# Phenotypic plasticity mediates colorectal cancer metastasis

**Mirjana Efremova** ( m.efremova@qmul.ac.uk )
Barts Cancer Institute https://orcid.org/0000-0002-8107-9974
**Samuel Ogden**
Barts Cancer Institute
**Nasrine Metic**
Barts Cancer Institute
**Elise Smith**
Barts Cancer Institute
**Alison Berner**
Queen Mary University of London https://orcid.org/0000-0002-1132-0275
**Ann-Marie Baker**
Institute of Cancer Research https://orcid.org/0000-0001-8905-9137
**Imran Uddin**
University College London
**Claude Chelala**
Barts Cancer Institute https://orcid.org/0000-0002-2488-0669
**Dayem Ullah**
Barts Cancer Institute
**Amina Saad**
Barts Cancer Institute
**Jo-Anne Chin Aleong**
Barts NHS Trust
**Trevor Graham**
Institute of Cancer Research https://orcid.org/0000-0001-9582-1597
**Hemant Kocher**
Barts Cancer Institute https://orcid.org/0000-0001-6771-1905

**Article**

**Keywords:**

**Posted Date:** February 1st, 2024

**DOI:** https://doi.org/10.21203/rs.3.rs-3846377/v1



**Additional Declarations:** There is **NO** Competing Interest.

# Phenotypic plasticity mediates colorectal cancer metastasis

Samuel Ogden[1#], Nasrine Metic[1#], Elise A Smith[1], Alison M Berner[1], Annie Baker[2], Imran Uddin[3,4], Cancer Tissue Bank[1], Trevor Graham[2], Hemant M Kocher[1], Mirjana Efremova[1*]

[1] Barts Cancer Institute, Queen Mary University of London, London, UK
[2] The Institute of Cancer Research, London, UK
[3] CRUK City of London Centre Single Cell Genomics Facility, UCL Cancer Institute, University College London, London, UK
[4] Genomics Translational Technology Platform, UCL Cancer Institute, University College London, London, UK.

# These authors contributed equally
*correspondence to m.efremova@qmul.ac.uk

## Abstract

A major challenge for inhibiting metastasis is the ability of cancer cells to reversibly switch states in response to microenvironmental cues along the metastatic cascade. The regulatory factors and signals from the microenvironment enabling colorectal cancer (CRC) cells to transition into an invasive state and to establish metastasis in the liver remain unknown. Using a combination of single-cell multiomics and spatial transcriptomics data from primary and metastatic CRC patients, we reveal putative metastasis-initiating cancer states with regenerative and inflammatory signatures, driven by transcription factors AP-1, NF-κB and YAP. We demonstrate the existence of an intermediate population with a hybrid regenerative and stem phenotype, indicating phenotypic transitions between stem and pro-metastatic cells. Our spatial analyses show localisation of the regenerative states at the invasive edge in primary CRC and in an immunosuppressive niche in liver metastasis, surrounded by immune and stromal cells that sustain these cells. We uncover putative ligand-receptor interactions driven by cancer-associated fibroblasts (CAFs), macrophages and CD8 T cells that activate the regenerative and inflammatory invasive phenotype in cancer cells. Together, our findings reveal regulatory and signalling factors that can be targeted to restrict transition into invasive states to impair metastasis.

# Introduction

Colorectal cancer is the third most common malignancy globally and the second leading cause of cancer-related death because of its high relapse rate[1]. Less than 1 in 5 people with metastatic CRC survive for 5 years following their diagnosis. Genomic divergence between primary and metastatic tumours from the same patients[2] is low and driver mutations tend to be clonal across metastases[3]. Furthermore, spatial mapping of mutations shows that many primary CRC (pCRC) clones are capable of invasion[4]. Collectively this points to a key role in CRC progression and therapy resistance for phenotypic plasticity - the ability of cells to undergo rapid phenotypic transitions in response to external signals and adapt to new microenvironments such as those encountered during metastasis[5,6,7]. Plasticity allows cancer cells to acquire and maintain phenotypic heterogeneity that enables them to become invasive, escape the immune system, colonise distant sites and evade therapy.

CRC has a cellular hierarchy resembling a healthy intestine[8], maintained by LGR5 expressing stem cells[9] that give rise to transient-amplifying progenitor cells which undergo differentiation into absorptive and secretory lineages. However, following tissue damage or loss of the stem cells, differentiated cells can dedifferentiate and replenish the impaired stem-cell niches to enable tissue repair[10,11]. The same phenomenon has been described in murine CRC models and organoids after ablation of the Lgr5+ stem cells. Furthermore, disseminating metastasis-initiating cells in mouse models were found to be predominantly Lgr5- cells that reacquire Lgr5+ stem-cell phenotype at metastatic sites to regenerate metastasis[12,13]. Chemotherapy was also shown to promote a transition from an LGR5+ stem-like state to an LGR5- drug-resistant state[14,15]. A recent study that for the first time characterised heterogenous cancer states in CRC metastasis, showed that liver metastases in some patients display progressive plasticity enabling differentiation into non-canonical squamous and neuroendocrine-like states[16]. Such reversible cell-state transitions indicate that cellular reprogramming is largely driven by epigenetic plasticity that can initiate new transcriptional programmes in response to external signals[17]. However, the regulatory factors and extrinsic signals enabling CRC cells to transition into an aggressive cell state, and the mechanisms that maintain this state, are poorly understood. This hinders efforts to improve prognostication, predict who will benefit from treatment, and develop new therapies.

Here, we sought to characterise the heterogeneous cancer cell states in primary and liver metastatic CRC using a combination of single-cell RNA-seq (scRNA-seq), single-nucleus multiomics and spatial transcriptomics data. We find putative metastasis-initiating cell states and identify gene regulatory networks and transcription factors driving those states. These pro-metastatic states are enriched at the tumour invasive front and surrounded by immunosuppressive immune and stromal cells. We reveal the spatial organisation in cellular niches of the metastatic liver and find ligand-receptor interactions potentially shaping and sustaining the invasive phenotype.

# Results

## Heterogeneous cancer cell states in primary CRC

To examine the cellular heterogeneity and composition of malignant cell states and the tumour microenvironment in pCRC, we integrated previously published scRNA-seq data[18–20] of 117 untreated CRC patients (Fig. 1a, Extended Data Fig. 1a). After quality control, we retained single-cell transcriptomes from 246,779 cells, including cancer, immune (T, natural killer - NK, B and myeloid cells) and stromal (cancer-associated fibroblasts (CAFs), endothelial and perivascular cells) cells (Fig. 1b, Extended Data Fig. 1b-c).

We first focused our analysis on the malignant cells, identified by inferring genome-wide single-cell somatic copy number variations (CNVs) with inferCNV. Integration of gene expression data was performed accounting for patient variability. Our analysis shows cell hierarchies reminiscent of those in normal tissues, with stem cells (*LGR5*) giving rise to TA cells (*MKI67*, *TOP2A*) which differentiate into distinct absorptive (Colonocytes - *SLC2A23*) and secretory lineages (Goblet - *MUC2*, Tuft - *LRMP* and Enteroendocrine - *PCSK1*) (Figure 1c-f, Extended Data Fig. 1d-i). In addition to the normal-like states, we identified a stem cell state absent in the healthy colon and characterised by upregulation of a range of WNT antagonists including *NOTUM*, *NKD1* and *APCDD1* (Extended Data Fig. 1d-g, Supplementary Table 1). Apc-mutant stem cells have been shown to secrete WNT inhibitors such as NOTUM to outcompete wild-type stem cells by driving their differentiation, thereby facilitating the outgrowth of Apc-mutant clones and development of premalignant adenomas[21]. In line with this, our analysis shows that the Stem NOTUM state is enriched in patients with mutations in the tumour suppressor *APC* (Extended Data Fig. 1j).

Among the cancer-specific states, gene enrichment analysis (GEA) reveals two subpopulations that share an expression profile with LGR5- regenerative stem cells (RSC) or revival colonic stem cells (revCSC), a plastic state identified in primary CRC[22,23], hereafter named regenerative cells (REC; marker genes: *LAMC2*, *EMP1*). In addition, we detect a hypoxic (*VEGFA*) state and an HLA high state (Fig. 1e,f, Extended Data Fig. 1g and Supplementary Table 2). Interestingly, a subset of RECs also upregulates interferon-alpha (IFN-α) and interferon-gamma (IFN-γ) target genes (Fig. 1e and Extended Data Fig. 1k), suggestive of an ongoing inflammatory response (hereafter named inflammatory regenerative cells - iREC). (i)RECs also upregulate epithelial-to-mesenchymal transition (EMT) signatures, but lack expression of mesenchymal markers and EMT transcription factors (Fig. 1e and Extended Data Fig. 1l) indicating that (i)REC are in a pEMT state, maintaining their epithelial identity. Furthermore, (i)REC are enriched for a CRIS-B signature, one of the five transcriptional subtypes of CRC inferred using patient-derived xenografts[24] that is associated with poor prognosis and EMT.

(i)RECs closely resemble a cancer state driving metastatic recurrence after surgical resection in a CRC mouse model[25], implicating them as metastasis-initiating CRC cells required for tumour regeneration. They also share an expression profile with a state enriched in early micrometastasis CRC mouse models[13] that upregulates a foetal intestinal signature and is characterised by high YAP signalling[21–24] (Fig. 1e). Consistent with this, we find that the (i)RECs are marked by activation of genes associated with foetal intestinal development, including *TACSTD2* (encoding cancer-associated trophoblast antigen TROP2) and *ANXA1*, as well as upregulation of YAP target genes (Fig. 1e and Extended Data Fig. 1g). The presence of an interferon-response signature in malignant cells suggests that IFN-γ signalling from T cells may activate this response in cancer cells that are in proximity of immune cells

and that disseminated cells may escape immune attack through the expression of immunomodulatory molecules[26]. In line with this, a similar interferon response module was recently revealed as a recurrent cancer state across 15 cancer types, spatially colocalising with T cells and macrophages[27]. Interestingly, we also find an intermediate state expressing both (i)REC and stem markers, potentially indicating a hybrid transition state between (i)REC and stem-like states. The Intermediate state also upregulates chemokines such as *CXCL2* and *CXCL3* (Fig. 1f and Supplementary Table 1).

Generally, different cancer cell states are present in all samples, but individual tumours display heterogeneity in the composition of cancer cell states (Fig. 1d and Extended Data Fig. 1h). Comparison between mismatch repair deficient (MMRd) and mismatch repair proficient (MMRp) tumours highlights their differences, with the HLA high state more abundant in MMRd tumours, whereas Stem, Stem NOTUM, Intermediate, and Tuft cell states more abundant in MMRp tumours (Extended Data Fig. 1m-q). Consistent with this, expression of cancer cell state signatures (Supplementary Table 3) in bulk tumours reveals higher expression of an HLA-high signature in MSI-H tumours (but not MSI-L), and lower expression of Stem NOTUM, Stem and Intermediate signatures in MSI-H tumours (Extended Data Fig. 1q). Moreover, MMRd tumours are enriched for intrinsic consensus molecular subtype 2 (iCMS2) signature, compared to MMRp tumours that have higher iCMS3[28] (Extended Data Fig. 1o). Therefore, genetically distinct tumour subtypes can affect the composition of cancer cell states.

## Metastasis-initiating cell states localise at the invasive tumour edge

Microenvironmental pressures within the primary tumour can drive cancer cells to adapt to different conditions and acquire pro-metastatic traits for invasion and colonisation of secondary organ niches. To better understand the influence of the tumour microenvironment (TME) on shaping distinct cancer cell states, we analysed the non-malignant cells by integrating 5 pCRC datasets[18–20,29]. We identified the majority of known immune cell types, including myeloid, mast, natural killer (NK), innate lymphoid (ILCs), T and B cells (Fig. 1b) and then resolved the heterogeneity of the TME cells by integrating and analysing each major cell type separately.

Within the stromal cell subpopulations, CAFs and endothelial cells were the major cell types (Extended Data Fig. 2a,b). Additionally, we find pericytes (*RGS5*, *ABCC9*, *PDGFRB*), vascular smooth muscle cells (*MYH11*, *ACTA2*, *TAGLN*) and enteric glial cells (*S100B*, *PLP1*). Among the CAFs, we find inflammatory CAFs termed C3+ iCAF (upregulating chemokines such as *CXCL12* and complement *C3*), ECM remodelling CAFs (upregulating *POSTN*, various collagen and matrix metalloproteinase genes) and contractile myofibroblasts (*ACTA2*, *TAGLN*) (Extended Data Fig. 2a). The inflammatory CAFs have been shown to interact with immune cells and orchestrate an immunosuppressive environment, whereas desmoplastic ECM CAFs remodel the ECM to facilitate cancer cell migration[30,31]. Furthermore, we observe fibroblasts that are present in the normal colon: bone morphogenetic protein (BMP)-producing CAFs (*CXCL14*) which produce BMPs to drive differentiation of epithelial cells[32] and GREM1+ CAFs that produce stem cell niche factors such as *RSPO3*[33] (Extended Data Fig. 2a). The endothelial cells are divided into four clusters, vascular stalk-like (*ACKR1*, *SELP*) and tip-like (*RGCC*, *KDR*) cells, lymphatic endothelial (*LYVE1*, *PROX1*) and proliferating endothelial cells (Extended Data Fig. 2b).

The myeloid compartment comprises tumour-associated macrophages (TAMs), monocytes, neutrophils and dendritic cells (DC) (Extended Data Fig. 2c). Within the TAMs, we identify immunosuppressive subpopulations, including SPP1+ TAMs marked by high expression of *MARCO*, *SPP1*, and *FN1* and upregulation of an angiogenic signature, and C1QC+ TAMs with high expression of complement C1Q genes (*C1QA*/*B*/*C*), *MS4A4A* and *TREM2* (Extended Data Fig. 2c,d). Both TAM subsets upregulate the expression signature of lipid-laden TAMs which are observed in multiple cancer types and have been shown to promote tumour progression[34] (Extended Data Fig. 2d). SPP1+ macrophages have been described as pro-metastatic and angiogenic TAMs potentially driven by hypoxic conditions in the tumour[35], whereas TAMs expressing *C1QC* and *TREM2* are known to induce T cell exhaustion and Treg infiltration[36,37]. We also find IL1B+ and NLRP3+ subpopulations, characterised by high expression of an inflammatory TAM signature (*IL1B*, *NLRP3*, *CCL3, CXCL3*)[37] (Extended Data Fig. 2d), which have previously been described as tissue resident because of their presence in normal tissue[18]. In kidney cancer, IL1B+ macrophages collocate with EMT-enriched tumours cells at the invasive edge[38]. In addition, we detect PLTP+LYVE1+ macrophages resembling perivascular macrophages which reside near blood vessels and play a role in restraining inflammation and tissue repair during fibrosis[39]. The monocyte subsets include FCN1+CD14+ monocytes and intermediate CD16+CD14+ monocytes (Extended Data Fig. 2c,e). In addition, we also identify conventional dendritic cells, cDC1 (*CLEC9A*, *XCR1*) and cDC2 (*CD1C*, *CLEC10A*), CCR7+LAMP3+ migratory DCs and LILRA4+ plasmacytoid DCs (Extended Data Fig. 2c). Migratory DCs are shown to be actively recruited during inflammation in the colon, secreting inflammatory cytokines, migrating to draining lymph nodes and mediating T cell activation[40].

T lymphocytes comprise diverse CD8+ and CD4+ T cells, spanning from naive to effector to exhausted states (Extended Data Fig. 2f). Specifically, the CD8+ cells are divided into effector memory and exhausted T cells, whereas in the CD4 compartment we observe naive, helper, follicular, Th17 and regulatory T cells. We also find T cells with a stress signature (*HSPA1A*, *HSPA1B*). Furthermore, we identify two subsets of NK cells, distinguished by expression of *XCL1*, *XCL2* and *GZMK* in NK1 and higher expression of granules (*PRF1*, *GZMB*, *GZMH*), *KIRD2L1*, *KIR3DL2* and *HAVCR2* in NK2. In addition, we also find NKT, ILC and gd T cells.

To comprehensively analyse the spatial organisation of CRC tumours and dissect the extrinsic signals promoting the invasive cell states, we used published spatial transcriptomics Visium data from 4 pCRC samples[41]. The tumour core and invasive edge annotations from the original publication were confirmed by manual assessment of the hematoxylin and eosin (H&E) staining (Fig. 2a,b). We performed differential expression analysis between the tumour core and the invasive edge to investigate whether different spatial structures have distinct transcriptional profiles (Supplementary Table 4). GEA reveals that while cell cycle and WNT signalling pathways are enriched among upregulated genes in the tumour core, the invasive edge is enriched with EMT, hypoxia, IFN-γ response, NFκB, angiogenesis and KRAS signalling (Fig. 2c, Extended Data Fig. 3a).

