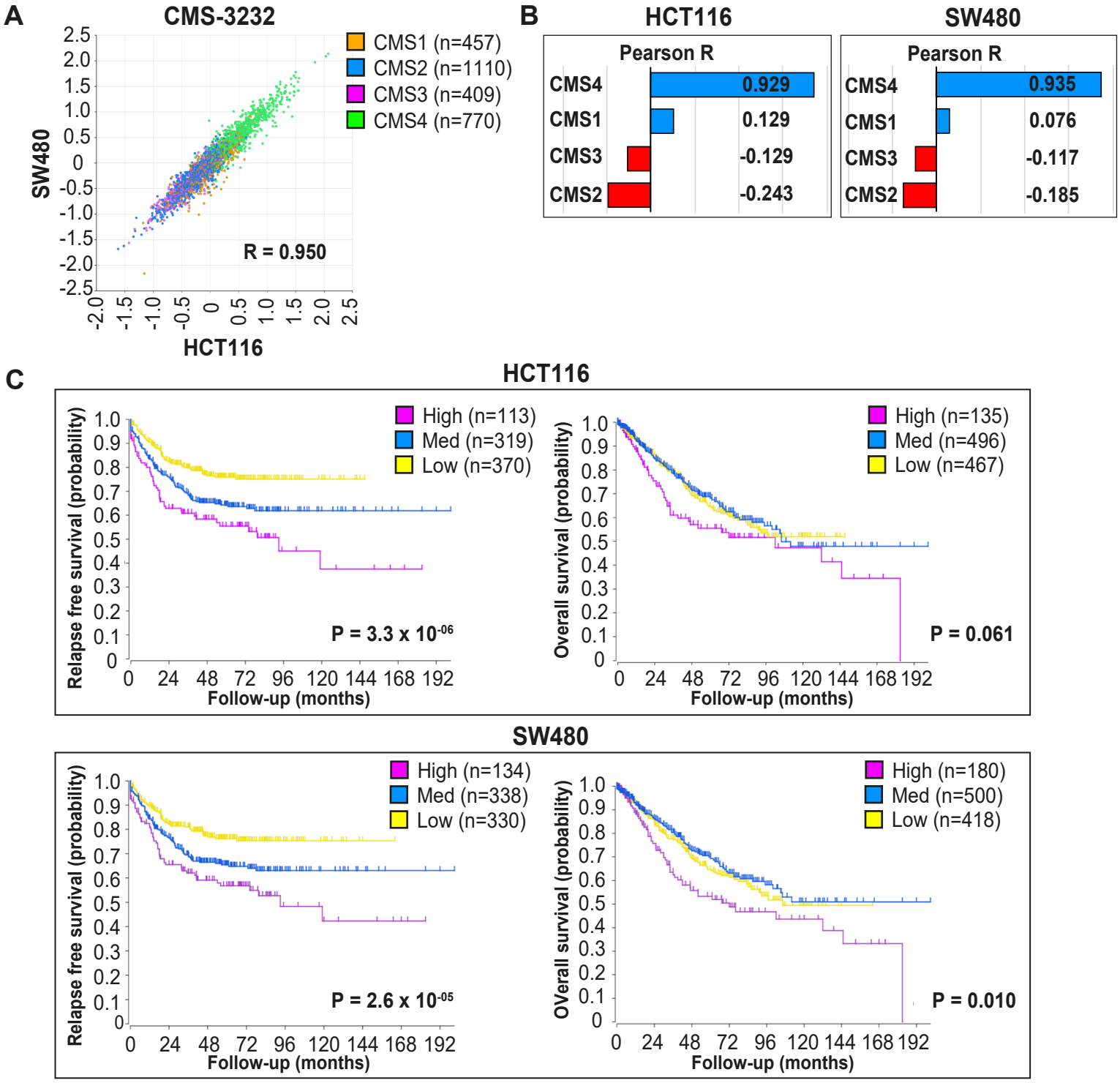

A
CMS-3232
CMS1 (n=457)
CMS2 (n=1110)
CMS3 (n=409)
CMS4 (n=770)
R = 0.950
SW480
HCT116
B
HCT116
Pearson R
CMS4 0.929
CMS1 0.129
CMS3 -0.129
CMS2 -0.243
SW480
Pearson R
CMS4 0.935
CMS1 0.076
CMS3 -0.117
CMS2 -0.185
C
HCT116
High (n=113)
Med (n=319)
Low (n=370)
Relapse free survival (probability)
Follow-up (months)
P = 3.3 x 10-06
High (n=135)
Med (n=496)
Low (n=467)
Overall survival (probability)
Follow-up (months)
P = 0.061
SW480
High (n=134)
Med (n=338)
Low (n=330)
Relapse free survival (probability)
Follow-up (months)
P = 2.6 x 10-05
High (n=180)
Med (n=500)
Low (n=418)
OVerall survival (probability)
Follow-up (months)
P = 0.010

