

Metastatic recurrence in colorectal cancer arises from residual EMP1⁺ cells

<https://doi.org/10.1038/s41586-022-05402-9>

Received: 13 December 2021

Accepted: 30 September 2022

Published online: 9 November 2022

 Check for updates

Adrià Cañellas-Socias^{1,2}, Carme Cortina^{1,2}, Xavier Hernando-Momblona^{1,2}, Sergio Palomo-Ponce^{1,2}, Eoghan J. Mulholland³, Gemma Turon¹, Lidia Mateo¹, Sefora Conti⁴, Olga Roman¹, Marta Sevillano^{1,2}, Felipe Slebe¹, Diana Stork¹, Adrià Caballé-Mestres¹, Antonio Berenguer-Llargo¹, Adrián Álvarez-Varela^{1,2}, Nicola Fenderico¹, Laura Novellasedmunt¹, Laura Jiménez-Gracia⁵, Tamara Sipka¹, Lidia Bardia¹, Patricia Lorden⁵, Julien Colombelli¹, Holger Heyn^{5,6}, Xavier Trepas^{4,7,8,9}, Sabine Tejpar¹⁰, Elena Sancho^{1,2}, Daniele V. F. Tauriello^{1,11}, Simon Leedham^{3,12,13}, Camille Stephan-Otto Attolini¹ & Eduard Batlle^{1,2,9}✉

Around 30–40% of patients with colorectal cancer (CRC) undergoing curative resection of the primary tumour will develop metastases in the subsequent years¹. Therapies to prevent disease relapse remain an unmet medical need. Here we uncover the identity and features of the residual tumour cells responsible for CRC relapse. An analysis of single-cell transcriptomes of samples from patients with CRC revealed that the majority of genes associated with a poor prognosis are expressed by a unique tumour cell population that we named high-relapse cells (HRCs). We established a human-like mouse model of microsatellite-stable CRC that undergoes metastatic relapse after surgical resection of the primary tumour. Residual HRCs occult in mouse livers after primary CRC surgery gave rise to multiple cell types over time, including LGR5⁺ stem-like tumour cells^{2–4}, and caused overt metastatic disease. Using *Emp1* (encoding epithelial membrane protein 1) as a marker gene for HRCs, we tracked and selectively eliminated this cell population. Genetic ablation of EMP1^{high} cells prevented metastatic recurrence and mice remained disease-free after surgery. We also found that HRC-rich micrometastases were infiltrated with T cells, yet became progressively immune-excluded during outgrowth. Treatment with neoadjuvant immunotherapy eliminated residual metastatic cells and prevented mice from relapsing after surgery. Together, our findings reveal the cell-state dynamics of residual disease in CRC and anticipate that therapies targeting HRCs may help to avoid metastatic relapse.

Surgical resection of primary CRC effectively cures most patients diagnosed with locoregional disease¹. However, around 5% of patients with AJCC stage I, 15% of stage II and 40% of stage III CRC will develop metastases in the following years¹. We and others have previously shown that the vast majority of genes that predict high risk of disease relapse in CRC are expressed by cells of the tumour microenvironment (TME), particularly by cancer-associated fibroblasts (CAFs)^{5–7}. To further investigate this finding, we mapped the poor-prognosis CRC transcriptome at the single-cell level. Using a large pooled transcriptomic cohort of primary CRC samples (Supplementary Table 1), we identified 2,530 genes that predicted disease relapse (Fig. 1a). Subsequently, the expression of this poor-prognosis gene set was analysed in two independent single-cell RNA-sequencing (scRNA-seq)

CRC datasets that included both tumour epithelial and microenvironment cells⁸ (Fig. 1b–d and Extended Data Fig. 1a–c). Supporting our previous findings, CAFs, endothelial cells and, to a lesser extent, myeloid cells expressed the highest levels of genes associated with poor prognosis (Fig. 1b and Extended Data Fig. 1a). However, detailed analysis of the recurrence gene set in specific cell populations purified from primary CRC samples from patients⁶ revealed that 99 out of the 2,530 genes were upregulated in epithelial tumour cells compared with TME cells (Fig. 1a and Supplementary Table 2). Indeed, these 99 recurrence-associated genes (hereafter, epithelial-specific high-risk gene set (EpiHR)), showed epithelial tumour cell-restricted expression patterns in the Samsung Medical Center (SMC) (Fig. 1d) and Katholieke Universiteit Leuven (KUL) (Extended Data Fig. 1c)

¹Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. ²Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Barcelona, Spain. ³Gastrointestinal Stem Cell Biology Lab, Wellcome Centre Human Genetics, University of Oxford, Oxford, UK. ⁴Institute for Bioengineering of Catalonia (IBEC), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. ⁵CNAG-CRG, Centre for Genomic Regulation, The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. ⁶Universitat Pompeu Fabra (UPF), Barcelona, Spain. ⁷Facultat de Medicina, Universitat de Barcelona, Barcelona, Spain. ⁸Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Barcelona, Spain. ⁹Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. ¹⁰Digestive Oncology, Department of Oncology, Katholieke Universiteit Leuven, Leuven, Belgium. ¹¹Department of Cell Biology, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands. ¹²Translational Gastroenterology Unit, John Radcliffe Hospital, University of Oxford, Oxford, UK. ¹³Oxford National Institute for Health Research Biomedical Research Centre, Oxford, UK.

✉e-mail: eduard.batlle@irbbarcelona.org

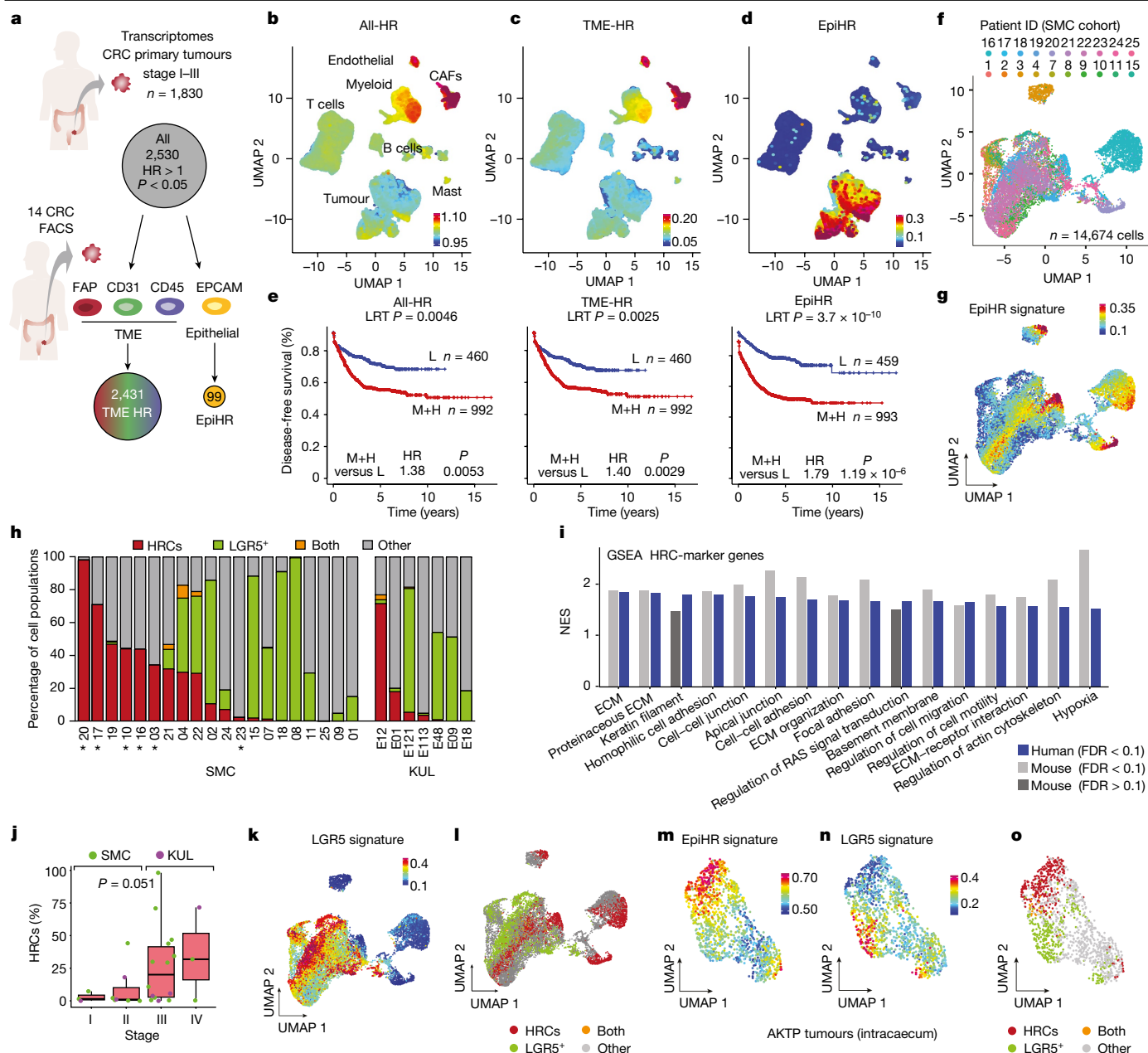


Fig. 1 | Identification of epithelial CRC cells associated with poor prognosis. **a**, Identification of TME-HR and EpiHR signatures in a metacohort of 7 pooled human stage I-III CRC datasets (Supplementary Table 1). **b-d**, UMAP layout of whole tumours (stoma + epithelium cells) belonging to the SMC dataset, coloured by all high-risk (All-HR) relapse (**b**), TME-HR (**c**) and EpiHR (**d**) genes. **e**, Kaplan-Meier survival curves showing relapse-free survival. Statistical analysis was performed using a two-sided Wald test. High (H), medium (M) and low (L) expression levels of All-HR, TME-HR and EpiHR gene signatures; LRT, likelihood-ratio test. **f,g**, UMAP layout of 14,674 CRC tumour cells (SMC cohort) coloured by patient ID (**f**) and by the expression of the EpiHR signature (**g**). **h**, The subpopulation composition of each patient. Patient ID is detailed. Patients with low WNT-signature scores are indicated by an asterisk. **i**, Selected Hallmark, GOSLIM and KEGG gene signatures enriched in HRCs compared with the rest of tumour cells in human and mouse CRC

samples. NES, normalized enrichment score. ECM, extracellular matrix. **j**, The proportion of HRCs in each clinical stage. The box plots show the median (centre line) first and third quartiles (box limits), and a maximum of $1.5 \times$ the interquartile range (whiskers). From left to right, $n = 3$, $n = 7$, $n = 14$ and $n = 3$ patients. Statistical analysis was performed using two-sided Kruskal-Wallis tests. **k**, UMAP analysis of tumour cells coloured by expression of the *LGR5* signature. **l**, UMAP analysis of tumour cells labelled according to their classification as HRCs, *LGR5*⁺, double-positive or other cells. **m-o**, Primary tumours were generated in the caecum of C57BL/6J mice by injecting syngeneic AKTP MTOs. UMAP plots depicting eGFP⁺ tumour cells dissociated from primary tumours coloured by expression levels of EpiHR signature (**m**), *LGR5* signature (**n**) and their classification as HRCs, *LGR5*⁺, double-positive or other cells (**o**). Colour scales in **b-d,g,k,m-o** indicate expression levels of the indicated gene signatures.

scRNA-seq cohorts. EpiHR expression levels predicted recurrence with an accuracy equivalent to the subset of TME poor-prognosis genes (Fig. 1e). In a multivariate analysis including the two signatures and clinical variables, the TME high risk (TME-HR) and EpiHR gene sets were independent prognostic factors (EpiHR: hazard

ratio (HR) (+1 s.d.) = 2.26, $P = 1.2 \times 10^{-7}$; TME-HR: HR (+1 s.d.) = 1.74, $P = 8 \times 10^{-4}$). The EpiHR signature was significantly associated with right-side colon cancer and to AJCC stages III-IV (Extended Data Fig. 1d). Moreover, it stratified CRC patients within each consensus molecular subtype⁹ into high and low risk of relapse (Extended Data

Fig. 1e). Thus, the EpiHR gene set encodes determinants of disease relapse with epithelial tumour-cell-specific expression.

Identification of EpiHR⁺ tumour cells

Representation of tumour epithelial CRC cells using uniform manifold approximation and projections (UMAP) analysis showed that 18 out of 27 CRCs contained cells labelled with the EpiHR gene set in proportions ranging from 1.4% to 98.1% (Fig. 1f–h and Extended Data Fig. 1f,g). We named this tumour cell population high-relapse cells (HRCs). An analysis of the EpiHR signature revealed a large number of highly correlated genes exhibiting overlapping expression in HRCs (gene clusters 1 and 3 in Extended Data Fig. 1h,i and Supplementary Table 2). The EpiHR contains only epithelial-specific genes, reflecting only a fraction of the HRC transcriptome. We further identified a core gene expression program upregulated in HRCs of most tumour samples (coreHRC; Extended Data Fig. 2a,b and Supplementary Table 2). HRCs belonging to different CRCs displayed a shared enrichment pattern of annotated gene sets, implying that they co-opt a similar phenotype and have equivalent functions (Extended Data Fig. 2c,d). Differential expression analysis of HRCs versus non-HRCs revealed prominent enrichment in genes related to hypoxia, cell-to-cell adhesion, extracellular matrix, actin cytoskeleton and regulation of cell migration (Fig. 1i). Among them (Supplementary Table 2), the collagen-sensing receptor DDR1, the integrins $\alpha 2$, $\alpha 3$ and $\beta 4$, and the protease PLAUR have been repeatedly associated with tumour cell invasion, extravasation and metastasis in multiple cancer types. Incidentally, we also discovered that the signature of basal-like pancreatic cancer cells¹⁰ marked both human and mouse HRCs, suggesting that they are akin to the most aggressive subtype of pancreatic cancer (Extended Data Fig. 2e and Supplementary Table 3). Further reinforcing our observations, most scRNA-seq samples containing abundant EpiHR⁺ cells were stage III and IV CRCs (Fig. 1j; $P = 0.051$ stage I + II versus stage III + IV).

Widespread evidence has demonstrated that CRC growth is driven by a subset of LGR5⁺ stem-cell-like tumour cells^{2–4}. However, our analyses revealed that HRCs represent a distinct cell population, shown by their mutually exclusive distribution in UMAP analyses (Fig. 1k–l and Extended Data Fig. 2f,g). Only one tumour sample exhibited a significant number of HRCs co-expressing the LGR5 signature (7.79% sample SMC04; Fig. 1h), whereas five others included a minimal fraction (<3%) of cells marked by both *LGR5* and the HRC gene programs. The expression patterns of several WNT/intestinal stem cell marker genes confirmed that HRCs were not LGR5⁺ stem-like tumour cells (Extended Data Fig. 2h–o). Indeed, a subset of CRCs exhibited marginal WNT target gene expression levels (Extended Data Fig. 2p,q) yet contained HRCs (Fig. 1h (asterisk)). Furthermore, HRCs represented a subset of cells in both epithelial intrinsic consensus molecular CRC subtypes iCMS2 and iCMS3¹¹ (Extended Data Fig. 2r).

We previously described that mice bearing mutations in *Apc*, *Kras*, *Tgfr2* and *Trp53* (AKTP mice) in LGR5⁺ intestinal stem cells (ISCs) develop human-like metastatic CRCs¹². We analysed CRCs generated by implantation of AKTP mouse tumour organoids (MTOs) into the caecum of c57BL/6 mice using scRNA-seq (Fig. 1m–o). Mirroring the observations in human tumour samples, we found that mouse CRCs contained abundant HRCs and that this population did not express the LGR5⁺ ISC-like expression program (Fig. 1m–o and Extended Data Fig. 2s). HRCs were enriched in similar gene categories in both species, including the pancreatic basal-like signature, implying functional equivalence (Fig. 1i and Extended Data Fig. 2s,t). Bulk RNA-seq analysis of our MTO biobank¹² showed elevated EpiHR and coreHRC gene signatures expression in AKTP MTOs derived from metastases compared with those established from primary CRCs (Extended Data Fig. 2u).

Dynamics of metastatic cell states

We developed a new mouse model of metastatic relapse that enabled us to investigate the contribution of HRCs to metastatic recurrence. In brief, we innovated classical needle-based orthotopic injections by relocating them to the tip of the caecum, which enabled complete surgical excision of singular invasive CRCs (Fig. 2a, Extended Data Fig. 3a–c and Supplementary Video 1). Dual eGFP–luciferase-labelled AKTP MTOs grew rapidly in the caecum of c57BL/6 mice, colonized the adjacent mucosa and generated invasive cancers (Extended Data Fig. 3a,b). Bioluminescence imaging (BLI) revealed that mice remained free of primary disease after surgical resection yet, over the following days, they relapsed in the form of liver metastases (Fig. 2b,c and Extended Data Fig. 3d,e). Occasionally, we also observed metastases in the mesenteric lymph nodes, lungs, peritoneum and diaphragm (Extended Data Fig. 3f–h). Primary tumour resection shortly after implantation cured all mice, whereas surgery at later points resulted in increased proportions of mice developing metastatic recurrences (Fig. 2b). In early primary tumour resection experiments (days 11–15), faint bioluminescence could be detected ex vivo in some livers immediately after surgery, implying the presence of residual disseminated tumour cells (DTCs). Light-sheet 3D fluorescence imaging revealed 3-to-10-cell micrometastases at the time of resection (Fig. 2d). On the basis of these observations, we established that 4–5 weeks after implantation is the optimal timepoint to enable a complete surgical resection of the primary CRC. We also developed a CRC relapse model based on the implantation of AKP MTOs. These triple-mutant CRCs exhibited delayed kinetics of metastatic recurrence after surgery (Extended Data Fig. 3i).

We next sought to profile tumour cells along the process of relapse. Isolation of residual tumour cells from large organs has been a major hurdle in cancer research¹³. We devised a tissue dissociation strategy that enriched for residual tumour cells from whole-liver samples (Extended Data Fig. 3j). In brief, during tissue preparation for fluorescence-activated single cell sorting (FACS), the vast majority of luciferase⁺ tumour cells were retained in 100 μ m filters after mild enzymatic digestions, whereas most parenchymal liver cells flowed through in these conditions (Extended Data Fig. 3k,l). By redigesting cells retained in the filter, we obtained a 400-fold eGFP–luciferase⁺ enriched metastatic tumour cell preparation (Extended Data Fig. 3m). This step enabled the purification of residual tumour cells from individual livers exhibiting absent or very low ex vivo bioluminescence (Extended Data Fig. 3n). Using this approach, we profiled by Smart-seq2 scRNA-seq 900 eGFP⁺ tumour cells derived from livers collected at different time points after implantation of MTOs as well as from their corresponding primary CRCs (Fig. 2e). We confirmed the presence of micrometastases, small metastases or macrometastases on the basis of bioluminescence measurements in the resected livers (Extended Data Fig. 3n).

A UMAP representation showed that tumour cells from primary tumours and metastases overlapped to a large extent (Fig. 2e,f and Extended Data Fig. 4). Hierarchical clustering analysis identified six cell clusters (Fig. 2g and Extended Data Fig. 4b,c). Cluster 0 included cells that expressed elevated levels of proliferation and biosynthesis-encoding genes (Fig. 2g and Extended Data Fig. 4b,c). LGR5⁺ ISC-like tumour cells occupied two different clusters depending on high (cluster 1) and low (cluster 2) expression levels of the Ki-67 proliferative signature¹⁴ (Fig. 2g and Extended Data Fig. 4c). Tumour cells of cluster 3 upregulated the differentiation marker *Krt20*. Clusters 4 and 5 were largely enriched in HRCs. Some cells in cluster 4 expressed *Krt20* suggesting that HRCs can also undergo differentiation (Fig. 2g and Extended Data Fig. 4b,c). Quantification of cell types revealed a dynamic distribution of cell populations across the metastatic relapse process (Fig. 2h). Primary CRCs and macrometastases exhibited similar distribution of cell populations, including proliferative cells, LGR5⁺ ISC-like cells, KRT20⁺ differentiated tumour cells and HRCs, most of which were also KRT20⁺ (Fig. 2h). By contrast, micrometastases were largely enriched in undifferentiated

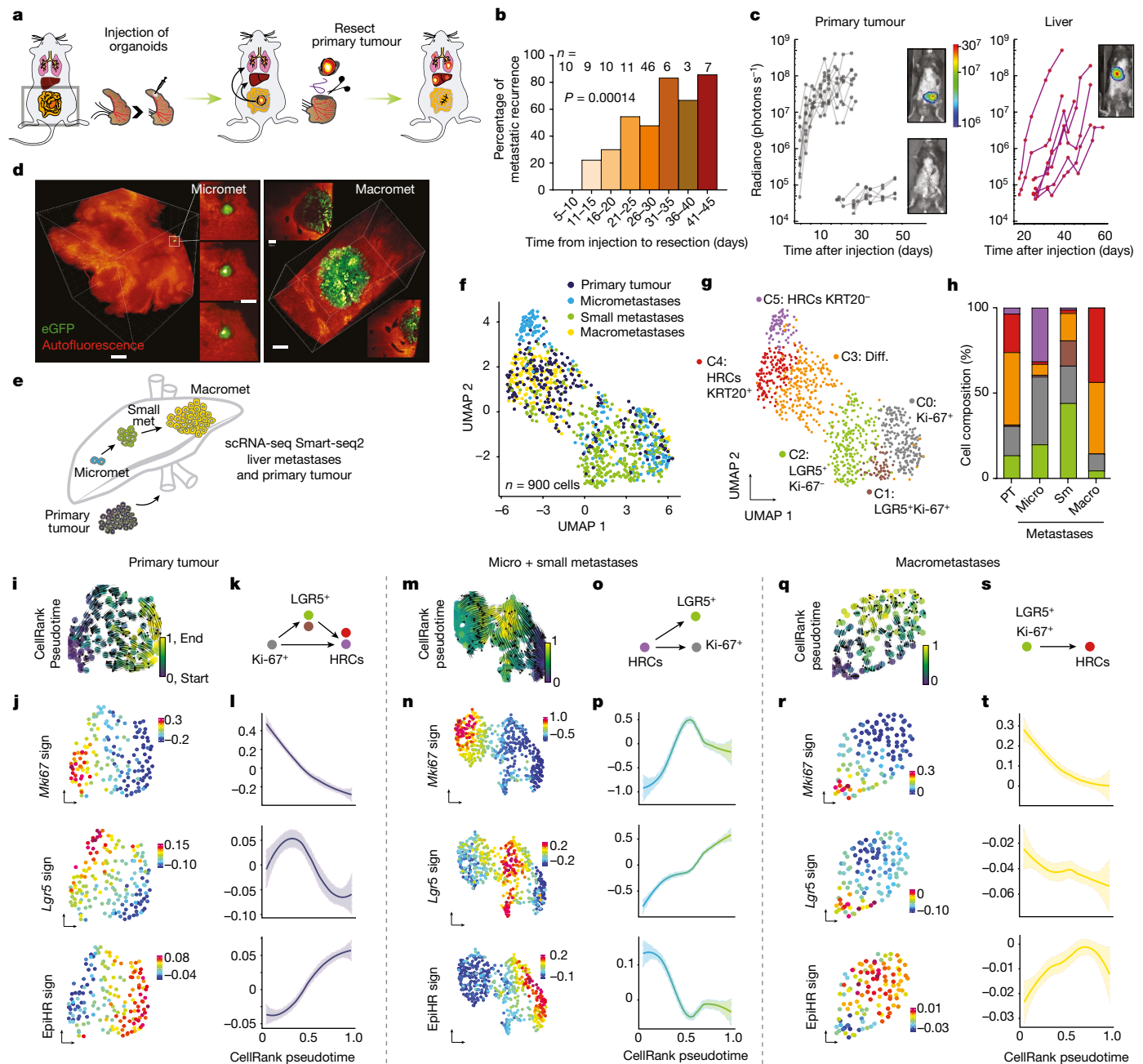


Fig. 2 | Spatiotemporal dynamics of CRC metastases resolved by scRNA-seq. **a**, Schematic of the mouse model of CRC metastatic relapse developed in this study. **b**, The percentage of metastatic recurrence depending on time to primary tumour resection. The number of mice is shown above the plot. The P value was calculated using a generalized linear model. **c**, Intravital BLI quantification (photons s^{-1}) of a representative experiment. The grey points and lines represent BLI in the lower thorax of individual mice and the purple points represent BLI in the liver. Representative images of BLI in the same mouse before surgery, after surgery and after liver metastasis formation are shown. Colour scale indicates intensity of radiance. **d**, Representative images of whole livers containing GFP-expressing tumour cells obtained using light-sheet microscopy. Scale bars, 300 μm (left, main), 50 μm (left and right insets) and 100 μm (right, main). Macromet, macrometastases; micromet, micrometastases. **e**, Illustration of the longitudinal single-cell RNA-expression analysis of tumour cells along the metastatic cascade. Met, metastases. **f**, **g**, UMAP layout of 900 tumour cells isolated from 7 different mice coloured

by metastatic stage (**f**) and Seurat clusters (**g**). Diff., differentiated. **h**, Seurat cluster composition by sample stage. PT, primary tumour; Sm, small. **i**, **m**, **q**, Vector fields representing RNA velocity projected onto UMAPs of primary tumours (**i**), micro + small metastases (**m**) and macrometastases (**q**), coloured by the pseudotime estimated for each cell with scVelo. **j**, **n**, **r**, UMAPs with cells separated in primary tumours (**j**), micro + small metastases (**n**) and macrometastases (**r**) and coloured by gene expression of *Mki67*, *Lgr5* and *EpiHR* gene signatures (sign). **k**, **o**, **s**, Schematics showing distinct hierarchical behaviour during the different stages of metastasis progression in primary tumours (**k**), micro + small metastases (**o**) and large metastases (**s**). **l**, **p**, **t**, Smoothed *Mki67*, *Lgr5* and *EpiHR* gene signature expression trends fitted with generalized additive models as a function of pseudotime in primary tumours (**l**), micro + small metastases (**p**) and large metastases (**t**). Colour scales in **i**, **m**, **q** indicate start to end of trajectory in pseudotime. Colour scales in **j**, **n**, **r** indicate, expression levels of the indicated gene signatures.

(KRT20⁺) HRCs and also contained abundant proliferative cells (Fig. 2h). Small metastases were mainly formed by LGR5⁺ ISC-like cells in both Ki-67⁺ and Ki-67⁻ states and contained fewer HRCs than micrometastases (Fig. 2h and Extended Data Fig. 4d).