To estimate the abundance of malignant and TME cells in each spot, we spatially mapped the fine-grained cell types/states defined by scRNA-seq data onto their spatial location using cell2location[42]. To identify cellular niches across all samples, we used SpatialDE2[43] on the cell2location spot-by-cell output to partition each Visium slide into cellular neighbourhoods. Joint analysis of all the Visium samples enabled us to identify recurrent patterns across

samples. Remarkably, our findings reveal spatial localisation of the (i)RECs at the invasive tumour edge (Fig. 2d and Extended Data Fig. 3b-e), further supporting the hypothesis that these states are potentially metastasis-initiating cells. In comparison, the stem cells are abundant at the tumour core (Fig. 2d and Extended Data Fig. 3c-f). The hypoxic cells are present both at the core and the invasive edge, suggesting a potential transitional state. To further explore differences in the activation of cellular signalling pathways we used TCGA reverse phase protein array data from bulk CRC tumours, which indicated greater abundance of active MEK1 (MAP2K1), ERK1/2 (MAPK1/3), and p38 MAPK (MAPK14) in tumours with high expression of (i)REC signatures (Extended Data Fig. 3g). Collectively, this suggests that MAPK signalling is associated with (i)REC states and may drive cellular transitions into (i)REC states at invasive fronts.

(i)RECs at the invasive edge colocalise with myofibroblasts and ECM CAFs (Fig. 2e-g and Extended Data Fig. 3h). In addition, (i)RECs are surrounded by immunosuppressive cells such as SPP1+ macrophages, neutrophils, CD8 Tex and Tregs which do not infiltrate in the tumour core (Fig. 2e-g and Extended Data Fig. 3i,j), indicating that the activation of interferon response pathways in (i)REC may be due to interactions with immune cells. (i)RECs are also in close proximity to perivascular cells (Fig. 2f), suggesting a connection with haematogenous or lymphatic dissemination. Near the border between the invasive edge and healthy colon, we also find abundance of C3 immunomodulatory CAFs known to be involved in recruitment and polarisation of immunosuppressive myeloid cells[30] (Fig. 2g).

## Cancer cell states are re-established in liver metastasis

Next, we sought to characterise the heterogeneity of cell states in CRC liver metastases. We generated single nucleus Multiome RNA+ATAC data from liver metastasis of 15 CRC patients, simultaneously profiling both mRNA and chromatin accessibility in single nuclei. 7 out of the 15 of the patients had previously received chemotherapy prior to surgical resection (Supplementary Table 5). After quality control, we retained 21,354 cells. Cells from different samples were integrated using the RNA modality and clustered into cell types including malignant, T, myeloid, stromal and endothelial cells, as well as liver-specific hepatocytes and cholangiocytes (Fig. 3a and Extended Data Fig. 4a-e). Clustering based upon the ATAC modality also resulted into these cell types (Extended Data Fig. 4b). The majority of cells were epithelial cells (Extended Data Fig. 4c,d), consistent with published CRC snRNA-seq datasets[44].

We isolated and analysed the transcriptome of the malignant cells, revealing a surprisingly similar tumour structure to pCRC (Fig. 3b, Extended Data Fig. 5a-e), demonstrating that upon disseminating to the liver, metastasis-initiating cells can recreate the primary tumour structure at distant sites. Cancer cell states were largely conserved between treatment-naive and chemotherapy-treated patients (Extended Data Fig. 5f,g). Overall, cancer cell states were present in all liver metastases (with exceptions for rarer, differentiated states such as Tuft and Enteroendocrine) but individual tumours showed variation in the proportions of cancer cell states (Fig. 3c). Similar to pCRC, pro-metastatic (i)REC states lack expression of mesenchymal markers but express pEMT genes (Extended Data Fig 5d and Supplementary Table 6). In addition, EpiHR and CRIS-B signatures are enriched in both iREC and REC, whilst INFα/γ hallmarks are more specific to iREC (Fig. 3d) which express higher levels of ISGs (Fig.

3d,e and Extended Data Fig. 5e). We did not observe an HLA-high cluster in liver metastases (Fig. 3e), but this could be due to the smaller dataset or lack of MMRd samples.

To confirm these results, we integrated our dataset with published liver metastasis scRNA-seq data from 14 additional patients[20,45,46]. In addition to our previously identified cell states, we find two recently described rare non-canonical states[16]: a neuroendocrine state and a stem-like state upregulating both squamous-like and endocrine development signatures (*KRT5*, *SP5*, *MBP4*, *PROX1*, *NR2F1*), present in only one and two patients respectively (Extended Data Fig. 5h-j).

Next, to investigate the TME in liver metastases, we integrated our data with publicly available scRNA-seq data from 26 additional patients[20,29], generating a dataset containing 98,312 transcriptomes after quality control (Extended Data Fig. 6a). We identified all major cell types in metastatic (mCRC) tumours which we previously described in pCRC (Extended Data Fig. 6b,c). This includes diverse TME subpopulations such as CAF populations (C3+ iCAFs, ECM CAFs), myofibroblasts, pericytes, vascular smooth muscle cells, endothelial cells and immune cells (Extended Data Fig. 6d-j). We also recovered liver-specific sinusoidal endothelial cells (*FCN3*, *FCN2* and *CLEC4G*), hepatocytes (*TTR, APOA1*), cholangiocytes (*SOX9*, *CLDN4*, *CLDN10*) (Extended Data Fig. 6e,f), and Kupffer cells (*CD5L, CETP, TIMD4*) which were marked by high expression of a resident tissue macrophage signature (Extended Data Fig. 6h,i).

## Transcription factors regulating malignant states

To predict transcriptional regulators of cancer cell states, we next investigated the accessible chromatin landscape of cancer cells in our mCRC dataset using paired snATAC-seq data. Peaks were called in each cancer cell state and a union peakset was formed of 82,491 accessible chromatin regions, the majority (80.7%) of which are distal to promoters (Extended Data Fig. 7a). To identify putative enhancers, we correlated the expression level of genes with the accessibility of distal chromatin regions (restricted to within 250 kb of a TSS) and identified 1,444 putative enhancer-gene linkages (PE-GLs) which we clustered using k-means clustering (Fig. 4a). More than 60% of the putative enhancers in PE-GLs overlap with a set of previously predicted enhancers in CRC organoid models[47] (Extended Data Fig. 7b) demonstrating the validity of this approach and its ability to identify novel putative enhancers. We identified cell-type specific PE-GLs potentially driving the expression of important marker genes such as for stem cells (*LGR5*, *ASCL2*), Colonocytes (*SLC26A3*), (i)REC (*EMP1*, *PLAUR*, *LAMC2*) and Hypoxia (*VEGFA*) (Fig. 4a-c and Extended Data Fig. 7c,d). GEA of genes in the PE-GLs k-means clusters reveals genes in (i)REC cell states (clusters 4 and 6) are enriched for hypoxia, EMT, MAPK, PI3K and TNFα signalling; whilst WNT signalling is enriched in Stem NOTUM (clusters 1 and 2) cells (Fig. 4c). Additionally, EpiHR and CRIS-B signatures are enriched in genes belonging to clusters 4 and 6 (Extended Data Fig. 7c). Therefore, PE-GLs likely establish important gene expression programs in cancer cell states.

Transcription factors (TFs) are key regulators of cell identity and function. To predict transcriptional regulators of cancer cell states, we used chromVAR to identify TF binding motifs that are differentially accessible across cancer cell states[48]. Interestingly, hierarchical clustering of the cancer cell states by motif accessibility shows that the Intermediate state

clusters closely with Hypoxia and has increased accessibility of TF motifs that are accessible also in (i)REC states and to a lower extent in the stem cells (Fig. 4d). This further suggests that the Hypoxia and Intermediate states are plastic and transitioning states. To further nominate TFs driving changes in gene expression in different cancer cell states, we identified TFs with the highest correlation between their gene expression and the chromatin accessibility of its cognate motif (Extended Data Fig. 7e). Motifs enriched in differentiated cell states include POU2F3 (Tuft), RFX3/6 (Enteroendocrine), HNF4A (Colonocytes) and CDX2 (Stem, Colonocytes) (Fig. 4d and Extended Data Fig. 7F). Several of these TFs have previously been identified to play a role in differentiation in the healthy colon[49–53]. Among the TFs most correlated with motif accessibility are AP-1 family members (FOS, FOSB, FOSL1, FOSL2, JUNB, JUND), NF-κB subunits (NFKB1, NFKB2, RELB) and LEF1 (Extended Data Fig. 7e). AP-1 and NF-κB motifs are enriched in (i)REC states, whilst LEF1 is enriched in Stem NOTUM (Fig. 4d and Extended Data Fig. 7f). AP-1 family members are regulated by MAPK signalling pathways[54] and TCF/LEF family members are regulated by WNT signalling[55], consistent with enrichment of MAPK or WNT signatures in genes associated with (i)REC or Stem NOTUM states respectively (Fig. 4c). *De novo* motif enrichment analysis of putative enhancers in PE-GLs also demonstrates enrichment of AP-1 and NF-κB in (i)REC, and the TCF/LEF motif enriched in Stem NOTUM PE-GLs (Fig. 4a and Supplementary Table 7), indicating that these transcription factors have important regulatory roles in the malignant cell states.

To further investigate the gene regulatory networks governing (i)RECs, we predicted both genes and accessible chromatin regions regulated by TFs in cancer cell states using SCENIC+[56]. SCENIC+ identified regulons for 22 out of 26 TFs (exceptions are: NR5A2, NEUROG3, HNF1B, RFX6) whose expression correlated with motif accessibility (Extended Data Fig. 7e-f). Focusing on AP-1 and NF-κB subunits as potential drivers of (i)REC states, accessible chromatin regions in FOS, FOSB, NFKB1 and RELB regulons are more accessible in (i)REC relative to stem-like cells, which feature greater accessibility of ASCL2, LEF1, HNF4A and CDX2 regulons (Fig. 4e). Importantly, chromatin regions in JUND and HNF4A regulons are bound by JUND and HNF4A in cell line models of CRC (Extended Data Fig. 7h), indicating that our computational analysis identifies true TF binding sites in CRC cells. Interestingly, we also find TEAD1 enriched in (i)RECs and TEAD4 in the Intermediate state (Extended Data Fig. 7h,i). YAP/TAZ are co-factors for the TEAD TF family and studies have shown cooperation between YAP/TAZ/TEAD and AP-1 at enhancers in different contexts[57–59]. This is in line with the upregulation of YAP target genes in the (i)RECs. In addition, *de novo* motif analysis of chromatin regions in the RELB regulon shows significant enrichment of the AP-1 motif in these regions (Supplementary Table 8), suggesting that AP-1 and NF-κB co-operate to establish the (i)REC state. Overall, our results indicate that YAP, AP-1 and NF-κB drive emergence of a regenerative foetal-like putative pro-metastatic state in CRC.

To corroborate this, we next assessed the expression levels of genes in FOSB, RELB, ASCL2 and LEF1 regulons across the cancer cell states. FOSB and RELB target genes are expressed highest in (i)REC, whereas LEF1 target genes are expressed highest in Tuft and Stem NOTUM states and ASCL2 target genes are highest in TA2 and Stem cells (Fig. 4f). This indicates that Stem NOTUM and Stem states are regulated differently, with Stem NOTUM cells being driven more by LEF1 (likely downstream of WNT signalling), and Stem cells driven more by ASCL2 activity, which is supported by expression levels of *ASCL2* and *LEF1* and accessibility of their binding motifs (Extended Data Fig. 7i). FOS family members heterodimerise with JUN family members, we therefore combined the target genes in FOS

and JUN family member regulons to create a unified AP-1 regulon (Supplementary Table 8). The CRC AP-1 regulon significantly overlaps ($P$ = 1.32e$^{-26}$, hypergeometric test) with experimentally determined AP-1 target genes identified in another gastrointestinal cancer, oesophageal adenocarcinoma (OAC)[61] (Supplementary Table 8). Experimentally determined AP-1 target genes are expressed higher in (i)REC states in both primary and metastatic sites (Fig. 4g and Extended Data Fig. 7l). In addition, the (i)REC marker genes *EMP1* and *LAMC2* are within the FOSB regulon (Supplementary Table 8). Consistent with this observation, expression of *EMP1* and *FOSB* positively correlates in bulk CRC RNA-seq data (Extended Data Fig. 7k).

## Cancer and TME cells organised in cellular niches in liver metastasis

To investigate the spatial organisation of the cancer cell states in liver metastasis and interrogate how cancer states and the TME interact, we studied patient-specific and shared patterns among samples using spatial transcriptomics data. We generated Visium data from three liver metastatic CRC samples for which we have paired Multiome data, and additionally analysed previously published Visium data of three liver metastatic CRC samples[62].

Liver metastases exhibit distinct histological growth patterns, reflecting different ways in which cancer cells interact with the surrounding liver parenchyma[63]. The desmoplastic growth pattern shown in sample LM4 is characterised by a desmoplastic capsule surrounding the metastatic tissue that consists of fibroblasts and extracellular matrix and contains dense infiltrates of immune cells, effectively encapsulating the cancer cells from the liver (Fig. 5a,d). In contrast to this, the cancer cells in the replacement growth pattern are arranged in plates in continuity with the hepatocyte plates creating a direct contact between hepatocytes and the invading tumour cells as observed in sample P13 (Fig. 5b,d). In line with this, sample LM4 with encapsulated growth pattern presents with an enrichment of immune cells and stromal cells at the metastatic border (Fig. 5e and Extended Data Fig. 8a,b). In contrast, sample P13 with replacement growth exhibits immune and stromal cells both within the tumour and at the normal liver site and is characterised by infiltration of immunosuppressive cells into the tumour (although the T cells are found in low abundance relative to other TME cells) (Fig. 5e and Extended Data Fig. 8a,b).

Interestingly, across all three samples capturing the tumour site and liver parenchyma (LM4, P13, P3), we detect a layered spatial organisation of stromal and myeloid cells from the liver parenchyma to the central tumour stroma. C3+ iCAFs are enriched at the liver site and the tumour-liver interface (Fig. 5e and Extended Data Fig. 8c), ECM CAFs with C1QC+ and inflammatory IL1B+ macrophages at the tumour-liver border and in the tumour site, whereas myofibroblasts and immunosuppressive SPP1+ macrophages infiltrate the tumour core (Fig. 5e and Extended Data Fig. 8a,c).

Joint analysis across all spatial liver metastasis samples with SpatialDE2 deciphers spatially segregated cancer states organised into distinct local neighbourhoods, each with a distinct composition of cell subtypes and local cellular interactions (Fig. 5f-h and Extended Data Fig. 8d). We show that the iREC subpopulation congregates with perivascular cells, ECM remodelling CAFs and myofibroblasts in a cellular niche (neighbourhoods 0, 5, Fig. 5h), further supporting the hypothesis that these cells are potentially metastasis-initiating cells. Similar to

the invasive edge in pCRC, iRECs promote establishment of an immunosuppressive niche comprising SPP1+ TAMs, inflammatory IL1B+ and NLRP3+ TAMs, neutrophils, exhausted CD8+ T cells and Tregs (mainly in neighbourhoods 0 and 5, and small abundance in neighbourhood 3 Fig. 5g,h). The stem cells reside in their own niche, however they also colocalise with the Intermediate state (neighbourhood 4), further implicating potential transitions between Stem NOTUM and Intermediate cells. Interestingly, in contrast to pCRC, the REC subpopulation is separate from iRECs and immune cells and colocalises with the Hypoxia state (neighbourhood 7), suggesting that the REC state potentially arises through cellular transitions from hypoxic cells and upregulates interferon signalling upon contact with immune cells to form iREC.

Although cellular neighbourhoods are often shared among samples, the joint analysis also captures spatial features that highlight inter-patient heterogeneity (Extended Data Fig. 8e). Cellular neighbourhoods 1 and 9 denote the liver parenchyma primarily composed of hepatocytes and Kupffer cells and they are specific for the replacement and desmoplastic growth factor patterns respectively (Fig. 5h). Neighbourhood 3 comprises immunomodulatory CAFs and immunosuppressive cells and is mostly enriched at the desmoplastic rim, whereas neighbourhood 2 is largely composed of immune cells and specific to samples CRC11 and CRC09 (Fig. 5f,h and Extended Data Fig. 8d,e).