Figure 4 - Figure supplement 1

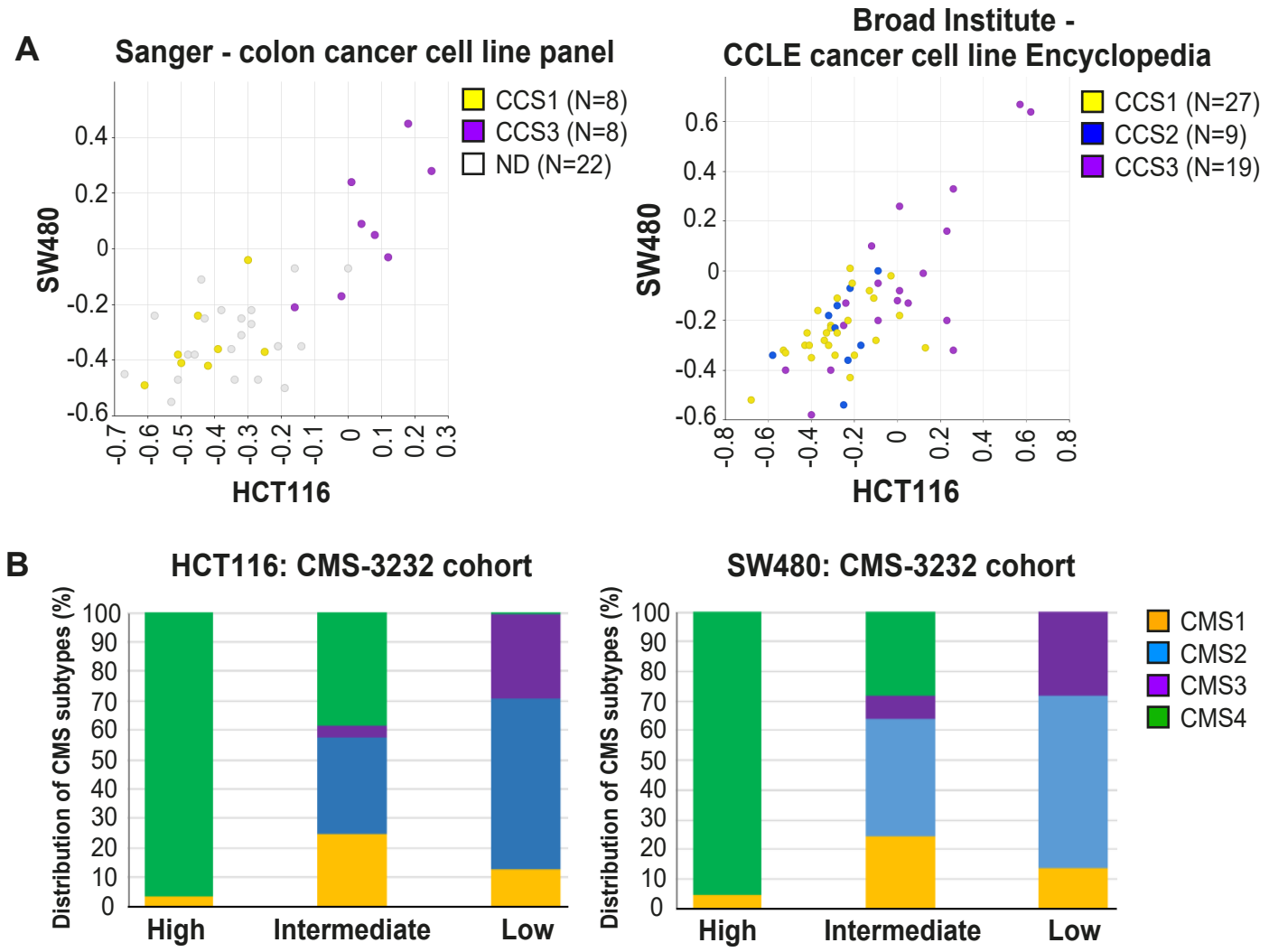


Figure 5

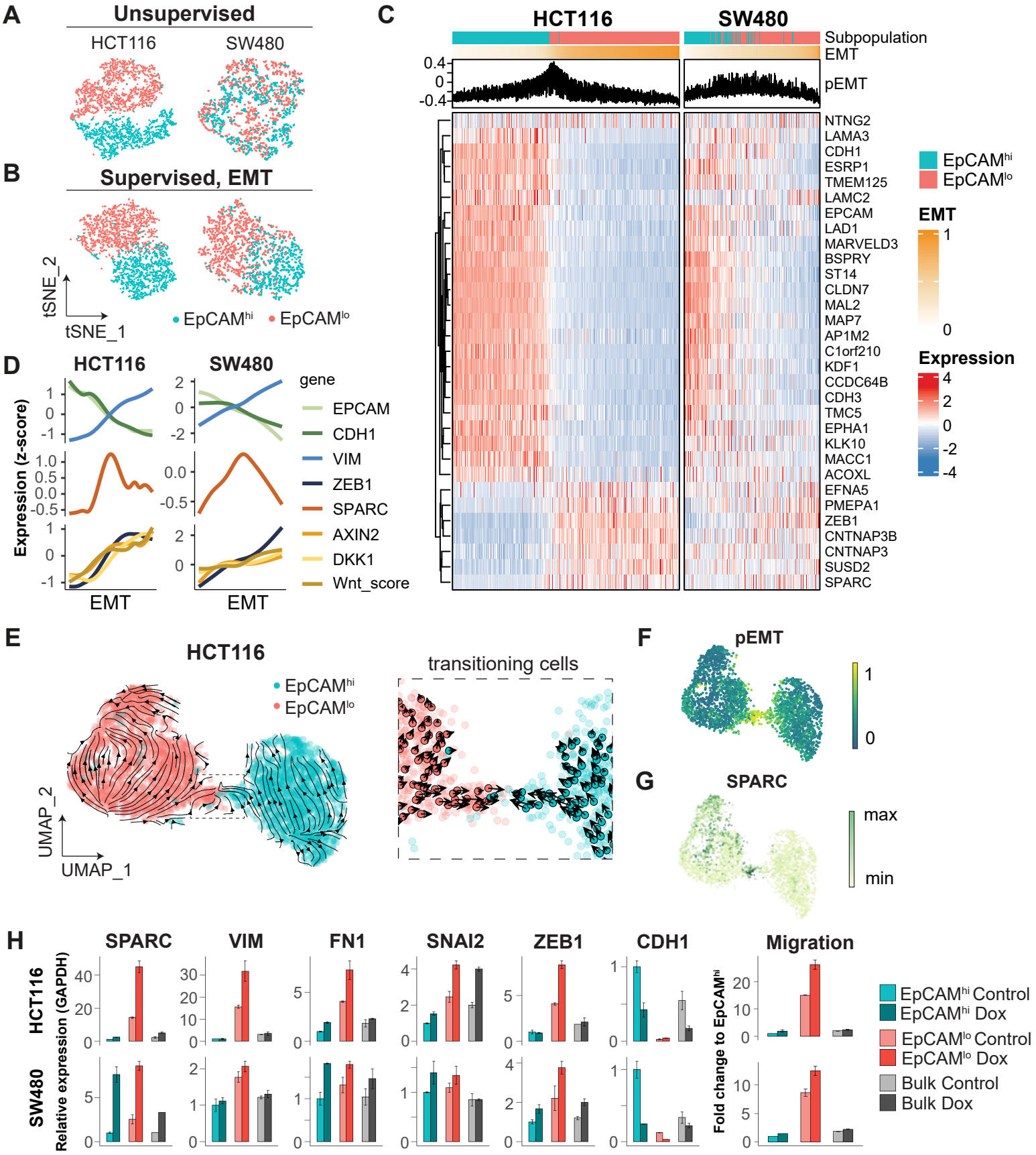

A
Unsupervised
HCT116
SW480
B
Supervised, EMT
tSNE_2
tSNE_1
EpCAMhi
EpCAMlo
C
HCT116
SW480
Subpopulation
EMT
pEMT
0.4
0
-0.4
NTNG2
LAMA3
CDH1
ESRP1
TMEM125
LAMC2
EPCAM
LAD1
MARVELD3
BSPRY
ST14
CLDN7
MAL2
MAP7
AP1M2
C1orf210
KDF1
CCDC64B
CDH3
TMC5
EPHA1
KLK10
MACC1
ACOXL
EFNA5
PMEPA1
ZEB1
CNTNAP3B
CNTNAP3
SUSD2
SPARC
EpCAMhi
EpCAMlo
EMT
1
0
Expression
4
2
0
-2
-4
D
HCT116
SW480
Expression (z-score)
gene
EPCAM
CDH1
VIM
ZEB1
SPARC
AXIN2
DKK1
Wnt_score
EMT
E
HCT116
EpCAMhi
EpCAMlo
UMAP_2
UMAP_1
transitioning cells
F
pEMT
1
0
G
SPARC
max
min
H
SPARC
VIM
FN1
SNAI2
ZEB1
CDH1
Migration
HCT116
SW480
Relative expression (GAPDH)
Fold change to EpCAMhi
EpCAMhi Control
EpCAMhi Dox
EpCAMlo Control
EpCAMlo Dox
Bulk Control
Bulk Dox

