Phenotypic plasticity and partial EMT underlie local invasion and distant metastasis in colon cancer.

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

[1]Department of Pathology and [4]Cancer Computational Biology Center and Department of Urology, Erasmus MC Cancer Institute, Erasmus University Medical Center, Rotterdam, The Netherlands. [2]Natural Product Research Center, Korea Institute of Science and Technology (KIST), Gangneung, Korea.[3]Department of Surgical Oncology, Cancer Centre, University Medical Centre Utrecht, Utrecht, The Netherlands. [5]Princess Máxima Center for Pediatric Oncology Utrecht, The Netherlands. [6]Cancer Research UK Beatson Institute, Glasgow, UK. [7]Department of Molecular Pathology, Oncode Institute, Netherlands Cancer Institute, Amsterdam, The Netherlands. [8]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 metastasizing carcinoma cell as it bestows it with the capacity of altering its own cellular identity when exposed to different environments along the invasion-metastasis cascade. However, the study of the mechanisms underlying phenotypic plasticity is hindered by the rarity of these events within primary lesions and by the lack of experimental study models. Here, we identified a subpopulation of colon cancer cells endowed with phenotypic plasticity: $CD44^{hi}$/$EpCAM^{lo}$ cells are highly motile, invasive, chemoresistant, and metastatic both *in vitro* and *in vivo*. Bulk and single-cell RNAseq analysis indicated that they encompass a broad spectrum of degrees of EMT activation, together with stem/progenitor-like feature, enhanced Wnt/β-catenin signaling, and a high correlation with the CMS4 subtype, accounting for ~25% of colon cancer cases with poor prognosis. Enhanced Wnt and the downstream upregulation of EMT transcription factors such as ZEB1 represent key events in eliciting phenotypic plasticity to cancer cells along the invasive front of primary colon carcinomas. Of note, different combinations of distinct sets of epithelial and mesenchymal genes define alternative transcriptional trajectories through which state transitions arise. Hybrid E/M stages are predicted to represent the origin of these differentiation routes through biologically-distinct cellular states and as such 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 phenotypic plasticity, possibly the most clinically relevant hallmark of cancer[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 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 features[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: *SNAI1*/Snail, *SNAI2*/Slug, *ZEB1*, *ZEB2*, *TWIST1*, and *TWIST2*) underlies the transient nature of these cellular modifications[8].

This "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"[9]. 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. However, 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[7,9]. However, the identity of these stromal factors and 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[10]. 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[11]. 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, by taking advantage of conventional colon cancer cell lines and of organoid-based *in vivo* models, we identified and extensively characterized EMT-competent and metastable mCSCs that underlie local invasion and distant metastasis, and whose expression signature is predictive of reduced disease-free survival among colon cancer patients.

Results

Identification and characterization of CD44^{high}EpCAM^{low} colon cancer cells.

To assess whether commonly employed colon cancer cell lines encompass distinct differentiated and more stem-like and/or EMT-positive 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 **Figure 1A** for HCT116 and SW480, the CD44/EpCAM combination best highlighted the presence of different cellular phenotypes with, as shown here below, distinct epithelial and more mesenchymal cellular morphologies. Notably, both HCT116 and SW480 revealed a continuum of different EpCAM and CD44 expression levels with a large CD44^{high}EpCAM^{high} cluster followed by a tail of gradually decreasing EpCAM and increasing CD44 levels. By applying specific gates, variable percentages of CD44^{high}EpCAM^{low} cells were found to feature the HCT116 (5.0±2.5%) and SW480 (16.7±13%) cell lines, respectively. The same CD44^{high}EpCAM^{low} cells were identified at similarly variable percentages in other commonly employed colon cancer cell lines (**Figure 1 – supplement 1**). Notably, within the same cell line, these percentages vary depending on the passage number and culture conditions. Here, we mainly focused on the HCT116 and SW480 lines as they are representative of the two main colon cancer subtypes, namely earmarked by MIN (microsatellite instability) and CIN (chromosomal instability; also known as MSS, microsatellite stable), respectively[12]. 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 further characterize the EpCAM^{lo} and EpCAM^{hi} colon cancer cells, they were sorted by FACS and cultured under normal conditions. While EpCAM^{hi} cells formed compact colonies with characteristic epithelial morphology, EpCAM^{lo} cells showed a more spindle- and mesenchymal-like appearance (**Figure 1B**). To investigate whether these morphological differences were reflected by more functional features, sorted EpCAM^{lo} and EpCAM^{hi} cells were analyzed by Transwell assay: a significantly increased migratory and invasive ability was observed in the EpCAM^{lo} fractions from both colon cancer cell lines when compared with their EpCAM^{hi} counterparts (**Figure 1C-D**). On the other hand, proliferation assays and cell cycle analysis (**Figure 1 – supplement 2A-B**) revealed that EpCAM^{lo} cells are less proliferative and enter the cell cycle at a slower rate, when compared with their more epithelial counterpart.

Hence, distinct EpCAM expression levels and the observed morphological and functional differences suggest the presence of distinct cellular identities within established colon cancer cell lines.

ZEB1 expression underlies EMT and phenotypic plasticity in EpCAM^{lo} colon cancer cells.

In agreement with the FACS results, *EpCAM* gene expression levels, as measured by RT-qPCR were lower in EpCAM^{lo} cells when compared with their epithelial counterparts in both cancer cell lines (**Figure 2A**). Accordingly, expression analysis of other EMT-associated marker genes and transcription factors (**Figure 2A**) showed significantly reduced mRNA levels of E-cadherin (*CDH1*), and increased expression of vimentin (*VIM*) in EpCAM^{lo} cells. Moreover, expression of the EMT-transcription factor *ZEB1* was significantly upregulated in EpCAM^{lo} cells from both lines, as also confirmed by immunofluorescence analysis (**Figure 2B-C**). Expression of other EMT-TFs such as *SLUG* (both cell lines) and *FOXC2* (HCT116) appeared to be increased in EpCAM^{lo} cells, although not at statistically significant levels (**Figure 2B)**.

To assess the role of *ZEB1* in the establishment and maintenance of the EpCAM^{lo} subpopulation, we employed dox-inducible shRNA lentiviral vectors to decrease its mRNA expression levels in HCT116 and SW480 (**Figure 2D**, by approx. 60% as measured by qPCR 72 hours after induction). Notably, FACS analysis revealed that *ZEB1* knockdown dramatically decreased the percentage of EpCAM^{lo} cells in both lines (**Figure 2E-F**).

Transcriptional regulation of *ZEB1* (and *ZEB2*) has been shown to be regulated by the miR-200 superfamily of micro RNAs (miR-200a, miR-200b, miR-200c, miR-141 and miR-429) which bind to specific 3'UTR sequences leading to EMT inhibition[13]. 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 (**Figure 2G**).

In view of their mesenchymal morphology, the increased motile and invasive features, and the EMT-related gene and miRNA expression patterns, EpCAM^{lo} colon cancer cells are likely to represent a plastic subpopulation of colon cancer cells likely to underlie local invasion and distant metastasis. Of note, although EMT has been shown to be functionally associated with stemness and resistance to cytotoxic therapy in breast cancer cells[6], apart from CD44 no other known previously employed cancer stem cell marker (i.e. CD133, CD24, and ALDEFLUOR) was found to be expressed in EpCAM^{lo} cells (data not shown).

To further investigate the plasticity of EpCAM^{lo} colon cancer cells, we assessed their capacity to differentiate into more epithelial cell types 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 **Figure 3A**, 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, *ECAD* 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). Likewise, *VIM* and *ZEB1* expression was decreased in 'late' vs. 'early' EpCAM^{lo} cells (**Figure 3B**). In addition, the migratory capacity of 'late' EpCAM^{lo} HCT116 cells was reduced to levels comparable with those of 'early' EpCAM^{hi} cells (**Figure 3C**).

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 **Figure 3D and supplement 1**, 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. clones 2F12 and 2E10, **Figure 3D**). A minority of the cells however, appears to have lost this plasticity and retains, even after extended culture, the EpCAM^{lo} phenotype (e.g. clone 2H12). In contrast, the majority of EpCAM^{hi} single cells retained their epithelial features with <1% switching to the EpCAM^{lo} state (**Figure 3D and supplement 1**).

These results highlight the metastable and cell-autonomous nature 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.

EpCAM^{lo} colon cancer cells are resistant to chemotherapeutic drugs

In view of the well-established correlation between EMT and therapy resistance[14], EpCAM^{lo} cells were cultured in the presence of two cytotoxic drugs commonly employed in the treatment of colon cancer, namely oxaliplatin and 5-fluorouracil (5-FU) and their viability compared with that of EpCAM^{hi} and bulk cells. To this aim, two consecutive assays were implemented, namely a metabolic activity assay (MTT) followed by a re-growth assay to measure the capacity of the treated cells to relapse upon withdrawal of the cytotoxic agents. EpCAM^{lo} cells

showed increased viability at all tested oxaliplatin (**Figure 4A** and **C**, left panels) and 5-FU (**Figure 4B** and **D**, left panels) concentrations. Likewise, the re-growth assay revealed that the chemo-resistant 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 the more sensitive EpCAM^{hi} counterparts (**Figure 4**, right panels).

Overall, the data indicate that two among the most commonly employed chemotherapeutic drugs in the treatment of colon cancer preferentially affect epithelial cells while the EMT-competent and stem-like EpCAM^{lo} cells are resistant.

EpCAM^{lo} colon cancer cells are highly metastatic in vivo.

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. After 4-8 weeks, mice were sacrificed and analyzed for liver metastases. EpCAM^{lo} cells from both lines resulted in significantly more liver metastases than with EpCAM^{hi} and bulk cells (**Figure 5A-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[9,15] (**Figure 5E**). FACS analysis of these liver metastases revealed predominant epithelial features with only a minority of EpCAM^{lo} cells, thus highlighting their striking plasticity and the key role played by MET in metastasis formation **(Figure 5 supplement 1).**

In vivo Zeb1 activation enlarges the EpCAM^{lo} subpopulation in primary colon cancers leading to lung and liver metastases

To validate the above observations 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 loss- and gain-of-function mutations at the *Apc*, *Kras*, and *Tp53* genes (*Apc$^{fl/fl}$*::*Kras$^{G12D/+}$*::*Trp53$^{fl/R172H}$*; AKP)[16]. Orthotopic transplantation of AKP organoids results in the establishment of a primary carcinoma in the caecum and subsequent metastasis development at distant organ sites, mainly liver and lungs[17]. We further modified the AKP

organoids by tagging them with GFP and a click beetle luciferase[18], and by making them doxycycline-inducible for the expression of mouse *Zeb1* (AKP-Z) (**Figure 6A**). 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 6B**).

Next, AKP-transplanted mice were continuously administered doxycycline in the drinking water during 8 weeks after transplantation and the incidence of lung and liver metastases evaluated in comparison with control (no dox) animals by taking advantage of the luciferase tag and bioluminescence imaging. As shown in **Figure 6C**, an overall higher photon/sec count was observed in AKP-Z induced mice compared to control animals. In order to obtain a comprehensive evaluation of the multiplicity of metastases, whole livers and lungs were resected and processed for β-catenin IHC analysis in 500 micron sections to facilitate tumor cells visualization by digital slide scanner (NanoZoomer Digital Pathology; NDP). Multiplicity of metastatic lesions was assessed by NDP view software. The total tissue area evaluated for the presence of metastatic lesions was marked using the NDP software and employed for data normalization. The results show an overall albeit not statistically significant increase in lung and liver metastases upon *Zeb1*-induction (**Figure 6D-E**). Combined analysis of lung and liver metastases shows that in the majority of the dox-treated animals both organs were affected by metastases (**Figure 6F**).

Overall, these results indicate that the *Zeb1* expression expands *in vivo* the subpopulation of EpCAM^{lo} cells in the AKP-derived primary tumors, thereby contributing to increased dissemination along the invasion-metastasis cascade. The admittedly marginal increase in lung and liver metastases is likely to result from the continuous induction of *Zeb1* expression and the consequent inhibition of mesenchymal-to-epithelial transitions essential for metastasis formation[7].

Gene expression profiling of EpCAM^{lo} colon cancer cells reveals underlying EMT, Wnt, and stemness pathways

In order to elucidate the mechanisms underlying plasticity, EMT, and resistance to therapy in EpCAM^{lo} colon cancer cells, RNAseq analysis was performed on the sorted subpopulations from the HCT116 and SW480 cell lines. Multidimensional scaling (MDS) showed a separation in the second dimension of EpCAM^{hi} and EpCAM^{lo} cells in

both cell lines (**Figure 7A**). 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 genes (n=34) were common to both lines (**Figure 7 – supplement 1A**). However, Ingenuity Pathway Analysis (IPA) revealed that the genes differentially expressed between EpCAM^{lo} and EpCAM^{hi} cells from both lines reflected common molecular and cellular functions including '*cellular movement*' (both cell lines: *VIM, ZEB1, SPARC, EPCAM* and *ESRP1*; HCT116: *LOXL2, ANK1, EGF*; SW480: *ALCAM, CXCL14, TGFB2*), '*cellular assembly and organization*' (HCT116: *LAMA3*, *ADAMTS2*, *OLFML2A*, *COL3A1*, *SERPINA1 and SERPINA5*; SW480: *SERPINB5, GREM1, COL1A2, ADAMTS8, ADAMTS10*), and '*drug metabolism*' (HCT116: *ABCA1*, *ABCG1* and *HAS2*; SW480: *ABCB1* and *ABCG2*)(**Figure 7 – supplement 1B)**. Likewise, pathway analysis of the combined EpCAM^{lo} expression profiles from HCT116 and SW480 revealed highly significant association with '*EMT regulation*' (*CDH1*, *ZEB1*, *TGFB2*, *WNT7A*, WNT10A, ESRP1), '*Wnt/β-catenin signaling*' (*CDH1, CDH2, TGB2, WNT10A, WNT7a*), '*human embryonic stem cell pluripotency*' (*NTRK2, TGFB2, WNT10A, WNT7A)*, '*IL8 signaling*' (*CR2, FLT1, GNG2, ITGB3*), and '*colorectal cancer metastasis*' (*CDH1, GNG2, TGFB2, WNT10A, WNT7A*) (**Figure 7B**).

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[19] and HCT116[20], respectively) predicted to underlie constitutive Wnt signaling activation. Nonetheless, Wnt signaling appears to be differentially activated 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[21,22], 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. To this aim, both cell lines were treated with CHIR99021 (Chiron), a potent and selective inhibitor of glycogen synthase 3β (GSK3β) leading to robust Wnt signaling activation. As shown by TopFLASH reporter assay and by FACS analysis of the treated cell lines, enhancement of Wnt signaling by Chiron led to an increase of the EpCAM^{lo} population by approximately 3-fold in the HCT116 cell line (**Figure 7C-D**). However, no significant increase was observed upon Chiron treatment in SW480 (**Figure 7D**). Nonetheless, 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[23,24] (**Figure 7E**).

To further validate the role of Wnt-signaling, 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 Ser45 del mutant allele[25]. 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 (**Figure 7F**). 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)[25] .

These results highlight the key role played by Wnt signaling in the establishment and maintenance of EpCAM^{lo} colon cancer cells through *ZEB1* upregulation and EMT induction.

EpCAM^{lo} cells localize at the invasion front of colon carcinomas and are associated with the CMS4 group of patients with shorter disease-free and overall survival.

The above data support the mCSC[7] and β-catenin paradox[9] models and provide novel insights in the cellular and molecular mechanisms underlying Wnt super-activation, EMT induction, and cell dissemination along the invasive front of colon carcinomas. 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[7,9]. 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 8**, co-localization of nuclear β-catenin and ZEB1 expression was found in 5 out of 25 cases investigated.