Additionally, to highlight the distinct transcriptional features that different cellular niches have, we performed differential expression analysis between the cellular neighbourhoods (Supplementary Table 9). GEA confirms that genes upregulated in iREC- and immune-enriched neighbourhoods (0, 2, 3, 5) are enriched for EMT, IFN-γ response, KRAS signalling and angiogenesis, whereas the stem-enriched niche (4) is enriched for WNT signalling pathway (Fig. 5i). Consistent with the identified spatial cellular niches, gene signature scores associated with EMT and interferon response are higher in the iREC and immune-enriched niches (Extended Data Fig. 8f,g). Conversely, these signature scores are inversely related to the WNT signalling signature, which is higher in the stem-enriched niche (Extended Data Fig. 8g-i).

To further corroborate our results, we also performed non-negative matrix factorization (NMF) on the cell2location spot-by-cell output and defined factors of co-occurring cell states. These results confirm the establishment of an immunosuppressive niche where iRECs colocalise with SPP1+, C1QC+ TAMs and CD8 Tex cells, inflammatory IL1B+ TAMs, and stromal cells such as myofibroblasts, ECM CAFs and pericytes (fact_8, Extended Data Fig. 8j). Furthermore, the Stem NOTUM cells congregate with the Intermediate state (fact_3), whereas the RECs are in close proximity to Hypoxia (fact_5), supporting our findings (Extended Data Fig. 8j).

## CAFs and immunosuppressive cells mediate metastasis-initiating cells

Given the close proximity of CAFs and immunosuppressive myeloid subpopulations with the iREC state in both primary and liver metastasis, we next investigated potential mediators of cellular crosstalk between these compartments. We first used CellPhoneDB[64] to identify enriched receptor-ligand pairs amongst the cell states residing in the spatial cellular niche

surrounding iRECs. Next, to determine which ligands potentially promote the inflammatory regenerative programme in cancer cells, we identified ligands that are predicted to induce the AP-1 and NF-κB (RELB) regulons using NicheNET[65]. Coupling these two computational methods, we efficiently linked transcriptional profiles of cancer cell states with potential upstream regulators within the spatial niche that induce expression of AP-1 and RELB regulons in iRECs.

Our results reveal a candidate list of CAF- and myeloid-derived ligands potentially activating AP-1 target genes, with corresponding receptors that are upregulated in the (i)RECs (Fig. 6a). GEA of the predicted ligands shows enrichment of EMT, MAPK signalling, inflammatory response and IFN-γ response, processes that are activated in iRECs relative to other cancer cell states (Fig. 3c and 6b). Several ligands expressed in ECM CAFs, myofibroblasts and pericytes have established roles in inducing EMT, invasion and immune evasion, including the cytokines and growth factors TGF-β (*TGFB2*/*3*)[66], *HGF*[67], fibroblast growth factors (*FGF1*/*2*/*7*)[68], VEGF (*VEGFA/B)*[69] and *IL6*[70] (Fig. 6a). In addition, ECM CAFs and myofibroblasts are likely involved in matrix remodelling through secretion of matrix metalloproteinases (*MMP2*) and collagens[71] (Fig. 6a). Another candidate CAF-secreted ligand is *IL-33* which has been shown to activate and maintain immunosuppressive TAMs[72,73]. Expression of CD39 (encoded by *ENTPD1*), which together with CD73 converts ATP to adenosine to prevent immune activation is also high in the CAFs[72].

Myeloid and T cells are predicted to induce an inflammatory phenotype in iRECs through activation of NF-κB, driven by proinflammatory genes such as IFN-γ (*IFNG*) and *CCL5* expressed in exhausted T cells, as well as *IL1B* and *TNF* expressed in inflammatory TAMs (Extended Data Fig. 9a). Furthermore, *APOE* is highly expressed in SPP1+ and IL1B+ TAMs and has been shown to induce expression of immunosuppressive factors such as *CXCL1* and *CXCL5*, through LDL receptor (*LDLR*) and NF-κB signalling[74]. In a similar way, *ANXA1* has been implicated in promoting immune suppression[75] and has also been associated with resistance to chemotherapy in colorectal cancer[76]. Other predicted ligands include chemokines such as CXCL2 and CXCL3 that can recruit neutrophils and myeloid-derived suppressor cells in the TME, contributing to establishment of an immunosuppressive environment and resistance to therapy[77].

Ligand-target links were inferred based on the regulatory potential scores computed by NicheNet and prior knowledge of ligand-target associations. Potential target genes of expressed ligand-receptor interactions in the iREC niche are summarised in Fig. 6c. Spatial transcriptomics data show spatial enrichment of ligands in the cellular neighbourhoods surrounding iRECs (neighbourhood 0,5) (Extended Data Fig. 9b) which confirms our predictions. To extend these observations into a larger population, we also analysed TCGA bulk RNA-seq data from 609 pCRC samples. These results show high correlation of ligands predicted to be secreted from the IL1B+ macrophages and exhausted CD8 T cells with the (i)REC signature in comparison to stem cell signatures, further supporting our findings (Fig. 6d).

Collectively, these results highlight ligands predicted to induce a regenerative and inflammatory pro-metastatic phenotype in iRECs and could lead to potential therapeutic strategies by targeting specific molecular mechanisms in the cellular niche that sustains the iREC state.

# Discussion

To form metastasis, cancer cells undergo rapid phenotypic transitions to leave the primary site, survive in circulation, adapt to new microenvironments and regenerate tumours at distant sites. Cellular plasticity - the ability of cells to reversibly change their phenotype as a response to external signals - provides cancer cells access to developmental or regenerative programs to adapt to new environments or stress.

Using single-cell multiomics data from patients with primary CRC and liver metastasis, here we show that the heterogeneous malignant cell states in pCRC are re-established in metastasis. We reveal states with regenerative and foetal signatures and high YAP signalling (iREC/REC) that are likely driving metastatic dissemination in the liver. Our analysis of joint expression and chromatin accessibility data identifies the transcription factors AP-1, NF-κB and YAP as regulators of these pro-metastatic cells, suggesting that pathways regulated by those TFs can serve as potential therapeutic targets to eliminate (i)RECs. While AP-1 and NF-κB have a well-established role in regulation of inflammatory responses[60], their role in driving regenerative pro-metastatic cancer states remains unknown. AP-1 and NF-κB have also been implicated as transcriptional regulators in foetal intestinal organoids[78], supporting emerging evidence that cancer progression often requires reacquisition of developmental transcriptional programs. Whether specific genomic mutations also facilitate greater epigenetic flexibility and access to previously restricted programs from developmental or regenerative origin to cancer cells remains to be investigated.

Interestingly, we show that a subset of the metastasis-initiating cells upregulates inflammatory genes, suggesting that interactions with immune cells may activate this response in cancer cells that are in proximity and that metastasis-initiating cells may use this response to escape immune attack. Mouse models of CRC micrometastasis show that iRECs could potentially be more sensitive to immune checkpoint blockade[25], therefore future efforts using immunotherapies should be explored.

The existence of a hybrid intermediate state expressing both REC and stem markers indicates a transition between (i)REC and stem states, presenting a significant challenge for targeting either population. Our patient-derived data is therefore in agreement with previous mouse studies demonstrating that the majority of metastases are seeded by LGR5- cells, however these pro-metastatic regenerative cells are able to transition back to LGR5+ cells and re-establish the cellular heterogeneity of primary tumours to form metastases. Interestingly, in both primary CRC and metastasis, we detect a subpopulation with upregulated hypoxic signatures that also expresses REC markers. At the invasive edge of primary CRC samples, we see the hypoxic state both at the core and the invasive edge, whereas at the liver metastatic site, it colocalises with the RECs. This suggests potential cellular transitions from the hypoxic state to RECs that can subsequently upregulate interferon signalling upon contact with immune cells. This is supported by hierarchical clustering of the cancer cell states by motif accessibility that shows shared motifs of the hypoxic state with RECs. Our results therefore indicate that in the dynamic intestinal epithelium, cellular transitions are complex, and plasticity may go beyond the stem-to-regenerative transition.

Our spatial transcriptomics analysis demonstrates spatially segregated cancer states organised into neighbourhoods with local cell-cell interactions, pointing to the role of the microenvironment in mediating and sustaining distinct cancer subpopulations. Despite the inter-patient heterogeneity, we reveal recurrent patterns among all patients. Specifically, we find that the phenotype of iREC cells is likely maintained by specific TME subpopulations, including CAFs and immunosuppressive and inflammatory myeloid and CD8 T cells. The ligand-receptor interactions predicted from our results to activate the invasive regenerative and inflammatory state could provide therapeutic strategies by targeting specific molecules in the cellular niche mediating the iREC state. Additionally, our analysis shows spatial organisation specific to distinct metastatic growth patterns, with clear differences between desmoplastic and replacement growth patterns. Considering that different histological growth patterns are a prognostic marker for patient outcome and might be associated with different immune responses and mechanisms of treatment[79,80], our results emphasise a need to further investigate how these growth patterns arise.

Non-genetic plasticity plays a crucial role in CRC initiation, progression, metastasis and resistance to therapy and represents a formidable challenge in cancer therapy. By identifying and characterising the distinct malignant states, their regulatory drivers and microenvironmental cues that maintain them in primary and metastatic CRC, our findings might lead to novel therapeutic opportunities to impair plasticity, by restricting transitions into invasive regenerative phenotypes or promoting transitions into immunosensitive states.

# Methods

## Human tissue samples

Metastatic colorectal cancer tissue was provided by the Barts Cancer Tissue Bank (Research Ethics Committee approval, 2014/LO/2031 (City and Hampstead) and renewed 2019/LO/1700, www.cancertissuebank.org; CTB approval 2020/05/QM/ME/P/FreshTissue and 2021/01/QM/EM/P/Blood&Tissue). We accessed archived frozen tissue samples and collected fresh tissue samples from the Royal London Hospital, Barts Health NHS Trust. Fresh tissue samples were flash frozen in a dry ice/ethanol bath and stored at -80°C.

## Tissue dissociation and single cell multiomics

10x Genomics Multiome technology was used to generate paired snRNA- and snATAC-seq data from the same cell. Nuclei were isolated using a method based upon the salty EZ-10 V2 method (dx.doi.org/10.17504/protocols.io.buxnnxme). Frozen patient tissue was cut in a sterile dish on dry ice into a rice sized section and the remaining tissue was stored at -80°C. Frozen tissue was then transferred into 300 mL Salty-Ez10 Lysis Buffer (10 mM Tris-HCl pH 7.4, 146 mM NaCl, 1 mM $CaCl_2$, 21 mM $MgCl_2$, 0.03% Tween-20, 1% BSA, 10% EZ lysis buffer [Sigma, NUC101-1KT], 1 U/mL RNase inhibitor [Merck, 3335399001], 1 mM DTT, nuclease free water) in a 1.5 mL tube and the sample was homogenised by stroking 15x with a douncer (Fisher Scientific, 13236679), keeping the sample on wet ice. 700 mL Salty-Ez10 lysis buffer was added and the sample was pipette mixed using a wide-bore pipette tip. The sample was then incubated for 3 minutes on wet ice, and was pipette mixed with wide-bore

tips twice during the incubation. The nuclei suspension was then passed through a 70 µm strainer (Fisher Scientific, 15346248) into a fresh 1.5 mL tube. Nuclei were then centrifuged at 500 RCF for 5 minutes at 4°C, before discarding the supernatant. Nuclei were resuspended in 500 mL WRB2 buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% Tween-20, 1% BSA, 0.01% digitonin [Thermo Fisher, BN2006], 1 U/mL RNase inhibitor [Merck, 3335399001], 1 mM DTT, nuclease free water), gently pipetting using wide-bore tips. Nuclei suspension was then passed through a 40 mm strainer (Sigma, BAH136800040) into a fresh 1.5 mL tube. Nuclei were then centrifuged at 500 RCF for 5 minutes at 4°C before discarding the supernatant. Nuclei were then washed by re-suspending in 500 mL WRB2 buffer using wide-bore pipette tips and the centrifugation step was repeated. Nuclei were then re-suspended using a wide-bore pipette tip in 10-40 µL 1X nuclei buffer (10x Genomics, 2000153) depending on the size of the pellet. Nuclei were counted by diluting 2 µL of the nuclei suspension 10-fold with WRB2 buffer. 10 µL of diluted nuclei were then added to 10 µL dead cell stain (2% Ethidium homodimer-1 [ThermoFisher, E1169], 30% glycerol, nuclease free water) and counted on a Countess II FL Automated Cell Counter (ThermoFisher; AMQAF1000). Downstream processing was performed immediately using Chromium Next GEM Single Cell Multiome ATAC + Gene Expression Reagent kits (10x Genomics, 1000285) and Chromium Next GEM Chip J Single Cell Kit (10x Genomics, 1000234). ATAC and gene expression libraries were sequenced on an Illumina NextSeq 550 or Illumina NovaSeq 6000.

## Analysis of primary CRC scRNA-seq datasets

Publicly available scRNA-seq data from primary CRC were combined from 4 studies[18–20,29]. Raw scRNA-seq counts were analysed using Scanpy[81] (v1.9.1). Quality control and initial filtering was done on each dataset separately before integrating them into a single dataset. Scrublet[81,82] (v0.2.3) was run per sample to identify potential doublets. The raw gene expression matrices were filtered using the following quality control criteria: (1) > 300 genes; (2) < 20% mitochondrial reads. Ribosomal and mitochondrial genes were discarded. The datasets were concatenated into a single gene expression matrix. The data were normalised with a scale factor of 10,000 and log1p-transformed. We extracted 2000 highly variable genes (HVGs) using the Seurat V3 method. The data were batch-corrected using scVI[83] (v0.16.4) on raw counts and HVG, aligning all datasets and patients, and correcting for unwanted sources of variation: mitochondrial and ribosomal percentages. We used default parameters (one hidden layer with size 128 and latent size 10). A neighbourhood graph (KNN) was built using the resulting 10 latent embeddings of all cells obtained from scVI to perform Leiden clustering, and UMAP visualisation.

To define major cell types, cells were clustered using the Leiden method (resolution parameter r = 0.2). Differentially expressed genes were identified for each cluster using Wilcoxon rank-sum test with Benjamini-Hochberg p-value correction in Scanpy. We selected differentially expressed genes with an adjusted p-value lower than 0.05 and a log2 fold-change higher than 1 and expression observed in a minimum of 10% of cells in a cluster. The transcriptomes were partitioned into 8 major cell types (epithelial, stromal, endothelial, T/natural killer (NK)/innate lymphoid cells (ILC), myeloid, mast, plasma and B cells) by comparing differentially expressed genes and canonical markers from the literature.

Subsequently, the same integration and clustering analysis was applied iteratively to the cells of each major cell type separately to identify and annotate cell states. For each cell, the cell cycle phase scores (G1, S, G2/M) were computed based on the expression of S and G2/M markers using *tl.score_genes_cell_cycle* using cell cycle genes identified previously[84]. In the immune and stromal cell analysis, clusters with high numbers of doublet cells were removed by checking for expression of markers of more than one cell type. For the T/NK/ILC subpopulations, two clusters were removed as they exhibited markers from myeloid and T cells and B and T cells respectively. For the myeloid cells, three clusters showing hybrid transcriptional signatures (B/Myeloid, T/Myeloid, Epithelial/Myeloid) and a high scrublet score were discarded. For stromal cells, four clusters exhibiting doublet signatures were excluded.
Cancer cell states in pCRC were identified by subsetting and re-clustering epithelial cells. Cell cycle phase scores were calculated as described above and the top 2000 HVGs were determined for the epithelial cells. The number of observed genes, percentage of mitochondrial and ribosomal reads, and S-phase and G2-phase scores per cell were regressed from log1p-normalised counts using *pp.regress* and then scaled using *pp.scale*. PCA was then computed on the top 2000 HVGs. The python implementation of Harmony[85] was then used to batch correct the data, using the *run_harmony* function with the patient of origin as the batch key. A neighbourhood graph was then constructed (*pp.neighbours*) from the corrected principal components followed by UMAP representation (*tl.umap*). Cells were then clustered (*tl.leiden, resolution=1.2*). Clusters that contained potential doublets were removed and the above steps were repeated to re-cluster the cells.

Primary tumour samples may contain normal colon cells. Therefore, copy number alterations were predicted from scRNA-seq data using the python implementation of inferCNV (https://github.com/icbi-lab/infercnvpy) to identify malignant cells. Normal epithelial cells in SMC, KUL[18] and Pelka *et al*.[19] datasets were used as a reference for inferCNV. Copy number alterations were inferred using bins of 100 genes, with a stepsize of 1 using the *inferCNV* function. CNV clustering was then performed using (*cnv.tl.pca*, *cnv.pp.neighbors, cnv.tl.leiden*). Normal reference cells were only present in 5/51 CNV clusters making up 23.0%, 8.8%, 4.0%, 0.16%, and 0.05% cells in each respective cluster. Epithelial cells present in 3 CNV clusters containing 23.0%, 8.8% and 4.0% reference cells were removed, resulting in 60,526 malignant cells.