Figure 5 - Figure supplement 1

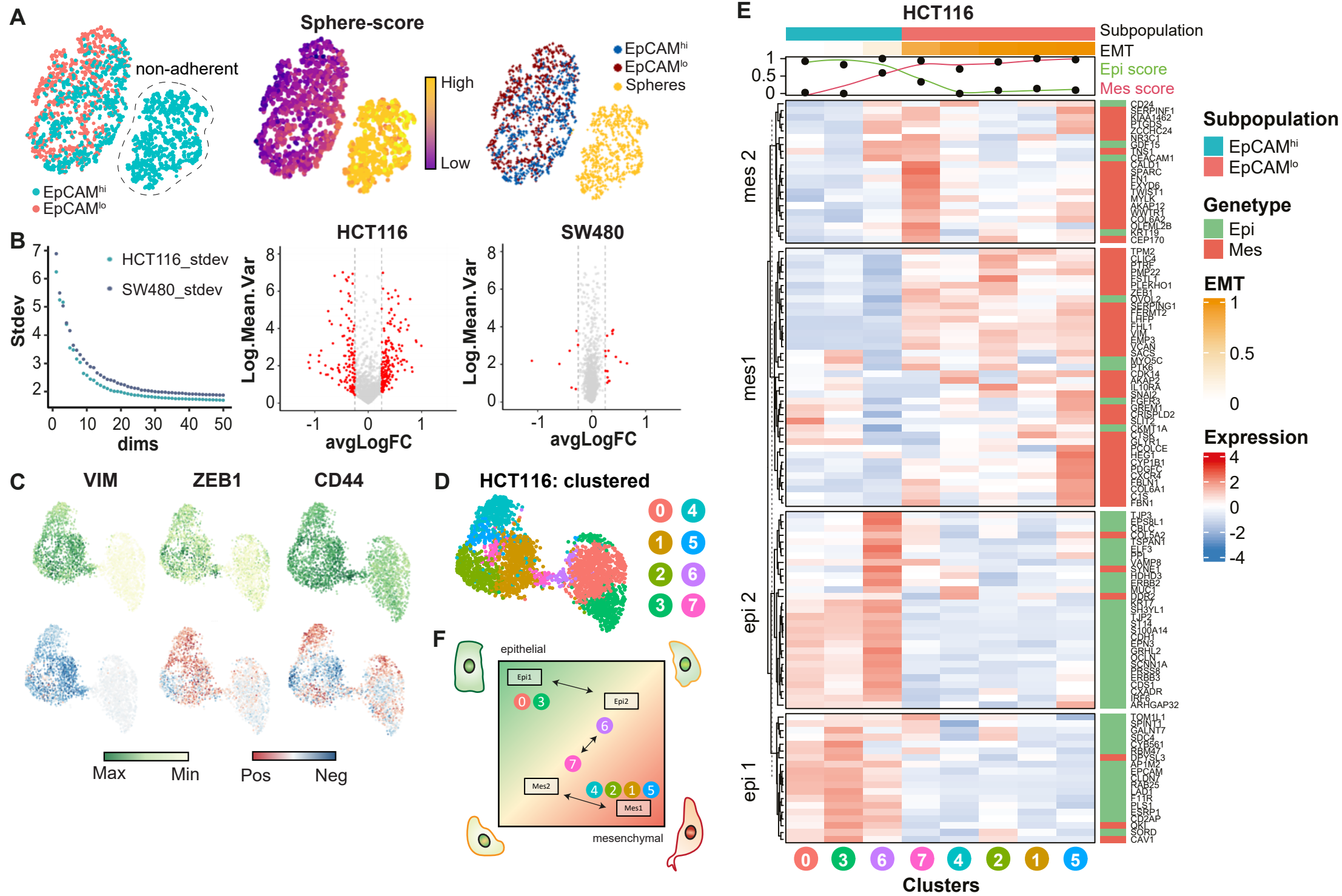