The trajectory-inference algorithm CellRank¹⁵ predicted diverse hierarchical organizations during metastatic progression. In primary CRCs, proliferative LGR5⁺ tumour cells gave rise to LGR5⁺ cells and HRCs (Fig. 2i–l). In macrometastases, the apex of the hierarchy was occupied by proliferative LGR5⁺ ISC-like cells, which generated HRCs over time (Fig. 2q–t). By contrast, the algorithm prognosticated that the cells that initiate the cellular hierarchy of micrometastases corresponded to undifferentiated HRCs in cluster 5 (Fig. 2m–p and Extended Data Fig. 4e). This cell population gave rise to LGR5⁺ ISC-like and proliferative tumour cell progeny (Fig. 2m–p), which are abundant in small metastases (Fig. 2h and Extended Data Fig. 4d).

We made equivalent observations in AKP CRCs (Extended Data Fig. 4f–p). Abundant HRCs characterized liver AKP lesions profiled at early time points (day 35 after implantation), whereas micrometastases collected later (day 70) contained a higher proportion of LGR5⁺ cells (Extended Data Fig. 4f–k). Micrometastatic LGR5⁺ cells in the AKP model expressed high levels of the signature of latent MEX3A⁺ cells (Extended Data Fig. 4l–n), which is in agreement with our previous study on the specification of this population in triple-mutant CRCs¹⁶. CellRank also predicted that HRCs are the origin of micrometastasis in AKP CRCs (Extended Data Fig. 4o,p).

Characterization of HRC features

A comparison of human and mouse scRNA-seq datasets revealed that a large subset of core genes was expressed consistently by HRCs of the two species (Extended Data Fig. 5a). Among them, we focused on *EMPI* because it was expressed at high levels in HRCs and exhibited a great degree of overlap with the expression of the EpiHR gene set in both human (Extended Data Fig. 5b,c) and AKTP mouse CRCs (Fig. 3a and Extended Data Fig. 5d). Echoing our results with the EpiHR signature, CellRank predicted that the cell origin of metastatic relapse in AKTP (Extended Data Fig. 5e–g) and AKP (Extended Data Fig. 5h,i) tumours expressed elevated *Emp1* levels.

Thus, we leveraged *Emp1* to track HRCs during disease relapse. To this end, we knocked-in an inducible caspase-9-tdTomato (iCT) cassette² into the *Emp1* locus of AKTP MTOs using CRISPR–Cas9 (Fig. 3b). Inspection of knockin MTOs revealed high tdTomato (TOM) expression in a subset of tumour cells (Fig. 3b). We next inoculated *Emp1*-iCT AKTP MTOs into the caecum of c57BL/6 mice. TOM expression in dissociated epithelial tumour cells measured by flow cytometry revealed heterogeneity in *Emp1* expression (Extended Data Fig. 5j). TOM^{high} cells purified by FACS showed substantial upregulation of *Emp1* expression (Extended Data Fig. 5k). By contrast, TOM^{low} cells were characterized by the expression of ISC-specific genes such as *Lgr5* and *Smoc2* (Extended Data Fig. 5k). Gene expression profiling confirmed elevated levels of EpiHR and coreHRC signatures in EMP1–TOM^{high} cells, whereas the WNT/LGR5⁺ ISC program was downregulated in these cells (Fig. 3c,d and Extended Data Fig. 5l).

Inspection of tissue sections evidenced that cancer cells invading the muscular layer were strongly labelled by the fluorescent reporter (Fig. 3e). In particular, isolated tumour buds and larger clusters in contact with the stroma at the edges on invasion fronts exhibited the highest TOM expression (Fig. 3e (inset) and Extended Data Fig. 6a–d). EMP1–TOM^{high} cell clusters were often found in proximity to peripheral blood vessels in mouse primary CRCs, suggesting a connection with haematogenous or lymphatic dissemination (Extended Data Fig. 6e). CRCs generated by implantation of AKP MTOs in the caecum also exhibited EMP1–TOM^{high} invasion fronts and tumour buds that overexpressed HRC-marker genes (Extended Data Fig. 6f–h).

We also examined livers of mice bearing AKTP tumours at various time points after orthotopic MTO implantation. Micrometastatic lesions

trapped within portal veins and liver sinusoids were populated entirely by EMP1–TOM^{high} cells (Fig. 3f,g and Extended Data Fig. 6i,j). Fitting the scRNA-seq analyses, we identified two EMP1–TOM⁺ subsets; one expressed KRT20 and was located mainly in the tumour cores, whereas the other was positioned at invasion fronts, lacked KRT20 and expressed the highest levels of TOM reporter (Extended Data Fig. 6k). Liver EMP1–TOM^{high} micrometastases were also KRT20 negative (Extended Data Fig. 6l). Notably, EMP1–TOM^{high} tumour buds and micrometastases were labelled with EPCAM and E-cadherin, implying that they retained an epithelial organization (Extended Data Figs. 6 and 7a). Analysis of epithelial-to-mesenchymal transition (EMT) master transcription factors showed equivalent expression levels in EMP1–TOM^{high} and EMP1–TOM^{low} cells (Extended Data Fig. 7b). scRNA-seq analysis also supported the lack of canonical EMT markers in human and mouse HRCs (Extended Data Fig. 7c,d). However, we noticed that the coreHRC signature genes *Lama3*, *Lamc2*, *Itga2* and *Plaur* belong to a recently characterized partial EMT gene expression module¹⁷. HRCs in mouse primary CRCs and micrometastases upregulated this partial EMT signature (Extended Data Fig. 4e). Notably, LAMC2 has previously been identified as a specific marker for tumour budding in multiple tumour types, including CRC¹⁸, and we confirmed that its expression is correlated with *EMPI* mRNA in samples from patients with CRC (Extended Data Fig. 7e–g). KRT17—a widely used marker of the basal pancreatic cancer subtype¹⁹—also marked EMP1^{high} invasion fronts and tumour buds (Extended Data Fig. 7f,g). Gene set enrichment analysis (GSEA) also revealed homophilic cell adhesion and apical junctions as central features of HRCs (Fig. 1i). Fittingly, *Emp1* encodes a component of the tight junctions^{20,21} and there were multiple other constituents of adherent junction, tight junction and desmosome complexes upregulated in the core HRC program, such as *Pcdh1*, *Dsc2*, *Cldn4* and *Jup* (Extended Data Fig. 7g and Supplementary Table 2). The latter encodes plakoglobin, which mediates the circulation of tumour cells as clusters and confers enhanced metastatic ability to breast cancer cells²².

To further investigate the relationship between *Emp1* and *Lgr5* expression, we engineered AKTP MTOs bearing both *Emp1*-iCT and *Lgr5*-eGFP knockin reporter cassettes (Extended Data Fig. 8a–f). Confocal imaging of dual-labelled MTOs showed a mutually exclusive expression pattern of the two reporters (Extended Data Fig. 8a) and quantitative PCR with reverse transcription (RT–qPCR) analysis confirmed the upregulation of *Emp1* and *Lgr5* in sorted TOM⁺ and eGFP⁺ cells, respectively (Extended Data Fig. 8b). Primary CRCs generated from inoculation of dual-labelled MTOs into the caecum also exhibited a mutually exclusive expression pattern of EMP1–TOM and LGR5–eGFP reporters (Fig. 3h and Extended Data Fig. 8c,d). EMP1–TOM^{high} cells were substantially enriched at tumour buds, which, by contrast, contained few LGR5–eGFP⁺ cells (Fig. 3h,i). RNA fluorescence in situ hybridization (FISH) analysis in human samples of patients with CRC also showed that *EMPI* expression was elevated at tumour invasion fronts, whereas *LGR5* marked the tumour cores in most cases (examples in Extended Data Fig. 8g–p).

Finally, we analysed the livers of mice bearing primary CRCs. LGR5–eGFP fluorescence was absent in DTCs and micrometastases, but was progressively gained during metastatic outgrowth, in a marked anti-thetic pattern to EMP1–TOM expression (Fig. 3j,k and Extended Data Fig. 8e,f). Together with the CellRank bioinformatic predictions, these observations suggest that HRCs are endowed with the ability to migrate and disseminate to foreign organs where they initiate metastatic outgrowth and subsequently give rise to non-HRC populations.

Determinants of the HRC population

The transcription factor YAP opposes the activity of the WNT pathway^{23,24} and a YAP-driven gene program has been associated with tumour cell plasticity^{25,26}, regeneration²⁷ and metastasis formation in CRC²⁸. Moreover, YAP promotes the conversion of LGR5⁺ CRCs cells to a fetal intestine-like progenitor state during chemotherapy^{16,25,29,30}. Extended

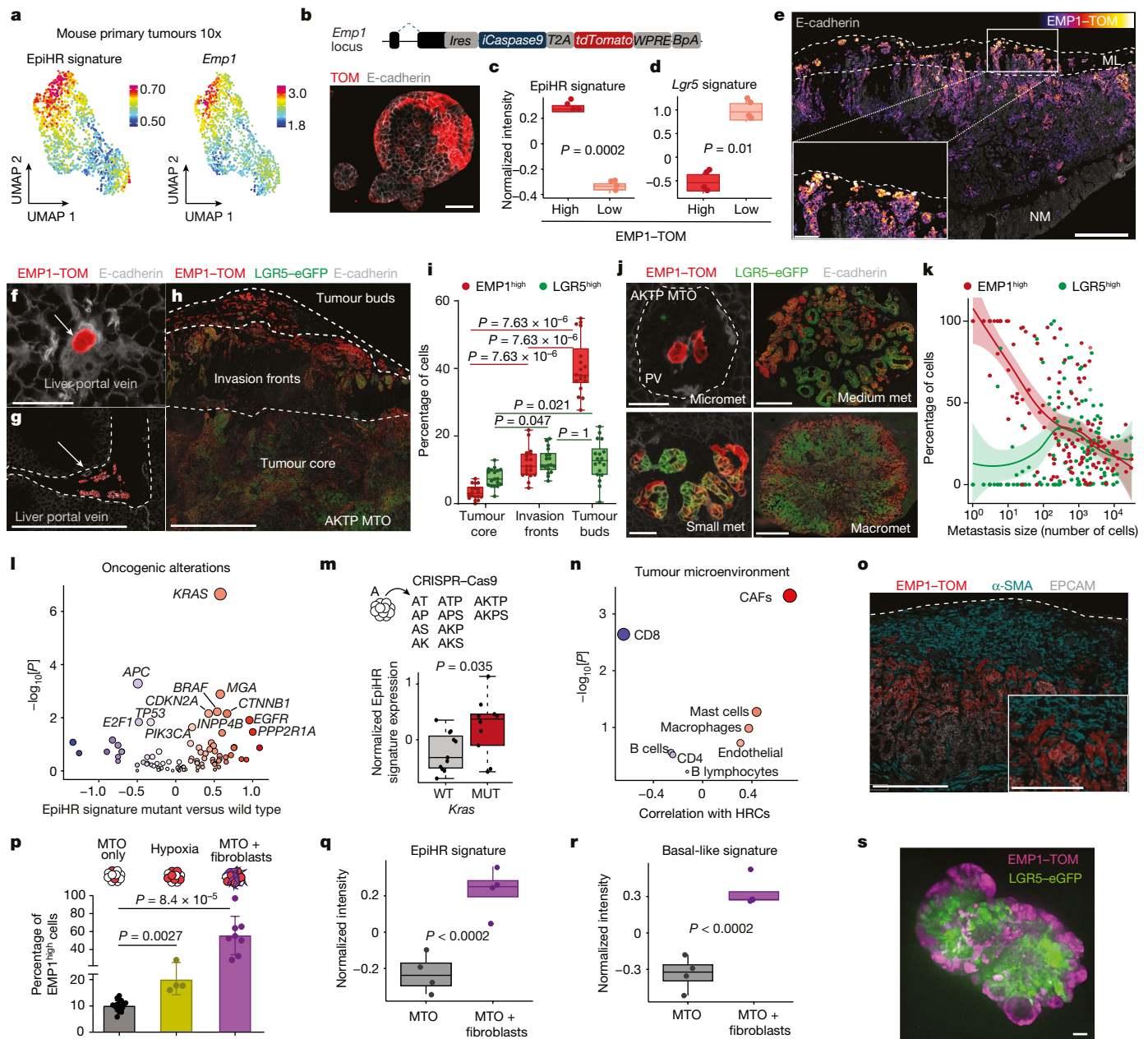


Fig. 3 | EMP1 marks cells enriched in invasion fronts and micrometastases.

a, UMAP showing EpiHR and *Emp1* expression in mouse primary tumours. Colour scale indicates expression levels. **b**, TOM and E-cadherin immunostaining in *Emp1*-iCT AKTP MTOs. Scale bar, 50 μ m. **c, d**, Expression (normalized intensity) of EpiHR (**c**) and *Lgr5* (**d**) signatures in EMP1^{high} and EMP1^{low} cells 4 weeks after implantation. $n = 4$ mice per condition. ROAST-GSA adjusted P values are shown. **e**, Representative images of EMP1-TOM⁺ cells in primary CRCs. The dashed lines encompass tumour buds and invasion fronts. ML, muscle layer; NM, normal mucosa. Scale bars, 500 μ m (main image), 50 μ m (inset). **f–h**, Representative images of EMP1-TOM⁺ and LGR5-eGFP⁺ cells in primary CRCs (**f**) and liver metastases (**g, h**). Scale bars, 500 μ m (**g, h**), 50 μ m (**f**). The dashed lines label the portal vein (**g**), and encompass the tumour buds, invasion fronts and tumour core (**h**). **i**, The percentage of LGR5-eGFP^{high} and EMP1-TOM^{high} cells in different CRC regions. The whiskers show the maximum and minimum values. $n = 855,330$ cells from 18 mice. Statistical analysis was performed using two-sided Wilcoxon signed-rank tests. **j**, Examples of liver metastases of increasing size. Scale bars, 50 μ m (micrometastases and small metastases), 250 μ m (medium metastases) and 500 μ m (macrometastases). **k**, The percentage of EMP1-TOM^{high} and LGR5-eGFP^{high} cells per metastasis size.

$n = 318,276$ cells from 137 liver metastases from 17 different mice. LOESS model with a 95% confidence interval. **l**, EpiHR levels versus driver mutations. The difference of median values (x axis) versus P values (Wilcoxon rank-sum test) (y axis) are shown. **m**, Normalized intensity of EpiHR signature expression in CTOs of different genotypes. $n = 6$ (wild-type (WT) *Kras*) and $n = 5$ (*Kras*^{G12D}) CTOs; 2 technical replicates per genotype. The P value was calculated using a two-sided t -test. **n**, The association between HRC frequency and TME cell populations in patients from the SMC cohort. Pearson correlation coefficients and P values are shown. **o**, Representative patterns of TOM, α -SMA and EPCAM in primary CRCs. The dashed lines delimit the caecum. Scale bars, 250 μ m (main) and 125 μ m (inset). **p**, The percentage of EMP1-TOM^{high} cells in the indicated conditions. Data are mean $\pm$ s.d. $n = 17$ (control), $n = 4$ (hypoxia) and $n = 8$ (fibroblast co-culture). Statistical analysis was performed using two-sided Wilcoxon rank-sum tests. **q, r**, EpiHR expression (normalized intensity) (**q**) or basal-like (**r**) signatures. $n = 4$ biological replicates. ROAST-GSA adjusted P values are shown. **s**, Representative confocal images of AKTP MTOs co-cultured with colon fibroblasts. Scale bar, 50 μ m. For the box plots in **c, d, i, m, q** and **r**, the median (centre line) and first and third quartiles (box limits) are shown; the whiskers in **m, q** and **r** show a maximum of $1.5 \times$ the interquartile range.

Data Fig. 9a,b shows the expression of the top 50 upregulated fetal intestinal progenitor genes from ref.³¹, and the broadly used YAP_22 target gene signature³². As a reference, we also show the expression of the top 50 HRC core genes (Extended Data Fig. 9c). Only 3 out of 22 YAP core signature genes were upregulated in HRCs of patients with CRC, whereas most bona fide YAP-target genes, including *CTGF* and *CYR61*, were not (Extended Data Fig. 9a). HRCs upregulated only a few canonical markers of the regenerative fetal-like intestinal population, but the majority of them were either not expressed or were expressed at low levels by HRCs (Extended Data Fig. 9b). We also found little overlap between coreHRC and YAP_22 or fetal intestinal progenitor signatures (Extended Data Fig. 9d). EMP1-TOM⁺ cells isolated from mouse primary CRCs were not enriched in YAP-target genes (Extended Data Fig. 9e,f). There are therefore only marginal similarities between YAP⁺/fetal intestinal progenitors and the HRC state. To obtain further insights into this question, we knocked down YAP levels using short hairpin RNAs in MTOs or blocked YAP-driven transcription by overexpressing an inducible dominant-negative TEAD transcription factor that inhibits both YAP and TAZ activity³³ (Extended Data Fig. 9g–m). These genetic manipulations decreased the levels of the canonical YAP-target genes *Ctgf* and *Cyr61*, but did not affect the expression of core HRC genes *Emp1* or *Lamc2* in MTOs, or the abundance of EMP1-TOM^{high} cells (Extended Data Fig. 9g–m). As a control for these experiments, we used chemotherapy (folfiri), which upregulated the YAP_22 signature, including *Ctgf* and *Cyr61*, and the intestinal fetal progenitor program, but did not induce the HRC state (Extended Data Fig. 9n,o).

Association analyses in the TCGA COAD cohort revealed a strong correlation between activating *KRAS* mutations and EpiHR expression levels (Fig. 3l). Using a series of mouse organoids with compound mutations in main CRC driver genes engineered by CRISPR–Cas9 (CRISPR-derived tumour organoids (CTOs)), we confirmed that genotypes containing *Kras*^{G12D} mutations exhibited upregulation of coreHRC and EpiHR gene signatures (Fig. 3m and Extended Data Fig. 10a,b). We next searched for associations between HRCs and TME composition using the scRNA-seq dataset of patients with CRC. This analysis revealed a direct correlation between CAF and HRC abundance (Fig. 3n). Consistent with this finding, α -SMA⁺ CAFs surrounded EMP1-TOM^{high} cells in primary AKTP CRCs (Fig. 3o). Co-culture of AKTP MTOs with CAFs augmented sixfold EMP1-TOM⁺ cell numbers (Fig. 3p), but *YAP* knockdown did not block this effect (Extended Data Fig. 10c). Expression profiling of MTO cells isolated from co-cultures demonstrated upregulation of EpiHR, coreHRC and basal pancreatic cancer signature levels (Fig. 3q,r and Extended Data Fig. 10d). Furthermore, an arrangement of tumour cell types reminiscent of the in vivo organization was evident—LGR5–eGFP cells were positioned at the organoid centre, whereas EMP1-TOM⁺ cells relocated to the boundaries in contact with the fibroblast population (Fig. 3s and Extended Data Fig. 10e). Fibroblasts surrounding MTOs appeared to be activated, as shown by the expression of α -SMA (Extended Data Fig. 10f). We also noticed that some organoids gained expression of the pancreatic cancer basal cell marker KRT17 (Extended Data Fig. 10g).

Relapse by residual EMP1⁺ cells

We next leveraged the inducible caspase-9 cassette inserted in the *Emp1* locus to perform cell-ablation experiments in intact tumours^{2,34} (Fig. 4a,b). Inoculation of mice with AP20187 (DIM) dimerized the chimeric caspase-9 expressed under the *Emp1* locus and specifically killed cells expressing the highest levels of EMP1-TOM reporter (Fig. 4c–e). This effect was reversible, shown by a slow but progressive recovery of the EMP1-TOM^{high} cell population at invasion fronts after cessation of treatment with DIM (Extended Data Fig. 10h–j). DIM was administered only during primary tumour growth, but treatment ceased the day before primary CRC resection (Fig. 4b). Macrometastases were not yet present when DIM treatment finished; thus, this experimental setting aimed to ablate EMP1^{high} cells in the primary tumour and possibly in incipient metastatic lesions. Notably, although DIM treatment did not affect primary

tumour growth (Fig. 4f), the majority of mice showed no signs of liver and lung metastatic recurrence and were disease-free at the experimental end points (Fig. 4g,h and Extended Data Fig. 10k). We made equivalent observations in AKTP CRC growing in nude mice, therefore ruling out that the genetic cell ablation strategy impaired metastatic progression through activation of the adaptive immune system (Extended Data Fig. 10l–n). When EMP1^{high} cell ablation started 1 week after primary CRC resection, we observed no changes in metastatic progression and all of the mice suffered metastatic relapse (Fig. 4i,j). EMP1^{high} cell ablation did not halt metastasis formation when MTOs were directly inoculated in the liver through the spleen (Fig. 4k,l). Reinforcing this observation, we found that intrasplenic inoculation of EMP1^{high} cells isolated from MTOs generated more and larger liver metastases compared with either LGR5^{high} cells or EMP1^{low}LGR5^{low} cells, but the difference was not substantial, suggesting that all of these tumour cell populations had metastasis-initiating ability in this assay (Extended Data Fig. 10o–r). Moreover, metastases produced by these three cell populations contained the other tumour cell types owing to extensive cell plasticity (Extended Data Fig. 10s). Overall, these data demonstrate that EMP1^{high} HRCs drive metastatic relapse after primary CRC resection, yet they are dispensable after metastatic seeding is completed. This model is further supported by the observation that HRC ablation just after surgery and during metastatic outgrowth decreased the number of small metastases significantly, but did not modify the frequency of large metastases (Extended Data Fig. 10t–w).

Despite the slow-growing metastatic lesions generated by the AKP CRCs, HRC ablation before primary tumour surgery also prevented metastatic recurrence in these triple-mutant models (Extended Data Fig. 11a–c). Moreover, we demonstrated that HRCs mediated recurrence in CRCs with mutations in *Smad4* (AKPS) (Extended Data Fig. 11d–f). To extend our observations beyond the colon-to-liver metastasis axis, we implanted AKTP MTOs into the rectum, which generated invasive rectal cancers that, as in humans, metastasized preferentially to the lungs (Extended Data Fig. 11g–k). Inspection of lung metastases revealed that, akin to liver metastases, micrometastases were mostly EMP1^{high}, whereas larger metastases decreased the percentage of EMP1^{high} cells (Extended Data Fig. 11i). Although we could not surgically resect mouse rectal tumours, ablation of EMP1^{high} cells caused a 20-fold reduction in lung metastasis burden in this model (Extended Data Fig. 11k).

Using a diphtheria toxin receptor (DTR)-based ablation strategy in CRC models, it was previously shown that LGR5⁺ CRC cells are necessary for liver metastasis formation³. To assess the role of LGR5⁺ tumour cells in our relapse models, we knocked-in a DTR cassette into the *Lgr5* locus of AKTP MTOs (Extended Data Fig. 11l,m). Inoculation of this MTO line into the caecum further validated our previous observations that invasion fronts, tumour buds and micrometastases seldom contain LGR5⁺ cells (Extended Data Fig. 11n–q). Importantly, treatment with diphtheria toxin (DT) before surgical removal of the primary CRC effectively eliminated LGR5⁺ cells (Fig. 4m–p), yet it did not prevent disease relapse and mice developed overt liver metastatic disease (Fig. 4q–s). Thus, LGR5⁺ cells are dispensable for dissemination and metastatic colonization. We validated these results using an independent AKTP MTO line engineered with an iCT cassette knocked-in into the *Lgr5* locus (Extended Data Fig. 11r–t). Again, effective ablation of LGR5⁺ cells in the iCT model (Extended Data Fig. 11u–w) neither altered CRC tumour growth (Extended Data Fig. 11x) nor prevented metastatic recurrence (Extended Data Fig. 11y). By contrast, ablation of LGR5⁺ cells after direct inoculation of MTOs into the liver through the portal vein halted metastasis formation (Fig. 4t,u), further supporting a requirement for LGR5⁺ cells during metastatic outgrowth³.