Distinct recurrent gene expression patterns underlie the recently proposed classification of human colon cancers in four Consensus Molecular Subtypes (CMS1-4)[26]. 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 mesenchymal state. Accordingly,

the CMS4 subtype was identified in tumor-derived organoids and cell lines, suggesting that CMS4 is an intrinsic property of the quasi-mesenchymal colon cancer cell[27]. 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[26] encompassing expression data relative to 3232 human colon cancers classified as CMS1-4. Expression of the HCT116 and SW480 signatures were highly correlated with each other (**Figure 9A**), with the CMS4 signature genes (n=143) (**Figure 9B**), and with the expression signature of colon cancer cell lines previously classified as mesenchymal-like (CCS3)[28] (**Figure 9 – supplement 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 (**Figure 9 – supplement 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 (**Figure 9C**).

In view of these results, we selected additional cell lines, previously classified according to the consensus molecular subtypes[29], to determine the relative abundance of the EpCAM^{lo} subpopulation (**Figure 1 - supplement 1**). Among these, three cell lines previously classified as CMS1 (RKO[30]) and CMS4 (NCI-H716[31] and COLO-320-HSR[32]) were found to encompass a majority of the EpCAM^{lo} cells. Of note, both NCI-H716 and COLO-320-HSR grow in suspension and have clear mesenchymal morphology. Whereas COLO-320-HSR was derived from a Dukes C carcinoma of the sigmoid[32], NCI-H716 was obtained from the ascites of a patient with poorly differentiated colon adenocarcinoma with peritoneal metastases[31]. As for the RKO cell line, it was derived from a microsatellite unstable (MSI-H) and undifferentiated colon cancer[30]. All three lines are known for their high metastatic potential, and for the downregulation of the *EPCAM* and *CDH1* epithelial markers, and increased expression of *ZEB1* and other EMT-TFs (ZEB2, *SNAI1*; data not shown).

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) analysis 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). Data were processed using the CellRanger Pipeline and imported in R for downstream analysis. First, the 'FindVariableGenes' function in Seurat[33] was performed to detect variably expressed genes across EpCAM^{lo} cells compared with the analogous genes across EpCAM^{hi} cells in both HCT116 and SW480. Subsequent differential expression analysis between the two subpopulations revealed EMT- and Wnt-related genes previously identified in the bulk RNA-seq analysis, as well as novel genes (not shown).

Next, to investigate the heterogeneity within the EpCAM$^{lo/hi}$ subpopulations, dimension reduction using the first 30 principal components was performed. After computation of tSNE (Barnes-Hut implementation), the EpCAM^{lo} and EpCAM^{hi} cells clustered in separate groups in HCT116 (**Figure 10A**), whereas the SW480 cells showed a partial overlap between the two subpopulations and a distinct EpCAM^{hi} cluster (**Figure 10 – supplement 1A**). Using a signature derived from our bulk RNAseq experiments (not shown), this was identified as the non-adherent subpopulation within the SW480 cell line ("spheres"). Nonetheless, substantial overlap between the EpCAM$^{hi/lo}$ subpopulations was retained upon subsequent exclusion of the non-adherent cells (**Figure 10B**). Unsupervised clustering using shared neighbor (SSN) modularity optimization revealed additional layers of heterogeneity as shown by the presence of distinct subclusters (n=8) within the HCT116 line: 3 of EpCAM^{hi} origin, and 5 EpCAM^{lo} (**Figure 10C**). In contrast, SW480 revealed several mixed clusters (n=7) with varying contributions from the two subpopulations (**Figure 10D**). Next, the dimension reduction was repeated in supervised fashion by taking advantage of the public available EMT gene list from the nCounter® PanCancer Progression Panel (Nanostring; n =180); out of these, 107 were found to overlap with our scRNAseq data. Using this 'EMT signature', the two subpopulations were clearly resolved both in HCT116 and SW480 (**Figure 10E-F**).

Next, we focused on distinct EMT stages in HCT116. To this aim, cells were mapped in the 2D EMT landscape by assigning individual scores for their epithelial and mesenchymal state. The EMT signature genes were subdivided into an epithelial (E; n=51) and mesenchymal (M; n=56) group based on the provided annotations (Nanostring). To improve on the zero values of some of the EMT input genes, these were first processed by Markov affinity-based

graph imputation of cells (MAGIC)[34]. The EMT signature genes were then scaled between zero and one to avoid that only few contribute to most of the score variance. Finally, an Epi- and Mes-score was computed by taking the average of the processed epithelial and mesenchymal lists (**Figure 10G**). Projection of the 8 unsupervised clusters on the E/M scores showed their coherent mapping on the 2D EMT landscape: the three main EpCAM^{hi} clusters (#0,4,6) map as "E", two EpCAM^{lo} clusters as "M" (#1,3), whereas an additional three EpCAM^{lo} clusters were intermediate and as such classified as partial EMT or hybrid E/M (#2,7, and, though to a lesser extent, #5; **Figure 10H**).

To further investigate the gene expression patterns underlying the different EMT states, we computed a heat-map where the EMT signature genes are grouped according to their average expression in the unsupervised clusters. This analysis revealed four sets of genes (mes1/2, and epi1/2) expressed in different combinations (**Figure 10 – supplement 1B**). Notably, *ZEB1* is expressed in all EpCAM^{lo} clusters, while other EMT-TFs such as *SNAI2* and *TWIST1* show expression in distinct EpCAM^{lo} clusters. Of note, the partial EMT clusters #5 and #7 showed distinct expression of these gene sets indicating alternative trajectories through which partial EMT cells can arise (**Figure 10 – supplement 1C**).

In view of the above results, distinct trajectories are likely to underlie the observed transcriptional heterogeneity. In an attempt to identify the source of this heterogeneity, a subsequent analysis was performed using Palantir[35]. Cells were projected on the Palantir diffusion map, which is particularly suited to capture dynamic trajectories. We projected the clusters on the diffusion map to evaluate the root clusters, likely to initiate a differentiation trajectory. As shown in **Figure 10I**, the partial EMT clusters (#2 and #7) map to the edges of the diffusion map, and are as such likely to initiate two distinct trajectories that proceed across different clusters of EpCAM^{lo} and EpCAM^{hi} cells.

Last, a high correlation was observed between the partial EMT clusters (#2,5,7) and the CMS4 subtype (**Figure 10 - supplement 10D**).

Altogether, these results reveal the existence of a considerable heterogeneity within the EpCAM^{lo} subpopulation of colon cancer cells. While EpCAM^{lo} cells generally associate with the mesenchymal phenotype, scRNAseq analysis portrays different stages of EMT, including partial EMT. These hybrid E/M stages features distinct transcriptional

signatures and are likely to underlie the phenotypic plasticity of colon cancer cells in view of their predicted role as origin of differentiation trajectories that proceed over biologically-distinct cellular states.

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[36]. 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 acquire stem-like features and as such 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[9]. As different levels of Wnt/β-catenin signaling are associated with distinct cellular outcomes[37], 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 with their more committed counterpart. Similar observations have been previously made in breast cancer and oral squamous cell carcinoma cell lines[10,11] which underlines their relevance for the study of phenotypic plasticity in a broad spectrum of malignancies.

Cellular and molecular analysis of the two colon cancer cell subpopulations revealed the increased invasive and metastatic capacity of EpCAM^{lo} cells and their distinct gene expression profiles when compared to their epithelial counterpart. In particular, EMT activation through *ZEB1* upregulation and enhanced Wnt signaling earmarks the genome-wide transcriptome of EpCAM^{lo} cells. Of note, *ZEB1* also plays unsuspected pleiotropic roles ranging from the modulation of oncogenic and oncosuppressive pathways, cell-fate determination, stemness, and cell plasticity[38] and is as such likely to play a central role in colon cancer invasion and metastasis. From this perspective, the recent debate on EMT as an essential requirement for metastasis[39] likely mirrors the complexity of transcription factors and their downstream targets involved in these processes. Nonetheless, the present data reinforce 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[19,20]. Further enhancement of Wnt signaling is likely to result in EMT activation through *ZEB1* upregulation and its conversion from a repressor to activator state[23]. On its turn, increased *ZEB1* expression was also shown to enhance Wnt, possibly in an autocrine stimulatory circuit[24]. 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[25] 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*[13,22]. 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[40] in the absence of Wnt-activating mutations[41]. As the epigenetic silencing and activation of the miR-200 locus is dynamically regulated by common EMT-inducers such as TGFβ and Wnt[22,40], 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[42]. From this perspective, a recent and elegant study by Morral et al.[43] has showed that a limited subset of colon cancer cells residing in defined niches adjacent to the stroma features high rDNA transcription and protein synthesis and that this tumor tissue architecture defines a stem cell hierarchy. Although similar differences in biosynthetic activities are likely to be less pronounced in immortalized cancer cell lines, GO and PA analysis of the combined EpCAM^{lo} signatures have revealed increased metabolic activities (e.g. amino acid, lipid, and carbohydrate metabolism) suggestive of increased protein synthesis.

Further validation of our results came from the bioinformatics mining of EpCAM^{lo} expression signatures from the HCT116 and SW480 cell lines. Although the two lists only show a very limited number of common genes (n=34 out of 152 and 353 differentially regulated genes in HCT116 and SW480, respectively (n=7 between the top 100 lists, among which *SUSD2* and *ZEB1*), pathway and GO analysis clearly show a common biological basis. Similar results are obtained when comparing the SW480 and HCT116 top-100 lists with the 342 genes differentially expressed between mesenchymal and epithelial cancer cell lines from the Broad Institute cell line collection[27]: 4 common genes with the SW480 list (among which *SUSD2* and *ZEB1*), and 5 with HCT116 (including *SUSD2*, *VIM*, and *ZEB1*). The upregulation of the *SUSD2* (sushi domain containing 2) gene is of interest in view of its capacity to promote

EMT through regulation of EpCAM expression and to correlate with poor overall survival, lymph nodes metastasis, and cisplatin resistance in high-grade serous ovarian cancer[44].

In confirmation of the functional overlap between the HCT116 and SW480 gene lists, both are highly correlated with the CMS4 signature[26] 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[45], several other studies including the present one have clearly demonstrated that cancer cells themselves can express mesenchymal genes[27].

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[46-49]. These hybrid E/M or partial EMT cells are thought to be highly metastable and as such endowed with increased invasive and metastatic capacity. Accordingly, our scRNAseq analysis of EpCAM^{lo} colon cancer cells has revealed distinct both fully M and 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 in specific combinations accompanies the transition between cellular states along distinct trajectories. Future studies will be needed to assess the functional and prognostic relevance of individual genes from within the epi- and mes-arrays.

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 M to E transitions 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.

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[50,51]. 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, Biolegend), 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[52]. 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[17]. 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)[53]. 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)[54] and the Homo sapiens GENCODE v23 annotation[55]. Raw counts were summed with the summarize overlaps function with union mode from the Bioconductor Genomic Alignments package[56] (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[57] was used to perform these calculations. We applied a Wald-test to identify statistical significant differently expressed genes. *P*-values were adjusted using the Benjamini-Hochberg[58] correction based on which a threshold value was set at < 0.01. Multidimensional scaling was performed after the reads 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[59]. 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[26], 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. Loupe cell browser was used to explore the samples and perform general quality checks including UMI distribution, reads per cell and sequencing saturation. Filtered gene-cell matrices were merged in R, and processed for downstream analysis using the Seurat package. Dimension reduction was performed using PCA and tSNE. CMS classification was done using the CMScaller Package, and epithelial and mesenchymal scores were computed using the Rmagic packages. Heatmaps were created using the Pheatmap package. Pseudotime analysis was performed in Python using Palantir.

Data accessibility

The RNA-sequencing data from this study have been submitted to the Gene Expression Omnibus (GEO)[60] database under the accession number GSE154927 and GSE154930for the bulk and single-cell RNAseq data, respectively.

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**, 759-767, doi:10.1016/0092-8674(90)90186-i (1990) DOI: 10.1016/0092-8674(90)90186-i.

2 Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57-70, doi:10.1016/s0092-8674(00)81683-9 (2000) DOI: 10.1016/s0092-8674(00)81683-9.

3 Varga, J. & Greten, F. R. Cell plasticity in epithelial homeostasis and tumorigenesis. *Nature cell biology* **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**, doi:10.3390/cancers11091368 (2019) DOI: 10.3390/cancers11091368.

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**, 704-715 (2008).

7 Brabletz, T., Jung, A., Spaderna, S., Hlubek, F. & Kirchner, T. Opinion: migrating cancer stem cells - an integrated concept of malignant tumour progression. *Nature reviews* **5**, 744-749 (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 Fodde, R. & Brabletz, T. Wnt/beta-catenin signaling in cancer stemness and malignant behavior. *Current opinion in cell biology* **19**, 150-158 (2007).

10 Gupta, P. B. *et al.* Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell* **146**, 633-644, doi:10.1016/j.cell.2011.07.026 (2011) DOI: 10.1016/j.cell.2011.07.026.

11 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**, 5317-5326, doi:10.1158/0008-5472.CAN-11-1059 (2011) DOI: 10.1158/0008-5472.CAN-11-1059.

12 Lengauer, C., Kinzler, K. W. & Vogelstein, B. Genetic instability in colorectal cancers. *Nature* **386**, 623-627, doi:10.1038/386623a0 (1997) DOI: 10.1038/386623a0.

13 Brabletz, S. & Brabletz, T. The ZEB/miR-200 feedback loop--a motor of cellular plasticity in development and cancer? *EMBO Rep* **11**, 670-677, doi:10.1038/embor.2010.117 (2010) DOI: 10.1038/embor.2010.117.

14 Dean, M., Fojo, T. & Bates, S. Tumour stem cells and drug resistance. *Nat Rev Cancer* **5**, 275-284, doi:10.1038/nrc1590 (2005) DOI: 10.1038/nrc1590.

15 Kirchner, T. & Brabletz, T. Patterning and nuclear beta-catenin expression in the colonic adenoma-carcinoma sequence. Analogies with embryonic gastrulation. *Am J Pathol* **157**, 1113-1121, doi:10.1016/s0002-9440(10)64626-3 (2000) DOI: 10.1016/s0002-9440(10)64626-3.

16 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, doi:10.1073/pnas.1701219114 (2017) DOI: 10.1073/pnas.1701219114.

17 Fumagalli, A. *et al.* A surgical orthotopic organoid transplantation approach in mice to visualize and study colorectal cancer progression. *Nat Protoc* **13**, 235-247, doi:10.1038/nprot.2017.137 (2018) DOI: 10.1038/nprot.2017.137.

18 Hall, M. P. *et al.* Click beetle luciferase mutant and near infrared naphthyl-luciferins for improved bioluminescence imaging. *Nat Commun* **9**, 132, doi:10.1038/s41467-017-02542-9 (2018) DOI: 10.1038/s41467-017-02542-9.

19 Nishisho, I. *et al.* Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients. *Science* **253**, 665-669, doi:10.1126/science.1651563 (1991) DOI: 10.1126/science.1651563.

20 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**, 10330-10334, doi:10.1073/pnas.94.19.10330 (1997) DOI: 10.1073/pnas.94.19.10330.

21 Lamouille, S., Xu, J. & Derynck, R. Molecular mechanisms of epithelial-mesenchymal transition. *Nat Rev Mol Cell Biol* **15**, 178-196 (2014).