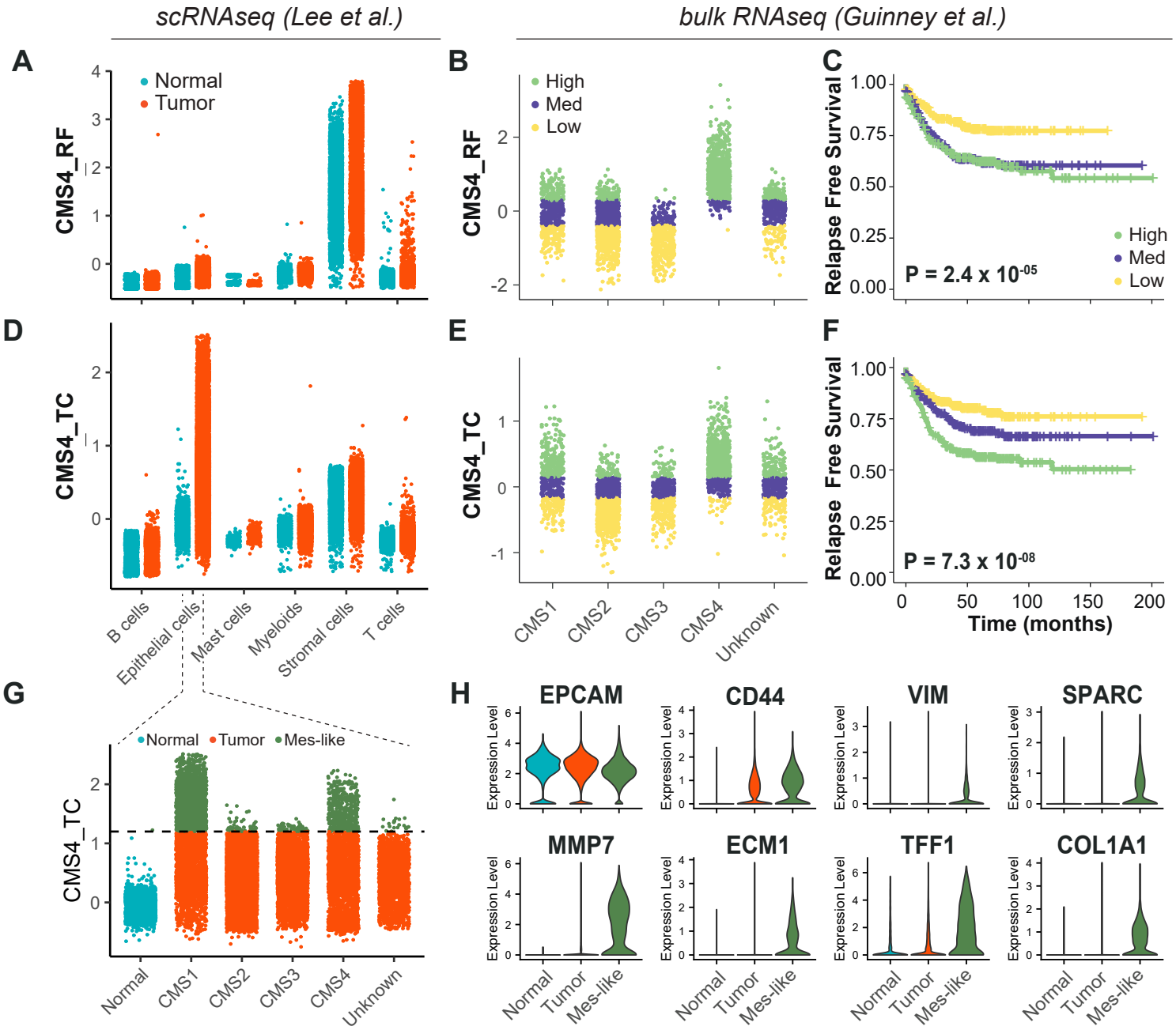