Neoadjuvant immunotherapy avoids relapse

We previously showed that mouse primary AKTP CRCs are models of microsatellite-stable/mismatch-repair-proficient CRCs characterized by a relatively low mutational burden, abundant stroma cells and

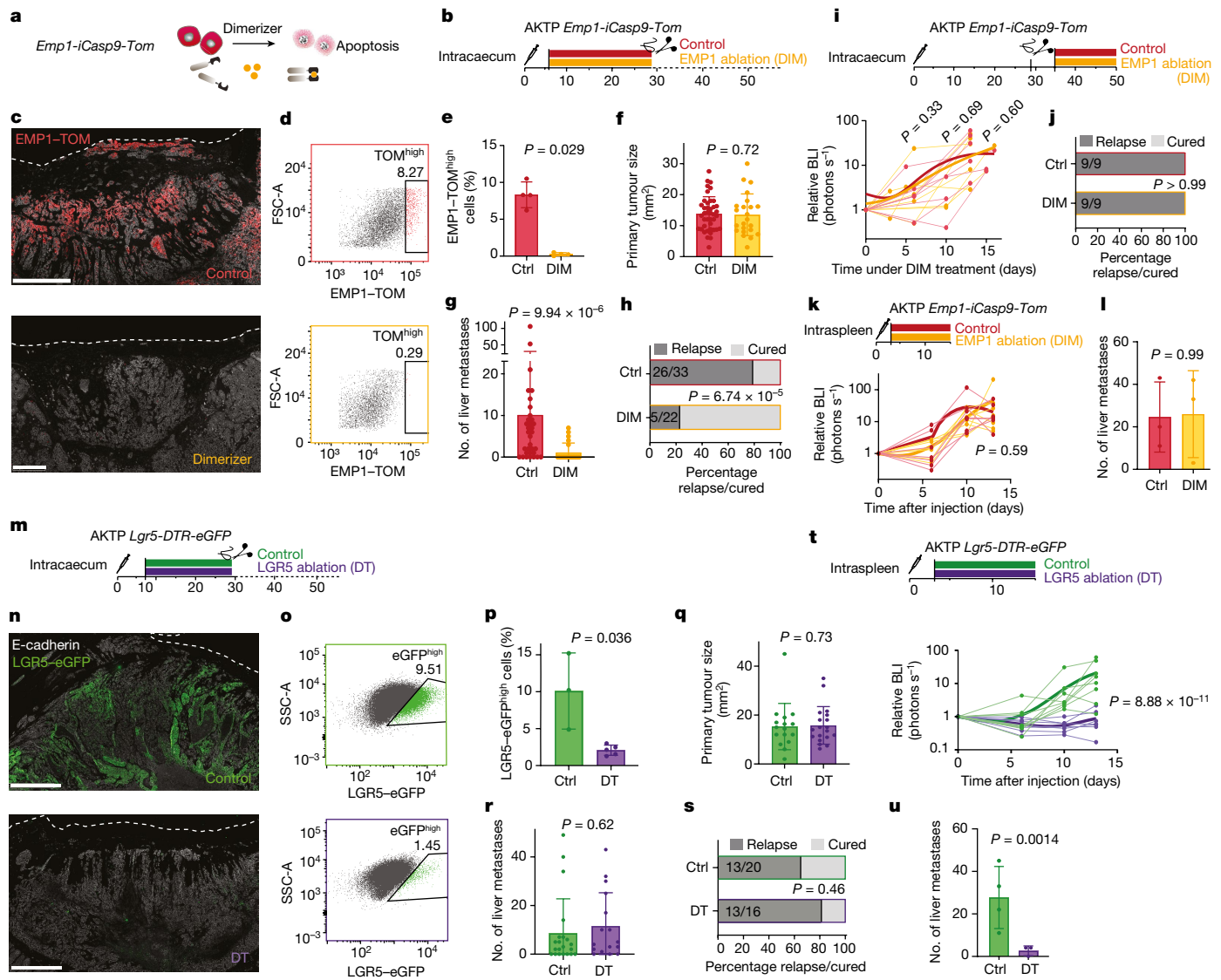


Fig. 4 | EMP1^{high} cells are the origin of metastatic relapse. **a**, EMP1-*iCasp9*-mediated cell ablation. **b**, Inducible EMP1⁺ cell ablation and surgery. **c**, EMP1-TOM⁺ cells in primary tumours after ablation. The lines mark the caecum borders. Scale bars, 500 μm (top) and 250 μm (bottom). **d**, Representative flow cytometry plot of TOM fluorescence in EPCAM⁺ cells. **e**, The percentage of EMP1-TOM^{high} tumour cells at the time of surgery. Data are mean ± s.d. Every dot represents a mouse. $n = 4$ in both groups. P values were calculated using two-sided Wilcoxon rank-sum tests. Ctrl, control. **f**, Primary tumour area after resection. Data are mean ± s.d. Each dot represents a mouse. $n = 33$ (control) and $n = 22$ (DIM) mice. Statistical analysis was performed using a linear model after Box-Cox transformation. **g**, Liver metastases at the experimental end points. Data are mean ± s.d. Each dot represents a mouse. n values are as described in **f**. **h**, The percentage of mice that relapse with liver metastases. P values were calculated using two-sided Fisher's tests. **i**, Late ablation of EMP1⁺ cells and metastatic growth on the basis of BLI monitoring. $n = 9$ mice per group. **j**, The percentage of mice that relapse with liver metastases. P values were calculated using two-sided Fisher's tests. **k**, Metastatic growth on the basis of BLI monitoring after ablation of EMP1⁺ cells 3 days after intrasplenic inoculation of *Emp1-iCasp9-Tom* organoids. $n = 3$ mice per each group. **l**, The

number of liver metastases in **k**. Data are mean ± s.d. $n = 3$ mice per group. **m**, The experimental timeline for **n–s**. **n**, Representative stainings of LGR5-DTR-eGFP⁺ cells in primary CRCs. Dashed lines outline the serosa. Scale bars, 100 μm. **o**, Representative flow cytometry plot of LGR5-eGFP fluorescence. **p**, The percentage of LGR5-eGFP^{high} tumour cells. Data are mean ± s.d. $n = 3$ (control) and $n = 5$ (DT) mice. The P value was calculated using a generalized linear model. **q**, Primary tumour area measured after resection. Data are mean ± s.d. $n = 20$ mice (control) and $n = 16$ (DT). P values were calculated using two-sided Wilcoxon rank-sum tests. **r**, Liver metastases at the experimental end points. Data are mean ± s.d. Each dot represents a mouse. n values are as described in **r**. **s**, The percentage of mice that relapse with liver metastases. P values were calculated using two-sided Fisher's tests. **t**, Experimental timeline and liver metastasis growth monitored by BLI after intrasplenic inoculation of MTOs. **u**, The number of liver metastases in **t**. Data are mean ± s.d. For **t** and **u**, $n = 7$ (control) and $n = 8$ (DT treated) mice. The points and lines of BLI measurements in **i**, **k** and **t** represent individual mice, the trend lines (bold) show a LOESS model and P values were calculated using a mixed-effects linear model. For **g**, **r**, **l** and **u**, P values were calculated using a generalized linear model.

T cell exclusion¹². Transplantation of AKTP MTOs in the caecum also gave rise to T-cell-excluded CRCs (Extended Data Fig. 12a). CAFs surrounded *Emp1*-expressing invasion fronts and tumour buds, but T cells

did not reach these structures and remained at the tumour periphery (Extended Data Fig. 12b). By contrast, liver micrometastases generated by these primary CRCs exhibited high CD3⁺ cell density (Fig. 5a). The

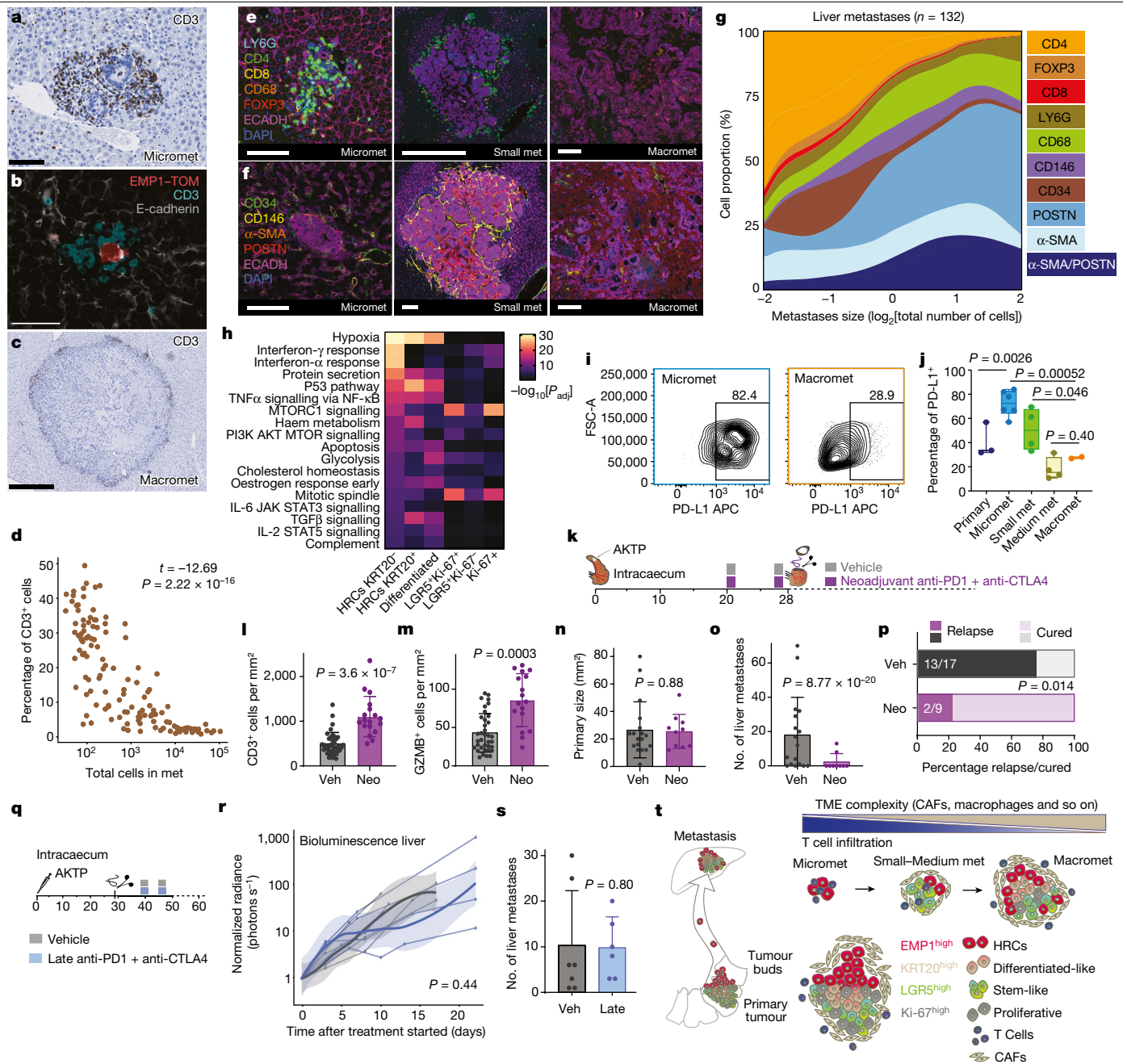


Fig. 5 | Neoadjuvant immunotherapy prevents metastatic relapse in CRC.

a, Representative images of CD3 immunohistochemistry on a liver micrometastasis. Scale bars, 100 μ m. **b**, tdTomato, CD3 and E-cadherin immunofluorescence showing Emp1-tdTomato⁺ liver micrometastatic cells surrounded by CD3⁺ cells. E-cadherin fluorescence is used for cell delimitation. Scale bar, 50 μ m. **c**, Representative images of CD3 immunohistochemistry on a liver macrometastasis. Scale bar, 500 μ m. **d**, The percentage of CD3⁺ T cells (of total TME cells) versus metastasis size. $n = 12$ mice, 133 metastases. Statistical analysis was performed using a linear model with mouse as a fixed effect. $t = -12.69$, $P = 2.22 \times 10^{-16}$. **e, f**, Representative examples of multiplex immunofluorescence of immune (e) and stromal (f) markers in metastases. Scale bars, 100 μ m (micro and macro), 500 μ m (small, e) and 100 μ m (small, f). **g**, Proportional stacked area graph showing TME cell types in metastases of different sizes; $n = 132$ metastases of 20 mice. **h**, Heat map showing Hallmark GSEA in the cell populations described in Fig. 2g. **i**, Representative flow cytometry contour plot showing PD-L1 expression in tumour cells of micro- and macrometastases. **j**, The percentage of PD-L1⁺ tumour cells determined using flow cytometry. The median (centre line), first and third quartiles (box limits) and the smallest and largest values (whiskers) are shown. $n = 3$ (primary tumour), $n = 6$ (micrometastases), $n = 4$ (small and medium metastases) and

$n = 2$ (macrometastases). Statistical analysis was performed using Wilcoxon rank-sum test. **k**, Neoadjuvant immunotherapy treatment. **l, m**, CD3⁺ (l) and GZMB⁺ (m) cell densities in primary CRCs. Data are mean $\pm$ s.d. $n = 19$ (control) and $n = 10$ (neoadjuvant (neo)). Statistical analysis was performed using a mixed-effects linear model. Veh, vehicle. **n**, Primary CRC area after resection. Data are mean $\pm$ s.d. Each dot represents a mouse. $n = 18$ (control) and $n = 10$ (treated). Statistical analysis was performed using a linear model. **o**, Liver metastases at the experimental end points. Data are mean $\pm$ s.d. Each dot represents a mouse. $n = 17$ (control) and $n = 9$ (treated). The P value was calculated using a generalized linear model. **p**, The percentage of mice that developed liver metastases at the experimental end points. n values are as described in o. The P value was calculated using a generalized linear model. **q**, Late immunotherapy. **r**, Liver metastasis growth measured by normalized BLI. The points and lines represent individual mice, and trend lines show a LOESS model. $n = 4$ (control) and $n = 5$ (late immunotherapy). Statistical analysis was performed using a linear model. **s**, Liver metastases generated by AKTP primary tumours. Data are mean $\pm$ s.d. Dots represent individual mice. $n = 7$ (control) and $n = 6$ (late immunotherapy). The P value was calculated using a generalized linear model. **t**, The proposed model for metastatic dissemination of CRC and TME co-evolution.

lack of T cell exclusion was evident and we could visualize T cell–HRC interactions in these early lesions (Fig. 5b). However, metastatic outgrowth was accompanied by a gradual decrease in CD3⁺ cells and the relocalization of T cells to the periphery (Fig. 5c,d). Multiplex immunohistochemistry analysis showed that most infiltrating T cells were CD4⁺FOXP3⁺ (examples are shown in Fig. 5e and quantification is shown in Fig. 5g and Extended Data Fig. 12c). T cell exclusion coincided with progressive CAF (α -SMA⁺ and/or POSTN⁺) and macrophage (CD68⁺) recruitment to the metastatic TME (Fig. 5f,g and Extended Data Fig. 12c).

scRNA-seq revealed elevated expression of interferon-alpha- and interferon-gamma-target genes in undifferentiated (KRT20[−]) HRCs (Fig. 5h and Supplementary Table 4), >95% of which belong to micrometastatic lesions, suggestive of an ongoing inflammatory response as HRCs reach the liver. Indeed, undifferentiated HRCs upregulate multiple interferon-response genes compared with the other tumour cell populations in primary CRCs and large metastases (Extended Data Fig. 12d). It was also apparent that HRCs in micrometastases expressed elevated *Cd274* (PD-L1) and *Ido1* levels, two negative regulators of the immune response (Extended Data Fig. 12d). Flow cytometry analysis confirmed that there was elevated PD-L1 expression on the surface of micrometastatic CRC cells and a progressive decline over subsequent outgrowth (Fig. 5i,j). On the basis of these findings, we hypothesized that, at the onset of organ colonization, the lack of a mature TME exposes HRCs to the adaptive immune system, yet disseminated cells bypass immune attack through the expression of immunomodulatory molecules. To test the susceptibility of micrometastatic disease to immunotherapy, we treated mice with anti-PD1 antibodies in combination with anti-CTLA4 antibodies in the neoadjuvant setting, that is, before the primary tumour was extirpated by surgery (Fig. 5k–p). This approach increased CD8⁺ cytotoxic T cell numbers in the primary CRC (Fig. 5l,m) but we did not observe an alteration in the tumour growth rate (Extended Data Fig. 12e) or curative effects on the primary disease at the experimental end points (Fig. 5n). Notably, however, most of these mice did not develop metastasis after surgical removal of the primary CRC and remained disease free (Fig. 5o,p). We obtained similar results with neoadjuvant anti-PD1 monotherapy (Extended Data Fig. 12f–i). By contrast, the same regime applied 2 weeks after surgery did not stop metastatic outgrowth (Fig. 5q–s), which is consistent with the failure of checkpoint immunotherapy observed in patients with metastatic microsatellite-stable CRC³⁵. Thus, during a temporal window after metastatic colonization, immunotherapy is effective in eliminating the residual disease and preventing subsequent metastatic relapse.

Discussion

HRCs represent a defined cell state in a large proportion of patient samples, implying a common origin and mechanism of metastatic recurrence across CRC genotypes and molecular subtypes. A pioneering study revealed that LGR5⁺ cancer stem cells are dispensable for primary CRC growth yet necessary for metastasis formation in experimental models³. Subsequently, it was proposed that metastases are initiated by disseminated differentiated tumour cells that, through plasticity, produce LGR5⁺ cancer stem cells upon reaching the liver³⁶. Another study showed that the adhesion molecule LICAM in LGR5⁺ and LGR5[−] cells is important to regenerate metastasis³⁷. We unequivocally show that HRCs are neither differentiated nor stem-like but, rather, co-opt a distinct state that enables migration and colonization of foreign organs. A defining feature of the HRC state is the upregulation of genes encoding cell-to-cell adhesion molecules. We therefore speculate that HRCs may extravasate as cell clusters and colonize foreign organs as oligocellular structures as previously shown for other cancer types^{22,38}. Reacquisition of the LGR5⁺ stem cell and proliferation programs occurs at a later phase and is necessary for metastatic outgrowth (Fig. 5t). Although ablation of HRCs in primary tumours prevents the vast majority of metastatic

relapses, it is formally possible that other tumour cell populations, including LGR5⁺ cells, may generate metastases if they reach foreign organs, as suggested by experiments of direct inoculation of tumour cells into the blood stream performed herein and elsewhere^{3,28,36,37}.

Our findings also reveal that residual metastatic cells lodged in foreign organs lack a mature TME and are susceptible to the attack of the adaptive immune system after immunotherapy treatment. This window of vulnerability could be exploited to prevent metastatic relapse. Our results support current efforts to use neoadjuvant immunotherapy in patients with early-stage CRC³⁹. This and other therapeutic strategies with the ability to eliminate HRCs may prevent disease relapse if applied before metastatic disease is overt.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-022-05402-9>.

- Amin, M.B. et al. The Eighth Edition AJCC Cancer Staging Manual: continuing to build a bridge from a population-based to a more “personalized” approach to cancer staging. *CA* **67**, 93–99 (2017).
- Shimokawa, M. et al. Visualization and targeting of LGR5⁺ human colon cancer stem cells. *Nature* **545**, 187–192 (2017).
- de Sousa e Melo, F. et al. A distinct role for Lgr5⁺ stem cells in primary and metastatic colon cancer. *Nature* **543**, 676–680 (2017).
- Cortina, C. et al. A genome editing approach to study cancer stem cells in human tumors. *EMBO Mol. Med.* **9**, 869–879 (2017).
- Calon, A. et al. Dependency of colorectal cancer on a TGF- β -driven program in stromal cells for metastasis initiation. *Cancer Cell* **22**, 571–584 (2012).
- Calon, A. et al. Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nat. Genet.* **47**, 320–329 (2015).
- Isella, C. et al. Stromal contribution to the colorectal cancer transcriptome. *Nat. Genet.* **47**, 312–319 (2015).
- Lee, H.-O. et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nat. Genet.* **52**, 594–603 (2020).
- Guinney, J. et al. The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**, 1350–1356 (2015).
- Raghavan, S. et al. Microenvironment drives cell state, plasticity, and drug response in pancreatic cancer. *Cell* **184**, 6119–6137 (2021).
- Joanito, I. et al. Single-cell and bulk transcriptome sequencing identifies two epithelial tumor cell states and refines the consensus molecular classification of colorectal cancer. *Nat. Genet.* **54**, 963–975 (2022).
- Tauriello, D. V. F. et al. TGF β drives immune evasion in genetically reconstituted colon cancer metastasis. *Nature* **554**, 538–543 (2018).
- Massagué, J. & Obenauf, A. C. Metastatic colonization by circulating tumour cells. *Nature* **529**, 298–306 (2016).
- Barriga, F. M. et al. Mex3a marks a slowly dividing subpopulation of Lgr5⁺ intestinal stem cells. *Cell Stem Cell* **20**, 801–816 (2017).
- Lange, M. et al. CellRank for directed single-cell fate mapping. *Nat. Methods* **19**, 159–170 (2022).
- Álvarez-Varela, A. et al. Mex3a marks drug-tolerant persister colorectal cancer cells that mediate relapse after chemotherapy. *Nat. Cancer* **3**, 1052–1070 (2022).
- Tyler, M. & Tirosh, I. Decoupling epithelial-mesenchymal transitions from stromal profiles by integrative expression analysis. *Nat. Commun.* **12**, 2592 (2021).
- Grigore, A. D., Jolly, M. K., Jia, D., Farach-Carson, M. C. & Levine, H. Tumor budding: the name is EMT. *Partial EMT. J. Clin. Med.* **5**, 51 (2016).
- Roa-Peña, L. et al. Keratin 17 identifies the most lethal molecular subtype of pancreatic cancer. *Sci. Rep.* **9**, 11239 (2019).
- Durgan, J. et al. SOS1 and Ras regulate epithelial tight junction formation in the human airway through EMP1. *EMBO Rep.* **16**, 87–96 (2015).
- Bangrow, T. et al. The epithelial membrane protein 1 is a novel tight junction protein of the blood-brain barrier. *J. Cereb. Blood Flow Metab.* **28**, 1249–1260 (2008).
- Aceto, N. et al. Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. *Cell* **158**, 1110–1122 (2014).
- Barry, E. R. et al. Restriction of intestinal stem cell expansion and the regenerative response by YAP. *Nature* **493**, 106–110 (2013).
- Cheung, P. et al. Regenerative reprogramming of the intestinal stem cell state via hippo signaling suppresses metastatic colorectal cancer. *Cell Stem Cell* **27**, 590–604 (2020).
- Vasquez, E. G. et al. Dynamic and adaptive cancer stem cell population admixture in colorectal neoplasia. *Cell Stem Cell* **29**, 1213–1228 (2022).
- Han, T. et al. Lineage reversion drives WNT independence in intestinal cancer. *Cancer Discov.* **10**, 1590–1609 (2020).
- Lupo, B. et al. Colorectal cancer residual disease at maximal response to EGFR blockade displays a druggable Paneth cell-like phenotype. *Sci. Transl. Med.* **12**, eaax8313 (2020).
- Heinz, M. C. et al. Liver colonization by colorectal cancer metastases requires YAP-controlled plasticity at the micrometastatic stage. *Cancer Res.* **82**, 1953–1968 (2022).

29. Solé, L. et al. p53 wild-type colorectal cancer cells that express a fetal gene signature are associated with metastasis and poor prognosis. *Nat. Commun.* **13**, 2866 (2022).
30. Ohta, Y. et al. Cell-matrix interface regulates dormancy in human colon cancer stem cells. *Nature* **680**, 784–794 (2022).
31. Mustata, R. C. et al. Identification of Lgr5-independent spheroid-generating progenitors of the mouse fetal intestinal epithelium. *Cell Rep.* **5**, 421–432 (2013).
32. Wang, Y. et al. Comprehensive molecular characterization of the hippo signaling pathway in cancer. *Cell Rep.* **25**, 1304–1317 (2018).
33. Yuan, Y. et al. YAP1/TAZ-TEAD transcriptional networks maintain skin homeostasis by regulating cell proliferation and limiting KLF4 activity. *Nat. Commun.* **11**, 1472 (2020).
34. Morral, C. et al. Zonation of ribosomal DNA transcription defines a stem cell hierarchy in colorectal cancer. *Cell Stem Cell* **26**, 845–861 (2020).
35. Le, D. T. et al. PD-1 blockade in tumors with mismatch-repair deficiency. *N. Engl. J. Med.* **372**, 2509–2520 (2015).
36. Fumagalli, A. et al. Plasticity of Lgr5-negative cancer cells drives metastasis in colorectal cancer. *Cell Stem Cell* **26**, 569–578 (2020).
37. Ganesh, K. et al. L1CAM defines the regenerative origin of metastasis-initiating cells in colorectal cancer. *Nat. Cancer* **1**, 28–45 (2020).
38. Padmanaban, V. et al. E-cadherin is required for metastasis in multiple models of breast cancer. *Nature* **573**, 439–444 (2019).
39. Chalabi, M. et al. Neoadjuvant immunotherapy leads to pathological responses in MMR-proficient and MMR-deficient early-stage colon cancers. *Nat. Med.* **26**, 566–576 (2020).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2022

Methods

MTOs and integration of cassettes using CRISPR

We previously described¹² the establishment of MTOs from primary tumours arising in genetically engineered mouse models with compound genetic alterations (*Apc*, *Kras*, *Trp53*, *Tgfbr2*). MTOs were cultured as described previously¹² and checked bimonthly for mycoplasma contamination. Small guide RNAs (sgRNA; 20 bp) were designed to cut 9–11 bp after the STOP codons using the <http://crispr.mit.edu> web tool and were cloned into the pX330-IRFP hSp-enhanced-Cas9 plasmid⁴. The sgRNA sequence for *Emp1* was AAATAAGCCGAATACGCTCA; and for *Lgr5* the sequence was GTCTCTAGTGACTATGAGAG. Then, 1 kb 5' and 3' homology arms were sequentially cloned into pShuttle vectors containing the inducible Caspase9-tdTomato cassette³⁴, a gift from T. Sato. The Ires-DTR-T2A-eGFP-WPRE-BGHpA sequence was cloned between the *Lgr5* homology arms flanking the gene stop codon. *eGFP* was cloned after the *Ires* sequence to generate the *Lgr5*-Ires-eGFP-WPRE-BGHpA donor. CRISPR-Cas9 knockin editing was performed as described previously⁴. AKTP-MTO-93 *Emp1-iCT* was generated from a single tdTomato⁺ cell (clone 14). MTO-93 *Emp1-iCT Lgr5-eGFP* was first generated from a single tdTomato⁺ cell (clone 49), then nucleofected with the *Lgr5-eGFP* construct and established from a pool of sorted eGFP⁺ cells. MTO-93 *Lgr5-iCT* was generated by a single tdTomato⁺ cell (clone 2). MTO-93 *Lgr5-DTR-eGFP* was generated by sorting a pool of eGFP⁺ cells. Triple-mutant AKP MTO-54¹² *Emp1-iCT* was generated by sorting a pool of tdTomato⁺ cells. CRISPR-Cas9 gene editing was subsequently used to introduce a *Smad4* mutation in AKP *Emp1-iCT* to generate AKPS MTOs as previously described⁴⁰. Correct integration of the knockin cassettes was checked by PCR and sequencing of the genomic regions. Correct expression of the reporter cassette was tested by RT-qPCR or by sorting cells from tumours as described in the text. A list of genotyping primers is provided in Supplementary Table 5.

Generation of CTOs

To study the effect of gene driver mutations on the HRC phenotype, we sequentially introduced mutations into *Apc*, *Kras*, *Trp53*, *Tgfbr2* and/or *Smad4* in normal mucosa colonic organoids using CRISPR-Cas9 as described elsewhere^{16,40,41}. Compound mutations were introduced into organoids derived from one c57BL/6J mouse, which enabled us to assess the effect of individual mutations on gene expression without other confounding variables. We named these organoids CTOs to distinguish them from MTOs, that were established from genetically engineered mouse models.

Lentivirus production and MTO infection

For bioluminescence tracking, MTOs were infected with a lentivirus encoding an eGFP-firefly luciferase fusion reporter construct under the control of the *Pgk1* promoter. For YAP inhibition experiments, AKTP organoids were infected with shControl (SHC002) or shYAP1 (TRCN0000095864, TRCN0000095865, TRCN0000095866, TRCN0000095867) Mission Sigma-Aldrich lentiviral constructs. We also used a pInducer20 eGFP-TEADi vector, a gift from R. Iglesias-Bartolome (Addgene plasmid, 140145), that blocks the activity of both YAP and TAZ³³.

Animal experimentation approval and maintenance

Experiments with mice were approved by the Animal Care and Use Committee of Barcelona Science Park under protocol CEEA-PCB-14-000053. Mice were maintained in a specific-pathogen-free (SPF) facility under a 12 h–12 h light–dark cycle, under controlled temperature and humidity (18–23 °C and 40–60%, respectively) and given ad libitum access to standard diet and water. All of the mice were closely monitored by the authors, facility technicians and by an external veterinary scientist responsible for animal welfare. We monitored primary tumour and metastasis growth using intravital bioluminescence at least once a week.