Cell clustering steps (starting from re-computing cell cycle scores and re-calling HVGs) were then repeated, resulting in 17 leiden clusters (0-16). Cluster 16 was removed because it potentially contained doublets (stromal markers, *COL1A2*, *COL3A1*). Clusters 5 and 12 contained secretory cells which were subsetted and re-clustered (resolution = 0.4) into Enteroendocrine, Goblet and Tuft cells. Cluster 10 contained both HLA-high and iREC cells, and was sub-clustered (resolution = 0.2) into HLA-high cells (express HLA genes and ISGs, but lack pEMT genes) and iREC (express ISGs and pEMT genes). Cluster 0 contained both Colonocyte and Hypoxia cells, and was subsetted and re-clustered (resolution = 0.4) into Colonocyte (*SLC26A3*) and Hypoxia (lack *SLC26A3* expression) cells.

The abundance of cancer cell states was compared between mismatch repair-proficient (MMRp) and mismatch repair-deficient (MMRd) pCRC tumours using Milo[83] (v1.6.0) in R. Milo was performed on Harmony-corrected scRNA-seq data, using default parameters except where specified. A k-nearest neighbour graph was constructed using $k$ = 25 and $d$ = 20, followed by defining neighbourhoods using prop = 0.1, $k$ = 25 and $d$ = 20 with

refinement_scheme set as 'graph'. Neighbourhood testing was performed using fdr.weighting set as 'graph-overlap'.

We scored the macrophage and monocyte subpopulations in pCRC and mCRC tumours for signatures derived from recurrent tumour-associated macrophage (TAM) and tumour-infiltrating monocyte subsets obtained from a single-cell RNA-seq analysis spanning over 15 tumour types (including CRC)[37]. In particular, we included signatures of lipid-laden TAMs, pro-angiogenic TAMs, inflammatory cytokine-enriched TAMs, interferon-primed TAMs, resident-tissue macrophages, classical monocytes, nonclassical monocytes and intermediate monocytes. We computed the gene signature scores for each transcriptome using Scanpy's *tl.score_genes* function.

## Analysis of single-nuclei Multiome data

Single-cell Multiome data was pre-processed using the 10x Genomics Cell Ranger ARC (v2.0) pipeline. Cellranger-arc count was used to align reads to hg38 (GRCh38) and to generate barcode counts. Filtered feature barcode matrices and ATAC fragment files were used for subsequent analysis.

The quality of snATAC-seq data was assessed using ArchR (v1.0.1)[86]. Cells (i.e. barcodes) were retained if transcription start site (TSS) scores were greater than 4 and the number of unique ATAC fragments was greater than 1500 per cell. Following snATAC-seq barcode filtering, quality control of snRNA-seq was performed for the same cell barcodes using Seurat (v4.1.0)[87]. Cells with less than 300 genes and more than 10% mitochondrial DNA reads were removed. Mitochondrial and ribosomal genes were removed. We observed ambient RNA contamination in our Multiome gene expression data which we decontaminated using decontX[88]. decontX also generates a decontamination score per nuclei and cells with a high decontamination score are potential doublets/low quality cells. We therefore removed nuclei with a decontX contamination score greater than 0.5.

Following decontamination of snRNA-seq data, decontaminated counts were loaded into an anndata object using Scanpy's (v1.9.1)[81] *read_mtx* function. Decontaminated counts were normalised (*pp.normalize_total* with scale factor of 10,000), log transformed (*pp.log1p*) and the top 2000 HVGs were identified (*pp.highly_variable_genes, flavor = 'seurat_v3', n_top_genes=2000, batch_key = 'Sample'*). scVI (v0.16.4)[83] was used to batch correct snRNA-seq data by sample (*model.SCVI.setup_anndata*, *model.learn*). 10 latent variables were used for batch correction, and the number of genes and percentage of mitochondrial reads were regressed out by inclusion as model covariates. A neighbourhood graph (*pp.neighbors*) was constructed from the latents and a UMAP (*tl.umap, min_dist=0.3*) was embedded followed by clustering of cells (*tl.leiden, resolution = 0.5)*. Cell types were annotated based upon the expression of known marker genes.

To cluster all cells based upon snATAC-seq data, Signac v1.5[89] was used. A peakset was formed by calling pseudobulk peaks on cell types annotated in snRNA-seq data. Signac functions *callPeaks(group.by = 'Cell_type')*, *keepStandardChromosomes (pruning.mode = "coarse")* and *subsetByOverlaps (ranges = blacklist_hg38_unified, invert = TRUE)* were used to call peaks and create a peakset. MACS2[90] was used for peak calling. Reads in peaks were then quantified using the *FeatureMatrix* function. The Signac pipeline was then run:

*FindTopFeatures(min.cutoff = 5)*, *RunTFIDF*, *RunSVD*, *RunUMAP*. Samples were then integrated using Signac: *SplitObject (split.by = 'Sample')*, *FindIntegrationAnchors(reduction = "rlsi", dims = 2:30)*, *IntegrateEmbeddings(dims.to.integrate = 1:30)*, *RunUMAP(dims = 2:30)*.

Following annotation of the major cell types, cancer cells were subsetted and analysed, using Scanpy and scVI to integrate the samples as described above. Cell cycle scoring was performed using *tl.score_genes_cell_cycle* using cell cycle genes identified previously[84]. The top 2000 HVGs were then re-identified and scVI batch correction was repeated, using 20 latent variables. The effect of the cell cycle was also used as a covariate in scVI. Doublets were removed by removing clusters that expressed marker genes for more than one cell type. Cancer cell states were annotated based upon the expression of marker genes, gene ontology analysis of the top genes in each cluster or by scoring cells for gene signatures (*tl.score_genes*).

Following analysis and clustering of the snRNA-seq data, the snATAC-seq data was further processed using ArchR. Peaks were called in each cell state and a union peakset was created using the *addGroupCoverages*, *addReproduciblePeakSet(cutOff = 1x10$^{-5}$)* and *addPeakMatrix* functions in ArchR; MACS2 was used for calling peaks.

To visualise snATAC-seq genome coverage, bigwig files were generated using *getGroupBW(tileSize = 50)* in ArchR and then visualised in the Integrative Genomics Viewer (IGV)[91]. Putative enhancer gene linkages (PE-GLs) were predicted in mCRC cancer cells based upon the correlation of gene expression with peak accessibility using *addPeak2GeneLinks(dimsToUse = 1:20, k = 100)*. The overlap of putative enhancers (pE) in PE-GLs with a set of chromHMM enhancers identified in CRC organoids[47] was determined using bedtools *intersect*[92].

To infer changes in TF activity, chromVAR deviations enrichment analysis[48] was calculated using ArchR. CIS-BP motif annotations were added to peaks using *addMotifAnnotations* and background peaks were calculated using *addBgdPeaks*. ChromVAR deviations were then calculated using *addDeviationsMatrix*.

For integrated analysis of transcription factor mRNA expression and motif enrichment analysis, decontaminated count data was loaded into an ArchR object using *addGeneExpressionMatrix*. TF mRNA expression was correlated with transcription factor binding motif accessibility using ArchR's *correlateMatrices* function.

*De novo* motif enrichment analysis was carried out using HOMER (v4.11)[93], *findMotifsGenome -size 200*.

To identify TF motifs enriched in topics and differentially accessible regions (DARs), a cistarget database was created using peaks in the mCRC cell state union peakset and the scenicplus public motif collection (v10nr_clust_public)[56]. DNA sequences of peaks in the union peakset were obtained using *bedtools getfasta*[92]. The create_cistarget_motif_databases python script (https://github.com/aertslab/create_cisTarget_databases) was then used to create ranking and scores databases.

SCENIC+ (v1.0.1)[56] was used to predict both chromatin regions (i.e. putative enhancers) and genes regulated by transcription factors in mCRC cell states. Rather than re-calling peaks with SCENIC+, peaks in the cancer cell state union peakset that was created with ArchR were used. Only cell barcodes that passed earlier QC filtering steps and genes expressed in a minimum of 60 cells were used. First a pycisTopic object was created using *create_cistopic_object_from_fragments*, using cancer cell barcodes and metadata, the blacklist peaks from ArchR hg38, and fragment files generated using 10x cellranger-arc. Topic modelling was used to identify variable open chromatin regions across cells, using *run_cgs_models*. 24 topics were selected and the model was added to the pycisTopic object using *add_LDA_model*. Putative enhancer regions for TF motif enrichment were identified from (1) regions assigned to topics and (2) differentially accessible regions. (1) Region-topic probabilities were obtained using *binarize_topics(method='ostu')* and *binarize_topics(method='ntop', ntop=3000*). (2) DARs were determined using *impute_accessibility(scale_factor=$10^6$)*, *normalize_scores(scale_factor=$10^4$)*, *find_highly_variable_features*, *find_diff_features*. Only open chromatin regions on known chromosomes were retained. TF motif enrichment was then performed against a custom cistarget database (see above) using *run_pycistarget*. A SCENIC+ object was then created (*create_SCENICPLUS_object*) from the cistopic object, log1p-normalised snRNA-seq data, and TF motif enrichment results generated using pyscistarget. TF-to-gene adjacencies were calculated (*arboreto_with_multiprocessing --method grnboost2*) with pySCENIC[94] and added to the SCENIC+ object using *load_TF2G_adj_from_file* to reduce SCENIC+ memory requirements. SCENIC+ was then run to identify TF regulons in the mCRC cell states using *run_scenicplus*(*biomart_host*="http://sep2019.archive.ensembl.org/").

## Integrated analysis of normal colon scRNA-seq datasets

scRNA-seq data of healthy colon epithelial cells[19,95,96] was integrated using Harmony in Scanpy as described for malignant cells in primary CRC datasets. Only cells with less than 10% mitochondrial reads and greater than 300 observed genes were retained. Covariates for the number of observed genes per cell and cell cycle effects (using the cell cycle difference score, calculated by subtracting G2/M score from S-phase score) were regressed out using scanpy's *pp.regress* function. 30 Harmony-corrected principal components were used to construct a neighbourhood graph. The resolution was set as 1.0 for leiden clustering.

We observed that expression of *LGR5* and other stem cell marker genes was low in stem cells in the healthy colon scRNA-seq. Therefore to identify a stem cell signature for comparing to our Multiome data from CRC liver metastases we identified a 'multiome' stem cell signature. snRNA-seq Multiome data of healthy colon epithelial epithelial cells[97] was integrated using Harmony in Scanpy. scATAC-seq quality control metrics were calculated using ArchR and scRNA-seq quality control metrics were calculated using Scanpy. Cells were retained with a TSS score > 5, number of ATAC fragments > 2000 and number of observed genes > 400. Ambient RNA was decontaminated using scAR[98]. Mitochondrial and ribosomal genes were removed and samples were integrated using Harmony as described above, but using 25 principal components. Cells were clustered using the leiden method with a resolution of 0.1. Epithelial cells were then sub-clustered, using 20 principal components and a resolution of 0.4 for leiden clustering. A stem cell cluster was then identified based on the expression of marker

genes (*LGR5*, *ASCL2*, *SMOC2*) and Seurat's *FindAllMarkers* function was used to identify a stem cell signature.

## Integration of published and Multiome liver metastatic CRC data

To build a comprehensive scRNA-seq reference dataset for TME of liver metastatic CRC tumours, we integrated our snRNA-seq data with publicly available scRNA-seq datasets[20,29] using scVI (v0.16.4). Each cell source was processed independently with Scanpy workflow (v1.9.1) before integrating them into a single batch-corrected dataset. While the counts were decontaminated for single-nuclei data, raw counts were used for single-cell data. We extracted 2000 HVGs using the Seurat V3 method. We log-normalised the raw counts using a scale factor of 10,000. The data were batch-corrected using scVI[83] (v0.16.4) on raw counts and HVGs only, aligning all datasets and patients using cell source and patient as categorical covariates, and correcting for unwanted sources of variation: mitochondrial and ribosomal percentages. We used the default parameters (1 layer of size 128, 10 latent variables). The resulting scVI latent space of size 10 was used to build a KNN graph (n_neighbors=15) and to perform UMAP visualisation and Leiden clustering (r=0.2). Major cell types were annotated based on differentially expressed genes for each cluster and expression of canonical makers from the literature. Stromal, endothelial, myeloid, T/NK/ILC and hepatocytes cells were re-analysed separately, repeating batch correction with scVI, dimensionality reduction and Leiden clustering to annotate fine-grained cell states.

Malignant cells from our Multiome data were integrated with malignant cells from published datasets[20,45,46] of liver metastatic CRC tumours using CCA in Seurat using each patient as a batch. We log-normalised the raw counts using a scale factor of 10,000 and extracted 2000 HVGs using the Seurat V3 method. Covariates for the number of observed genes per cell and cell cycle effects were regressed out. Mitochondrial and ribosomal genes were removed. Cells were clustered using the default louvain method with a resolution of 1, using 30 principal components. Cancer cell states were annotated based upon the expression of marker genes, gene ontology analysis of the top genes in each cluster or by scoring cells for gene signatures.

## Differential gene expression in sc/sn-RNA-seq datasets

Differentially expressed genes were identified in sc/sn-RNA-seq datasets using a Wilcoxon rank-sum test using Seurat's *FindAllMarkers* (logfc.threshold > 0.25, min.pct = 0.1) function.

## 10X Visium spatial transcriptomics library preparation

10 µm sections were taken from fresh frozen liver metastatic CRC samples from three patients. The tissue permeabilisation time was optimised using a Visium Spatial Tissue Optimization Slide & Reagent Kit (10x Genomics, 1000193), following the manufacturer's instructions (Rev D). A 6 minute permeabilisation time was selected.

Spatial transcriptomic libraries were created using a Visium Spatial Gene Expression Slide & Reagent Kit (10x Genomics, 1000187), following the manufacturer's instructions (Rev E). Libraries were sequenced on an Illumina NovaSeq 6000.

## Visium spatial transcriptomics analysis

Using 10X Space Ranger (v.1.3), count matrices for each sample were generated. FASTQ files were aligned to the human genome reference version GRCh38-2020-A. Automated tissue detection was performed, and the spot locations were aligned to the fiducial border spots in the H&E slide image to select spots located on the tissue. This generated data containing 1192 (CRC08), 937 (CRC09), and 1774 (CRC11) spots. Each Visium sample was processed independently using Scanpy (v1.9.1). Basic filtering was performed by discarding ribosomal and mitochondrial genes. We filtered out genes expressed in less than 3 spots and spots with fewer than 5 genes expressed. Segmentation was performed using the H&E image to approximate the number of nuclei per spot using Squidpy (v1.2.2) pipeline[99]. Downloaded publicly available 10X Visium data were also processed in the aforementioned manner: four primary CRC samples capturing the tumour core and the invasive edge derived from one patient[41] and three liver metastatic CRC samples derived from three patients[29]. In primary CRC Visium samples, normal spots were excluded.

We visualised the spatial distribution of signatures derived from 41 meta-programs obtained from a single-cell RNA-seq analysis of 24 tumour types (including CRC)[100]. We computed the gene signature scores for each spot using Scanpy's *tl.score_genes* function.

To assign cell types and cancer cell states annotated by our scRNA-seq analysis to spots, we used the deconvolution-based method cell2location (v0.1)[42]. Leveraging our annotated scRNA-seq reference, cell2location estimates the abundance of each cell type at each spot. Briefly, cell2location estimated cell type signatures from our raw count scRNA-seq dataset, removing genes expressed in less than 5 cells. Gene expression profile at each spot was decomposed into a weighted linear combination of cell type signatures. Each Visium sample was analysed separately. Additionally, to infer common patterns across a given tumour site (either primary CRC or liver metastatic CRC), we performed a joint inference by integrating and normalising Visium data across four primary CRC samples and six liver metastatic CRC samples respectively. We used raw spatial mRNA counts, filtered to genes shared with the scRNA-seq data. Cell2location uses priors on the tissue and experiment quality, such as the number of cells per spot. We determined the average number of nuclei per spot upon nuclei segmentation analysis (n=3) and set the regularisation of within-experiment variation in RNA detection sensitivity ($\alpha = 20$) while the remaining hyperparameters were set to default. The model was trained for 30,000 iterations using GPU acceleration. We visualised the cell abundance and the absolute amount of mRNA, which represents the amount of mRNA contributed by each cell type in each location.