22 Ghahhari, N. M. & Babashah, S. Interplay between microRNAs and WNT/beta-catenin signalling pathway regulates epithelial-mesenchymal transition in cancer. *Eur J Cancer* **51**, 1638-1649, doi:10.1016/j.ejca.2015.04.021 (2015) DOI: 10.1016/j.ejca.2015.04.021.

23 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**, 19204-19209, doi:10.1073/pnas.1108977108 (2011) DOI: 10.1073/pnas.1108977108.

24 Sanchez-Tillo, E. *et al.* ZEB1 and TCF4 reciprocally modulate their transcriptional activities to regulate Wnt target gene expression. *Oncogene* **34**, 5760-5770, doi:10.1038/onc.2015.352 (2015) DOI: 10.1038/onc.2015.352.

25 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, doi:10.1038/s41598-019-54890-9 (2019) DOI: 10.1038/s41598-019-54890-9.

26 Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat Med* **21**, 1350-1356, doi:10.1038/nm.3967 (2015) DOI: 10.1038/nm.3967.

27 Vellinga, T. T. *et al.* Collagen-rich stroma in aggressive colon tumors induces mesenchymal gene expression and tumor cell invasion. *Oncogene* **35**, 5263-5271, doi:10.1038/onc.2016.60 (2016) DOI: 10.1038/onc.2016.60.

28 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**, 614-618, doi:10.1038/nm.3174 (2013) DOI: 10.1038/nm.3174.

29 Linnekamp, J. F. *et al.* Consensus molecular subtypes of colorectal cancer are recapitulated in in vitro and in vivo models. *Cell Death Differ*, doi:10.1038/s41418-017-0011-5

10.1038/s41418-017-0011-5 [pii] (2018) DOI: 10.1038/s41418-017-0011-5

10.1038/s41418-017-0011-5 [pii].

30 Brattain, M. G. *et al.* Heterogeneity of human colon carcinoma. *Cancer Metastasis Rev* **3**, 177-191, doi:10.1007/bf00048384 (1984) DOI: 10.1007/bf00048384.

31 Mathur, A. *et al.* FGFR2 is amplified in the NCI-H716 colorectal cancer cell line and is required for growth and survival. *PLoS One* **9**, e98515, doi:10.1371/journal.pone.0098515 (2014) DOI: 10.1371/journal.pone.0098515.

32 Quinn, L. A., Moore, G. E., Morgan, R. T. & Woods, L. K. Cell lines from human colon carcinoma with unusual cell products, double minutes, and homogeneously staining regions. *Cancer Res* **39**, 4914-4924 (1979).

33 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821, doi:10.1016/j.cell.2019.05.031 (2019) DOI: 10.1016/j.cell.2019.05.031.

34 van Dijk, D. *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**, 716-729 e727, doi:10.1016/j.cell.2018.05.061 (2018) DOI: 10.1016/j.cell.2018.05.061.

35 Setty, M. *et al.* Characterization of cell fate probabilities in single-cell data with Palantir. *Nature biotechnology* **37**, 451-460 (2019).

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

37 Gaspar, C. & Fodde, R. APC dosage effects in tumorigenesis and stem cell differentiation. *The International journal of developmental biology* **48**, 377-386 (2004).

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

39 Maheswaran, S. & Haber, D. A. Cell fate: Transition loses its invasive edge. *Nature* **527**, 452-453, doi:10.1038/nature16313 (2015) DOI: 10.1038/nature16313.

40 Davalos, V. *et al.* Dynamic epigenetic regulation of the microRNA-200 family mediates epithelial and mesenchymal transitions in human tumorigenesis. *Oncogene* **31**, 2062-2074, doi:10.1038/onc.2011.383 (2012) DOI: 10.1038/onc.2011.383.

41 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**, 1130-1134 (1998).

42 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**, 7654-7663 (2011).

43 Morral, C. *et al.* Zonation of Ribosomal DNA Transcription Defines a Stem Cell Hierarchy in Colorectal Cancer. *Cell stem cell* (2020).

44 Xu, Y. *et al.* SUSD2 promotes cancer metastasis and confers cisplatin resistance in high grade serous ovarian cancer. *Experimental cell research* (2018).

45 Calon, A. *et al.* Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nat Genet* **47**, 320-329, doi:10.1038/ng.3225 (2015) DOI: 10.1038/ng.3225.

46 Jolly, M. K. *et al.* Implications of the Hybrid Epithelial/Mesenchymal Phenotype in Metastasis. *Frontiers in oncology* **5**, 155 (2015).

47 Aiello, N. M. *et al.* EMT Subtype Influences Epithelial Plasticity and Mode of Cell Migration. *Developmental cell* **45**, 681-695 e684 (2018).

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

49 Pastushenko, I. & Blanpain, C. EMT Transition States during Tumor Progression and Metastasis. *Trends in cell biology* **29**, 212-226 (2018).

50 Yu, M. *et al.* Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. *Science* **339**, 580-584, doi:339/6119/580 [pii]

10.1126/science.1228522 (2013) DOI: 339/6119/580 [pii]

10.1126/science.1228522.

51 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**, 138-145, doi:10.1016/j.ebiom.2016.01.007

S2352-3964(16)30005-6 [pii] (2016) DOI: 10.1016/j.ebiom.2016.01.007

S2352-3964(16)30005-6 [pii].

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

53 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).

54 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:bts635 [pii]

10.1093/bioinformatics/bts635 (2013) DOI: bts635 [pii]

10.1093/bioinformatics/bts635.

55 Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-1774, doi:22/9/1760 [pii]

10.1101/gr.135350.111 (2012) DOI: 22/9/1760 [pii]

10.1101/gr.135350.111.

56 Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80, doi:gb-2004-5-10-r80 [pii]

10.1186/gb-2004-5-10-r80 (2004) DOI: gb-2004-5-10-r80 [pii]

10.1186/gb-2004-5-10-r80.

57 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:s13059-014-0550-8 [pii]

10.1186/s13059-014-0550-8 (2014) DOI: s13059-014-0550-8 [pii]

10.1186/s13059-014-0550-8.

58 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, doi:https://doi.org/10.2307/2346101 (1995) DOI: https://doi.org/10.2307/2346101.

59 Meinders, M. *et al.* Sp1/Sp3 transcription factors regulate hallmarks of megakaryocyte maturation and platelet formation and function. *Blood* **125**, 1957-1967, doi:blood-2014-08-593343 [pii]

10.1182/blood-2014-08-593343 (2015) DOI: blood-2014-08-593343 [pii]

10.1182/blood-2014-08-593343.

60 Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207-210 (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. Using specific gates, cells were divided in a large $CD44^{high}EpCAM^{high}$ cluster, together with a considerably smaller $CD44^{high}EpCAM^{low}$ subpopulation. To ensure good separation from the large $EpCAM^{high}$ cluster and maximal sorting purity, $EpCAM^{low}$ cells were gated as $CD44^{high}$ events ≤60% of the EpCAM fluorescence intensity of the left border of the $CD44^{high}EpCAM^{high}$ gate, and sorted from ≤50% of that value. The percentages of $CD44^{high}EpCAM^{low}$ cells within each cell line are indicated. 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 $CD44^{high}EpCAM^{high}$ and $CD44^{high}EpCAM^{low}$ cells from HCT116 (upper images) and SW480 (lower images) cells. Scale bar: 100 µm.

C. Transwell migration assay of $CD44^{high}EpCAM^{high}$ (black bar) and $CD44^{high}EpCAM^{low}$ (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 ($P < 0.05$).

D. Invasion assay of $CD44^{high}EpCAM^{high}$ (black bar) and $CD44^{high}EpCAM^{low}$ (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 ($P < 0.05$).

Figure 1

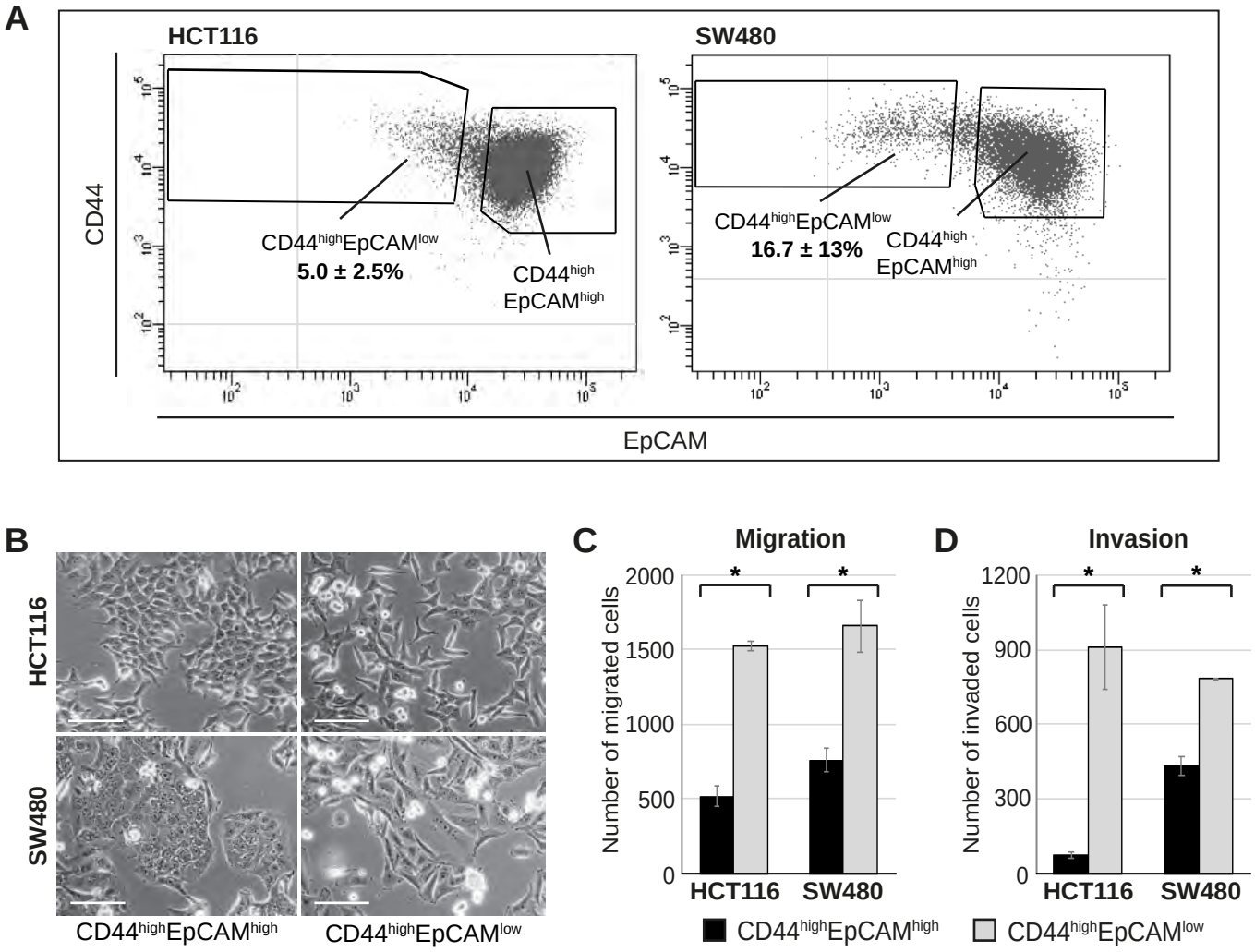

A
HCT116
SW480
CD44
EpCAM
CD44highEpCAMlow
5.0 ± 2.5%
CD44high
EpCAMhigh
CD44highEpCAMlow
16.7 ± 13%
CD44high
EpCAMhigh
B
HCT116
SW480
CD44highEpCAMhigh
CD44highEpCAMlow
C
Migration
Number of migrated cells
2000
1500
1000
500
0
HCT116
SW480
D
Invasion
Number of invaded cells
1200
900
600
300
0
HCT116
SW480
CD44highEpCAMhigh
CD44highEpCAMlow

Figure 1 – supplement 1

Table containing panel of commonly used colon cancer cell lines analyzed for their CD44 and EpCAM expression by flow cytometric analysis. Per cell line the percentage $CD44^{high}EpCAM^{low}$ cells is presented. If indicated by Lindeman et. al CMS classification of the cell line is shown.

Figure 1 - figure supplement 1

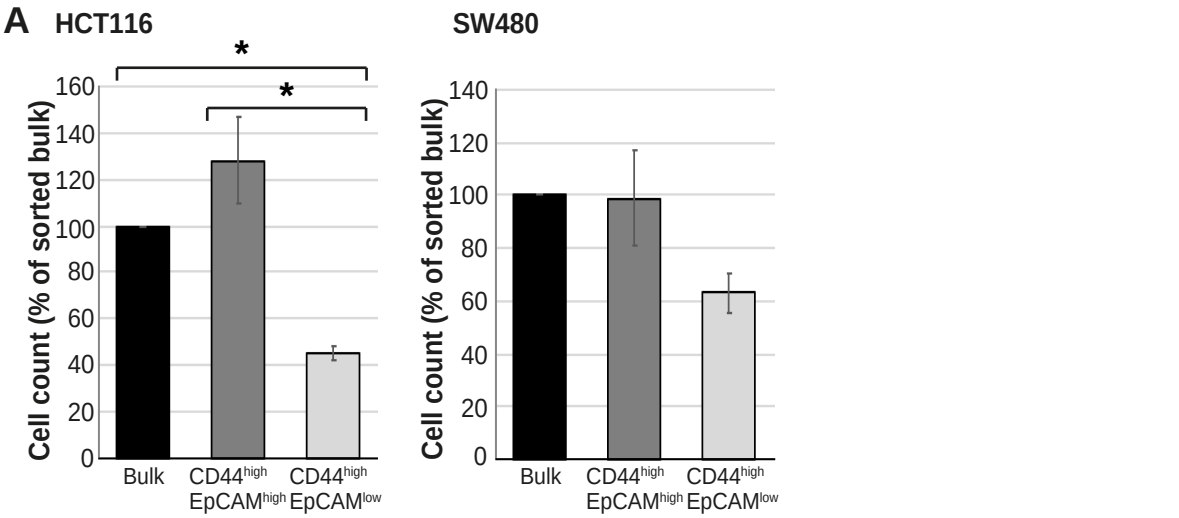
	% $CD44^{high}EpCAM^{low}$	Linnekamp et al.
HCT116 - early culture	5.0 ± 2.5	CMS1
HCT116 - late culture	0.32 ± 0.18	
SW480 - early culture	16.7 ± 13	
SW480 - late culture	0.34 ± 0.24	
RKO	100	CMS1
COLO320-HSR	99.6	CMS4
NCI-H716	66.0 ± 21.21	CMS4
SW620	12.0 ± 4.9	no subtype
CAR-1	0.15 ± 0.07	CMS4
HT29	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

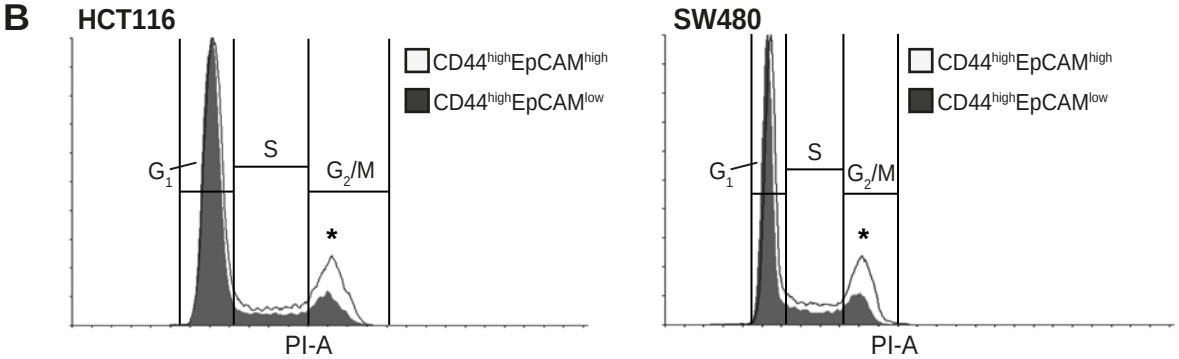
Figure 1 – supplement 2

A. Cell proliferation assay. Sorted bulk, $CD44^{high}EpCAM^{high}$ and $CD44^{high}EpCAM^{low}$ 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.