Inoculation of MTOs into the caecum and rectum

For all injections, c57BL/6J mice were purchased from Janvier Labs at six weeks of age and injected at 7 to 9 weeks of age. Sex always matched the origin of the tumour. Intracaecum injections were used to generate primary tumours. Organoids were collected and incubated for 30 min with cold HBSS to break down BME (Bio-Techne, 3533-010-02) without disrupting their structure. Cells were then counted and resuspended in 70% BME in HBSS for injection (0.1×10^6 cells in 10 μ l per mouse). Full organoids were injected with a 30G syringe into the submucosal wall of the distal caecum while looking through a binocular lens. We introduced a significant modification to previous protocols⁴² by moving the injection site to the apex of the caecum, which enabled posterior surgical resection. For liver colonization studies, we used intrasplenic injections of organoids as described before¹². Rectal injections were performed according to a previously described procedure⁴³. Primary CRC and metastasis growth were monitored using bioluminescence. Experiments were terminated when tumour bioluminescence measurements reached saturation or if the animals presented signs of suffering or poor health status according to protocols approved by the animal experimentation committee of the Barcelona Science Park.

Primary tumour resection

Mice were anesthetized with isoflurane and placed into dorsal recumbency. The abdomen was shaved and sterilized with povidone-iodine surgical solution. A small midline incision—slightly to the left—was performed to open the skin and peritoneum and expose the abdominal area. We placed a sterile surgical drape on top of the abdomen with a circular hole above the incision and sprawled the caecum over the drape using cotton swabs and saline to keep it hydrated. After confirming the presence of a primary tumour, Kelly forceps were used to first knot the surgical suture into the caecum wall, in between the ileocecal junction and the primary tumour. This provided a grip for subsequent caecum ligation. After ligation, the apical caecum containing the primary tumour was excised and any remaining caecal tissue was trimmed. After resection and organ fixation, we measured primary tumour size using callipers. A video of the surgery, which usually lasted 5–10 min, is provided (Supplementary video 1). The fitness of the mice was monitored weekly throughout the experiment. Mice were euthanized four weeks after resection and metastases were scored macroscopically.

Pharmacological treatments

For iCaspase-9-inducible ablation experiments, animals were treated with dimerizer (AP20187, MedChem Express, HY-13992) through intraperitoneal injection at 2.5 mg per kg three times per week (*Emp1-iCasp9*) or 5 mg per kg every day (*Lgr5-iCasp9*). For DTR-inducible ablation, mice were treated with 16.7 μ g per kg of DT (Sigma-Aldrich, 322326) three times per week. For immunotherapy experiments, 2 shots of 250 μ g of anti-mouse CTLA-4 (C2444, Leinco) and/or anti-mouse PD1 (P372, Leinco) were administered through intraperitoneal injection 5 days apart.

Tumour dissociation for flow cytometry

Primary tumours and microdissected liver metastases were chopped with razor blades. Subsequent enzymatic digestion was performed with 200 U ml⁻¹ collagenase IV (Sigma Aldrich, C5138) in HBSS (Lenovo) for 30 min at 37 °C, in a shaking water bath or a gentleMACS Dissociator (Miltenyi Biotec). Digested tissue fragments were then filtered through 100 μ m and 40 μ m meshes, washed and treated for 5 min with ammonium chloride. Single-cell preparations were first blocked with anti-CD16/32 (93, eBioscience) and then stained with APC anti-EPCAM (BioLegend, 118214) and BV605 anti-CD45 (BioLegend, 103155) antibodies at 1:200 concentration. In experiments measuring PD-L1 expression in tumour cells, cells were stained with FITC anti-EPCAM (Santa Cruz,

G8.8, 53532) and APC anti-PD-L1 (BD, 564715, MIH5) antibodies at 1:200 concentration. Finally, cells were resuspended in HBSS with 0.5% FBS and DAPI (Sigma Aldrich, D9542).

Collection of mouse primary CRC and metastasis samples

Metastatic seeding in the CRC relapse models is not synchronized; as it occurs in patients, tumour cells are shed from the primary tumour continuously during the weeks until the day of resection. As a result, metastases expand for different periods in the liver and lungs, and exhibit different sizes at the experimental time points depending on when HRCs colonized the organ. We therefore decided to classify according to size instead to time. Four different CRC samples were collected—primary, micrometastases, small metastases and macrometastases. Micrometastases and small metastases were collected from liver of mice 29 to 31 days after primary tumour implantations (that is, at the time of primary CRC resection). Micrometastasis samples were DTCs collected from livers with absent or residual bioluminescence *ex vivo* in which metastases were not visible. For small metastases, metastatic nodules were visible but small in size (<1.5 mm). Macrometastases samples were derived from metastatic nodules larger than 4 mm in livers from mice 64 to 66 days after primary tumour implantation. Primary tumour samples were paired with micro, small or macrometastasis samples and were collected from both time points (29 to 31 and 64 to 66 days after primary tumour implantation).

Isolation of liver residual DTCs

For isolation of low numbers of tumour cells from mice without visible metastases, whole livers were thoroughly minced with razor blades. After an initial 30 min digestion with 200 U ml⁻¹ collagenase IV (Sigma-Aldrich, C5138), samples were filtered through 100 µm meshes. Although most cells flowed through, a small fraction of the sample—highly enriched for tumour cells—was retained in the filter (Extended Data Fig. 3k,l). The filters were then laid into a six-well plate and covered with HBSS containing 200 U ml⁻¹ collagenase IV, 200 µg ml⁻¹ dispase II (Sigma-Aldrich, D4693), 40 µg ml⁻¹ DNase I (Sigma-Aldrich, 10104159001) and 13 µM rock inhibitor (MedChem Express, HY-10583). After redigestion in a water bath for 30 min at 37 °C, most cells now seeped through the filters. The protocol continued with washing, ammonium chloride treatment and antibody staining as described above. DAPI⁺eGFP⁺CD45⁻ cells were gated to sort tumour cells.

Histology and tissue staining

Standard haematoxylin and eosin and antibody staining were performed on 4 µm tissue sections using standard procedures, as described previously¹². Details are provided in Supplementary Table 6. The following secondary antibodies were used: donkey anti-goat conjugated to Alexa 488/568/647 (Life Technologies A11055, A11057, A21447), donkey anti-rabbit conjugated to Alexa 488/568/647 (Life Technologies A21206, A10042, A31573) and donkey anti-mouse conjugated to Alexa 488/568/647 (Life Technologies A-21202, A10037, A31571) at room temperature. Digital scanned bright-field and fluorescent images were acquired with a NanoZoomer-2.0 HT C9600 scanner (Hamamatsu Photonics). All of the images were visualized with the NDP.view 2 U123888-01 software (Hamamatsu Photonics) with a gamma correction set at 1.8 in the image control panel.

Co-culture of MTOs and colon fibroblasts

To address the role of fibroblasts in inducing HRCs, we seeded 5,000 AKTP *Emp1-ICT* MTO cells with or without 100,000 mouse colon fibroblasts in BME (Bio-Techne, 3533-010-02) and supplemented them with MTO cell culture medium¹² without noggin and galunisertib. Co-cultures were analysed 7 days after by flow cytometry and confocal microscopy. For Fig. 3s and Extended Data Fig. 10e, MTOs and mouse colon fibroblasts were also co-cultured in hanging drops as previously described⁴⁴.

In situ hybridization analysis of samples of patients with CRC

Paraffin-embedded tissue sections (thickness, 2–3 µm) of human CRC primary tumours were air-dried and further dried at 60 °C over-night before any staining. To compare *EMP1* and *LGR5* expression, sections were hybridized with RNAscope Probe Hs-LGR5 (311021, Bio-Techne, R&D Systems) in the C1 channel, a custom-made RNAscope Probe Hs-EMP1 (Bio-Techne, R&D Systems, 895051-C2) in the C2 channel and an Alexa568-conjugated antibody against E-cadherin. To compare the expression of *EMP1* with other HRC-marker genes, sections were hybridized using the RNAscope Probe Hs-EMP1 mRNA in channel 1 (Bio-Techne, R&D Systems, 895051-C1) and stained with the relevant antibodies. FISH probes were detected using the RNAscope Multiplex Fluorescent Detection Reagents Kit v2 (Bio-Techne R&D Systems, 323110).

Multiplex immunofluorescence

Multiplex immunofluorescence staining was performed on 4-µm-thick formalin-fixed paraffin-embedded sections using the OPAL protocol (Akoya Biosciences) on the Leica BOND RXm autostainer (Leica Microsystems). Six consecutive staining cycles were performed using the following primary antibody–Opal fluorophore pairings. Stroma panel: CD34 (1:3,000, ab81289; Abcam) with Opal 520; CD146 (1:500, ab75769; Abcam) with Opal 570; α-SMA (1:1,000, ab5694; Abcam) with Opal 620; periostin (1:1,000; Abcam, ab227049) with Opal 690; and E-cadherin (1:500; Cell Signaling, 3195) with Opal 650. Immune panel: (1) LY6G (1:300; BD Pharmingen, 551459) with Opal 540; (2) CD4 (1:500; Abcam, ab183685) with Opal 520; (3) CD8 (1:800; Cell Signaling, 98941) with Opal 570; (4) CD68 (1:1,200, Abcam, ab125212) with Opal 620; (5) FOXP3 (1:400; Cell Signaling, 126553) with Opal 650; and (6) E-cadherin (1:500; Cell Signaling, 3195) with Opal 690. Primary antibodies were incubated for 60 min and detected using the BOND Polymer Refine Detection System (Leica Biosystems, DS9800) according to the manufacturer's instructions, substituting DAB for the Opal fluorophores, with a 10 min incubation time and without adding haematoxylin. Whole-slide scans and multispectral images (MSI) were obtained on the Akoya Biosciences Vectra Polaris system. Batch analysis of the MSIs from each case was performed using the inForm (v.2.4.8) software provided. Finally, batch-analysed MSIs were fused in HALO (Indica Labs) to produce a spectrally unmixed reconstructed whole-tissue image, ready for analysis.

Statistical analyses multiplex immunofluorescence

The Halo Image Analysis Platform software was used for quantification of cell phenotypes within metastases and primary tumours. Two matrices of counts (distinguishing immune and stromal panels) with the total number of cells per metastasis/primary (in rows) assigned to each cell population (in columns) were obtained. Multiple positives were present in both immune and stromal panels. These cases observed in the immune panel represented only 0.7% of the total assignments and were removed from the analysis. For the stromal panel, α-SMA and POSTN double positives were retained and labelled in a distinct category α-SMA/POSTN. The rest of the multiple positives, comprising 7% of the total cells, were amalgamated and labelled as stromal others. For all metastases, information regarding metastasis size (log₂-transformed total number of cells), metastatic burden (defining micro, small and macrometastases) and organ site were considered for subsequent analysis.

For the proportional stacked area graph, only the percentages of CD4, FOXP3, CD8, LY6G and CD68 for the immune panel, and the percentages of CD146, CD34, POSTN, α-SMA and α-SMA/POSTN for the stromal panel were considered. The cell composition of all measured metastases (in percentages) was averaged out using the R function *aggregate*, taking as grouping elements both metastasis size and metastatic burden. Linear mixed-effects models were fitted independently for every cell population using the CLR-transformed values as the response variable,

Article

the metastatic burden and the metastasis size as fixed effects, and the tumour ID as a random effect to consider the dependence between metastases for the same tumour.

Quantification of tdTomato and eGFP fluorescence intensity

For the quantification of percentages of EMP1^{high}-tdTomato high and LGR5^{high}-eGFP^{high} in tumour sections, we used the Halo Image Analysis Platform. In brief, the epithelial tumour area was classified apart from the stroma, background and necrosis using a random-forest algorithm. Single cells were detected and tdTomato/eGFP fluorescence intensity was measured for every cell. In primary tumours, TOM^{high} and eGFP^{high} tumour cells were defined as the cells in the 90th percentile of each sample. In liver metastases, TOM^{high} and eGFP^{high} tumours cells were defined as the cells in the 90th percentile for all metastases measured. For *Emp1-iCT* liver metastases (Extended Data Fig. 6j) and *Lgr5-DTR-eGFP* liver metastases (Extended Data Fig. 11q), tumour cell fluorescence intensity was analysed using ImageJ with a custom-made macro. Tumour cells were first detected and isolated using E-cadherin to create a mask. TdTomato or eGFP intensity was calculated for every pixel inside the masked area. We then plotted the percentage of fluorescence high and low pixels as a function of the area of the metastases (measured in pixels).

Gene expression analysis using RT-qPCR

RT-qPCR and microarrays were used to compare subpopulations of EMP1^{high} and EMP1^{low}, and LGR5^{high} and LGR5^{low} cells dissociated from MTO-93 organoids grown in vitro or in vivo. Subpopulations were defined in flow cytometry as the top/bottom 10% in fluorescence expression. RNA from 2,000 cells was extracted and retrotranscribed to cDNA as described previously⁴⁵. To analyse changes in gene expression, RT-qPCR was performed using 5 ng of cDNA per each qPCR well. qPCRs were performed using the TaqMan Universal PCR Master Mix (Applied Biosystems, 4369016) or the PowerUp SYBR Green Master Mix (Applied Biosystems, 100029284) in triplicates, according to the manufacturer's instructions. Gene expression levels were normalized using the housekeeping genes *PPIA* or *B2M*. The following TaqMan assays were used: *TdTomato-BGHPA* (custom-made probe; F, GGGCATGGCACCAGCAGCACC; R, CCTACTTGACAGCTCGTC CATGCC), *MmPpia* (Mm002342430_g1), *MmB2m* (Mm00437762_m1), *eGFP* (Mr04097229mr), *MmEmp1* (Mm00515678_m1), *MmLgr5* (Mm0043889_m1), *MmSmoc2* (Mm00491553_m1), *MmLamC2* (Mm00500494), *Ctgf* (Mm01192033_g1). The following Sybr primers were used: *MmYap* (F, ACCCTCGTTTGGCATGAAC; R, TGTGCTGG GATTGATATCCGTA), *MmCyr61* (F, AGGTCTGCGCTAAACAACCTCA; R, ATATTCACAGGGTCTGCCTTCT).

Western blotting

Cells were collected in Cell Recovery Solution (Corning, 354257), pelleted and lysed in RIPA buffer (Tris-HCl pH 7.4, 150 mM NaCl, 0.5% Sodium deoxycolate, 0.1% SDS, 1 mM EDTA, 10 mM NaF, 1 mM PMSF, 1% Triton X-100) supplemented with protease and phosphatase inhibitors (Sigma-Aldrich). Primary antibodies were incubated overnight at 4 °C at the following dilutions: YAP/TAZ (Cell Signalling, CS93622, 1:1,000) and vinculin (Sigma Aldrich, 9264, 1:20,000). Anti-Rabbit IgG HRP (NA934V) and anti-mouse IgG HRP (NXA931V) secondary antibodies were diluted 1:5,000 and incubated for 1 h at room temperature with the PVDF membranes. Membranes were visualized using Hyper Processor (Amersham Pharmacia Biotech).

RNA-seq and analysis

We used RNA-seq to profile CTOs with different genotypes and to profile the effects of chemotherapy treatment on MTOs. For the first dataset, 5,000 cells from CTOs with multiple genotypes (AT, AP, AS, AK, ATP, APS, AKP, AKS, AKPT, AKPS) were seeded in BME (Bio-Techne, 3533-010-02) and cultured in full stem cell medium⁴. Organoids were collected

7–10 days after seeding. For chemotherapy experiments, MTO54 organoids were dissociated with TrypLE express (Thermo Fisher Scientific, 12604039) and single cells were resuspended in BME in six-well plates in complete stem cell medium. Then, 2 days later, the medium was removed and fresh medium containing folfiri (5FU at 50 $\mu\text{g } \mu\text{l}^{-1}$ (Sigma-Aldrich, F6627) plus SN38 at 100 nM (MedChem, HY-13704)) or control (MTO stem cell medium¹²) was added for 2 days.

RNA was extracted and sequenced as we previously described¹⁶. RNA-seq reads from datasets (CTOs or chemotherapy treatment) were aligned with STAR (v.2.5.2)⁴⁶ using the default parameters to the *Mus musculus* reference genome built with annotations version GENCODE_mmusculus_vM25. SAM files were converted to BAM and sorted using Sambamba (v.0.7.1)⁴⁷. Count matrices were generated using the R (v.4.0.5) package Rsubread (v.2.4.3)⁴⁸ with the GENCODE_mmusculus_vM25 custom annotation. Data from parental versus metastatic MTOs were processed as described previously¹². Gene-wise differential expression in the chemotherapy dataset between controls and folfiri treatment was performed using the R package DESeq2 (v.1.30.1)⁴⁹. Normalized values for plots were obtained using the rlog function of the same package. Signature scores were defined as the scaled mean of all genes in the signature after scaling the expression matrix. Comparison of signature scores between conditions was assessed using *t*-tests.

Microarray expression analyses

Microarrays (GeneChip Mapping 250K Nsp Assay Kit, Affymetrix) were used to compare EMP1^{high} and EMP1^{low} cells from MTO-93 *Emp1-iCT* primary tumours grown for 4 weeks; and to compare MTO-93 *Emp1-iCT* MTOs grown in vitro alone or in combination with mouse colon fibroblasts for 7 days. Samples were processed with oligo (v.1.46.0)⁵⁰ (fit-ProbeLevelModel: background = TRUE, normalize = TRUE, target = "core", method = "plm"). Raw cel files were normalized using the RMA method⁵¹ (with the default parameters). Probe sets were annotated with Clariom_S_Mouse_HT-na36-mm10-transcript Affymetrix databases. Standard quality controls were considered to identify abnormal samples⁵². No samples were excluded due to quality issues.

Differential expression analysis of EMP1^{high} versus EMP1^{low} data was performed using a linear model with empirical shrinkage (limma R package)⁵³, taking into account the paired data setting. Differential expression analysis of co-culture data was performed using the same regression method, but considering as adjustment variable the Eklund metric pm.iqr⁵⁴ to reduce the influence of technical variation. Benjamini–Hochberg false-discovery rate (FDR) correction was used for multiple-comparison correction. Gene set analysis was used to explore the enrichment in custom gene sets. limma's rotation-based approach for enrichment⁵⁵ was considered to represent the null distribution. The maxmean enrichment statistic proposed in ref. ⁵⁶, under restandardization, was used for competitive testing. Gene signature (Supplementary Table 7) z-scores⁵⁷ were used to measure pathway activity. For doing so, normalized expression was adjusted for biological replicate, centred and scaled gene-wise according to the mean and the s.d. computed across samples. Gene signature z-score was summarized by taking the average of its constituent genes. Furthermore, a global signature was computed using all genes and used for a priori centring of signature scores. This strategy has been proved to be useful to avoid systematic biases. Only the expression of the most variable probe sets per gene were considered for gene set analyses.

10x mouse scRNA-seq analysis

Cell Ranger⁵⁸ (v.4.0.0) was used to align reads to a custom refdata-gex-mm10-2020-A transcriptome including the *eGFP* and luciferase genes. Gene expression was analysed using Seurat (v.4.0.3)^{59–62}. A total of 1,330 cells with <20% mitochondrial content and >3,000 detected genes were considered. Ribosomal reads (17% of the total) were removed. Mitochondrial content was regressed out during SCT normalization⁶³.

SCT-transformed counts were smoothed using MAGIC (v.2.0.3)⁶⁴. Gene signature expression was summarized by taking the average MAGIC score of its constituent genes. HRCs and LGR5⁺ cell populations were defined by having a score above the 75th percentile. FindAllMarkers was used to identify differentially expressed genes from raw counts. The HRC cell population was compared to the rest to identify mouse HRC markers. Testing was limited to genes detected in >10% of cells and showing >0.25 log₂-transformed fold difference. Functional enrichment analysis was performed using the GSEA preranked⁶⁵ method ranking genes by the log₂-transformed average fold change. The significance threshold was set at 5% Benjamini–Hochberg FDR.

To select for marker genes to track HRCs in mouse tumours, we computed the correlation scores for all genes with the EpiHR signature in the SMC human dataset and the mouse 10x dataset. We used MAGIC score for genes and signature scores as defined previously.

Mouse Smart-seq2 scRNA-seq analysis

Smart-seq2 reads were aligned to the UCSC_GRCm38.mm10 genome with zUMIs⁶⁶ and analysed using Seurat^{59–62} (v.4.0.3). Four technical batches of AKTP at different stages were merged into a single object and two technical batches of AKP micrometastases were merged into another object. A total of 1,057 cells with <20% mitochondrial content and >20,000 UMIs were considered for AKTP, whereas 414 cells with <20% mitochondrial content and >100,000 UMIs were isolated from AKP micrometastases. Ribosomal reads were removed. Mitochondrial content was regressed out during SCT normalization. SCT-transformed counts were further imputed and smoothed using MAGIC (v.2.0.3)⁶⁴. Gene signature expression (Supplementary Table 7) was summarized by taking the average MAGIC score of its constituent genes. Non-epithelial cells were removed and normalized again. To improve the integration of the four AKTP batches, the IntegrateData function was used with precomputed anchors based on 3,000 features. The integrated dataset was SCT normalized, smoothed with MAGIC (v.2.0.3)⁶⁴ and clustered with the FindClusters Seurat function (resolution = 1.2). FindMarkers was used to identify differentially expressed genes from raw counts. CellRank¹⁵ was used to uncover the cell-state dynamics of CRC metastases from RNA velocity estimates^{67,68}. Gene expression or signature expression was represented as a function of latent time with R⁶⁹. Moreover, the integration, normalization, imputation and trajectory analysis were performed independently for the subset of cells collected from primary tumours, incipient metastasis and macrometastases.

Creation of CRC transcriptomic meta-cohort

Public CRC transcriptomic datasets were downloaded from the Gene Expression Omnibus (GEO)⁷⁰ and NCI GDC commons⁷¹, preprocessed and homogenized into a unique meta-cohort, including 1,830 samples from TCGA⁷², GSE38832 (ref. ⁷³), GSE44076 (ref. ⁷⁴), GSE33113 (ref. ⁷⁵), GSE14333 (ref. ⁷⁶), GSE39582 (ref. ⁷⁷) and GSE37892 (ref. ⁷⁸) (Supplementary Table 1). The last four datasets include disease-free survival information with a median follow-up of 3.7 years and clinical information across all datasets is gender, age, stage and location of primary tumour (Supplementary Table 1). When not available, MSI status was imputed using the transcriptomic signature reported in ref. ⁷⁹ through density-based nonparametric clustering^{80,81}. Signature scores were computed as the scaled mean of the genes in the signature after scaling the expression matrix.

Microarray data were processed separately using RMA⁸². Information about sample processing and hybridization was retrieved from the CEL files. For TCGA, the Legacy version was used for expression with clinical annotation from October 2016. Genes were annotated using the Ensembl Biomart database (GRCh37)^{82,83}. Duplicated samples across platforms were removed from the GA dataset, as well as samples from locations other than the colon or rectum. RSEM⁸⁴ expression values were already log₂-transformed and quantile-normalized. Samples

TCGA-A6-2679-01A and TCGA-AA-A004-01A were excluded owing to an anomalous expression distribution among all samples.

Each microarray series was corrected by Eklund metrics⁵⁴, centre and scanning date using a mixed-effect linear model⁸⁵. Age, gender, stage, site and MSI were also included in these models. TCGA sets were corrected for occurrences of combinations of centre and plate identifier (random effect). For microarrays, probe sets were summarized at the gene level using the first principal component from probe sets in that gene. This component was then centred and scaled to the weighted mean of the means and s.d. values of the probe sets. The sign of the component was changed to the sign of the probe set contributing the most. All of the datasets were merged after gene-wise standardization to the GSE39582 series according to the distribution of gender, age, MSI and stage using undersampling: a subsample of the same number of patients and the same distribution according to these clinical variables was selected from the GSE39582 series for each dataset. Expression values were truncated to the maximum and minimum values observed in the reference dataset.

Gene screening for association with relapse

Each gene in the metacohort was evaluated for linear association of its expression with recurrence using a frailty Cox proportional hazards model^{86,87}, including dataset and technical variables. Statistical significance was assessed using a Wald test. HR values and their corresponding 95% confidence intervals were computed as a measure of association. The 2,530 genes with HR > 1 and $P < 0.05$ were included in the all_HR signature.

We used the GSE39397 dataset⁶, which includes expression profiles of epithelial cancer cells (EPCAM⁺), CAFs (FAP⁺), leukocytes (CD45⁺) and endothelial cells (CD31⁺) isolated by FACs from dissociated primary CRCs ($n = 14$), to classify genes according to their expressions in these populations. The EpiHR signature contains genes from the allHR signature, which are upregulated (fold change > 1; $P < 0.05$) in EPCAM⁺ cells compared with the three TME populations (FAP⁺, CD31⁺ and CD45⁺). AllHR genes that did not pass this cut-off comprised the TME-HR signature.

Signature scores were computed as the scaled mean of the genes in the signature after scaling the expression matrix. The association with recurrence of the EpiHR signature in the whole dataset as well as the subclasses of consensus molecular subtype samples was assessed as described above. The likelihood ratio test comparing the EpiHR and AllHR P value was computed using the drop1R function. Kaplan–Meier plots were generated using the survfit and plot functions.

Association between clinical variables and the EpiHR signature in the CRC metacohort was assessed by fitting a linear model for each variable of interest independently. Technical factors (dataset and centre) were included as covariates.

Association between oncogenic alterations and the EpiHR score

Annotations of oncogenic alterations for the TCGA samples were downloaded from ref. ⁸⁸. A Wilcoxon test comparing the expression of mutated versus wild-type samples was performed independently for every alteration. The difference in expression median values was used as a measure of the impact of each mutation in the gene expression.