We used SpatialDE2 (v1.1.1.dev103+g78da0ac)[43] on raw mRNA counts to identify tissue regions in two Visium samples capturing the tumour core and the invasive edge in primary CRC (A1 and C1) and two Visium samples capturing different growth patterns in liver CRC metastasis (P13 and LM4), as this method takes into account spatial information. Briefly, the model is based on a Bayesian hidden Markov random field and leverages a graph representation of Visium data, assigning a cluster label to each spot based on its gene expression profile and its neighbouring spots. Each sample was analysed separately. We computed spatially variable genes on the slide and retained those with an adjusted p-value lower than 0.001. We used the top 2000 spatially variable genes to construct a graph which

embeds the spatial relationships among spots. SpatialDE2 determined regions in each Visium slide with a spatial smoothness parameter s (s=2 for A1, s=0.5 for B1, s=0.7 for C1 and D1 and P13, s=0.2 for LM4). Clusters were merged to identify spatial clusters matching histological annotations.

We identified differentially expressed genes between the two spatial clusters denoting the invasive edge and the tumour core using Wilcoxon rank-sum test with Benjamini-Hochberg p-value correction in Scanpy. We selected differentially expressed genes with an adjusted p-value lower than 0.05 and a log2 fold-change higher than 0.25 and expression observed in a minimum of 10% of spots in a spatial cluster. We performed gene set enrichment analysis of the differentially expressed genes at the invasive edge and the tumour core using GSEApy (v0.10.8)[101], and 'MSigDB_Hallmark_2020' and 'KEGG_2021_Human' gene sets.

To identify spatial cellular neighbourhoods, the subsequent analyses were applied separately to Visium data from primary CRC samples and to Visium data from liver metastatic CRC samples. Joint analysis of the Visium samples from the same tumour site enabled us to identify common patterns across samples. To identify colocalization of cell types and to build cellular neighbourhoods, we used the mRNA counts contributed by each cell state in each spot estimated by cell2location to partition each Visium slide into distinct cellular niches using SpatialDE2 (s=0.1 for joint primary CRC analysis, s=1.2 for joint liver metastatic CRC analysis). We built cellular neighbourhoods based on the estimated cell type abundance profiles of the spot itself and the surrounding neighbours, with the underlying assumption that spots having similar cell type abundance profiles will be grouped together. We leveraged the 5% percentile of the posterior distribution of the mRNA counts estimated by cell2location, i.e. the number of mRNA molecules contributed by each cell state in each spot.

Additionally, to validate the identified cellular neighbourhoods, we used the non-negative matrix factorisation (NMF) module from cell2location to identify spatial co-occurrence of cell types. NMF decomposes cell type abundance estimates from cell2location into factors of cell types that colocalise. This model assumes an additive decomposition, entailing that multiple factors can co-exist at a single spot. The model was trained for a range of {5, ..., 20} factors. We chose the decomposition into 7 factors for the joint analysis of the primary CRC analysis, and 10 factors for the joint liver metastatic CRC analysis as these configurations captured a reasonable number of microenvironments without splitting the cell states into many distinct factors.

We identified differentially expressed genes between the spatial cellular neighbourhoods using Wilcoxon rank-sum test with Benjamini-Hochberg p-value correction in Scanpy. We selected differentially expressed genes with an adjusted p-value lower than 0.05 and a log2 fold-change higher than 0.25 and with expression observed in a minimum of 10% of spots in a spatial neighbourhood. We performed gene set enrichment analysis of the differentially expressed genes in each cellular neighbourhood using GSEApy (v0.10.8)[101], and 'MSigDB_Hallmark_2020' and 'KEGG_2021_Human' gene sets.

## Spatially resolved ligand-receptor interactions analysis

Spatial analysis of liver CRC metastases from patient samples identified cellular niches of segregated cancer cell states and cell subpopulations of the tumour microenvironment. Two different computational methods, CellPhoneDB (v3.1.0, database v4.0.0)[64] and NicheNet (v2.0.0)[65], were leveraged to investigate cell-cell interactions in the cellular neighbourhood surrounding the iREC cancer cell state. First, to predict the potential ligand-receptor interactions between cell types of the tumour microenvironment (senders) and the iREC cancer cell state (receiver), CellPhoneDB was performed on the identified cellular neighbourhood surrounding the iREC cancer cell state. The ligand-receptor interactions were inferred using our single-cell transcriptomics dataset of liver CRC metastases, by focusing on the cell types and cancer cell states that were congregating in the cellular neighbourhood. Ligand-receptor interactions satisfying the following criteria were selected: (1) all ligands and receptors were expressed in at least 10% of the cells of each cell state; (2) the ligand-receptor interactions between two cell states were inferred using the statistical analysis method in CellPhoneDB with a p-value threshold of 0.05; (3) ligand-receptor interactions were pruned based on mean expression levels. Second, the predicted interactions were further filtered using NicheNet, retrieving ligand-receptor interactions whose downstream TF is active, as inferred by SCENIC+. Combining TF activity to ligand-receptor interactions using NicheNet highlighted the relevant interactions that activate downstream signalling in the responder cell state. Specifically, ligand-receptor pairs known to induce AP-1 regulon and NF-κB regulon (RELB) expression in the iREC state were investigated. We considered genes positively regulated by AP-1 and NF-κB respectively. Inferred ligands were ranked according to the prior potential score, i.e. how well a ligand induces the expression of target genes of the AP-1 regulon and the NF-κB regulon respectively. Additionally, ligands were not only prioritised based on their potential score but also according to the ligand and receptor gene expression. We computed the average scaled gene expression values of ligands in senders and of corresponding receptors in the iREC state and other cancer cell states. A final set of relevant ligand-receptor interactions satisfying the following criteria were retrieved: (1) common ligand-receptor interactions between NicheNet and CellPhoneDB analyses that were statistically significant in CellPhoneDB analysis; (2) interactions inferred by NicheNet that were not present in the CellPhoneDB database (3) the average scaled expression of the corresponding receptor was higher in the iREC state relative to other cancer cell states. A circos plot using R package circlize (v0.4.15) was designed to highlight the main target genes for the predicted set of ligands. Ligands were assigned to senders by computing ligand specificity. The ligands that were expressed in more than one sender were labelled as common ligands.

We performed gene set enrichment analysis of the predicted ligands in the cellular neighbourhood colocalising with iREC cancer cell state using GSEApy (v0.10.8), and 'MSigDB_Hallmark_2020' and 'KEGG_2021_Human' gene sets.

We computed spatial enrichment of potential ligands predicted to influence the iREC phenotype in the cellular neighbourhood $C_k$ containing iREC using the spatial gene expression of liver metastatic 10X Visium samples. The spatial neighbourhood analysis was carried out on all liver metastatic samples together based on the cell type abundance, while spatial ligand enrichment was computed for each liver metastatic sample separately. We log-normalised the raw counts using a scale factor of 10,000. We implemented the ligand spatial enrichment of ligand *t* in the cellular neighbourhood $C_k$ (odds ratio) as the ratio of the odds of the ligand *t* being expressed by the odds of the other genes being expressed. The odds of ligand *t* being expressed in spots belonging to $C_k$ were calculated by dividing the number of spots in $C_k$

expressing ligand *t* by the number of spots that are not part of $C_k$ expressing ligand *t*. For each ligand *t*, a 2-by-2 contingency table (2 regions: part of $C_k$ and not part of $C_k$; 2 categories: ligand *t* and other genes) was built. Significance was assessed with a Chi-square test using scipy.stats.chi2_contingency (SciPy v1.8.1)[102] and by adjusting p-values for multiple testing with Benjamini-Hochberg correction method using statsmodels.stats.multitest.multipletests (statsmodels v0.13.2). A ligand *t* was considered statistically significantly enriched in neighbourhood $C_k$ with adjusted p-value lower than 0.05 and with a positive log odds ratio. We computed the 95% confidence interval for the log odds ratio.

## Analysis of cancer cell state signatures in TCGA CRC bulk tumours

TCGA bulk RNA-seq count data was downloaded for colon and rectal cancers using the TCGAbiolinks R package[103]. Count data was then normalised using variance stabilising transformation (VST) using DESeq2[104] and VST counts were z-scored. Samples from patients with multiple samples in the dataset were removed.

Primary cancer cell state signatures were obtained using Seurat's *FindAllMarkers* function, on 'cell subtype' level annotations for all cells in the primary CRC dataset (e.g. cell states/subtypes for epithelial, myeloid cells…). The top 50 differentially expressed genes ranked by $\log_2$ fold change were then used as a gene signature, and each tumour was scored for expression of the cancer cell state signature by calculating the mean of z-scored VST counts for genes in the signature. For ligands expressed by tumour microenvironment cells, each tumour was scored for expression of the ligands in the same manner. Scores for cancer state signatures and ligand expression values were then correlated using the Pearson method. To investigate cancer state gene signature expression in MSS/MSI subtypes, only TCGA bulk RNA-seq samples with MSS/MSI annotations were used.

To investigate differential abundance of proteins and phospho-peptides, we accessed TCGA reverse phase protein array (RPPA) data using cBioPortal and compared the top tertile of bulk TCGA tumours by expression of REC or iREC signature to the bottom tertile.

## Gene enrichment analysis

We performed gene enrichment analysis (GEA) on differentially expressed genes using GSEApy (v0.10.8)[101].

## ChIP-seq analysis

ChIP-seq data for JUND (GSE32465)[105] and HNF4A (GSE49402)[106] generated from HCT-116 and LoVo CRC cell lines respectively were downloaded from the NCBI sequence read archive. Reads were trimmed using trimmomatics (*LEADING:5 TRAILING:5 SLIDINGWINDOW:4:15 MINLEN:20*)[107] and mapped to hg38 (GRCh38, refdata-cellranger-arc-GRCh38-2020-A-2.0.0) using bowtie2[108]. Samtools[109] was used to filter reads, keeping high quality (q30) reads, mapping to known chromosomes. Reads aligned to blacklist regions were removed using bedtools *intersect*, and duplicates were marked using Picard *MarkDuplicates*.

Peaks were called using MACS2 *callpeak* using input or IgG controls and the following parameters: *-q 0.01 -g hs -f AUTO -B --SPMR --call-summits*. The summit file was extended +/- 250bp using bedtools *slop*. Bedgraphs were converted into bigwig files for visualisation using UCSC tools *bedGraphToBigWig*. Deeptools[110] *computeMatrix* and *plotHeatmap* were used to generate heatmaps showing TF ChIP-seq signal at chromatin regions identified in SCENIC+ JUND and HNF4A regulons.

## Data availability

Sequencing data generated in this study have been deposited at ArrayExpress under the accession numbers: E-MTAB-13651, E-MTAB-13652 and E-MTAB-13655. Processed count matrices and metadata will be made available at ArrayExpress.

## Code availability

Code will be made available at github.

## Members of the Cancer Tissue Bank

Claude Chelala, Dayem Ullah, Jo-Anne ChinAleong, Amina Saad

# Acknowledgments

This work was supported by Bart Charity Lectureship (grant MGU045), Cancer Research UK Career Establishment Award (RCCCEA/100003) and Cancer Research UK City of London Award (C7893/A26233). Cancer Tissue Bank is supported by Cancer Research UK Centre of Excellence award to Barts Cancer Institute, Queen Mary University of London (C/A). Researchers are grateful to many staff members of Cancer Tissue Bank (in particular to Mr Rory Smith, Ms Sarah Mueller) and Consultant Surgeons (Ajit Abraham, Deepak Hariharan, Vincent Yip) for their contribution to sample and data collection. We are very grateful to the UCL Single Cell Genomics facility, the Barts Cancer Institute (BCI) Pathology core facility and the BCI Microscopy core facility. Work at the CRUK City of London Centre Single Cell Genomics Facility and Cancer Institute Genomics Translational Technology Platform was supported by the CRUK City of London Centre Award [C7893/A26233]. This research utilised Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT http://doi.org/10.5281/zenodo.438045.

# Declaration of Interests

The authors declare no competing interests.

# Figure legends

**Figure 1. Integrated scRNA-seq analysis reveals heterogeneous cancer cell states in pCRC.**

**a.** Experimental outline and overview of the computational workflow. **b.** UMAP representation of cell types present in pCRC. **c.** UMAP representation of malignant pCRC cell states. **d.** Proportions of cancer cell states present in pCRC datasets. **e.** Heatmap showing GEA of differentially expressed genes (DEGs) in cancer cell states for the indicated signatures. Interferon/MHC-II, hypoxia, and EMT-II[100], CRIS[24], EpiHR and coreHRC[25], revCSC[23], RSC[111], fetal[112], YAP[113], pEMT[114], CMS2 and CMS3[115], iCMS2 and iCMS3[28], all other signatures are from MSigDB Hallmarks. Genes in each signature are listed in Supplementary Table 2. **f.** Dotplot showing the scaled mRNA expression levels of the indicated marker genes, interferon stimulated genes (ISGs) and differentiation marker *KRT20* expression in pCRC cell states. The size of the dot indicates the fraction of cells in each state in which expression of the gene is detected.

**Figure 2. Spatial neighbourhoods surrounding cancer cell states in pCRC**

**a.** H&E staining of primary colorectal cancer sample A1, manual annotations and clustering annotations based on spatial gene expression. **b.** H&E staining of primary colorectal cancer sample C1, manual annotations and clustering annotations based on spatial gene expression. **c.** Gene enrichment analysis of upregulated genes in the invasive edge and the tumour core. **d.** Abundance of cancer cell states across spatial locations of representative samples A1 and C1 capturing the tumour core and invasive edge. Cell abundance (colour represents intensity, size represents score) per spot assigned to each cell type and cancer cell state is estimated by cell2location. **e.** Identification of spatial neighbourhoods shown in representative samples A1 and C1. Spatial cellular neighbourhoods were deciphered by joint modelling of four Visium samples to infer common patterns across samples using SpatialDE2. **f.** Dotplot representing the average cell abundance (dot size and colour) for each cell state, per neighbourhood, and normalised between 0 and 1 per cell state. **g.** Estimated cell type abundances for distinct stromal cells and immune subpopulations colocalising with REC and iREC cancer cell states.

**Figure 3. Cancer cell states are re-established in liver mCRC.**

**a.** UMAP representation of cell types in mCRC Multiome (paired snRNA-seq + snATAC-seq) data, based upon the RNA modality. **b.** UMAP representation of cancer cells in mCRC Multiome data, based upon the RNA modality. **c.** Boxplot showing the proportion of cancer cell states in each mCRC patient sample. **d.** Heatmap showing GEA of differentially expressed genes in mCRC cell states for the indicated signatures. Genes in the signatures are listed in Supplementary Table 2. **e.** Dotplot showing the scaled mRNA expression levels of the indicated marker genes, ISGs and differentiation marker *KRT20* in mCRC cell states.

**Figure 4. Regulation of cancer cell states.**

**a.** Putative enhancer-gene (PE-GL) linkages (n = 1444) in mCRC cell states. Left heatmap shows chromatin accessibility of putative enhancers. Right heatmap shows mRNA expression of genes linked to putative enhancers. K-means clustering was performed on the chromatin accessibility data. Selected transcription factor (TF) *de novo* motifs enriched in accessible regions of each k-means cluster were identified using Homer (Supplementary Table 7) and are shown on the left. Motifs are annotated using the top annotation from Homer. **b.** Genome browser view of chromatin accessibility at the EMP1 locus in the indicated mCRC cell states. PE-GL linkages are shown and chromHMM enhancers[47]. **c.** GEA of genes in k-means clusters shown in Fig. 5a using KEGG pathways and MSigDB Hallmarks. **d.** Heatmap showing chromVAR motif deviation z-scores for TFs in mCRC cell states. Only statistically significant motifs are shown (Wilcoxon, FDR < 0.05). TFs are annotated based on the DNA binding

domain (DBD)[116]. **e.** Accessibility of chromatin regions in the indicated SCENIC+ regulons. **f.** Z-scored mRNA expression of genes in the indicated SCENIC+ regulons across mCRC cell states. **g.** Z-scored mRNA expression levels of AP-1 target genes[61] across mCRC cell states.