B. Cell cycle analysis of $CD44^{high}EpCAM^{high}$ and $CD44^{high}EpCAM^{low}$ cells in HCT116 (left panel) and SW480 (right 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: $CD44^{high}EpCAM^{high}$, grey graph: $CD44^{high}EpCAM^{low}$. Asterisks show the significant ($P < 0.05$) differences between $CD44^{high}EpCAM^{high}$ and $CD44^{high}EpCAM^{low}$ cells in G_1 and G_2M-phases.

Figure 1 - figure supplement 2





	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 2

A. RT-qPCR expression analysis of epithelial (*EPCAM* and *CDH1*) and mesenchymal (*VIM*) markers in sorted $CD44^{high}EpCAM^{high}$ (black bars) and $CD44^{high}EpCAM^{low}$ (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 ($P < 0.05$).

B. RT-qPCR expression analysis of EMT transcription factors (*ZEB1, ZEB2, TWIST, FOXC2, SLUG* and *SNAIL*) in $CD44^{high}EpCAM^{high}$ (black bars) and $CD44^{high}EpCAM^{low}$ (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 ($P < 0.05$).

C. Immunofluorescence (IF) analysis of $CD44^{high}EpCAM^{high}$ and $CD44^{high}EpCAM^{low}$ 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.

D. 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.

E. Bar graph of flow cytometric analysis of HCT116 and SW480 shCT- and shZEB1-transfected cell lines using antibodies against CD44 and EpCAM. Cells were induced with 1 µg/mL doxycycline for 72 hrs before analysis. Each bar represents the relative mean ± SD of three independent experiments.

F. Representative analysis of one experiment of the FACS analysis as shown in **E**.

G. RT-qPCR expression analysis of the members of the miRNA 200 family (miR-200a, miR-200b, miR-200c, miR-141 and miR-429) in $CD44^{high}EpCAM^{high}$ (black bars) and $CD44^{high}EpCAM^{low}$ (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 $P < 0.001$.

Figure 2

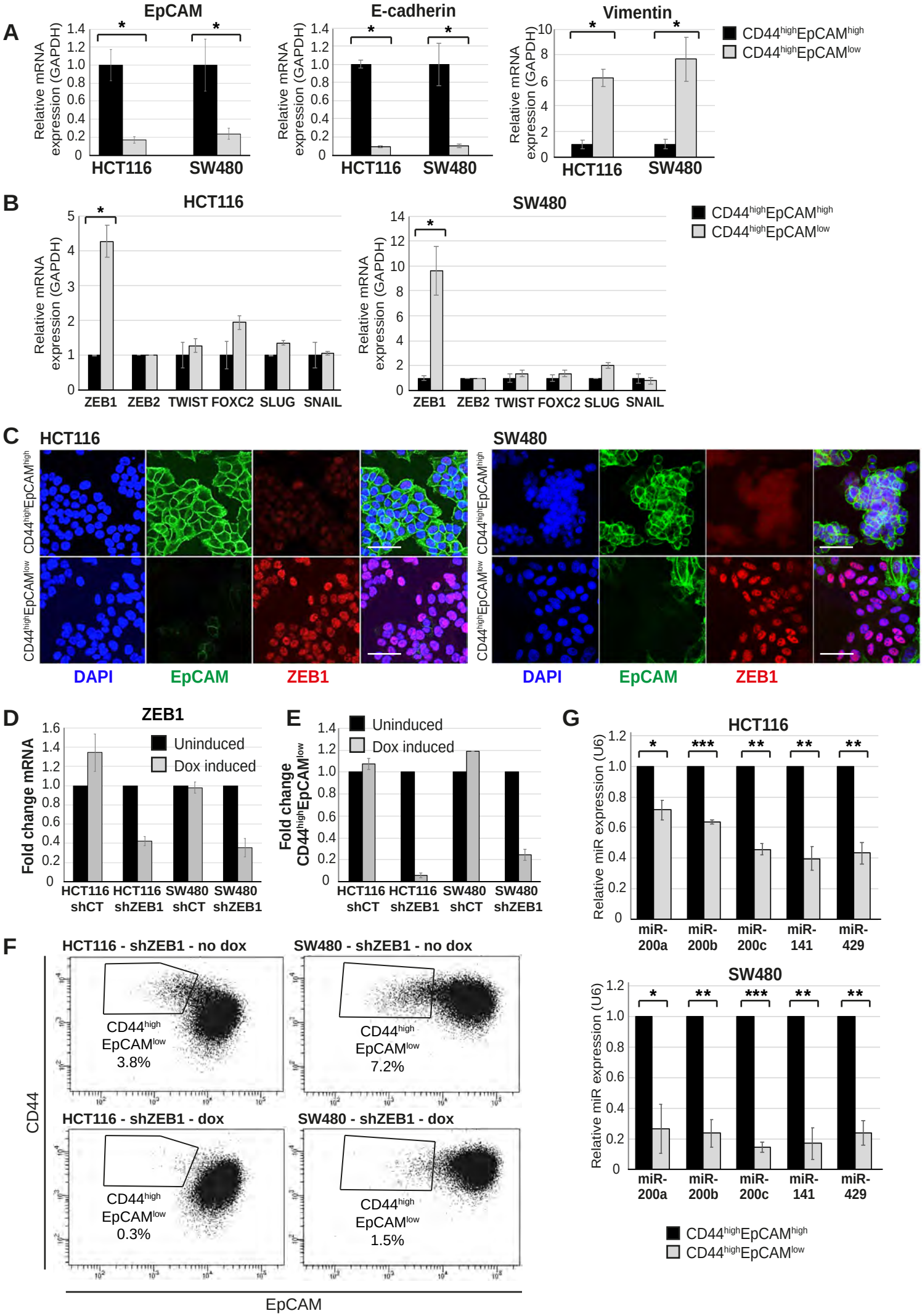

A
EpCAM
E-cadherin
Vimentin
Relative mRNA expression (GAPDH)
HCT116
SW480
CD44highEpCAMhigh
CD44highEpCAMlow
B
HCT116
SW480
ZEB1
ZEB2
TWIST
FOXC2
SLUG
SNAIL
C
DAPI
EpCAM
ZEB1
D
ZEB1
Fold change mRNA
Uninduced
Dox induced
HCT116 shCT
HCT116 shZEB1
SW480 shCT
SW480 shZEB1
E
Fold change CD44highEpCAMlow
G
Relative miR expression (U6)
miR-200a
miR-200b
miR-200c
miR-141
miR-429
F
HCT116 - shZEB1 - no dox
SW480 - shZEB1 - no dox
HCT116 - shZEB1 - dox
SW480 - shZEB1 - dox
CD44high EpCAMlow 3.8%
CD44high EpCAMlow 7.2%
CD44high EpCAMlow 0.3%
CD44high EpCAMlow 1.5%
CD44
EpCAM

Figure 3

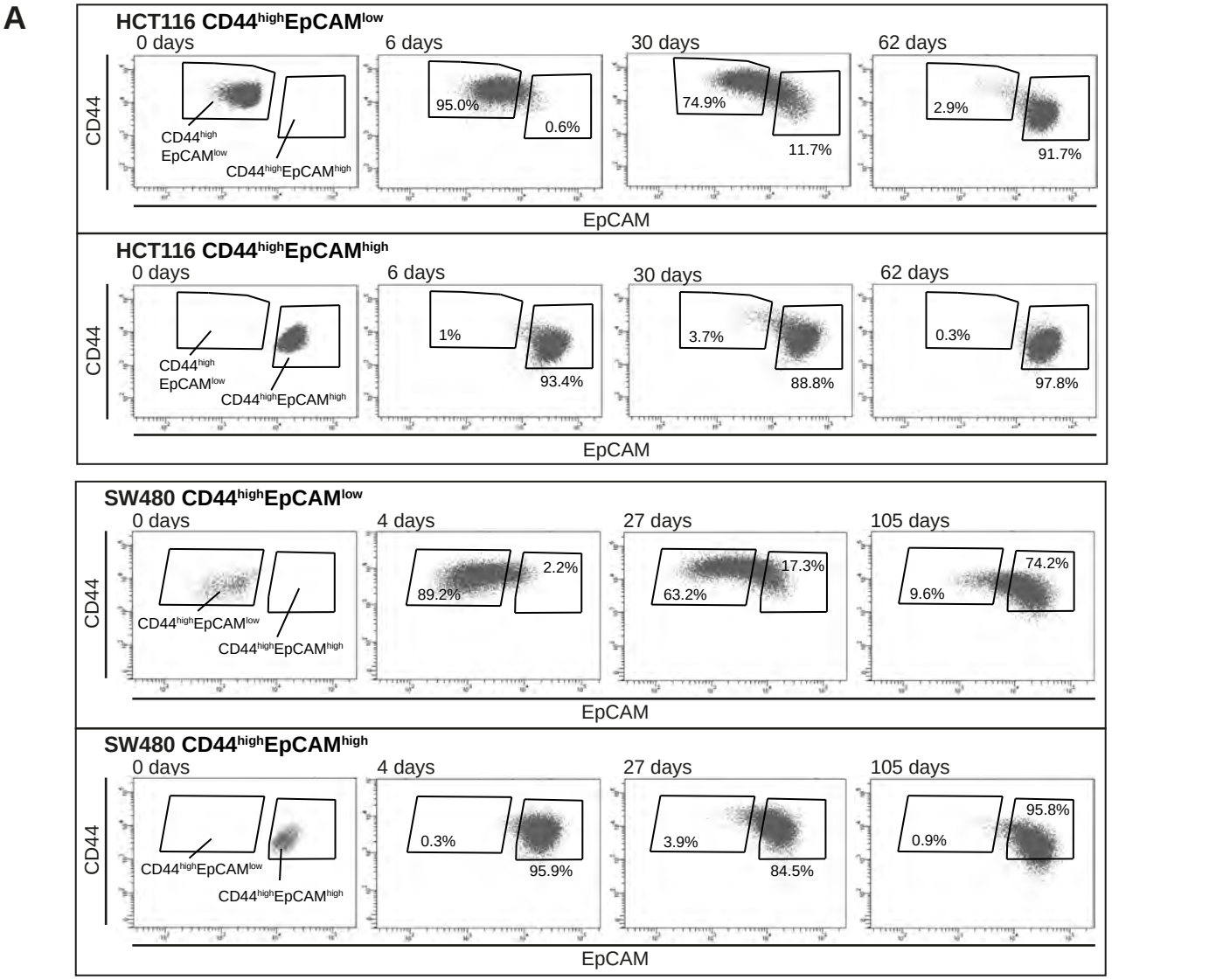
Analysis of plasticity of CD44^{high}EpCAM^{high} and CD44^{high}EpCAM^{low} cells from HCT116 (**A – upper panel**) and SW480 (**A – lower panel**). CD44^{high}EpCAM^{high} and CD44^{high}EpCAM^{low} cell fractions were sorted and plated in culture. At different time points, as indicated, cells were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

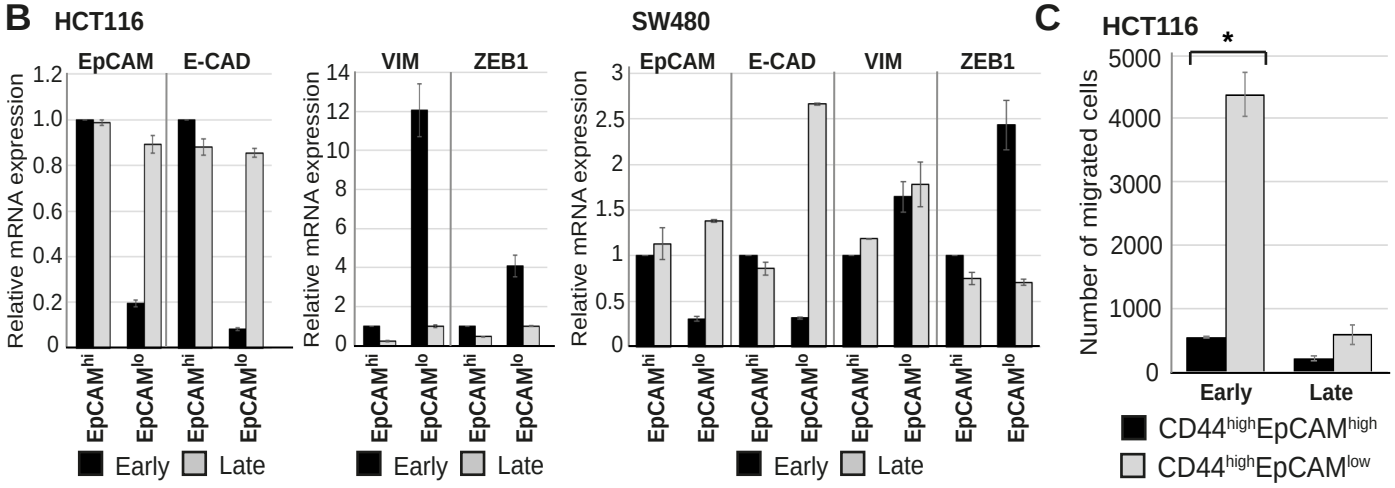
B. RT-qPCR expression analysis of *EPCAM*, *CDH1*, *VIM*, and *ZEB1* in 'early' (e.g. cells that were used shortly after FACS sorting) and 'late' (e.g. cells that were cultured for an extended period of time before performing the experiment) sorted CD44^{high}EpCAM^{high} and CD44^{high}EpCAM^{low} cell from HCT116 and SW480 cell lines. *GAPDH* was employed for normalization purposes. Each bar represents the mean ± SD of two independent experiments.

C. Transwell migration assay of 'early' and 'late' CD44^{high}EpCAM^{high} (black bar) and CD44^{high}EpCAM^{low} (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 ($P < 0.05$).