Patient 10x single-cell analysis

Count matrices were downloaded from ArrayExpress (E-MTAB-8107 for samples EXT001, EXT002, EXT003, EXT009, EXT010, EXT011, EXT012, EXT013, EXT014, EXT018, EXT019, EXT020, EXT021, EXT022, EXT023, EXT024, EXT025, EXT026, EXT027 and EXT028) and GEO (GSE132465 for all Samsung Medical Center tumour (SMC-T) samples)⁸. The remaining EXT samples were processed as referred in E-MTAB-8107 and deposited in ArrayExpress under accession number E-MTAB-9934. Cells with mitochondrial content higher than 20%, less than 1,000 counts, more than 6,000 or less than 200 genes were discarded. Ribosomal genes

Article

were also removed from the matrix to avoid technical biases during normalization. Samples with less than 500 cells that were not classified as core tumour were discarded from further analyses. The Korea (SMC samples) and Leuven (EXT samples) cohorts were processed independently according to the recommendations of the R package Seurat v.3⁶²: samples were combined and normalized using the SCT-transform function regressing mitochondrial percentage, with the method “glmGamPoi”, min.cells = 1 and return.only.genes=FALSE to keep the maximum number of genes. Dimensionality reduction and visualization were performed using RunPCA and RunUMAP, with 26 principal components. Expression was imputed and smoothened using the MAGIC algorithm⁶⁴. The expression of the *EPCAM* gene was used to define the connected components corresponding to epithelial cells.

The epithelial component of each cohort was processed as follows: cells with less than 1,000/3,000 (SMC/KUL) counts and less than 200/1,250 genes detected were removed from the dataset. No further normalization was applied. RunPCA, RunUMAP, FindNeighbors and FindClusters were applied, with 5/7 principal components and resolution of 0.7. Expression was imputed and smoothened using the MAGIC algorithm. Signature scores were computed as the mean value of the MAGIC expression per cell for all genes in the signature. The EpiHR and *Lgr5* cell populations were defined as cells with the corresponding signature expression above the 75th percentile. Population markers were found using FindMarkers. Functional enrichment was computed through the GSEA implementation in ref.⁸⁹ with genes ordered by fold change. Samples in the SMC and KUL datasets were annotated according to their iCMS class¹¹.

To identify the core gene expression program upregulated in HRCs, we computed the correlation scores for all genes with the EpiHR signature in the SMC and KUL human dataset. The resulting list was ranked by the average correlation across all samples (Supplementary Table 2). The coreHRC signature was defined as the top 100 genes in the SMC dataset.

Association of HRCs and TME populations

We classified clusters according to the expression of known markers of microenvironment components⁸. For each population we estimated the association between its percentage per sample and the percentage of HRCs. Spearman correlation coefficient and *P* value were used to assess said association.

Clustering of EpiHR genes according to expression correlation

We computed the Pearson correlation coefficient for all pairwise combinations of genes in the EpiHR signature. We then applied hierarchical clustering (hclust in R with method ‘complete’) and defined 6 clusters using the cutree function. After visual inspection, we decided to merge the three clusters with higher correlation, resulting in cluster 1 (Extended Data Fig. 1h).

Ethics oversight

Samples of primary CRC from patients used for IF and ISH analysis were obtained from the Hospital Clinic de Barcelona-IDIBAPS Biobank (B.0000575), which is integrated in the Spanish National Biobanks Network. Samples were donated by patients under informed consent and they were processed according to standard operating procedures with the appropriate approval of the Ethics and Scientific Committee of the Hospital Clinic de Barcelona (register code: HCB/2020/1478) and according to the guidelines of the European Network of Research Ethics Committees, following European, national and local laws.

Statistics

No statistical test was used to determine sample size upfront. Instead, sample size was determined empirically according to previous knowledge of the variation in the experimental set-up^{4,12,34}. A minimum of four mice were quantified in each experiment and for each condition. For the majority of in vitro experiments, we used $n > 3$ according to

previous experience with similar experiments. Additional information is provided in the Reporting summary. Automated blind quantifications and blind data analysis were performed whenever possible. The sample size typically results in a standard error <25% of the mean. No data from in vitro or in vivo experiments were excluded, except for the CRC relapse model, in which a small fraction of mice, typically 1 in every 10 in each experiment, were not included in the follow up due to invasion into the ileocaecal junction, which impeded successful surgery. Occasionally, mice bearing CRCs were euthanized 1–2 days after surgery owing to significant weight loss or unhealthy aspect, according to protocols approved by the animal experimental committees. Data distributions were assumed to be normal but this was not formally tested. Transformations were applied whenever needed. For percentage data, normality was assumed for values far from 0 or 100. Statistical analyses were performed using the R software package (v.4.0.5) and GraphPad Prism (v.7.03).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data relevant to this study are available from the corresponding author on reasonable request. Expression arrays and RNA-seq data are available at the Gene Expression Omnibus (GEO). Accession numbers for gene expression sequencing experiments reported in this paper are GEO: GSE190055 (arrays, EMP1^{high} versus EMP1^{low} AKTP tumour cells), GSE208139 (arrays, MTOs co-cultured with fibroblasts), GSE207974 (RNA-seq chemotherapy) and GSE207668 (RNA-seq CTOs). Count matrices for single-cell RNA-seq experiments were deposited at Array-Express under accession number E-MTAB-11284 (10x AKTP primary tumours), E-MTAB-11302 (Smart-seq metastatic progression) and E-MTAB-11981 (Smart-seq AKP micrometastases). Additional metadata and processed data files, including UMAP embeddings and gene signature scores, are available at Synapse (syn35000645). Source data are provided with this paper.

40. Matano, M. et al. Modeling colorectal cancer using CRISPR-Cas9-mediated engineering of human intestinal organoids. *Nat. Med.* **21**, 256–262 (2015).
41. Drost, J. et al. Sequential cancer mutations in cultured human intestinal stem cells. *Nature* **521**, 43–47 (2015).
42. Céspedes, M. V. et al. Orthotopic microinjection of human colon cancer cells in nude mice induces tumor foci in all clinically relevant metastatic sites. *Am. J. Pathol.* **170**, 1077–1085 (2007).
43. Chen, Y.-C. et al. Gut fecal microbiota transplant in a mouse model of orthotopic rectal cancer. *Front. Oncol.* **10**, 568012 (2020).
44. Conti, S. et al. CAFs and cancer cells co-migration in 3D spheroid invasion assay. *Methods Mol. Biol.* **2179**, 243–256 (2020).
45. Gonzalez-Roca, E. et al. Accurate expression profiling of very small cell populations. *PLoS ONE* **5**, e14418 (2010).
46. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
47. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
48. Liao, Y., Smyth, G. K. & Shi, W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.* **47**, e47 (2019).
49. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
50. Carvalho, B. S. & Irizarry, R. A. A framework for oligonucleotide microarray preprocessing. *Bioinformatics* **26**, 2363–2367 (2010).
51. Bolstad, B. M. et al. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (eds Gentleman, R. et al.) (Springer, 2005).
52. Fridlyand, J. Microarray Data Analysis. in *Selected Works in Probability and Statistics* (ed Dudoit, S.) https://doi.org/10.1007/978-1-4614-1347-9_15 (Springer, 2012).
53. Ritchie, M. E. et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
54. Eklund, A. C. & Szallasi, Z. Correction of technical bias in clinical microarray data improves concordance with known biological information. *Genome Biol.* **9**, R26 (2008).
55. Wu, D. et al. ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics* **26**, 2176–2182 (2010).
56. Efron, B. & Tibshirani, R. On testing the significance of sets of genes. *Ann. Appl. Stat.* **1**, 107–129 (2007).

57. Lee, E., Chuang, H. Y., Kim, J. W., Ideker, T. & Lee, D. Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.* **4**, e1000217 (2008).
58. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
59. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
60. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
61. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
62. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
63. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
64. van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729 (2018).
65. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
66. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. zUMIs—a fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience* **7**, giy059 (2018).
67. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
68. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).
69. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2020).
70. Barrett, T. & Edgar, R. [19] Gene Expression Omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol.* **411**, 352–369 (2006).
71. Grossman, R. L. et al. Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).
72. Muzny, D. M. et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
73. Tripathi, M. K. et al. Nuclear factor of activated T-cell activity is associated with metastatic capacity in colon cancer. *Cancer Res.* **74**, 6947–6957 (2014).
74. Sanz-Pamplona, R. et al. Aberrant gene expression in mucosa adjacent to tumor reveals a molecular crosstalk in colon cancer. *Mol. Cancer* **13**, 46 (2014).
75. Kemper, K. et al. Mutations in the Ras-Raf axis underlie the prognostic value of CD133 in colorectal cancer. *Clin. Cancer Res.* **18**, 3132–3141 (2012).
76. Jorissen, R. N. et al. Metastasis-associated gene expression changes predict poor outcomes in patients with dukes stage B and C colorectal cancer. *Clin. Cancer Res.* **15**, 7642–7651 (2009).
77. Marisa, L. et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med.* **10**, e1001453 (2013).
78. Laipe, S. et al. A seven-gene signature aggregates a subgroup of stage II colon cancers with stage III. *OMICS* **16**, 560–565 (2012).
79. Jorissen, R. N. et al. DNA copy-number alterations underlie gene expression differences between microsatellite stable and unstable colorectal cancers. *Clin. Cancer Res.* **14**, 8061–8069 (2008).
80. Azzalini, A. & Menardi, G. Clustering via nonparametric density estimation: the R package pdfcluster. *J. Stat. Softw.* **57**, 1–26 (2014).
81. Azzalini, A. & Torelli, N. Clustering via nonparametric density estimation. *Stat. Comput.* **17**, 71–80 (2007).
82. Smedley, D. et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.* **43**, W589–W598 (2015).
83. Drost, H. G. & Paszkowski, J. Biomart: genomic data retrieval with R. *Bioinformatics* **33**, 1216–1217 (2017).
84. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinform.* **12**, 323 (2011).
85. Bates, D., Mächler, M., Bolker, B. M. & Walker, S. C. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
86. Therneau, T. M., Grambsch, P. M. & Pankratz, V. S. Penalized survival models and frailty. *J. Comput. Graph. Stat.* **12**, 156–175 (2003).
87. Therneau, T. coxme: mixed effects Cox models. R package version 2.2-3 www.cran.r-project.org/package=coxme. Oikos (2012).
88. Sanchez-Vega, F. et al. Oncogenic signaling pathways in the Cancer Genome Atlas. *Cell* **173**, 321–337 (2018).
89. Mootha, V. K. et al. PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).

Acknowledgements We thank all of the members of the laboratory for their support and discussions; the staff at the IRB Barcelona core facilities for biostatistics, histopathology, functional genomics and advanced digital microscopy, as well as the flow cytometry, animal facilities of the UB/PCB, and the CRG genomic unit for assistance; the staff at the HCB-IDIBAPS Biobank for sample and data procurement. Sample collection of this work was also supported by the Xarxa de Bancs de Tumors de Catalunya (XBTC) sponsored by Pla Director d'Oncologia (PDO). A.C.-S., G.T. and L.J.-G. have held FPU fellowships from Spanish Ministry of Universities and A.A.-V. has held a La Caixa predoctoral fellowship. S.C. holds a FPI fellowship from the Spanish Ministry of Economy and Competitiveness (MINECO). H.H. is a Miguel Servet (CP14/00229) researcher funded by the Spanish Institute of Health Carlos III (ISCIII) and the Agencia Estatal de Investigación (AEI) and FEDER (SAF2017-89109-P). This work has been supported by ERC advanced grants 884623 (residualCRC to E.B.) and 883739 (Epifold to X.T.); LCF/PR/HR19/52160018 from La Caixa foundation; PID2020-119917RB-I00 from the Spanish MICINN; and CRUK Accelerator Award C7932/A26825 (ACRCelerate), in collaboration with AECC (grant GEACC19006BAT_2021). E.B., X.T. and H.H. are supported by the Fundació La Marató de TV3 (201903-30-31-32). H.H. received support for the project PID2020-115439GB-I00- funded by MCINN/AEI/10.13039/501100011033. S.L. was supported by a Wellcome Trust Senior Clinical Research Fellowship (206314/Z/17/Z), E.J.M. is funded by the Lee Placito Medical Research Fund (University of Oxford). IRB Barcelona and IBEc are recipients of a Severo Ochoa Award of Excellence from MINECO. Single-cell profiling of CRC samples was supported by the Belgian Federation against Cancer grant nos 2018-127 and 2016-133 and by a grant from Fondation Roi-Baudouin. S.T. is supported by a Fundamental Clinical Researchers KU Leuven grant and Foundation against Cancer grant for this work. The Magdalena Socias Moyá fund supports metastasis research at IRB Barcelona.

Author contributions E.B. and A.C.-S. conceived the study, coordinated experiments and wrote the manuscript. A.C.-S. designed and performed key experiments including the profiling of residual disease, the analyses of tumour buds and micrometastases, genetic ablation studies and immunotherapy experiments in the CRC relapse model. A.C.-S., C.C., G.T., F.S. and D.S. generated and characterized MTO knock-in lines. A.C.-S., X.H.-M., S.P.-P. and G.T. developed the CRC relapse model. A.C.-S., X.H.-M. and S.P.-P. performed all mouse work. L.M. and C.S.-O.A. analysed scRNA-seq data. C.S.-O.A. and A.B.-L. analysed human CRC transcriptomic datasets. A.C.-M., L.M. and C.S.-O.A. performed statistical analyses. C.C. and T.S. performed IF and imaged organoids in vitro. A.C.-S., C.C. and O.R. generated and characterized YAP-KD models. A.C.-S., C.C., O.R., S.C. and X.T. performed in vitro co-cultures. A.C.-S. and A.A.-V. quantified immunofluorescence stainings. A.C.-S. and N.F. developed the method to purify residual tumour cells in whole livers. M.S. performed IF and IHC. E.J.M. and S.L. performed multiplex IF. L.N. performed chemotherapy experiments. L.B. and J.C. performed 3D light-sheet imaging. L.J.-G., C.C., P.L. and H.H. provided support with scRNA-seq experiments. S.T. generated scRNA-seq data from CRC patient samples. D.V.F.T. and D.S. generated MTO and CTO biobanks. E.S. provided strategic support and helped with figures and manuscript writing. E.B. supervised the study.

Competing interests The authors declare no competing interests.

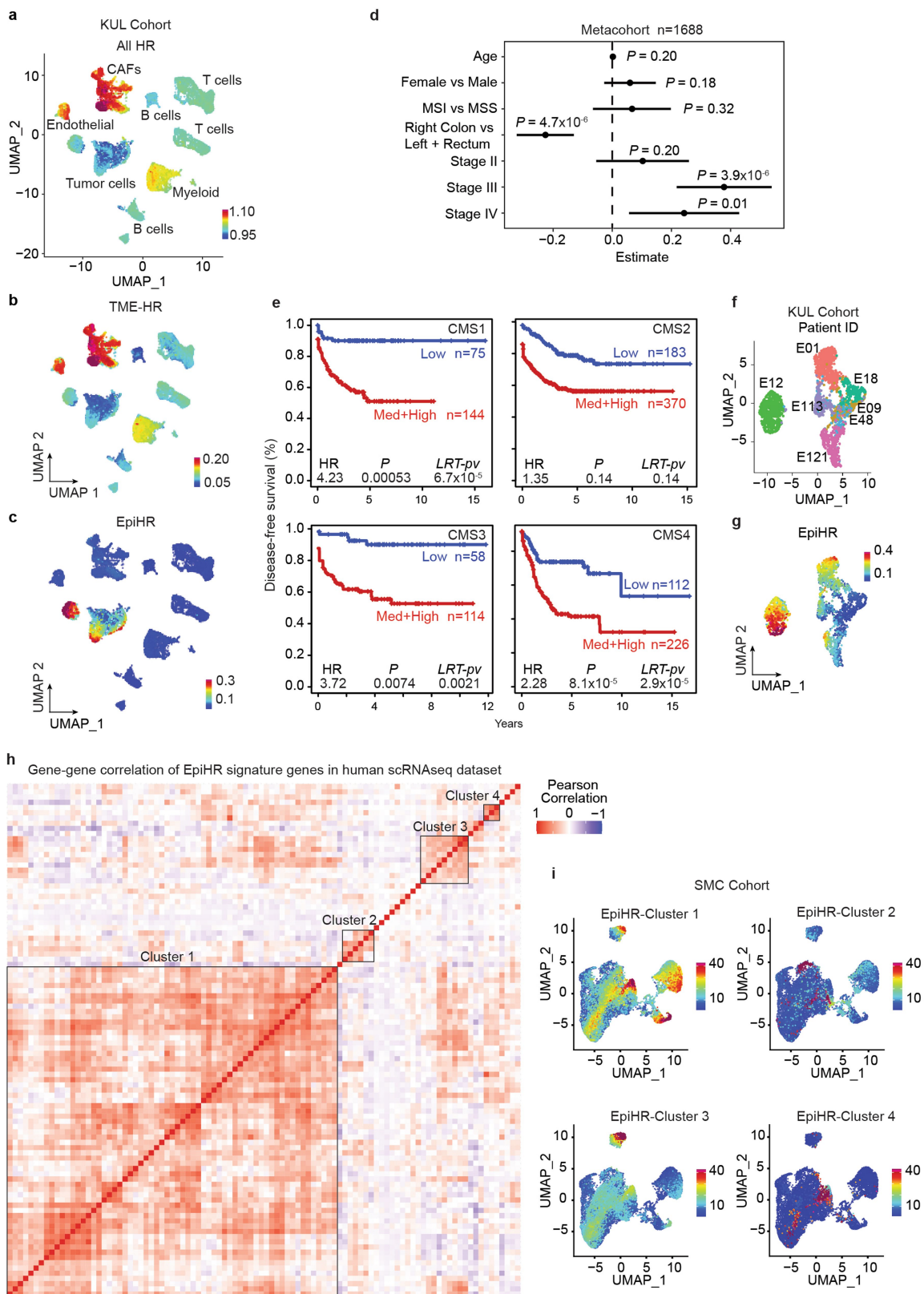
Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-022-05402-9>.

Correspondence and requests for materials should be addressed to Eduard Batlle.

Peer review information *Nature* thanks Hugo Snippert, Itay Tirosh, Nicola Valeri and the other, anonymous, reviewers for their contribution to the peer review of this work.

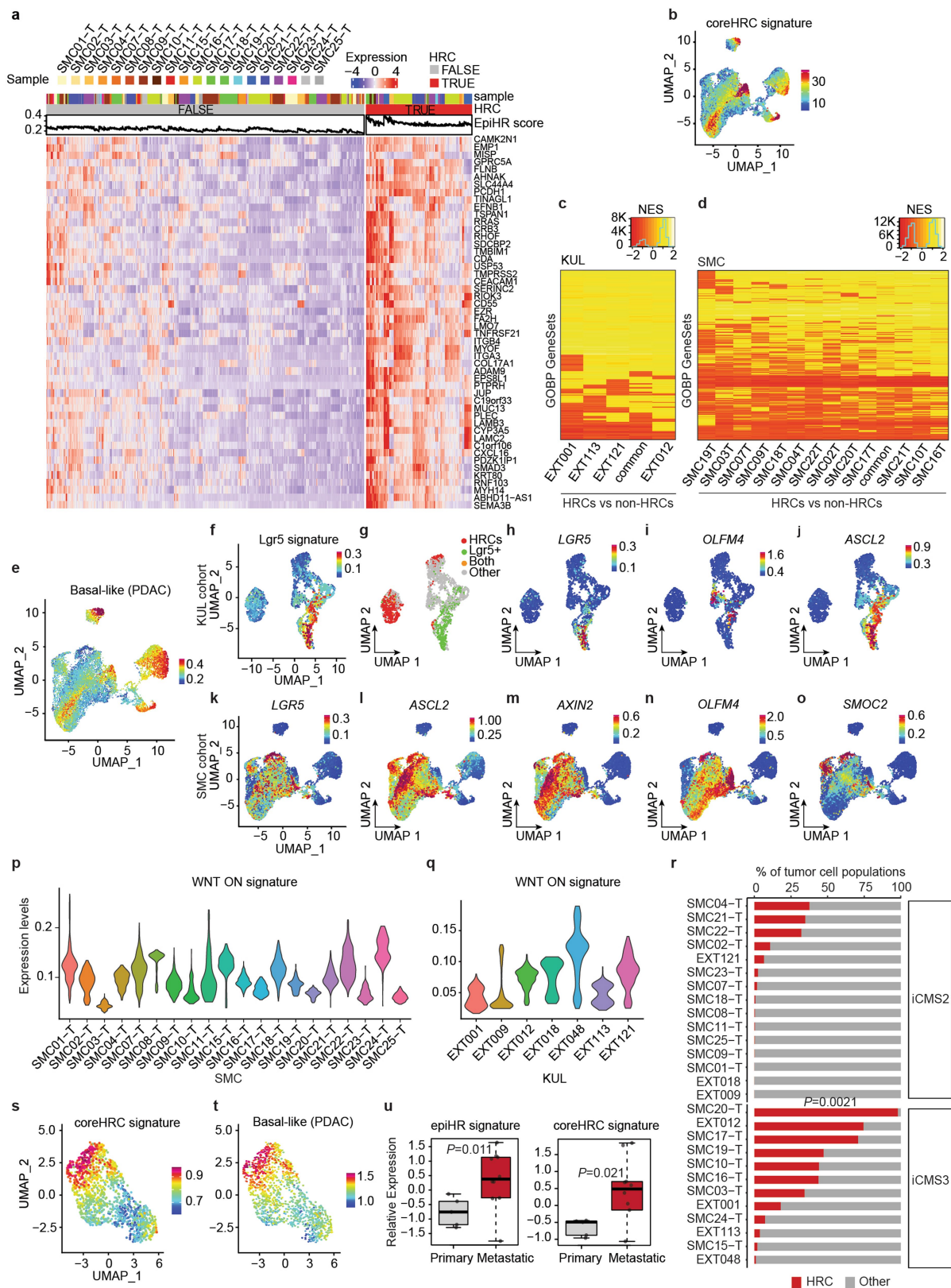
Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | The EpiHR gene set marks a defined tumour cell population across CRCs. **a-c**, UMAP layout of whole tumours (stroma + epithelium cells) from 7 CRC patients in the KUL dataset. Coloured by **(a)** gene expression of all high hazard ratio genes (AllHR), **(b)** tumour microenvironment-specific HR genes (TME-HR), and **(c)** epithelial-specific HR genes (EpiHR). **d**, Association between clinical variables and the EpiHR signature in the CRC meta cohort was assessed by fitting a linear model for each variable independently. Technical factors (dataset and centre, as described in extended methods) were included as covariates. Lines show the left and right confidence intervals. n = 1688 patients. **e**, Kaplan-Meier survival curves indicating relapse-free

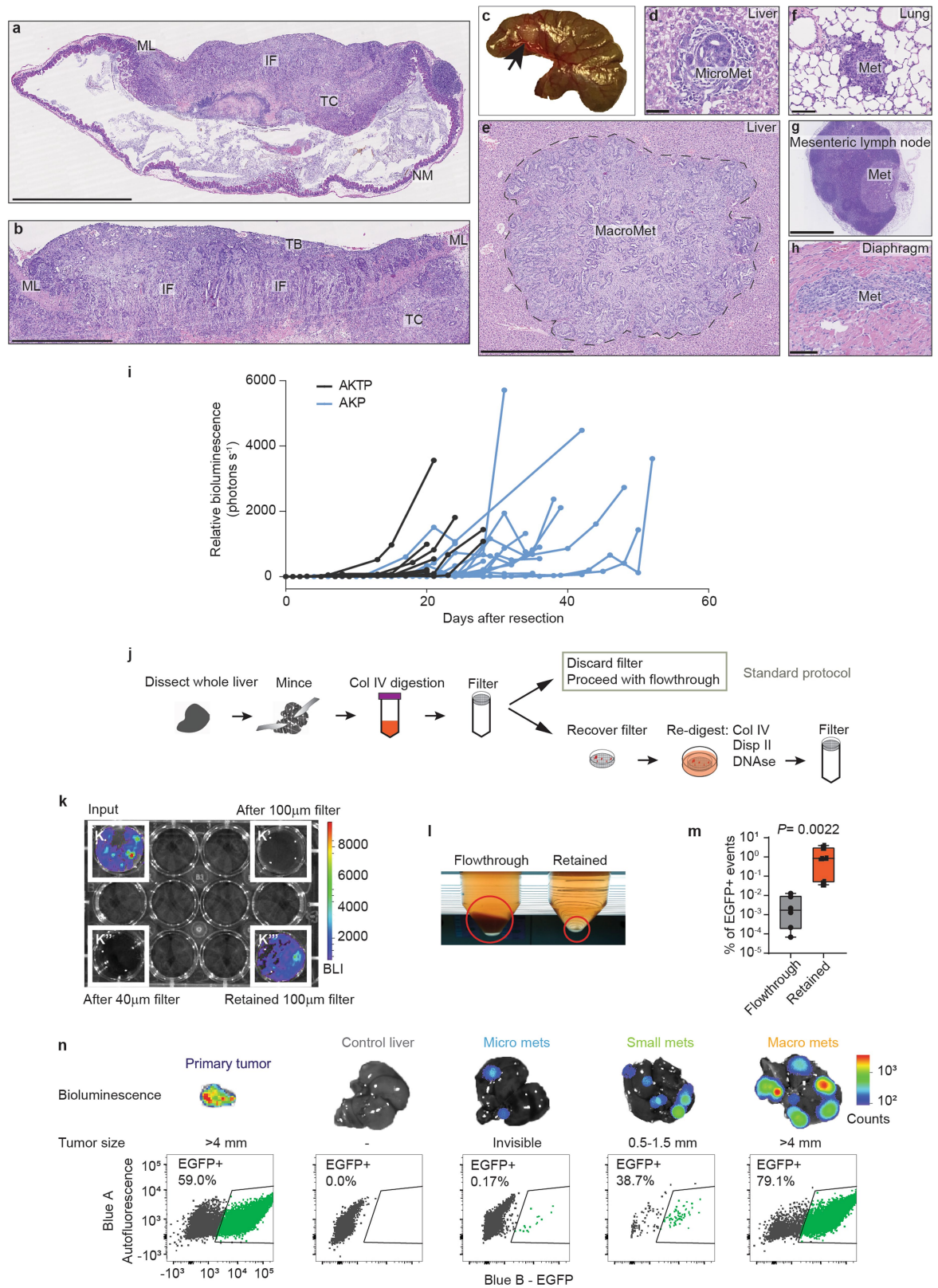
survival according to EpiHR gene signature expression for CRC patients classified by CMS. Two-sided Wald test. **f-g**, UMAP layout of 2718 CRC tumour cells from the KUL cohort coloured by **(f)** patient ID and **(g)** expression of the EpiHR signature. **h**, Heatmap showing Pearson correlation scores in gene expression among EpiHR signature genes in patients from the SMC cohort. Note that most genes belong to one coherent subset (Cluster 1). Gene lists are detailed in Supplementary Table 2. **i**, UMAP layout of human CRC tumour cells coloured by the expression of genes belonging to Clusters 1, 2, 3 and 4 identified in **(h)**.



Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Characterization of HRC features. **a**, Heatmap showing scaled expression of the top 50 most correlated genes with the EpiHR signature across SMC patients. Tumour cells are divided as non-HRCs (left, FALSE) and HRCs (right, TRUE). The EpiHR signature score for each individual cell is plotted above the heatmap. **b**, UMAP layout of CRC tumour cells coloured by the expression of the coreHRC signature. The coreHRC signature is defined as the top 100 genes with better correlation with the EpiHR signature. **c-d**, Heatmap showing Normalized Enrichment Scores (NES) for Gene sets in Gene Ontology Biological Processes (GOBP) in HRCs from different patients in the KUL (**c**) and SMC (**d**) cohorts. Only GOBP gene sets with NES scores above 0.5 are shown. Gene sets and patients are ordered by hierarchical clustering. **e**, UMAP layout of human CRC tumour cells in the SMC cohort painted with the Basal cell state signature in Pancreatic Ductal Adenocarcinoma (PDAC) by Raghavan et al.¹⁰. **f**, UMAP layout of tumour cells from the KUL cohort showing the expression of the *Lgr5* signature. **g**, UMAP of same tumour cells labelled according to their classification into HRCs, *Lgr5*⁺, double positive or other

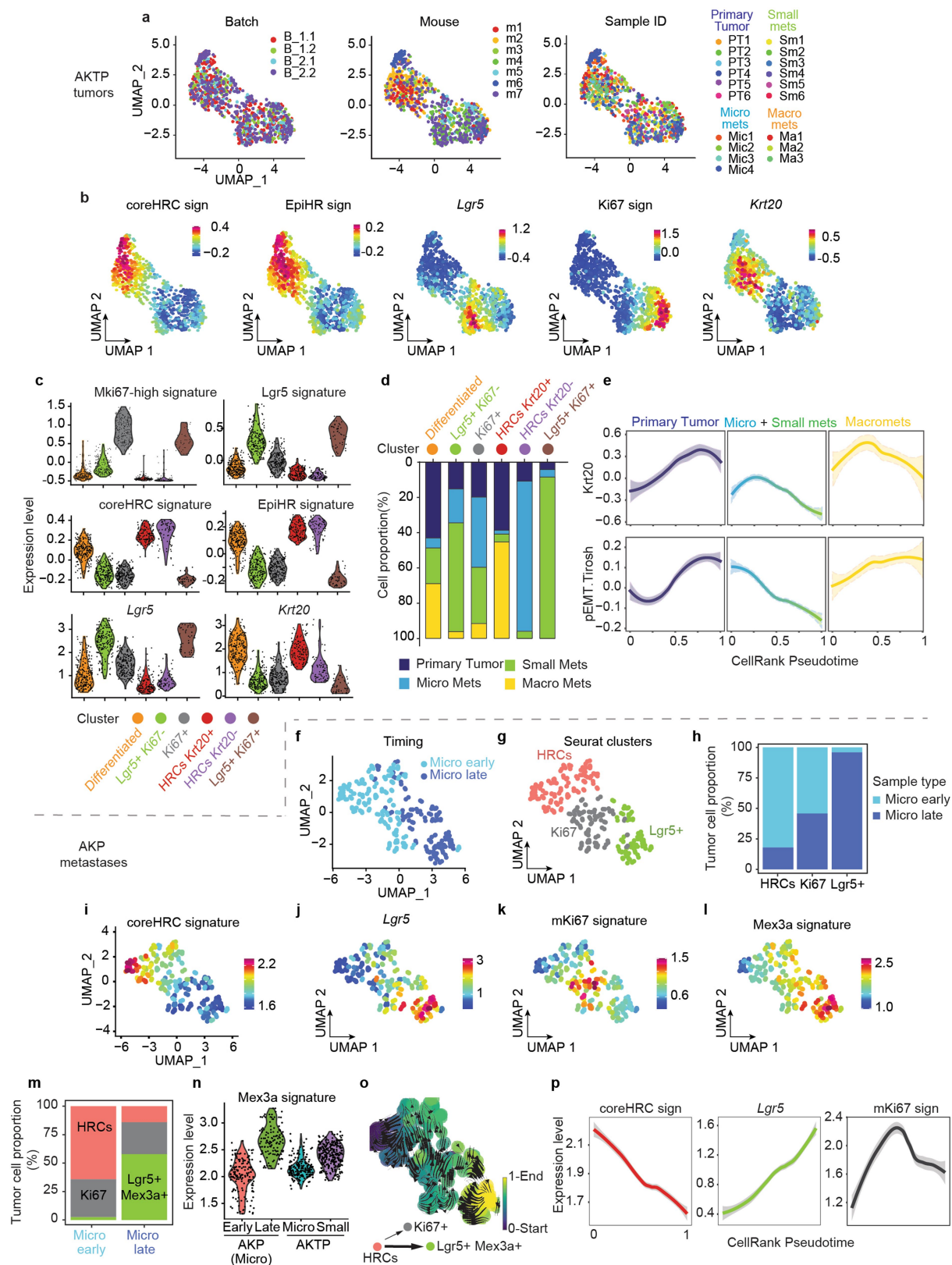
cells. **h-j**, UMAP layout of same tumour cells showing gene expression levels of canonical intestinal stem cell genes *LGR5*, *OLFM4* and *ASCL2*. **k-o**, UMAPs of tumour cells in the SMC dataset showing gene expression levels of canonical intestinal stem cell genes *LGR5*, *ASCL2*, *AXIN2*, *OLFM4*, and *SMOC2*. **p,q**, Violin plots showing WNT-ON signature expression levels in epithelial tumour cells from patients in the SMC (**p**) and KUL (**q**) cohorts. **r**, Barplot quantifying the HRC composition of each patient (combined SMC and KUL datasets). Patients are classified as iCMS2 or iCMS3 according to Joanito et al.¹¹. Two-sided Kruskal-Wallis test. **s,t**, UMAPs of mouse CRC AKTP tumour cells coloured according to (**s**) the coreHRC signature and (**t**) the Basal cell state signature in Pancreatic cancer by Raghavan et al.¹⁰ **u**, Gene expression levels of EpiHR (left) and coreHRC (right) signatures in MTOs derived from primary tumours or from liver metastases. Boxes represent the first, second (median) and third quartiles. Whiskers indicate maximum and minimum values. Welch two-sided t-test. n = 5 (primary) 10 (metastatic).



Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Analysis of the CRC relapse mouse models and purification of DTCs. **a**, Representative micrograph of hematoxylin- and eosin (HE)-stained adenocarcinoma with subserosal invasion (T4) generated by injection of an AKTP MTO in the mouse caecum. Tumour centre (TC), invasive fronts (IF), muscle layer (ML) and normal mucosa (NM) are indicated. Scale bar, 2.5 mm. **b**, Representative image of a different T4 tumour penetrating the muscle layer (ML) and reaching the serosa layer. TB: Tumour buds. Scale bar, 1 mm. **c**, Picture of a caecum 21 days after injection and imaged at the time of surgery showing a primary tumour (arrow) in the distal part. **d-e**, Haematoxylin-Eosin (HE) staining of micrometastases and large metastases observed in the liver of orthotopic isografted mouse. Scale bars, 50 μ m and 1 mm, respectively. In **e**, tumoral tissue is surrounded by dashed lines. **f-h**, HE staining of lung (**f**), lymph node (**g**) and diaphragm (**h**) metastases from orthotopic isografted mice. Scale bars, 100 μ m (**f** and **h**) 1 mm (**g**). **i**, Graph showing liver longitudinal BLI measurements (photons per second), normalized to the day of primary tumour resection. Points and lines represent individual mice. $n = 9$ (AKTP), 24 (AKP) mice. **j**, Schematic representation of a novel tissue-dissociation strategy that enables recovery of DTCs from livers. Whole livers are dissected and minced thoroughly. After a mild collagenase IV digestion, samples are filtered through 100 μ m meshes. The filter retained sample is highly enriched in

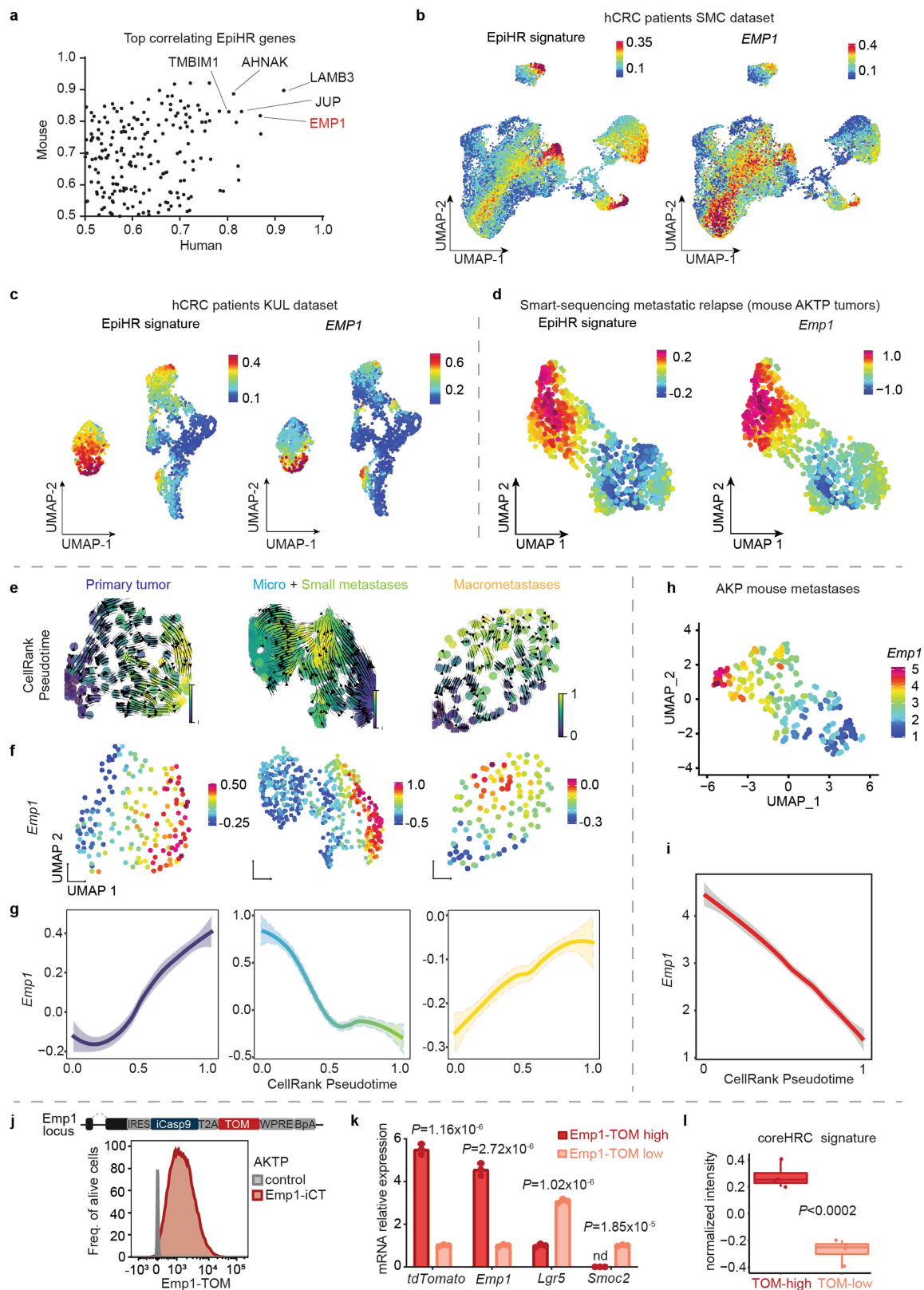
tumour cells. Remaining tissue in the filter is re-digested with a stronger enzymatic cocktail to fully digest it, and then re-filtered. **k**, Representative bioluminescent image of a whole liver sample containing luciferase+ tumour cells before enzymatic digestion (B, input), after filtering through 100 μ m (B') and 40 μ m (B'') meshes (previous protocol), and after recovering and re-digesting tissue retained in the 100 μ m filter (B'''). **l**, Image showing the large cell pellet containing liver cells after 1 mild digestion and the small pellet in the retained and re-digested sample enriched in DTCs. **m**, Percentage of GFP+ cells measured by flow cytometry in samples with 1 round of digestion compared to re-digested samples. Boxes represent the first, second (median) and third quartiles. Whiskers indicate maximum and minimum values. Paired two-sided Wilcoxon test on percentages. $n = 6$ independent paired samples examined in 2 independent experiments. **n**, Representative bioluminescent images, tumour burden and flow-cytometry plots of the 4 different stages analysed by single-cell Smart-sequencing described in Fig. 2. Micrometastases samples were DTCs collected from livers with absent or low bioluminescence in which metastases were not visible. For small metastases samples, metastatic nodules were visible but small in size (<1.5 mm). Macrometastases samples were metastatic nodules larger than 4mm.



Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Additional description of residual AKTP and AKP metastatic cells. **a**, UMAPs of colorectal primary tumours and liver metastases at different stages (micro, small and large) coloured according to sequencing batch, mouse ID, and sample ID. **b**, UMAPs showing the expression levels of coreHRC, EpiHR, and mKi67 gene signatures and *Lgr5* and *Krt20* genes. **c**, Violin plots showing expression of relevant genes used to define the 6 different Seurat clusters. **d**, Fraction of cells (y axis) from each Seurat cluster (x axis) present in the different sample types: Primary Tumour, micro-, small- and macro- metastases according to the indicated color code. Note the “HRCs Krt20-” are mostly exclusive from micro metastases samples, whereas *Lgr5*+ cells are highly enriched in small metastases samples. **e**, Smoothed *Krt20* gene and partial EMT gene signature¹⁷ expression trends fitted with Generalized Additive Models as a function of pseudotime in primary tumours, micro+small

and large metastases. **f, g**, UMAP of AKP liver micrometastases coloured according to timing of profiling and Seurat clusters. **h**, Barplot showing proportion of different Seurat tumour cell types captured in AKP early vs late micrometastases. **i-l**, UMAPs showing the expression levels of the coreHRC, mKi67 and *Mex3a* gene signatures and *Lgr5* mRNA in AKP metastases. **m**, Barplot showing Seurat cluster distribution across AKP early and late micrometastases. **n**, Violin plots showing expression levels of the *Mex3a* signature¹⁶ in AKP early and late micrometastases versus AKTP micro and small metastases. **o**, Vector fields representing RNA velocity projected on UMAPs of AKP micrometastases, coloured by the pseudotime estimated for each cell with scVelo. **p**, Smoothed coreHRC, mKi67, and *Lgr5* gene signature expression trends in the early and late AKP micrometastasis dataset fitted with generalized additive models as a function of CellRank pseudotime.

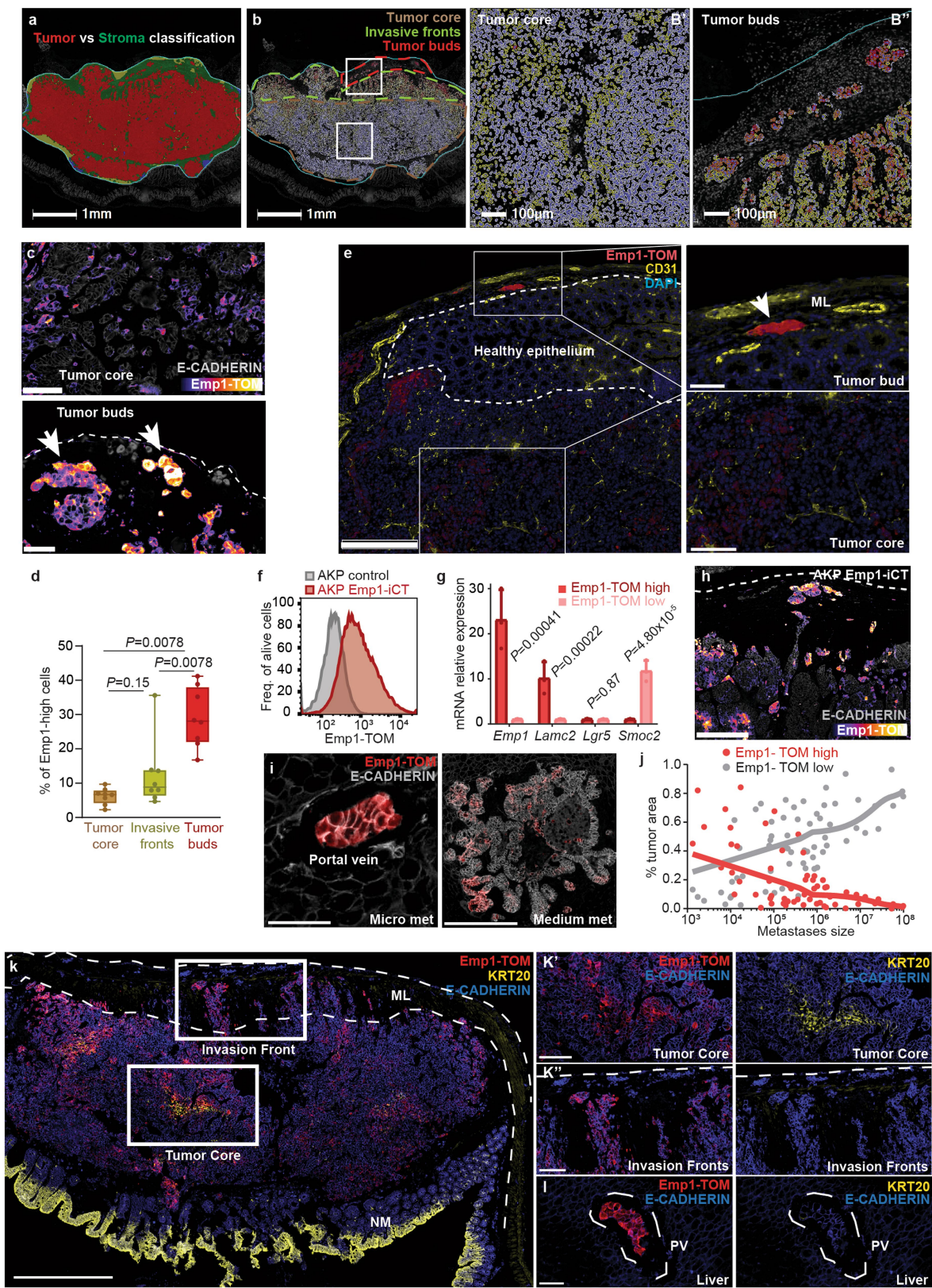


Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Epithelial membrane protein 1 (EMPI) marks HRCs.

a, Scatter plot showing the correlation value between individual genes in the human SMC cohort (x axis) and in mouse primary tumours (y axis) with the EpiHR signature. Genes with correlation scores higher than 0.8 in both datasets are highlighted. **b**, UMAP of tumour cells from CRC patients in the SMC dataset coloured according to the expression of EpiHR signature (left) and of *EMPI* gene (right). **c**, As in **b**, for CRC tumour cells from the KUL datasets. **d**, UMAP representation of Smart-sequencing single cell data of AKTP mouse tumour cells along metastatic relapse sequence coloured by the EpiHR signature (left) and *Emp1* gene (right). **e**, Vector fields representing RNA velocity projected on AKTP primary CRC, micro+small and macro metastases UMAPs, coloured by the pseudotime estimated for each cell with scVelo. **f**, AKTP tumour cell UMAPs coloured by *Emp1* gene expression. **g**, Smoothed *Emp1* gene expression trends fitted with Generalized Additive Models as a

function of pseudotime in AKTP primary tumour, micro+small and macro metastases samples. **h**, UMAP representation of AKP micrometastases coloured by *Emp1* gene expression. **i**, Smoothed *Emp1* gene expression trends fitted with Generalized Additive Models as a function of pseudotime in AKP micrometastases samples. **j**, Representative flow cytometry plot of TOM expression in *wild-type* and *Emp1-ICT* AKTP MTOs. **k**, Relative mRNA expression of indicated marker genes in *Emp1-TOM^{high}* and *Emp1-TOM^{low}* sorted cell populations from *Emp1-ICT* AKTP MTOs *in vitro*. Two-sided t-test after normalizing by *Ppia*. n = 3 technical replicates. Mean $\pm$ SD. **l**, Boxplot showing normalized intensity of coreHRC signature expression in *Emp1-TOM^{high}* and *Emp1-TOM^{low}* cells dissociated from primary tumours 4 weeks post-implantation. Box plots have whiskers of maximum 1.5 times the interquartile range; boxes represent first, second (median) and third quartiles. n = 4 mice per condition. ROAST-GSA adjusted p-values are shown.

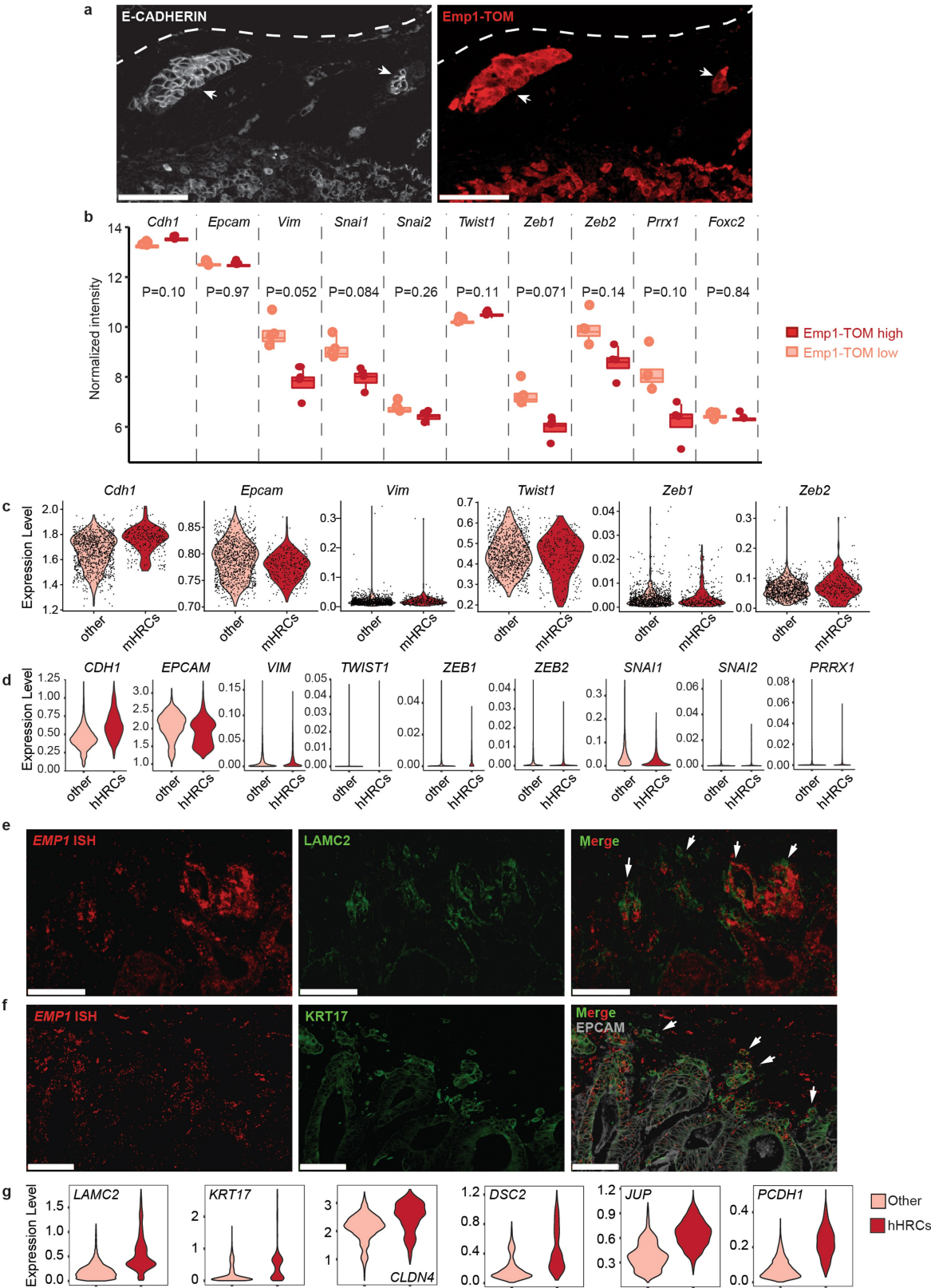


Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | HRCs are enriched in invasion fronts and micrometastases.

a, Primary tumour outlined by cyan line and coloured in 4 different regions identified with HALO image analysis classifier (tumour-red, stroma-green, background-yellow, necrosis-blue). Scale bar, 1 mm. **b**, TOM cell intensity analysis in the tumour area after segmentation into individual cells. B' and B'' show magnified regions corresponding to tumour core (B') and invasion fronts + tumour buds (B''). Scale bars, 1 mm (B), 100 μ m (B' and B''). **c**, Representative immunostaining of TOM and E-CADH in the tumour core and in tumour buds of primary tumours derived from Emp1-iCT MTOs 4 weeks post implantation in the caecum. TOM fluorescence is shown with mpl-inferno LUT. The dashed line delimits the caecum edge. Arrows point to tumour buds. Scale bars, 100 μ m (tumour core) 50 μ m (tumour buds). **d**, Quantification of Emp1-TOM^{high} (defined as cells in percentile 90 for TOM expression) in the tumour core (submucosal area), invasion fronts (inside muscular layer) and isolated glands (over muscular layer). Boxes represent the first, second (median) and third quartiles. Whiskers indicate maximum and minimum values. Two-sided Wilcoxon test on percentages. n = 8 mice. **e**, Immunofluorescence of TOM, CD31 and DAPI in primary tumours. Amplified insets show the tumour core and invasive glands intermingled in mucosal layers (ML) next to blood vessels. Dashed lines outline healthy intestinal epithelium. Scale bars, 250 μ m, 100 μ m (tumour core) and 50 μ m (tumour buds). **f**, Representative flow cytometry plot