**Figure 5. Spatial niches of cancer cell states in liver metastasis.**

**a.** Desmoplastic growth pattern: H&E staining of publicly available liver metastatic sample LM4, manual annotations and clustering annotations based on spatial gene expression. Sample LM4 captures the desmoplastic rim separating the tumour and the liver parenchyma. The dashed line in black denotes the desmoplastic rim. **b.** Replacement growth pattern: H&E staining of publicly available liver metastatic colorectal cancer sample P13, manual annotations and clustering annotations based on spatial gene expression. In the replacement growth pattern, tumour cells are in direct contact with hepatocytes. The dashed line in black denotes the tumour-liver border. **c.** H&E staining of liver metastatic colorectal cancer sample CRC11. **d.** Abundance of cancer cell states across spatial locations of three representative samples (LM4, P13, CRC11). Cell abundance (colour represents intensity, size represents the score) per spot assigned to each cell type and cancer cell state is estimated by cell2location. **e.** Estimated cell type abundances for distinct stromal subpopulations in LM4, P13 and CRC11. **f.** Spatial cellular neighbourhoods in LM4, P13 and CRC11. Spatial cellular neighbourhoods were deciphered by joint modelling of six Visium samples to infer common patterns across samples using SpatialDE2. **g.** Estimated cell type abundances for distinct immune subpopulations in the immunosuppressive niche in representative samples LM4, P13 and CRC11. **h.** Dotplot representing the average cell abundance (dot size and colour) for each cell state, per neighbourhood, and normalised between 0 and 1 per cell state. 0 and 5 are the cellular neighbourhoods containing iREC. Niche 7 is the cellular neighbourhood containing REC. Niche 4 is the cellular neighbourhood containing Stem NOTUM. Cellular neighbourhoods 1 and 9 denote the liver parenchyma, depending on the growth pattern. **i.** GEA of upregulated genes in the spatial cellular neighbourhoods.

**Figure 6. Spatially resolved cell-cell interactions in the cellular neighbourhood surrounding iRECs.**

**a.** Heatmaps summarising the inferred spatial cell-cell interactions mediated by stromal and myeloid cells in the cellular neighbourhood containing iRECs with iRECs as the receiver, using CellPhoneDB and NicheNet. Specifically, we identified potential upstream ligand-receptor pairs which can induce the AP-1 regulon program in the neighbouring pro-metastatic phenotype. Z-score of the gene expression of selected potential ligands in each cell type of the cellular neighbourhood (top panel) and z-score of gene expression of corresponding receptors in cancer cell states (bottom panel). In both heatmaps, the x-axis denotes ligand-receptor interactions, with the ligand in bold and receptor in grey for sender cells (top panel) and the ligand in grey and the receptor in bold for cancer cell states (bottom panel). **b.** GEA of ligands predicted to activate genes in the AP-1 regulon in iRECs. **c.** Circos plot depicting links between predicted ligands from stromal and myeloid cells and target genes of the AP-1 regulon, as inferred by NicheNet. Links denote the regulatory potential scores between ligands and target genes of the AP-1 regulon, as predicted by NicheNet. **d.** Correlation in TCGA bulk CRC RNA-seq data of the expression of the indicated pCRC cancer cell state signatures and ligands expressed in TME subpopulations that potentially drive the expression of AP-1 and RELB regulons in CRC malignant cells. Ligands are shown in Fig. 6a (AP-1) and Extended Data Fig. 9a (RELB). Common AP-1 and common RELB are ligands whose expression is shared in more than one TME subpopulation.

# References


1. Xi, Y. & Xu, P. Global colorectal cancer burden in 2020 and projections to 2040. *Transl. Oncol.* **14**, 101174 (2021).
2. Martínez-Jiménez, F. *et al.* Pan-cancer whole-genome comparison of primary and metastatic solid tumours. *Nature* **618**, 333–341 (2023).
3. Makohon-Moore, A. P. *et al.* Limited heterogeneity of known driver gene mutations among the metastases of individual patients with pancreatic cancer. *Nat. Genet.* **49**, 358–366 (2017).
4. Ryser, M. D. *et al.* Minimal barriers to invasion during human colorectal tumor growth. *Nat. Commun.* **11**, 1280 (2020).
5. Gupta, P. B., Pastushenko, I., Skibinski, A., Blanpain, C. & Kuperwasser, C. Phenotypic Plasticity: Driver of Cancer Initiation, Progression, and Therapy Resistance. *Cell Stem Cell* **24**, 65–78 (2019).
6. Pérez-González, A., Bévant, K. & Blanpain, C. Cancer cell plasticity during tumor progression, metastasis and response to therapy. *Nat Cancer* **4**, 1063–1082 (2023).
7. Househam, J. *et al.* Phenotypic plasticity and genetic control in colorectal cancer evolution. *Nature* **611**, 744–753 (2022).
8. Dalerba, P. *et al.* Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat. Biotechnol.* **29**, 1120–1127 (2011).
9. Barker, N. *et al.* Crypt stem cells as the cells-of-origin of intestinal cancer. *Nature* **457**, 608–611 (2009).
10. Zhou, J. & Boutros, M. Intestinal stem cells and their niches in homeostasis and disease. *Cells Dev* **175**, 203862 (2023).
11. Tian, H. *et al.* A reserve stem cell population in small intestine renders Lgr5-positive cells dispensable. *Nature* **478**, 255–259 (2011).
12. Fumagalli, A. *et al.* Plasticity of Lgr5-Negative Cancer Cells Drives Metastasis in Colorectal Cancer. *Cell Stem Cell* **26**, 569–578.e7 (2020).
13. Heinz, M. C. *et al.* Liver Colonization by Colorectal Cancer Metastases Requires YAP-Controlled Plasticity at the Micrometastatic StageCellular Determinants of Metastatic Outgrowth. *Cancer Res.* OF1–OF16 (2022).
14. Kobayashi, S. *et al.* LGR5-positive colon cancer stem cells interconvert with drug-resistant LGR5-negative cells and are capable of tumor reconstitution. *Stem Cells* **30**, 2631–2644 (2012).
15. Posey, T. A. *et al.* Loss of LGR5 through Therapy-induced Downregulation or Gene Ablation Is Associated with Resistance and Enhanced MET-STAT3 Signaling in Colorectal Cancer Cells. *Mol. Cancer Ther.* **22**, 667–678 (2023).
16. Moorman, A. R. *et al.* Progressive plasticity during colorectal cancer metastasis. *bioRxiv* (2023) doi:10.1101/2023.08.18.553925.
17. Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer. *Science* **357**, (2017).
18. Lee, H.-O. *et al.* Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nat. Genet.* **52**, 594–603 (2020).
19. Pelka, K. *et al.* Spatially organized multicellular immune hubs in human colorectal cancer. *Cell* **184**, 4734–4752.e20 (2021).
20. Che, L.-H. *et al.* A single-cell atlas of liver metastases of colorectal cancer reveals reprogramming of the tumor microenvironment in response to preoperative

chemotherapy. *Cell Discov* **7**, 80 (2021).
21. Flanagan, D. J. *et al.* NOTUM from Apc-mutant cells biases clonal competition to initiate cancer. *Nature* **594**, 430–435 (2021).
22. Gil Vazquez, E. *et al.* Dynamic and adaptive cancer stem cell population admixture in colorectal neoplasia. *Cell Stem Cell* **29**, 1612 (2022).
23. Qin, X. *et al.* An oncogenic phenoscape of colonic stem cell polarization. *Cell* **186**, 5554–5568.e18 (2023).
24. Isella, C. *et al.* Selective analysis of cancer-cell intrinsic transcriptional traits defines novel clinically relevant subtypes of colorectal cancer. *Nat. Commun.* **8**, 1–16 (2017).
25. Cañellas-Socias, A. *et al.* Metastatic recurrence in colorectal cancer arises from residual EMP1 cells. *Nature* **611**, 603–613 (2022).
26. Gocher, A. M., Workman, C. J. & Vignali, D. A. A. Interferon-γ: teammate or opponent in the tumour microenvironment? *Nat. Rev. Immunol.* **22**, 158–172 (2022).
27. Barkley, D. *et al.* Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. *Nat. Genet.* **54**, 1192–1201 (2022).
28. 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).
29. Wu, Y. *et al.* Spatiotemporal Immune Landscape of Colorectal Cancer Liver Metastasis at Single-Cell LevelSpatial and Cellular Landscape of CRLM. *Cancer Discov.* (2022).
30. Davidson, S. *et al.* Single-Cell RNA Sequencing Reveals a Dynamic Stromal Niche That Supports Tumor Growth. *Cell Rep.* **31**, 107628 (2020).
31. Elyada, E. *et al.* Cross-Species Single-Cell Analysis of Pancreatic Ductal Adenocarcinoma Reveals Antigen-Presenting Cancer-Associated Fibroblasts. *Cancer Discov.* **9**, 1102–1123 (2019).
32. McCarthy, N. *et al.* Distinct Mesenchymal Cell Populations Generate the Essential Intestinal BMP Signaling Gradient. *Cell Stem Cell* **26**, 391–402.e5 (2020).
33. Harnack, C. *et al.* R-spondin 3 promotes stem cell recovery and epithelial regeneration in the colon. *Nat. Commun.* **10**, 1–15 (2019).
34. Kloosterman, D. J. & Akkari, L. Macrophages at the interface of the co-evolving cancer ecosystem. *Cell* **186**, 1627–1651 (2023).
35. Nasir, I. *et al.* Tumor macrophage functional heterogeneity can inform the development of novel cancer therapies. *Trends Immunol.* **44**, 971–985 (2023).
36. Mulder, K. *et al.* Cross-tissue single-cell landscape of human monocytes and macrophages in health and disease. *Immunity* **54**, 1883–1900.e5 (2021).
37. Ma, R.-Y., Black, A. & Qian, B.-Z. Macrophage diversity in cancer revisited in the era of single-cell omics. *Trends Immunol.* **43**, 546–563 (2022).
38. Li, R. *et al.* Mapping single-cell transcriptomes in the intra-tumoral and associated territories of kidney cancer. *Cancer Cell* **40**, 1583–1599.e10 (2022).
39. Chakarov, S. *et al.* Two distinct interstitial macrophage populations coexist across tissues in specific subtissular niches. *Science* **363**, (2019).
40. Liu, J., Zhang, X., Cheng, Y. & Cao, X. Dendritic cell migration in inflammation and immunity. *Cell. Mol. Immunol.* **18**, 2461–2471 (2021).
41. Ozato, Y. *et al.* Spatial and single-cell transcriptomics decipher the cellular environment containing HLA-G+ cancer cells and SPP1+ macrophages in colorectal cancer. *Cell Rep.* **42**, 111929 (2023).
42. Kleshchevnikov, V. *et al.* Cell2location maps fine-grained cell types in spatial transcriptomics. *Nat. Biotechnol.* **40**, 661–671 (2022).

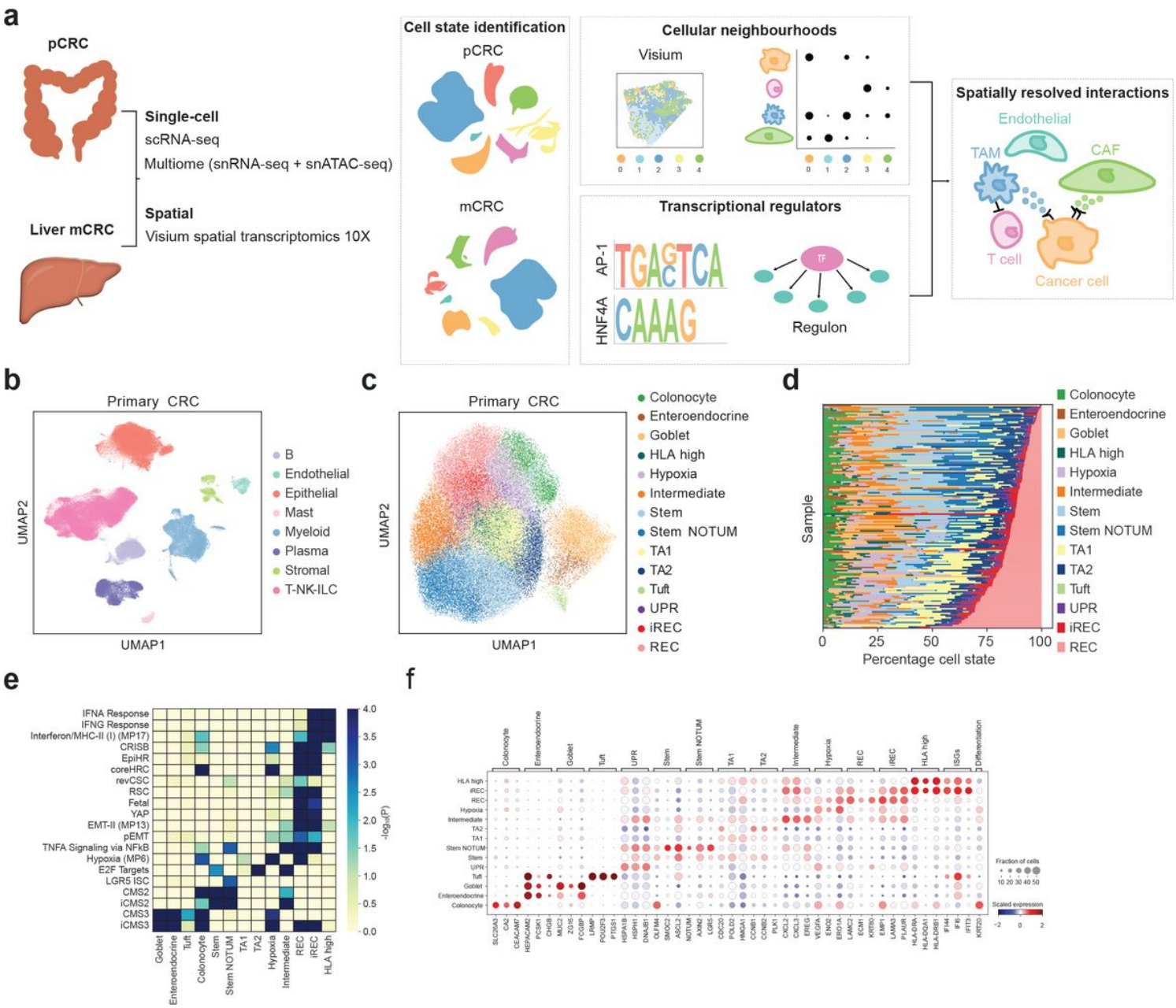
43. Kats, I., Vento-Tormo, R. & Stegle, O. SpatialDE2: Fast and localized variance component analysis of spatial transcriptomics. *bioRxiv* 2021.10.27.466045 (2021) doi:10.1101/2021.10.27.466045.
44. Becker, W. R. *et al.* Single-cell analyses define a continuum of cell state and composition changes in the malignant transformation of polyps to colorectal cancer. *Nat. Genet.* **54**, 985–995 (2022).
45. Wang, F. *et al.* Single-cell and spatial transcriptome analysis reveals the cellular heterogeneity of liver metastatic colorectal cancer. *Sci Adv* **9**, eadf5464 (2023).
46. Sathe, A. *et al.* Colorectal Cancer Metastases in the Liver Establish Immunosuppressive Spatial Networking between Tumor-Associated SPP1+ Macrophages and Fibroblasts. *Clin. Cancer Res.* **29**, 244–260 (2023).
47. Della Chiara, G. *et al.* Epigenomic landscape of human colorectal cancer unveils an aberrant core of pan-cancer enhancers orchestrated by YAP/TAZ. *Nat. Commun.* **12**, 2340 (2021).
48. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
49. Chen, L. *et al.* A reinforcing HNF4–SMAD4 feed-forward module stabilizes enterocyte identity. *Nat. Genet.* **51**, 777–785 (2019).
50. Huang, Y.-H. *et al.* POU2F3 is a master regulator of a tuft cell-like variant of small cell lung cancer. *Genes Dev.* **32**, 915–928 (2018).
51. Wu, X. S. *et al.* OCA-T1 and OCA-T2 are coactivators of POU2F3 in the tuft cell lineage. *Nature* **607**, 169–175 (2022).
52. Gehart, H. *et al.* Identification of Enteroendocrine Regulators by Real-Time Single-Cell Differentiation Mapping. *Cell* **176**, 1158–1173.e16 (2019).
53. Freund, J.-N., Duluc, I., Reimund, J.-M., Gross, I. & Domon-Dell, C. Extending the functions of the homeotic transcription factor Cdx2 in the digestive system through nontranscriptional activities. *World J. Gastroenterol.* **21**, 1436–1443 (2015).
54. Yang, S.-H., Sharrocks, A. D. & Whitmarsh, A. J. MAP kinase signalling cascades and transcriptional regulation. *Gene* **513**, 1–13 (2013).
55. MacDonald, B. T., Tamai, K. & He, X. Wnt/beta-catenin signaling: components, mechanisms, and diseases. *Dev. Cell* **17**, 9–26 (2009).
56. Bravo González-Blas, C. *et al.* SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat. Methods* **20**, 1355–1367 (2023).
57. Zanconato, F. *et al.* Genome-wide association between YAP/TAZ/TEAD and AP-1 at enhancers drives oncogenic growth. *Nat. Cell Biol.* **17**, 1218–1227 (2015).
58. Park, J. *et al.* YAP and AP-1 Cooperate to Initiate Pancreatic Cancer Development from Ductal Cells in Mice. *Cancer Res.* **80**, 4768–4779 (2020).
59. He, L. *et al.* YAP and TAZ are transcriptional co-activators of AP-1 proteins and STAT3 during breast cellular transformation. *Elife* **10**, (2021).
60. Taniguchi, K. & Karin, M. NF-κB, inflammation, immunity and cancer: coming of age. *Nat. Rev. Immunol.* **18**, 309–324 (2018).
61. Ogden, S. *et al.* Oncogenic ERRB2 signals through the AP-1 transcription factor to control mesenchymal-like properties of oesophageal adenocarcinoma. *NAR Cancer* **5**, zcad001 (2023).
62. Wu, Y. *et al.* Spatiotemporal Immune Landscape of Colorectal Cancer Liver Metastasis at Single-Cell Level. *Cancer Discov.* **12**, 134–153 (2022).
63. Fernández Moro, C. *et al.* An idiosyncratic zonated stroma encapsulates desmoplastic