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

Figure 3


A
HCT116 CD44highEpCAMlow
0 days
6 days
30 days
62 days
CD44
EpCAM
CD44high EpCAMlow
CD44highEpCAMhigh
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
2.2%
89.2%
17.3%
63.2%
74.2%
9.6%
SW480 CD44highEpCAMhigh
0.3%
95.9%
3.9%
84.5%
95.8%
0.9%


B
HCT116
SW480
EpCAM
E-CAD
VIM
ZEB1
Relative mRNA expression
EpCAMhi
EpCAMlo
Early
Late
C
HCT116
Number of migrated cells
Early
Late
CD44highEpCAMhigh
CD44highEpCAMlow

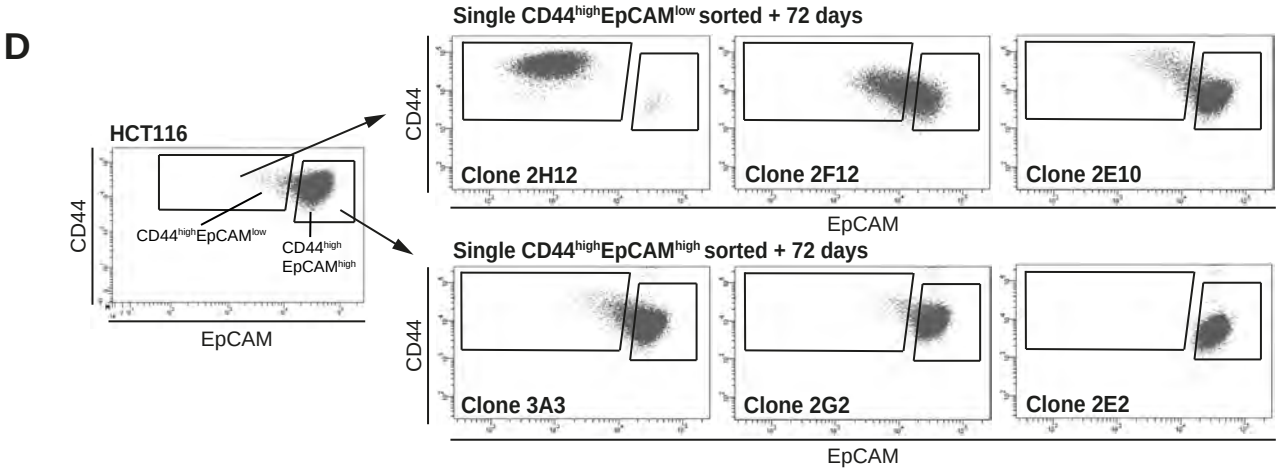

D
HCT116
CD44
EpCAM
CD44highEpCAMlow
CD44high EpCAMhigh
Single CD44highEpCAMlow sorted + 72 days
Clone 2H12
Clone 2F12
Clone 2E10
Single CD44highEpCAMhigh sorted + 72 days
Clone 3A3
Clone 2G2
Clone 2E2

Figure 3 – supplement 1

Table with selection of individual single cell sorted CD44^{high}EpCAM^{high} and CD44^{high}EpCAM^{low} 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 CD44^{high}EpCAM^{high} and CD44^{high}EpCAM^{low} gates are presented in the bar graph. Total success rate of clones grown out of single cells: HCT116 CD44^{high}EpCAM^{high} 15 clones out of 24 sorted cells, HCT116 CD44^{high}EpCAM^{low}: 31/48, SW480 CD44^{high}EpCAM^{high}: 5/36, SW480 CD44^{high}EpCAM^{low}: 11/36.

Figure 3 - figure supplement 1

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

Figure 4

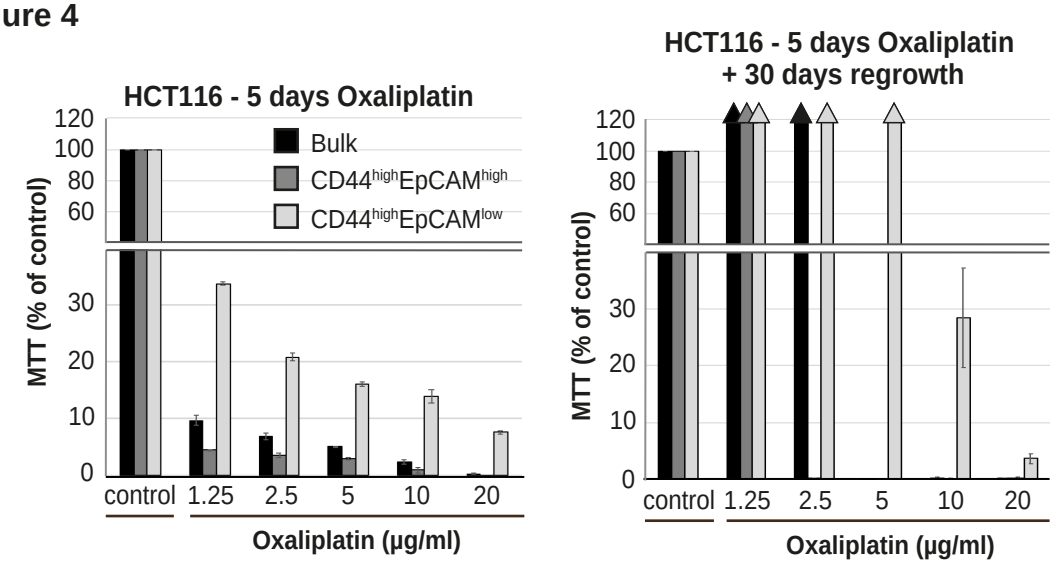
A-D. Bulk, CD44^{high}EpCAM^{high} and CD44^{high}EpCAM^{low} HCT116 and SW480 cells were sorted and plated to recover and expand for four days. 10^5 cells per well were plated in in triplicate in a 96-well plate and left to attach. After 24 h, standard cell culture medium was changed for medium containing 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), CD44^{high}EpCAM^{high} (dark grey bars) and CD44^{high}EpCAM^{low} (light grey bars) HCT116 cells. Bars represent the average ± SD of the O.D. corrected for the background measurements.

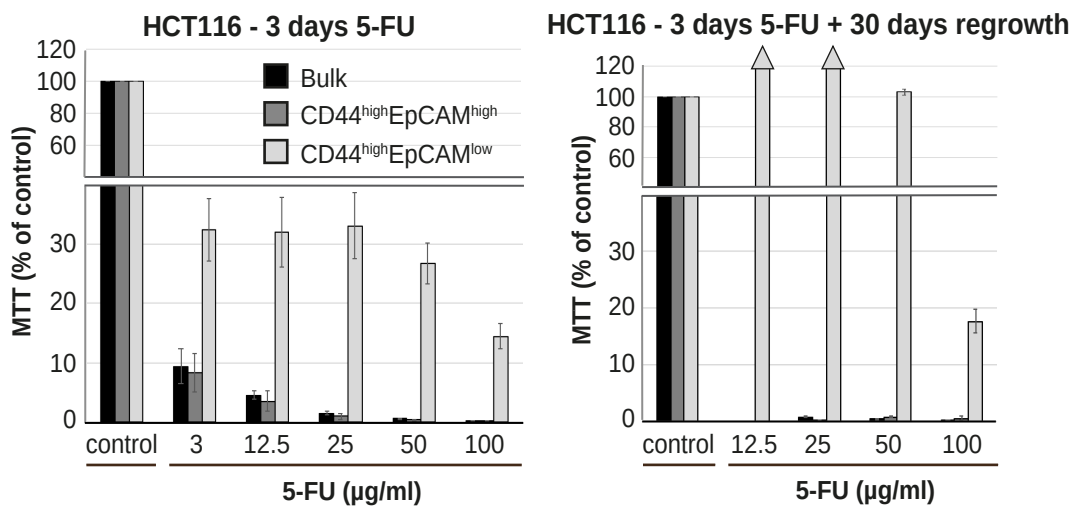
C-D. Chemo sensitivity assays using oxaliplatin (**C**) or 5-FU (**D**) in bulk (black bars), CD44^{high}EpCAM^{high} (dark grey bars) and CD44^{high}EpCAM^{low} (light grey bars) SW480 cells. Bars represent the average ± SD of the O.D. corrected for the bachground measurements.

Figure 4

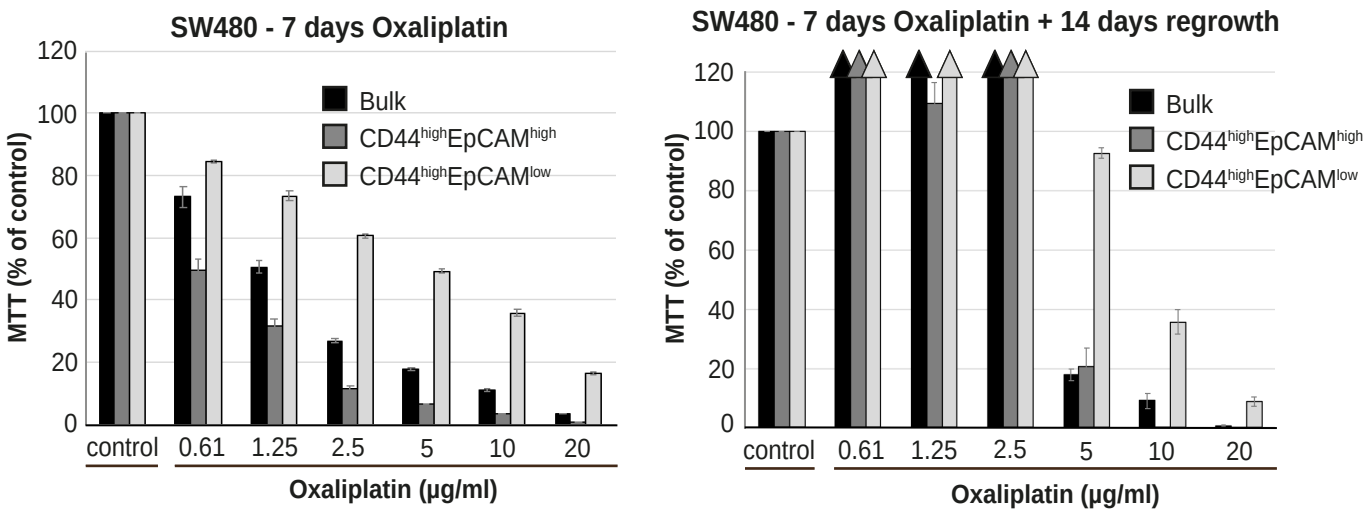
A



B



C



D

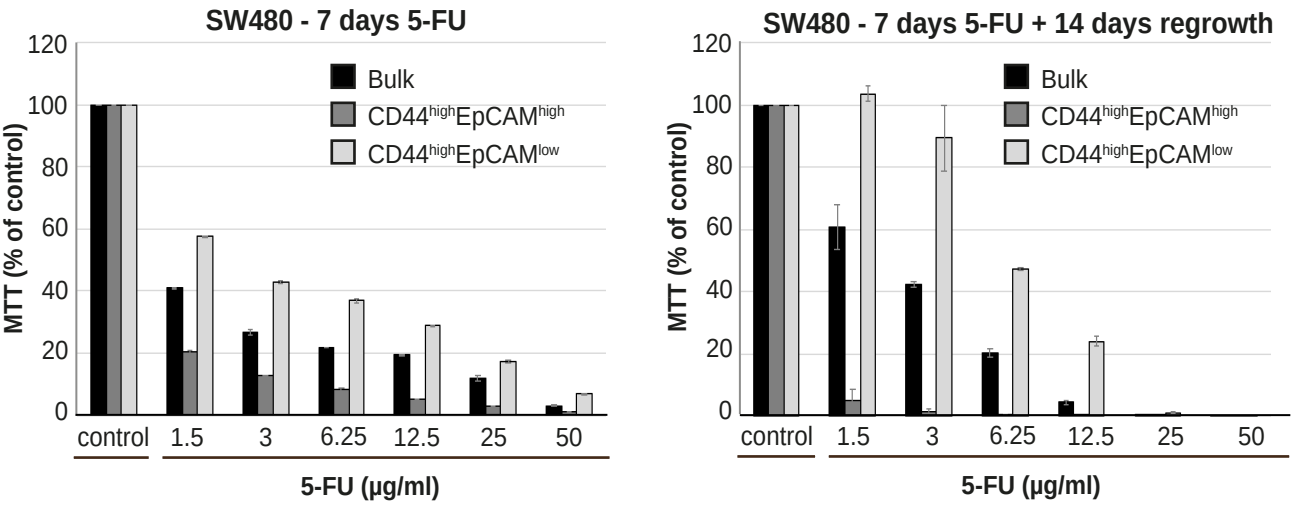


Figure 5

A. Intrasplenic injection of bulk (n=8), CD44^{high}EpCAM^{high} (n=9), and CD44^{high}EpCAM^{low} (n=7) HCT116 cells in NSG mice. For each transplantation experiment, 2 x 10^4 cells were injected in the spleen of a recipient NSG mouse. Four 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).

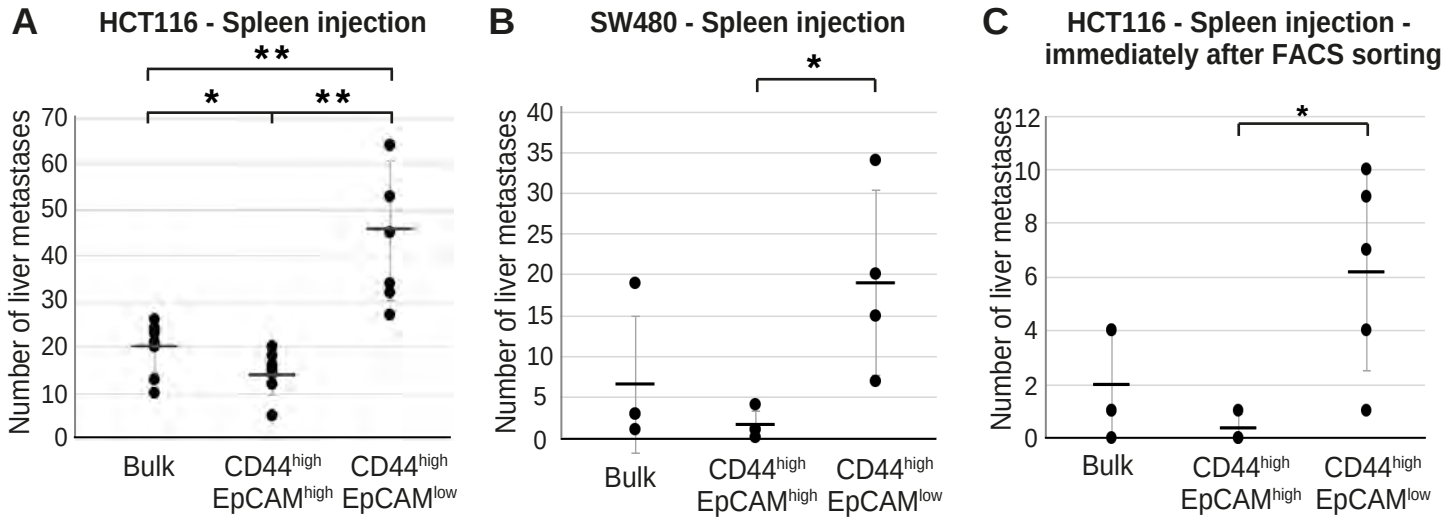
B. Intrasplenic injection of bulk (n=4), CD44^{high}EpCAM^{high} (n=4) and CD44^{high}EpCAM^{low} (n=4) SW480 cells in NSG mice. Mice were sacrificed eight weeks after injection of 2 x 10^4 cells in the spleen. Asterisks indicate significant differences ($P < 0.05$).