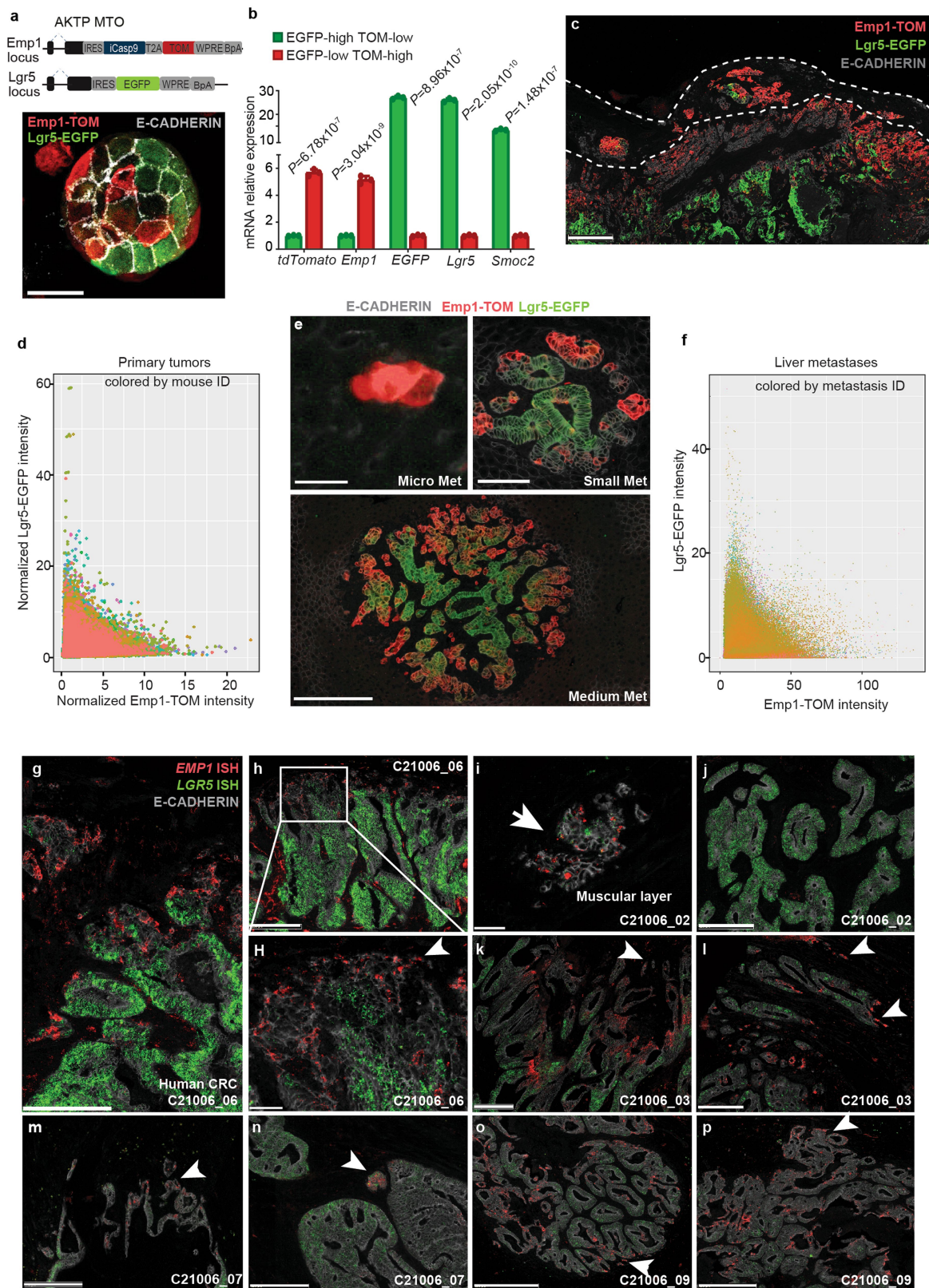
of TOM expression in *wild-type* and Emp1-iCT AKP MTOs. **g**, Relative mRNA expression of indicated genes in Emp1-TOM^{high} and Emp1-TOM^{low} sorted cell populations from Emp1-iCT AKP MTOs. Two-sided t-test after normalizing by *Ppia*. n = 3 technical replicates. Mean $\pm$ SD. **h**, Representative immunostaining for TOM and E-CADHERIN in Emp1-iCT AKP tumours implanted in the caecum 4 weeks post-implantation. Emp1-TOM fluorescence is shown with an mpl-inferno LUT. Dashed lines delimit the edge of the caecum. Scale bar: 250 μ m. **i**, Representative images of TOM and E-CADHERIN staining in micro (left) and medium (right) size metastases. Scale bars: 50 μ m and 250 μ m. **j**, Percentage of tumour area containing TOM-high and TOM-low fluorescent pixels versus metastases size (in pixels). Each dot represents an individual metastasis. **k**, TOM, KRT20 and E-CADHERIN staining in primary tumours generated by Emp1-iCasp9-tdTomato AKTP MTOs. Dashed lines encompass invasion fronts and tumour buds. KRT20 staining is observed in normal mucosa (NM) and to a lesser extent in the tumour core. Tumour cell clusters invading the muscular layer (ML) express high levels of TOM and no KRT20. Amplified insets show an example of tumour core (K') and invasion fronts (K'') with TOM (left) and KRT20 (right) stainings. Scale bars, 500 μ m (k) and 100 μ m (K' and K''). **l**, Immunofluorescence of TOM and E-CADHERIN (left) and KRT20 and E-CADHERIN (right) in a cluster of tumour cells that enter the liver through a portal vein (PV, delimited with dashed lines). Scale bar, 50 μ m.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | HRCs retain an epithelial phenotype. **a**, Immunostaining of E-CADHERIN and TOM in Emp1-ICT primary tumours 4 weeks post-implantation of MTOs. Arrows point at examples of E-CADHERIN+ invasion fronts and tumour buds. Dashed lines show the caecum edge. Scale bars, 100 μ m. **b**, Boxplot showing normalized expression of genes related to EMT in Emp1-TOM^{high} versus Emp1-TOM^{low} cells. Box plots have whiskers of maximum 1.5 times the interquartile range; Boxes represent first, second (median) and third quartiles. P-value for differential expression with Linear Model for Microarray Analysis (limma). n = 4 biological replicates. **c-d**, Violin plots showing expression of selected EMT-related genes in HRCs versus the rest of other cells in mouse epithelial primary tumour cells (**c**) and human tumour cells from the SMC

cohort (**d**). Genes present in **b** not shown (Snai1 and Snai2) were undetected in (**c**). **e**, Representative example of *EMPI* mRNA FISH combined with LAMC2 immunofluorescence on human primary CRC tissue section showing an overlapping pattern of expression of *EMPI* and LAMC2 in invasion fronts and tumour buds (arrows). Scale bar, 100 μ m. **f**, Representative example of *EMPI* mRNA FISH combined with KRT17 and EPCAM immunofluorescence on human primary CRC tissue sections showing an overlapping pattern of expression of *EMPI* and KRT17 in invading fronts and tumour cell clusters (arrows). Scale bars, 100 μ m. **g**, Violin plots showing enrichment of *LAMC2*, *KRT17* and several cell-to-cell adhesion genes in HRCs (SMC cohort).

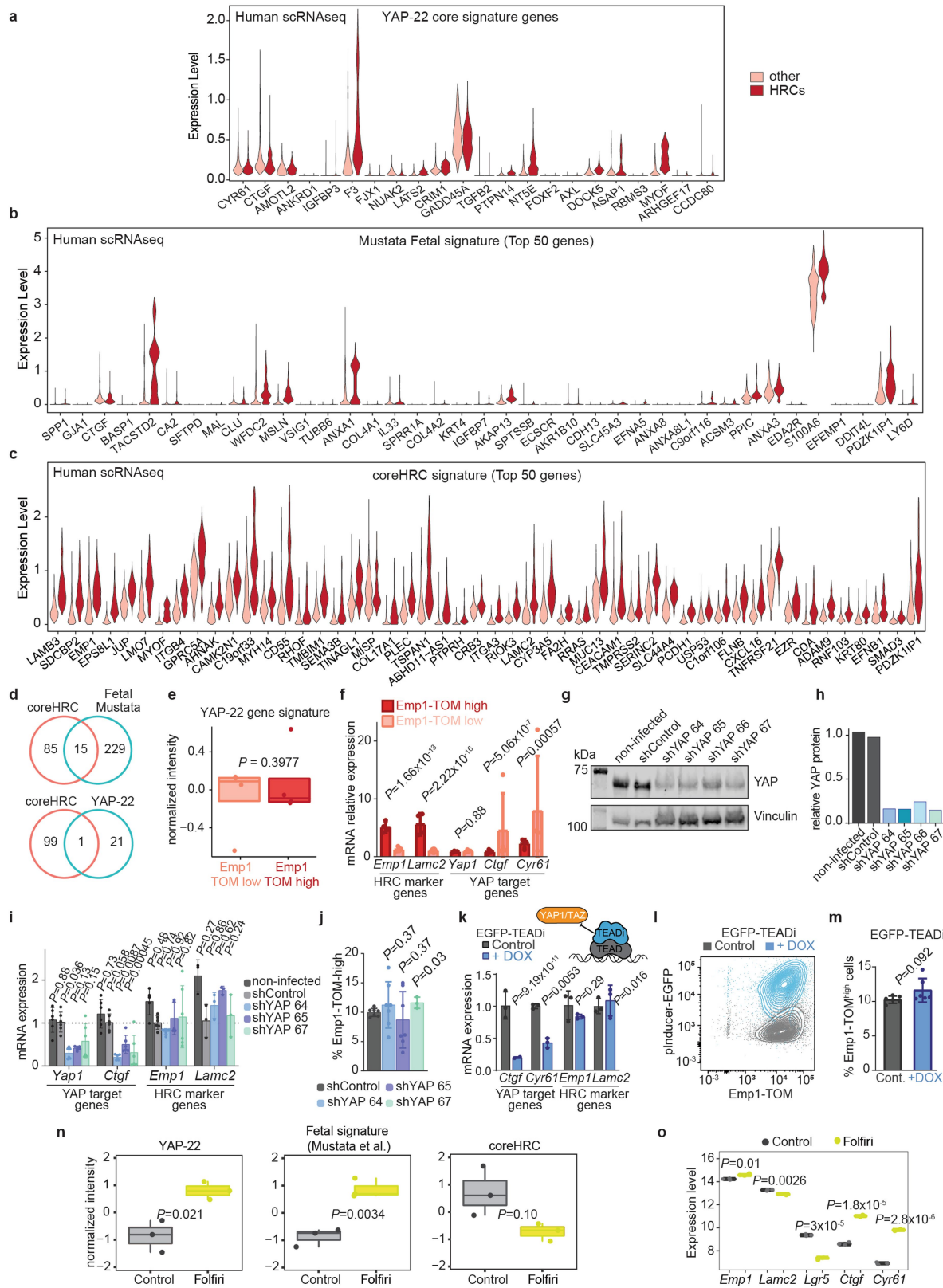


Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Emp1 and Lgr5 mark distinct tumour cell populations.

a, Emp1-iCasp9-tdTomato and Lgr5-EGFP alleles introduced in AKTP MTOs. Confocal imaging of TOM, EGFP and E-CADHERIN immunostaining in edited MTOs. Single z-plane. Scale bar, 10 μ m. **b**, Relative mRNA expression of indicated marker genes in EGFP-high/TOM-low and EGFP-low/TOM-high sorted cells dissociated from subcutaneous AKTP Emp1-iCT Lgr5-EGFP tumours. Two-sided t-test after normalizing by *PPIA*. Mean $\pm$ SD. n = 3 technical replicates. **c**, Immunostaining of TOM, EGFP and E-CADHERIN in Emp1-iCT Lgr5-EGFP primary tumours 4 weeks post-implantation of MTOs in the caecum. Dashed lines encompass tumour buds. Scale bar, 250 μ m. **d**, Scatter plot showing normalized Emp1-TOM intensity versus normalized Lgr5-EGFP intensity in 855,330 cells from 18 different primary tumours. Note the absence of double

positive cells (TOM and EGFP high). **e**, Representative immunofluorescence staining of TOM, EGFP and E-CADHERIN in liver metastases of increasing size (micro, small, medium) generated from the mouse CRC relapse model. Scale bars, 25 μ m (micro) 100 μ m (small) 250 μ m (medium). **f**, Scatter plot showing TOM intensity versus EGFP intensity in 318,276 cells from 137 different liver metastases. Note the absence of double positive cells (TOM and EGFP high). **g-p**, Examples of dual *EMPI* and *LGR5* mRNA ISH combined with E-CADHERIN immunofluorescence on human primary CRC tissue sections demonstrating a mutually exclusive pattern of expression of *EMPI* and *LGR5*. Note that *EMPI* expression is elevated in invasion fronts and tumour cell buds (white arrows). Scale bars, 500 μ m (l, p) 250 μ m (g, h, i, j, m, n, o) 50 μ m (H', k).

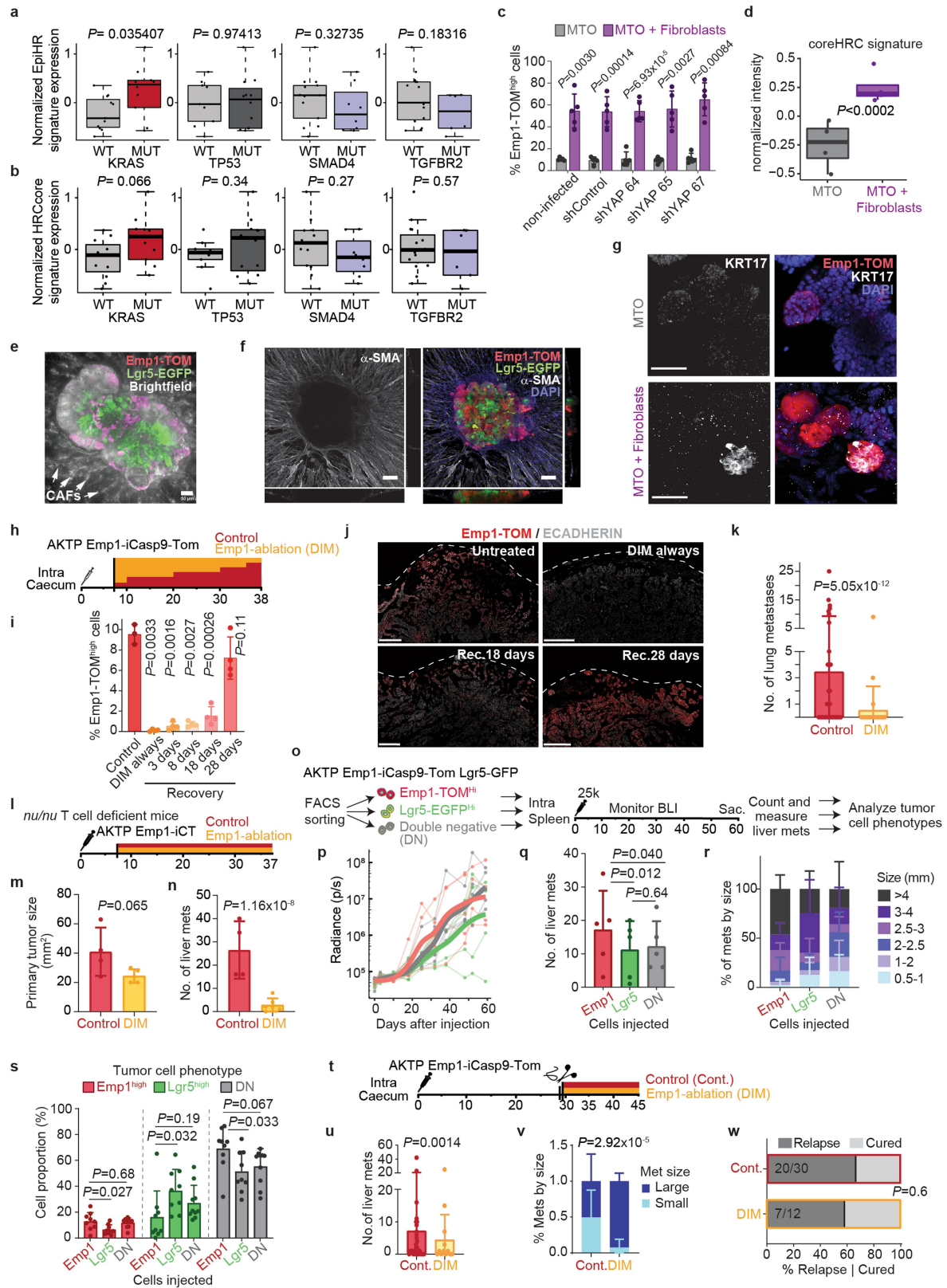


Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | YAP/TAZ signalling is not required for HRC

specification. a-c, Violin plots comparing the expression of genes belonging to the YAP-22 core signature³² (**a**), the top 50 genes from the Fetal intestine progenitor signature³¹ (**b**) and the top 50 genes of the coreHRC signature (**c**) in HRCs vs other cells in the SMC scRNA-seq cohort. **d**, Venn Diagram showing genes that overlap between the coreHRC signature and YAP-22 or Fetal intestine progenitor signatures. **e**, Boxplot showing normalized intensity of YAP-22 signature expression in Emp1-TOM^{high} and Emp1-TOM^{low} cells dissociated from primary tumours 4 weeks post-implantation. Box plots have whiskers of maximum 1.5 times the interquartile range; boxes represent first, second (median) and third quartiles. $n = 4$ mice per condition. ROAST-GSA adjusted p-values are shown. **f**, Relative mRNA expression measured by RT-qPCR in Emp1-TOM^{high} and Emp1-TOM^{low} sorted cell populations from AKTP Emp1-iCT primary CRCs. Two-sided t-test after normalizing by *PPIA*. $n = 4$ biological replicates. Mean $\pm$ SD. **g**, Western blot for YAP and VINCULIN in non-infected, shControl and shYap infected AKTP organoids. **h**, Western blot quantification of YAP normalized by Vinculin. **i**, Relative mRNA expression (mean $\pm$ SD) in MTOs infected with shControl plasmid compared to uninfected MTOs and MTOs infected with three different shYAP plasmids. Analysed with a mixed effects linear model after normalizing by *PPIA* housekeeping gene. $n = 2$ biological replicates with 3 technical replicates. **j**, Percentage (mean $\pm$ SD)

of Emp1-TOM^{high} cells in organoids infected with shControl or shYAP plasmids. Two-sided Wilcoxon rank-sum test. $n = 3$ (sh67) 7 (all other) measurements examined over 4 independent experiments. **k**, Relative mRNA expression (mean $\pm$ SD) in MTOs infected with pInducer GFP-TEADi plasmid treated or untreated with doxycycline (DOX). GFP+ cells were sorted in DOX treated organoids, whereas alive cells were sorted in untreated MTOs. $n = 3$ technical replicates. Analysed with a mixed effects linear model after normalizing by *PPIA* housekeeping gene. **l**, Representative flow cytometry plot showing Emp1-TOM fluorescence versus pInducer GFP-TEADi fluorescence in TEADi MTOs untreated or treated with DOX. **m**, Quantification (mean $\pm$ SD) of Emp1-TOM^{high} in TEADi MTOs untreated or treated with DOX for 5 days. Two-sided Wilcoxon rank-sum test. $n = 2$ biological replicates with 3 technical replicates. **n**, Boxplot showing expression levels (normalized intensity) of YAP-22, Fetal and coreHRC signature genes in control MTOs versus MTOs treated with chemotherapy (folfiri) for 4 days. Boxes represent the first, second (median) and third quartiles. Whiskers indicate maximum and minimum values. $n = 3$ biological replicates per condition. Two-sided t-test. **o**, Boxplot showing the expression levels of relevant genes in control MTOs versus MTOs treated with chemotherapy (folfiri) for 4 days. Boxes represent the first, second (median) and third quartiles. Whiskers indicate maximum and minimum values. $n = 3$ biological replicates per condition. Two-sided t-test.

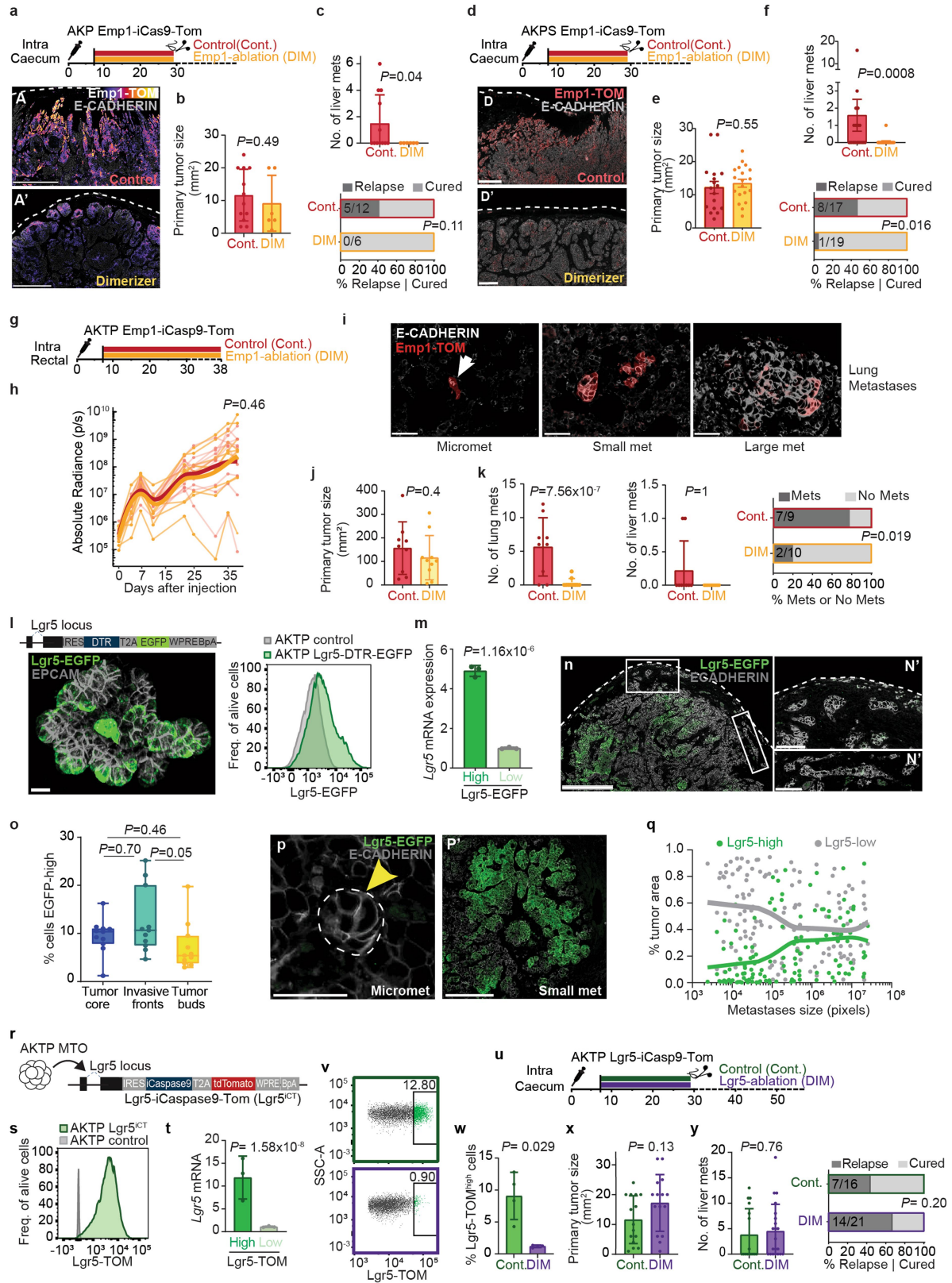


Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | KRAS mutations and CAFs specify the HRC population responsible for metastatic relapse. a-b,

Normalized intensity of EpiHR and coreHRC signature expression in CTOs grouped by gain of function mutation in *Kras* g12d and loss of function mutations in *p53*, *Smad4* and *Tgfb β 2*. Box plots have whiskers of maximum and minimum values; boxes represent first, second (median) and third quartiles. n = 6 (WT) 5 (MUT) CTOs; 2 technical replicates. P-values for two-sided T-tests. **c**, Percentage of Emp1^{high} tumour cells (defined as the top 10% of the TOM population in control MTOs, mean $\pm$ SD) in parental (non-infected), control shRNA or shRNAs targeting YAP1. n = 5 biological replicates. P-value for two-sided t-test. **d**, Normalized intensity of the coreHRC signature expression in control MTOs versus MTOs co-cultured with colon fibroblasts. Box plots have whiskers of maximum 1.5 times the interquartile range; boxes represent first, second (median) and third quartiles. n = 4 biological replicates. ROAST-GSA adjusted p-value is shown. **e**, Representative images of MTOs Emp1-iCT Lgr5-EGFP co-cultured with colon fibroblasts. Maximum intensity projection of confocal stacks, step 4 μ m, z stack 120 μ m. Scale bar, 50 μ m. **f**, Immunostaining of α -SMA Emp1-iCT Lgr5-EGFP MTOs co-cultured with colon fibroblasts for 2 days. Scale bars, 100 μ m. **g**, Immunostaining of KRT17 in 4-days grown MTO Emp1-iCT organoids: fibroblast co-cultures and organoids alone control cultures. Scale bars, 50 μ m. **h**, Ablation by dimerizer (DIM) treatment and surgery schedule of mice with AKTP Emp1-iCT primary tumours to assess the recovery of HRCs upon treatment cessation. **i**, Percentage (mean $\pm$ SD) of Emp1-TOM^{high} cells (defined as top 10% in control animals) in untreated mice versus mice treated with DIM, with treatment discontinued at various timepoints post-injection. Two-sided T-test. n = 3 (control) 4 (rest) mice. **j**, Representative immunostainings showing effective Emp1-TOM^{high} cell ablation in DIM-treated primary tumours and recovery upon treatment cessation. Dashed lines delimitate the caecum edge. Scale bars, 250 μ m. **k**, Lung metastases (mean $\pm$ SD) generated by MTO Emp1-iCT up to one month after primary tumour resection, treated with vehicle or DIM as in Fig. 4a. Each