liver metastases and originates from injured liver. *Nat. Commun.* **14**, 5024 (2023).
64. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat. Protoc.* **15**, 1484–1506 (2020).
65. Browaeys, R., Saelens, W. & Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods* **17**, 159–162 (2019).
66. Derynck, R., Turley, S. J. & Akhurst, R. J. TGFβ biology in cancer progression and immunotherapy. *Nat. Rev. Clin. Oncol.* **18**, 9–34 (2021).
67. Lau, E. Y. T. *et al.* Cancer-Associated Fibroblasts Regulate Tumor-Initiating Cell Plasticity in Hepatocellular Carcinoma through c-Met/FRA1/HEY1 Signaling. *Cell Rep.* **15**, 1175–1189 (2016).
68. Xie, Y. *et al.* FGF/FGFR signaling in health and disease. *Signal Transduct Target Ther* **5**, 181 (2020).
69. Goel, H. L. & Mercurio, A. M. VEGF targets the tumour cell. *Nat. Rev. Cancer* **13**, 871–882 (2013).
70. Lee, J. W. *et al.* Hepatocytes direct the formation of a pro-metastatic niche in the liver. *Nature* **567**, 249–252 (2019).
71. Saw, P. E., Chen, J. & Song, E. Targeting CAFs to overcome anticancer therapeutic resistance. *Trends Cancer Res.* **8**, 527–555 (2022).
72. Vijayan, D., Young, A., Teng, M. W. L. & Smyth, M. J. Targeting immunosuppressive adenosine in cancer. *Nat. Rev. Cancer* **17**, 765 (2017).
73. Andersson, P. *et al.* Molecular mechanisms of IL-33-mediated stromal interactions in cancer metastasis. *JCI Insight* **3**, (2018).
74. Kemp, S. B. *et al.* Apolipoprotein E Promotes Immune Suppression in Pancreatic Cancer through NF-κB-Mediated Production of CXCL1. *Cancer Res.* **81**, 4305–4318 (2021).
75. Araújo, T. G. *et al.* Annexin A1 as a Regulator of Immune Response in Cancer. *Cells* **10**, (2021).
76. Onozawa, H. *et al.* Annexin A1 is involved in resistance to 5-FU in colon cancer cells. *Oncol. Rep.* **37**, 235–240 (2017).
77. Ozga, A. J., Chow, M. T. & Luster, A. D. Chemokines and the immune response to cancer. *Immunity* **54**, 859–874 (2021).
78. Pikkupeura, L. M. *et al.* Transcriptional and epigenomic profiling identifies YAP signaling as a key regulator of intestinal epithelium maturation. *Sci Adv* **9**, eadf9460 (2023).
79. van Dam, P.-J. *et al.* Histopathological growth patterns as a candidate biomarker for immunomodulatory therapy. *Semin. Cancer Biol.* **52**, 86–93 (2018).
80. Höppener, D. J. *et al.* Enrichment of the tumour immune microenvironment in patients with desmoplastic colorectal liver metastasis. *Br. J. Cancer* **123**, 196–206 (2020).
81. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
82. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst* **8**, 281–291.e9 (2019).
83. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
84. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
85. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

86. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
87. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
88. Yang, S. *et al.* Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol.* **21**, 57 (2020).
89. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).
90. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
91. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
92. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
93. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38**, 576–589 (2010).
94. Kumar, N., Mishra, B., Athar, M. & Mukhtar, S. Inference of Gene Regulatory Network from Single-Cell Transcriptomic Data Using pySCENIC. *Methods Mol. Biol.* **2328**, 171–182 (2021).
95. Elmentaite, R. *et al.* Cells of the human intestinal tract mapped across space and time. *Nature* **597**, 250–255 (2021).
96. Smillie, C. S. *et al.* Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. *Cell* **178**, 714–730.e22 (2019).
97. Hickey, J. W. *et al.* Organization of the human intestine at single-cell resolution. *Nature* **619**, 572–584 (2023).
98. Sheng, C. *et al.* Probabilistic machine learning ensures accurate ambient denoising in droplet-based single-cell omics. *bioRxiv* 2022.01.14.476312 (2022) doi:10.1101/2022.01.14.476312.
99. Palla, G. *et al.* Squidpy: a scalable framework for spatial omics analysis. *Nat. Methods* **19**, 171–178 (2022).
100. Gavish, A. *et al.* Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours. *Nature* **618**, 598–606 (2023).
101. Fang, Z. *GSEApy: Gene Set Enrichment Analysis in Python*. (Github).
102. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
103. Colaprico, A. *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **44**, e71 (2016).
104. 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).
105. Gertz, J. *et al.* Distinct Properties of Cell-Type-Specific and Shared Transcription Factor Binding Sites. *Mol. Cell* **52**, 25–36 (2013).
106. Yan, J. *et al.* Transcription Factor Binding in Human Cells Occurs in Dense Clusters Formed around Cohesin Anchor Sites. *Cell* **154**, 801–813 (2013).
107. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
108. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
109. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,

2078–2079 (2009).
110. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–91 (2014).
111. Vasquez, E. G. *et al.* Dynamic and adaptive cancer stem cell population admixture in colorectal neoplasia. *Cell Stem Cell* **29**, 1213–1228.e8 (2022).
112. Mustata, R. C. *et al.* Identification of Lgr5-Independent Spheroid-Generating Progenitors of the Mouse Fetal Intestinal Epithelium. *Cell Rep.* **5**, 421–432 (2013).
113. Serra, D. *et al.* Self-organization and symmetry breaking in intestinal organoid development. *Nature* **569**, 66–72 (2019).
114. Tyler, M. & Tirosh, I. Decoupling epithelial-mesenchymal transitions from stromal profiles by integrative expression analysis. *Nat. Commun.* **12**, 1–13 (2021).
115. Sveen, A. *et al.* Colorectal Cancer Consensus Molecular Subtypes Translated to Preclinical Models Uncover Potentially Targetable Cancer Cell Dependencies. *Clin. Cancer Res.* **24**, 794–806 (2018).
116. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).
117. Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D. & Marioni, J. C. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nat. Biotechnol.* **40**, 245–253 (2021).

# Figures

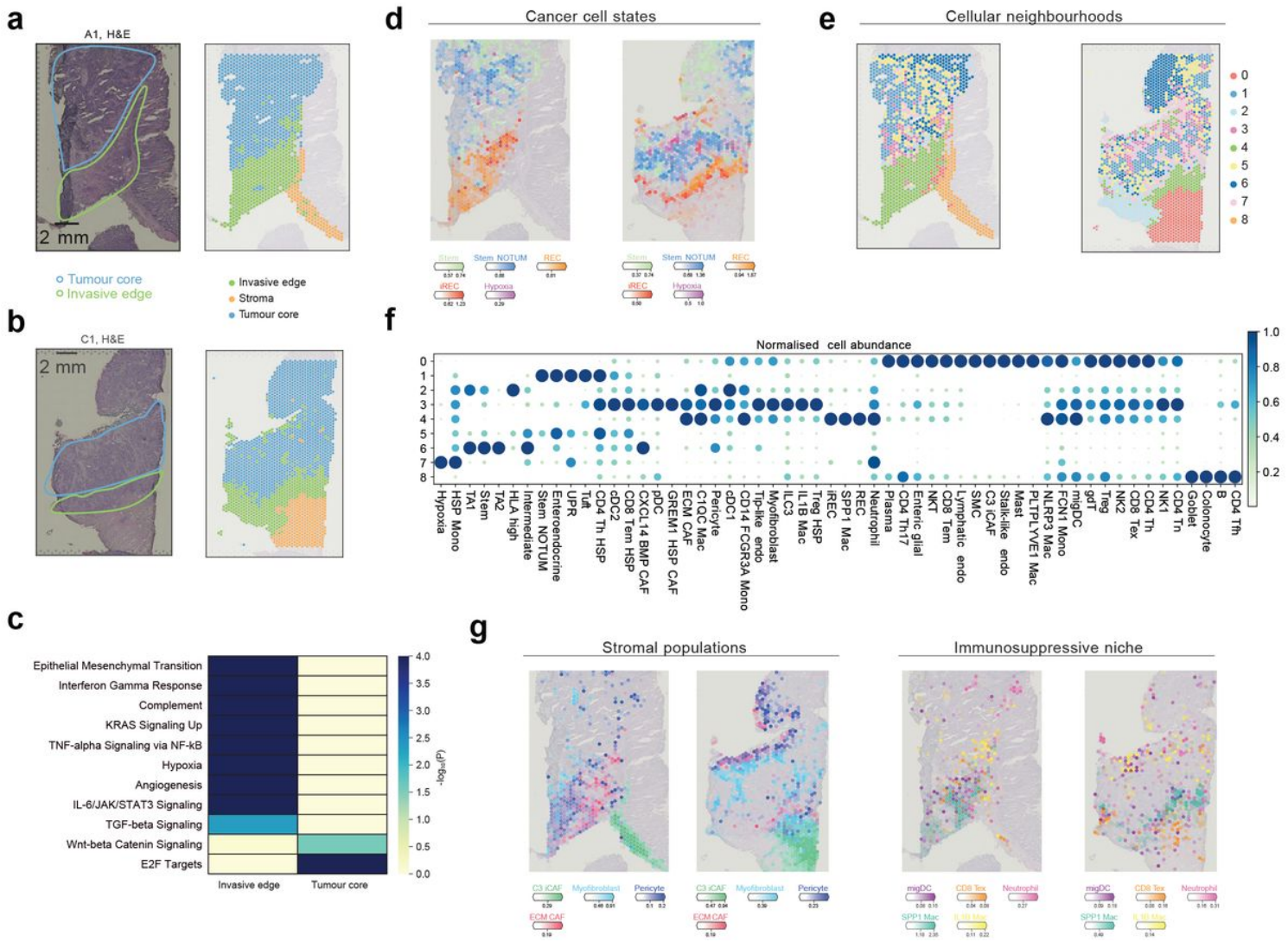



## Figure 1

Integrated scRNA-seq analysis reveals heterogeneous cancer cell states in pCRC.

a. Experimental outline and overview of the computational workflow. b. UMAP representation of cell types present in pCRC. c. UMAP representation of malignant pCRC cell states. d. Proportions of cancer cell states present in pCRC datasets. e. Heatmap showing GEA of differentially expressed genes (DEGs) in cancer cell states for the indicated signatures. Interferon/MHC-II, hypoxia, and EMT-II100, CRIS24, EpiHR and coreHRC25, revCSC23, RSC111, fetal112, YAP113, pEMT114, CMS2 and CMS3115, iCMS2 and iCMS328, all other signatures are from MSigDB Hallmarks. Genes in each signature are listed in Supplementary Table 2. f. Dotplot showing the scaled mRNA expression levels of the indicated marker

genes, interferon stimulated genes (ISGs) and differentiation marker KRT20 expression in pCRC cell states. The size of the dot indicates the fraction of cells in each state in which expression of the gene is detected.

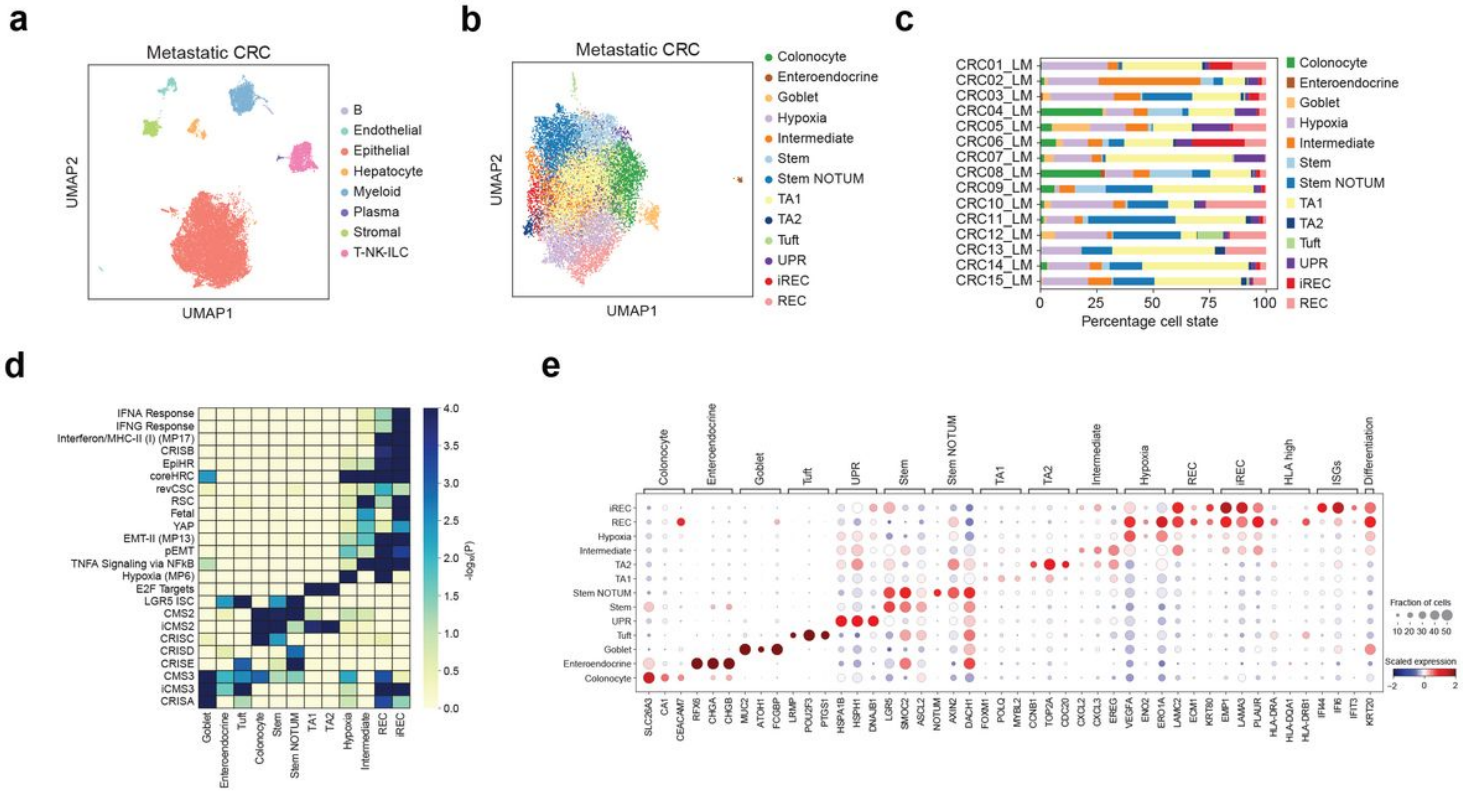



## Figure 2

Spatial neighbourhoods surrounding cancer cell states in pCRC

a. H&E staining of primary colorectal cancer sample A1, manual annotations and clustering annotations based on spatial gene expression. b. H&E staining of primary colorectal cancer sample C1, manual annotations and clustering annotations based on spatial gene expression. c. Gene enrichment analysis of upregulated genes in the invasive edge and the tumour core. d. Abundance of cancer cell states across spatial locations of representative samples A1 and C1 capturing the tumour core and invasive edge. Cell abundance (colour represents intensity, size represents score) per spot assigned to each cell type and cancer cell state is estimated by cell2location. e. Identification of spatial neighbourhoods shown in representative samples A1 and C1. Spatial cellular neighbourhoods were deciphered by joint modelling of four Visium samples to infer common patterns across samples using SpatialDE2. f. Dotplot representing