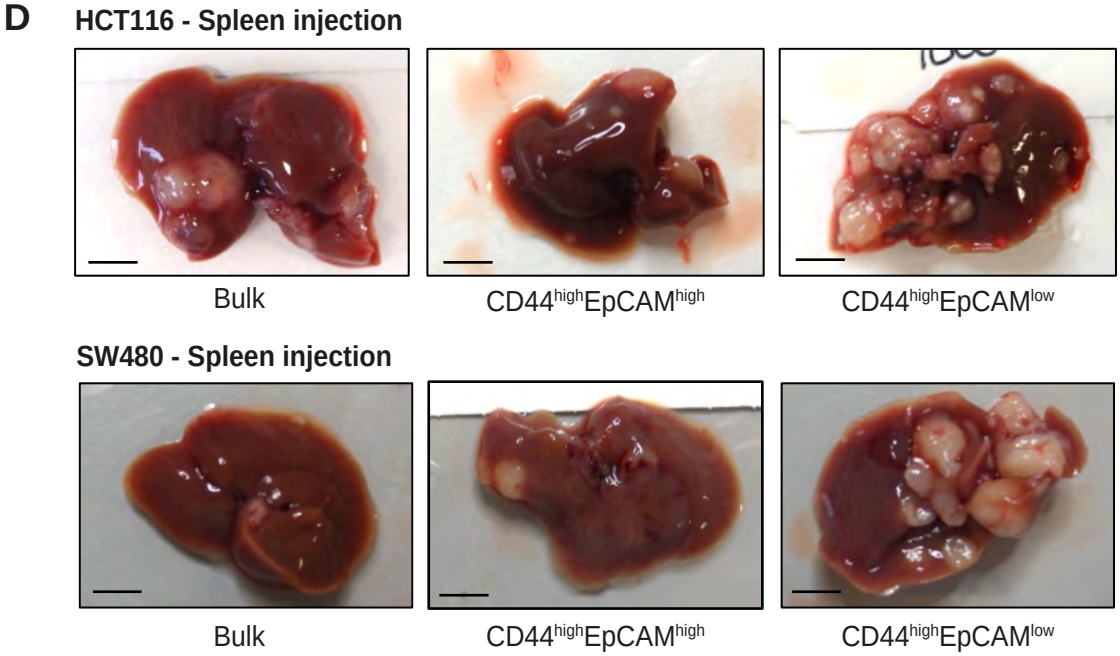
C. Intrasplenic injection of bulk (n=5), CD44^{high}EpCAM^{high} (n=3) and CD44^{high}EpCAM^{low} (n=5) HCT116 cells. 10^4 bulk, CD44^{high}EpCAM^{high} and CD44^{high}EpCAM^{low} cells were sorted and immediately injected in the spleen of NSG mice. Mice were sacrificed four weeks after injections and individual tumors were counted. Asterisks show significance of $P < 0.05$.

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

E. Hematocylin 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.

Figure 5


A
HCT116 - Spleen injection
Number of liver metastases
0
10
20
30
40
50
60
70
*
**
**
Bulk
CD44high EpCAMhigh
CD44high EpCAMlow
B
SW480 - Spleen injection
Number of liver metastases
0
5
10
15
20
25
30
35
40
*
Bulk
CD44high EpCAMhigh
CD44high EpCAMlow
C
HCT116 - Spleen injection - immediately after FACS sorting
Number of liver metastases
0
2
4
6
8
10
12
*
Bulk
CD44high EpCAMhigh
CD44high EpCAMlow


D
HCT116 - Spleen injection
Bulk
CD44highEpCAMhigh
CD44highEpCAMlow
SW480 - Spleen injection
Bulk
CD44highEpCAMhigh
CD44highEpCAMlow

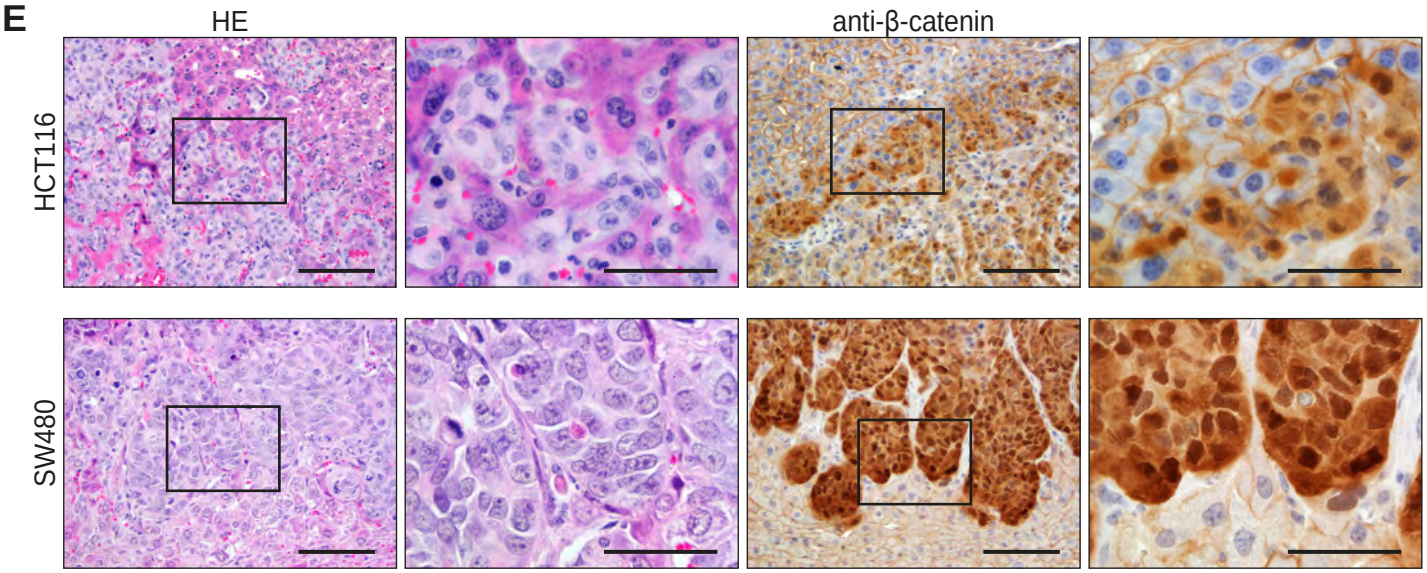
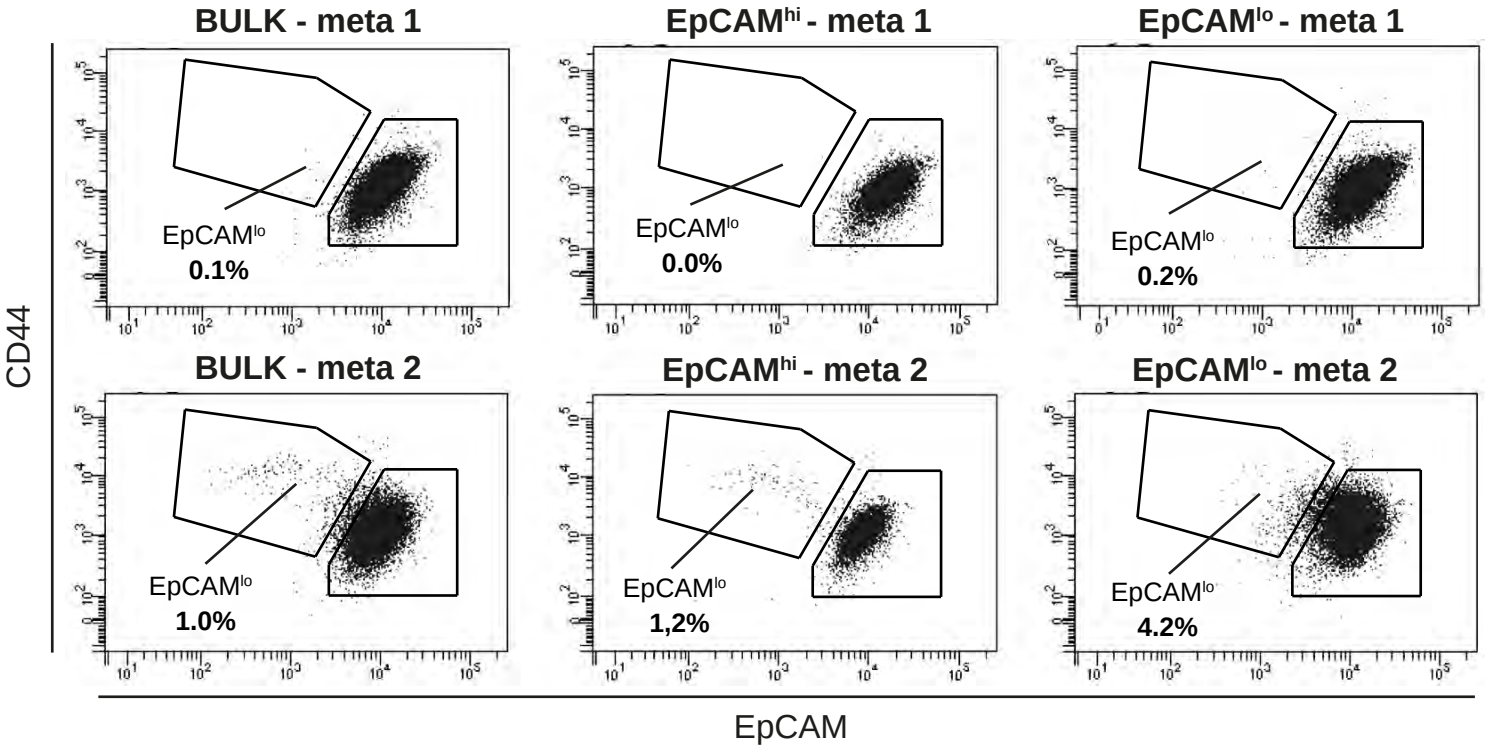

E
HE
anti-β-catenin
HCT116
SW480

Figure 5 – supplement 1

FACS analysis of liver metastases obtained by spleen injection of HCT116 bulk, EpCAM^{hi}, and EpCAM^{lo} cells. **A.** Representative FACS plots. **B.** Quantification of EpCAM$^{hi/lo}$ percentages of all liver metastases as determined by FACS.

Figure 5 - figure supplement 1

A



B

	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^{lo}	0.1	1.0	0.0	0.1	0.0	0.0	1.2	0.2	1.2	4.2	1.1	51.9
% EpCAM^{hi}	99.1	95.3	98.8	99.2	99.7	99.6	97.7	98.4	91.0	83.2	96.8	25.3

Figure 6

A. Microscopic (top panels) analysis of AKP-Z organoids tagged with GFP and click beetle luciferase. Bottom graph 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.

B. 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 doxycylin. Graphs show one representative graph of one control (no doxycycline treated mouse) and two of primary tumors treated with doxycycline.

C. IVIS imaging of liver and lung tissues derived from non-doxycylin treated mice (black; n=4) and doxycycline treated mice (grey; n=5). Doxycycline treatment was started 2 weeks after transplantation and was continued during the rest of the experiment. Mice were sacrificed 8 weeks after transplantation.

D. Quantification of the number of liver metastases in uninduced (black; n=4) and 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.

E. Quantification of the number of lung metastases in uninduced (black; n=4) and 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.

F. Correlation of number of liver (**Figure 6D**) and lung (**Figure 6E**) metastasis per mouse (doxycycline uninduced: n=4; doxycycline induced: n=5).

Figure 6

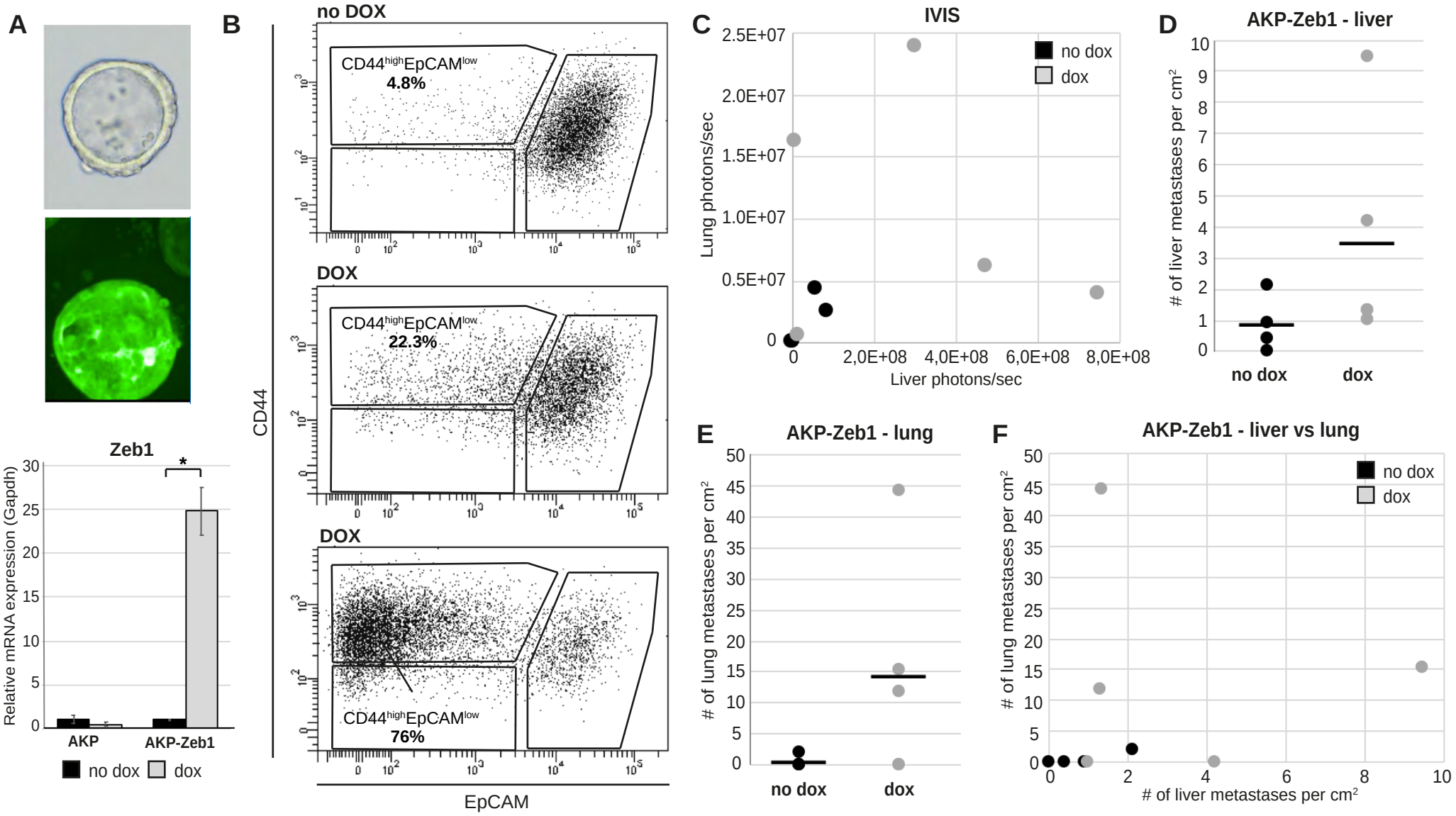

A
Zeb1
Relative mRNA expression (Gapdh)
AKP
AKP-Zeb1
no dox
dox
B
no DOX
CD44highEpCAMlow
4.8%
DOX
CD44highEpCAMlow
22.3%
DOX
CD44highEpCAMlow
76%
CD44
EpCAM
C
IVIS
Lung photons/sec
Liver photons/sec
no dox
dox
D
AKP-Zeb1 - liver
# of liver metastases per cm²
no dox
dox
E
AKP-Zeb1 - lung
# of lung metastases per cm²
no dox
dox
F
AKP-Zeb1 - liver vs lung
# of lung metastases per cm²
# of liver metastases per cm²
no dox
dox

Figure 7

A. Multidimensional scaling (MDS) analysis of RNAseq profiles of $CD44^{high}EpCAM^{high}$ and $CD44^{high}EpCAM^{low}$ cells from the HCT116 and SW480 lines. Red: HCT116, black: SW480, circle: $CD44^{high}EpCAM^{high}$, triangle: $CD44^{high}EpCAM^{low}$.

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 HCT116 (left panel) and SW480 (right panel) cultures. Graphs show percentage of cells within the $CD44^{high}EpCAM^{high}$ and $CD44^{high}EpCAM^{low}$ 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 (Kim, Kwon et al., Scientific Reports, 2019).