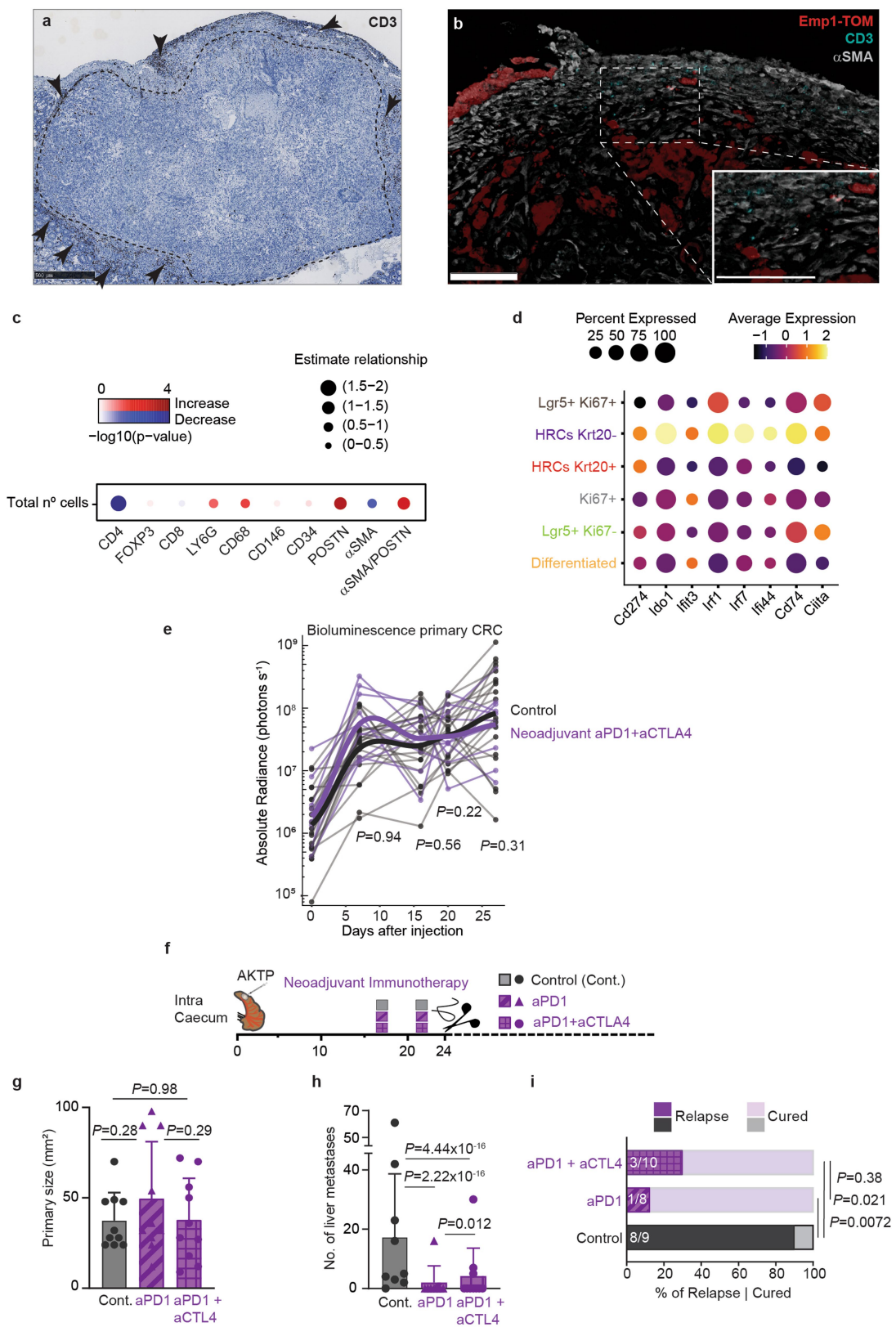
dot is a mouse; n = 34 (control) 29 (DIM). P-value for generalized linear model with negative binomial family. **l**, Inducible ablation schedule of nude mice (*nu/nu*) implanted with AKTP Emp1-iCT primary tumours. Resection was not possible due to local spreading of tumours to neighbouring tissue. **m**, Primary tumour area (mean $\pm$ SD) measured at sacrifice. Each dot is a mouse; n = 4 (control), 5 (DIM) mice. P-value for linear model. **n**, Liver metastases (mean $\pm$ SD) generated by MTO Emp1-iCT. Each dot is a mouse; n = 4 (control), 6 (DIM) mice. P-value for generalized linear model. **o**, Schematics of an experiment to analyse the potential of Emp1+, Lgr5+ or double negative (DN) cells to colonize the liver and generate metastases. 25,000 FACS-sorted Emp1-TOM-high, Lgr5-EGFP-high or double negative cells were injected intrasplenically. **p**, Metastatic growth measured by BLI. **q**, Liver metastases (mean $\pm$ SD) generated by Emp1-high, Lgr5-high or double negative cells. Each dot is a mouse; n = 5 mice. P-value for generalized linear model. **r**, Distribution of liver metastasis diameters (mean $\pm$ SD). n = 5 mice per condition. **s**, Percentage (mean $\pm$ SD) of Emp1-TOM^{high}, Lgr5-EGFP^{high} and double negative tumour cells in metastases generated by the injection of Emp1-TOM^{high}, Lgr5-EGFP^{high} and double negative cells, n = 9 (Emp1 and Lgr5) and 10 (DN) mice. Two-sided t-test. **t**, Experimental setup showing inducible HRC ablation after surgery of primary AKTP CRCs. **u**, Liver metastases (mean $\pm$ SD) generated by MTO Emp1-iCT in mice treated with vehicle or DIM 1 day after primary tumour resection and until experimental endpoints. Each dot is a mouse; n = 30 (control) 12 (DIM) mice P-value for generalized linear model. **v**, Percentage (mean $\pm$ SD) of small (diameter equal or smaller than 1 mm²) and big metastases (bigger than 1 mm²) in mice treated with vehicle or DIM 1 day after primary tumour resection and until experimental endpoints. Mixed effects linear model after boxcox transformation with mouse as random effect, n = 20 (control) and 7 (DIM) mice. **w**, Percentage of mice that developed liver metastases in control and Emp1-ablated mice. Analysed with a generalized linear model.



Extended Data Fig. 11 | See next page for caption.

Extended Data Fig. 11 | Metastatic relapse in different mouse CRC models arises from HRCs. **a**, Inducible ablation and surgery schedule of mice with AKP Emp1-iCT primary tumours. Panels A and A' show immunostaining of TOM and E-CADHERIN demonstrating effective ablation of Emp1-high cells in primary CRCs. Dashed lines delimitate the caecum edge. Scale bars, 500 μ m. **b**, Primary tumour area (mean $\pm$ SD) measured after resection. Each dot is a mouse, n = 12 (control) and 6 (DIM) mice. P-value for linear model after boxcox transformation. **c**, Liver metastases (mean $\pm$ SD) generated by MTO AKP Emp1-iCT up to one month after primary tumour resection. Each dot is a mouse, n = 12 (control) and 6 (DIM) mice. P-value for generalized linear model with negative binomial family. Bottom panel indicates the percentage of mice that developed liver metastases in the same experiment. Analysed with a two-sided fisher test. **d**, Inducible ablation and surgery schedule of mice with AKPS Emp1-iCT primary tumours. Panels D and D' show immunostaining of TOM demonstrating effective ablation of Emp1-TOM^{high} cells in DIM-treated primary tumours. Dashed lines delimitate the caecum edge. Scale bars, 250 μ m. **e**, Primary tumour area (mean $\pm$ SD) measured after resection. Each dot is a mouse, n = 17 (control) and 19 (DIM) mice. P-value for linear model after boxcox transformation. **f**, Liver metastases (mean $\pm$ SD) generated by MTO AKPS Emp1-iCT up to one month after primary tumour resection. Each dot is a mouse, n as in panel **e**. P-value for generalized linear model with negative binomial family. Bottom panel indicate the percentage of mice that developed liver metastases in the same experiment. Analysed with a two-sided fisher test. **g**, Inducible ablation schedule of mice implanted with AKTP Emp1-iCT MTOs in the rectum. **h**, Longitudinal intravital BLI quantification of AKTP MTOs implanted in the rectum. **i**, Representative TOM and E-CADHERIN immunostainings of lungs from mice bearing AKTP rectal tumours. Lung metastases of increasing size are shown. Note that TOM expression is higher in micrometastases and progressively reduced. Scale bars, 50 μ m. **j**, Primary rectal tumour area (mean $\pm$ SD) measured at sacrifice. Each dot is a mouse, n = 9 (control) and 10 (DIM) mice. P-value for linear model after boxcox transformation. **k**, Lung (left panel) and liver (middle panel) metastases (mean $\pm$ SD) generated by MTO Emp1-iCT injected in the rectum. Each dot is a mouse, n as in panel **j**. P-value for generalized linear model with Poisson family. Right panel shows the percentage of mice that developed metastases in the same experiment. Analysed with a two-sided fisher test. **l**, CRISPR-Cas9 targeting strategy to

introduce an DTR-GFP cassette into the Lgr5 locus of MTOs. Confocal imaging of immunostaining for EGFP and EPCAM in Lgr5-DTR-EGFP organoids. Scale bar, 30 μ m. Right panel shows a representative flow cytometry plot of EGFP expression in wild-type and Lgr5-EGFP organoids. **m**, Relative *Lgr5* mRNA expression (mean $\pm$ SD) of Lgr5-EGFP^{high} versus -low cells isolated from Lgr5-DTR-EGFP subcutaneous tumours. n = 3 biological replicates. Two-sided t-test normalizing to *B2M*. **n**, Immunofluorescence showing EGFP and E-CADHERIN in primary tumours. Insets (N' and N'') correspond to invasion fronts and tumour buds lacking EGFP expression at higher magnification. Scale bars, 500 μ m (D) and 100 μ m (D' and D''). **o**, Quantification of Lgr5-EGFP^{high} cells (defined as cells in percentile 90 for EGFP expression) in the tumour core, invasion fronts and tumour buds. Boxes represent the first, second (median) and third quartiles. Whiskers indicate maximum and minimum values. Paired two-sided Wilcoxon test on percentages. n = 11 mice. **p**, Representative images of Lgr5-EGFP staining in micro (P) and small (P') metastases. Dashed lines and the yellow arrow surround a micrometastasis. Scale bars: (F) 50 μ m; (F') 250 μ m. **q**, Percentage of tumour area containing Lgr5-EGFP^{high} and Lgr5-EGFP^{low} cells versus metastases size. Each dot represents an individual metastasis. **r**, CRISPR-Cas9 targeting strategy to introduce an iCaspase-9-TOM cassette into the LGR5 locus of AKTP MTOs. **s**, Representative flow cytometry plot of TOM expression in Lgr5-iCasp9-tdTomato organoids. **t**, Quantification of *Lgr5* mRNA (mean $\pm$ SD) by RT-qPCR in Lgr5-TOM^{high} and Lgr5-TOM^{low} cells dissociated from primary tumours grown for 4 weeks. n = 3 primary tumours. Analysed with a mixed effects linear model. **u**, Timing of inducible ablation and surgery in mice implanted with AKTP Lgr5-iCasp9-TOM primary tumours. **v**, Representative flow cytometry plot of Lgr5-TOM fluorescence in controls versus dimerizer-treated mice. DAPI-/EPCAM+ cells are shown. **w**, Percentage (mean $\pm$ SD) of Lgr5^{high} tumour cells (defined as the top 10% of the TOM+ population) in control and treated mice. n = 4 mice each group. Two-sided Wilcoxon test. **x**, Primary tumour area measured after resection. n = 15 mice each group. Mean with SD, p-value of linear model after boxcox transformation. **y**, Liver metastases counted at experimental endpoints after primary tumour resection. n = 16 (control) and 21 (Lgr5-ablation) mice. Mean $\pm$ SD. Analysed with a linear model with negative binomial family. Left panel show the percentage of mice that developed liver metastases in control and Lgr5-ablated tumours in the same experiment. Two-sided Fisher test.



Extended Data Fig. 12 | See next page for caption.

Extended Data Fig. 12 | Immune checkpoint immunotherapies prevent metastatic relapse. **a**, Representative image of CD3+ cell distribution in primary AKTP CRC showing T cell exclusion. Arrows point to T cells located at the tumour periphery. **b**, Representative immunostaining of Emp1-TOM, CD3 and α -SMA in primary tumours. Scale bars, 100 μ m. **c**, Dotplot summarizing regression models applied to multiplex immunofluorescence data. Effects of the total number of cells on the composition of every cell population are represented by different point sizes (defining the magnitude of the effect) and colours (showing both the sign of the effect in blue(-)/red(+) and the statistical significance by color intensity). **d**, Dotplot showing examples of interferon response genes across 6 tumour scRNA-seq cell clusters as defined in Fig. 2g. **e**, Bioluminescence monitoring of the effect of the neoadjuvant immunotherapy regime used in Fig. 5k on primary tumour growth. Points and lines represent individual mice, trend lines (bold) show a LOESS model. n = 19 (control) 10

(PD1+CTLA4) mice. Mixed effects linear model with data normalized to time 0 and mouse as random effect. **f**, Schematics of an experiment comparing metastatic relapse in untreated mice and mice treated with neoadjuvant treatment with anti-PD1 monotherapy or anti-PD1+/anti-CTLA4 combined therapy. **g**, Primary tumour area (mean $\pm$ SD) measured after resection in the experiment described in **f**. Each dot is an individual mice. n = 10 mice each group. Linear model. **h**, Liver metastases (mean $\pm$ SD) generated by AKTP primary tumours up to one month after primary tumour resection in the experiment described in **f**. Each dot is an individual mice. n = 9 (control), 8 (PD1), 10 (PD1/CTLA4). Generalized linear model of Poisson family. **i**, Percentage of mice that developed liver metastases or remained metastases-free at experimental endpoints (4 weeks after resection) in control and immunotherapy-treated tumours. n = 9 (control), 8 (PD1), 10 (PD1/CTLA4). Generalized linear model with beta-binomial distribution.

Corresponding author(s): Eduard Batlle

Last updated by author(s): 13/09/22

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- | | | |
|-------------------------------------|-------------------------------------|--|
| ☐ | ☒ | The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ | The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| ☐ | ☒ | A description of all covariates tested |
| ☐ | ☒ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ | For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| ☒ | ☐ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ | Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	Organoid and Histological images acquisition: Nikon NIS-Elements (Nikon A1 confocal), NanoZoomer-2.0 HT C9600 scanner, Leica Inverted SP5 confocal microscope, Custom lightsheet microscope (MacroSPIM), Flash4.0 v2 camera (Hamamatsu, Japan), GeneChip Scanner GCS3000 (Affymetrix/ThermoFisher Scientific) Flow Cytometry and cell sorting: BD FACS DIVA V8.0.1 Real Time PCRs: StepOne software v2.3. , QuantStudio™ Real-Time PCR Software Intravital bioluminescence imaging (BLI): Living image x4.5.2
Data analysis	We used R (v4.0.5), RStudio (v1.1.383) and FIJI/ImageJ (v1.53d), and have custom codes that are available upon request. In addition, we used QuPath (v0.1.2), GraphPad Prism (v7.03), FlowJo (v10.7), BD FACS DIVA, Sambamba (v0.7.1), Living image x4.5.2, StepOne software v2.3, QuantStudio™ Real-Time PCR Software, HALO image analysis platform

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Expression arrays and RNA-seq data are available at Gene Expression Omnibus (GEO). The accession number for gene expression sequencing experiments reported in this paper are GEO: GSE190055 (Arrays Emp1-high vs. Emp1-low AKTP tumor cells), GSE208139 (Arrays MTOs co-cultured with fibroblasts), GSE207974 (RNAseq chemotherapy) and GSE207668 (RNAseq CTOs). Count matrices for single cell RNAseq experiments were deposited in ArrayExpress under accession number E-MTAB-11284 (10X AKTP primary tumors) E-MTAB-11302 (Smart-seq metastatic progression) and E-MTAB-11981 (Smart-seq AKP micromets). Additional metadata and processed data files, including UMAP embeddings and gene signature scores, are available at Synapse (syn35000645).

Previously published single cell RNA sequencing data of CRC patient samples that were reanalyzed here are available at GEO under accession codes GSE132465, GSE132257, GSE144735 and at Array Express under code E-MTAB-8107. Previously published bulk RNA sequencing and microarray expression data of CRC patient samples reanalyzed here are available at GEO under accession codes GSE14333, GSE39582, and at TCGA Research Network (<http://cancergenome.nih.gov/>) under code TCGA-COAD.

Materials generated in this paper are available to the community upon reasonable request, subject to restrictions imposed by MTAs. All other materials are readily available from indicated standard commercial sources, or within the paper and its supplementary information files.

The following figures have associated source data: Fig1 h,l,i; Fig2 b,c,h; Fig3 c,d,l,k,l,m,n,p,q,r; Fig4 e,f,g,h,i,j,k,l,p,q,r,s,t,u; Fig5 d,g,h,j,l,m,n,o,p,r,s; ExtFig1 d; ExtFig2 u; ExtFig3 i,m; ExtFig4 d,h,m; ExtFig5 k,l; ExtFig6 d,g,j; ExtFig7 b; ExtFig8 b; ExtFig9 h,i,j,k,m,n,o; ExtFig10 a,b,c,d,l,k,m,n,p,q,r,s,u,v,w; ExtFig11 q,t,w,x,y; ExtFig12 c,e,g,h,i.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical test was used to determined sample size upfront. Instead sample size was determined empirically according to previous knowledge of the variation in the experimental setup. We used at least 4 mice per group for the in vivo settings, which is sufficient to detect meaningful differences based on historical record with similar experiments. For the majority of in vitro experiments, we used n=3 according to previous experience with similar experiments. The sample size typically results in standard error <25% of the mean.
Data exclusions	No data from in vitro or in vivo experiments were excluded, except for the CRC relapse model, where a small fraction of mice, typically 1 in every 10 in each experiment, were not included in the follow up due to invasion into the ileocaecal junction, which impeded successful surgery. Occasionally, mice bearing CRCs were sacrificed 1-2 days after surgery due to significant weight loss or unhealthy aspect.
Replication	Images of whole livers with micro and macro metastases (Fig.2d) are representative of two independent experiments. Micrographs in Fig.3b,e,o,s; Fig. 5b; Ext.Fig.6e,h,k,l; Ext.Fig.7a; Ext.Fig.8a; Ext.Fig.10e,f,g,j; Ext.Fig.11l,n and Ext.Fig.12a,b are representative examples of results observed consistently across n>3 organoids or mice in at least 2 independent experiments. Micrographs in Fig.4c,n and Ext.Fig.10j are representative examples of results observed consistently across n>3 mice in at least 2 independent experiments. In these experiments additional validation of the results is provided by flow cytometry analyses (Fig.4d,o; Ext.Fig.10i). Micrographs in Fig.3f,g,h,j; Fig.5a,c,e,f; Ext.Fig.6a,b,c,i; Ext.Fig.8c,e; Ext.Fig.11p are representative examples of immunofluorescences which have been quantified (Fig.3i,k; Fig. 4i, k; Ext.Fig.6d,j; Ext.Fig.8d,f; Ext.Fig.11q). Micrographs in Ext.Fig.11a,d,i are representative examples of results observed in at least 5 mice analyzed in one experiment. The western blot in Extended Figure 9g was repeated in two independent experiments with similar results. ISH analyses of human CRC (Ext.Fig.7e,f and Ext.Fig.7g-p) were only performed once and images show representative staining patterns present in each section. All the rest of results were replicated at least in two independent experiments.
Randomization	For in vivo experiments involving treatments (DIM or immunotherapy), animals were distributed in treatment groups containing mice with similar tumor burden according to BLI measurements. For the comparison of metastasis formation capacity between different cell phenotypes (Extended data Figure 11o), the different tumor cell population were injected in an intercalated order to avoid biases as a result of cell viability. For in vitro experiments, all samples were analyzed equally with no sub-sampling; therefore, there was no requirement for randomization.
Blinding	Treatments (DIM or immunotherapy) were assigned after inoculation of the MTOs in the primary site. For primary tumor surgery we were blinded to the treatment arm and the scientist measuring primary tumors and counting the number of metastasis was blinded to the treatment arms. Quantification of multiplex IHC in Figure 5 was performed in a blinded manner. In the rest of in vivo and in vitro experiments, data collection and analyses were not performed blinded to the conditions of the experiments.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Only commercial antibodies have been used:
For IF/IHC, the following list was used:

- Primary antibodies: Rabbit anti-TdTomato (Rockland, 600-401-379), Goat anti-TdTomato (Sicgen, Ref.: AB8181-200), Mouse anti-E-cadherine (BD, Ref.: 610182), Goat anti-GFP (Abcam, ab773), Rabbit anti-CD31 (Abcam, ab28364), Rabbit anti-KRT20 (Sigma HPA024309), Rabbit anti-KRT17 (Sigma, HPA000453), Mouse anti-aSMA (Abcam, ab7817), Rabbit anti-CD3 (Abcam, ab16669), Goat anti-Epcam (R&D, AF960), Mouse anti-LAMC2 (Sigma, AMAB91098). Envision FLEX antibody diluent (Agilent, K8006) was used as diluent

- Secondary antibodies: donkey anti-goat conjugated to Alexa 488/568/647 (Life Technologies A11055, A11057, A21447), donkey anti-rabbit conjugated to Alexa 488/568/647 (Life Technologies A21206, A10042, A31573) and donkey anti-mouse conjugated to Alexa 488/568/647 (Life Technologies A-21202, A10037, A31571).

For immunotherapy experiments, 250 µg of anti-mouse CTLA-4 (C2444, Leinco) and anti-mouse PD-1 (P372, Leinco) diluted at 1 mg/ml were administered via intraperitoneal injection. Validation reported on supplier's website.

Flow Cytometry antibodies: and APC anti-mouse CD326 (Ep-CAM) (Biolegend, ref 118217), BV605 anti-mouse CD45 (Biolegend, 103155) and APC anti-mouse PD-L1 (BD, 564715) were used at concentration of 1:200.

Validation

IF/IHQ and WB and flow cytometry Antibodies were validated extensively by the provider (including westernblot analysis), used in multiple studies previously and confirmed by the authors for specific localization.
Extra validation was performed by comparing expression in internal controls (known negative populations).

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)

Mouse tumor organoids (MTOs) were derived from genetic models in our lab. They are described in Tauriello et al. Nature 2018.
HEK293T cells were purchased from [ATCC Number: CRL-3216]

Authentication

MTOs were genotyped by PCR at least once a year using specific primers for key mutations.

Mycoplasma contamination

Mycoplasma contamination was tested for (bi-monthly) and was always negative.

Commonly misidentified lines (See [ICLAC](#) register)

No commonly misidentified cell lines according to the ICLAC database were used in the study.

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals

Injection and transplantation of MTOs was performed in c57BL/6J hosts aged 7-9 weeks, sex matched with the organoid to be injected (males).
All mice were closely monitored by authors, facility technicians (during treatments and experiments) and by an external veterinary scientist responsible for animal welfare. Mice were maintained in a specific-pathogen-free (SPF) facility with a 12-h light-dark cycle, under controlled temperature and humidity (18-23°C and 40-60% respectively) and given ad libitum access to standard diet and water.

Wild animals	No wild animals were used in the study.
Field-collected samples	No field collected samples were used in the study.
Ethics oversight	All experiments with mouse models were approved by the Animal Care and Use Committee of Barcelona Science Park (CEEA-PCB) and the Catalan government.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	Tissue sections for ISH and IHC (Extended Data Figures 7e,f and 8g-p) were of stage II-IV primary colon adenocarcinomas obtained from patients that were diagnosed and underwent surgery at Hospital Clinic de Barcelona (n=10). Patients were both male and female, with an age range of 46-74 years old. None of the patients had received neoadjuvant treatment before collection of the samples.
Recruitment	Samples used for ISH and IHC in the study (Extended Data Figure 7e,f and Extended Data Figure 8g-p) correspond to archived FFPE material collected for the purpose of diagnosis. Patients donated their tissues for research purposes under informed consent.
Ethics oversight	Tissue sections from patients used for ISH and immunofluorescence were provided by the Hospital Clinic de Barcelona-IDIBAPS Biobank (B.0000575), integrated in the Spanish National Biobanks Network and they were processed following standard operating procedures with the appropriate approval of the Ethics and Scientific Committees of the Hospital Clinic de Barcelona (registration code:HCB/2020/1478). Human subjects involvement in this study was restricted to biospecimen collection for diagnosis and no interventions were performed on them.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

- ☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☒ All plots are contour plots with outliers or pseudocolor plots.
- ☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Organoids were trypsinized for 5-10 minutes at 37 degrees with pipetting and shaking until they disaggregated into single cells. After washings, they were resuspended in FACS buffer (HBSS 0.5% FBS) containing DAPI $\mu\text{g/ml}$. Primary tumours and micro-dissected liver metastases were chopped with razor blades. Subsequent enzymatic digestion was performed with 200 U/ml collagenase IV (Sigma Aldrich) in HBSS (Lenovo) for 30-40 min at 37 °C, in a shaking water bath or a gentleMACS Dissociator (Miltenyi Biotec). Digested tissue fragments were then filtered through 100- and 40-μm meshes, washed and treated for 5 minutes with ammonium chloride. Single-cell preparations were first blocked with anti-CD16/32 (clone 93; eBioscience) and then stained with APC anti-Epcam (118214; Biolegend) and BV605 anti-CD45 (103155, Biolegend) antibodies for 20 minutes on ice. Finally, cells were resuspended in HBSS with 0.5% FBS and DAPI $\mu\text{g/ml}$ (Sigma Aldrich). For isolation of low numbers of tumour cells from mice without visible metastases, whole livers were thoroughly minced with razor blades. After an initial 30-minute digestion with 200 U/ml collagenase IV, samples were filtered through 100 μm meshes. Although most cells flowed through, a small fraction of the samples –highly enriched for tumour cells– was retained in the filter Extended Figure 5c). Filters were then laid into a 6-well plate and covered with HBSS containing collagenase IV, Dispase II and DNase I. After re-digestion in a water bath for 30 minutes at 37 °C, most cells now seeped through the filters. The protocol continued with washing, ammonium chloride treatment and antibody stainings.
Instrument	Flow cytometry analysis and cell separation were performed in a FACSAriaFusion flow cytometer (Beckton Dickinson).
Software	For cell sorting, BD FACS diva software was used. Data were analyzed using FlowJo software v10.7.
Cell population abundance	The enrichment of TOM+ or GFP+ cells in the sorted samples was determined by expression analysis of marker genes as shown in ExtFig 5k, 6g, 8b, 9f, 11m, 11t.
Gating strategy	Flow cytometry analysis on organoids (ExtFig5j, ExtFig6f, ExtFig11l, s): To analyze Emp1 and Lgr5 population in mouse organoids we gated Cells/Single cells/Living cells.

Flow cytometry on in vivo primary tumors (Figure 4d-e, F4o-p, and ExtFig11v-w: We gated Cells/Single cells/Living cells/Epcam+CD45- cells/Tomato+ cells. We defined the tomato-high population as the 10% brightest cells.
Flow cytometry for 10X and smart-seq single cell profiling experiments: We gated DAPI-CD45-GFP+ cells.
Flow cytometry in livers with residual DTCs (ExtFig4m): We gated GFP+ cells in all events.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.