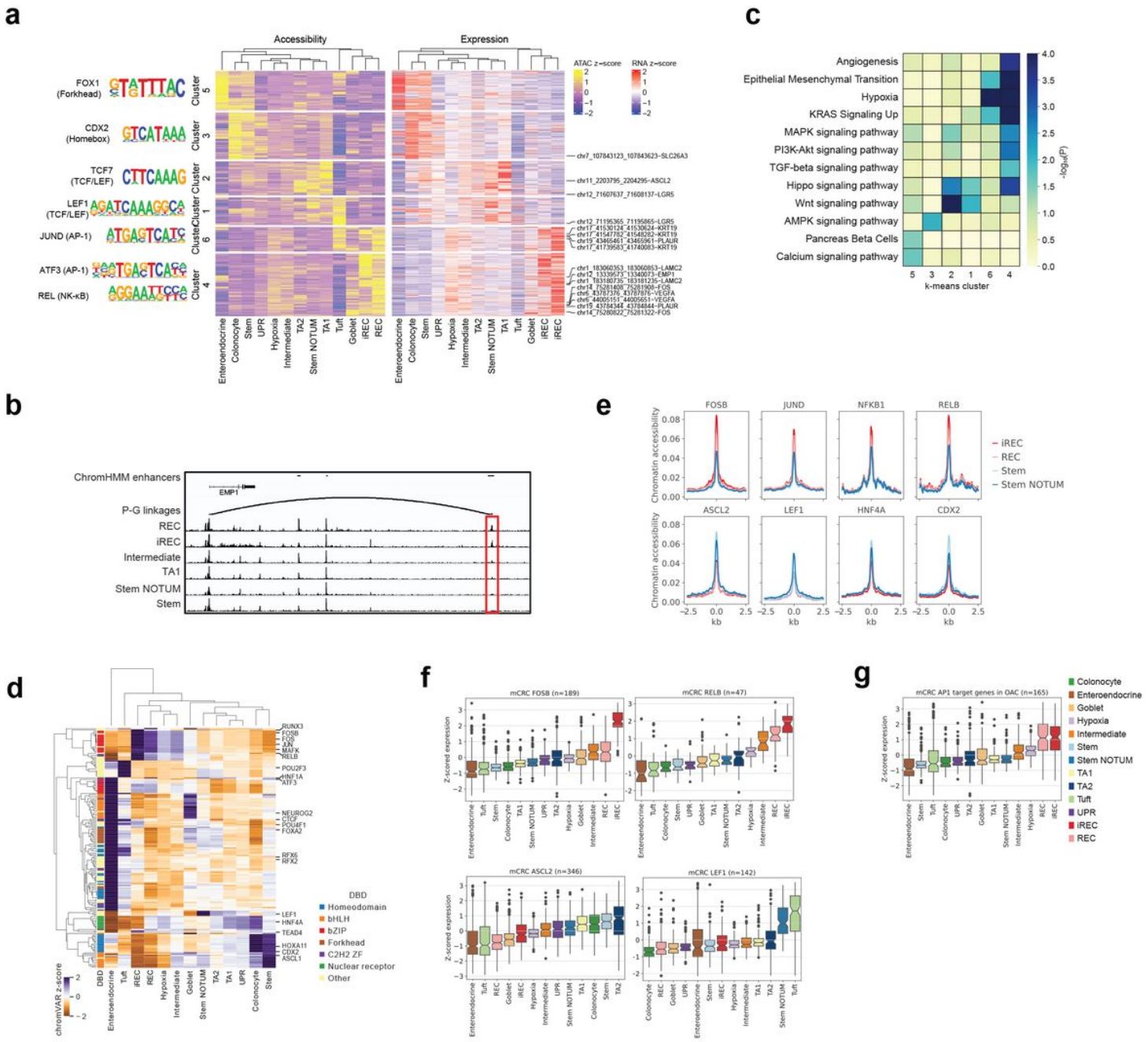
the average cell abundance (dot size and colour) for each cell state, per neighbourhood, and normalised between 0 and 1 per cell state. g. Estimated cell type abundances for distinct stromal cells and immune subpopulations colocalising with REC and iREC cancer cell states.




## Figure 3

Cancer cell states are re-established in liver mCRC.

a. UMAP representation of cell types in mCRC Multiome (paired snRNA-seq + snATAC-seq) data, based upon the RNA modality. b. UMAP representation of cancer cells in mCRC Multiome data, based upon the RNA modality. c. Boxplot showing the proportion of cancer cell states in each mCRC patient sample. d. Heatmap showing GEA of differentially expressed genes in mCRC cell states for the indicated signatures. Genes in the signatures are listed in Supplementary Table 2. e. Dotplot showing the scaled mRNA expression levels of the indicated marker genes, ISGs and differentiation marker KRT20 in mCRC cell states.

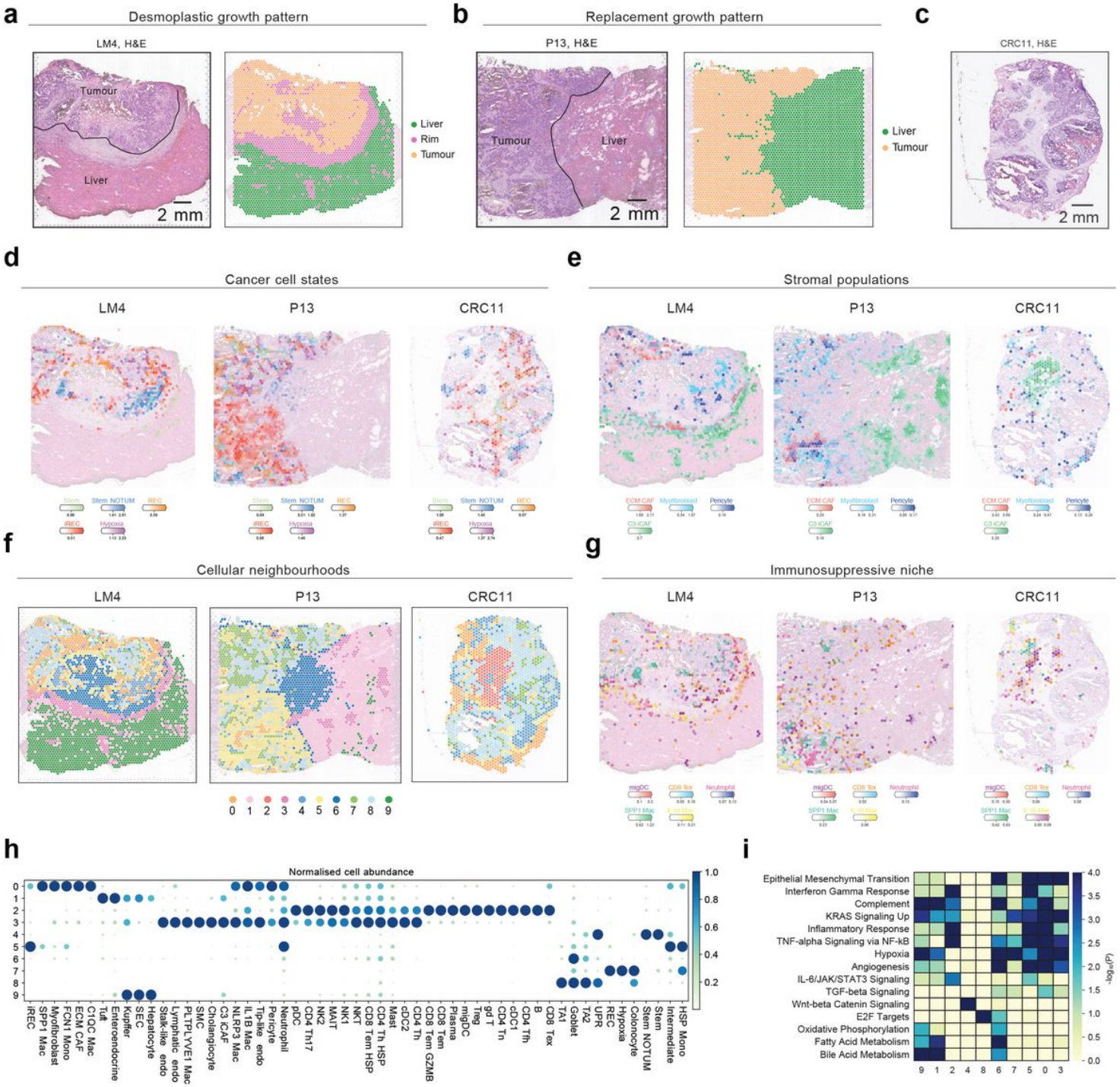



## Figure 4

Regulation of cancer cell states.

a. Putative enhancer-gene (PE-GL) linkages (n = 1444) in mCRC cell states. Left heatmap shows chromatin accessibility of putative enhancers. Right heatmap shows mRNA expression of genes linked to putative enhancers. K-means clustering was performed on the chromatin accessibility data. Selected transcription factor (TF) de novo motifs enriched in accessible regions of each k-means cluster were identified using Homer (Supplementary Table 7) and are shown on the left. Motifs are annotated using the top annotation from Homer. b. Genome browser view of chromatin accessibility at the EMP1 locus in the indicated mCRC cell states. PE-GL linkages are shown and chromHMM enhancers47. c. GEA of genes in k-means clusters shown in Fig. 5a using KEGG pathways and MSigDB Hallmarks. d. Heatmap showing

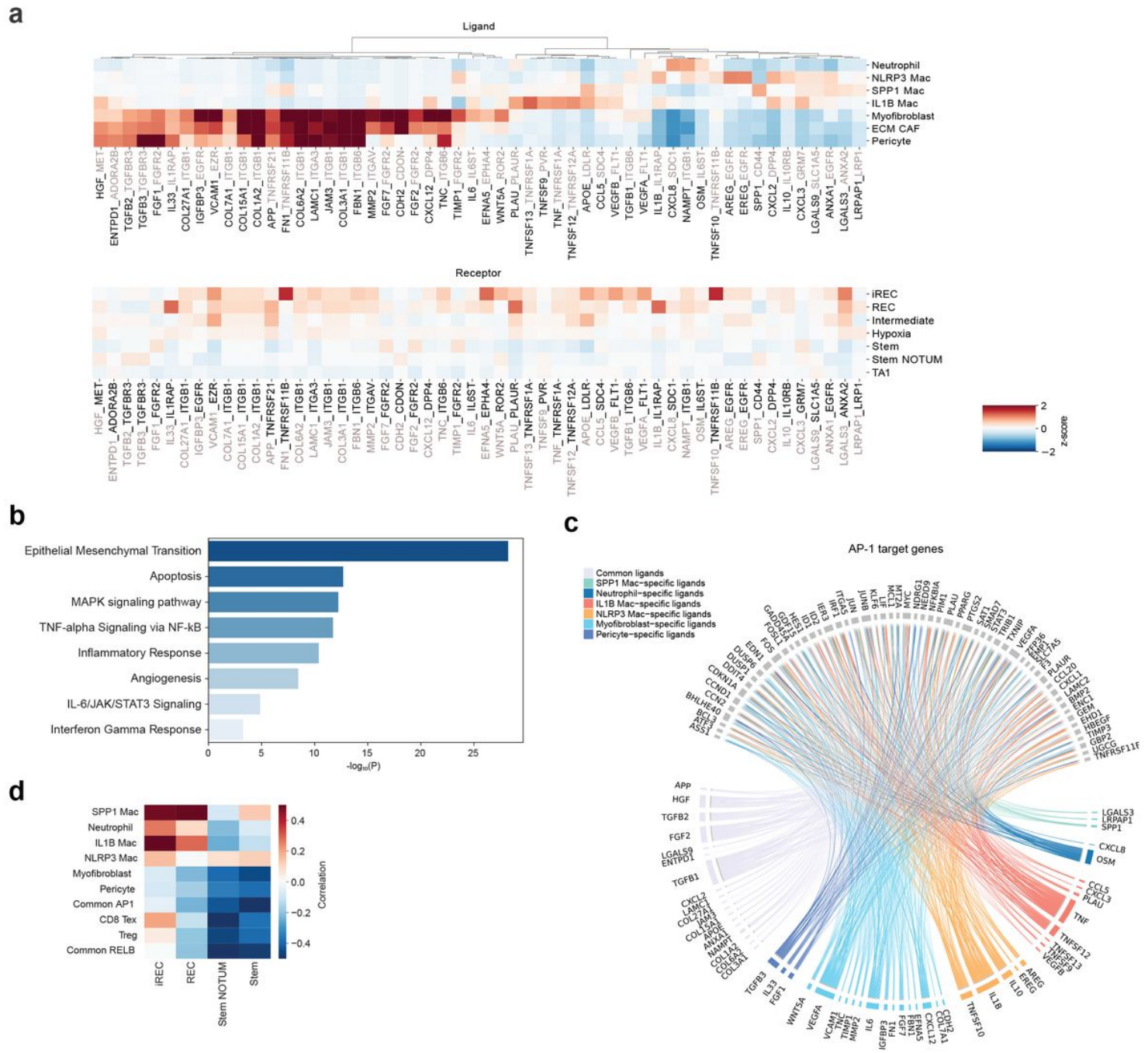
chromVAR motif deviation z-scores for TFs in mCRC cell states. Only statistically significant motifs are shown (Wilcoxon, FDR < 0.05). TFs are annotated based on the DNA binding domain (DBD)116. e. Accessibility of chromatin regions in the indicated SCENIC+ regulons. f. Z-scored mRNA expression of genes in the indicated SCENIC+ regulons across mCRC cell states. g. Z-scored mRNA expression levels of AP-1 target genes61 across mCRC cell states.




## Figure 5

Spatial niches of cancer cell states in liver metastasis.

a. Desmoplastic growth pattern: H&E staining of publicly available liver metastatic sample LM4, manual annotations and clustering annotations based on spatial gene expression. Sample LM4 captures the desmoplastic rim separating the tumour and the liver parenchyma. The dashed line in black denotes the desmoplastic rim. b. Replacement growth pattern: H&E staining of publicly available liver metastatic colorectal cancer sample P13, manual annotations and clustering annotations based on spatial gene expression. In the replacement growth pattern, tumour cells are in direct contact with hepatocytes. The dashed line in black denotes the tumour-liver border. c. H&E staining of liver metastatic colorectal cancer sample CRC11. d. Abundance of cancer cell states across spatial locations of three representative samples (LM4, P13, CRC11). Cell abundance (colour represents intensity, size represents the score) per spot assigned to each cell type and cancer cell state is estimated by cell2location. e. Estimated cell type abundances for distinct stromal subpopulations in LM4, P13 and CRC11. f. Spatial cellular neighbourhoods in LM4, P13 and CRC11. Spatial cellular neighbourhoods were deciphered by joint modelling of six Visium samples to infer common patterns across samples using SpatialDE2. g. Estimated cell type abundances for distinct immune subpopulations in the immunosuppressive niche in representative samples LM4, P13 and CRC11. h. Dotplot representing the average cell abundance (dot size and colour) for each cell state, per neighbourhood, and normalised between 0 and 1 per cell state. 0 and 5 are the cellular neighbourhoods containing iREC. Niche 7 is the cellular neighbourhood containing REC. Niche 4 is the cellular neighbourhood containing Stem NOTUM. Cellular neighbourhoods 1 and 9 denote the liver parenchyma, depending on the growth pattern. i. GEA of upregulated genes in the spatial cellular neighbourhoods.




## Figure 6

Spatially resolved cell-cell interactions in the cellular neighbourhood surrounding iRECs.

a. Heatmaps summarising the inferred spatial cell-cell interactions mediated by stromal and myeloid cells in the cellular neighbourhood containing iRECs with iRECs as the receiver, using CellPhoneDB and NicheNet. Specifically, we identified potential upstream ligand-receptor pairs which can induce the AP-1 regulon program in the neighbouring pro-metastatic phenotype. Z-score of the gene expression of selected potential ligands in each cell type of the cellular neighbourhood (top panel) and z-score of gene expression of corresponding receptors in cancer cell states (bottom panel). In both heatmaps, the x-axis denotes ligandreceptor interactions, with the ligand in bold and receptor in grey for sender cells (top

panel) and the ligand in grey and the receptor in bold for cancer cell states (bottom panel). b. GEA of ligands predicted to activate genes in the AP-1 regulon in iRECs. c. Circos plot depicting links between predicted ligands from stromal and myeloid cells and target genes of the AP-1 regulon, as inferred by NicheNet. Links denote the regulatory potential scores between ligands and target genes of the AP-1 regulon, as predicted by NicheNet. d. Correlation in TCGA bulk CRC RNA-seq data of the expression of the indicated pCRC cancer cell state signatures and ligands expressed in TME subpopulations that potentially drive the expression of AP-1 and RELB regulons in CRC malignant cells. Ligands are shown in Fig. 6a (AP-1) and Extended Data Fig. 9a (RELB). Common AP-1 and common RELB are ligands whose expression is shared in more than one TME subpopulation.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ExtendedDataFigures.pdf
- SupplementaryTables.xlsx