Figure 7

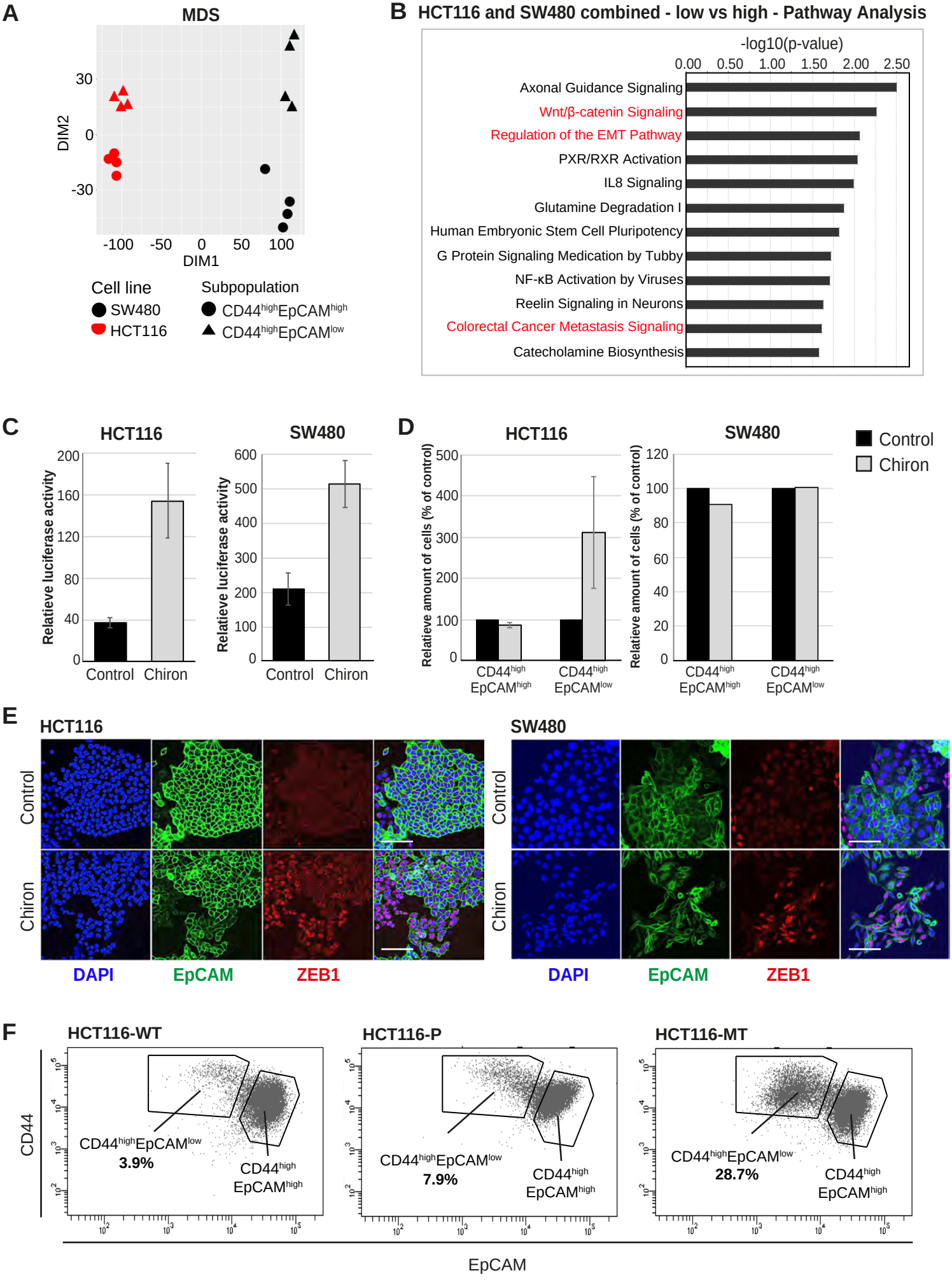

A
MDS
DIM2
DIM1
30
0
-30
-100
-50
0
50
100
Cell line
SW480
HCT116
Subpopulation
CD44highEpCAMhigh
CD44highEpCAMlow
B
HCT116 and SW480 combined - low vs high - Pathway Analysis
-log10(p-value)
0.00
0.50
1.00
1.50
2.00
2.50
Axonal Guidance Signaling
Wnt/β-catenin Signaling
Regulation of the EMT Pathway
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)
CD44high EpCAMhigh
CD44high EpCAMlow
Control
Chiron
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 7 – supplement 1

A. Gene list of commonly differentially expressed genes in $CD44^{high}EpCAM^{low}$ versus $CD44^{high}EpCAM^{high}$ cells in HCT116 and SW480. A total of 152 and 353 differentially regulated genes were identified between the $CD44^{high}EpCAM^{high}$ and $CD44^{high}EpCAM^{low}$ 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 7 - 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

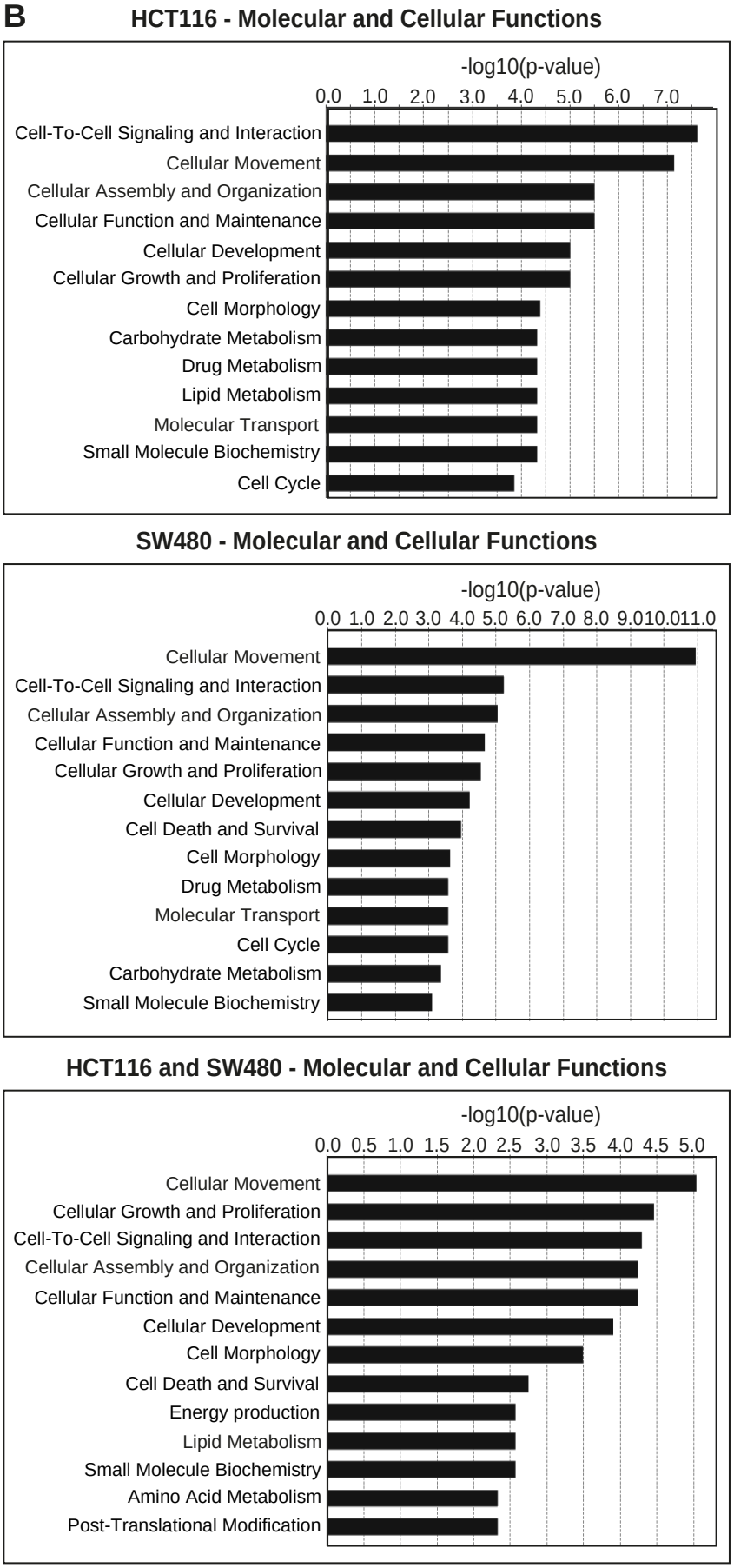


Figure 8

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 8

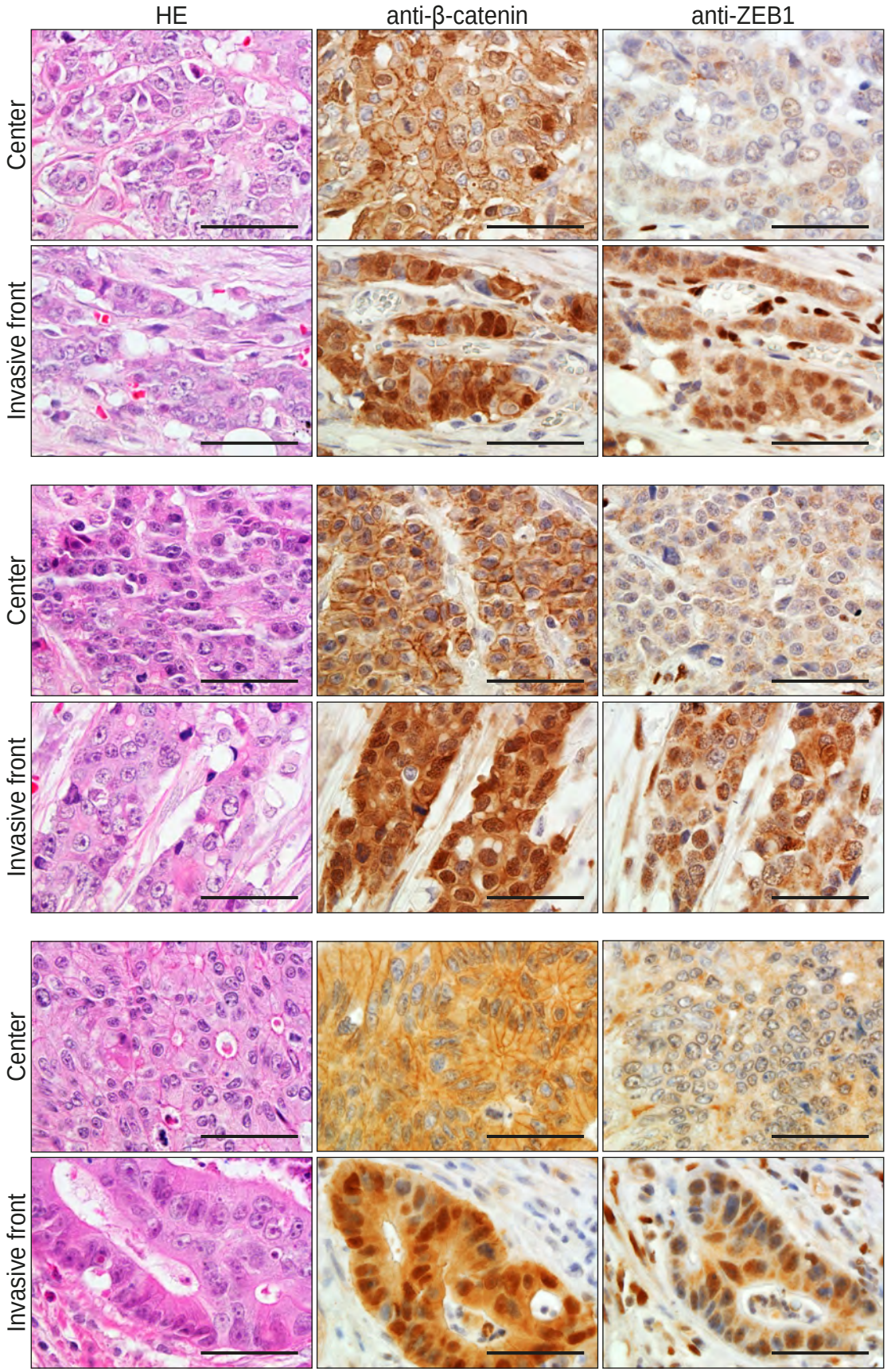

HE
anti-β-catenin
anti-ZEB1
Center
Invasive front
Center
Invasive front
Center
Invasive front

Figure 9

A. Correlation of meta-gene expression values of the signatures derived from CD44^{high}EpCAM^{low} HCT116 and SW480 cells in the CMS-3232 human colon cancer cohort[26].

B. Correlation of meta-gene expression values of the signatures derived from CD44^{high}EpCAM^{low} 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 CD44^{high}EpCAM^{low} 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 9

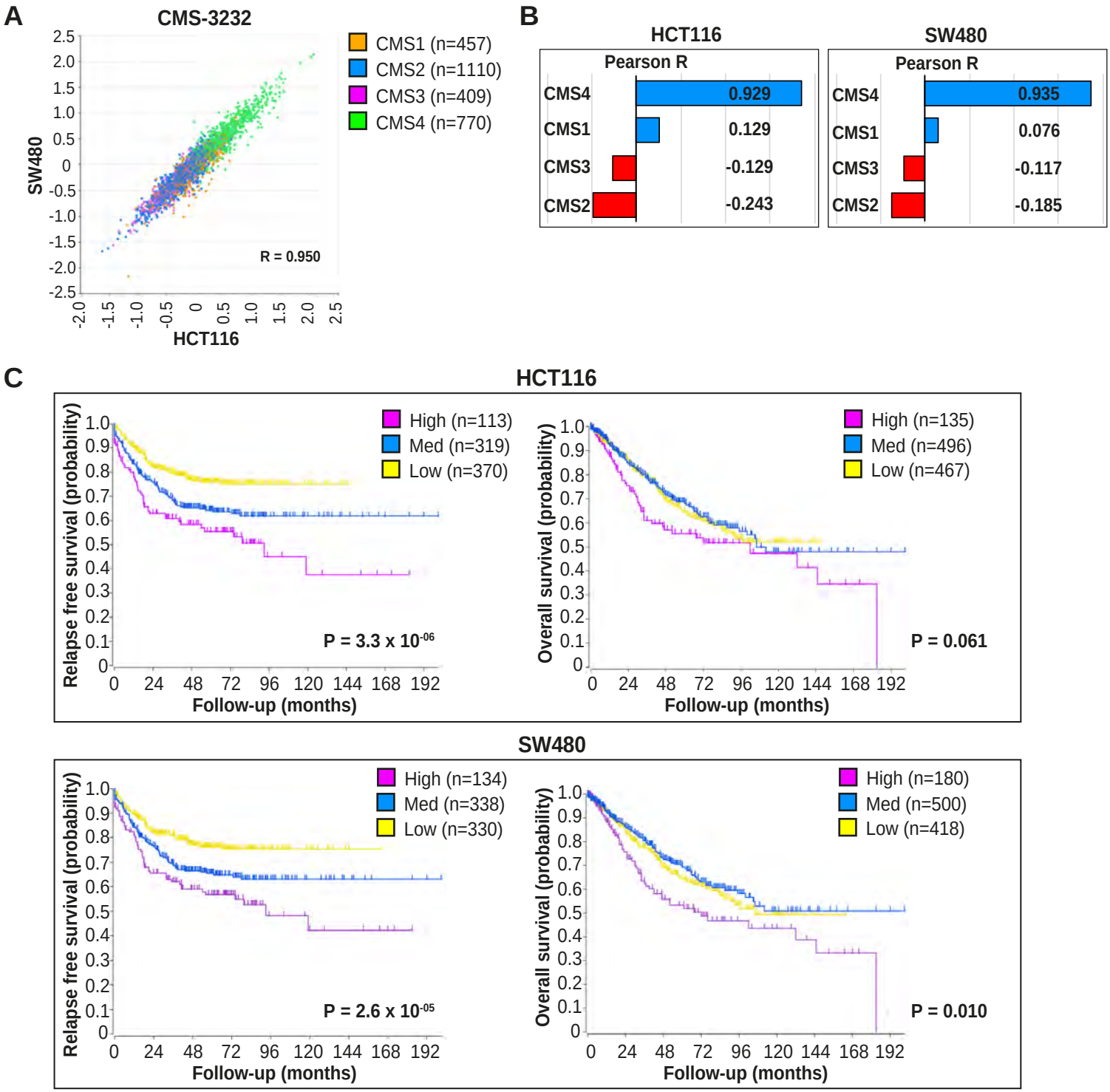

A
CMS-3232
SW480
HCT116
R = 0.950
CMS1 (n=457)
CMS2 (n=1110)
CMS3 (n=409)
CMS4 (n=770)
B
HCT116
SW480
Pearson R
CMS4
CMS1
CMS3
CMS2
0.929
0.129
-0.129
-0.243
0.935
0.076
-0.117
-0.185
C
HCT116
Relapse free survival (probability)
Follow-up (months)
High (n=113)
Med (n=319)
Low (n=370)
P = 3.3 x 10-06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

Figure 9 – supplement 1

A. Correlation of meta-gene expression values of the signatures derived from CD44^{high}EpCAM^{low} 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 [28].