Figure 6


scRNAseq (Lee et al.)
bulk RNAseq (Guinney et al.)
A
Normal
Tumor
CMS4_RF
B
High
Med
Low
CMS4_RF
C
Relapse Free Survival
P = 2.4 x 10-05
High
Med
Low
D
CMS4_TC
B cells
Epithelial cells
Mast cells
Myeloids
Stromal cells
T cells
E
CMS4_TC
CMS1
CMS2
CMS3
CMS4
Unknown
F
Relapse Free Survival
P = 7.3 x 10-08
Time (months)
G
Normal
Tumor
Mes-like
CMS4_TC
Normal
CMS1
CMS2
CMS3
CMS4
Unknown
H
EPCAM
CD44
VIM
SPARC
MMP7
ECM1
TFF1
COL1A1
Expression Level
Normal
Tumor
Mes-like

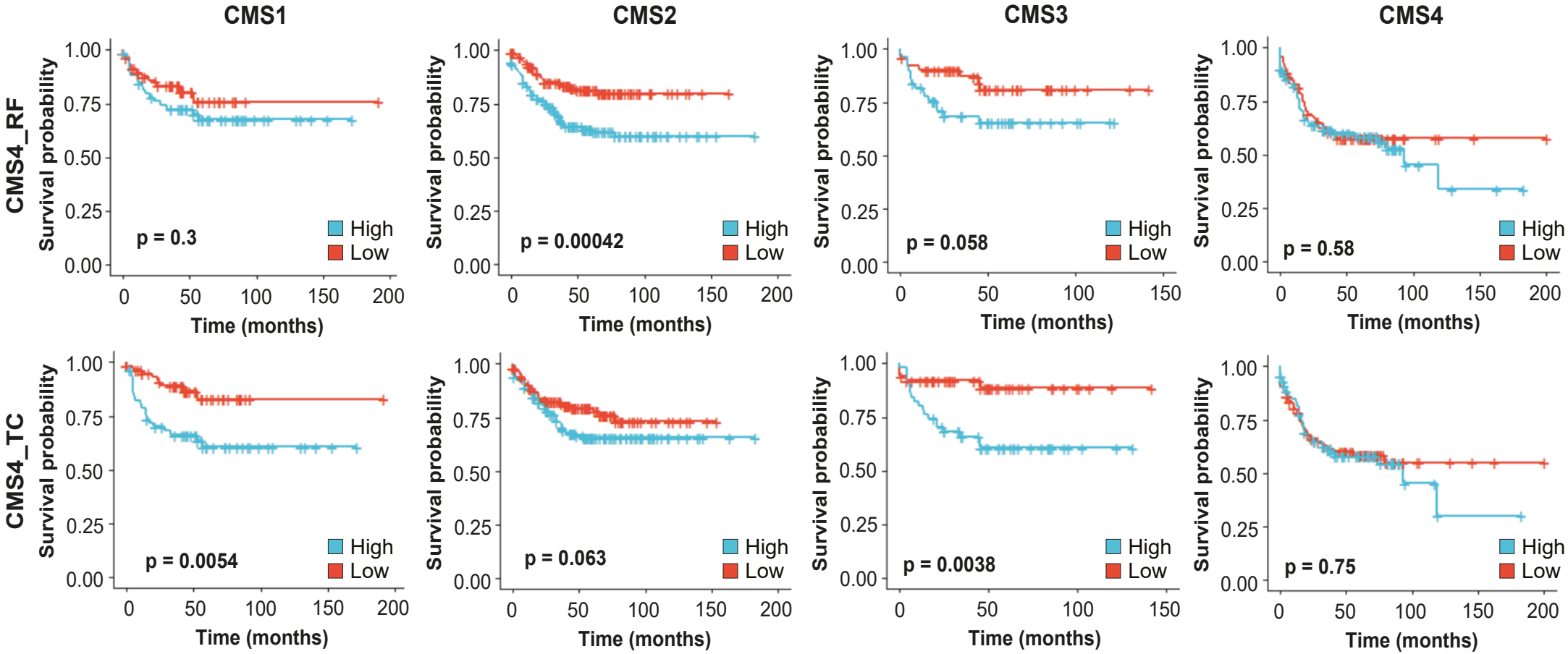



Figure 6 - Figure supplement 1