B. The gene sets identifying the CD44^{high}EpCAM^{low} 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 LOW, INTERMEDIATE, HIGH expression subgroups.

Figure 9 - figure supplement 1

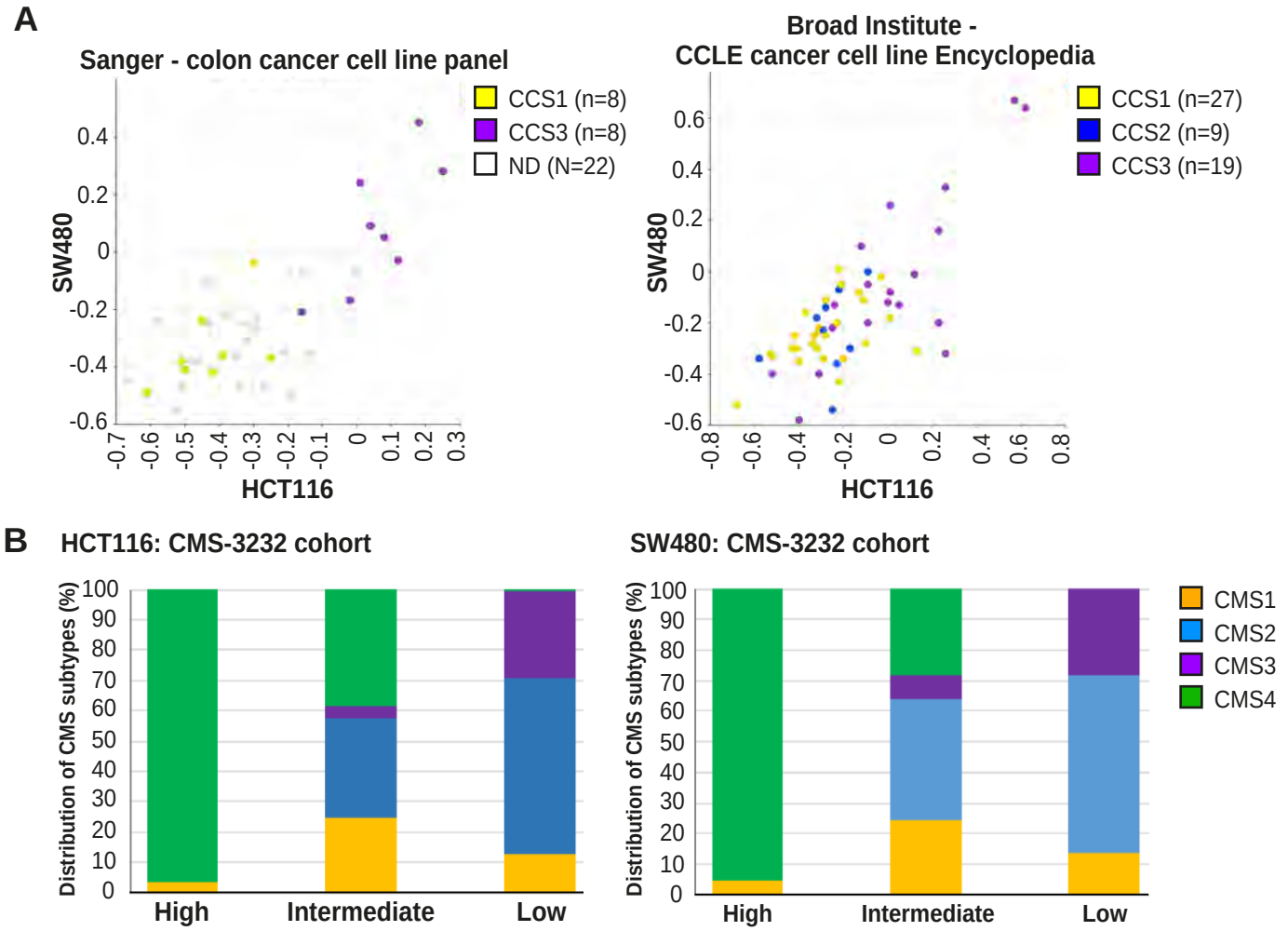


Figure 10

A. tSNE of HCT116 cells based on the variable expressed genes across $CD44^{high}EpCAM^{low}$ and $CD44^{high}EpCAM^{high}$ populations.

B. tSNE of SW480 cells based on the variable expressed genes across $CD44^{high}EpCAM^{low}$ and $CD44^{high}EpCAM^{high}$ populations after exclusion of the non-adherent cells using a signature derived from bulk RNAseq experiments.

C. Unsupervised clustering of HCT116 using shared nearest neighbour (SNN) modularity optimization. SNN revealed additional layers of heterogeneity as shown by the presence of 8 distinct clusters: 3 of $EpCAM^{hi}$ origin, and 5 $EpCAM^{lo}$.

D. Unsupervised clustering of SW480 using SNN modularity optimization. SW480 revealed several mixed clusters (n=7) with varying contributions from the two subpopulations.

E. tSNE of HCT116 cells based on EMT signature from Nanostring EMT panel (N = 107 genes).

F. tSNE of SW480 cells based on EMT signature from Nanostring EMT panel (N = 107 genes).

G. Projection of HCT116 cells based on their epithelial and mesenchymal score computed by Markov Affinity-Based Graph Imputation of Cells (MAGIC).

H. Projection of the unsupervised clusters derived in **C** on the E/M scores on a 2D EMT landscape.

I. Palantir graph embedded tSNE of the E/M clusters (2, 5, 7) modelling dynamic plasticity trajectories of differentiated cells.

Figure 10

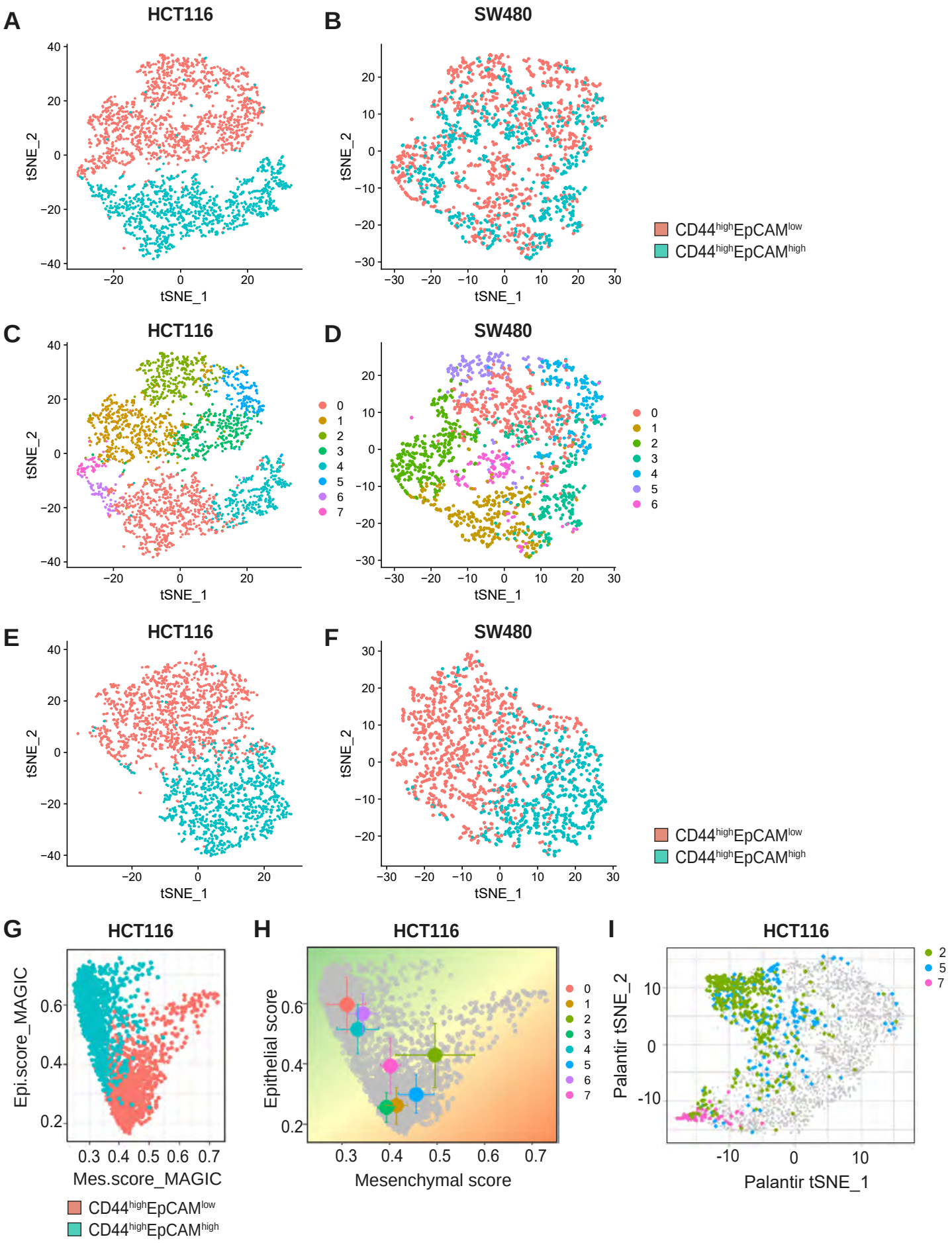

A
HCT116
B
SW480
tSNE_1
tSNE_2
CD44highEpCAMlow
CD44highEpCAMhigh
C
HCT116
D
SW480
0
1
2
3
4
5
6
7
E
HCT116
F
SW480
G
HCT116
Epi.score_MAGIC
Mes.score_MAGIC
H
HCT116
Epithelial score
Mesenchymal score
I
HCT116
Palantir tSNE_1
Palantir tSNE_2

Figure 10 – supplement 1

A. tSNE of SW480 cells based on variable expressed genes across $CD44^{high}EpCAM^{low}$ and $CD44^{high}EpCAM^{high}$ populations.

B. Heatmap grouping the genes derived from the EMT-signature according to their average expression in the unsupervised clusters. This analysis revealed four sets of genes (mes1/2, and epi1/2) expressed in different combinations.

C. Graph indicating alternative trajectories through which partial EMT cells can arise.

D. Classification of single HCT116 cells into categories based on the expression of CMS classifier genes identifying each of the four molecular subtypes.

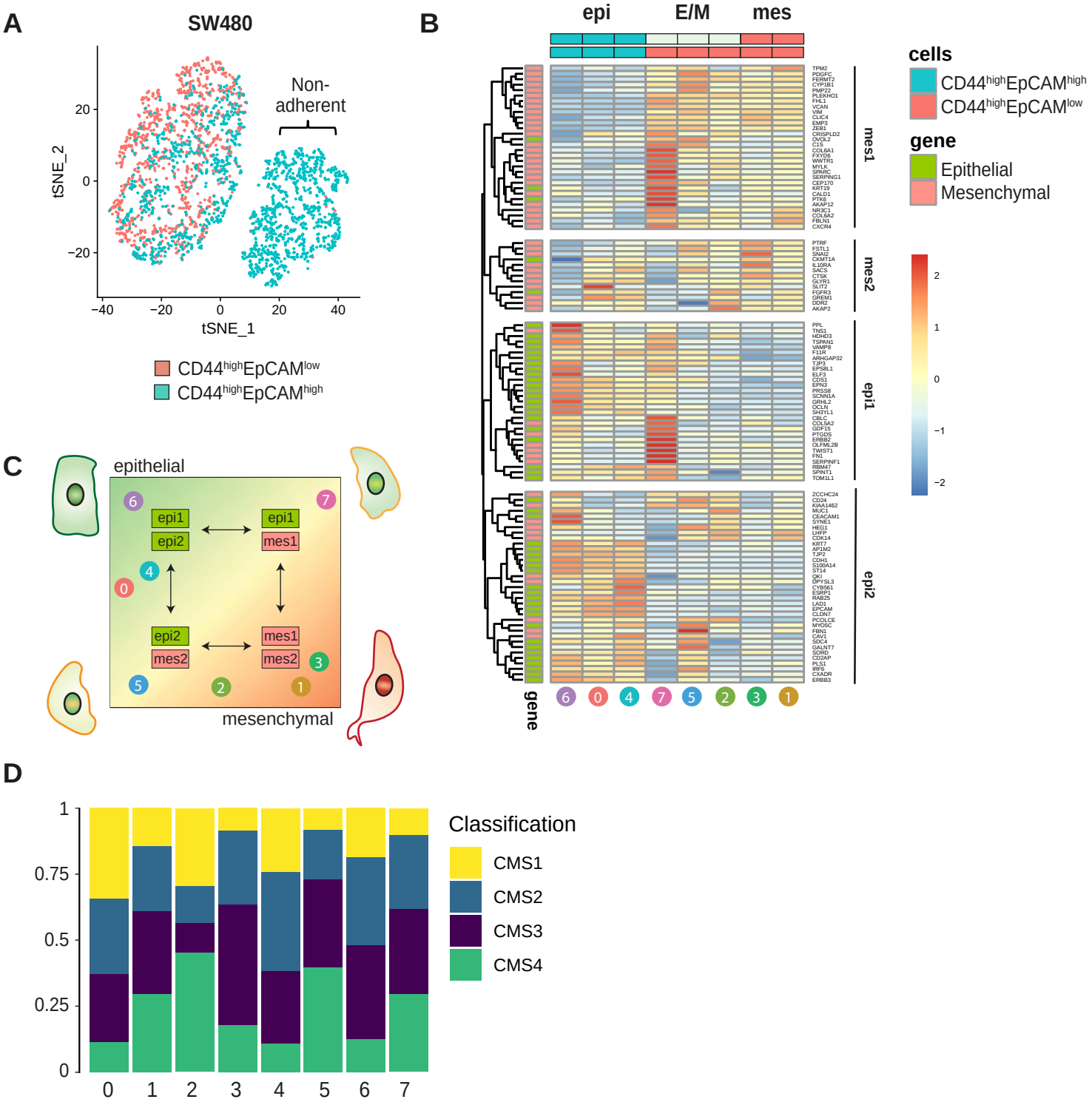



Figure 10 - figure supplement 1